Exploratory Data Analysis and Prediction of International Coffee Prices with a Variety of Algorithms of Machine Learning

Simon Atehortua

ID 501094927

Data Analytics, Big data, and Predictive Analytics, Toronto Metropolitan University

CIND 820D1H: Big Data Analytics Project

Prof. Tamer Abdou, PhD

December 5, 2022

Abstract

Coffee is one of the greatest drinks that humanity could have ever discovered. The prices of coffee are expressed in US cents per pound of green coffee and many factors affect the prices. The ICO (International Coffee Organization), will collect daily prices from the New York, Germany, and France futures exchanges and will set up a price based on these factors and a variety of conditions.

The idea of this project and after doing an exploratory analysis of the data is to predict the prices of the ICO market given the prices in Europe, more exactly in the countries of Germany and France, as well as in the United States of America of a different variety of Coffee that is grouped as follows: Colombian Mild Arabicas (Colombian Excelso UGQ screen size 14, Colombian Excelso European preparation screen size 15), Other Mild Arabicas (Costa Rica hard bean, Mexico Prime washed, Honduras high grown, Guatemala prime washed, El Salvador Strictly High grown, Guatemala hard bean, Honduras High grown European preparation) Brazilian (Brazil Santos ¾ screen size 14/16, Brazil Santos ⅔ screen size 17/18, Brasil Santos ¾ screen size 14/16) and Robustas (Vietnam grade 2, Indonesia EK grade 4, Uganda Standard, Côte d'Ivoire grade 2).

All data for this project comes from Federación Nacional De Cafeteros de Colombia (The Colombian National Coffee Growers), which is the only authorized entity by the government of Colombia, that is responsible for exporting, buying, and setting the prices of coffee, the website (https://federaciondecafeteros.org/wp/coffee-statistics/?lang=en), section Coffee prices, area, and

production have an excel file with the data required during this project. The data will be cleaned and checked for any inconsistencies or errors that can alter our results, after that, there will be an Exploratory Data Analysis to check the behavior of the time series data, and to possibly discover new information that could help find important and relevant information about our study.

Through a variety of different Machine Learning regression algorithms like linear regression, SVM (Support Vector Machine), KNN regressor (K-Nearest Neighbors), Decision Tree regressor, and Neural Networks, I will intend to predict the future price of the coffee set by the International Coffee Organization.

A variety of tools will be used for this project: Jupyter Notebook, Python, R studio, Microsoft Excel, Github, and Google Documents.

During this project, a variety of questions have to be formulated in order to obtain the best answer to our results. An example of these questions, among many others, are: which is the most accurate algorithm to predict the ICO prices of coffee and why the selection of this algorithm? What will be the future price in the short term given that the model has been adequately trained and formulated? Why is this investigation relevant to the student and how it can help to develop his knowledge not only in exploratory data analysis but as well in machine learning algorithms?

Keywords: coffee, exploratory data analysis, regression analysis, prediction, machine learning algorithms

References

Federación Nacional de Cafeteros de Colombia. (2022, septiembre 16). Precios, área y producción del café. Bogota; https://federaciondecafeteros.org/app/uploads/2020/01/Precios-%C3%A1rea-y-producci%C3%B3n-de-caf%C3%A9-3.xlsx.

Statistics Committee. (2021, March 29). sc-106e-rules-indicator-prices.pdf. London; International Coffee Organization.