Sustainable Cocoa Farming Prediction

McGill University

[redacted]

[redacted]

Introduction

Chocolate is a snack that has a long, rich history, dating back to indigenous mesoamerican rituals where cocoa was used to create chocolate flavored drinks in 175 BC (Barras, 2018). To this day, the chocolate market continues to grow every year, having amassed a total of 238.5 billion dollars in revenue worldwide in 2023 alone (Statista, n.d.). The popularity of chocolate makes it an attractive product to sell for many companies, with notable players in the industry like Nestle and Mondelez making millions of dollars off chocolate sales alone (Statista, n.d.). However, chocolate is a product that requires the harvesting of millions of cacao pods. Often, chocolate producers can sell their chocolate at low, competitive pricing because its production is unethical for the environment and its labor. To have this competitive pricing, these companies source their cocoa from suppliers that sell at the cheapest rate. They can only get these prices by employing unethical means; often, child labour is used with children trafficked to work in cacao farms from as young as 10 years old. In an article by the Washington Post, the largest chocolate producers in the world refuse to identify the farms at which they source from. Yet, over 2 million children are used as labor to extract cocoa in farms (Whoriskey & Siegel, 2019). In addition, cocoa farms require large pieces of land to grow the cacao trees and take a long time to yield cocoa. It takes a cacao tree an entire year to produce the cocoa required for only half a pound of chocolate. As such, the WWF estimates that 70% of Africa's illegal deforestation is caused by the destruction of tropical forests to create cocoa farms (WWF, 2017). Given this, this paper aims to use statistical analysis on a dataset of chocolate bars and their ratings to create an ethical sourcing plan for chocolate producers' best rated chocolate bars. This will be achieved by predicting the best locations to source cocoa, then utilizing that prediction to find ethical cocoa farms in that location.

Data Description

The data in the chocolate dataset included 1,759 observations. These observations had 9 different features that describe the chocolate bar's details such as the company that sells the chocolate bar, its location, and the origin of the cocoa bean both specifically, with the exact town for example, and more generally, with just the country it was sourced from. It also included the REF number for the chocolate bar, the date that the review was left, and the rating given to the chocolate bar. Lastly, the cocoa percent of the bar and the type of bean were also included as features in the dataset.

To get better insight into the distributions of the data, visualizations were created with these features to see how they're distributed. The majority of chocolate bars have a cocoa percent of around 70%, which is considered dark chocolate (Appendix 1). The average cocoa percentage was 71%, and this feature had many outliers since the distribution ranges from 40% chocolate to 100% pure dark chocolate (Appendix 2). The most common rating given to chocolate bars was around 3.5 stars (Appendix 3). The average rating was around 3.1 stars, and had very few outliers, meaning the majority of ratings were between 2.8 and 3.5 stars (Appendix 4). The reviews in the dataset ranged from 2006 to 2017, and the average review year was 2013 (Appendix 5 & 6). The correlations between these three variables were quite weak, the strongest being between cocoa percent and rating, whose negative correlation shows that the stronger the cocoa percent, the lower it will be rated (Appendix 8).

As mentioned above, the target variable focused on is the Broad Bean Origin, or the location where the cocoa bean was sourced for each chocolate bar. Because the bean origin is a categorical variable, or non-numeric, it cannot be ran as a target variable with linear regressions and correlations. Instead, multicollinearity and heteroskedasticity, which require linear

regressions to check for, were done using Rating as the predictor. The best linear regression output was between Rating and Specific Bean Origin, with a low P-value and relatively high R-squared, potentially flagging this as a more linear relationship. Out of all the variables, only Company and Company Location were not heteroskedastic. In terms of multicollinearity, there was an extremely high collinearity issue between the REF number and the Review Date for the chocolate bars, with a correlation of 99% (Appendix 8). As such, only the review date was used in the subsequent testing and modeling as it is easier to interpret and yields better insights.

Lastly, in order to be able to make predictions on the best location to source cocoa beans, the null values in broad bean origin were removed from the dataset.

