Final Project: Predictive Chocolate Rating Analysis

Presented to

Professor Juan Serpa

By

Michelle Barabasz- 261152119

MGSC 661

McGill University - Desautels Faculty of Management

Introduction

In the realm of chocolate, understanding the intricacies of a chocolate bar's success is crucial. This project will dive into the complexities of chocolate ratings, framing a classification problem: predicting whether a chocolate bar will receive a high or low rating. In determining this, features such as cocoa percentage, bean type and broad bean origin will be examined to determine their relevance in predicting a chocolate bar's rating, while exploring the efficacy of various classification algorithms—Logistic Regression, Random Forest, and Clustering.

The core focus of this chocolate analysis project lies in constructing a robust classification model to predict whether a chocolate bar will receive a high or low rating. This predictive tool holds immense value for chocolate producers and enthusiasts, offering a deeper understanding of the factors contributing to a chocolate's rating. Beyond classification, the project utilizes tree-based methods, such as Decision Trees and Random Forest, to discern vital features influencing chocolate ratings across diverse geographic regions. By uncovering these critical elements, the project aims to provide a strategic roadmap for chocolate producers to tailor their offerings effectively. Additionally, the exploration extends into unsupervised learning, employing techniques like K-Means Clustering, to reveal distinct flavor profiles among chocolate bars. This clustering approach offers insights to aid producers in product development and marketing strategies.

The goal is to understand chocolate ratings through advanced data analysis techniques. By understanding the interplay of features and leveraging diverse algorithms, this project aims to equip stakeholders in the chocolate industry with actionable insights for refining strategies, optimizing flavors, and meeting the evolving expectations of chocolate enthusiasts.

Data Description

The dataset comprises data from 1795 chocolate bar ratings, with details such as the percentage of cocoa in the chocolate bar, the origin of the bean used, and the country of the chocolate's company. for each rating. Right away, one identifier variable, relating to the reference number was removed from the dataset as it provided no meaningful insights into the creation of the model. There were also 2 missing values which were removed. After this removal of these, the

dataset contained 8 variables that could potentially be used in the development of the above-mentioned models.

Exploratory Data Analysis

To gain a better understanding of each variable, exploratory data analysis was conducted to assess their frequency, distribution, and their relationships with the final rating. First, the target variable, rating, was examined. The frequency distribution of the ratings (see Exhibit 1 in Appendix) showed that the most frequent ratings of chocolate bars tend to be in the middle range between 1 and 5, specifically, the rating 3.5 had the greatest number of occurrences (almost 22% of all occurrences). Additionally, the summary statistics further confirmed that the mean rating value is 3.12 which is less than the median (3.25), indicating a slight left-skew.

Following this, the distribution of the independent numerical variables was examined. This was done by looking at the summary statistics and a histogram for each of the two numerical variables, the review year and the cocoa percentage. In examining the summary statistics for the review year variable, it is seen that the scope of the dataset extends from 2006 to 2017. The median review year is 2013, suggesting a central tendency around this year, and the mean review year is 2012, indicating a slight left-skew (see Exhibit 2 in Appendix).

For the cocoa percentage variable, the statistical summary indicates a range from a minimum of 42% to a maximum of 100%. The average cocoa percentage is about 71.7% and the distribution is slightly right skewed.

The correlation between the two numerical variables, review date and cocoa percentage was also analysed, and it was revealed that with a value close to zero, there is a very weak positive correlation between these two variables, suggesting that changes in review date are not strongly associated with changes in the percentage of cocoa.

For the independent categorical variables, a different approach was used. To specify, the categorical variables that were examined during this step included company (originally called Company (Maker-if known), specific bean origin or bar name, company location, bean type, and

broad bean origin. For each categorical variable, the top 5 categories were identified in terms of frequency and visualized using bar plots. Percentage distributions were also examined. It was determined that the top five chocolate companies are Soma, Bonnat, Fresco, Pralus, and A. Morin which collectively account for approximately 80% of the observations. Madagascar, Peru, Ecuador, Dominican Republic, and Venezuela emerge as the most frequent origins or bar names, making up nearly 100% of the dataset, with Madagascar alone representing almost 28%. The majority of the chocolate samples originate from the United States (63.48%), followed by France (12.98%), Canada (10.32%), the United Kingdom (7.99%), and Italy (5.24%). In numerous instances, the bean type field was left blank. To simplify interpretation, these instances were updated to unknown. Bean types primarily fall into four categories: Trinitario (55.54%), Criollo (26.17%), Forastero (9.58%), and Forastero (Nacional) (5.45%). All other bean types contribute to the remaining 3.26%. Noteworthy broad bean origins include Venezuela (24.94%), Ecuador (22.49%), Peru (19.23%), Madagascar (16.90%), and the Dominican Republic (16.43%), reflecting the diversity of the dataset. These frequencies were also visualized using histograms (see Exhibit 3 in Appendix).

