Kavya Chaturvedi

Professor Jonathan Hanke

SML310 Final Project

Overview

Poor labor conditions within corporate supply chains, including forced labor, is a pervasive and crucial issue for companies, investors, and policy makers. The International Labor Organization estimates that 24.9 million people are victims of forced labor globally, and forced labor in the private economy generates USD $150 billion in illegal profits each year. This project attempts to understand which types of companies are most likely to engage in corporate practices that are particularly nontransparent and risky for labor conditions.

This project uses benchmarking data from KnowTheChain, a resource that publishes individual company scorecards and industry rankings on metrics related to social and labor conditions within corporate supply chains. For hundreds of large international companies, KnowTheChain researches company activities, disclosures, and policies and assigns the company a score and ranking on their corporate practices.

Using random forest models, combinations of decision trees use characteristics of the company to predict the KnowTheChain benchmark final score, as well as the 7 individual sub-scores that make up the final score. These random forest models are able to predict the KnowTheChain score with varying accuracy levels for different scores and sub-scores, and these predictions differ across industry, suggesting that information on general characteristics of companies may be able to predict some types of risky corporate behaviors, but not all. However, the results show that in most cases, approximately half of the variation in scores *can* be

explained by these characteristics, providing interesting insight for policymakers, investors, and corporations.

Related Work

While there is academic literature dedicated to the use of forced labor practices in corporate supply chains, much of this literature is qualitative research into specific sectors, or specific companies. This research is surely valuable, as it helps stakeholders, including KnowTheChain, understand which corporate practices create conditions for forced labor, and which industries are most high risk that should be analyzed further. Many of the indicators that KnowTheChain includes in their methodology is based on this body of academic literature. For example, case study based research has shown that forced labor is enabled when workers are hired through recruitment agencies, which often take advantage of disadvantaged workers and traffic the workers to warehouses. Based on this, KnowTheChain's research includes a sub-score for Recruitment practices, giving higher scores to companies that hire their workers directly and transparently.

Despite the valuable qualitative research being rich in this field, there is little quantitative research into corporate practices related to forced labor in their supply chains. The quantitative research that exists in this field is mostly non-academic, and done for the benefit of investors and companies themselves, like the work of KnowTheChain.

Relevant Data

The two datasets used in this project both come from KnowTheChain, a public resource that uses a detailed methodology to grade company's efforts to mitigate risks of forced labor within their supply chain. In this paper, two datasets are used from KnowTheChain's 2020/2021 review, for both the food and beverage sector (42 companies) and the information and

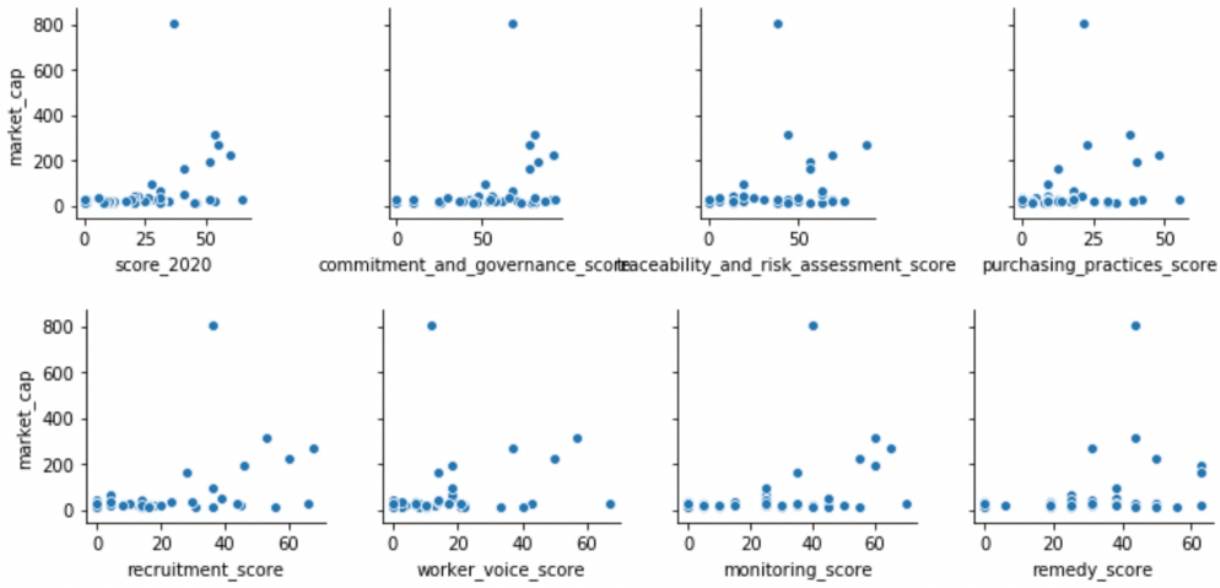
communication technology sector (48 companies). KnowTheChain's benchmarking methodology is based on the UN Guiding Principles on Business and Human Rights, and is reviewed every two years before their research is refreshed. According to their website, KnowTheChain's methodology is developed with the support of industry leaders, including "sector-specific consultations with stakeholders from civil society, investors, and business".