Model Selection & Methodology

The type of model selected for this analysis was Random Forest. Random Forest was chosen because this tree-based model works well for both numerical and classification problems, notably when the data has a mix of both numerical and categorical variables. Since the chocolate dataset had both numerical values, such as the cocoa percent, and categorical values, such as the broad bean origin, using Random Forest was the best type of model to be able to capture these differences and bring them together to be able to make a prediction. Before selecting a final model, some Random Forest models were created as tests using different predictors and target variables. Even though the ultimate goal is to be able to predict the bean origin, test models were created using many variables as predictors to see which yielded the best results. Firstly, to see the difference between using a numerical target variable and a categorical one, a model was created using rating as the target variable. Then, models using four different categorical target variables were also tested: Broad Bean Origin, Company, Company Location, and Specific Bean Origin. Broad Bean Origin was tested as the target variable, including different predictors in the model

to see which would yield the best results. Each classification model was compared using the Out-of-Bag (OOB) error rate estimate. When running a numerical Random Forest, your results yield a mean squared of residuals, which is the amount of error between the predicted value and the actual value. In this case, while the numerical Rating model had better overall performance with an MSR of 19%, the Broad Bean Origin classification model was the best out of all the classification models with an OOB estimate of 76%. Lastly, boosting models were created for each of these two models. Individual boosting trees are weak learners, but all together they become an ensemble, turned into very strong models. This can be especially useful to avoid overfitting. The numerical boosting model had an extremely low MSE and only 5% (Appendix 10). However, boosting in this case can't be used to improve the classification model since Bernoulli models require the target variable to be only 2 categories.

In order to deliver better managerial insights to chocolate companies and draw richer conclusions that can have a positive impact on the planet and its people, the bean origin model was selected as the final model even though the numerical model was better in terms of prediction accuracy.

Results

The final model used, as mentioned above, is a classification, Random Forest model that predicted where would be the best place to source cocoa beans given the following features: the type of chocolate (Dark versus Milk), the rating the chocolate bar should have, given those already sold, the current year, the company selling the chocolate bar, and its location. The bean type was not included in the final model because it had many missing values. This means there was little data to be based on for imputation. It would have been alright to make assumptions for a small number of missing values, but with so many, there was the risk of skewing the data if all

the missing fields were filled with the same value. Given this, the final OOB Error Rate of the model was around 76%, whose most important predictors were the company and cocoa percent, and least important predictor was the year of review. The information that was given to the model to select a cocoa bean location was the following:

1. Ambrosia, a company located in Canada, producing a 70% dark chocolate bar, whose rating would like to be 4.5 stars in 2023
2. Arete, a company located in the USA, producing a 80% dark chocolate bar, whose rating would like to be 4 stars in 2023
3. Soma, company located in Canada, producing a 40% milk chocolate bar, whose rating would like to be 4.2 stars in 2023
4. Batch, a company located in the USA, producing a 30% milk chocolate bar, whose rating would like to be 4.7 stars in 2023

With four different cases, the model was able to predict the following locations as the best for sourcing the cocoa beans needed, given the details of the chocolate bar being produced:

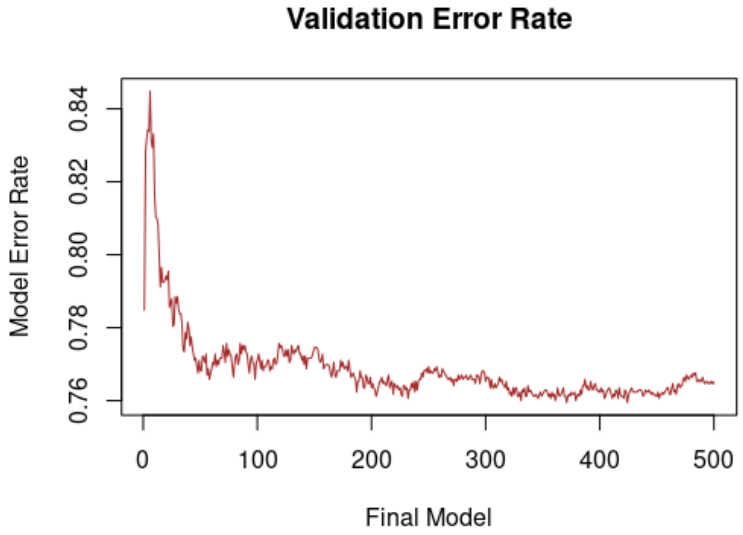
Table 1. Model Prediction Results.

Chocolate Bar	**Location to Source from**
1	Colombia
2	Peru
3	Sao Tome
4	Sao Tome

The results show that for a Canadian company like Ambrosia, the best location to source their cocoa beans from is Colombia. Similarly, for American companies such as Arete, they should source from Peru. Lastly, both Canadian and American companies producing milk chocolate

should source their beans from Sao Tome. In order to validate if these predictions were accurate, out-of-bag testing and error rates were used. The model performed better with a lower number of trees. It also overall performed only alright, with error rates of around 75%, as shown in the figure below.