Following the distribution of the independent variables, their relationships with the target variable was examined. For the numerical variables, linear regression was run to see if these two variables contribute significantly to determining the rating, and both variables seemed to be statistically significant in relation to rating. A scatter plot was also developed for each variable and rating but nothing of value could be concluded as the scatter plots showed no patterns (see Exhibit 4 in Appendix).

For the categorical variables, a chi-squared test was run on each variable in relation to the target variable and it was revealed that only two categorical variables significantly contributed to rating which were company and the country of the company.

Model Selection and Methodology

After getting a clearer understanding of the dataset, preprocessing could be done to filter out the key variables that would be relevant in building the models.

Data Preprocessing

After completing the exploratory data analysis, it became clear that many variables in the dataset were not contributing at all to predicting the rating of a chocolate bar. From the categorical variables, broad bean origin, specific bean origin or bar name, and bean type were all removed. Additionally, the two numerical variables were standardized to ensure that all variables contribute equally to the model, preventing variables with larger scales from dominating the model due to their larger range of values.

Classification

In the development of the logistic regression classification model, the primary objective was to predict chocolate ratings based on diverse features. The response variable, rating, underwent a binary transformation, classifying ratings exceeding 3.5 as 1 (indicating positive reviews) and the rest as 0. To accommodate categorical variables like company and the company's location, the logistic regression approach incorporated one-hot encoding, generating dummy variables for each unique category. These dummy variables were subsequently added to the dataset. The logistic regression model's formulation included numerical features, such as cocoa percentage and the review year, in addition to the dummy variables representing both company and location. The model was trained comprehensively on the entire dataset, and predictions were derived from the logistic regression probabilities.

The random forest classification model was also employed to predict chocolate ratings, leveraging a different approach. The target variable, rating, underwent a binary transformation similar to the logistic regression model. One-hot encoding was applied to handle categorical variables like company and company location, creating dummy variables for each category. These dummy variables were integrated into the dataset, contributing to the development of the random forest model. Unlike logistic regression, random forest is an ensemble method that incorporates multiple decision trees. A consideration was made to include the other categorical variables that were removed earlier but the additional increase in accuracy score was not enough for the complexity that the model would bring in. The utilization of random forest in this context provided a more robust and accurate prediction of chocolate ratings, highlighting its efficacy in handling the intricacies of the dataset.

Clustering

The final model was developed through a systematic approach to analyze the chocolate dataset, employing a combination of feature engineering, machine learning, and clustering techniques. The initial preprocessing stage involved the creation of dummy variables for categorical features, just as in both models above. A Random Forest model, configured with 500 trees, was then employed to assess feature importance. In the process of deciding which features to utilize for the clusters, consideration was given to the top 10 features. However, a selection was made to focus on a subset of 6 distinct features that were deemed to be of particular interest. These included review year, cocoa percentage, company Soma, company Bonnat, company location France, and company location U.S.A.

The core of the analysis involved applying k-means clustering to this refined subset with the aim of unveiling inherent patterns and groupings within the data. An elbow plot was utilized to determine the optimal number of clusters, which was found to be four. The resulting clusters offered valuable insights into the distribution of observations across the identified groups. The cluster assignments were then integrated back into the original dataset.

Results

The final models built using logistic regression and random forest algorithms achieved accuracy scores of 85% and 90% respectively. The random forest model, which inherently captures complex relationships in the data, outperformed the logistic regression model in terms of predictive accuracy. The top five features contributing to the prediction in the random forest model were the year of review, cocoa percentage, and the presence of companies Soma, Idilio (Felchlin), and Amedei.

The results of the clustering model showed four clusters, and their cluster means (see Exhibit 5 in the Appendix) can be interpreted to understand the key features in each cluster. Cluster 1 has chocolates with a below-average cocoa percent and older review dates. The chocolates in this cluster are primarily from the USA and France. The presence of companies Soma and Bonnat is relatively low. Cluster 2 has chocolates with a slightly below-average cocoa percent and more

recent review dates. All chocolates in this cluster are from the USA. There is no presence of chocolate from either company Soma or Bonnat, or chocolates from France. Cluster 3 has chocolates with a below-average cocoa percent and the most recent review dates. The chocolates in this cluster are primarily from France. The presence of chocolates from company Soma is relatively high compared to other clusters, while the company Bonnat has a low presence. There are no chocolates from the USA in this cluster. Cluster 4 has chocolates with a significantly above-average cocoa percent and average review dates. The chocolates in this cluster are primarily from the USA with a lower presence from France. The presence of chocolates from companies Soma and Bonnat is relatively low.