For each company, the 2020/2021 datasets include the name of the company, its market capitalization, its country and region, and subindustry (for the food and beverage dataset), and various scores for the company. For the purpose of this analysis, the categorical variables have been one-hot encoded to allow them to be used in the random forest models. The main, composite score is a grade that incorporates seven sub-scores equally, each of which are determined by research of publications from the company. The seven indicators are 1) Commitment and Governance, 2) Traceability and Risk Assessment, 3) Purchasing Practices, 4) Recruitment, 5) Worker Voice, 6) Monitoring, 7) Remedy. The specific research and grading parameters has also been made public by KnowTheChain for each indicator. For example, in order to grade the Commitment indicator, KnowTheChain looks at five sub-indicators and analyzes in great detail their public commitment to addressing forced labor, their supplier code of conduct, collaboration and engagement with other stakeholders, and more.

Two limitations of the dataset are the selection bias for inclusion into the datasets and the small volume. First, KnowTheChain studies companies that are among the largest (by market capitalization) in high-risk industries for forced labor. The implication of this selection bias is that the scores may be lower than other smaller companies in different industries, and so the models may not be able to extrapolate to companies of a different size or a different industry. Second, because of the intensive research that goes into including each company in the dataset,

the datasets end up being quite small, with less than 50 companies in each industry's dataset. This small sample size prevents us from doing a test-train split, which may impact the accuracy of the model. Additionally, the small sample size forces us to create quite shallow decision trees in order to avoid overfitting, which also may cause us to sacrifice accuracy.

Exploratory data analysis helps to prove that the selection of observation and small sample size may affect our analysis of this data. For example, graphed below is one of the covariates, `market_cap` (which represents the size of the company's market capitalization, in billions of US dollars), vs the composite score and each of the subset scores. We can see that the majority of companies are clustered around the lower end of market capitalizations. We also observe that there is varying levels of correlation between market cap and each of the scores within the sub-scores, which will contribute to the model's ability to predict the scores based on covariate information.