Figure 1. Out-of-Bag Validation Error Rate



This means the accuracy of the model is only around 25%. Given this, the model has much room for improvement, which could be achieved in several ways; having a more robust dataset, including more predictors in the final model, and including Bean Type in the model with complete data.

Insights and Conclusions

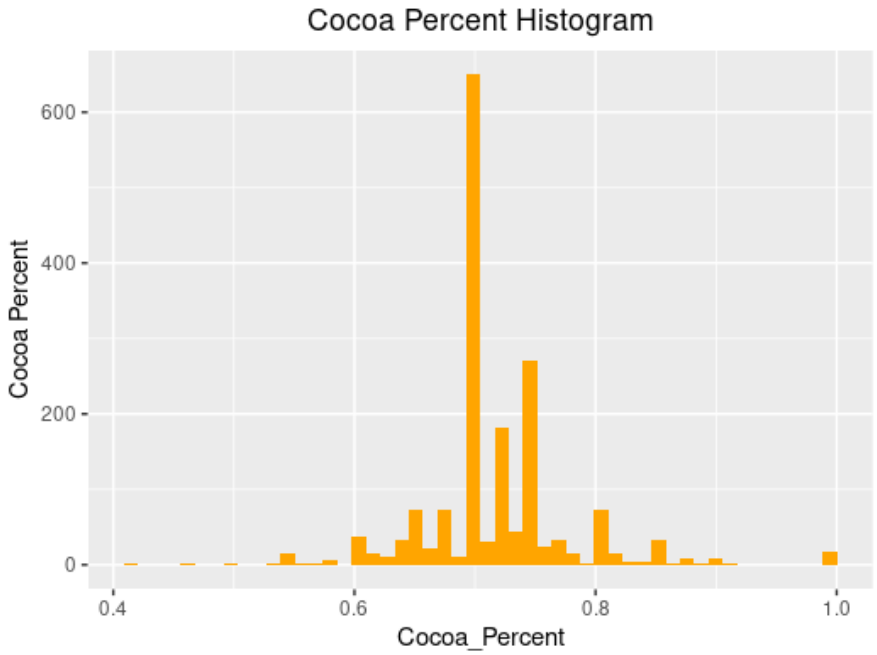
The results given above can be used to not only flag the best locations to source cocoa beans when these companies want high ratings, it also gives insight into locations where companies should look to try and find ethical and sustainable cocoa farms. As sustainability becomes more and more sought after in companies, there are many sustainable farms to choose from.

Firstly, in Sao Tome, Daarnhouwer is a certified ethical farm that produces many nuts and beans, including cocoa beans. Their cocoa production employs over 1000 community members, including over 400 women (Daarnhouwer, n.d.). They utilize their premiums to give back to the communities on Sao Tome, by building infrastructure like kindergartens. They have a strict code of conduct and social responsibility policy, as well as being part of the Sustainable Nut Initiative (SNI, 2021). Similarly, in Colombia, companies like GoodSAM work directly with indigenous and small local farmers to create a "direct trade" initiative, where there is no middle man between them and the farmers, buying all their viable crops and reinvesting profits into the community. They ensure that the traditional indigenous means of working the land are preserved to further ensure viability (Yu, 2022). Lastly, the Colpa de Loros Cooperative in Peru is one of the largest sustainable and ethical cocoa cooperatives, sourcing out of the local cocoa farmers in the Ucayali Rainforest. They work with many partners to create sustainable production and conservation projects to combat the severe deforestation in the region (Ramirez, 2023).

