Dark chocolate Secrets

TONY XU

Tony Xu

261173499

Dark Chocolate Rating: The Key to Wonka Factory

Introduction:

The global cocoa and chocolate market, valued at approximately $60 billion and growing annually by 5%, is witnessing increased popularity, especially in developing nations. This surge in interest positions chocolate as a vital snack sector for food companies and entrepreneurs. A notable trend is the rising demand for high-percentage cocoa chocolates, prized by enthusiasts for their health benefits. Consequently, companies are dedicating more research to this niche.

This project aims to analyze the factors contributing to exceptional dark chocolate using data-driven methods, specifically targeting the dark chocolate niche. The findings will guide companies in sourcing cocoa beans, selecting top manufacturers, and determining the optimal cocoa solid content for their products. It will conduct this finding using Decision Tree and PCA to help interested stakeholders better start their chocolate ventures.

Data Description and EDA:

The dataset encompasses 1795 unique chocolates, each with detailed information and ratings, covering the years 2006 to 2017. The median year represented is 2013. Among the 416 unique chocolate-making companies featured, 'Soma' stands out with the highest frequency of 47 instances. The companies are spread across 60 countries, with the USA leading in representation. Chocolate ratings range from 1 to 5 and exhibit a fairly normal distribution, as depicted in the attached graphs. In terms of bean types, there are 41 identified varieties, but about half of these entries are missing. The dataset also highlights 100 different bean origins, with Venezuela as the most common source.

From an initial examination, several observations emerge:

The dark chocolate sector demonstrates robust competition. With a multitude of companies involved, the largest, 'Soma', constitutes a mere 2.6% of the total sample. This diversity suggests a competitive yet accessible market for newcomers, supported by the presence of many smaller-scale companies.

Geographic concentration is evident, with 43% of these companies based in the USA and France trailing with less than 10%. This indicates that the USA may possess the largest chocolate market, potentially offering more room for growth. This is supported by a graph from Mordor Intelligence in the Appendix, which might be indicative of a strong talent and production infrastructure within the USA.

The extensive variety in bean origins nations (100 unique entries) offers supply chain resilience but also presents challenges in selecting the most fitting beans for different markets. The subtle taste variations of beans from different regions could significantly impact the final product's rating. Preliminary data also suggests a complex relationship between cocoa percentage and chocolate ratings. A higher cocoa content does not consistently lead to higher ratings. Further analysis is anticipated to shed light on these two aspects.

The Appendix contains further Exploratory Data Analysis (EDA) graphs and additional insights from Mordor Intelligence, providing a broader context to these findings.

Models:

Two classification models were developed for the purpose of categorizing chocolates. One model predicts whether a chocolate falls into the lowest graded category (rating < 2.5), while the other predicts whether a chocolate achieves a higher rating (rating > 3.5). The overarching objectives of these two models are twofold: 1) To identify and avoid the potential pitfalls associated with producing subpar chocolate.2) To investigate the combinations of factors that lead to the production of the highest-rated chocolates.

The preprocessing of the datasets involves several key steps:

1) . Conversion of continuous rating values into categorical labels, distinguishing between low and high ratings.

2) Removal of the 'Company' column and the introduction of a new feature 'company_count' denoting the count of entries associated with each chocolate company. A similar process is applied to the 'broad_origin' column, resulting in the creation of the 'origin_count' feature. These steps are taken to address the challenge of dealing with a high number of categorical variables.

3) The calculation of the average rating score for each chocolate's place of origin and creation company is performed and added as a new feature. This serves to highlight the quality of cocoa beans from specific geographical regions.

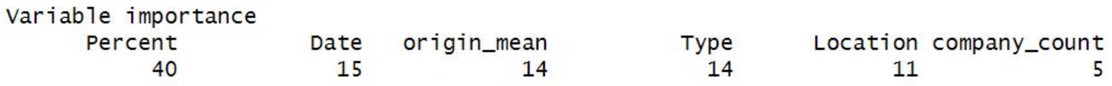
4) The above two steps were repeated for the ‘location’ column in the PCA scenario to address high number of potential categorical variables.

