Predicting Grape Suitability Base on Weather Data

By: George Jieh

Non-Technical Overview

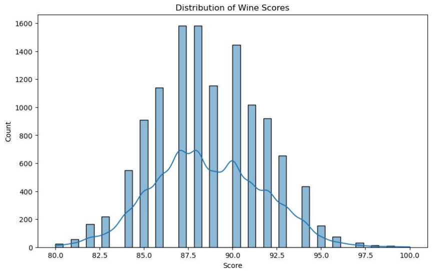
- Problem Statement
  - Climate change is heavily impacting the wine industry by causing locations that are fit to traditionally produce certain wines to be no longer suitable for those grapes. This causes decrease in quality and make wine grape growth cycles increasing hard to get right by vintners. This results in crop waste.
- Solution
  - Develop a model that can predict a location’s suitability for certain wine grape types.
- Impact
  - Reduce crop waste.
  - Discover new areas for viticulture.
  - Increase supply of quality wine, which in turn can make them more accessible.

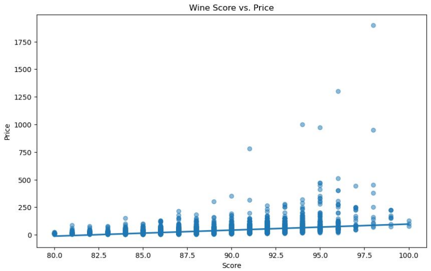
The Dataset

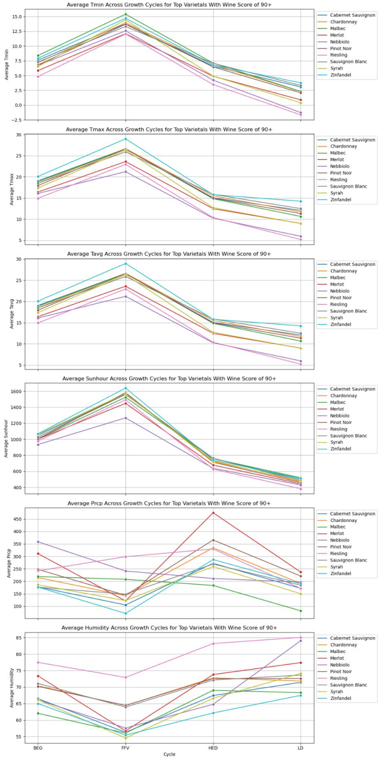
- Dataset is a combination of many dataset.
- The wine data is a combination of a data set from Kaggle that someone else scraped from Wine Enthusiast website and from my own scraping of data from Wine Enthusiast with my own custom scrapper.
- The weather data initially was a dataset I downloaded directly from NOAA that I needed to write custom parsing script to clean, then later on switched to a script that requests NOAA API for better accuracy, and eventually move on to World Weather Online API for better efficiency. These weather APIs all need latitude and longitude information, so that was fetched through Bing Maps API.
- The combined dataset contains name of wine, name of winery, price, wine score, location details, temperature data, precipitation, sun exposure data, and humidity.
- During data cleaning process I needed to reorganize column order, change column names, drop bad data, denote North and South hemisphere, and combine months into grape growth phases to reduce column amounts.
- The grape growth phases are Budburst, Early Growth, Flowering, Fruit Set, Verasion, Harvest, Early Dormancy, and Late Dormancy. Some of the months for each phase overlap, so I combined some of these phases as well.

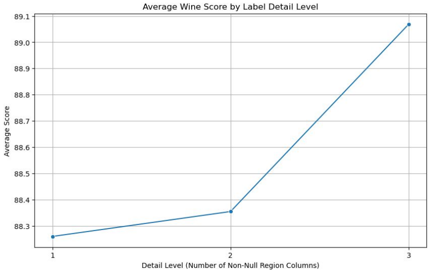
Exploratory Data Analysis

- Detailed labeling seems to be associated with higher quality and price.
- Wine score distribution gathers around 87 to 88 points.
- While there is a trend of higher prices for higher-scoring wines, there is a substantial number of reasonably priced wines with high scores, indicating that quality is accessible at various price points.
- High scoring wines have ample amount of rain during BEG and FFV and tends to have the least amount of rain during LD and HED.
- Tempure for high quality wines during FFV tends to go the not too hot but not to cold route, but tends to have the lowest temperatures in all other cycles. The same could be said for sun exposure as well.
- Humidity seems to be all over the place too, with Riesling being able to withstand high humidity the most. However this is most likely because Riseling have many different styles and one of the most expensive styles is late harvest, which encourages Botrytis, a type of grape fungal growth, that requires high humidity right before sunrise and enough heat to dry off the dew quickly during the rest of the day.









Baseline Models and Evaluation Metrics

- Variables used are al the climate data. The model is judged based on how close it is at predicting the wine score and wine variety.
- For the framework of the problem we are trying to solve, wines that are blends and not a specific grape type were removed or in the case of Bordeaux, have their data incorporated into Cabernet Sauvignon and Merlot.
- Ran preliminary logistic regression, random forest, gradient boosting classifier, and neural network with tensorflow.
- All results are subpar with very low accuracy, ranging between 20 to 30%.

Next Steps

- Improve model accuracy.
  - Fine tuning of the dataset
  - Fine tuning of the model
- Production version of the project would be a local hosted interface where the user can type in the address of a location and the model will recommend what wine grapes to grow there.