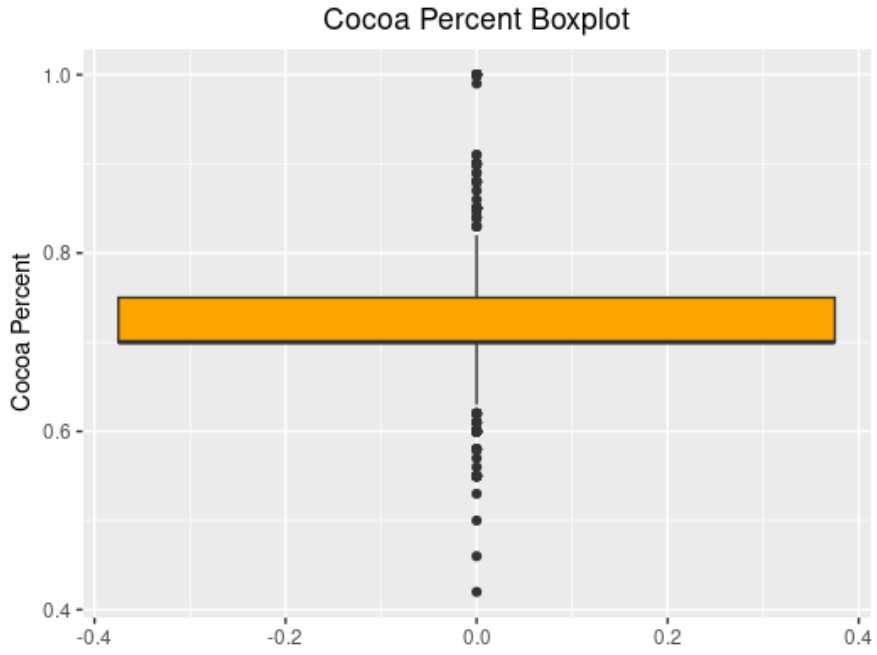
By selecting to source from farms such as these, chocolate companies can help reduce their negative environmental impact by ensuring that their cocoa is not causing mass deforestation or detrimental impact to soil. It can help them certify that ethical labor is being used, and ensure at least that their chocolate is fair trade. Nowadays, consumers are hyper aware of these issues, and seek to find ethical brands to buy from. As such, moving towards ethical farms can generate larger consumer bases and sustain long-term profitability in comparison to competitors. By using the model created, companies can look for the right places to source from, and help combat the human destruction of earth's resources and biodiversity.

Appendix

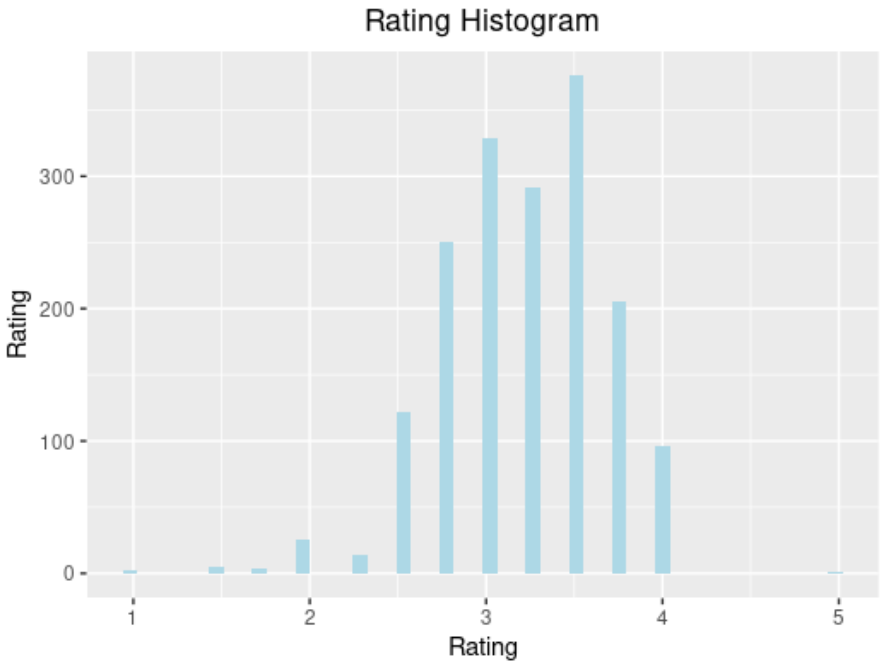
Appendix Figure 1. Cocoa Percent Histogram



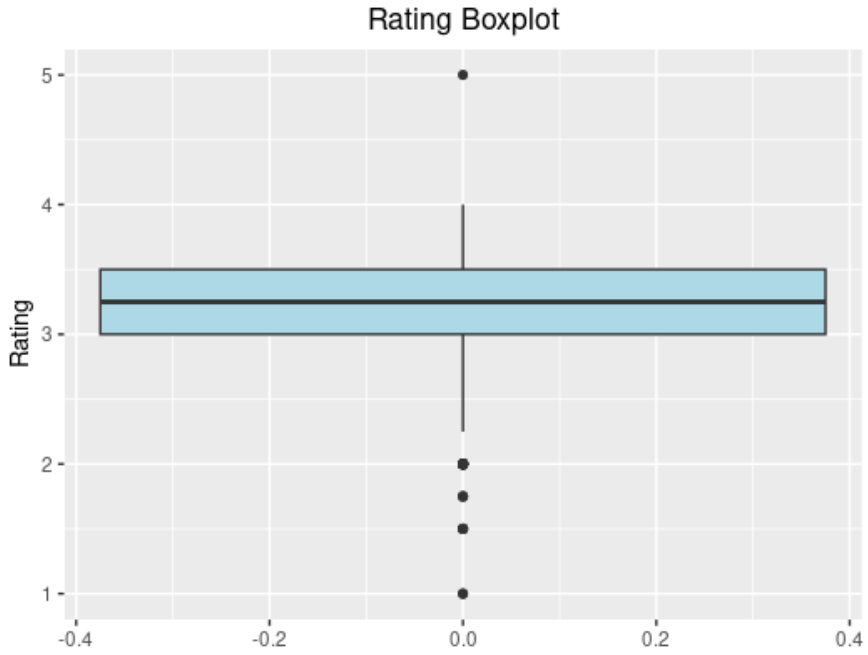
Appendix Figure 2. Cocoa Percent Boxplot



