CX4240 Final Project Report

Jacob Schroeder (903584648)

2024-04-15

Introduction

Dom Pérignon is a renowned brand of vintage Champagne produced by the French Champagne house Moët & Chandon, which is a part of the luxury goods company LVMH (Louis Vuitton Moët Hennessy). It is named after Dom Pérignon, a Benedictine monk who made important contributions to the production and quality of Champagne in the 17th century.

Dom Pérignon Champagne is known for its exceptional quality and prestige. It is made exclusively from Chardonnay and Pinot Noir grapes grown in the Champagne region of France. Each bottle is produced as a vintage Champagne, meaning it is made from grapes harvested in a single year, and only the best grapes are selected for production.

Dom Pérignon Champagne is characterized by its complexity, depth, and elegance. It undergoes a meticulous process of production, including hand harvesting, gentle pressing of the grapes, and aging in the cellars for several years before release. The final product typically exhibits a fine balance of fruitiness, minerality, and a creamy texture, with delicate bubbles.

Dom Pérignon is often regarded as one of the finest and most prestigious Champagnes in the world, sought after by collectors and enthusiasts alike for special occasions and celebrations. Its distinctive bottle design and iconic label contribute to its status as a symbol of luxury and excellence in the world of Champagne.

Motivation

The short answer here is curiosity. Dom Pérignon is not produced every year. Only certain years produce a grape of high enough quality that they release a batch of champagne. The last vintage was released in 2013. Can we use different weather data along with machine learning classifiers to predict which new years will yield Dom Pérignon vintages.

Data

Data is collected from two sources. The first source is where all of the necessary features were collected from. This is all of the weather data collected from the Open-Meteo API. Data collected from the source include three different parameters for every day of the year from January 1, 1940 until December 31, 2023. The labels were webscraped from the Dom Perignon Wikipedia Page.

Features

The three features used in the classification model were:

- Average daily temperature at 2 meters in altitude
  - Temperature is really important for grapevines. In Champagne, where it's usually cool, grapes grow well, with a good balance of acidity and sugar. But big swings in temperature can mess up the timing of important stages like when buds appear, flowers bloom, grapes change color, and when they're ready to pick. If it gets too hot or too cold, there might be fewer grapes, they might

taste different, or the vines could get stressed or harmed. Warmer weather can also mean grapes are ready to harvest earlier, which can change how the wine tastes compared to the traditional style of the region.

- Total daily precipitation
  - Rain is important for helping grapevines grow by giving them enough water. But too much rain, especially when the grapes are flowering or starting to form, can cause problems like bad pollination, more diseases, and making the grapes taste less flavorful. On the other hand, if there's not enough rain for a long time, it can stress out the vines, make fewer grapes grow, and make the grapes taste sweeter. Getting the right amount of rain at the right time is really important for keeping the grapes healthy and making good wine in Champagne.
- Mean shortwave radiation
  - Sunlight is really important for grapevines because it helps them make sugars and other stuff they need to ripen. Having enough sunlight makes the grapes ripen evenly, gives them good color, and makes them taste better. But if there's too much sunlight and heat, it can burn the grapes and make them ripen unevenly, which isn't good for the wine. Also, changes in clouds and weather can change how much sunlight reaches the grapes, which can affect how they grow and taste.

Other factors impact grape production and the resulting wine; however, for the purposes of this study we decided to focus on only these three. Further analysis could involve using extreme values rather than averages as those can have dramatic impacts on growth.

Labels

The labels for this data are binary. Either a year produces a Dom Pérignon vintage or a year does not. For the purposes of building the classification model, these will be represented as boolean values:

- `True` - a year produced a Dom Pérignon vintage
- `False` - a year did not produce a Dom Pérignon vintage

Labels were only obtained for the years 1940 through 2013. After 2013, announcements have not been made as it usually takes approximately 10 years for the champagne house to release a vintage. Data from 2014 through 2023 is purely a prediction and will only be validated once the next few vintages get announced.

Pre-processing

Pre-processing was critical for this model. For each of the above mentioned features every year had 365 individual values that corresponded to that year. This made the original dataset have a dimensionality of $3 \times 84 \times 365$. For analysis, the goal was to reduce the dimensionality of this to 3×84. This was achieved using Linear Discriminant Analysis.

To accomplish this process, the scikit-learn library was employed. Utilizing the `LinearDiscriminantAnalysis` class and implementing the `svd` solver, three different models were built and trained. One for each of the features which allowed for the data to be reduced down to a single value per year for each of the three features.

LDA has some drawbacks in the form of its assumptions. LDA assumes that the data has a Gaussian distribution and that the covariance matrices of the different classes are equal. This was determined to be accurate. It also assumes that the data is linearly separable, meaning that a linear decision boundary can accurately classify the different classes. This is not entirely the case with the data; however, there are few enough outliers that we decided to proceed with this method.

Supervised Learning Models

The ultimate aim of this project was to correctly classify the data into two classes. We decided to use three different methods of classification to determine which model is most effective in this scenario. The models that we used were:

- Support Vector Machines (SVM)
- Random Forest
- Logistic Regression

To determine the best model, we looked at both training and testing accuracy. Using scikit-learn's function `train_test_split` the data was broken into two groups with 80% used for training while the remaining 20% was reserved for testing.

Support Vector Machines (SVM)

Random Forest

Decision Tree

Logistic Regression with Gradient Descent