Detecting Food Safety Risks and Human Trafficking Using Interpretable Machine Learning Methods

by

Jessica H. Zhu

B.S. Operations Research and Chinese, U.S. Military Academy

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June, 2019

Author . . . .
Sloan School of Management
May 18, 2017

Certified by . . . .
Dr. Lin Li
Technical Staff, MIT Lincoln Laboratory
Thesis Superivisor

Certified by . . . .
Prof. Y. Karen Zheng
Associate Professor of Operations Management
Thesis Supervisor

Accepted by . . . .
Prof. Dimitris Bertsimas
Boeing Professor of Operations Research
Co-Director, Operations Research Center

THIS PAGE INTENTIONALLY LEFT BLANK

Detecting Food Safety Risks and Human Trafficking Using Interpretable Machine Learning Methods

by

Jessica H. Zhu

Submitted to the Sloan School of Management
on May 18, 2019 in partial fulfillment of the
requirements for the degree of
Master of Science in Operations Research

Abstract

Black box machine learning models have allowed researchers to design accurate models using large amounts of data, at the cost of interpretability. Model interpretability not only improves user buy-in, but in many cases provides users with important information. Especially in the case of the classification problems addressed in this thesis, the ideal model should not only provide accurate predictions, but should also inform users of how features affect the results.

My research goal is to solve real-world problems and compare how different classification models affect the outcomes and interpretability. To this end, this thesis is divided into two parts: food safety risk analysis and human trafficking detection. The first half analyzes the characteristics of supermarket suppliers in China that indicate a high risk of food safety violations. Contrary to expectations, supply chain dispersion, internal inspections, and quality certification systems are not found to be predictive of food safety risk in our data. The second half focuses on identifying human trafficking advertisements, specifically sex trafficking, hidden amongst online classified escort service advertisements. We propose a novel but interpretable keyword detection and modeling pipeline that is more accurate and actionable than current neural network approaches. The algorithms and applications presented in this thesis succeed in providing users with not just classifications but also the characteristics that indicate food safety risk and human trafficking.

Thesis Supervisor: Dr. Lin Li
Technical Staff, MIT Lincoln Laboratory

Thesis Supervisor: Prof. Y. Karen Zheng
Associate Professor of Operations Management

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Syngenta Foundation's Sourcing and Safety Management of Agricultural Products in China grant No. 6936419 and the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Syngenta Foundation nor the Under Secretary of Defense for Research and Engineering.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

First and foremost, I would like to thank my advisors, Professor Karen Zheng and Dr. Lin Li, for their guidance and support. I am especially thankful for their dedication to these two projects and flexibility in allowing me to pursue my research interests. I would not have been able to explore such a large breathe of operations research topics without them. They have been an immense influence on my growth here at MIT as a student and researcher.

I would also like to thank Lincoln Laboratories, the Operations Research Center, Group 52, and Mr. John Kuconis for the opportunity to pursue my interests in operations research. I would not be here without your belief in my abilities and worth. I feel especially fortunate that I have had the opportunity to work with Group 52 and especially the HDDN team. Their dedication to solving real-world problems have been an inspiration to me as a researcher. In addition, their valuable guidance and input throughout my two years were a significant reason for the success of the human trafficking project.

This experience would not have been the same without my friends. The friends I met along the way and in the ORC made MIT into a truly memorable experience. They have been amazing sounding boards throughout this chapter and I've learned so much from them. Thank you to them and my long-distance friends for helping me stay sane.

Finally, I would like to thank my parents and sisters for their love and support. Their encouragement throughout life has given me the strength to pursue this journey.

Contents

List of Figures

List of Tables

Chapter 1

Thesis Overview

This thesis covers two disparate projects that both use interpretable machine learning methods to analyze large amounts of data. Chapter 2 discusses our research on the detection of food safety risks in Chinese suppliers. This project analyzes a group of suppliers from a leading Chinese supermarket. The chapter discusses the background on food safety research, data sources used, hypotheses, machine learning methods applied, predictive results, and implications to our collaborator and food safety in China. We ultimately discover that certification systems and other supply chain characteristics are inconsequential to reducing a supplier's likelihood of failure in national food safety exams in our data.

Chapter 3 follows a similar structure. It discusses our work on detecting human trafficking advertisements from online escort service ads. It discusses previous human trafficking detection research, data used, pipelines tested, predictive results, applications to organization detection, and implications of our model. In this project, we develop an unsupervised keyword detection pipeline that can be used to train supervised models that accurately identify suspected human trafficking advertisements.

Chapter 2

Predicting Food Safety Violations

2.1 Section Overview

2.1.1 Motivation

More than 500,000 food safety violations were uncovered from over 15 million inspections in China in 2016 [2]. Despite sweeping reforms to Chinese food safety standards and inspections the following year, the scandals have continued. In March 2019, over one million pounds of pork were seized by U.S. border agents in New York over suspicions of swine flu contamination [3]. China has suffered numerous food safety scandals across all products, dairy, meat, vegetable, oils...etc, ever since scrutiny increased after joining the World Trade Organization in 2001.

Poor food safety regulation in China is not just a national issue; it affects the international community. China is the U.S.' fourth largest supplier of agricultural imports. In 2017, the U.S. imported $4.5 billion worth of agricultural products from China [4]. Yet despite the high costs of undetected food safety violations, very few of these agricultural imports are actually inspected. In 2015, only 2.2% of all imported seafood were examined [5]. However, food safety violations in the United States have caused 50 million people, or one in six people, to fall ill and three thousand to die annually [6]. In addition, the costs of food safety recalls on average were $10 million dollars in direct costs to the company [7]. Given these gaps in government level food safety inspections, a push for quality and traceability certifications has taken hold to mitigate food safety risks starting at the beginning of the supply chain.

Supplier quality and traceability certifications increase consumer confidence in product safety, especially as food supply chains become increasingly complex in a global economy. However, significant start-up investments are required in order to implement mechanisms that comply with standards. In the case of traceability systems, suppliers must also invest in technology for testing, recording, storing, and transferring product information from its beginning at a farm to its end of life with the consumer. In addition, suppliers often need additional training in order to learn how to successfully implement and maintain food safety standards. These investments are expected to be cost-effective solutions for reducing food safety risks.

Although a significant amount of research indicates that investments in food safety certifications are generally cost-effective and beneficial to suppliers and retailers in the long run [8], there are few studies to date on whether these measures actually improve food safety. Using data provided by a leader in the grocery industry and quality management systems in China that has developed its own rigorous supplier quality and traceability certification system, we analyze how effective these measures are in improving food safety.

2.1.2 Objective

Our research addresses the following questions:

1. How effective are quality certification systems in improving food safety?
2. What supplier characteristics are predictors of food safety risks?

2.1.3 Approach

We apply a data-driven, analytical approach to investigate the effectiveness of supplier certification systems and the characteristics of suppliers with high risks of food safety. Combining data from the Chinese Food and Drug Administration (CFDA) with supply chain data from a leading supermarket in China, we model the likelihood of a supplier being at risk of food safety failures. We characterize each supplier as a vector of features describing its supply chain composition: the number of farms, factories, products, etc. Most importantly, we factor in the results of our collaborator's internal supplier evaluations: grades and certification status. We then use interpretable classification modeling techniques to understand the characteristics of high risk suppliers and if the company's internal quality and traceability certification system improves food safety.

2.1.4 Contributions

Our results show that the company's internal certification systems may not be as effective in ensuring food safety as consumers and retailers alike expect. Our collaborator's certification system does not reduce the risk of a supplier failing food safety tests. We do observe that it does reduce the chance of a supplier being sampled by the CFDA in the market, potentially because if the supplier failed the company's internal internal certification, then the company would source fewer products from this supplier. Furthermore, supply chain characteristics, such as distance between a supplier's farms and factories, also are not found to be significant influencers of food safety risk. We did find that suppliers located in regions with stronger governance failed CFDA tests more frequently. This could imply that governments in regions with weaker governance tend to identify fewer problems due to lax control. These results suggest that further changes in the CFDA's governance and our collaborator's quality management system are needed to truly improve food safety.

2.2 Background

2.2.1 Trends in Food Safety Quality Management

Over the past two decades, numerous international, national, and local level legislation have been written recommending, and even requiring, some degree of traceability and quality assurance. Simultaneously, many international third party quality assurance systems have been developed, to include benchmarks by Global Food Safety Initiative (GFSI), Global Good Agricultural Practice (GlobalG.A.P), and International Food Standar (IFS). In fact, quality assurance systems have become standard business practice for food suppliers in many regions, like the U.K. [9], and traceability systems are rapidly becoming standard as well. For example, in China, beginning in 2001, Shanghai began requesting that vendors provide information on their products [10]. In 2002, Beijing also began requiring a low level of traceability information for food products [10]. In 2009, China took a significant national step to improve food safety by passing their Food Safety Law. More recently in 2015, a sweeping revision of this law was passed to require a state-owned food traceability system. However, changes are slow and it was not until 2017 that implementing regulations were passed [10]. It is yet unclear if these changes have in fact improved the safety of Chinese food products.

Nevertheless, these systems are used to assure customers that products and processes are consistently delivered [9]. They can take the forms of privatized international standards, like those mentioned above, government regulations, like in China, or proprietary systems that are often maintained by large retail food chains [9]. Suppliers that meet the standards are often then awarded with certification labels that inform customers that their products are of the expected quality (e.g. chemical-free, traceable, or, most importantly, safe).

Quality certification systems have become increasingly popular in the global economy. They ensure that retailers and suppliers comply with best practices and food safety standards via education and inspections. In addition, suppliers have an incentive to participate because quality certification systems are expected to help improve market access, improve product quality, and even potentially improve operational efficiency [9]. However, suppliers may incur high sunk costs to adopt the system and often times also pay inspection fees in order to become certified [8]. As a result, large suppliers often adopt the standards and gain certification more easily, while also benefiting more from the economies of scale than small and medium sized suppliers [8].

The documentation of production processes required by quality assurance systems often corresponds to traceability certifications. In 1998, in conjunction with the growth in quality certification systems, new attention was drawn to food traceability systems as a method to ensure food safety [11]. Traceability, as defined by Moe, is "the ability to trace the history, application or location of an entity, by means of recorded identifications" and is essential to quality management [11].

The purpose of food traceability systems are primarily three fold: improve food quality, improve recall efficiency, and offer a business advantage. Traceability systems allow stakeholders to identify the life history of a product: where and how it was farmed, transported, and processed [12]. This history is not sufficient in reducing food safety risks. Rather, the information must be used in conjunction with a quality assurance system to identify poor practices and prevent unsafe foods from entering the market [13]. In the event of a food recall, traceability systems also facilitate the identification of products of concern and more importantly, they allow companies to find the origin of the problem and resolve it at the source [12][13]. These characteristics of traceability systems are particularly important in a country, like China, with significant problems in food safety such that in 2007 they had twice as many food recalls as the United

States [14].

Retailers and suppliers have an incentive to implement quality and traceability certification systems despite high initial investment because of the prospective business advantage. These systems are expected to reduce transaction costs between buyers and sellers through the implementation of best practices [9]. Regattieri et al., posits that an effective and efficient traceability system can "significantly reduce operating costs and can increase productivity" [15]. Various studies have also found that certain consumers are willing to pay a premium for quality assured foods. A survey of Chinese consumers found that they were willing to pay a premium for product traceability, although they would prefer governmental or private quality assurance certification [16]. These results were corroborated in another survey published in 2010 of citizens in Jiangsu, China that found that 32% of respondents opted for certifiably traceable foods and 68% of those consumers were willing to pay for traceability [17]. Therefore, improving food safety via quality and traceability inspection systems is expected to be economically beneficial for retailers and suppliers.

Given these benefits, numerous quality assurance frameworks and related technology have been developed to increase food safety over the years. Roth et al. proposes a six part framework, the "six Ts": traceability, transparency, testability, time, trust, and training [14]. Deloitte recommends a similar framework that is composed of initiating business with formal documentation, due diligence and selection of suppliers, contracting and on-boarding of food safety specifications, ongoing monitoring, and formal termination and off-boarding [18]. Essentially, these and other frameworks all recommend that companies have clear records of their suppliers' activities, conduct training to ensure suppliers comply with company standards, and repeatedly verify that suppliers are meeting these standards. They are achieved through traceability, education, and inspections, respectively.

These efforts at improving food safety must first overcome significant challenges, especially in China. First, Chinese suppliers have less financial incentive to participate in certification systems. 90% of Chinese farms are smaller than 2.5 acres [19] and as previously discussed, it is more difficult and less financially beneficial for small suppliers to implement and maintain quality and traceability assurance systems. Second, despite the expected long term financial benefits, it is difficult to convince suppliers to shift practices and abide by new standards and traceability systems. China's market renders systems without short term positive impact unheeded [14]. Instead, suppliers bend to economic pressure to use cost-cutting measures to ensure profit, potentially resulting in noncompliance and food safety violations. Finally, local administrators, until recently, have been disincentivized to enforce food safety compliance. As discussed by Roth et al.,"if local governments close all the companies that violate food safety regulations, a lot of workers will lose their jobs."[14]. As a result, food inspections might not be as rigorous or accurate as needed. Given these challenges, the presence of inspection and quality certification systems, even in tandem, do not guarantee a reduction in food safety violations.

These systems will only be successful in improving safety if they are implemented in conjunction with a shift in attitude through training and incentive structures [20]. A low cost traceability and quality assurance system would make certification accessible to China's numerous small suppliers. China's dispersed small enterprises and high worker turnover also require a shift in individual behaviors. This can be achieved through facilitating collaboration with regulations, training, and incentives so that group norms converge upon an industry standard [20][14]. Likewise, traceability systems paired with quality assurance inspections that allow failure costs to be allocated to the sourcing producer can offer a strong financial incentive to motivate suppliers to implement and follow product quality standards [13]. Many of these characteristics are present in our collaborator's quality management system.

2.2.2 Collaborator Quality Management System

It is unsurprising that given the expected benefits of food traceability and other supply chain quality management systems, top retail stores have implemented their own systems. Our collaborator has implemented a state-of-the-art certification and traceability system on a subset of their grocery suppliers. These suppliers are provided training on best farming and processing practices. Their products are then labeled and certified as traceable if the supplier is in accordance with the company's internal quality standards. The company has created these internal quality standards by adapting well-established international ones, such as Good Agricultural Practices (GAP) and Good Manufacturing Practices (GMP), and leveraging their own experiences working with the vendors. Customers can search online or simply scan a product package's QR code to learn more about the product sources and processing. Our collaborator provides customers with information on the distributor company, packaging date, factories involved, inspection reports, additional certifications, transportation processes, and more.

In order to become certifiably traceable, suppliers must undergo rigorous quality and traceability inspections and training sessions every six months at all levels of the supply chains. These inspections and training sessions are in addition to the annual inspections non-traceable suppliers already undergo. A supplier who receives lower than a "B" score is considered to have failed the inspection and must undergo additional testing to regain certification. Weak points are also discussed with the supplier and corrective actions are recommended after each inspection. We leverage the results of these inspections and other supply chain information provided by our collaborator to analyze the impact that their certification and traceability system has had on the food safety risks of their suppliers.

2.2.3 Food Safety Risk Identification

The food industry has high hopes that traceability and quality management systems will have a dramatic impact on improving food safety. A survey on companies who have implemented and certified a new quality management system found that most believed there was an improvement in the food supply chain, including simplification of quality control and reduction of errors [21]. In a case study of a cheese company that implemented a traceability system, they were found to have successfully used the system to product the authenticity of their brand[15]. It required passing along a slight increase in product cost to customers but because of the system, customers were able to check the origin and production process of their purchase. In addition, the manufacturer was able to check production progress and rapidly implement recall strategies as needed. However, the researchers did not analyze the effectiveness of this safety assurance in improving safety.

To the best of our knowledge, few studies have used data-driven modeling techniques to predict food safety risks and evaluate quality management systems. This is likely due to the unavailability of data. Only a handful of studies exist to date. A 2015 study on dairy farm's cattle welfare (which is often linked to food safety) used a dataset of only 24 dairy farms to train a decision tree model that classifies the welfare of over six thousand dairy farms [22]. Although they created synthetic data using SMOTE (further discussed in section 1.4), to correct for the unbalanced data, it is unlikely that there is sufficient data to verify the model accuracy. With more success, another study evaluated food safety in dairy products with 86% accuracy over a test set of 6000 samples [23]. They had significantly more data, although it was also enhanced with synthetic data using SMOTE, and were able to apply neural networks on a balanced data set to predict the presence of contaminants. Although the features used are not specifically explained in this study, there is no indication that supply chain information is included. Even if it was, due to the neural network approach, no conclusions can be drawn on the characteristics that indicate food safety violations. A recent study did use supply chain features to predict food safety risks at the manufacturer level [24]. They analyzed data on 900 companies from publicly available Chinese websites involved in food exports. This data included information on the number and output volumes of upstream suppliers. Using Heckman's sample selection model, they found that high supply chain dispersion and weak local governance are predictors of higher risk manufacturers, and manufacturers located in regions with weak governance are sampled less [24]. This analysis does not include characteristics on traceability or internal quality management.

2.3 Data Overview

The data we used in our analysis are supply chain and inspection data from our collaborator, location based data from [24], and food product sampling data published by the CFDA.[1] We provide a detailed description of this data in this chapter.

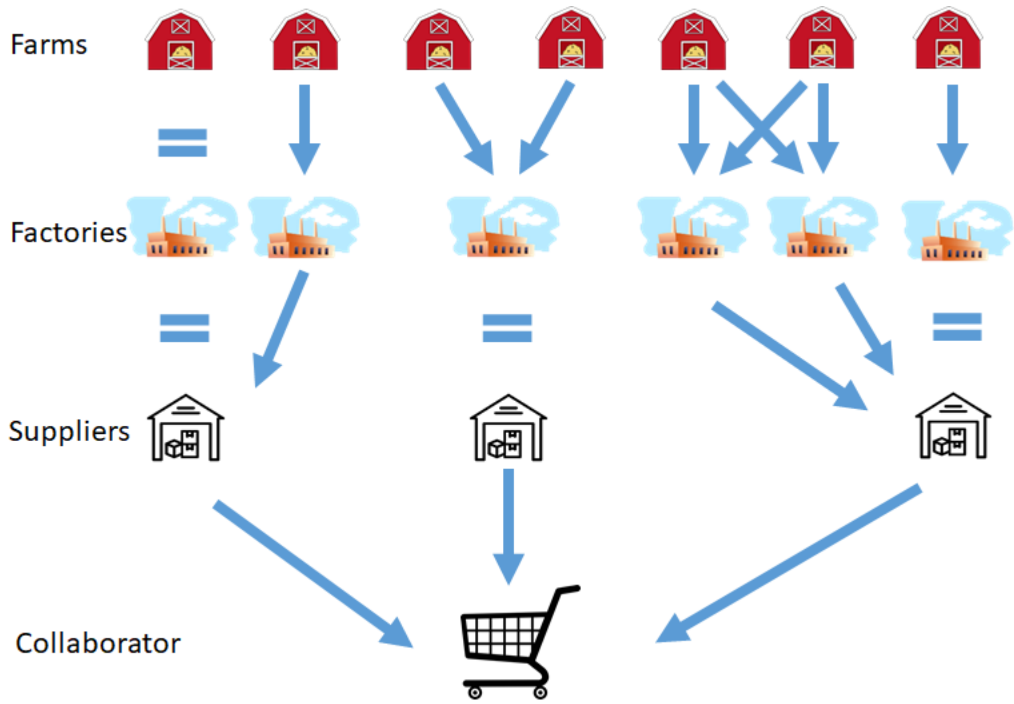
2.3.1 Collaborator Data

We collaborated with a top Chinese grocery retailer to collect data on their supply chain and internal inspections. Resulting from the small to medium-sized farms characteristic of China, their supply chain is quite complex with many small farms feeding into multiple factories and suppliers. A simplified visualization is shown in Figure 2.1. The far left chain shows the simplest chain, where a farm (or distributor) can also process its own product as a factory (or processor), and be the final supplier to the retailer, our collaborator. In addition, another chain of farms and factories can sell to that same supplier. Likewise, multiple farms can supply to the same factory, which may also be the direct supplier to the retailer. Finally, farms can supply to multiple factories, where one may also function as the direct supplier to the retailer. These are just a few examples of the multitude of variations found in our collaborator's supply chain. Our

[1] The CFDA data was made available via a project by Retsif Levi, Qiao Liang, Nicholas Renegar, Qi Yang, Run Zhou, and Weihua Zhou. Combining Multiple Information Sources for Informing Food Safety Regulation in China. Working Paper. February 2019.

analysis focuses on the supplier level, which includes any suppliers that also farm or process their own products.

Figure 2.1: Supply Chain From Farms, to Factories, to Suppliers, to the Retailer



Along with an annual internal inspection on all suppliers, our collaborator also conducts additional inspections on the farms and factories of a subset of self selected suppliers to certify their products as high quality and traceable. Henceforth, we will refer to these inspections as regular or certification inspections. A complete record within this data provides information on each supplier's name, location, products, farms, factories, and internal inspection results. We do not have information on why a supplier may have failed an internal inspection, but we do know when and which type of inspection it failed. This data includes inspections dating back to 2011 and up to March 2018. For the sake of completeness and consistency, we focus on suppliers with data from 2014 and after. From this data we can observe when suppliers are certified or not by the retailer.

For our study we focus on suppliers that were inspected for meat, aquatic, vegetable, fruit, tea, egg, and nut products by the CFDA. These are some of the most common product categories in the CFDA data with a non trivial number of failures. We apply a neural network based food categorization model designed by another MIT research team [25] to map the product names in the supply chain data into product types. We find that these suppliers may also sell products to our collaborator outside of our main categories of interest. We annotate this characteristic by including an "other" product category. This results in a dataset of over three thousand suppliers.

2.3.2 CFDA Data

Our dependent variable, level of food safety risk, is derived from the results of CFDA food safety tests. The CFDA periodically samples food products from the market and tests them against quality standards [26]. Since 2016, these test results have been published online [26]. These results were collected by an MIT research team from publicly available Chinese government websites [25]. The version used in our analysis covers all published CFDA test results as of October 2018 from the state-level CFDA, all 34 province-level or municipality CFDAs, and 335 prefecture-level CFDAs. It includes records dating back to 2014. This research effort has resulted in a dataset describing over two million unique tests.

Each data record describes the name and type of the tested product (e.g. vegetable or aquatic), manufacturer and sampled location, the production date, the website announcement date, and the test results. If a failure occurred, the test result also includes the cause of failure. From the data, we found that although our collaborator has a lower failure rate than the average across all CFDA tests collected, it has a higher than average failure rate compared to its major competitors, as depicted in Figure 2.2. On the other hand, although the test frequency of certain product types, like meat products, are significantly higher for the collaborator, the failure rates are not. This is visualized in Figure 2.3 and 2.4. The failure rates across categories have distinct differences. Despite our collaborator's rigorous internal testing, its food safety test performance does not appear to be consistently better than either its competitors' or the average supplier's performance.