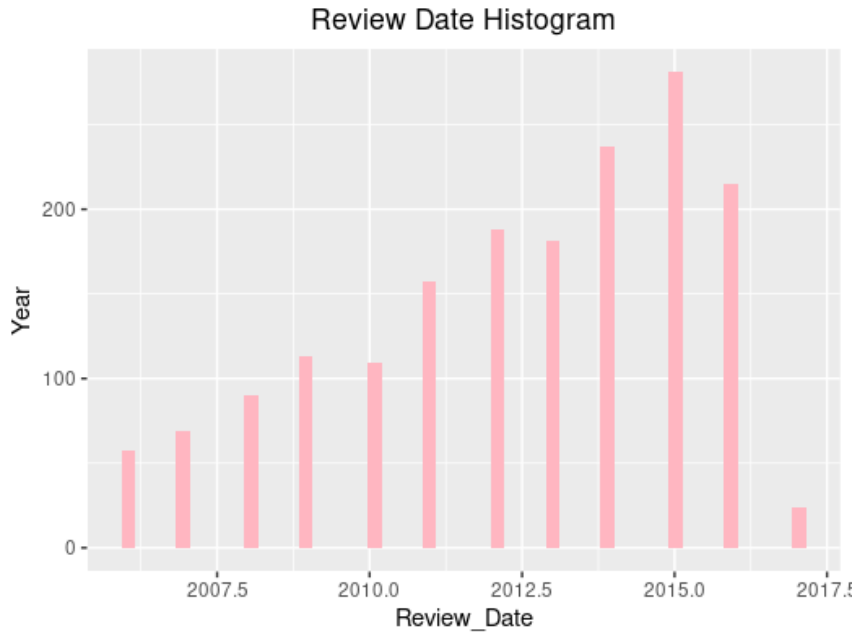
Appendix Figure 3. Rating Histogram



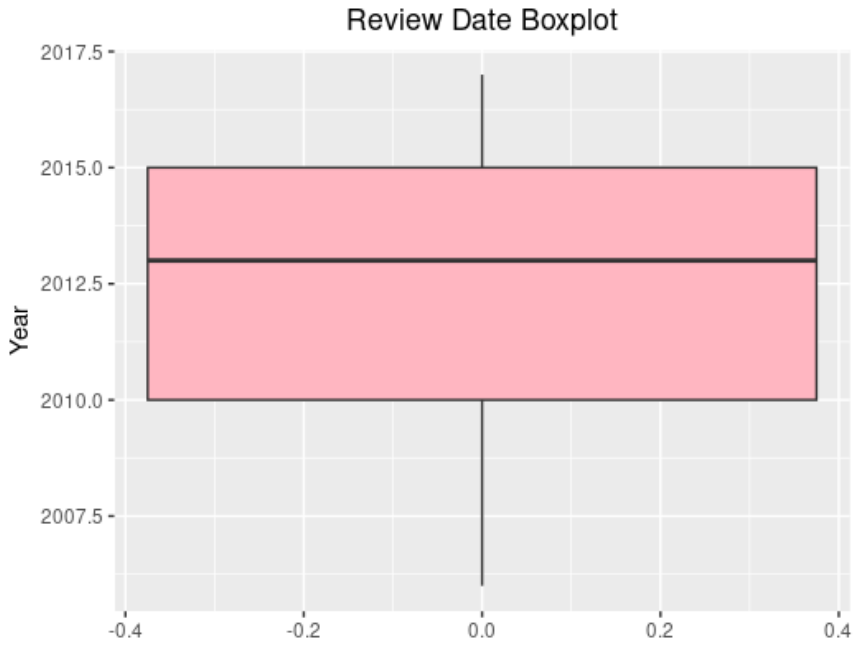
