Wine Taste Forecaster

Domain Background

Wine traders, consumers, and producers spend millions of dollars to target a specific tasting profile. Experiments on making wine are tremendously difficult due to the number of factors affecting the wine taste including weather, soil, water, temperature and mineral compositions. Production of wine can take years and certain type of taste only appear after its maturity. This process may take up to 20 years. It is therefore important that the stakeholders know in advance what the wine will taste like.

This project aims to create a forecaster on wine taste and rating based on location, time, temperature, type of grapes used, and rainfall.

Taste and quality of wine

The result of this forecaster will be useful to winemakers who can set the price and release the wine based on the quality expected. Market speculator are able to forecast the quality of wine before they mature. Consumers are able to know which wine will have the characteristics that they desire.

Blending and import of grapes

Producers usually import grapes from different regions in order to match the taste profile that they desire. This predictor will give them an insight into the quality of the grapes in each region. This will be especially useful for producers in a non-traditional wine region who rely heavily on imports.

Maturity of wine

Ageing of wine is one of the most important decisions that all the stakeholders need to make. It is absolutely critical that wine is aged ideally. Ageing too long, the wine will turn sour. Ageing too little, the wine would have yet to reach the optimum maturity.

The decisions are traditionally made through tasting. However, this is an unreliable and expensive way to draw the conclusion. The result of this model will help the users to minimize cost and maximize profit.

Problem statement

The goal of the project is to create a forecaster which takes in information as listed in the input array and output a taste index, taste description, wine price, and wine score

Inputs

- Weather data
    - Average Temperature (daily)
        - Over the year before production
        - Over the year of production
    - Average Rainfall (daily)
        - Over the year before production
        - Over the year of production
- Location data
    - Country of production
    - Province of production
    - Vineyard name
    - GPS location of the vineyard
- Raw material data
    - Grape types

Outputs

- Wine Taste index array (to be used to display as a chart/paragraph with a webapp)
- Price
- Quality rating by Winemag

Solution Statement

The tasks involved are following.

1. Parse and clean all the training data
    a. Wine data is moved into a wine_df dataframe then
    b. the latitude/longitude data is extracted using geopy and appended to wine_df.
    c. Year of production is parsed from the description text

d. Temperature data downloaded from kaggle database
    i. Temp data is grouped by latitude/longitude and year of production.
    ii. Resultant df will have 365 columns of max/min temperature per day of each year for each latitude/longitude region.
    iii. Weather data is appended to each wine by closest location, year of production, year prior to production
    iv. Each weather data column will result in 730 additional column for each wine in wine_df (2 years of daily temp)
    v. Name column [tempi] where i is number of day from the date of origin (1st Jan of year before production)

2. Rainfall data is scraped from the worldclimatedata website[11] using a selenium script.
    a. Rainfall data is grouped by latitude/longitude and year of production.
    b. Resultant df will have 365 columns of rainfall per day of each year for each latitude/longitude region.
    c. Rainfall data is appended to each wine by closest location, year of production, year prior to production
    d. Each rainfall data column will result in 730 additional column for each wine in wine_df (2 years of daily rainfall)
    e. Name column [rainii] where i is number of day from the date of origin (1st Jan of year before production)
3. Tokenization and filter out unrelated keys
    a. In wine_df['description'] column, the description of wine has to be tokenized and counted for the key word of interest
        i. Use nltk library to filter out unrelated words, remove prefix ,and suffix
        ii. Tokenize the word with countVectorizer tokenizer
        iii. Use top 20-30 flavouring description that appear most frequently and count them against the description
        iv. Append taste note to wine_df as additional columns. Name the columns[taste index i] where i is the key.
4. One hot encode the columns [country, designation, province, region_1, region_2, variety, winery] replace null with zero. Then append this to wine_df.
5. Plot correlation and potentially remove some X column
6. Split data into train and test set using sklearn.train_test_split and add label to each column.
    a. Y columns = [price, point, taste index i]
    b. X columns = [tempi, raini, country, designation, province, region_1, region_2, variety, winery]
7. Create a model to train on the list
    a. Models in consideration for categorical Y [taste index]
        i. Pytorch
        ii. SKLearn
            1. SVA
            2. SVA linear

3. K- nearest
            4. Random-forest
            5. Linear-regression
        iii. XGBoost
        iv. LightGBM
    b. Model for forecasting a variable include
        i. Pytorch
        ii. SKLearn
            1. Lasso
            2. Elasticnet
            3. Ridgeregression
        iii. XGBoost
        iv. LightGBM
    c. Determine which model would perform best
        i. For pytorch, test around 10 different configuration
    d. Potentially combine the model with ensemble
8. Create a RESTApi with API gateway and aws lambda
9. Create a webapp to interact with the gateway and gather input-output