In the modeling phase, Decision Trees are chosen as the preferred algorithm due to their inherent interpretability. The primary focus of this project is not centered on achieving the highest classification accuracy, hence Random Forest is intentionally omitted. Optimal complexity parameters are selected to fine-tune the final Decision Tree models.

In addition, Principal Component Analysis (PCA) was utilized to illustrate the features that have a correlation with chocolate ratings and the magnitude of these correlations. Unlike previous models that distinguished between lower and higher quality chocolates based on certain factors, this model centers on the overall rating score. The primary aim of employing PCA here is to offer a concise overview of the key features that most significantly influence the chocolate ratings.

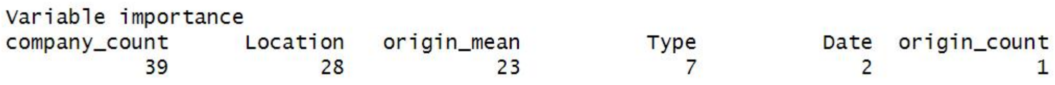
The resulting Decision Trees and PCA results, accompanied by feature importance scores and PCA graph, are attached in the Appendix.

Results, Recommendations to Stakeholders:



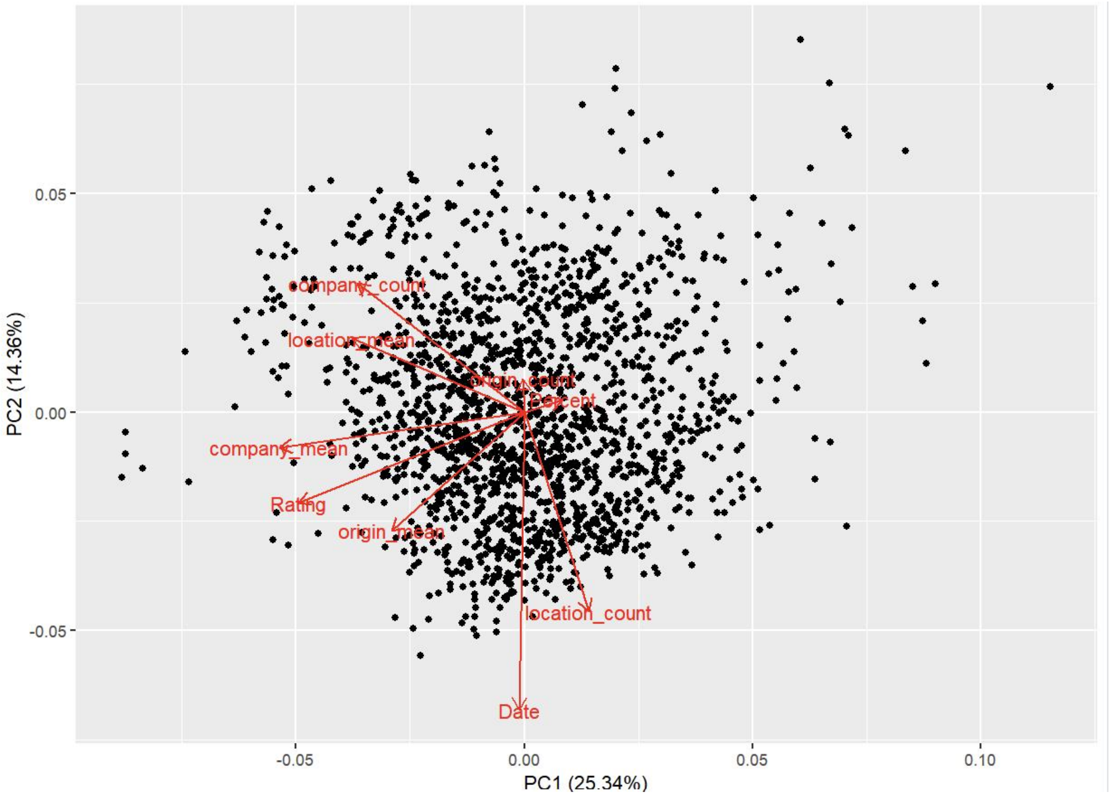
```
Variable importance
      Percent          Date   origin_mean          Type      Location company_count
           40            15            14            14            11             5
```

