DataCamp

Data Science Certification

Case Study on

["Coffee Aftertaste Quality Analysis"]

Capstone Project

Submitted by	Submitted to
Shusant Sapkota	DataCamp
shusant.sapkota@gmail.com	

Submitted on: July 27, 2021

1. Motivation

The coffee is prepared from roasted seeds of berries belonging to Coffea species [1]. In 2019/20, the global coffee consumption is estimated to be 167.59 million bags [2]. Finland is the country where an individual consumes 12kg of coffee on average followed by Norway 9.9kg per annum [3]. The quality of coffee is measured/rated in ten areas, they are Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Cupper Points, and Sweetness.

Aroma is a fragrance we get from the brewed coffee, Flavor is the taste i.e., the overall perception of coffee, Body also termed as the mouthfeel is the tactile impression on the palate when coffee coats tongue, Acidity refers to briny sensation on the tongue tip or tart taste near the back of the tongue, Sweetness is the smoothness and mildness of coffee [4]. The Aftertaste is the final sensory experience when tasting a coffee, it is highly dependent upon other sensory quality attributes of coffee as mentioned above. Furthermore, it seems the quality is also impacted by moisture content, and bean defects [4] [5].

This project aims in finding the major quality coffee attributes that affect the overall Aftertaste of coffee. The Dataset provided by DataCamp is used in analyzing the factors affecting Aftertaste. Finally, the project aims to prepare a machine learning model to predict the Aftertaste value based on the factors that affect it. The findings of this project can be used to improve the Aftertaste of a coffee and a machine learning model can be used to predict the Aftertaste from the known attributes.

2. Success Criteria

Dataset

The dataset has been provided by DataCamp. It has 43 input features and a target variable namely Aftertaste. The columns can be categorized into three sub-categories

i. Quality Measures: Aroma, Flavor, Aftertaste, Acidity, Body, Balance, Uniformity, Clean Cup, Sweetness, Moisture, Defects
ii. Bean Metadata: Processing Method, Color, Species
iii. Farm Metadata: Owner, Country of Origin, Farm Name, Lot Number, Mill, Company, Altitude, Region

Out of these features, the quality measures are analyzed about their impact on Coffee Aftertaste and model fitting. Besides this, the impact of categorical data is also studied.

Metrics

The metrics are used to evaluate the performance of our trained model. R2-score is used for evaluating the performance of our model. It is calculated using the following formula.

$$R^2 = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma\left(y - \bar{y}\right)^2}$$

Where ȳ – mean

ỹ - predicted value

The higher the value or R2, the higher is the accurate performance of the model. Most of the Machine Learning algorithm libraries provided by Scikit-Learn consists of method 'score' which takes the input of features and target and returned the corresponding R2 score of the model.

3. Analysis Plan

This is a supervised regression problem because the target variable 'Aftertaste' is a continuous value. The following steps will be followed for solving this supervision problem:

i. Initially, the dataset is loaded, initial insight is observed, and NULL value distributions in columns are analyzed.
ii. Exploratory data analysis (EDA) is performed to discover initial insights into the data.
iii. Make appropriate visualization analysis about features concerning target variables to discover the important features.
iv. The final dataset is prepared with the most important features. It is divided into training and testing sets.
v. The training set is used in training several models of different Supervised ML Algorithms. The training phase is done in two ways, with more features and fewer features.
vi. The different models with fewer and more features in training are compared against each other in terms of training and testing accuracy. Based on this, the best model is selected.
vii. Discuss the metrics and study ways of improvements in the final model.

9. Outcome

From the Exploratory analysis, the Aftertaste of coffee is largely dependent upon these attributes Flavor, Acidity, Cupper Points, Balance, Aroma, Body, Uniformity, Clean Cup, Sweetness, Moisture, and Category Two Defects. When these attributes are visually analyzed, Moisture Category Two Defects, Uniformity, Clean Cup, and Sweetness have minimal impact in Aftertaste. Hence, Flavor, Acidity, Cupper Points, Balance, Body, and Aroma affect the Aftertaste of coffee largely. The Flavor feature has a correlation of 0.9 with Aftertaste. While visualizing the relationship between Aftertaste and Flavor, the data points densely lie along a positively skewed line. Thus, Flavor is the most important feature which impacts significantly on the Aftertaste of Coffee. This level of importance is followed by Balance and Cupper Points.

A dataframe with these final features is selected along with the target variable. This frame is divided into training and testing sets. The training set is used in training six different machine learning algorithms. Each model is hyper-tuned manually. These trained models are evaluated by both training set and testing set using the R-square approach. Most the model gets overfitted and the Decision Tree model have the largest gap between training and testing accuracies whose respective values are 99.9% and 55.5%. This problem is minimal in Linear Regression with a training accuracy of 79.3% and testing accuracy of 78.5%.

In the second case, we take only four features dropping Aroma and Body. Thus, the features that are taken are Flavor, Balance, Cupper Points, and Acidity. Again, the models are trained and evaluated similarly. The problem of overfitting is minimal in most of the models except the Decision Tree. The training and testing accuracies in the Decision Tree are 98.3% and 64.3% respectively. Though, the testing accuracy increases in the Decision Tree model, it is still enormously underperforming in unseen data. The Support Vector Machine performs better than others. Its training and testing accuracies are 82.7% and 81.7% respectively and this model is saved for future purposes.

10. Future Work

Future enhancements of this project are listed below:

i. Though we develop the model. But it is not deployed in a real-world scenario. It can be deployed through a proper website/app and many people like research scientists, coffee sellers, farmers, and mills can use this deployed model easily for predicting and analyzing the overall coffee Aftertaste.

ii. Since we have tested the model performance using training and testing data set which were made by splitting the initially given dataset. We do not know how our analysis and model are relevant in a real-world case. Detail research can be done on this analysis and model performance by making a new dataset of the latest coffee data.

iii. The neural network can be used in training the new model. Since the neural network can deal with large data and can find the most complex pattern within data. Thus, we can increase the feature space and train the model.

Furthermore, the dataset has 44 columns of different data. In this project, only the factors affecting the Aftertaste of coffee are analyzed. In the future, the target variable can be changed to any other like Category Two Defects, Category One Defects, etc. and the features impacting them can be studied. There are many other features of Farm metadata which include Country of Origin, Region, Producer, etc. These sorts of data can be used in making advanced analyses of coffee production scenarios based upon these scenarios. These future findings might be beneficial for coffee exporters, sellers.

11. References

[1] National Coffee Association USA, "What is Coffee?," NCAUSA, [Online]. Available: https://www.ncausa.org/About-Coffee/What-is-Coffee. [Accessed 23 July 2021].

[2] International Coffee Organization, "Coffee Market Report," October 2020. [Online]. Available: https://www.ico.org/news/cmr-1020-e.pdf. [Accessed 21 July 2021].

[3] K. Bernard, "WorldAtlas," 6 August 2020. [Online]. Available: https://www.worldatlas.com/articles/top-10-coffee-consuming-nations.html. [Accessed 23 July 2021].

[4] Espresso and Coffee Guide, "What is Coffee Quality?," Espresso and Coffee Guide, 2006. [Online]. Available: https://espressocoffeeguide.com/quality-coffee/. [Accessed 23 July 2021].

[5] G. Oden, "Java Presse," [Online]. Available: https://www.javapresse.com/blogs/enjoying-coffee/how-to-taste-coffee-aftertaste. [Accessed 23 July 2021].