Figure 2.2: Comparison of CFDA Failures in Chinese Supermarket Chains

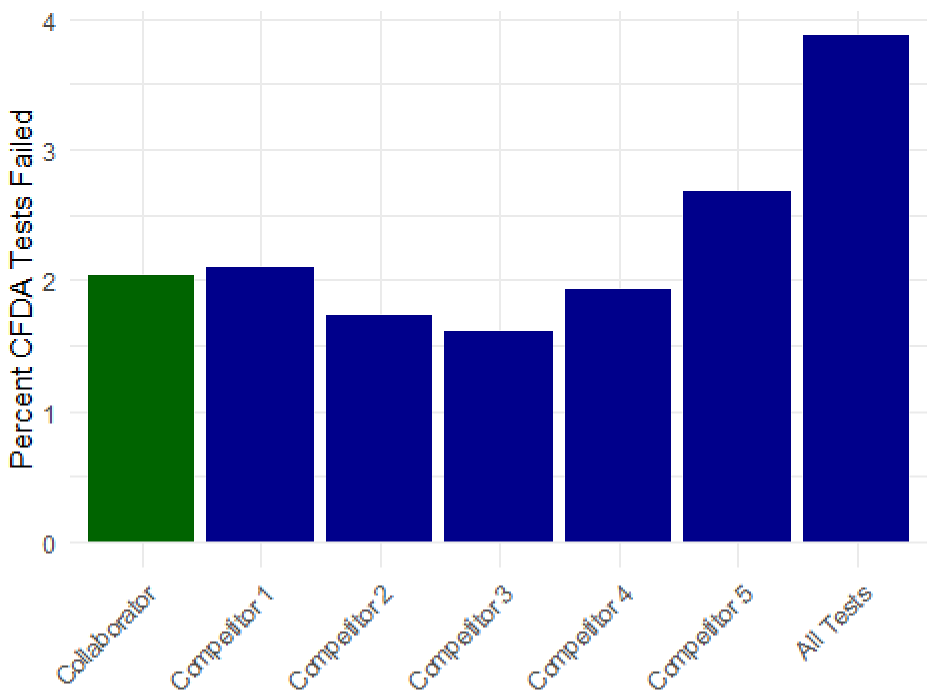


Figure 2.3: Comparison of CFDA Test Frequency Across Product Categories

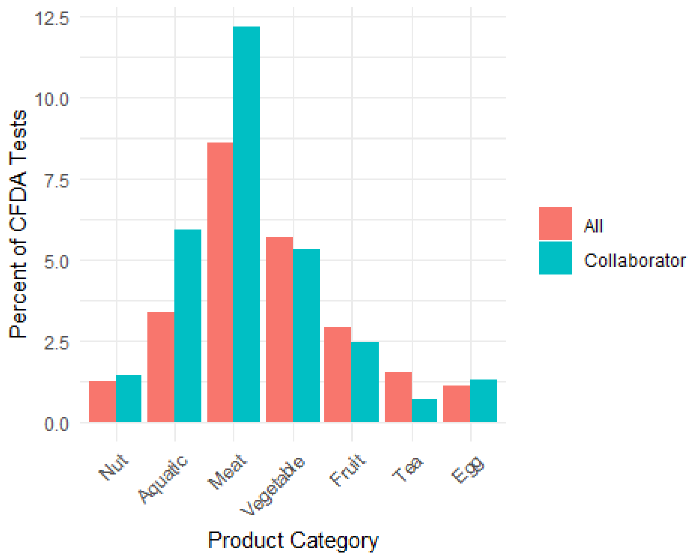
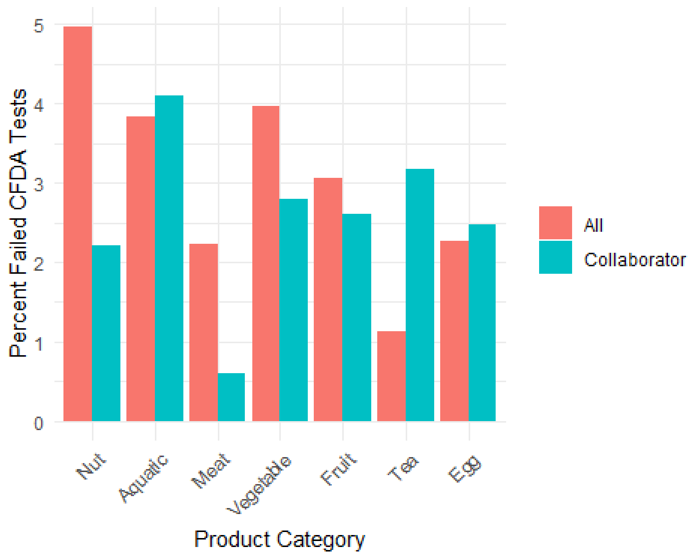


Figure 2.4: Comparison of CFDA Test Failures Across Product Categories



2.3.3 Location Based Data

In addition to the CFDA data, we also analyze data on each prefecture's demographics (GDP per capita and population) and governance quality. Quality of governance is scored by a misconduct ranking on a 0 to 5 scale and a 4 dimensional transparency score as introduced in [24].

The misconduct ranking is calculated by identifying the number of misconduct cases reported between 2003 and 2015 and scoring each prefecture based on the depth of the misconduct in the higher-ranks of governance. A prefecture is ranked 5 if its mayors, party secretaries, and subordinates were all engaged in misconduct cases, 1 if only subordinates had, and 0 if no cases were reported. For suppliers whose location is only given at the provincial level, we take the average of the misconduct rankings of other prefectures in its province. Figure 2.5 presents the misconduct ranking of the suppliers analyzed.

Figure 2.5: Misconduct Rankings Across China

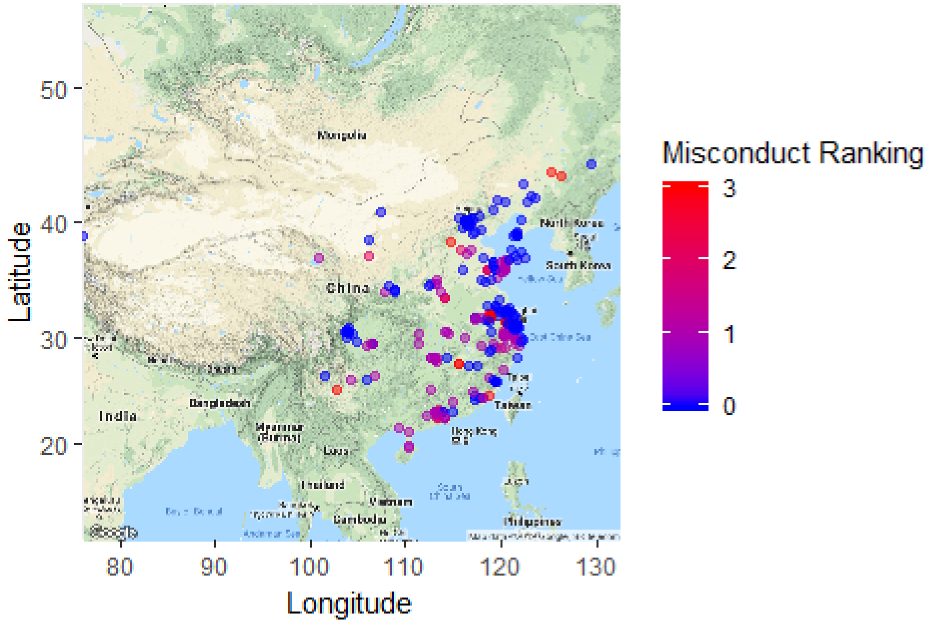
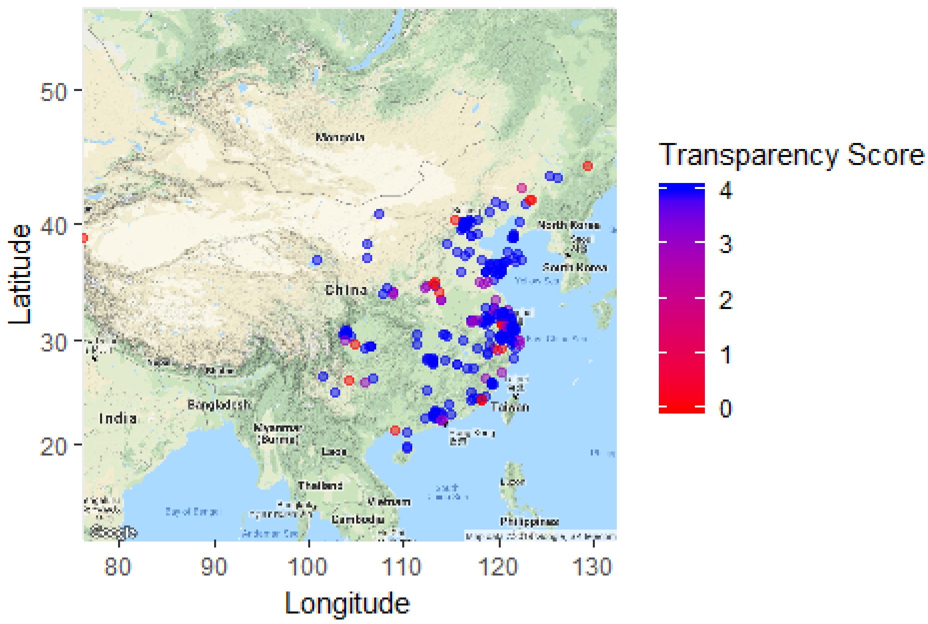


Figure 2.6: Transparency Rankings Across China



Transparency is measured by the presence of various components on the government agency website in question, in this study, the CFDA. These components indicate to what extent each prefecture discloses and solicits food safety information. They examine if a supplier black list, complaint forms, test results, and/or food safety knowledge are published. A higher transparency score indicates that the government is more transparent and hence stronger in food safety governance. Figure 2.6 presents the transparency score of all the suppliers analyzed.

2.3.4 Final Dataset

To compile our final dataset, we created a cross-sectional data structure for our analysis due to the low number of failures per supplier in the CFDA data. Each observation in the data corresponds to a unique supplier. We searched the CFDA data and found the tests associated with each of our collaborator's suppliers. We included tests from all sampled locations including when the supplier's products were sampled from our collaborator's or other retailers' stores, or on site from the supplier's farms/factories. In addition we searched the CFDA data for entries where our collaborator is listed as a sampled location and extracted all corresponding suppliers who were not included in the supply chain data shared by our collaborator. In this dataset, we define each supplier to include any farm or factory that directly provides products to our collaborator. For the farms and factories that are separately sampled by the CFDA, we linked their CFDA test results with their associated supplier per the supply chain data. This is to ensure that no CFDA test results are considered multiple times.

By comparing the announcement date of the CFDA tests for each supplier to our collaborator's internal inspection data, we labeled whether or not a supplier was certified traceable at the time of the CFDA test. For each supplier, we also computed the average grade of the internal inspections the year before the CFDA tests. In addition, we calculated the average distance between a given supplier and its associated farms and factories. Finally, we used the location of the supplier to match it to the demographic and governance data previously described. We exclude additional transparency measurements due to collinearity between government agencies.

Our final dataset includes the following information for each supplier:

- Number of times it has been tested by the CFDA
- Number of times it has failed a CFDA test
- Percent of CFDA tests completed/failed while it was labeled a certified traceable supplier
- Number of products it supplies to our collaborator
- Number of farms under the supplier working with our collaborator
- Number of factories under the supplier working with our collaborator
- Number of different food categories (e.g. fruit, vegetable, meat, tea) it supplies
- What food categories it supplies (meat, aquatic products, fruit, vegetable, nut, tea, egg, or other)
- Average distance between the supplier and its network of farms and factories
- Average grade of both regular and certification inspections the year before each CFDA test

- Total number of regular or certification inspections conducted the year before each CFDA test
- Whether it was ever certified
- Average GDP and GDP per capita of the supplier's location
- Average population of the supplier's location
- Misconduct ranking of the supplier's location
- Transparency score of the prefecture's CFDA website
- Length of time it was a regular and/or a certified supplier with our collaborator (Age)

This results in 25 explanatory variables that can be used in our analysis. We have 679 suppliers with internal test results who supply products in the categories of interest. Among them, 313 suppliers also have CFDA tests. These 313 suppliers have a total of 11,485 CFDA tests. We conduct our predictive analysis of food safety risk on these 313 suppliers and use the additional 366 suppliers to factor in potential sampling bias from the CFDA.

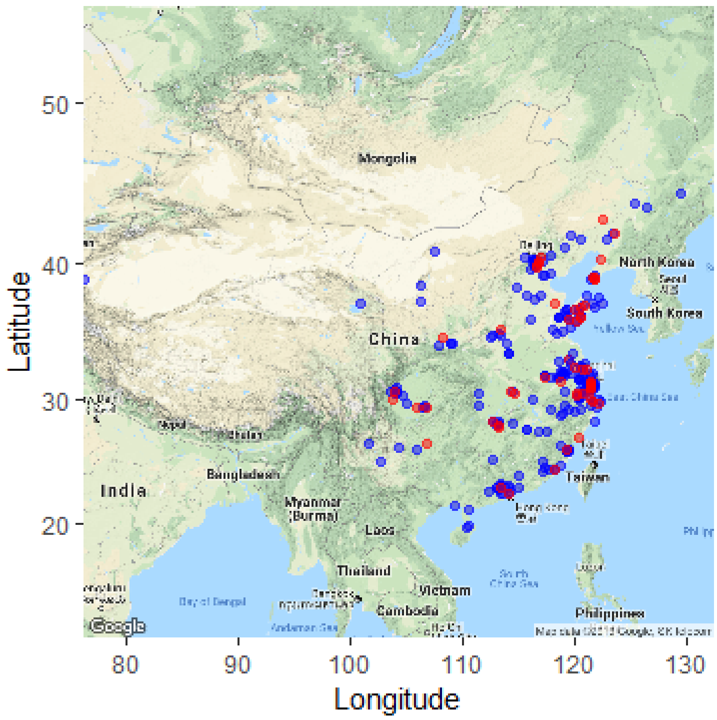
Our dependent variable is the risk level of a supplier having a food safety violation, given that it has been sampled by the CFDA. It is a binary bin that represents if a supplier has failed a CFDA test: 1 if it has had one or more failures and 0 otherwise. Henceforth, we will refer to suppliers with a CFDA failure as high-risk suppliers and suppliers that have passed all their CFDA tests as low-risk suppliers. Out of the 313 suppliers with CFDA tests, we find that suppliers have an average failure rate of 2.1% with 18.5% suppliers having at least one failure. Figure 2.7 maps our collaborator's suppliers and colors it by their risk level. It shows that food safety failures are not more common in one region over another.

It is important to note that as with most real world problems, due to inconsistent internal data gathering, about .5% of the total data is missing and 18% of the suppliers with CFDA tests are missing some information. In addition, our data may be biased because we only have data from our collaborator's perspective but are including CFDA tests from all retailers. Suppliers may provide different products to other retailers or have additional farms and factories not inspected by our collaborator. As a result, the farms, factories, and products that our collaborator has tested internally might not correspond with what the CFDA tested. In addition, the product quantities, age, number of farms, and number of factories are only proxies for the true values. These are only the numbers pertinent to our collaborator and do not reflect the total products, age, farms, or factories a supplier has. For instance, a supplier could have existed for longer than its age under our collaborator. These are also static quantities and do not reflect growth or reductions over time. The data reflects the reported status as of March 2018. This data demonstrates the complexities in evaluating food safety systems and conducting risk analysis.

2.4 Hypotheses and Expectations

The primary purpose of this study is to analyze the impact of traceability and supply chain characteristics on food safety. To this end, we present our hypotheses and preliminary data

Figure 2.7: Suppliers Across China



Red points designate high risk suppliers; blue points designate low risk suppliers.

analysis in this chapter.

2.4.1 Hypothesis

We present the following hypotheses used to structure our analysis.

Hypothesis 1:

Certified suppliers have a lower risk of failing CFDA tests because potential risk points should have been identified and remedied from rigorous internal inspection processes.

We expect a lower failure rate amongst certified suppliers. Our collaborator's certification system is in line with the best practices recommended by traceability and quality management literature and developed based on well-established international standards such as GAP and GMP. Suppliers who are certified as traceable have more rigorous inspections and higher quality requirements than the regular suppliers. Although these are nonrandom inspections, the rigorous inspections should motivate the certified suppliers to develop better quality management processes than the regular suppliers. Note that these suppliers could have been tested on products from farms or factories whose production process have not been certified by our collaborator and are supplied to other retailers. This may cause a bias so that the reduction in failure rate due to our collaborator's certification system is smaller than what would otherwise be expected. However, we hypothesize that the processes and management practices resulted from our collaborator's certification system will benefit all products produced by the certified suppliers.

Hypothesis 2

Suppliers with more internal inspection failures have higher failure rates in the CFDA tests.

We hypothesize that in general, internal failures correlate to more CFDA failures. Quality management and traceability systems are expected to be able to identify and mitigate food safety risks. Internal inspection failures point to the presence of risk factors that may lead to food safety violations. Since other retailers are not privy to our collaborator's internal inspection results, a supplier can continue to sell their products to other retailers after failing a regular or certification inspection. There is a higher likelihood that the CFDA also identifies food safety violations in these products because potential failure points have already been identified. Therefore, if the regular and certification inspections are effective, then the failures should be predictive of CFDA test failures.

Hypothesis 3

Suppliers with a more dispersed supply chain have a higher failure rate in the CFDA tests.

Motivated by the results in [24], our final hypothesis is that supply chain dispersion will have a negative effect on food safety; greater dispersion is associated with higher risks. We measure dispersion in a number of dimensions: the average distance between a supplier and its farms/factories, the number of farms/factories a supplier works with, and the variety of products a supplier supplies. The longer the distance, and the more farms/factories involved, the more likely contamination and hazards may occur along the supply chain. Similarly, if a supplier works with a large variety of products, it is likely less centralized and has more potential points of failures in its supply chain. On the other hand, the opposite effect could occur if the suppliers with more dispersion are simply larger, more established companies. We account for this factor by controlling for the product quantity supplied to our collaborator.

Controls

In order to test our hypotheses, we control for a number of additional features. First, we control for product types. As we show in the following section, there are differences in sampling and failure rates across categories. This is likely due to the varying levels of concerns and resulting emphasis the CFDA puts on certain products, like meat and eggs. We also control for demographic information using GDP per capita and population size. Suppliers in prefectures with high GDP and population may be tested more because of the greater affluence in that area. In addition, following results in [24], we capture the strength of governance in the prefecture where a supplier is located, measured by the prefecture's misconduct ranking and the transparency score of the prefecture's CFDA website. Finally, we control for the age and size of the supplier and the number of CFDA tests it has received. We expect that suppliers with more CFDA tests are more likely to have at least one failure occur. Likewise, we expect older suppliers and suppliers with higher product quantities to be more likely to fail because they are also more likely to have been sampled multiple times.

Preliminary Analysis

We provide a general overview of the features relevant to our predictive analysis in this section.

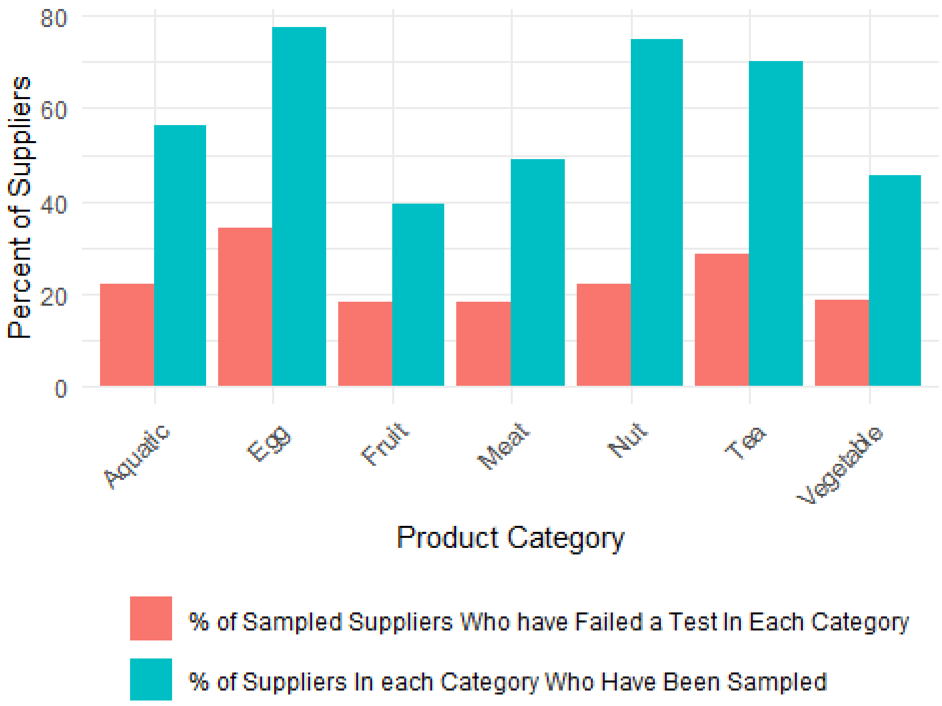
Many of our collaborator's suppliers in the categories of interest supply a variety of product types. Figure 2.8 presents what fraction of the suppliers sampled are in each product category and among those with failures, the fraction belonging to each category (i.e. number of suppliers in category X out of total number of suppliers who have been sampled or failed a CFDA test). Our data shows that most of the suppliers analyzed are meat, vegetable, aquatic, and/or fruit product suppliers. The failures are also equally distributed across all supplier categories (i.e. categories with more suppliers sampled also have more failures). However, egg suppliers are an anomaly. They make up a disproportionate number of the suppliers with failures given their small sample size.

Figure 2.8: Percent of Suppliers in Each Category

In addition, one can see in Figure 2.9 that sampling rates of suppliers vary across categories. The average sampling rate is 58.9%, but there is a standard deviation of 15.2%. Out of the categories examined, only our collaborator's fruit and vegetable suppliers have less than 50% of the categories' suppliers sampled. However, failure rates (i.e. number of suppliers who failed in a category out of total number of suppliers in that category) are more similar across categories, with a mean of 23.2% and standard deviation of 6.1%. Only egg and tea suppliers have more than a 25% CFDA failure rate.

The features we extract from the available data mostly do not have a statistically significant relationship to our dependent variable, that a supplier has a high risk of food safety violations. However, there are a few variables that are significantly different between the suppliers that have been sampled by the CFDA and those that have not been sampled (679 suppliers in total), and between the suppliers with and without failures (conditional on being sampled; 313 suppliers

Figure 2.9: Supplier Tests and Failures Out of Each Category



in total). Statistics for particular variables are summarized in Tables 2.1 and 2.2. Variables not included are not significantly different between the groups based on either chi-squared tests or t-tests with a p-value of less than .1. We include a few variables that may be of interest to readers but are not significant.

