Machine Learning Engineer Nanodegree

Capstone Project Proposal

Predicting Wine Quality

Chikka Udaya Sai

June 8th , 2018

Domain Background

Wine has become an important part of human life and often considered as a luxury product. The role of wine is significant in human's life because it is an important source of nutrition and the art of wine making has its own importance. Wine sector contributes a lot of economy to most of the countries like UK , Portugal etc., There are a lot of evidences (Evidence is given in the references) which suggests that moderate consumption of wine helps people to live longer , protecting them against certain cancers and improve mental health.

From the past , wine was used to treat various health conditions . Monks lived longer than other people because of their regular and moderate consumption of wine. This fact is also proven scientifically . For all this to happen , the quality of wine has to be great and wine makers give utmost importance to quality of wine. Producing a good quality of wine is not an easy task . There are so many factors influence the quality of wine but most of the factors related to the chemical composition involved in the manufacturing of wine. There is a research going on to produce the best quality of wine. Some of them I found interesting are:

https://www.researchgate.net/publication/259922445_Proficiency_test_on_FTIR_wine_analysis

https://www.emeraldinsight.com/doi/abs/10.1108/17511061111186514

http://www.guadoalmelo.it/en/winemaking-high-quality-artisanal-production/

https://health.gov/dietaryguidelines/2015/guidelines/appendix-9/

The reason behind my interest in this project is to know how Machine Learning algorithms perform on real – world applications. Our model is well suited for the field in testing the quality of wine.

Problem Statement

The aim of this project is to predict the quality of wine based on the chemical composition involved in the manufacturing of wine . If the quality of wine is good , then my model will give output as **1** else **0.**

The problem is a binary classification problem which have two possible outputs ('1' for Good quality of wine and '0' for bad quality of wine).

I am going to use 11 features (chemical composition involved in the manufacturing of wine) in the data as predictor variables and Quality (Quality rating) as the target variable.

For this purpose , I am modifying the data of Quality attribute to 1 and 0 depending on the quality rating of wine greater than 5 or not.

In the project , I am going to use various Machine Learning Algorithms to predict the quality of wine and compare their performance and finally declare my final model.

Datasets and Inputs

The dataset that I am working is downloaded from **UCI Machine Learning Repository**. It contains chemical composition of both red wine and white wine but I am working only with white wine data .

Dataset URL :

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Citation:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Available at: https://www.sciencedirect.com/science/article/pii/S0167923609001377?via%3Dihub

The data is open – sourced and can be used for research purpose with proper citation included.

There are 12 features and 4898 instances of data . The feature that I am interesting is Quality (Target Feature) . All data in the dataset are of numerical type and continuos except Quality which is discrete (It has values in between 1 to 10) . There are no missing values in the data. I am going to use all this data for my project.

The features are :

1. Fixed Acidity : Measure of total concentration of tartaric acid that are present in the wine which are produced in the fermentation process.
2. Volatile Acidity : It is the acidity produced in the wine which is caused by the bacteria which gives a characteristic flavour & aroma to wine.
3. Citric Acid : Amount of Citric acid added to the wine to increase the acidity of wine and to give fresh flavour to it.
4. Residual Sugar : Amount of sugar that are left after the completion of fermentation step. The sweetness of wine depends on this amount.
5. Chlorides : Amount of chloride present in the wine. This is added to increase the saltiness.
6. Free Sulfur dioxide : Amount of free sulfur dioxide present in the wine . It is added to wine to preserve it from bacteria.
7. Total Sulfur dioxide : Amount of total sulfur dioxide (free sulfur dioxide + other sulphate content) present in the wine.
8. Density : It is the density of wine.
9. pH : It is a measure of acidity present in the wine . If acidity is more , pH is low.
10. Sulphates : The amount of sulphate content present in the wine. This is added to the wine to stabilize it.
11. Alcohol : It is the volume of alcohol present in the wine.
12. Quality : It is the measure of quality of wine from 1 to 10.

Solution Statement :

The solution that I am going to give to this problem is to predict the quality of wine based on the composition involved in the manufacturing of wine. For this , I am going to change the feature Quality to **1** if Quality rating is greater than 5 else **0** if Quality rating is less than or equal to 5 .

Benchmark Model :

The Benchmark model that I am taking is to predict the class of target variable as **Good Quality 1** irrespective of the chemical composition of the wine . But this will give very bad results and we'll use SL algorithms to achieve reasonable accuracy. The reason why I choose this as a Benchmark model because most of the wines in the data are good quality and no wine manufacture wants to prepare bad quality wine which is a loss to his company because nobody wants to buy the bad quality wine. Sometimes , due to several reasons Bad quality wine is formed.

Evaluation Metrics:

The evaluation metric that I am using to measure the performance of our model F – score . Accuracy and F – score are good metrics to measure the performance of classification model . F- score takes account of both precision and recall and accuracy tells how many of our predictions are right. I didn't choose accuracy score as my metric because my data is unbalanced.

Therefore , I am choosing F – score as my metric because I want to select the model based on the balance between Precision and Recall. In addition to that , I will also check training and testing time .

Project Design :

The project is composed of different steps as follows:

Data Exploration:

First task is to read the dataset and perform visualizations on it to get some insights about the data.

Data Pre-processing:

After Data Exploration , I want to apply Outlier and Missing value treatment and standardize the data to make it suitable for Machine Learning Algorithms.

Model Selection:

First I want to choose a Benchmark model which will always predict the quality of wine as good quality . The accuracy will be low and then I want to test 3 classification algorithms (Logistic Regression , Decision Tree and Adaptive Boosting Classifier) on the data and compare the performance of them against their training time , testing time and F – score . I want to select the best model among them.

Model Tuning:

After selecting the model , I want to use Grid Search technique to tune Hyper parameters and to increase the best model efficiency to little bit further.

Summary:

I want to summarize the results by comparing the performance of models that I used in the project and suggesting improvements how model can be improved.