Appendix Figure 4. Rating Boxplot



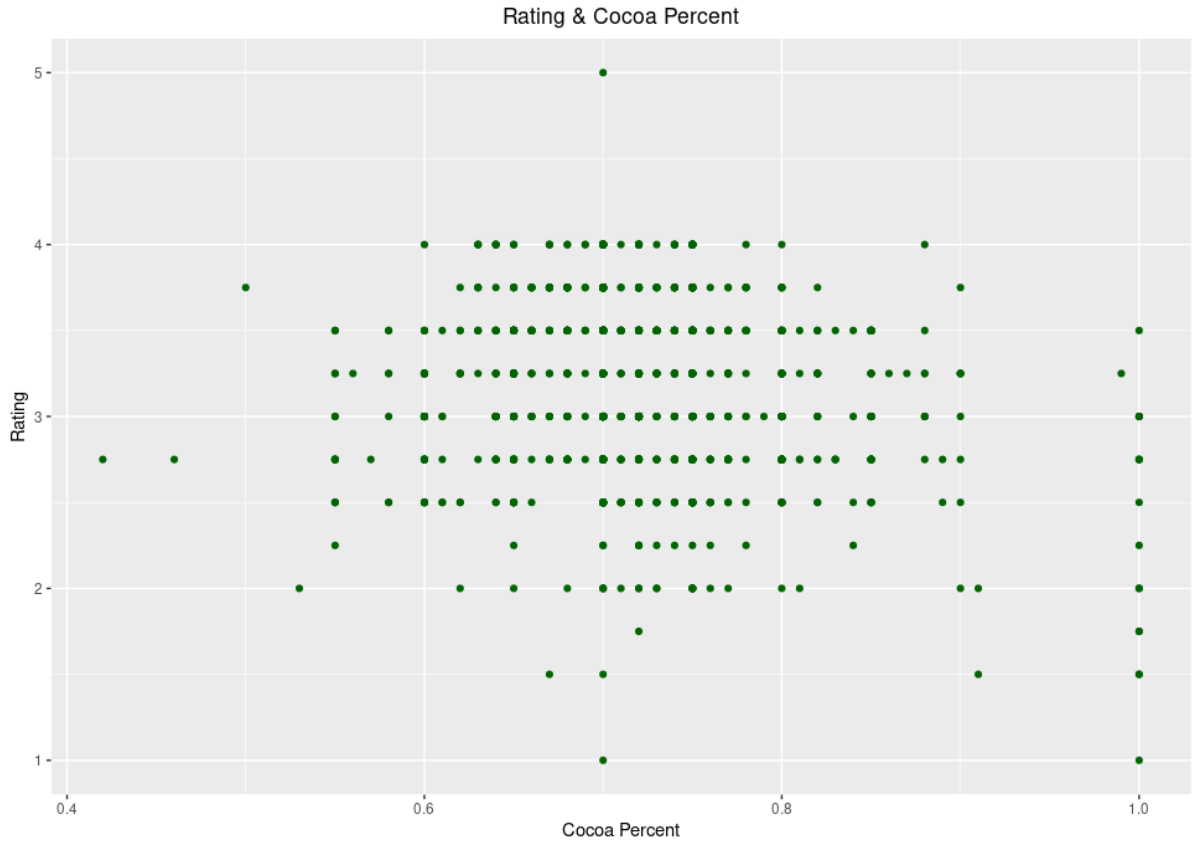
Appendix Figure 5. Review Date Histogram