Between the suppliers with and without CFDA tests (Table 2.1), transparency scores, number of egg suppliers, product quantity, average distance between farms and factories, internal inspection grades, and demographic data are significantly different. Unsurprisingly, a supplier that has been sampled is more likely to come from a more transparent location with a larger population, and higher GDP per capita. This bias may result from the fact that more prosperous locations have better governance and thus more tests and/or more well published tests. Intuition and our data also supports that suppliers with CFDA tests have significantly higher product quantities. It is interesting to note that suppliers that are not sampled have higher internal failure rates for both regular and traceability tests. It also appears that the CFDA tests egg suppliers, which have a history of fake egg scandals, more rigorously than other categories. At the same times, it does not test categories like meat, which is also rife with scandal, as often, in proportion to the number of suppliers that exist in each category. We further explore the potential biases more in our Heckman selection model discussed in Section 2.6.3.

Only five out of our twenty-five features have a significant difference between the suppliers with and without CFDA failures (given they have at least one CFDA test). The misconduct ranking, number of egg suppliers, number of certified suppliers, supplier age, and number of CFDA tests are significantly different. As hypothesized, the suppliers with CFDA failures have more CFDA tests and have been a supplier with our collaborator longer than those without. It is interesting to note that the misconduct rankings between these two groups are significantly

Table 2.1: Statistical Significance of Difference Between Sampled and Non-Sampled Suppliers

Feature	Average: Sampled	Average: Not Sampled	P-value
Transparency Score	3.47	3.19	<.001 ***
Is an Egg Supplier	12.1%	3.00%	<.001***
Distance To Farms/Factories	.573	1.44	<.001***
Percent Regular Inspections Failed	14.3%	36.0%	<.001***
Percent Certification Inspections Failed	1.9%	61%	<.001***
Prefecture's GDP per capita	44800	32400	<.001***
Product Quantity	6.02	3.99	.037 **
Prefecture's Population	546	483	.019**
Number of Factories	.936	.675	.133
Age as Collaborator's Supplier	3.07	2.57	.165
Misconduct Ranking(0-5)	.601	.704	.474
Certified At Least Once	40.3%	51.1%	.909

***: $p < .01$; **: $p < .05$; *: $p < .10$
p values are derived from chi-squared tests or two sided t tests among the 679 suppliers

different. Suppliers without failures are in locations with more misconduct, on average. Egg suppliers are also more prevalent in the population of suppliers with failures. Contrary to our expectation, a larger fraction of suppliers with failures were certified as traceable during at least one CFDA test, than the suppliers who passed all the tests. Internal inspection grades do not differ significantly between the two groups. Outside the scope of Table 2.2, we also did not observe a significant difference between failure rates when a supplier is certified versus when it is not.

We additionally investigate the differences between the suppliers who become certified traceable and the ones that do not pass the certified traceable inspections. We compare the characteristics of the suppliers who pass the inspection and become certifiably traceable (26 suppliers) versus those that attempt but do not pass the inspection (3 suppliers). We define the entry certification inspection as the treatment. Suppliers that never attempt traceability certifications are not included in this comparison. Despite the small sample size, a few features are found to statistically significantly different between the certified and attempted-certify suppliers. These results are shown in Table 2.3.

The certified suppliers are in locations with less traceability, include more aquatic suppliers, and have longer distances to their farms and factories. In addition, the CFDA failure rates and regular internal inspection grades are significantly different between the two groups, both before and after the entry level certification inspection, and on average. The certified suppliers have significantly more regular internal inspection failures before the entry inspection but fewer failures after compared to suppliers who do not pass the inspection. However, certified suppliers are more likely to fail a CFDA test than the attempted-certify suppliers, though this failure rate does drop after certification. In addition, after receiving certification, suppliers are also less likely to be sampled in a different prefecture than its manufacturer. It is interesting to also note that, in line with current research, the suppliers who pass or attempt to pass certification

Table 2.2: Statistical Significance of Difference Between Suppliers With and Without Failures

Feature	Average		P-value
	Failures	No Failures	
Number of CFDA Tests	92.7	23.9	$< .001$***
Misconduct Ranking(0-5)	.414	.643	.015**
Is an Egg Supplier	22.4%	9.80%	.015**
Age as Collaborator's Supplier	3.88	2.89	.027**
Certified during a CFDA Test	43.1%	12.9%	.063*
Product Quantity	9.15	5.31	.141
Transparency Score	3.53	3.46	.175
Number of Factories	1.48	.812	.178
Prefecture's Population	613	531	.242
Prefecture's GDP per Capita	47600	44200	.228
Percent Certification Inspections Failed	.027	.017	.387
Distance to Farms/Factories	.731	.537	.516
Percent Regular Inspections Failed	.153	.140	.748

***: $p < .01$; **: $p < .05$; *: $p < .10$
p values are derived from chi-squared tests or two sided t tests among the 313 suppliers that have been sampled

inspections are all larger suppliers with significantly more farms, factories, product quantity, and variety than suppliers who never attempt traceability certification.

Table 2.3: Statistical Significance of Difference Between Suppliers That Have Passed Versus Failed the Entry Certification Inspection

Feature	Feature's Mean Value		P-value
	Certified	Attempted-Certify	
Failure Rate	.011	.000	$<.001$ ***
Regular Inspections Failed (BT)	24.9%	0.00%	$<.001$ ***
Regular Inspections Failed (AT)	9.80%	20.7%	$<.001$ ***
Aquatic Supplier	38.5%	0%	$<.001$***
CFDA Failure Rate (AT)	.009	.000	$<.001$***
CFDA Failure Rate (BT)	.018	.000	.002***
Distance between Farms and Factories	.829	$<.001$	.025**
Transparency Score	3.58	4.00	.069*
Test and Manufacturer Prefecture Different (AT)	75.5%	82.1%	.076*
Test and Manufacturer Prefecture Different (BT)	85.6%	82.5%	.627
Product Quantity	19.5	14.3	.284
Number of Factories	2.62	1.67	.394
Number of Farms	4.54	5	.909

BT: Before Treatment; AT: After Treatment
***: $p < .01$; **: $p < .05$; *: $p < .10$
p values are derived from two sided t tests of the differences between the certified (26) and attempted-certify (3) suppliers; or their corresponding 1785 (499 before treatment) and 135 (40 before treatment) CFDA tests, respectively

These preliminary statistics demonstrate that some of our hypotheses may not hold within our data and that it is nontrivial to model whether a supplier would fail a CFDA test.

2.5 Methodology

Due to the small size of our dataset, we were limited in the complexity of the techniques we could use to predict food safety risks. At the outset, we tested a variety of models, including support vector machines, linear discriminant analysis, naive Bayes classifiers, hierarchical clustering, classification and regression trees (CART), probit regressions, and Heckman's sample selection. We also applied Synthetic Minority Over-sampling Technique (SMOTE) to create more balanced data (between the number of suppliers with and without CFDA failures) for model training. In this and the results chapter we will focus our discussions on SMOTE, CART, probit, and Heckman's sample selection model because they perform significantly better than the other models.

2.5.1 SMOTE

Since its publication in 2002, SMOTE has laid "the foundation for learning from imbalanced datasets" [27]. Numerous extensions and variations have since been developed to create synthetic minority data for a variety of situations. In our study we apply an R implementation [28] of the original SMOTE algorithms described in [29] and summarized here.

SMOTE improves the classification of minority classes in imbalanced data. It allows one to over-sample the minority class and under-sample the majority class. Unlike previous algorithms which over-sample the minority class by replication, leading to over-fitting, SMOTE creates synthetic minority data. It over-samples the minority class by taking k (in our case, $k = 5$) nearest neighbors for a given minority data sample, finding the difference between the features of it and a randomly chosen neighbors, multiplying this difference by a random number between 0 and 1, and adding it to the feature vector. SMOTE repeats this sampling and perturbation algorithm to create minority data samples according to the amount of over-sampling desired. For instance, over-sampling by 200% creates two new synthetic minority samples by separately perturbing a sample along the vectors of two different nearest neighbors. SMOTE also allows one to under-sample the majority class by removing samples until the new majority class is a certain percentage of the original minority class' sample size. Depending upon the percentage of over and under sampling, the resulting dataset may have more or fewer samples in the minority class than in the original data.

With slight variation, a similar technique can be used for categorical variables. In the case of mixed categorical and continuous variables, like our dataset, SMOTE calculates the nearest neighbors by first calculating the median of standard deviations of the continuous features in the minority class. If the categorical variables differ between the sample and its potential nearest neighbors, then the previously calculated median is included in calculating the Euclidean distance between samples. After the k nearest neighbors are determined, the synthetic categorical features are assigned the majority occurring values amongst the nearest neighbors while the continuous variables are calculated in the original fashion.

By creating synthetic minority classes, SMOTE creates more general decision regions than the small, specific regions that result from replication of minority classes. Because data is only perturbed by a factor between 0 and 1, this method does limit the synthesized data to be no more or less than the extreme values of the real data. Yet this approach has proven to be successful in improving the classification of the minority class and has been applied to problems in a variety of applications, such as: text classification, time series, and bioinformatics, to name a few [27]. We have also found it used in the agricultural industry, as previously discussed in predicting cattle welfare and dairy product safety [22][23].

Due to the small sample size and the low failure rates in our data, we apply SMOTE to expand our minority class of high-risk suppliers. We use a five-fold cross validation approach, such that the model was trained on 80% of the data and tested on 20%. A range of over and undersampling percentages from none to 1000% was applied to the training set. We then built a CART or probit model on this synthetic training dataset. We do not use synthetic data on Heckman's sample selection model because it would cause undesired effects on the estimation of the selection model. The test set also was not over or under sampled. We used it to validate our model and calculate the AUC (area under the ROC curve), accuracy, and confusion matrix. We then calculated the precision, recall, and F-1 through summing the confusion matrix across all five validated test sets for a conservative estimate of model performance on the entire test set. For each specification of features and sampling levels, we iterated this procedure one hundred times in order to calculate a confidence interval of our model accuracy.

2.5.2 CART

Decision trees are a commonly used and interpretable method of classification and prediction in a variety of contexts, including supply chain management. Dani recommends using classification techniques, like classification and regression trees (CART), and regressions as possible methods to predict supply chain risk, which is a requirement for "an effective proactive risk management process" [30]. CART is a popular decision tree methodology first discussed by Breiman, et al. in *Classification and Regression Trees*. We apply an R implementation of CART called Recursive Partitioning (rpart) [31].

Rpart splits nodes along features that maximize impurity reduction. We use the Gini information index as an impurity function, where it is defined as $f(p) = p(1-p)$. Impurity is defined as $I(A) = \sum_{i=1}^{C} f(p_{iA})$ across C classes. Therefore rpart is maximizing the following function:

$$\Delta I = p(A)I(A) - p(A_l)I(A_L) - p(A_R)I(A_R)$$

p_{iA} is the proportion of a node, A, that belongs to class i in the data, and R/L are the right and left splits of said node A [31]. Rpart continues splitting nodes in order to maximize impurity reduction until there is only one sample, no difference in splits, or it has reached a maximum pre-determined depth. At this point, the algorithm prunes the tree. It simplifies the tree according to a complexity parameter that represents a minimum level of improvement that a split must achieve to be included in the tree [32].

CART is popular because it is relatively easy to interpret and implement [33]. It also requires fewer assumptions than regression models. It performs well with nonlinear, multi-modal data as it is non-parametric [32]. As a result, it often has better predictive performance than naive regression models. However, the model performance is significantly dependent on how well the complexity parameter is tuned, because it determines the simplicity of the tree and the possibility of overfitting. In addition, with the addition of interaction terms and transformations, regression models can have comparable performance to CART models [33].

2.5.3 Probit Model

Probit models are another commonly used method of risk prediction first proposed by Bliss in 1934 to model medicine dosage mortality [34]. It has since expanded to other fields as a method of predicting various binary responses.

Along with assuming binary dependent variables, independent observations, and little multicollinearity between variables, a probit model assumes a normal distribution of errors. Therefore, a probit regression results from assuming

$$\Phi^{-1}(\pi) = \beta_1 + \beta\mathbf{x}$$

[35]. Φ denotes the cumulative probability function for $N(0,1)$ and π is the probability that $Y = 1$, where Y is the dependent variable. The β's can then be estimated using maximum likelihood estimation.

After an initial model is built, we choose relevant features using stepwise model selection. Features are removed or added based on the extent to which they improve the model versus the cost of increasing model complexity. This is measured by difference in Aikake information criterion (AIC), where $AIC = -2ln(\hat{L}) + 2k$. $\hat{L}$ is the maximum likelihood of a given model, and k is the number of parameters. Features are changed in order to minimize this score.

In order to improve model accuracy and the prediction of the minority class, the initial features must be carefully considered. Inclusion of interaction and transformed terms as well as close attention to a probit model's parametric assumptions can lead to sufficient improvements such that less interpretable machine learning models are unnecessary [33]. Probit models, like CART, also do not rely on as much data as more complex, but often times, more accurate machine learning models like random forests and neural networks.

Although stepwise selection does reduce the model size, it only reduces it to a local minimum of AIC. Because of its stepwise characteristics, different starting features will result in different final models. In our analysis, we experimented with a range of starting variables to include interactions and transformations of features. For example, we included the interaction between the number of farms and product quantity, as a supplier with a large number of products and large number of farms may affect risk differently from a supplier with few farms and few products. In addition, we tested interactions of the number of internal tests on test grades, and the product quantity on number of factories. We also took the log of the population and gdp per capita and experimented with both the square root and logs of the number of CFDA tests and supplier age.

2.5.4 Heckman's Sample Selection Model

Neither probit nor CART account for sample selection biases. They assume that suppliers were sampled by the CFDA at random. We employ Heckman's sample selection model to account for potential selection biases that may result from non-random sampling by CFDA officials or from suppliers removing themselves (or being removed) from the market prior to being sampled (e.g. recognizing a food safety risk and proactively exiting the market) [36].

We estimate the Heckman's selection model using maximum likelihood estimation (MLE) rather than the original two-step approach. Although the maximum likelihood estimation is less computationally flexible than a two-step approach [37], it is more efficient [38]. We follow the framework used in [24] and formulate the selection and outcome equations accordingly:

$$S_i^* = \gamma Z_i + \epsilon_i^S$$

$$R_i^* = \beta X_i + \epsilon_i^R$$

.

S_i and R_i are the likelihoods of being sampled by the CFDA and having a CFDA inspection failure, respectively, for supplier i, where $S_i = 1$ and $R_i = 1$ mean that supplier i was sampled and had a failure. S_i^* and R_i^* are the latent variables such that $S_i = 1$ if $S_i^* \geq 0$ and $R_i = 1$ if $R_i^* \geq 0$, and both S_i and $R_i = 0$ otherwise. γ and β correspond to the vector of coefficients for the independent variables Z_i and X_i. The error terms are represented by ϵ_i^S and ϵ_i^R, such that a nonzero correlation, ρ, between the two indicates the presence of sample selection biases. These error terms are assumed to jointly follow a bivariate normal distribution with mean 0, standard deviations σ_S and σ_R, and covariance equal to $\rho\sigma_S\sigma_R$. Ultimately, using maximum likelihood estimation, our model results from the following:

$$\max_{\gamma,\beta,\rho,\sigma_S\sigma_R} \mathcal{LL} \equiv \sum_{i\epsilon\{i:S_i=0\}} \log \mathbb{P}(S_i = 0) +$$

$$\sum_{i\epsilon\{i:R_i=1,S_i=1\}} \log \mathbb{P}(R_i = 1, S_i = 1) + \sum_{i\epsilon\{i:R_i=0,S_i=1\}} \log \mathbb{P}(R_i = 0, S_i = 1)$$

We implemented this estimation by adapting the code developed in [24] and experimented with various combinations of features. We trained a model on 70% of the data and then tested its prediction accuracy on the remaining 30%. We iterated this procedure one hundred times per model to develop confidence intervals for the model coefficients and accuracy. We tested if there is significant sample selection bias by using a likelihood ratio test. Specifically, we compare the log-likelihood of the Heckman sample selection model to the sum of the log-likelihoods of independently constructed selection and outcome probit models that assume $\rho = 0$. The log likelihood ratio is calculated as follows:

$$\mathcal{LLR} = \mathcal{LL}_h - (\mathcal{LL}_s + \mathcal{LL}_o)$$

$\mathcal{LL}_h$ is the log likelihood of the Heckman model and $\mathcal{LL}_s$ and $\mathcal{LL}_o$ are the log likelihoods of the independently constructed selection and outcome models. $2 \times \mathcal{LLR}$ follows a χ^2 distribution, so can be used to calculate its p-value.

Although Heckman selection model is more generalizable, it is more dependent on the model being correctly specified than a regular regression [39]. However, even if no selection bias is detected, by using a Heckman sample selection model we are able to gather a better, more generalized understanding of the underlying interactions between features,the odds of being sampled, and the risk of food safety failures.

2.5.5 Model Assessment

These models can all be evaluated using a variety of prediction measurements, to include accuracy, F-1, prediction, recall, and AUC. Accuracy equals

$$\frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, FN are true positives, true negatives, false positives, and false negatives, respectively. Positives indicate high-risk suppliers; negatives indicate low-risk suppliers. In an unbalanced dataset like ours, a high accuracy but useless model can be gained by simply predicting all suppliers as low risk. Alternatively, F-1 is useful in measuring how well a model predicts the minority class. For models using unbalanced data, there is a tradeoff between high accuracy and high F-1. F-1 is the harmonic mean of prediction and recall, such that

$$\text{F-1} = 2 \times \frac{precision \times recall}{precision + recall}$$

Precision is a measurement of how many of the predicted positive class are correct: $\frac{TP}{TP + FP}$. Recall is a measurement of how many of the actual positive class in the data are predicted correctly: $\frac{TP}{TP + FN}$. Finally, AUC (Area under the ROC curve) is a measurement of how well a model discriminates between binary classes. The ROC curve is a plot of the true positive rate against the false positive rate at various propensity score thresholds. AUC can be interpreted as the expectation that a randomly selected positive sample will be predicted to have a higher probability than a randomly selected negative sample (e.g. a model with an AUC of .7 has a 70% chance of predicting higher probabilities for high-risk suppliers than for low-risk suppliers). Like accuracy, AUC is also skewed by imbalanced data. Depending on a stakeholder's goals, any of these measurements of model quality can be useful. For our study we focus on models that best discriminate the minority from the majority class, i.e. have a high F-1.

2.6 Results

In this chapter, we discuss the results of our CART, probit, and Heckman models. The AUC, F-1, accuracy, precision, and recall of the best models using these three techniques are provided in Table 2.4. For comparison, we also provide the prediction results of our CART and probit models without using SMOTE. As the table shows, over- and under- sampling the minority and majority classes, respectively, do improve prediction of the minority class. However, even with the addition of synthetic data, our CART and probit models still perform worse at minority class predictions than the outcome model using the Heckman framework.

Table 2.4: Food Safety Prediction Results

	Mean (StDev)				
Model	AUC	F1	Accuracy	Precision	Recall
CART	.678 (.037)	.440 (.029)	.673 (.026)	.323(.024)	.694(.054)
CART -no SMOTE	.651 (.070)	.438 (.053)	.833 (.013)	.583 (.057)	.352 (.053)
Probit 1	.786 (.011)	.496 (.016)	.751 (.011)	.397 (.015)	.662 (.244)
Probit 2	.774 (.012)	.507 (.018)	.788 (.010)	.446 (.019)	.588 (.026)
Probit -no SMOTE	.764 (.018)	.347 (.024)	.832 (.006)	.623 (.044)	.24 (.019)
Heckman Outcome	.772 (.049)	.541 (.061)	.850 (.018)	.505 (.132)	.594 (.134)

Results are the means and standard deviations across one hundred iterations.

2.6.1 CART

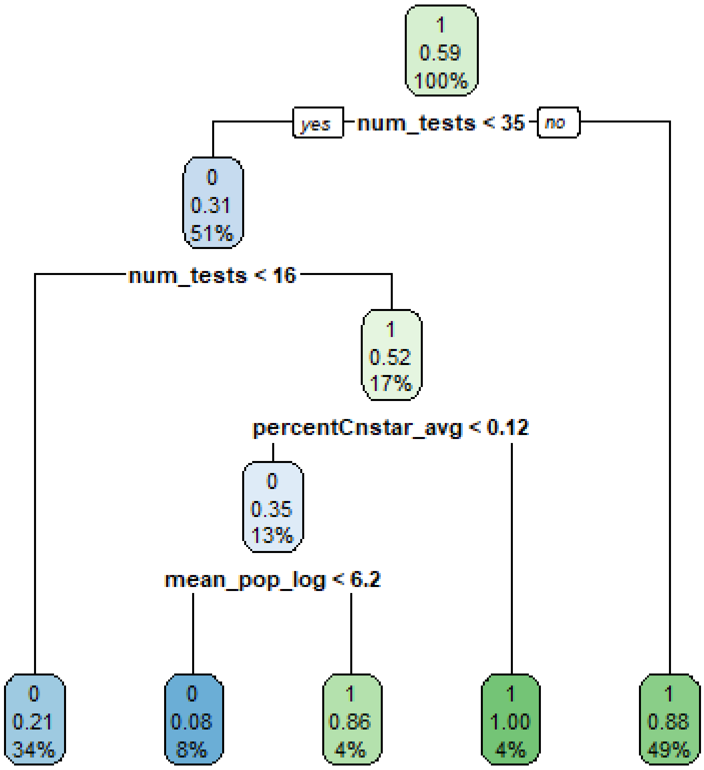
We experimented with a variety of potential features and over/under-sampling ranges to train our CART models. A small amount of synthetic data slightly improves the AUC and F-1 scores, while too much over- or under-sampling results in the model predicting mostly the majority class. Ultimately, the best model is the one that over-samples the minority class by 100% (doubling its size) and under-samples the majority class to be 140% of the original minority class. The best CART model is still 7% worse than the best probit model when comparing F-1 scores. However, this F-1 score is less than 50% which indicates it is not a particularly useful model. It can be intuitively interpreted to represent that if the majority and minority class were split 50-50, our model would not do better than random guessing. The model's recall is on average 69.4%. This indicates that it is able to correctly predict 69.4% of the minority class. However, its precision is 32.3%, indicating that it is over classifying suppliers to the minority class.

Across one hundred iterations with different training/testing splits, the variables used to build the models in order of average importance (i.e. the greatest total decrease in impurity for all splits) are number of CFDA tests, percent of CFDA tests completed while a supplier is certified, percent regular inspections failed, age of a supplier, prefecture's population, age as a certified supplier, average distance to farms and factories, number of certification inspections every year, prefecture's misconduct ranking, prefecture's GDP per capita, number of regular inspections every year, number of product types, percent of certification inspections failed, and whether it is an egg supplier. This is the only model in which our collaborator's internal inspection outcomes play a significant role in affecting the risk levels of the suppliers.

