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.