Classification/Predictions and Conclusions

The predictive models built using logistic regression and random forest algorithms achieved accuracy scores of 85% and 90% respectively. The top five features contributing to the prediction in the random forest model were the year of review, cocoa percentage, and the presence of companies Soma, Idilio (Felchlin), and Amedei. The random forest model outperformed the logistic regression model, indicating its superior ability to capture complex relationships in the data. Additionally, the clustering analysis revealed four distinct clusters in the data. These clusters were differentiated by factors such as cocoa percent, review date, and the presence of specific companies and countries of origin.

The random forest model's superior performance suggests that future predictive modeling efforts might benefit from focusing on tree-based methods, especially when dealing with complex datasets with potential non-linear relationships and interactions between variables.

The importance of features like the year of review and cocoa percentage in the random forest model suggests that these factors significantly influence the binary rating. This information could be useful for companies looking to improve their ratings.

The clustering results provide valuable insights into the segmentation of chocolates. Understanding these segments can help companies tailor their products and marketing strategies to the preferences of different segments. For example, one segment prefers chocolates with a

high cocoa percent and is less concerned about the company of origin. Another segment prefers chocolates from specific companies and countries.

The presence of companies like Soma, Idilio (Felchlin), and Amedei in the top features suggests that these companies have a significant influence on the binary rating. This could be due to various factors such as the quality of their products, branding, customer perception, etc. Other companies might look into the practices of these influential companies for potential areas of improvement.

In conclusion, the analysis of the chocolate dataset using machine learning models and clustering algorithms provided valuable insights. The predictive models, particularly the random forest model, were able to accurately predict the binary rating of chocolates based on features such as the year of review, cocoa percentage, and the presence of specific companies. The clustering analysis further revealed distinct segments in the chocolate market, differentiated by factors such as cocoa percent, review date, and the presence of specific companies and countries of origin. These findings can guide companies in tailoring their products and marketing strategies to cater to the preferences of different segments, potentially improving their ratings and market position. Future predictive modeling efforts might benefit from focusing on tree-based methods, especially when dealing with complex datasets with potential non-linear relationships and interactions between variables.

Appendix

Exhibit 1: Distribution of Rating Variable

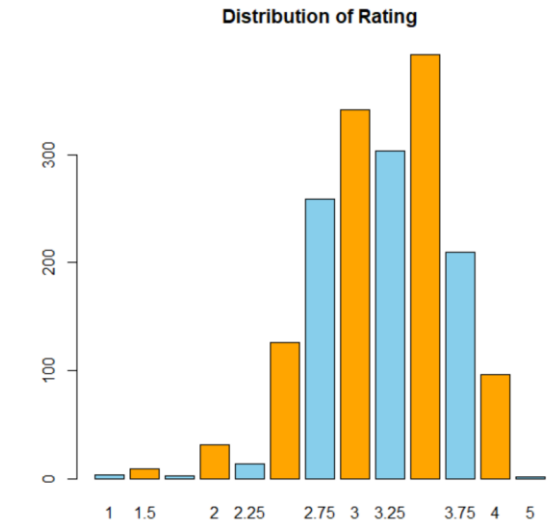

Exhibit 2: Distribution of Review Year and Cocoa Percentage Variables

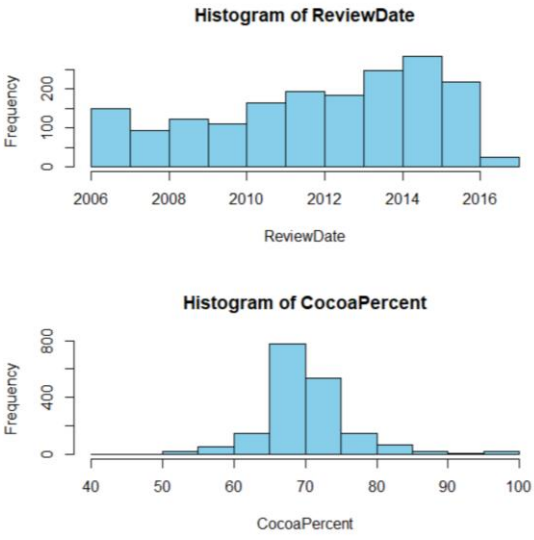

Exhibit 3: Bar Plots for Frequencies of Top 5 Categories of Each Categorical Variable