We depict the single best predictive model (highest F-1) out of all one hundred trials in Figure 2.10. This tree shows, as an example, that 49% of suppliers have more than 35 tests, and if a supplier has more than 35 CFDA tests, then it has an 88% chance of being high risk. Each node shows the predicted classification (high-risk, 1, or low-risk, 0), the probability of being high-risk, and the percent of suppliers that belong in that node.

The CART models are not consistent across various training and testing splits, as noted by the discrepancy between the list of important variables across one hundred trials previously

Figure 2.10: Best CART Model



Note: "num_tests": number of CFDA tests; "percentCnstar_avg": average number of our collaborator's regular inspections failed; "mean_pop_log": log of the supplier's prefecture's population

described and this "best" model. For example, while the model depicted in Figure 2.10 classifies, in accordance with our hypothesis, suppliers with more regular internal inspections failures as high-risk suppliers, this often is not the case in the other 99 models trained using the same features. The variables used and the direction of splits are not consistent.

Along with the low prediction scores, the inconsistency of the CART models indicates that despite its interpretability and minimal assumptions, it is not a good technique for prediction with this small and unbalanced dataset.

2.6.2 Probit Model

Our probit models performed significantly better than the CART models in both AUC and F-1. Following the same process as the CART models, we tested models built with various combinations and interactions of our variables over a range of over- and under-sampling percentages. The best probit model required more synthetic data than the CART model. It results from data with the minority class over-sampled by 200% and the majority class under-sampled to be 180% of the original minority class. It is interesting to note that the probit models built without any synthetic data have significantly worse F-1 performance, but significantly better

AUC scores, than the CART models without synthetic data (Table 2.4).

Our best models are built using the variables and results outlined in Table 2.5. Regardless of whether or not synthetic data is used, the same variables are significant. Probit 1 uses a natural log transformation of the number of tests and supplier age. It is the model with the highest AUC amongst all models we trained. Including the square root of both the number of tests and the supplier's age results in a model with the highest F-1 score, Probit 2. Observe from Table 2.4 that the AUC of both models indicate that more than 75% of the time suppliers with CFDA failures are predicted to have a higher risk than suppliers without failures. However, Probit 1's F-1 score is less than 50% and Probit 2's F-1 score is marginally greater than 50%. These low F-1 scores are a result of their disproportionately low precision scores. Both models are over assigning suppliers to the positive, high risk, class. This indicates that these models are not good predictors of high risk suppliers in practical applications.

Table 2.5: Regression Results of Probit Models

	Value	Standard Error	p-value
Probit 1			
Misconduct Ranking	-.283	.153	.019 **
Egg Supplier	.410	.332	.025 **
Log(Number of CFDA Tests)	.510	.065	.000 ***
Log(Supplier Age)	.052	.047	.022 **
Probit 2			
Misconduct Ranking	-.261	.153	.027 **
Egg Supplier	.410	.330	.027 **
Sqrt(Number of CFDA Tests)	.177	.033	.000 ***
Sqrt(Supplier Age)	.237	.126	.021 **
Probit - no SMOTE			
Misconduct Ranking	-.101	.113	.048 **
Egg Supplier	.410	.300	.044 **
Sqrt(Number of CFDA Tests)	.137	.011	.000 ***
Sqrt(Supplier Age)	.085	.095	.041 **

***: $p < .01$; **: $p < .05$; *: $p < .10$
Values are the mean of the estimated coefficients across one hundred iterations.

Despite this poor predictive performance, our model is useful in understanding the data. The positive coefficient assigned to egg suppliers supports the characteristics we found in our chi-squared test. Egg suppliers have a higher risk of failures. Unsurprisingly as well, number of CFDA tests and supplier age also have positive coefficients. The more CFDA tests and the longer a supplier exists, the more likely it has had a CFDA failure. The square root and natural log transformations of these two variables indicate that increasing ages and tests have diminishing effects on the likelihood of failure.

The coefficients on misconduct rankings are also aligned with the results of our chi-squared test. They consistently have a negative coefficient. This indicates that suppliers in locations with more reports of misconduct will have a lower probability of failing CFDA tests, which may indicate that less corrupt governments are more effective in detecting problematic products and suppliers.

The lack of inclusion of supply chain, internal inspection, and certification features in our best model indicate that they may not be as influential on reducing or predicting food safety risks as researchers have previously expected. While the results of our probit models are more consistent than the CART models, they still have weak predictive capability. In the following section, we discuss our overall best model.

2.6.3 Heckman's Sample Selection Model

The outcome model built using a maximum likelihood estimation of Heckman's Sample Selection model has the best prediction results overall, with an average F-1 score of 54.1%. This 4 to 5% increase in F-1 score is improved primarily as a result of increased precision; the model has fewer false positives while maintaining a comparable number of true positives. However, its average AUC is not better than our probit models. Neither of these differences are statistically significant.

The likelihood ratio test indicates that there is no sample selection bias in our data. Although sampling prediction is not a study objective, we observe that the predictive accuracy of the selection model is quite low. However, joint estimation of the selection model with the outcome model does improve risk level classification accuracy. The inclusion of statistically insignificant variables in our selection model also improves this accuracy. These results are shown in Table 2.6. From these results, we observe that greater transparency in a supplier's prefecture, GDP per capita in a supplier's prefecture, and supplying non-meat, vegetable, aquatic, fruit, nut or tea products increase the likelihood of a supplier being sampled. Larger prefecture populations, and more regular and traceable internal test failures decrease the likelihood of a supplier being sampled by the CFDA. A possible interpretation of the relationship between internal inspection failures and sampling rate is that our collaborator successfully stopped sourcing from a supplier with inspection failures. This implication does not account for situations where other retailers remain unaware of a supplier's potential food safety problems and continue sourcing products from the supplier. These observations warrant future research.

The best outcome model contains the same variables as in our probit model. In addition, similar relationships between the variables and food safety risk are estimated. Egg suppliers, suppliers with a longer history with our collaborator, and suppliers with more CFDA tests have higher risk of failure, while suppliers in locations with more misconduct are associated with a lower chance of CFDA failure. Unlike the probit models, the regression results of the Heckman selection's outcome model are not all significant. However, the inclusion of the insignificant variables does increase predictive accuracy. Most importantly, like the probit models, the Heckman's sample selection models also demonstrate that our collaborator's quality and traceability certifications, internal inspections, and supply chain dispersion are not useful predictors of food safety risk.

Table 2.6: Regression Results of Best Heckman Sample Selection Model

	Value	Standard Error	p-value
Selection regression (likelihood of a supplier being tested)			
Misconduct Ranking	-.081	.196	.658
FDA Transparency	.106	.054	.078*
Log of Population	-.189	.055	.001***
Log of GDP per Capita	.258	.054	.000***
Number of Farms	.050	.029	.133
Number of Factories	.001	.048	.631
Variety Count	-.058	.109	.508
Other Product Supplier	.642	.161	.000***
Aquatic Supplier	.288	.158	.105
Percent Certification Inspections Failed	-1.42	.655	.052*
Percent Regular Inspections Failed	-1.06	.208	.000***
Outcome regression (likelihood of a supplier having at least one failure)			
Egg Supplier	.466	.308	.185
Number of CFDA Tests	.006	.001	.000***
Misconduct Ranking	-4.12	89.3	.960
Supplier Age	.069	.038	.120

***: $p < .01$; **: $p < .05$; *: $p < .10$
Values are the mean of the estimated coefficients across one hundred iterations.
Likelihood ratio: .177; p-value = .5518

2.7 Discussion

Our probit and Heckman models perform significantly better at predicting high risk suppliers than our CART model. The F-1 and AUC of our probit and Heckman's sample selection models are comparable. They are not significantly different from one another. Since our Heckman's sample selection model relies on fewer generalizations and assumptions, we consider it the best and most realistic model.

Regardless of which is the best model, in practical application, a few conclusions result from this analysis. Contrary to our hypotheses, quality certification, internal inspections, and supply chain dispersion do not affect food safety risk in our data. We find that our control features are the only characteristics that are significant in predicting CFDA failures. It is surprising that more misconduct in a prefecture is consistently related to lower risk of failures in both our probit and Heckman models. This may indicate that weaker governance is associated with less effective detection of food safety problems by local governments. Alternatively, this result could also be interpreted to mean that these prefectures have had more attempts to uncover misconduct and hence, they actually have stronger governance and accordingly better food safety. Further analysis is necessary to make that determination.

Unlike previous studies [24], we do not identify a relationship between the supply chain dispersion of a supplier and its food safety risk level. We had hypothesized that greater supply chain dispersion, represented by distance, variety, number of farms, and number of factories would be related to higher risks of food safety. These features are not found to be significant in any of our models. This indicates that they do not play a significant role in helping to predict CFDA failures of suppliers in our data. However, this result may also be biased because we only factor in the supply chain characteristics known to our collaborator. Suppliers may have more farms and factories that our collaborator does not source from and as a result does not have data on.

Of primary interest to our collaborator, we find that their internal inspections do not reduce the risk of a supplier having CFDA test failures. Internal inspections, both certification and regular ones, are not found to be significantly associated with the suppliers' risk levels. However, we do observe that internal inspection failures reduce the chances of a supplier being sampled by the CFDA. These results counter the current beliefs of traceability and quality management experts.

The scope of our analysis is limited by the amount and quality of data available. As demonstrated by our use of SMOTE, more data will improve analysis. Many of the suppliers that our collaborator provided supply chain data for did not have corresponding internal inspection results. With more consistent data collection of internal inspections, we could triple the size of the data analyzed from 313 suppliers to 1012 suppliers. With more CFDA data, we could also conduct more accurate analysis of supplier risks through analyzing the data as a panel dataset. It would be more useful to analyze the likelihood of a supplier failing a single CFDA test given its internal inspection results and characteristics immediately prior, than to analyze the aggregate risk of failures, as we did in our study. It would also be interesting to analyze if the results of CFDA tests affect future internal test results of a supplier.

There is still a significant amount of opportunity for future exploration and improvement in order to generalize these conclusions to other quality management systems. The CFDA began a new wave of reforms to align its national standards closer to international standards in 2017 [40]. With more data, we could isolate analysis to CFDA tests performed after 2017 and test if the increased standardization offers any improvement in detection or reduction of the effects of misconduct and transparency on failures. Also, if we had our collaborator's specific internal inspection features and results, we could test if there are specific aspects of the inspections that are more useful than others in predicting risk. It would additionally be beneficial if we received our collaborator's internal product inspection data. We could then directly relate it to the supply chain and inspection data, rather than relying on CFDA test results which are much more sparse per supplier. Due to the data limitations of this study, we can only draw conclusions on our collaborator's certification system, and must be careful not to over-generalize our results to other retailers' or third party traceability and quality inspection systems.

Chapter 3

Identifying Human Trafficking

3.1 Section Overview

3.1.1 Motivation

Modern day slavery, also known as human trafficking, exploits more people now than ever before in human history [41]. Human trafficking is the "act of recruiting, harboring, transporting, providing, or obtaining a person for compelled labor or commercial sex acts through the use of force, fraud, or coercion."[42] Despite slavery being outlawed, it remains a global problem and affects an estimated 40 million victims worldwide in the form of human trafficking [43]. In just the United States, l8,524 cases of human trafficking cases and 10,708 victims were identified in 2018 [44].

This thesis focuses on sex trafficking, which is estimated to make up 79% of human trafficking cases and generates an estimated annual profit $99B globally [45] [46]. Sex trafficking is characterized by individuals who commit commercial sex acts under threat of force, fraud, or coercion, or anyone under 18 years old [47]. Like most illegal activity, identifying and interdicting sex trafficking is a difficult problem for law enforcement agencies. Countering sex trafficking has become even more difficult in recent decades because it has moved from the streets to obfuscated online classified advertisements and the dark web.

It comes as no surprise that combating human trafficking is a "key Defense Department mission" in the United States [48]. In support of this mission, DARPA began the Memex program in 2015 [48]. Memex's goal is "to move forward the state of the art in content indexing and web searching on the Internet". This program has opened the path to developing tools that have proven useful in helping law enforcement counter human trafficking. One of these tools is TellFinder which provides users with visualizations of personas identified in archived web data by their similar attributes, like phone numbers or images. It then flags content with high risk human trafficking indicators [49]. However, current technology can not efficiently, automatically, and accurately identify these trafficking indicators.

Unfortunately, sex trafficking investigations are resource and time intensive activities. In Florida, a 2019 sex trafficking case across ten spas that resulted in more than 200 charges [50] took seven months and over $400,000 worth of detective work to build [51]. Suspicious activity only came to law enforcement's attention after a health inspection, despite many publicly available reviews (on Yelp and Google for instance) indicating that the massage parlors were actually fronts for brothels. Even with all of this detective work, the spa owners' may only be charged with prostitution solicitation rather than human trafficking despite many clear indicators that the sex workers were being manipulated [50].

Successful applications of the tools resulting from Memex can significantly aid law enforcement in identifying and investigating human traffickers. In January 2019, with the help of technology that scheduled and tracked prostitution dates from online posts, law enforcement officers succeeded in seizing about 500 websites and indicting six people for running a global sex trafficking organization in the U.S., Canada, and Australia. This organization logged more than 30,000 customer phone numbers [52]. However, an additional resource is needed that could identify advertisements, contacts, locations, and ultimately organizations of suspected sex traffickers from online ads. This would allow law enforcement to be even more efficient and effective with their resources.

The difficulty in building such a platform is that identifying sex trafficking is a nontrivial problem. The advertisements are often hidden amongst legal escort service and voluntary (albeit, illegal in most of the U.S.) prostitution. These ads are full of non-standard English grammar structures and emoticons. Furthermore, human trafficking ad identification operates in an adversarial environment where traffickers are obfuscating text and using coded keywords, like the global sex trafficking case previously mentioned [53], to describe services. Although Backpage.com, a former major platform for sex ads, has been shutdown, human trafficking ads have since resurfaced on other platforms. Without the consolidation of ads on one site, manual online data combing has become even more difficult [54]. As a result, given the time intensiveness of labeling advertisements, a useful platform must be able to identify trafficking ads even as obfuscation techniques change. It must be a generalizable model that does not depend on characteristics specific to the training dataset, like emails and phone numbers, but rather adapts to identify new keywords as the language used in human trafficking ads transforms. This would allow law enforcement to spend less time sifting through data and would provide them with starting points for future investigations. This study furthers the development of such a tool for combating sex trafficking.

In the following sections we discuss our work in developing a pipeline to improve upon current sex trafficking detection technology.

3.1.2 Objective

The objective of our work is to answer the following questions:

1) Can we build an accurate and interpretable model for detecting sex trafficking advertisements?

2) How can we identify keywords of sex trafficking ads even as language transforms?

3.1.3 Contributions

We develop a text based pipeline using natural language processing and interpretable predictive algorithms that performs better at classifying human trafficking ads than all known models, to include models trained using known human trafficking keywords. Our pipeline also has better predictive performance than the results of a published deep multimodal network model approach that uses both the pictures and text of the same data as this study [55]. Although we only have non-dynamic data, we demonstrate an opportunity for accurate and unsupervised keyword identification. Our pipeline detects structures in human trafficking advertisements and narrows down keyword lists that distinguish human trafficking advertisements. Finally, we demonstrate that our pipeline can be successfully applied to outside data to detect suspected human trafficking organizations. Unlike current state-of-the-art models, not only does our pipeline allow for efficient and accurate human trafficking detection, it is also interpretable and could allow for keyword identification even as language transforms.

3.2 Literature Review

The role of social networking sites and online ads in facilitating human trafficking was unclear in 2010 [56]. Today, there is no question that online platforms are being exploited by human traffickers. However, as Laterno writes, technology can be used to further efforts to combat it as well. His comprehensive report of online human trafficking suggests data scraping, natural language processing, and facial recognition as technology that can be leveraged to identify victims more quickly [56]. Many of these technologies and more have come to fruition almost a decade later.

These developments are outlined in a recent review of the relationship between technology and human trafficking in [47]. For example, human trafficking detection programs include PhotoDNA, Spotlight, and Traffic Jam. PhotoDNA compares photos to those in a repository of confirmed child exploitation cases and automatically reports matches to law enforcement. Spotlight searches the internet for advertisements promoting sexual acts [47]. Traffic Jam, developed by Marinus Analytics (our data provider), combines facial recognition, natural language processing, and network analysis to identify human trafficking ads that may be linked to an input photo or phone number [57][58]. Both Spotlight and Traffic Jam provide descriptive information on the suspected victims in addition to contact or location information from the advertisements. They use machine learning techniques and linguistic properties to improve data scraping, overcome text obfuscation, and identify high risk advertisements [47]. Ultimately, Pendergrass finds that while technology has significantly aided traffickers in ensnaring and exploiting victims, new developments, especially in machine learning, have also reduced the manual labor required of law enforcement to identify human trafficking victims.

Many of these technological developments leverage linguistic cues in advertisements. Research has repeatedly proven that advertisement language often contain human trafficking signals. In one of the first studies to apply data analytics to online human trafficking, researchers analyzed advertisements in the Adult section of Dallas' Backpage site for the week leading up to the 2011 Super Bowl [56]. Using natural language processing, they were able to find potential keyword indicators of trafficking. However, researchers were unable to confidently verify that the suspected ads were actually human trafficking using these methods [56].

Building from this study, researchers have primarily detected suspected sex trafficking advertisements from escort service advertisements using sets of pre-determined attributes. Most of these studies also do not have truth data. For example, Kennedy, the cofounder and president of Marinus Analytics, developed a methodology for detecting and visualizing patterns in trafficking movements via text analytics. Kennedy narrowed down a database of advertisements scraped from Backpage by using various characteristics, like keywords, language, websites, phone numbers, and locations, that map to indicators of human trafficking: being underage, shared management, and movement. This pipeline then allowed a user to conduct queries and visualize the related advertisements and their metadata (eg. posting time, frequency, and location) [59].

Similarly, Silva et al., designed a system for identifying the prostitution networks of possible underage sex trafficking victims [60]. However, like Kennedy's research, the utility of this tool is also dependent on information known a priori because there was no verification data[60]. A more detailed content analysis of advertisements posted on Hawai'i Backpage created an index of human trafficking indicators: inconsistent ages, inconsistent aliases, movement, shared management, third party posting, advertised nationality, and potential restricted movement [61]. However, this analysis found that out of the 1436 advertisements analyzed, 82% of the ads contained one or more indicators. It is unlikely that the true prevalence rate is this high. This shows that the presence of any given indicator can not be seen as proof of sex trafficking but only a flag to be raised for further investigation [61].

Researchers have also found that not only do human traffickers use coded words and phrases, they also use coded emoticons [62]. An ontology of emoticons, keywords, and phrases that are indicators of human trafficking were compiled from interviews with law enforcement and individuals involved in combating sex trafficking by Whitney et al. Emoticons used in advertisements with keyword/phrase indicators of human trafficking were then compared with advertisements without indicators using hypothesis testing and logistic regressions. This exploratory study not only found that emoticons are a useful indicator of human trafficking but that they may be used independent of keyword indicators [62]. Although this discovery provides more leads to potential victims, it adds another layer of noise for accurately narrowing down human trafficking investigations. Hultgren et al., suggests researchers can use a knowledge management approach to update keyword ontologies as successes occur to maintain system accuracy [63]. However the manual identification and updating of keywords that is suggested is laborious in itself. These studies are all dependent upon the accuracy and availability of known indicators of human trafficking. Automated detection of sex trafficking advertisements and indicators would be a significant improvement to current technology but is yet relatively un-researched.

One example of a supervised modeling approach to human trafficking detection is a study by Dubrawski et al. They compare three methods of feature selection and test a tenfold cross-validation random forest classifier on a dataset of 37,000 unique advertisements where 40% of the ads contained phone numbers of known traffickers and the remaining 60% were randomly selected from a set of unlabeled escort service advertisements [64]. These three models use keyword/phrases gathered from interviews with law-enforcement, regular expression extractions of personal identifying physical and operational characteristics (e.g. ethnicity or url), and natural language processing features selected from the top 300 principal component analysis words from a bag of words representation of all 16 million advertisements. The model trained using NLP selected features has significantly better predictive performance than both the keywords and regular expression based models. This suggests that although experts have identified discriminatory keywords and phrases, the NLP selected features are able to identify more subtle indicators.

Dubrawski et al.'s has excellent predictive results (F-1 of 73.7%). However, their training data is an unrealistic representation of real data because the prevalence rate of human trafficking is much lower. This would make model training much more difficult. In addition, they built their NLP features using all 16 million advertisements, which means it used both the words in the training and testing set. To achieve unbiased results, they should have selected features only from the set of training advertisements. Nevertheless, this study demonstrates that unsupervised NLP features may be an improvement from detecting human trafficking advertisements using pre-determined indicators.

Tong et al. similarly uses natural language processing, combined with computer vision, techniques to detect suspected human trafficking advertisements. They built a rigorously annotated dataset of 10,000 advertisements to train their deep multimodal network model, named Human Trafficking Deep Network (HTDN). HTDN uses a language network built using word embeddings trained on one million unlabeled ads outside of the annotated set coupled with a vision network using advertisement images. They find that HTDN performs significantly better than baseline models built using random forest, logistic regression, and linear SVM with 108 keywords, average trafficking vectors, 108 informative words, or bag of words as features. They report an upper bound F-1 per human performance metrics of 73.7%. HTDN results in an F-1 of 66.5% [55]. Although the HTDN pipeline performs better than all the other more simplistic methods, it does not allow for any interpretability. Law enforcement would have to accept the results at face value as they can not decipher how features impact the results. In addition, this approach is completely dependent upon having a well annotated training set, which significantly detracts from the automation of the entire process. Using the same training data, we propose a more interpretable and more accurate method that is less dependent on labeled data.

3.3 Data

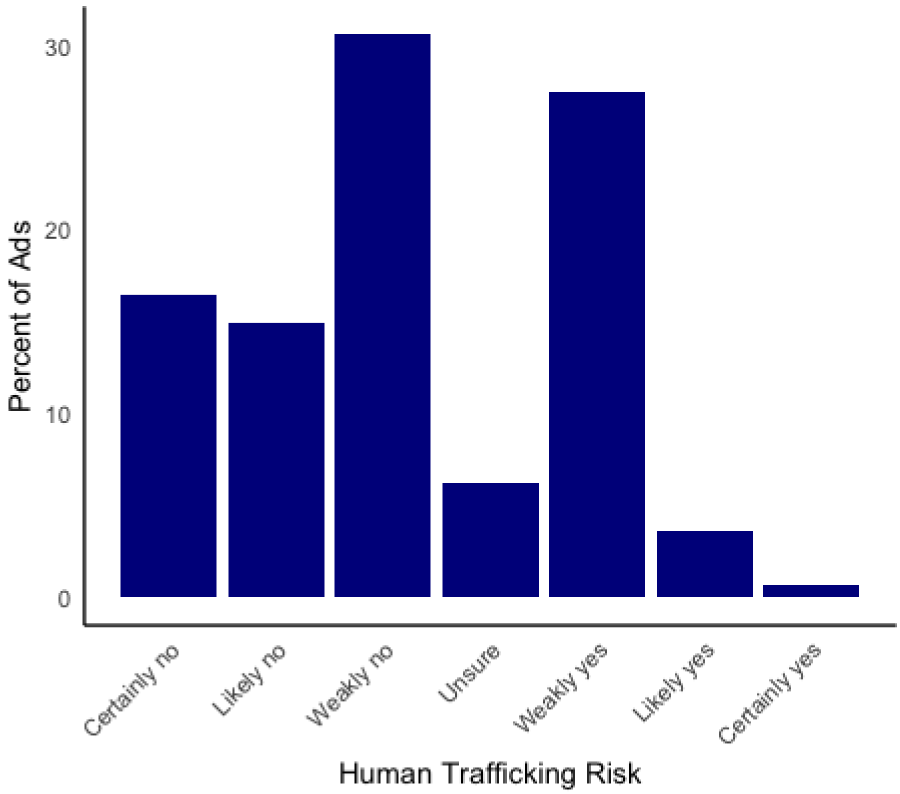
In this section we present the datasets used to train our language models and classification models. Both were provided by Marinus Analytics.