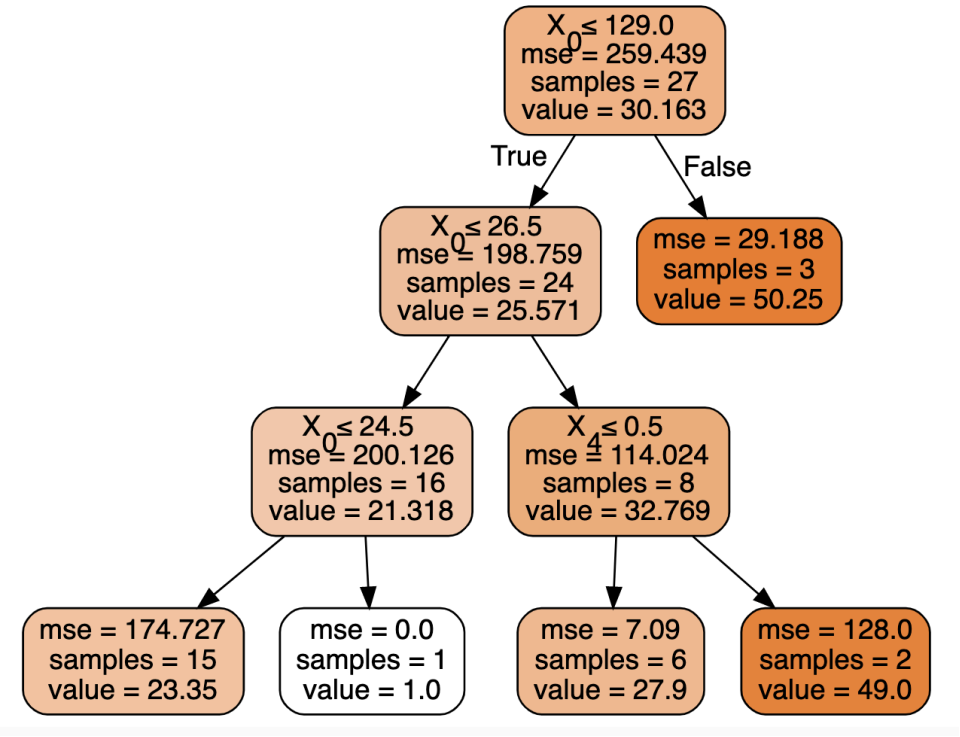
Analysis

In order to model this data, I will use random forest regressor estimators for each of the eight scores available on both datasets. The goal of each of these models is to use characteristics of the company (market cap, region, and subindustry if available) to predict the scores and compare the scores of the models. A random forest regressor model is used in this paper for two reasons. First, our outcome of interest – the scores – are continuous variables, and so the regressor must be used in order to predict a continuous variable. Second, random forest models are preferable to decision trees in this case to control over-fitting. Random forests use many decision trees over sub-samples of the dataset and average across the models to prevent over-fitting, a problem that is particularly significant given the small sample sizes in this case.

A few choices with regards to model parameters are made for all the random forest models in this project. First, the number of estimators (meaning the number of decision trees in the random forest) is set to 100, the default in the sci-kit learn package. Additionally, the maximum depth for the decision trees is set at 3 and the minimum number of samples required to split an internal node is set at 5. Once again, these decisions were made to prevent overfitting within the decision trees, and the number was arrived at after trial and error of picking out trees from the forest and attempting to prevent overfit trees.

Below is an example decision trees pulled out of the random forest model randomly. We can see that the nodes in these two trees are being split based off of the first covariate and fourth covariates, in this case representing the market cap and presence in the region of Australia. There does seem to be large variance within some leaf nodes, such as the leaf node on the extreme left, with the lowest values but many samples. Other nodes, such as the leaf node second from the left, are overfit with just one sample in that group. However, while these problems of some leaf nodes being overfit while others are underfit are present in this particular decision tree, we

expect that the random forest of repeated decision trees will help to eliminate some of the concerns around fit.



Results

Random forest models were made for the 8 scores of the two datasets, for the Food & Beverage company dataset and the Information & Communication Technology company dataset. For each model, a score was computed, which returns the coefficient of determination, R^2, of the prediction. R^2 measures the amount of proportion of the variance in the outcome of interest (scores) that is explained by the independent variables (company characteristics). Below is a chart that shows the R^2 values for each of the models. Importantly, these numbers are not the scores for the industries themselves, but rather the percentage of variation in the scores that can be explained by our models with our explanatory variables.

As we can see, the models with the highest R^2 values, meaning that the scores can be most explained by our independent variables, differ across and between our two datasets. For the

Food & Beverage dataset, the Monitoring and Recruitment scores are the best explained, with a R^2 value over 0.57%. In other words, for those scores, 57% of the variance of those scores can be explained by our model, which is a substantial amount. The lowest score for the Food & Beverage dataset is the Remedy score, which is 0.4568, meaning that only 45.7% of the variance in the Remedy scores can be explained by our model. For the Information & Communication Technology Score, we see much more variation amongst the scores. The highest score was in the Worker Voice category, at 60.9%, and the lowest score being 37.1% for Purchasing Practices.