Datasets and inputs

1. The tasting data set used is from the Kaggle wine dataset version 4[1] by the magazine Winemag[2].
    a. This includes 150,929 datasets. However, after removing Nan, we are left with 125,870 datasets.
    b. The data is formatted in csv with the following column [country, description, designation, point, price, province, region_1, region_2, variety, winery] with the unique categorical values in table 3.1.
    c. The wine score is give between 0 and 100 according to winemag rating
    d. Price is given in USD

	country	designation	province	region_1	region_2	variety	winery
count	150924	105195	150924	125870	60953	150929	150929
unique	47	30621	454	1236	18	632	14809

Table 3.1 Unique categorical value of each index

2. Temperature data is from Kaggle Global Warming database by @berkeleyearth[3].
    a. For This dataset, there are 5 columns [datatime, max daily temp, min daily temp, latitude, longitude]
    b. This can be mapped to the wine dataset using the nearest latitude/longtitude data
3. GPS coordinates is extracted from Geopy library 2018[10]

    a. The library is installed by pip3 using ```pip3 install geopy```
4. Rainfall matrix is scraped from the Climateknowledgeportal website[11].

Evaluation Metrics

- Taste note
    - F1 score
        - There is no indication that the data is balanced which means accuracy can be misleading.
        - Binary classification data
        - F1 is a harmonic mean of precision and recall which are both good indicators of the performance of the model
- Score/Price
    - Relative Absolute Error (RAE) and Mean Absolute Error (MAE)
        - This is suitable since there are likely to be a lot of outliers in both wine score and wine price. We would like those to not affect the actual model too much.
        - Wine price is affected strongly by other factors such as fashion, marketing, and general market conditions. It is expected that there will be a high number of outliers

Benchmark Model

I have found various projects that has similarities to this proposal.

1. Robinson, S., (2019)[4] research is using the same Kaggle data set to determine the price of wine. However, It is using the bag of word description to forecast the price and she is not sharing the end result so the model needs to be reproduced to calculate the evaluation metrics.
2. cortez, P., (2019)[8] aims to forecast wine price using weather data on a linear regression model and comes up with a matrix of price probability table for each vineyard.
3. Freecodecamp(2018)[6] article tries to understand what makes wine taste good based on its chemical characteristics using wine dataset from UCI (Uciedu, 2019).
4. Olivier goutay, (2018)[9] forecasts the 5 groups of wine quality based on its description. Which the author claim has over 97% efficiency using random forest classifier.

After reading all the articles, I come up with the following benchmarks

1. Wine description forecast should be better than a linear regression classifier using sklearn. (Model will be built as a benchmark)
2. Wine group of quality should be at least 97% accurate according to research 4.

3. Wine prices have MAE of 4 according to the 3 results presented in research 1.

Reference

1. Kaggle. 2019. Wine Reviews 130k wine reviews with variety, location, winery, price, and description. [ONLINE] Available at: https://www.kaggle.com/zynicide/wine-reviews. [Accessed 31 July 2019].
2. Winemag. 2019. Wine Enthusiast. [ONLINE] Available at: http://www.winemag.com/?s=&drink_type=wine. [Accessed 31 July 2019].
3. Kaggle. 2019. Climate Change: Earth Surface Temperature Data. [ONLINE] Available at: https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data. [Accessed 31 July 2019].
4. Robinson, S., 2019. Predicting the price of wine with the Keras Functional API and TensorFlow. Predicting the price of wine with the Keras Functional API and TensorFlow, [Online]. 1, 1. Available at: https://medium.com/tensorflow/predicting-the-price-of-wine-with-the-keras-functional-api-and-tensorflow-a95d1c2c1b03 [Accessed 31 July 2019].
5. Vonvino. 2019. Artificial Vintelligence: AI Gets Taste of Wine Industry. [ONLINE] Available at: https://vonvino.com/artificial-intelligence/. [Accessed 31 July 2019].
6. Freecodecamp. 2018. How to Use Data Science to Understand What Makes Wine Taste Good. [Online]. [31 July 2019]. Available from: https://www.freecodecamp.org/news/using-data-science-to-understand-what-makes-wine-taste-good-669b496c67ee/
7. Uciedu. 2019. Uciedu. [Online]. [31 July 2019]. Available from: https://archive.ics.uci.edu/ml/datasets/wine quality
8. cortez, P., 2019. Decision Support Systems. Modeling wine preferences by data mining from physicochemical properties, 47/4, 547-553.
9. Olivier goutay. 2018. Medium. [Online]. [31 July 2019]. Available from: https://towardsdatascience.com/wine-ratings-prediction-using-machine-learning-ce259832b321
10. https://geopy.readthedocs.io/en/stable/
11. https://climateknowledgeportal.worldbank.org/download-data