3.3.1 Language Model Dataset

We use a set of over 2.5 million unique escort service ads that were scraped over a six month period in 2017 from the now defunct *Backpage.com.* These advertisements represent activities across the United States and Canada. Each ad is composed of all textual information that is displayed on the webpage, to include titles and emojis, and is labeled with IDs and locations. Although previous processing efforts have tried to remove ad reviews, due to the unstructured nature of this data many still remain in the data.

These ads are not annotated, so it is unknown how many of them are suspected to be tied to human trafficking organizations. Nevertheless, they provide an accurate, albeit unknown, representation of the true distribution and language of human trafficking in escort service ads.

Figure 3.1: Distribution of Human Trafficking Risk in Trafficking-10k Ads



3.3.2 Trafficking-10k Dataset

The Trafficking-10k dataset also is the data used in the previously discussed HTDN model in [55]. It is an annotated set of ten thousand advertisements that were randomly sampled from a larger set of escort ads scraped from *Backpage.com* at an unknown time[55]. The ads do not overlap with those in the Language Model Dataset. But like the Language Model Dataset, they also represent escort service ads from across the United States and Canada [55]. Although the annotators and the multi-modal HTDN model use images from the ads to aid classification, this study does not include images in its analysis.

These ads were rigorously annotated by subject matter experts on a 7 degree scale of likelihood of being human trafficking, with the middle level being unsure [55]. More information on the annotation methodology is described by Tong et al. The distribution of advertisements and scores is depicted in Figure 3.1. As depicted, most of the ads are not found to be human trafficking related. However, there is also a lot of uncertainty in classification. In fact, annotators are mostly uncertain even for the suspected human trafficking ads. As a result, because there are few certain high-risk ads, we approach this as a binary classification problem between not-suspected and suspected human trafficking ads. Ads that are suspected to be human trafficking are considered to be high-risk and ads that are unlikely to be human trafficking are considered to be low-risk.

3.4 Methodology

Given the level of human interaction and burden of proof required in human trafficking investigations, black box approaches are not ideal. Law enforcement officers require justification behind their actions. In addition, they do not have the resources to read through the millions of ads that are posted everyday. An ideal methodology would not only identify features used to classify human trafficking but be able to do so with minimal supervision. We develop a pipeline that can do exactly that. After pre-processing, we use unsupervised NLP features on a bag of words representation of each ad to train interpretable models (classification and regression trees, random forest, and binomial logistic regressions).

3.4.1 Pre-Processing

In order to focus on textual features we conduct a rigorous pre-processing of the advertisements to remove unnecessary or overly specific information using regular expressions. We cleaned up utf-8 characters that were mangled during crawling and striped HTML tags from the text. We also removed ad ID codes and locations. Next, we cleaned obfuscated words. The most common and easily replaced obfuscations were words whose characters were separated by spaces and asterisks. We then identified and replaced phone numbers, emails, costs, and times with filler words indicating the original purpose (e.g. "phonenumber" , "email"). We replaced all remaining numbers with a filler word as well. Finally, we removed all emojis, websites, and image references. For future research, we developed an alternate pathway to keep or use filler words for the emojis, websites, and images, but did not experiment with that implementation in this study. Finally, we tokenized all the punctuation. We did not conduct stemming because words like "girl" versus "girls" have significantly different implications in the case of human trafficking scenarios. This procedure was implemented on both the Language Model and Trafficking-10k Dataset. In the Trafficking-10k dataset, we also removed non-unique and the "unsure" class ads. It is unclear from Tong et al.'s study if the "unsure" class was included in the analysis and if so, if those ads were grouped with the minority (high-risk) or majority (low-risk) class. In addition, there are some non-unique advertisements that are annotated with different scores. As a result, to reduce the noise in our data, we neither include the "unsure" ads nor the non-unique ads. This leaves 9108 advertisements from the Trafficking-10k dataset for training and testing our models.

3.4.2 Phrase Detection

After pre-processing, a phrase detection and replacement algorithm is applied to the advertisements iso that phrases will be treated as a single token in later models. Phrase detection is applied because many of the indicators used in previous literature were not just singular words, but phrases like "new in town", indicating movement or a minor, or "no outcall", indicating restricted movement [47]. A word like "outcall" or "town" on its own is not a good indicator of human trafficking without the entire phrase. In addition, regardless of whether or not phrases are selected for model building, the inclusion of these phrases changes the language model estimations for the surrounding words and reduces multicollinearity between commonly neighboring words. Individual words may have lower perplexity scores in the language model if they are part of a phrase, but if the phrase is interpreted as a token, the scores of the combined words, and the surrounding context will be different.

The phrase detection algorithm is run on the Language Model Dataset of 2.5 million escort service advertisements to create a phrase dictionary. The phrase detection algorithm was created by the HDDN team at Lincoln Laboratory. It identifies repeated multiword units from the text and considers them to be phrases if they meet a count threshold and weighted pointwise mutual information (PMI) minimum. Weighted PMI equals the frequency of a multiword unit in a text multiplied by its PMI score. PMI measures the probability of mutual occurrence between tokens in the multiword unit given the rest of the corpus. Frequency is the number of occurrences of the multiword unit divided by the total number of words in the text. After this dictionary of phrases is completed, only the phrases that are within a minimum and maximum length requirements are kept. The resulting set is the final phrase dictionary.

We then concatenate and replace the detected phrases in both datasets, thereby transforming them into “words”, before creating the language model and training the classification models. Only exact matches are replaced. A more robust approach would build a phrase dictionary that takes into account obfuscations and minor variations.

We experiment with varying degrees of granularity in phrase detection by varying the minimum and maximum lengths of phrases and repeated occurrence thresholds. We test pipelines using phrases ranging between three to seven words and occurrence cutoffs ranging from ten to twenty. The most accurate classification models ultimately use phrases of three to six words and require a twenty occurrence minimum.

3.4.3 Language Characteristics

After processing, there are a total of 1,959,339 unique words in the Language Model dataset. The ads are in nonstandard English, with a small percentage also in foreign languages. The processed advertisements on average have 68 words and 436 characters per ad with a standard deviation of 51 and 343, respectively, due to some extremely long advertisements. Our most useful phrase detection algorithms were able to identify 2128 unique phrases. Each phrase is 3 to 6 words long, with a median length of 3 words. Each ad contains on average 4.5 phrases with a standard deviation of 3.8.

In the Trafficking-10k dataset, there are a total of 9,227 unique words. Similarly to the Language Model dataset, these ads have on average 75 words and 489 characters with a standard deviation of 48 and 376, respectively. There are also on average 3.6 phrases found in each ad with a standard deviation of 2.7.

The word count and character count are significantly different at $p < .001$ between the high- and low-risk advertisements (ignoring ads that are classified as “unsure”). On average low risk ads have more words and characters. The statistical difference found in average word count likely results from extreme outliers in the low risk advertisements, as can be seen in Figure 3.2. The phrase count between classes is not significantly different ($p = .8112$). This implies that the phrases detected in the Language Model dataset likely are not associated with risk indicators. The distribution of the phrase counts across various risk levels is depicted in Figure 3.3. These statistics demonstrate that high-risk and low-risk ads have similar language structures.

Figure 3.2: Word Count Across Risk Levels

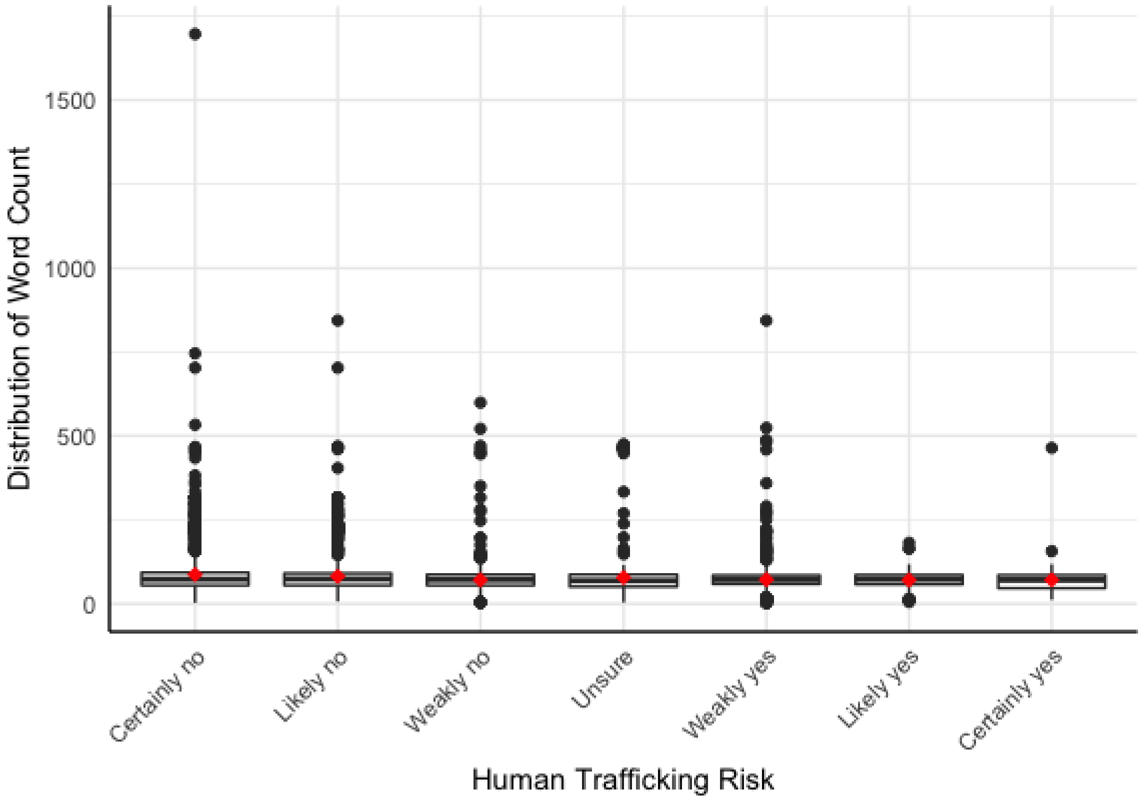
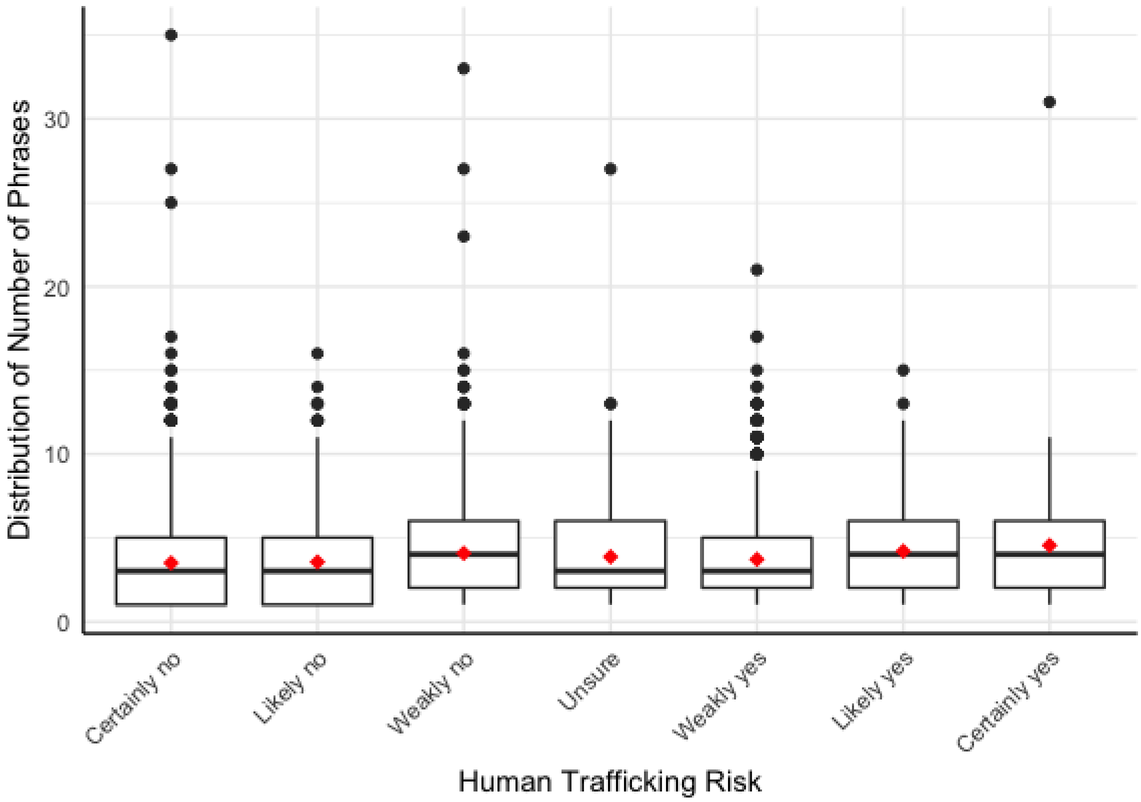


Figure 3.3: Phrase Count Across Risk Levels



3.4.4 Feature Selection Overview

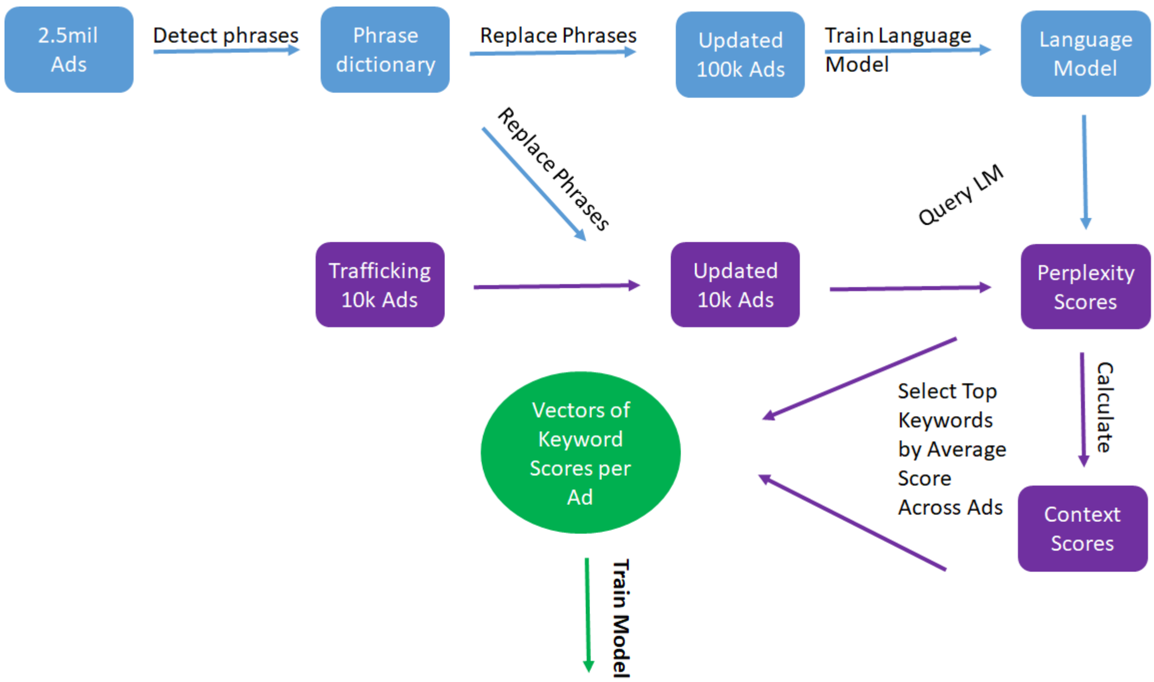
Features are selected using natural language processing (NLP) characteristics calculated from a language model trained on the 2.5 million ad dataset after pre-processing and phrase detection is completed. This language model is queried to evaluate perplexity and context scores of the words and ads in the processed Trafficking-10k dataset. We then experiment with perplexity, context, frequency, and TFIDF scores to select words for training the supervised model. A visualization of this process is included in Figure 3.4. We further detail our features selection pipeline in the following subsections.

Language Model

The keyword selection pipeline selects words using the results of a language model query. Language models are useful for predicting words that should occur. It is most commonly used to help machines identify words from noisy input, like in speech, handwriting, spelling, or translations [65]. In addition, they allow for topic independent keyword detection [66] and topic signature detection [67]. Given these benefits, we hypothesize that language models are also useful for identifying words that we do not expect to occur – the words and phrases that are indicators of human trafficking advertisements. At the time of writing, we are not aware of any similar applications of language modeling in human trafficking ad detection.

We train a 5-gram language model using the ads from the Language Model dataset and use this model to discover these unexpected words. These words are then used as features to detect

Figure 3.4: Feature Selection Pipeline



Blue: Language model processing; Purple: Trafficking-10k processing; Green: Modeling

differences between the language used in legal escort service ads and suspected human trafficking ads. We assume that if the language model ads are primarily not potential human trafficking ads, then the words that are out of context are expected to be indicators of human trafficking. This is a valid assumption because previous researchers have found a human trafficking prevalence rate of 12% in escort service ads [64] and have primarily relied on contextual clues to identify human trafficking [59][61][62].

We estimated, filtered, and queried our language model using the KenLM Language Model Toolkit. The output was processed using a variant of code written by Dr. Michaeel Kazi. KenLM efficiently finds the probabilities and backoff penalties of n-grams (sequences of n words) from a language model [68]. It implements modified Kneser-Ney smoothing with interpolation to estimate perplexity scores of each word [69]. The complete pipeline used to estimate the language model is shown in Figure 3.5 which is originally from [1]. We summarize the methodology as discussed in [69] and [1] below.

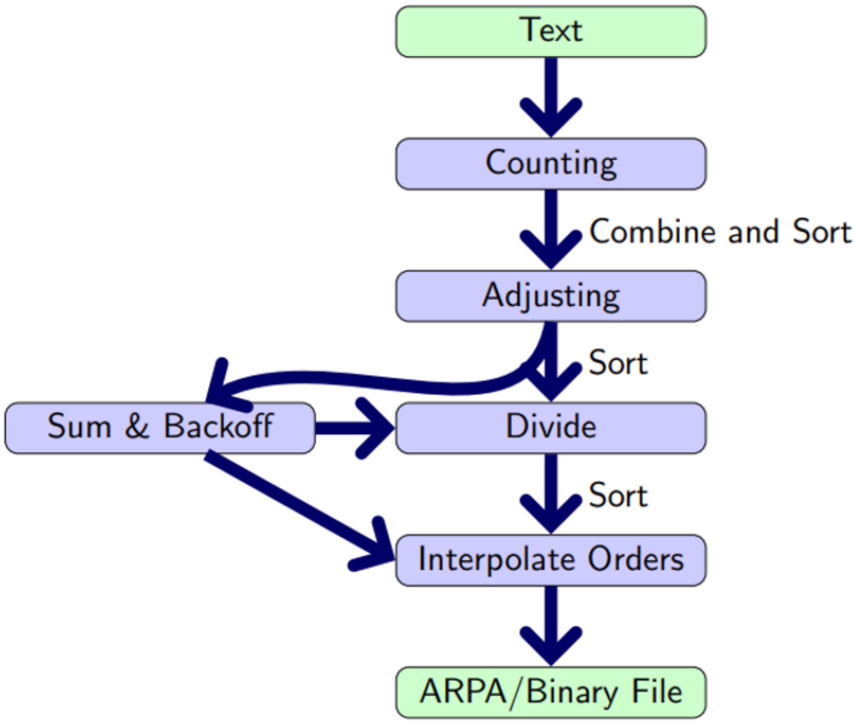
Language model of order n estimate $P(w_n|w_1^{n-1})$, the probability a sequence of n words (w_n) occurring given the previous $n-1$ words (w_1^{n-1}). To calculate this, the first step is to count all n-grams in the corpus. These counts, c, are then replaced with adjusted counts, a, per:

$$a(w_1^n) = \begin{cases} c(w_1^n), \text{ if } n = N \text{ or } w_q =< s > \\ |v : c(vw_1^n) > 0|, \text{ otherwise} \end{cases}$$

where v represents the number of unique words prior to w_1^n.

Smoothing statistics $t_{n,k}$ (the number of n-grams with adjusted count k) and discount $D_n(k)$ are also calculated at this time.

Figure 3.5: KenLM Pipeline from Heafield et al.'s Presentation [1]



$$t_{n,k} = |\{w_1^n : a(w_1^n) = k\}| \text{ for } k \in [1:4]$$

$$D_n(k) = k - \frac{(k+1)t_{n,1}t_{n,k+1}}{(t_{n,1}+2t_{n,2})t_{n,k}} \text{ for } k \in [1,3]$$

$$D_n(0) = 0 \text{ and } D_n(k) = D_n(3) \text{ for } k \geq 3$$

Next, KenLM normalizes the probabilities by computing pseudo probability, u, and backoff penalty, b, for unobserved events:

$$u(w_n|w_1^{n-1}) = \frac{a(w_1^n) - D_n(a(w_1^n))}{\sum_x a(w_1^{n-1}x)}$$

$$b(w_1^{n-1}) = \frac{\sum_{i=1}^3 D_n(i)|\{x : a(w_1^{n-1}x) = i\}|}{\sum_x a(w_1^{n-1}x)}$$

The final probability, p, is then calculated such that

$$p(w_n|w_1^{n-1}) = u(w_n|w_1^{n-1}) + b(w_1^{n-1})p(w_n|w_2^{n-1})$$

and

$$p(w_n) = u(w_n) + b(\epsilon)\frac{1}{|vocabulary|}$$

where ϵ denotes an empty string[69].

Using KenLM's methodology for computing these probabilities, we efficiently query the language model procure perplexity scores, PP for each Trafficking-10k ad, W, where $PP(W) = P(w_1w_2...w_N)^{\frac{-1}{N}}$ [65].