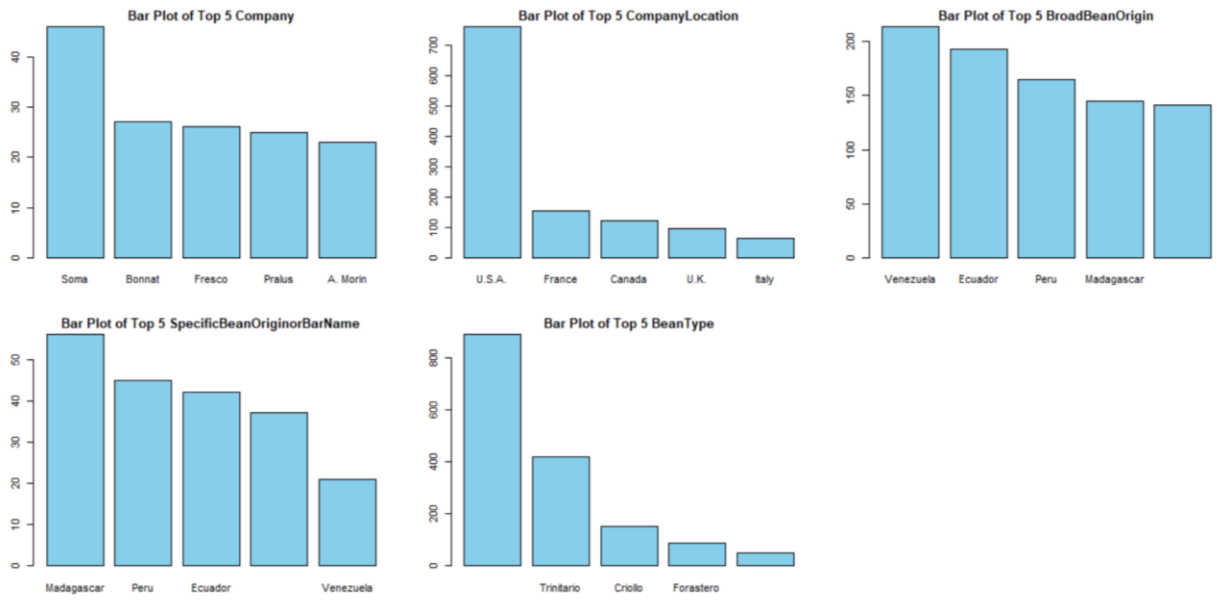

Exhibit 4: Scatter Plot of Each Numerical Variable with Rating

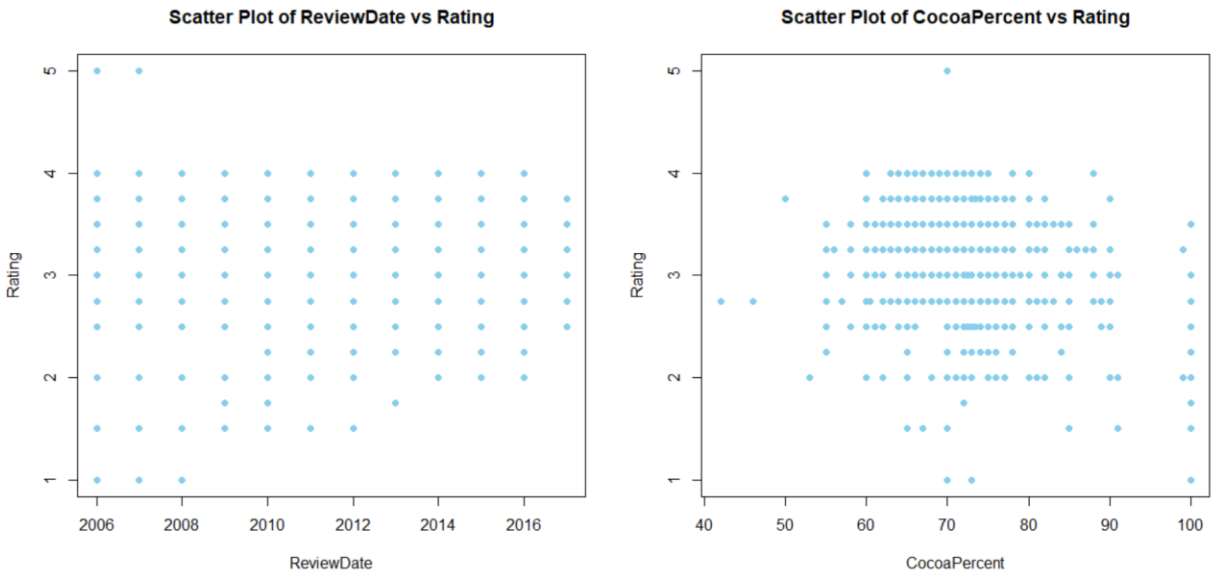

Exhibit 5: Cluster Means

```
Cluster means:
   ReviewDate CocoaPercent CompanySoma CompanyBonnat CompanyLocationFrance CompanyLocationU.S.A.
1 -1.50417172  -0.53825024  0.01754386   0.035087719           0.20467836           0.2690058
2 -0.28447020  -0.09735809  0.04500978   0.019569472           0.07436399           0.4050881
3  0.01275924   2.05760528  0.01604278   0.005347594           0.04812834           0.3582888
4  0.87304784  -0.20045235  0.01859230   0.005312085           0.05179283           0.5272244
```