Appendix Figure 6. Review Date Boxplot



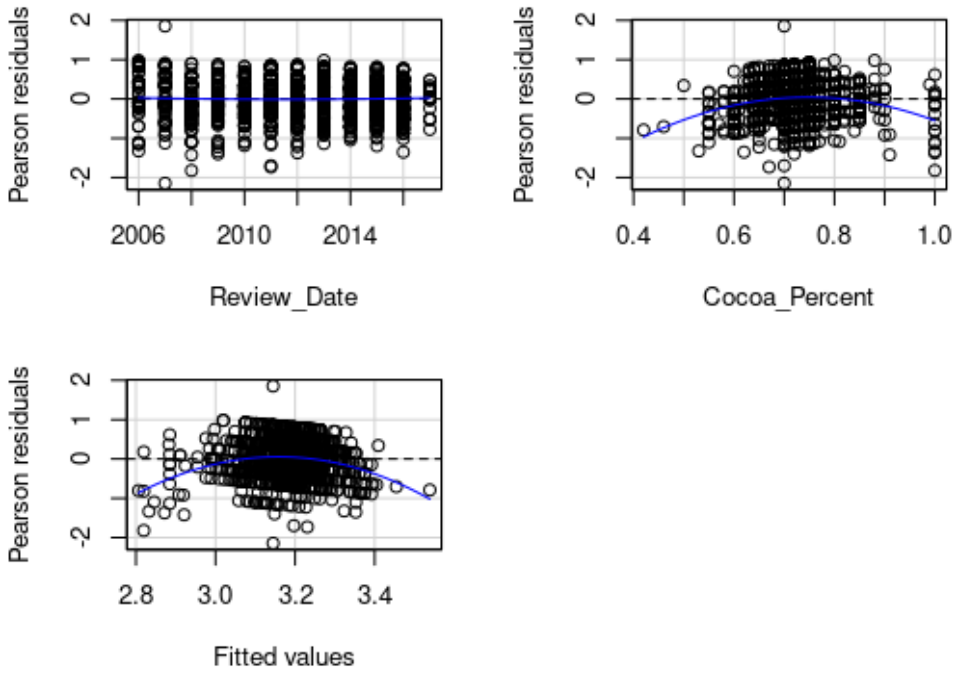
Appendix 7. Scatter Plot Between Cocoa Percent and Rating



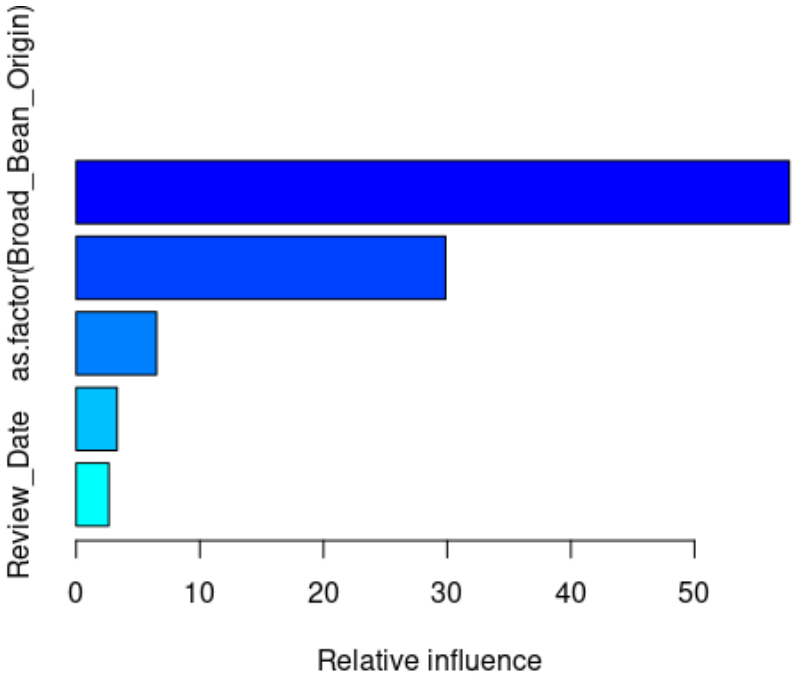
Appendix 8. Correlation Matrix Table

	REF	Review Date	Cocoa Percent	Rating
REF	1.00	0.99	0.04	0.08
Review Date	0.99	1.00	0.04	0.08
Cocoa Percent	0.04	0.04	1.00	-0.15
Rating	0.08	0.08	-0.15	1.00