This results in each token (word or phrase) in each ad being assigned a perplexity score (with a maximum possible score of 0). More negative scores are more "perplexing" – they occur in unlikely n-grams. We also calculate a "context" score to estimate the likelihood a token is in a real sentence. The context score is an average of the perplexity scores of the k tokens to the left and k tokens to the right of the token in question. We use $k = 5$. Observe that this does not include the perplexity score of the token in question. Similar to perplexity, we can interpret tokens with lower context scores to represent tokens that are less likely to be a part of a real sentence because their surrounding context is more unexpected. In addition, KenLM calculates the total perplexity score per ad that is equal to the sum of the scores for all the tokens in that ad and the total number of tokens that were out of the vocabulary (oov) of the language model.

The Trafficking-10k tokens using the baseline language model, with no phrases included, has an average perplexity score of -1.73 with a standard deviation of 1.59 and an average context score of -1.58 with a standard deviation of .74. The perplexity and context scores are only correlated with a Pearson coefficient of .37. There are on average 3.47 tokens out of vocabulary but there is a standard deviation of 32.0. This is because of about 30 ads with over 100 tokens found to be out of vocabulary. This inconsistency resulted primarily from pre-processing error that kept some html tags. It exemplifies the difficulties in working with non-regular text.

Using the language model and phrase detection that allowed for the most accurate detection of human trafficking risks, the perplexity score on average is -2.03 with a standard deviation of 1.57 and an average context score of -1.84 with standard deviation .71. The perplexity and context scores under this model are slightly less correlated with a correlation coefficient of .35. This language model has fewer tokens out of vocabulary, with an average of 1.39 and a standard deviation of 2.63. Under a t-test, the differences between perplexity, context, and oov in the baseline and best language model results are all significantly different at a p-value $< .001$.

Feature Selection and Representation

After training and querying the language model, the next step in our feature selection pipeline is to calculate the average perplexity or context score for each unique token (word or phrase) in the Trafficking-10k ads. We then experiment with choosing the tokens with the highest or lowest average scores across varying ranges. The lowest scores correspond to the most perplexing or out of context tokens. We additionally experiment with the usefulness of phrase detection algorithms by comparing three methods: selecting the top words without phrases, selecting the top tokens with phrases, and selecting the top words and the top phrases separately. These selected tokens become the keyword features used in our model.

Each ad is represented by a vector of these keywords. Although the keywords were selected by context or perplexity scores, we experiment with different numerical representations. The vectors were represented with either the TFIDF, frequency, perplexity, or context scores of each keyword in the ad. These representations test whether given a list of keywords, number of occurrences (frequency), level of strangeness (perplexity/context), or a combination of the two (TFIDF), will result in the most accurate predictions.

We also experiment with additional filters to remove unimportant words. First, we remove tokens that are too short, which we define as words that are fewer than three characters long.

These are "words" that likely resulted from processing error, where we did not identify letters that were separated from one another. We also run models with only tokens that are both low in context and low in TFIDF score in order to further reduce the potential noise caused by overly common words. The final filter removes all non-sparse terms, which we define as words in less than 2% of all documents.

In addition to these tokens, we include features describing language characteristics of the advertisements: number of phrases, number of words, number of characters, percent of words that are phrases, total perplexity, total out of vocabulary (oov) words, and the sum of the NLP (e.g. perplexity, tfidf or context) features.

We compared this pipeline to two simpler techniques, where all but the sparse tokens are kept. After feature selection, each ad is represented as a vector of the frequency or TFIDF of the tokens that remain, as calculated by R's tm package [70]. This matrix of word vectors is used to train our models. The level of sparsity removed is a parameter that can significantly affect accuracy and utility of the model. Keeping too many sparse words may result in too much noise and a long list of "keywords". On the other hand, removing too many sparse words may result in removing actual keywords and subsequently reducing model accuracy. As a result, we experiment with varying level of sparsity to tune the model. We find that models using TFIDF scores and removing words that are in less than 2% of documents have consistently better predictive performance than the corresponding frequency based models so we will focus our discussion on the TFIDF-based model results.

3.4.5 Human Trafficking Detection

In order to build a model for human trafficking detection, we applied this feature selection pipeline to the Trafficking-10k data. The multiclass risk annotations were converted to a binary system of high- vs low-risk, with the mid-level "unsure" class removed. We trained the models using a five-fold cross validation approach. The words/phrases used were only those that were identified in the training dataset. As a result, the testing data so often missed words in the training data. These missed words were added as having a score of 0 to the testing set. Using this data we experimented with multiple machine learning methods. We applied clustering methods, but due to their poor initial performance, we instead focused our modeling efforts on binary logistic regressions, classification and regression trees, and random forests.

We applied binary logistic regressions (logit) for its simplicity and interpretability. Not only does it classify the ads but it also allows for a clear ranking in likelihood of human trafficking. Logit is especially useful as the coefficients can be easily interpreted to understand whether a word increases or decreases the likelihood of an ad being human trafficking. In addition, it is often better than probit at modeling the effects of extreme independent variables, where one feature may significantly impact classification. This is important in an application like human trafficking, where the presence of a particular keyword may be a near guarantee that an ad is human trafficking related.

Unlike probit (as discussed in Chapter 2.5), logit assumes a logistic link function such that:

$$\log(\frac{\pi}{1-\pi}) = \beta_1 + \beta_2 x$$

[35]. Like probit, logit also assumes independent observations and little multicollinearity between variable. We propose that these assumptions are valid because we have removed duplicated and advertisements. We further ensure that there is no multicollinearity between variables by only including one set of language model features in the model, only perplexity, context, TFIDF, or context. Because of the large number of features, we do not apply stepwise model selection to reduce the number of features. We also do not include interaction terms as that would significantly reduce the automation of the model building. However, the inclusion of phrases in later models, while also reducing multicollinearity, could be interpreted to represent the interaction of the words within the phrases.

Classification and regression trees (CART) were also used for similar reasons to those discussed in Chapter 2.5. It is interpretable and requires fewer assumptions than regression analysis. Yet, large trees are sometimes still difficult to interpret and may cause it to be difficult to decipher the effect of keywords. Like the aforementioned section, we implement R's Rpart.

Finally, we also train models using random forests because we have a suitable amount of data (50 times more than the food safety project). We use an R implementation of Breiman and Cutler's random forests from the package "randomForest" [71]. Our forest is built using the default 500 trees.

Random forest is the least interpretable of the aforementioned methods. The output model allows back end analysis of the importance of features but does not allow users to decipher whether a keyword is a high or low risk indicator. Its primary benefit, as Breiman wrote, is that it is more robust to noise than regular trees and regressions and is less likely to overfit the data [72]. Although random forests provide out-of-bag estimation, in order to have comparable experiments as the other algorithms, we still apply five-fold cross validation and tune within each fold for the number of variables randomly sampled at each split.

Additional techniques are applied to gain more interpretable and predictive results. We experiment with further reducing the keyword list by choosing the most "important" words per the results of the best random forest model. We reduce this list by choosing features with a mean decrease in Gini index greater than the average decrease across all features. In order to understand precisely how these features affect human trafficking risk, we build a logit model using these reduced features. We hypothesize that this model allows us to reduce most of the noise from unimportant words while still keeping the important, desired keywords.

These models provide a verification that our unsupervised keyword selection methodology is informative. Tokens are selecting using a language model built on generic advertisements and chosen by likelihood of occurrence. Rather than manually collecting the attributes that are indicative of human trafficking, our methodology can discover the keywords far less laboriously. Most importantly, our keyword detection pipeline indicates that keywords can be selected, accurately and without supervision, by examining out of context and rare occurring words. They are discovered solely based on the results of NLP characteristics. If keywords are correctly identified, the models will be able to make accurate predictions of human trafficking risk. Therefore, with this methodology, users can apply interpretable, automatic, and data driven methods of selecting features for predicting human trafficking.

3.5 Results

3.5.1 Best Model Overview

Our interpretable language model pipeline has significantly better predictions than not just the unimodal HTDN model but the multimodal HTDN model as well. For the most part, the more words that are included, the better the model performs. However, there are diminishing returns in model improvement and in some cases we find that too many words causes too much noise, especially in the logistic regression models. We tested all models with up to 1000 tokens and 100 phrases, and also tested a few pipelines with up to 1500 tokens. We found that using over a thousand tokens offered insignificant improvements to predictive performance (F-1). 1000 words is about 5 to 6% of the unique tokens in any given training set (80% of the full data).

Our best pipeline, Random Forest and Logistic Regression Model (RFLM) has the best predictive performance out of all the pipeline variations we experimented with. We assume that the best model is the one that is most applicable to end users: it provides a precise list of keywords and high predictive capability. We also assume that a model that is better at identifying high risk advertisements is better than one that is more precise in its identifications, given equivalent F-1 scores. As a result, we determine that the RFLM to be the best model.In addition, we find that it has a significantly higher F-1 score than not just the unimodal HTDN model, but also the multimodal HTDN model with a p-value $< .001$ using a one-sample, one-sided t-test (since no standard deviation was given for the HTDN models). This is shown in Table 3.1.

Table 3.1: Top Model Results

Method-Features	Average # of Features	F-1(StDev)	Recall(StDev)
HTDN-Unimodal	N/A	.658	.623
HTDN-Multimodal	N/A	.665	.622
Human baseline	N/A	.737	.709
RFLM	**223**	**.667(.003)**	**.729(.006)**

StDev (Standard Deviation) is calculated from one hundred iterations of the pipeline across different random splits
RFLM: random forest with logistic regression model trained with low context words
HTDN and human baseline models are results from [55].

We will further discuss RFLM's feature selection technique, key findings, comparable pipelines, and alternative modeling methods in the following sections. The results of additional pipeline variations are in the Appendix in Table A.1.

Feature Selection

RFLM is the pipeline with the highest F-1, recall, and interpretability out of all our experiments. However, it is the most computationally complex pipeline of the ones we experimented with. It applies a two phase modeling technique. First, it begins by selecting the 1000 and 25 lowest context words and phrases, respectively. Then it reduces the list of keywords by choosing only words that are three or more characters long while keeping the sparse words. Next, it adds in three language features: the total oov, total TFIDF, and total perplexity in each ad. Finally, it trains a random forest model. Using the model output, it further reduces the list of keywords by only keeping tokens that have an above average decrease in Gini Index in the random forest model. This results in an average of 220 tokens along with the language characteristics as features for the next modeling phase. The logistic regression model is trained on these remaining features and is the final model used to predict accuracy and identify human trafficking indicators.

Including phrases in our language model does improve the overall detection model. They consistently perform better than the comparable models without phrases even though the phrases are not consistently found to be statistically significant in the final logistic regression model. After testing a range of phrase lengths from three to seven words and cutoffs of between ten to twenty occurrences, phrases using three to six words and a minimum threshold of twenty occurrences causes the best improvement in model performance. RFLM keeps on average three out of the twenty five initial phrases after the random forest phase. Although some of them like the phrase "I love what I do" (a positive indicator of human trafficking) are significant when used, none of the phrases are significant and present consistently across trials. This indicates that most of the phrases are not clear indicators of human trafficking though they still help in identifying the other keywords by influencing context scores and reducing multi-collinearity in the models.

Only a few overall language features were found to be useful features in our pipeline. Number of characters, phrases, and words, despite having statistically significant differences between the two classes, ultimately add too much noise and decrease model performance across the majority of the modeling pipelines. Total perplexity, total TFIDF scores, and words out of vocabulary (oov) are kept in RFLM, although oov is not statistically significant. More perplexing words surprisingly decrease the risk of an ad being human trafficking at a significance level on average of $< .03$ and higher total TFIDF scores reduce the risk of an ad being classified as human trafficking with a p-value of $< .001$.

Key Findings of RFLM

As previously discussed, the RFLM pipeline allows users to identify a list of about 220 keywords each iteration. Over the course of a five-fold cross validation test, we identify 257 keywords total, as not all the same keywords are used in each model. To identify the list of true keywords, we take the average of the coefficient values and statistical significance of each word across the five-folds logistic regression model output. The coefficient values of a logistic regression describes the amount and direction of influence a feature has on human trafficking. Features with positive coefficient values are high-risk indicators, while features with negative coefficient values are low-risk indicators. The statistical significance of a feature describes the probability that the true coefficient value is 0. The coefficient is assumed to be 0 and the significance is assumed to be 1 for words that are not included in an iteration.

Although the majority of the words are not significant across all models, most of them are usually human trafficking indicators. 61 (24%) of these words (and 0 phrases) are found to be significant at a p-value of $< .1$ on average. 182 of all the words and phrases (71%) are high risk indicators, while 48 (79%) of the statistically significant words are high risk indicators. This implies that once non-important words are removed (e.g. those that have below average decrease in Gini index in a random forest model), most remaining low context words are potential high-risk indicators.

Previous literature have discussed that race, youthfulness, and restricted movement are known to be indicators of human trafficking. Our keyword list include such words and also identifies low risk indicators. Although not many low risk indicators in total are identified in our keyword list, the ones that are identified are informative and distinctly different from the human trafficking indicators. As demonstrated in the keyword examples shown in Tables 3.2 and 3.3, RFLM is able to identify known indicators of human trafficking and separate them from legal or voluntary sex work, while also picking up potentially new keywords.

Table 3.2: Select Human Trafficking Keywords

Keyword	Beta	Significance
Asian	.563	<.001
Korean	.459	<.001
Young	.381	<.001
Japanese	.310	<.001
Girls	.271	<.001
Chinese	.198	.03
Incalls	.187	<.001
Tight	.165	.001
Petite	.161	<.001
Slim	.159	.005

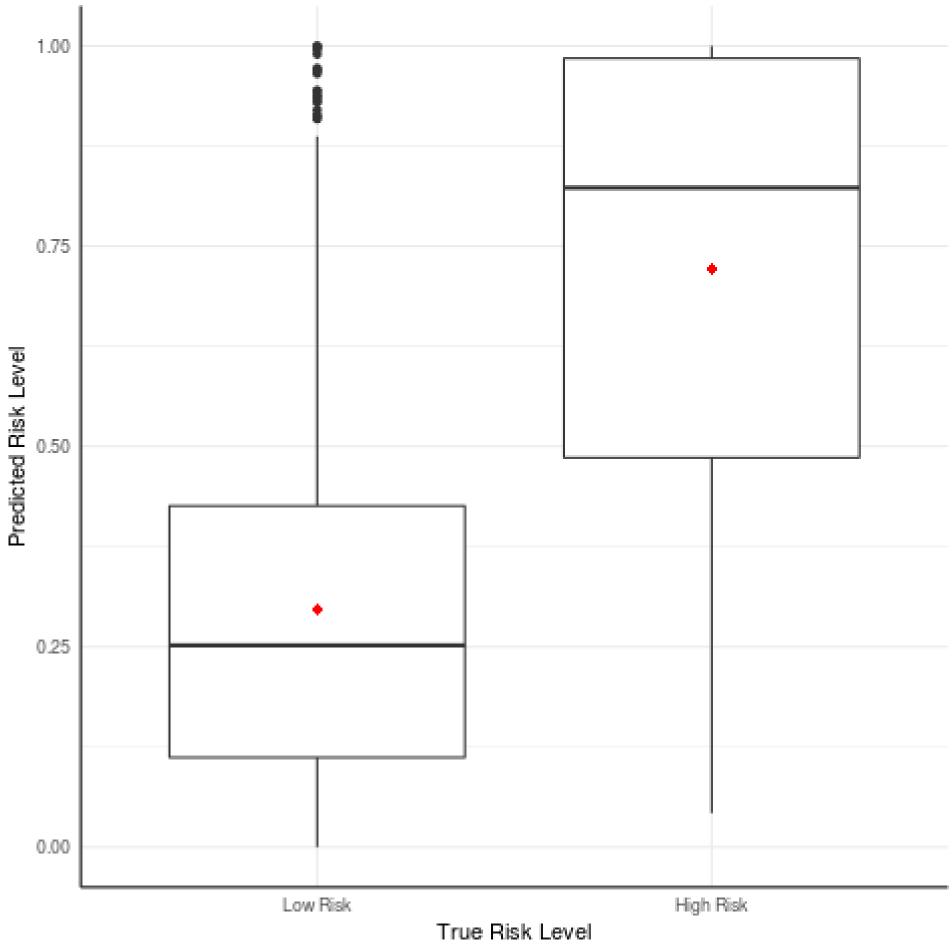
Table 3.3: Select Non Human Trafficking Features

Keyword	Beta	Significance
Sex	-.644	<.001
Details	-.396	<.001
Mature	-.188	.001
Sensual	-.187	<.001
Woman	-0.174	.003

Our final model is still 6% away in F-1 score from achieving the maximum possible accuracy which is defined by the human baseline from [55]. As shown in the box plot of predicted risk levels in Figure 3.6, there is a clear distinction between the predicted probabilities and the true binary risk levels. In addition, Figure 3.7 demonstrates that our model is especially accurate in classifying ads that human annotators are not completely certain of and may otherwise spend more time on. The high predictive capacity and interpretability of our model has the potential to significantly improve efficiency in human trafficking detection.

The mis-classifications in our model seem to arise primarily from the ads that were “certainly yes”. This is shown in Figure 3.7. An ideal model would have consistently rising probabilities as the risk levels increase across the six classes. Since we trained a binary classifier, these probabilities do not naturally extend to the multi-class annotations. Most of our high risk advertisements are annotated as "weakly yes" and accordingly our model is best at predicting these high-risk ads correctly. The miss-classifications of the "certainly yes" ads demonstrate that the human trafficking indicators are not consistent across classes. However, there are not enough ads that are "certainly yes" for our model to learn the characteristics that discriminate it from the low-risk ads. Therefore, more data on "certainly yes" and other higher risk ads are needed in order to have more accurate binary predictions.

Figure 3.6: Predicted Risk Level Against Binary True Risk Level

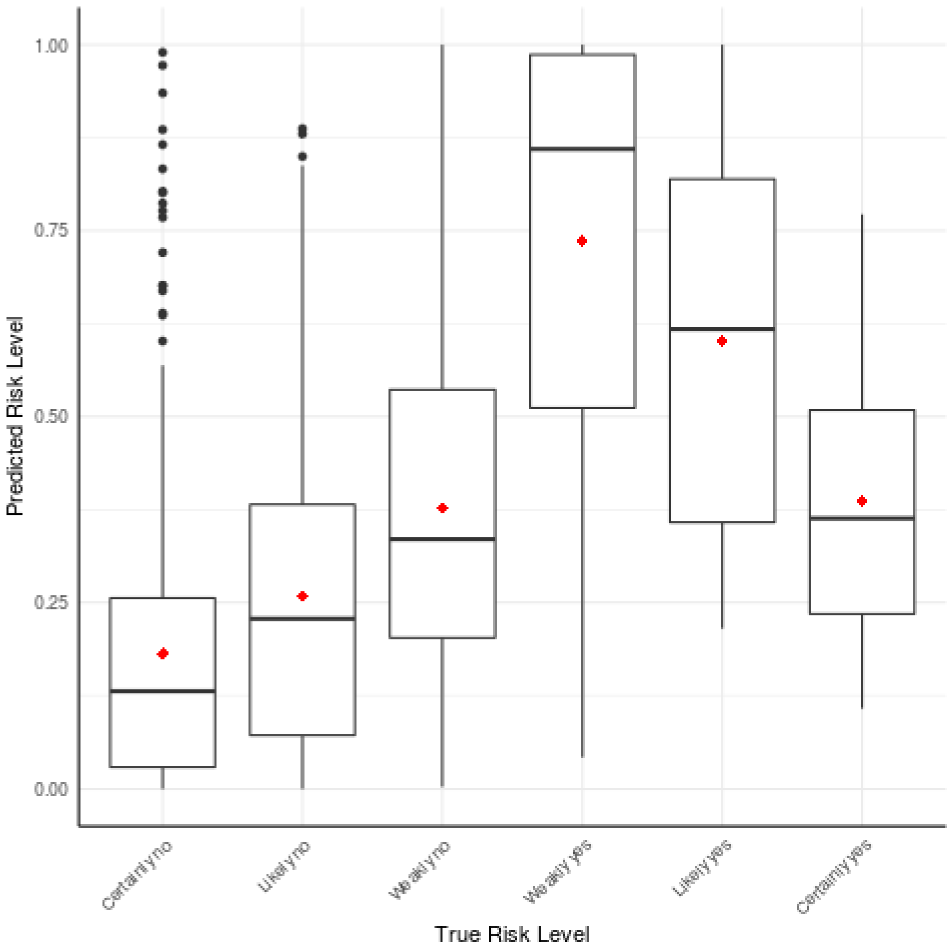


3.5.2 Alternative Pipelines Overview

In addition to RFLM, we discover three additional pipelines that have comparable F-1 scores that are also significantly better than both the unimodal and multimodal HTDN models with a p-value $< .001$ across the board. Their predictive results are shown in 3.4. None of these four models are significantly different from one another in F-1 score at p-values $< .05$. However, we do find that as complexity of the model increases, so does the recall. As a result, RFLM is the best model, but at the cost of increased complexity.

As can be seen in Figure 3.8 and the previously mentioned F-1 t-tests, the four top feature selection techniques do not have significantly different predictive performance. However, RFLM has significantly higher recall scores than all the other models, with a p-value of $< .001$. In addition, it relies on the fewest number of keywords in its final model (on average 220 as oov,

Figure 3.7: Predicted Risk Level Against Actual Risk Level



plex, and total TFIDF score are also kept). Its downside is that it relies on starting with 1000 tokens and 25 phrases, and even after filtering, users are still left with 955 features to build the random forest. As a result, if no labeled data is available, it would not be the best model. Its keyword list is too long to be practically applicable. Instead, we would consider the Low Context Logistic Regression Model (LCLM) the best because it has the next highest significantly different recall rate and also has the second smallest list of keywords for users to contend with (332 on average).

LCLM uses 332 tokens on average, total perplexity, and oov words to predict human trafficking risk levels. Like RFLM, it starts by choosing the 1000 lowest context words and 25 lowest context phrases. It then removes all words that are too short and sparse. This final list of 332 words and phrases and its language characteristics are used to train a logistic regression model. LCLM, like RFLM, is more computationally expensive but uses significantly fewer words than TFIDFS.

TFIDFS is our least complex pipeline. It is a simple bag of words logistic regression model. It uses all words that are in more than 2% of the advertisements and longer than three characters, which results in a keyword list of 414 words. The TFIDFS model uses fewer words than the the Low Context High TFIDF (LCHT) model.