Variable Importance Score from Low Rating Classification



```
Variable importance
company_count      Location   origin_mean          Type          Date  origin_count
           39            28            23             7             2             1
```

Variable Importance Score from High Rating Classification



PCA Graph

The PCA graph enables us to identify which features have the strongest positive and negative correlations with the rating column. It reveals that the two primary factors potentially boosting a chocolate's rating are 1) the historical average rating of the chocolate company and 2) the average rating of the cocoa beans' origin. The graph also indicates a lesser yet notable influence from the number of chocolates produced by the company and the country of production. Interestingly, cocoa percentage shows a negative correlation with the rating, but the strength of this correlation is relatively weak.

From the analysis of two decision trees in the chocolate industry, several key insights emerge that correspond nicely with the insights from the PCA analysis:

1) . Despite chocolate lovers expressing a preference for high cocoa content, the data

reveals that chocolates with cocoa percentages above 91% are more likely to receive lower ratings. This suggests a critical threshold in cocoa content affecting consumer satisfaction.

2) The mean score of cocoa beans from different origins is a significant predictor of chocolate quality. A lower mean score from a bean's origin correlates with a higher likelihood of a lower product rating, underscoring the importance of sourcing high-quality beans from regions with a strong reputation for excellence.

3) Contrary to expectations, there isn't a direct correlation between the number of products a company offers and higher chocolate ratings. However, location plays a pivotal role. Notably, U.S.-based companies, despite operating in the world's largest chocolate market, often face challenges in achieving high ratings.

4) While the type of beans is an important factor, it does not rank among the top determinants of chocolate quality in either decision tree model, indicating other factors might play more significant roles.

Based on these insights, for those looking to launch a dark chocolate brand, the following strategies are recommended:

1) . Opt for establishing the company in smaller, yet renowned markets like Belgium or Switzerland, rather than following major players.

2) . Aim for a balanced cocoa percentage, avoiding extremes.

3) . Prioritize sourcing cocoa beans from regions with proven quality records.

4) . For large food companies, it might be more effective to start with a niche brand to build reputation, rather than leveraging the name of existing major chocolate brands.

In conclusion, this project demonstrates the power of data-driven approaches in crafting new products. By leveraging insights from the model, businesses can make informed decisions to enhance product quality and market appeal. This methodology not only aids in identifying key factors influencing consumer preferences but also provides a strategic roadmap for new and established chocolate brands alike.

=

Appendix:

