DataCamp

Data Science Certification

Case Study on

[“Coffee Quality Dataset Analysis”]

Capstone Project

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

Submitted on: July 20, 2021

1. Motivation

The coffee is prepared from roasted seed of berries belonging to Coffea species. In 2019/20, more than 166 million bags of coffee are consumed. Finland is the country where an individual consumes 12kg of coffee in average followed by Norway 9.9kg per annum. 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 fragrance we get from the brewed coffee, Flavor is the taste of coffee, Body also termed as mouthfeel is 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 tongue, Sweetness is the smoothness and mildness of coffee. The Aftertaste is final sensory experience when tasting a coffee. It is highly dependent upon other sensory quality attributes of coffee as mentioned above. Different researches show that, the quality is also impacted by moisture content, and bean defects.

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

Metrics

The metrics is 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, $\bar{y}$ – mean

$\tilde{y}$ - predicted value

The higher the value or R2, higher is the accurate performance of 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, after loading dataset unnecessary features and columns with most null values are dropped.
ii. Exploratory data analysis (EDA) is performed to discover initial insights of the data.
iii. Make appropriate visualization analysis about features with respect to target variables. From this appropriate feature are selected
iv. The features are used in training several models of different algorithms. The training phase is done in two ways, with more features and less features.
v. The different models are compared against each other in terms of training and testing accuracy. Best model is selected.
vi. Discuss the merits of and improvements of the model.