LCHT uses the 1500 and 25 lowest context words and phrases, respectively that overlap with

Table 3.4: Alternative Model Results

Method	Average # of Features	F-1(StDev)	Recall(StDev)
TFIDFS	416	.667(.003)	.599(.003)
LCLM	334	.667(.003)	.606(.003)
LCHT	569	.667(.004)	.605(.005)
RFLM	223	.667(.003)	.729(.006)

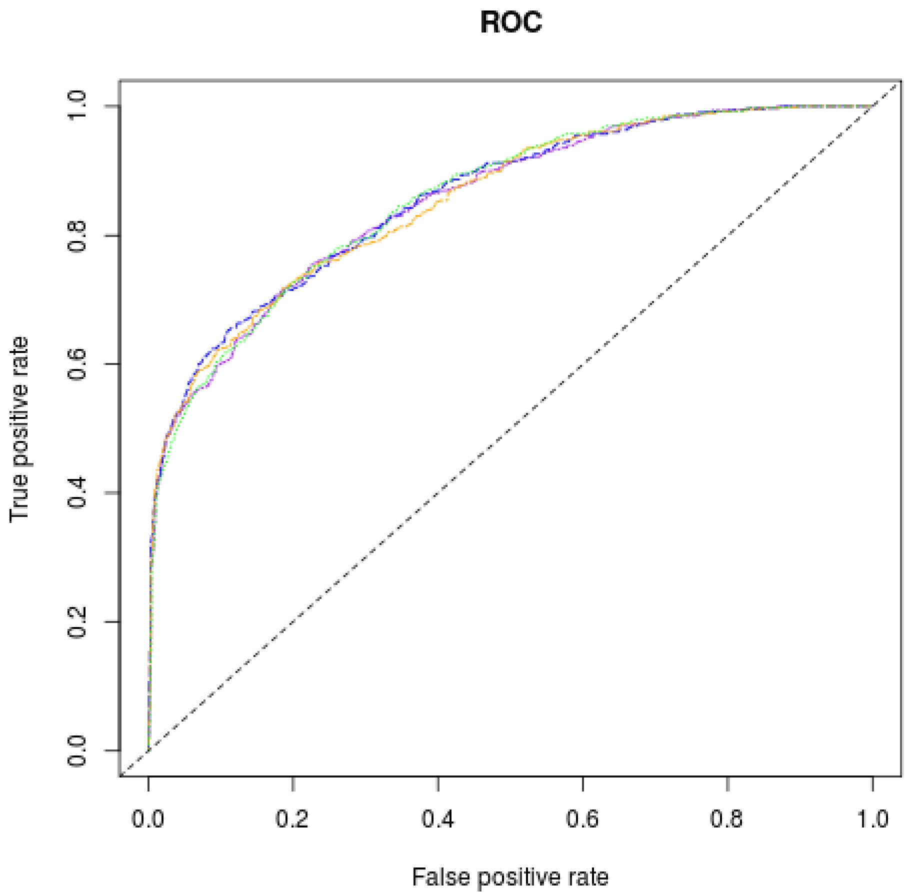
StDev (Standard Deviation) is calculated from one hundred iterations of the pipeline across different random splits
TFIDFS: logistic regression model trained with all non-sparse words
LCLM: logistic regression model trained with low context but non-sparse words
LCHT: logistic regression model trained with low context and high TFIDF words
RFLM: random forest with logistic regression model trained with low context words

the 600 and 25 highest TFIDF scored words and phrases, respectively. It, like the other models, does not include words that are too short but does include its language characteristics (total perplexity, oov, and TFIDF). It keeps the sparse words because removing sparse words significantly worsens predictions. LCHT does have a significantly higher recall rate thatn TFIDFS, which indicates that the additional sparse words may allow the model to be less granular in classifications. All four pipelines represent ads by a vector of their keyword's TFIDF scores.

Phrase detection improves model performance in all but the TFIDFS pipeline. The phrases are too sparse to be included as a keyword and do not affect the TFIDF score of the remaining words, so TFIDFS' model results are exactly the same with and without phrases. However, for the other three models, phrase detection does significantly influence the words selected. After querying the Trafficking-10k ads with a language model built using phrase detection, there are significant changes to the average perplexity and context scores of each word. Using t-tests, we find that the differences between the results of a language model with and without phrase detection are significantly different to a p-value $< .001$. Like RFLM, LCLM and LCHT are trained using phrases lengths of three to six words and a minimum threshold of twenty occurrences. These phrases are rarely found to be consistently significant in the final models.

Ultimately, our models demonstrate that keywords can be identified by selecting unexpected tokens, especially those that are less likely to be in cogent sentences (low contextual features), but are not too rare. Simply modeling using non-rare words can have relatively accurate performance but focusing further on low context words appears to be more accurate. Low context tokens are ones that are surrounded by tokens that are not in the expected n-grams. Tokens in the expected context have very poor predictive performance when used in models (< 1% F-1 see Table A.1) . Both "perplexing" and not "perplexing" words have similar performance with enough phrases and words included. If phrases are not included, words of high perplexity are not good indicators for classifying human trafficking. These words are too rare to be useful. We also find that low perplexity tokens are often times too common to be useful as well. Therefore, if training data is unavailable to run the RFLM pipeline, we find that focusing on low context but non-sparse words, like LCLM will provide a useful keyword list.

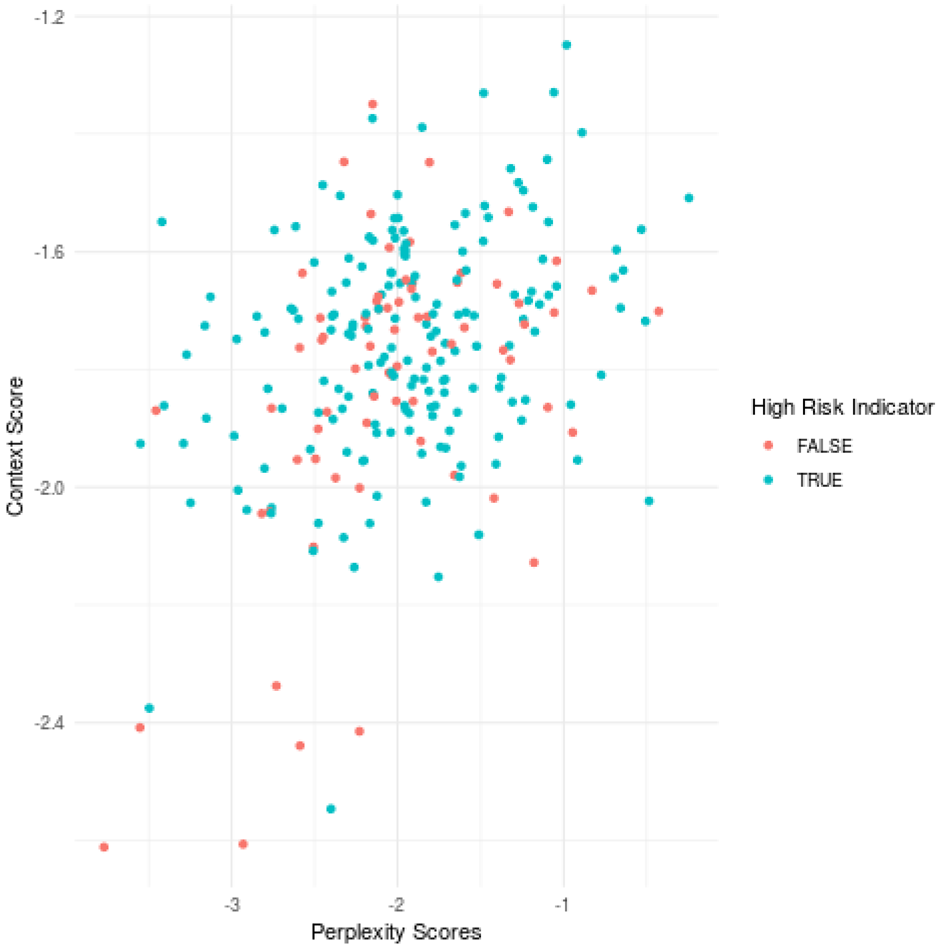
Figure 3.8: ROC by Feature Selection



Purple = RFLM, Blue= TFIDF, Orange = LCLM, Green = LCHT

Although selecting words using a language model may be better than selecting words by frequency and TFIDF, we did not find a direct correlation between a word's average context or perplexity score across documents and its risk indication, as shown in Figure 3.9. This figure shows the average context and perplexity scores for key words used in our models and codes the token as either a high or low risk indicator based on their coefficient from RFLM's logistic regression. There is no visible pattern in this graph. In fact, the correlation coefficient between context scores and β values is .122. However, this is a higher correlation than that of TFIDF scores and risk, which is .013. Despite this lack of correlation, TFIDF is still the most useful characteristic, over count, perplexity, or context scores to represent the word vectors. This demonstrates the complexity in keyword identification and the necessity for supervised models. Although, a keyword list can be created, language features are not directly related to human trafficking risk levels. As a result, models are the only way to accurately separate high-risk from low-risk indicators.

Figure 3.9: Language Model Scores and Risk Level



Modeling Methods

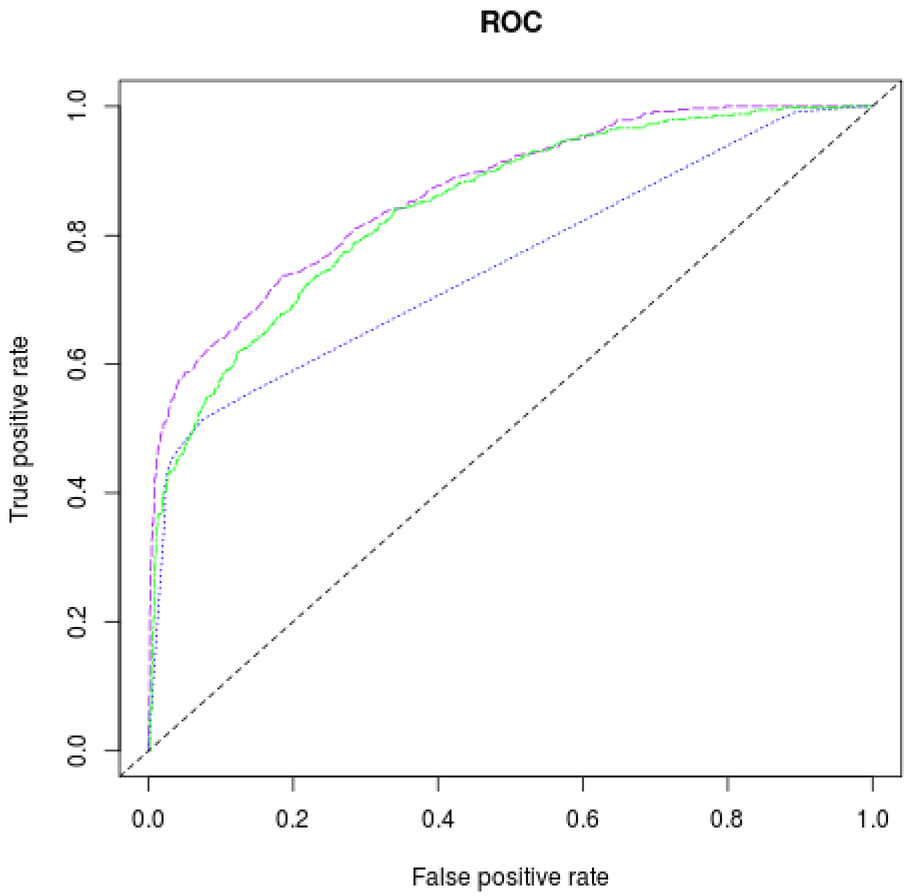
Our experiments found that logistic regression and random forest models have the best predictive performance. A comparison of the performance of the modeling techniques discussed in our methodology section is shown in the ROC graphs in Figure 3.10. It demonstrates the superior performance of random forests and logistic regression in this application. All three models shown were built using the same 1000 low context words and 25 low context phrases. The random forest model has higher sensitivity (true positive rate) initially but otherwise appears to have similar results as the logistic regression model. On the other hand, as exemplified in the ROC graph, the CART models consistently have the worst predictive performance.

Despite the number of potential features, after pruning, the CART models result in very simple models. One example is shown in Figure 3.11. These results show that although CART may capture a few key indicators, it is unable to capture the greater complexities and more obfuscated terms in human trafficking ads. Certain features do not make it into the final model despite being human trafficking indicators, because they do not provide sufficient improvement in probability to warrant the increase in tree depth. This indicates that CART can not discriminate effects when given a large number of potential keywords and a small minority class. However, with more training data, CART model may be able achieve more comparable performance.

Random forest's performance without dimensionality reduction is on par with logistic regressions. As shown in the ROC curve, random forest models initially have a higher true positive rate at lower false positive rates, so have a slightly higher AUC (Area Under the Curve). However, across five-fold cross validation tests, random forest models usually have lower recall than logistic regression models. This indicates that random forest models are less adept at identifying which ads are high risk than logistic regressions. The random forest models do tend to have higher precision. This indicates that of the ads that are predicted to be high risk, they are more likely to be correct than the logistic regression models. These results can be seen in Table A.1. In addition, the random forest models have generally lower F-1 scores than logistic regression models across five-fold cross validation tests. As a result, we would only recommend a random forest model if precision is of greater importance than clarity or recall.

Using the same data, logistic regressions do have a higher F-1 than random forests. This may indicate that the data does in fact generally fulfill parametric assumptions; the frequencies of various words are linearly increasing or decreasing risk indicators . Logistic regression models are also more interpretable. Users are able to understand exactly how features influence risk level via the beta values and significance test. With our best model, we are able to identify a list of about two hundred keywords and sort them to demonstrate key indicators of human trafficking. As a result, although in the ROC graph in Figure 3.10 the random forest model may have the better curve and AUC, we conclude that logistic regressions provide the better model in the holistic context of human trafficking detection.

Figure 3.10: ROC by Modeling Method



Green = Logit, Blue = CART, Purple = RF Using the 1000/25 Words/Phrases in the Least Context

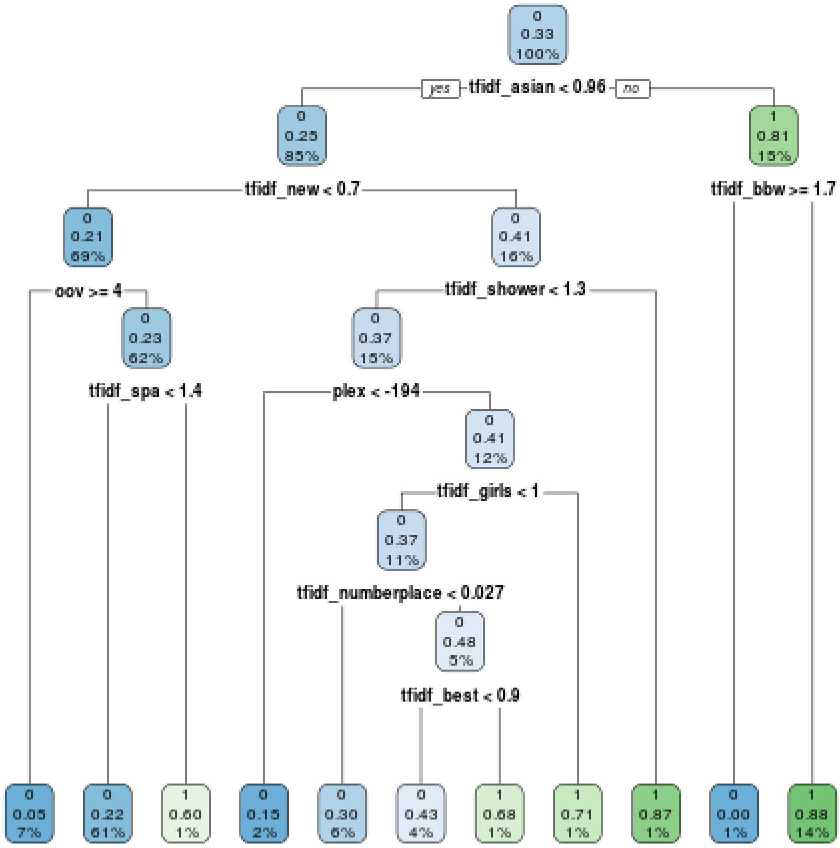
3.6 Application to Organization Detection

Our language modeling pipeline can be even more beneficial when applied in conjunction with organization detection. In this section, we discuss the results of RFLM when applied to known human trafficking organization advertisements and unknown organization detection methods. We further verify our model's accuracy by checking if the detected organizations involve the movement of persons.

3.6.1 Application to Known Human Trafficking Organizations

First, we applied our model to advertisements from known human trafficking organizations from outside the Trafficking 10k dataset. We extracted advertisements from another set of *Backpage.com* advertisements that matched locations related to two major human trafficking organizations. One is from the infamous bust of a sex trafficking ring that supplied "johns" such as Patriots owner Robert Kraft and was sourced from massage parlors across Florida. These advertisements were identified by matching the names of the offending massage parlors that were reported in the news [51]. The second case was a ring that is being indicted in Oregon

Figure 3.11: Best CART Model Results



but spanned the United States, Canada, and Australia and operated under the guise of escorts. We matched this ring with the advertisements using contact and web information that were disclosed by the U.S. Justice Department [52]. This resulted in 53 advertisements from the Florida case and 437 advertisements in the Oregon Case. Although these organizations are suspected to be human traffickers by law enforcement, at the time of writing, the organizers have not yet been convicted for human trafficking. In addition, it is important to consider that not all the advertisements tied to these organizations are necessarily advertisements for sex related work; they may also engage in entirely legal or voluntary work.

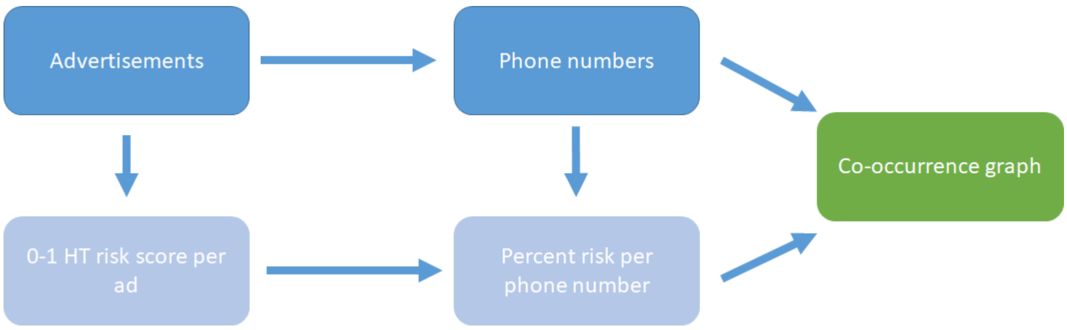
While the Florida case is very localized, the Oregon case clearly involves significant human movement across the United States and Canada. This movement is a key indicator of human trafficking. Using RFLM we are able to detect that all of the Oregon case advertisements are high risk, despite none of the ads being in the Trafficking 10k set. However we only detect that 12 out of the 53 (22%) of the Florida case advertisements are high risk human trafficking advertisements. The average risk across all the advertisements are .973 and .333, respectively for the Florida and Oregon cases. It is expected that not all advertisements are considered high risk since the locations are also licensed massage parlors.

Upon manually verifying the ads, we find that the high risk ads include many of the known indicators of human trafficking, like youth, “new” and “slim”, and ethnicity while the low risk ads do not. The lowest risk Florida ad simply describes the massage parlor's prices, services, and contact information. Unlike the other ads there are no descriptions of the "masseuse". As the majority of the Florida case advertisements were not detected to be human trafficking, organization detection would be significantly beneficial. It would allow users to link the ads that are low risk to the ads that are high risk and thus gather more information on the potential personas involved. Therefore, if the advertisements were coupled into an organization detection pipeline, users would be able to identify that even though certain advertisements are low risk, the entire Florida and Oregon case rings are likely to be human trafficking related.

3.6.2 Application to Unknown Organizations

With the intuition that phone number matches across advertisements are indicators of an organization, we construct a phone number co-occurrence network across 124,856 *Backpage.com* advertisements supplied by Marinus Analytics using code developed by the Lincoln Laboratory HDDN team. We scored these advertisements using RFLM and used the average raw probability score for a given phone number as the percent risk that any given phone number node is human trafficking. This pipeline is shown in 3.12 and the final network is shown in Figure 3.13. This network is made up of 1574 nodes, 2342 edges, with an average of 2.98 degrees. The nodes have an average probability of 39.9% of being human trafficking, with a standard deviation of 26.3%. 30.6% (482) of the nodes are classified as high risk (probability of over 50%).

Figure 3.12: Graph Creation Pipeline



From this network, one can see that there are is one cluster that is heavily considered (very red) to be human trafficking but there are no clusters where all nodes are high risk of being human trafficking. There are also many small organizations that appear to be made up of advertisements with a low risk of being human trafficking as well as isolated high risk nodes. As certain organizations may be more discrete in cross referencing contact information, organization detection algorithms using template matching like the ones developed by Lin et al. [73] would be useful in determining if these nodes are part of a larger, but obfuscated organizations. This would allow more accuracy in detecting human trafficking organizations. For the sake of this study, we focus on the known organizations detected by phone number co-occurrence.

We can further split the network by the low risk (probability $< .5$) and high risk (probability $>= .5$) nodes. These networks are shown in Figures 3.14a and 3.14b. From these graphs, one can see that there high risk nodes are less common but there is one complicated web of high risk nodes that we could also see in the full network graph. The high risk nodes otherwise appear prone to star clusters, where one node centers a group of protruding nodes. However, there are also a few star clusters found in the low risk network. We do not find significant correlation between the percent risk of a node being human trafficking and various other node attributes, like degree and betweenness centrality the distinctions are not yet sufficiently clear between the low and high risk graphs to make conclusive judgments about the structural characteristics of suspected human trafficking organizations.

As demonstrated by the Florida cases, human trafficking organizations may include legal or obfuscated advertisements alongside their detected sex trafficking advertisements. As a result, a more useful graph is one that connects the suspected organization with any related but unsuspected nodes as in Figure 3.14d. The resulting cluster provides user with a fuller understanding of the organization in question, potential suspects, and likely contacts. In this network we only kept connected components of three or more nodes and also connected all the remaining nodes to its original neighbors in the full network.

Using this methodology we detect a total of 18 organizations with an average probability of .672 of being human trafficking. These 18 organizations encompass 35,047 ads that users may otherwise have had to manually sift through. The various detected organizational sizes and probabilities are shown in table 3.5

Table 3.5: Detected Organizational Size and Probability of Being Human Trafficking

Number of Ads	Number of Nodes	Probability
7029	116	.439
882	27	.782
6016	253	.451
9405	204	.796
654	38	.487
1292	22	.730
470	14	.522
762	10	.723
526	8	.573
483	9	.713
661	8	.885
623	9	.516
487	9	.677
717	5	.847
2905	22	.298
464	8	731
751	3	.993
920	5	.926

As shown, after factoring in the connected low risk nodes, not all of these advertisements have a probability of $> .5$ of being human trafficking. This demonstrates a potential weakness in our model. However, we can verify the efficacy of this methodology in detecting human trafficking organizations by checking if they, like the Oregon case, are related to significant movements of persons. The lowest probability cluster (2905 advertisements, 22 nodes, and probability of .298) we find to have minimal movement. It is centered in Florida. Although this organization may be like the Florida case, it does remove one significant indicator of human trafficking and resultantly justifies the low score that our model assigned.