Figure 1: Score Distribution

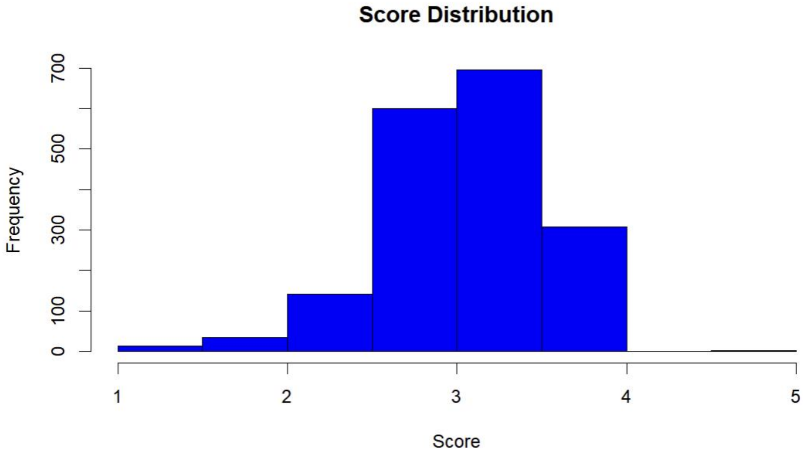


Figure 2: Coca Percentage Distribution

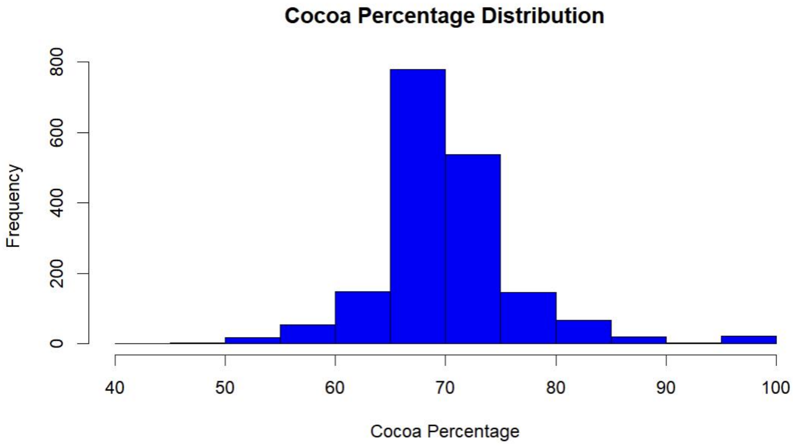


Figure 3: Coca Percentage Distribution and Rating Scatter plot

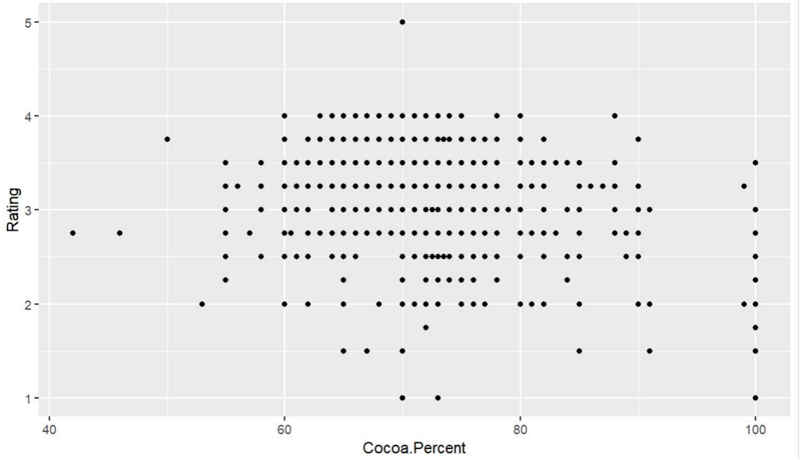
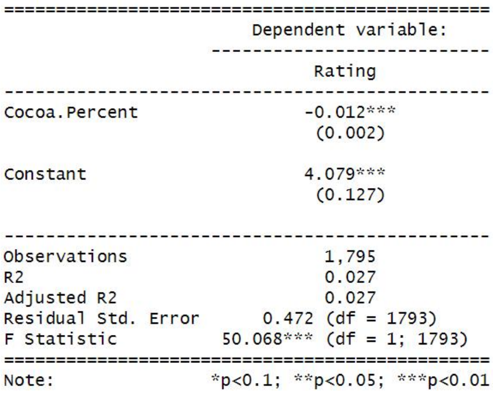