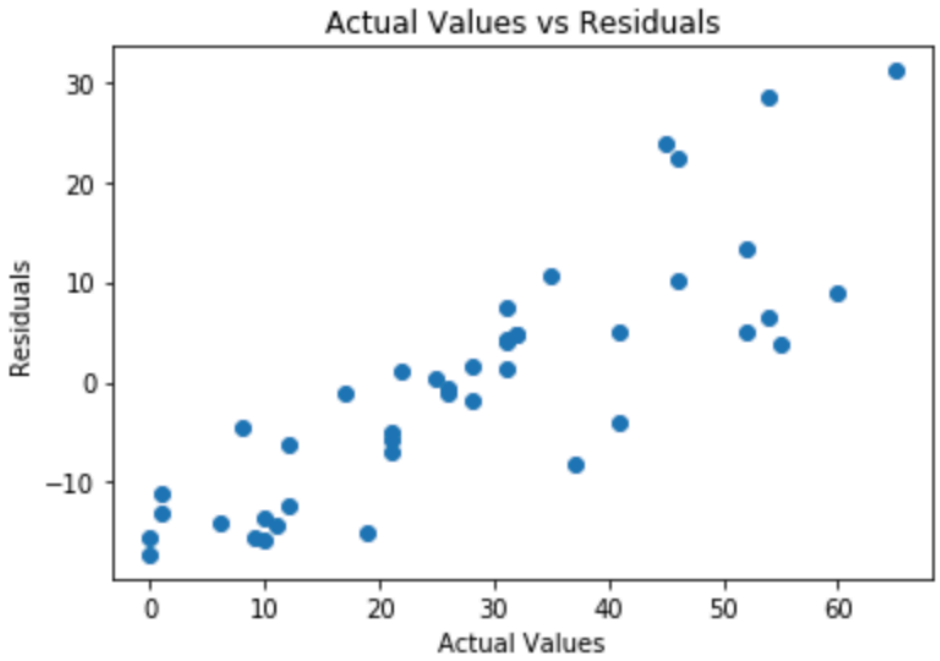
Score	Food & Beverage Score	Info & Communication Technology Score
General	0.5393	0.5471
Commitment and Governance	0.5469	0.5309
Traceability and Risk Assessment	0.4801	0.5363
Purchasing Practices	0.547	0.371
Recruitment	0.5701	0.5584
Worker Voice	0.5366	0.6093
Monitoring	0.5792	0.5261
Remedy	0.4568	0.4891

Interestingly, there are also substantial differences between the industries' R^2 values for certain scores. There can be some speculation about why some scores are lower between the two industries. For example, perhaps the reason why a much smaller percentage of the variation in the Purchasing Practices category can be explained with the Information and Communication Technology sector is because technology companies vary much more in their purchasing practices since their materials are much more sophisticated, and so further information about the company is required in order to predict their scores. Similarly, perhaps because the ICT sector is more manufacturing-dominated with established supply chains, the Worker Voice score is more easily predicted because all technology companies may have more direct means of communicating with their workers (meaning lower variance within that industry). Clearly, these explanations are simply educated guesses, and further research would be helpful to understand

what the exact differences are between these industries that could be driving the differences across the scores.