On the other side of the coin, we mapped one of the high probability large clusters (9405 ads, 204 nodes, and probability of .796) and detect a significant amount of movement across all of the United States. As shown in this example, RFLM did indeed succeed in detecting a likely human trafficking organization.

Like the previously discussed Florida and Oregan case advertisements, the ads associated with this cluster include commonly recognized indicators of human trafficking: youth, race, and movement. On the other hand, when looking at the nodes that are not considered to be part of the high risk connected network, we result in the network shown in Figure 3.14c. It has an average probability of .269 of being human trafficking. It is also very dispersed but still has a few large clusters.

Upon further examination of the two largest clusters (110 and 80 nodes) we find that their ads truly do not have many indications of human trafficking. In fact the 80 node cluster is associated with escort service reviews rather than the advertisements themselves. This demonstrates the difficult in accurately web scraping and processing the data because these reviews should not have made it into the final dataset since our training data does not have any reviews.

The 110 node cluster is clearly part of a larger organization per the websites and businesses referenced in the ads. However, this organization is centered in Ottowa and Montreal, which is not a significant amount of personnel movement to indicate human trafficking. In addition, the ads mostly describe one woman, and although they describe her appearance, they do not describe race or age. They also do not have indicators of restricted movement or recent movement. Furthermore, unlike the previously identified human trafficking advertisements, these ads have minimal emoji usage and other text obfuscation techniques. Their language appears to be closer to standard English. As a result, these indicators, or lack thereof, further justifies our model in identifying this cluster as a low risk human trafficking organization.

3.7 Discussion

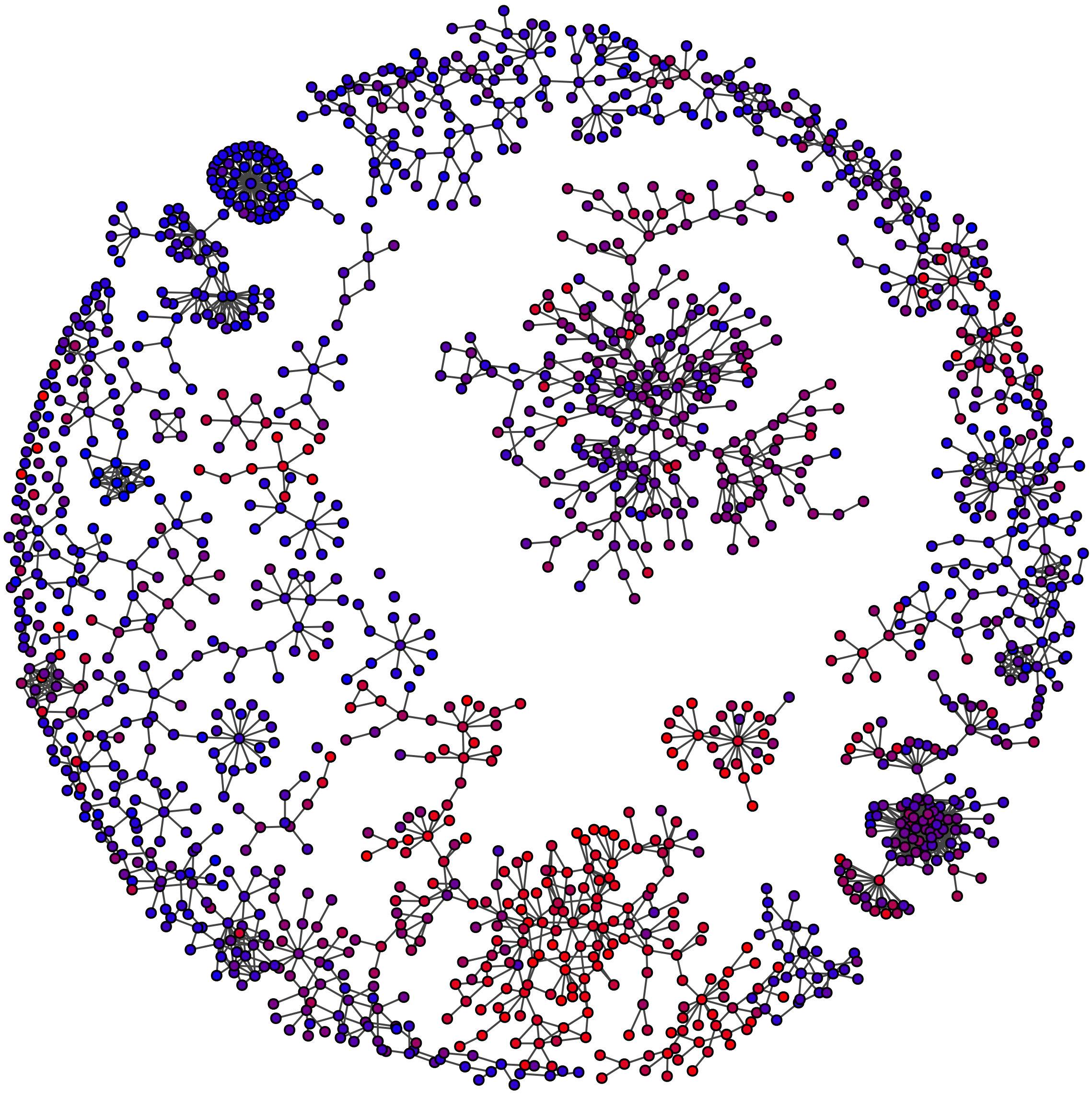
The results of this study offer four main contributions. Most importantly, our human trafficking pipeline can significantly improve efficiency in human trafficking detection and performs on par with human experts. Its predictive performance is better than the unimodal HTDN by more than one standard deviation. It can also have better performance than the multimodal neural network model, HTDN [55]. It achieves all this without factoring in personally identifiable information, emojis, and images. With a combination of language modeling, phrase detection, random forests, and logistic regressions, we are able to accurately identify keywords of high risk human trafficking ads.

Second, we discover that applying phrase detection before training the language model also improves keyword detection and prediction accuracy, though few phrases are actually used in the final predictive model. Additionally, our model allows for the automatic detection of keywords. By simply focusing on low context and non-sparse words in the Trafficking-10k set, users can isolate a set of only two hundred likely human trafficking keywords. This is a manageable list that subject matter experts can easily review to identify potential changes in human trafficking indicators. Our pipeline allows law enforcement to identify human trafficking ads and discover new human trafficking indicators without the painstaking process of reading and analyzing every ad.

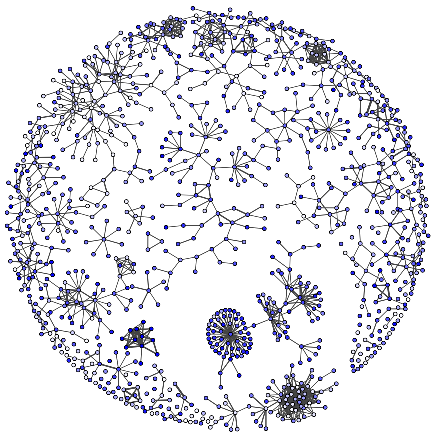
Finally, we demonstrate that our pipeline can be successfully used to detect human trafficking organizations. If our human trafficking advertisement detection pipeline is combined with an organization detection algorithm like the one discussed in [73], it could be used to verify that an organization is human trafficking or be used to detect potentially missed sex trafficking organizations.

There are still many ways in which this pipeline could be improved. Further fine tuning of phrase length and word thresholds would likely improve accuracy. Tuning the various model's hyper-parameters and especially reducing features of the logistic regression using AIC scores would also improve accuracy and reduce the final keyword list. Additionally, as previous studies [62] and the Florida trafficking ring involving Robert Kraft [74] have shown, emojis should be included in our language model because they are often used as code for human trafficking. Furthermore, the ideal human trafficking detection model would be able to identify keywords as they change. Although we are able to identify a set of one thousand words, this is likely far too many to be of practical use to users. Instead to use our dimensionality reduction pipeline, there must be training data that is continuously updated with new human trafficking ads that have been identified. The language model would also need to be retrained on new escort service advertisements periodically. Further study should also be conducted on ads over time to understand how quickly human traffickers adapt their coded language. This would inform how regularly the model and data should be updated. Even though Backpage is now shut down, human traffickers have simply gone to less centralized online platforms to advertise.[54] With a web trawling platform, finding and analyzing these ads could become an automated process. As the online presence of human trafficking continues to grow, an accurate automated organization detection pipeline is needed to conceptualize the millions of ads posted. This combined with a web trawling and keyword detection pipeline would be a significant aid to law enforcement in combating a hidden multi-billion dollar industry.

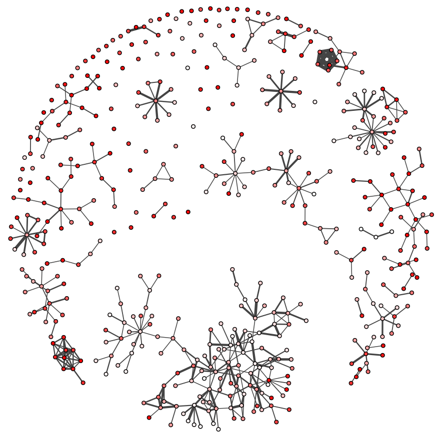
Figure 3.13: Full Co-occurrence Network



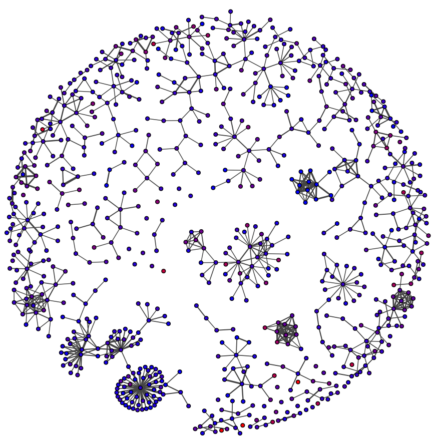
More red=higher risk; more blue=lower risk



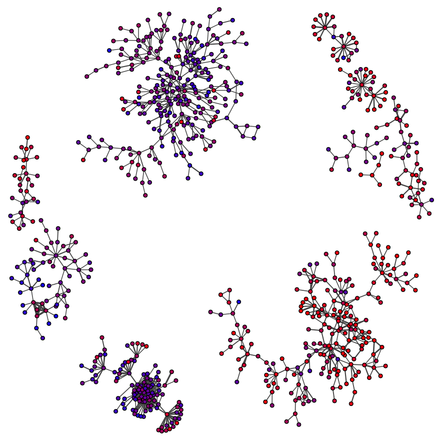
(a) Low Risk Network



(b) High Risk Network



(c) Nodes not in High Risk with Connections Network



(d) High Risk Network with Connections

Figure 3.14: Subsets of the Full Network

Chapter 4

Conclusion

This research demonstrates the strength of regression based modeling despite present day's hype for neural networks and deep learning. Careful feature selection and pre-processing was necessary to achieve applicable results. Feature selection was a non-trivial task. However, our methodologies are generalizable and easily understood by potentially less technical end users. With an interpretable maximum likelihood Heckman model, we were able to glean a better understanding of the indicators (or lack thereof) of food safety risks. With a simple logistic regression and NLP, we were able to out-perform a multimodal deep learning based model to detect human trafficking advertisements. These models provide end users with an understanding of "why" that a neural network simply can not provide. As a result, it leaves an opportunity for users to take actionable steps in accordance to model results: address food safety risks or investigate a human trafficking organization.

Appendix A

Human Trafficking Detection Model Results

Table A.1: Model Results

Method-Features	# of Words/Phrases	F1	Accuracy	Precision	Recall
HTDN-Unimodal		.658	.788	.698	.623
HTDN-Multimodal		.665	.800	.714	.622
Human baseline		.737	.840	.767	.709
RF-Low Context	1000/25	.665	.811	.832	.553
RF-Low Context	414/0	.648	.798	.792	.548
RF-High Context	1000/25	.003	.662	.002	.003
RF-High Perplexity	1000/25	.593	.794	.896	.442
RF-Low Perplexity	1000/25	.594	.792	.881	.448
RF-No Sparse	414/0	.639	.806	.868	.505
CART-Low Context	1000/25	.603	.789	.825	.475
CART-No Sparse	414/0	.613	.789	.809	.494
Logit-Low Context	1000/25	.667	.784	.697	.639
Logit-No Sparse (TFIDFS)	414/0	.670	.800	.762	.597
Logit -Low Context and No Sparse and No Short(LCLM)	1000(665)/25(1)	.674	.799	.749	.612
Logit -Low Context + High TFIDF + No Short (LCHT)*	1000+600 (330)/25(1)	.674	.802	.761	.605
RF+Logit - TFIDF*	414(110)/0	.655	.801	.793	.557
RF+Logit - Low Context*	1500(328)/25(2)	.661	.727	.570	.787
RF+Logit - Low Context +No Sparse*	1500(167)/25(0)	.665	.792	.730	.610
RF+Logit - Low Context + No Short (RFLM)*	1000(218)/25(3)	.662	.745	.600	.734

* The parentheses denote the number of entities remaining on average after dimensionality reduction

**These are the five-fold cross validation results across models using the same random splits

***Although RFLM does not have the highest results in this split, across a hundred different random samples it out performs all the other models.

Bibliography

[1] Kenneth Heafield, Ivan Pouzyrevski, Jonathan H Clark, and Philip Koehn. Scalable modified Kneser-Ney language model estimation, August 2013.

[2] David Stanway. China uncovers 500,000 food safety violations in nine months, December 2016.

[3] Humeyra Pamuk. U.S. seizes 1 million pounds of pork from China on swine fever concerns, March 2019.

[4] The People's Republic of China: U.S.-China trade facts.

[5] Imported seafood safety: FDA and USDA could strengthen efforts to prevent unsafe drug residues, September 2017.

[6] Peter Cassell. Statement from FDA commissioner Scott Gottlieb, M.D., on developments in the romaine outbreak investigation, recent outbreaks and the use of modern tools to advance food safety, June 2018.

[7] Tyco Integrated Security. Recall: The food industry's biggest threat to profitability, October 2012.

[8] Maki Hatanaka, Carmen Bain, and Lawrence Busch. Third-party certification in the global agrifood system. *Food Policy*, 30:354–369, 2005.

[9] Erin Holleran, Maury E Bredahl, and Lokman Zaibet. Private incentives for adopting food safety and quality assurance. *Food Policy*, 24:669–683, 1999.

[10] Riccardo Berti and Mariagrazia Semprebon. Food traceability in China. *European Food & Feed Law Review*, pages 522–531, 2018.

[11] T. Moe. Perspectives on traceability in food manufacture. *Trends in Food Science & Technology*, 9:211–214, 1998.

[12] Linus U. Opara. Traceability in agriculture and food supply chcain: A review of basic concepts, technological implications, and future prospects. *Food, Agriculture & Environment*, 1:101–106, 2003.

[13] S. Andrew Starbird and Vincent Amanor-Boadu. Do inspection and traceability provide incentives for food safety. *Journal of Agricultural and Resource Economics*, 31:14–26, 2006.

[14] Aleda V. Roth, Andy A Tsay, Madeleine E Pullman, and John V. Gray. Unraveling the food supply chain: Strategic insights from China and the 2007 recalls. *Journal of Supply Chain Management*, 44:22–39, 2008.

[15] A. Regattieri, M. Gamberi, and R. Manzini. Traceability of food products: General framework and experimental evidence. *Journal of Food Engineering*, 81:347–356, 2007.

[16] David L. Ortega, H. Holly Wang, Laping Wu, and Nicole J. Olynk. Modeling heterogeneity in consumer preferences for select food safety attributes in China. *Food Policy*, 36:318–324, 2011.

[17] Lingling Xu and Linhai Wu. Food safety and consumer willingness to pay for certified traceable food in China. *Journal of the science of food and agriculture*, 90:1368–1373, 2010.

[18] Deloitte. From farm to fork: Addressing food safety risks along the supply chain, February 2019.

[19] Tracie Mcmillan. How china plans to feed 1.4 billion growing appetites, February 2018.

[20] Yong-Sheng Liu, Rong Yu, and Xiang-Xiang Lin. Food supply chain safety risk prevention and control: Based on the behavioral perspective. *Journal of Service Science and Management*, 5:263–268, 2012.

[21] Dominik Zimon. The impact of quality management systems on the effectiveness of food supply chains. *Technology, Education, Management, Informatics Journal*, 6:693–698, 2017.

[22] C. Krug, M.J. Haskell, T.Nunes, and G. Stilwell. Creating a model to detect dairy cattle farms with poor welfare using a national database. *Preventive Veterinary Medicine*, 122:280–286, 2015.

[23] Ruifang Zhang, Lin Zhou, Min Zuo, Qingchuan Zhang, Mingwen Bi, Qingyu Jin, and Zelong Xu. Prediction of dairy product quality risk based on extreme learning machine, 2018.

[24] Shujing Wang. *Improving Behavioral Decision Making in Operations and Food Safety Management*. PhD thesis, Massachusetts Institute of Technology, 77 Massachusetts Avenue Cambridge, MA 02139, September 2018.

[25] Retsef Levi, Qiao Liang, Nicholas Renegar, Qi Yang, Run Zhou, and Weihua Zhou. Combining multiple information sources for informing food safety regulation in china. Technical report, Massachusetts Institute of Technology and Zhejiang University, February 2019. Working Paper.

[26] Yasheng Huang, Retsef Levi, Stacy Springs, Shujing Wang, and Yanchong Zheng. Risk drivers for economically motivated food adulteration in china's farming supply chains. Technical report, Massachusetts Institute of Technology, 2017. Working Paper.

[27] Alberto Fernandez, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, April 2017.

[28] L. Torgo. *Data Mining with R, learning with case studies.* Chapman and Hall/CRC, 2010.

[29] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O.Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[30] Samir Dani. *Chapter 4: Predicting and Managing Supply Chain Risks*, chapter 4, pages 53–66. Springer, 2009.

[31] Terry M. Therneau and Elizabeth J. Atkinson. An introduction to recursive partitioning using the RPART routines. *Mayo Foundation*, February 2018.

[32] Roger J Lewis. An introduction to classification and regression tree (CART) analysis. *Annual Meeting of the Society for Academic Emergency Medicine*, 2000.

[33] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, pages 826–833, August 2010.

[34] C.I. Bliss. The method of probits. *Science*, 79:38–39, 1934.

[35] Annette J Dobson. *An Introduction to Generalized Linear Models: Second Edition.* Chapman & Hall/CRC, 2002.

[36] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47:154–161, January 1979.

[37] James J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5:475–492, October 1976.

[38] Shenyang Guo and Mark W Fraser. *Sample Selection and Related Models*, chapter 4, pages 85–125. Sage Publishing, 2014.

[39] Statacorp. Heckman - Stata.

[40] Jennifer Clever and FAS Beijing Staff. Food and agricultural import regulations and standards - narrative, January 2018.

[41] BBC. Modern slavery, 2014.

[42] U.S. Department of State. Trafficking in persons report 2013, 2013.

[43] ILO. Global estimates of modern slavery: Forced labour and forced marriage, 2017.

[44] National Human Trafficking Hotline. Hotline statistics, 2018.

[45] Global Initiative to Fight Human Trafficking. Global report on trafficking in persons, February 2009.

[46] ILO. Profits of poverty: The economics of forced labour, 2014.

[47] Melisa A Pendergrass. The intersection of human trafficking and technology. Master's thesis, Utica College, May 2018.

[48] Seth Daire. Memex helps find human trafficking cases online, May 2015.

[49] Eric Hal, David Schrol, and William Wright. Tellfinder: Discovering related content in big data, 2015.

[50] Lori Rozza, Tom Jackman, and Mark Berman. Surveillance cameras, suitcases, and billionaires: How an investigation into massage parlors unfolded in florida, March 2018.

[51] Rick Jervis. How Florida police snared nearly 300 - including Robert Kraft - at spas used for sex trafficking, March 2019.

[52] District of Oregon U.S. Attorney's Office. Nationwide sting operation targets illegal asian brothels, six indicted for racketeering, January 2019.

[53] 18 U.S.C. §§ 2, 371, and 1952(a)(3). U.S. v ZongTao Chen, Weixuan Zhou, Yan Wang, Ting Fu, Chaodan Wang, November 2018.

[54] Ryan Tarenelli. Online sex ads rebound, months after shutdown of backpage, November 2018.

[55] Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. Combating human trafficking with deep multimodal models. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 Long papers)*, pages 1547–1556, 2017.

[56] Mark Latonero. Human trafficking online, September 2011.

[57] Marinus Analytics. Amazon Rekognition helps Marinus Analytics fight human trafficking.

[58] Jocelyn Canas-Moreira Larry Lavarez. A victim-cenered approach to sex trafficking cases, November 2015.

[59] Emily Kennedy. Predictive patterns of sex trafficking online, April 2012.

[60] Daniel Ribeiro Silva, Andrew Philpot, Abhishek Sundararajan, Nicole Marie Bryan, and Eduard Hovy. Data integration from open internet sources and network detection to combat underage sex trafficking. *Proceedings of the 15th Annual International Conference on Digital Government Research*, pages 86–90, June 2014.

[61] Michelle Ibanez and Daniel D Suthers. Detection of domestic human trafficking indicators and movement trends using content available on open internet sources. *2014 47th Hawaii International Conference on System Sciences*, 2014.

[62] Jessica Whitney, Murray E Jennex, Aaron Elkins, and Eric Frost. Don't want to get caught? don't say it: The use of emojis in online human sex trafficking ads. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[63] Maris Hultgren, John Persano, Murray E Jennex, and Cezar Ornatowski. Using knowledge management to assist in identifying human sex trafficking. *Proceedings of the 49th Hawaii International Conference on System Sciences*, 2016.

[64] Artur Dubrawski, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking*, 1:65–85, April 2015.

[65] Daniel Jurafsky and James H Martin. *Speech and Language Processing*, chapter Chapter 4: Language Modeling with N-grams. Prentice Hall, 2016.

[66] Tatsuya Kawahara and Shuji Doshita. Topic independent language model for key-phrase detection and verification. *1999 IEEE Interational COnference on Acoustics, Speech, and Signal Processing*, pages 685–688, 1999.

[67] Xiaohua Zhou, Xiaohua Hu, and Xiaodan Zhang. Topic signature language models for ad hoc retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19(9):1276–1287, September 2007.

[68] Kenneth Heafield. KenLM: faster and smaller language model queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July 2011.

[69] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August 2013.

[70] Ingo Feinerer and Kurt Hornik. *tm: Text Mining Package*, 2018. R package version 0.7-5.

[71] Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R News*, 2(3):18–22, 2002.

[72] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, January 2001.

[73] Lin Li, Olga Simek, Angela Lai, Matthew Daggett, Charlie K. Dagli, and Cara Jones. Detection and characterization of human trafficking networks using unsupervised scalable text template matching. In *2018 IEEE International Conference on Big Data*, pages 3111–3120, December 2018.

[74] Nicholas Kulish, Frances Robles, and Patricia Mazzei. Behind illicit massage parlors lie a vast crime network and modern indentured servitude. *The New York Times*, March 2019.