Table 1: Rating Regression Against Coca Percentage

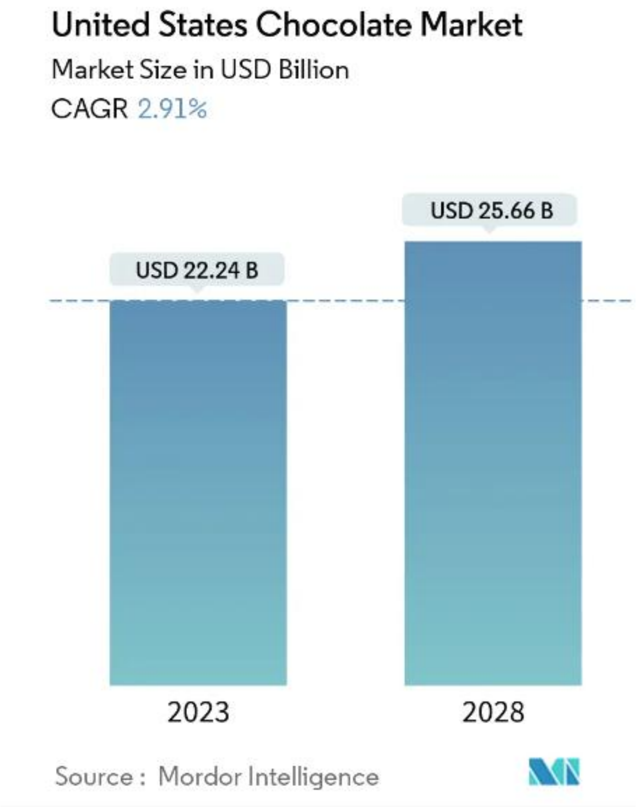


```
===============================================
                        Dependent variable:
                    ---------------------------
                               Rating
-----------------------------------------------
Cocoa.Percent                 -0.012***
                               (0.002)

Constant                       4.079***
                               (0.127)

-----------------------------------------------
Observations                    1,795
R2                              0.027
Adjusted R2                     0.027
Residual Std. Error       0.472 (df = 1793)
F Statistic           50.068*** (df = 1; 1793)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Figure 4: US Chocolate Market Size



Source : Mordor Intelligence

Figure 5: Decision Tree for low category

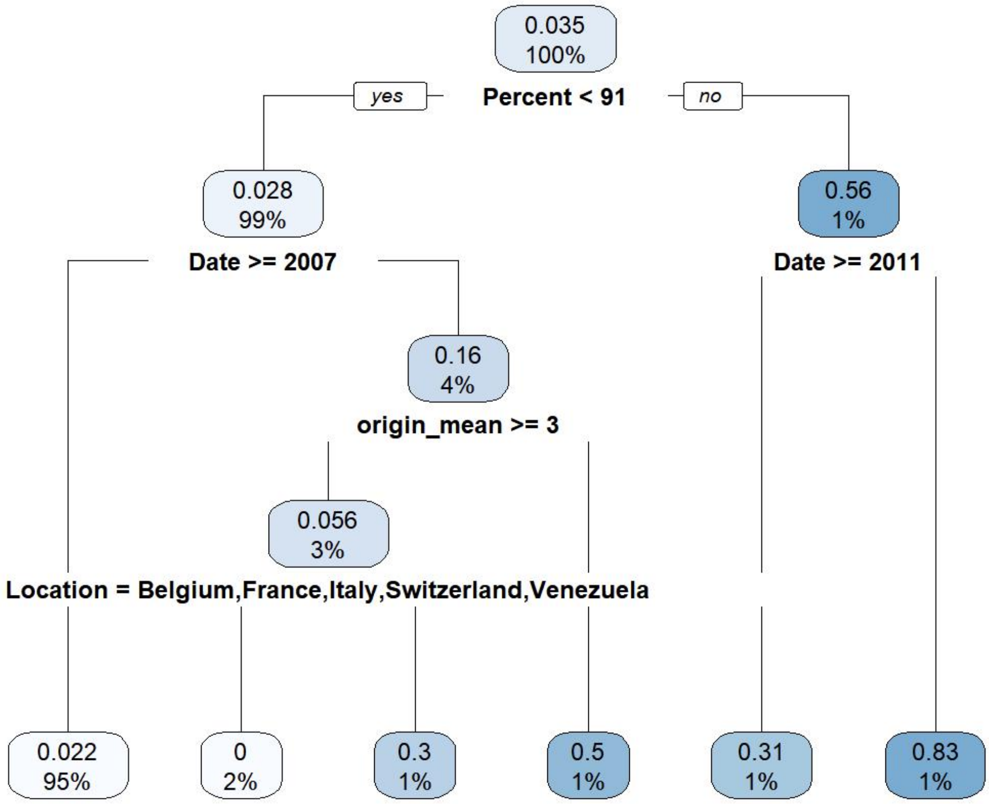


Figure 6: Decision Tree for high category

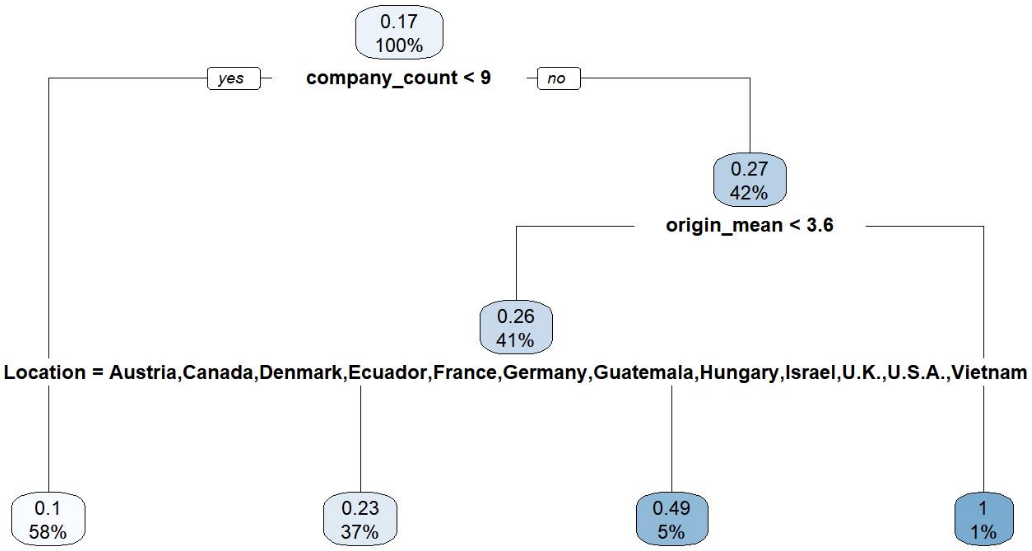
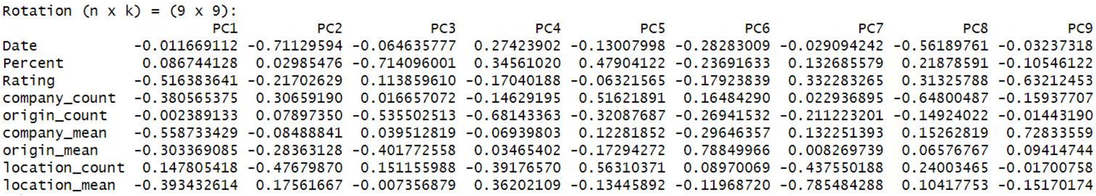

0.17
100%
yes
company_count < 9
no
0.27
42%
origin_mean < 3.6
0.26
41%
Location = Austria,Canada,Denmark,Ecuador,France,Germany,Guatemala,Hungary,Israel,U.K.,U.S.A.,Vietnam
0.1
58%
0.23
37%
0.49
5%
1
1%

Table 2: PCA results