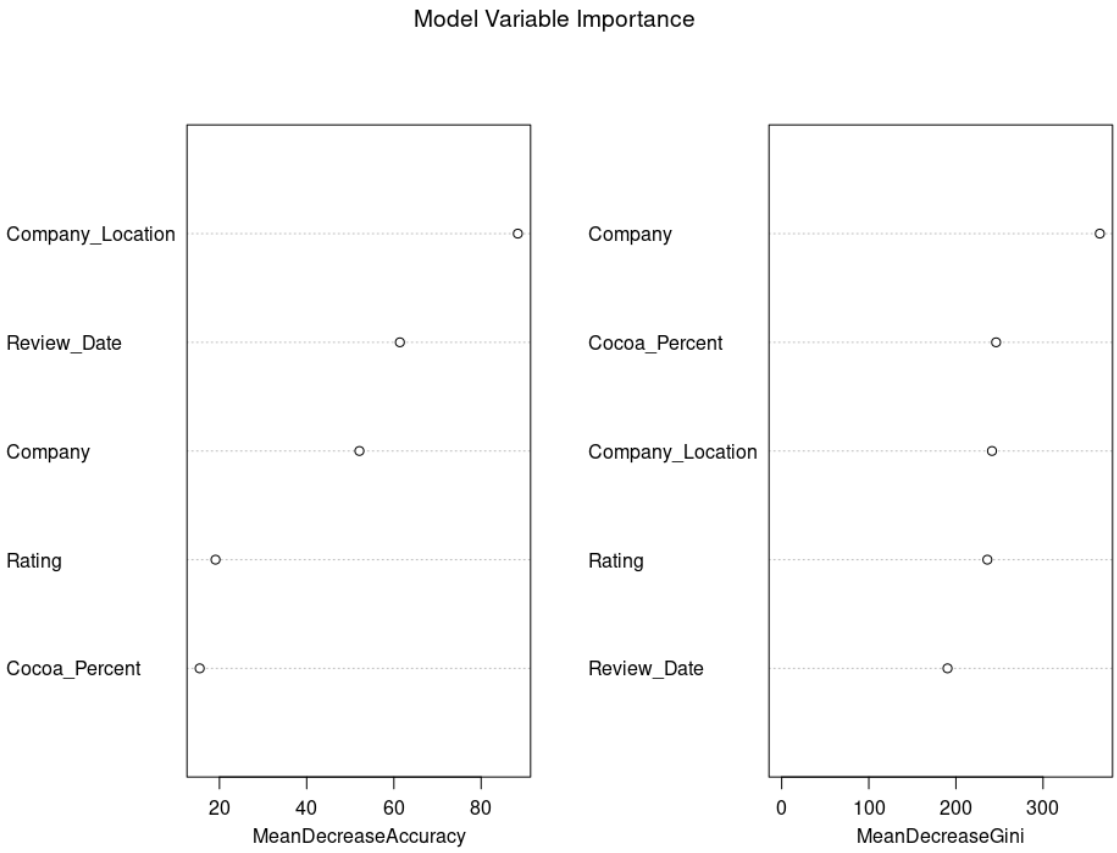
Appendix 9. Residual Plots



Appendix 10. Boosting on Test Numerical Model



Appendix 11. Variance Importance Plot for Final Model



References

Barras, C. (October 20, 2018). *World's oldest chocolate was made 5300 years ago-in a South American rainforest*. Science. https://www.science.org/content/article/world-s-oldest-chocolate-was-made-5300-years-ago-south-american-rainforest#:~:text=Ancient%20pots%20shift%20the%20nexus%20of%20chocolatemaking%20from%20Central%20America%20to%20modern%20Ecuador&text=Our%20love%20affair%20with%20chocolate,Ecuador%20some%205300%20years%20ago.

Mondelēz. (January 31, 2023). *Mondelēz International Reports Q4 And Fy 2022 Results*. Retrieved from https://ir.mondelezinternational.com/news-releases/news-release-details/mondelez-international-reports-q4-and-fy-2022-results

Ramirez, E. (June 30, 2023). *Peruvian producers are banking on cacao with the aroma of sustainability*. Alliance of Bioversity & CIAT. https://alliancebioversityciat.org/stories/peruvian-producers-are-banking-cacao-aroma-sustainability

Statista. (November, 2023). *Chocolate Confectionery - Worldwide*. Retrieved from https://www.statista.com/outlook/cmo/food/confectionery-snacks/confectionery/chocolate-confectionery/worldwide#revenue

Statista. (March, 2023). *Sales of Nestlé's confectionery sector worldwide from 2010 to 2022, by segment*. Retrieved from https://www.statista.com/statistics/236101/global-sales-of-the-confectionery-sector-of-nestle-by-segment/

Whoriskey, P. & Siegel, R. (June 5, 2019). *Cocoa's child laborers*. The Washington Post. https://www.washingtonpost.com/graphics/2019/business/hershey-nestle-mars-chocolate-child-labor-west-africa/

World Wildlife Fund. (2017). *Bittersweet: chocolate's impact on the environment*. Retrieved from https://www.worldwildlife.org/magazine/issues/spring-2017/articles/bittersweet-chocolate-s-impact-on-the-environment

Yu, D. (August 4, 2022). *'Great Things Are Coming': Colombian Cocoa Farmers' Rugged Winding Road To Upend The Unequal Food System And Their Hope For Prosperity*. Forbes. https://www.forbes.com/sites/douglasyu/2022/08/04/great-things-are-coming-colombian-cocoa-farmers-rugged-winding-road-to-upend-the-unequal-food-system-and-their-hope-for-prosperity/?sh=39142edb5110