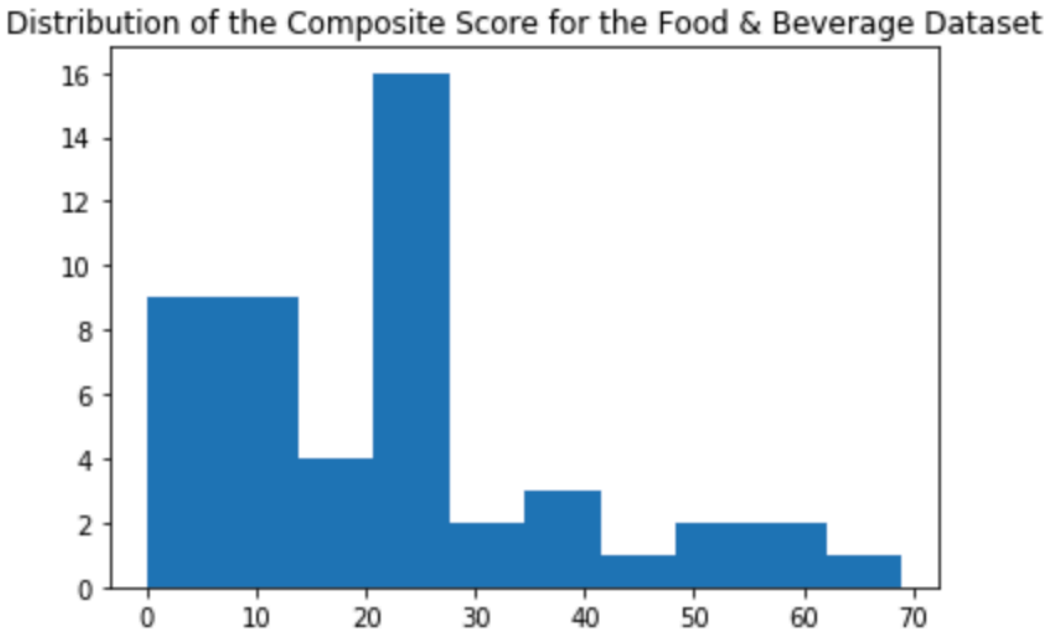
Generally, the R^2 scores across the board are quite low, with all but one of the scores below 60%. These scores suggest that, in general, more contributes to a company's labor practices in their supply chain than just the size, region, and subindustry, which is an intuitive conclusion. Furthermore, it is quite impressive to have a model explain roughly 50% of the variation for such a complex dependent variable.

However, further analysis can be done to understand why these scores are not higher for our models. In order to answer this question, we can graph the actual y-values vs the residuals (the actual y values minus the model's predicted values), as seen below.



Above is the residuals vs actual values for the first random forest model, for the Food & Beverage database's Composite Score. As we can see from this graph, this is quite a strong correlation between the residuals and the actual values of our dataset. The residuals seems to start in the negative values when the actual y values are low, meaning the predicted values are

overestimates, and the predicted values are underestimates of the actual high scores because of the very high residuals. This error could occur because of the distribution of the scores. As we can see from the distribution of the composite scores for the Food & Beverage dataset on the histogram below, the scores are concentrated in the 20s. Accordingly, around the 20s, the residuals are near zero. In addition, there are very few observations with high composite scores, making it difficult for us to extrapolate to data points at the upper extreme and increasing the residuals in that region.



Conclusion

In this paper, random forest regressor models were used to predict company scores on corporate practices related to forced labor in their supply chains. Using data from KnowTheChain's corporate benchmarking databases, a model was made for one composite score and seven sub-scores within two datasets, for companies within the Food & Beverage industry and the Information & Communication Technology industry. Each of these models was then scored with its R^2 value, which indicates how much of the variance of the outcome variable (the

company's score) is explained by the explanatory variables. For most of the models, approximately 50% of the variation could be explained by just a few company characteristics such as region, market cap, and subindustry, which is quite significant. We also observed interesting differences among the seven sub-scores included in the dataset, and striking differences in the model's score between the two industries. While some of these differences can be explained intuitively, further research would also help to explain this variability.

There were also significant limitations to the models created in this study due to the small sample size found in this dataset. For example, decisions were made about parameters of model fit to avoid overfitting and help the model remain generalizable. Additionally, the data within the dataset was very skewed. Because the observations all came from large companies within high-risk sectors, the scores were all quite low, meaning that the models may not accurately predict companies of a different size or different industry. Further research can enhance these findings by expanding to larger datasets as research becomes more readily available or expanding this work into a longitudinal study to compare company scores over time.