```
Rotation (n x k) = (9 x 9):
                        PC1         PC2          PC3         PC4         PC5         PC6          PC7         PC8         PC9
Date           -0.011669112 -0.71129594 -0.064635777  0.27423902 -0.13007998 -0.28283009 -0.029094242 -0.56189761 -0.03237318
Percent         0.086744128  0.02985476 -0.714096001  0.34561020  0.47904122 -0.23691633  0.132685579  0.21878591 -0.10546122
Rating         -0.516383641 -0.21702629  0.113859610 -0.17040188 -0.06321565 -0.17923839  0.332283265  0.31325788 -0.63212453
company_count  -0.380565375  0.30659190  0.016657072 -0.14629195  0.51621891  0.16484290  0.022936895 -0.64800487 -0.15937707
origin_count   -0.002389133  0.07897350 -0.535502513 -0.68143363 -0.32087687 -0.26941532 -0.211223201 -0.14924022 -0.01443190
company_mean   -0.558733429 -0.08488841  0.039512819 -0.06939803  0.12281852 -0.29646357  0.132251393  0.15262819  0.72833559
origin_mean    -0.303369085 -0.28363128 -0.401772558  0.03465402 -0.17294272  0.78849966  0.008269739  0.06576767  0.09414744
location_count  0.147805418 -0.47679870  0.151155988 -0.39176570  0.56310371  0.08970069 -0.437550188  0.24003465 -0.01700758
location_mean  -0.393432614  0.17561667 -0.007356879  0.36202109 -0.13445892 -0.11968720 -0.785484288  0.10417753 -0.15170174
```