Machine Learning Engineer Nanodegree

Using Supervised Learning to predict the quality of Redwine

Durgaprasad Rajani

June 27th 2018

Proposal

Domain Background

In today life we need everything that meets standards, otherwise we won't consider to take it. So, In this I am going to find the quality of red wine.Good quality of wine makes health good.In some cases the quality does not meet standards.The companies should maintain the belief of the customers.The belief can be achieved from the quality of the product.

Every company invest lot of money for preparing wine.So, it is crucial for them to make a quality of wine.Redwine also preferable by some doctors.There is a belief that it lowers the chances of occurring heartattack or heartstrokes.And I explore for the dataset that what features make a good quality of wine.I got the dataset from this Link:https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

Problem Statement

The Redwine quality can be decided by many of the factors that are used for preparing the redwine.Those ingredients proportions can change the wine quality.By considering all those ingredients we can define which of those can significantly effect the wine quality.The main purpose of this is find the most related features with the quality and use the to find the wine quality by saying it is good or bad.And for that we can consider it as a classification problem.

Features and Description:

- fixed acidity – It is non-volatile acid in the wine

- volatile acidity – It is amount of acetic acid in the wine
- citric acid – It is the amount of citric acid in the wine
- residual sugar – It is the amount of sugar remained after the fermentation process
- chlorides – It is the salt content in wine
- free sulfur dioxide – It is amount of free sulfur dioxide present in wine as gas form
- total sulfur dioxide – It is the amount of free and bound forms of sulfur dioxide
- density – It is the density of water that is close to that of water depending on the percent of alcohol and sugar content
- pH – It is the value for representing how much it is acidic or basic nature
- sulphates – It is the amount of sulphates in wine
- alcohol – It is the percent of alcohol content in wine
- quality – It the quality of the wine

By observing the above features we can say that alcohol content can effect the quality of wine the higher the alcohol the better the wine.The feature residual sugar can effect majorly because too much sugar content can make wine sweet which is not a good wine.The citric acid feature is also show impact on determining the wine quality.If it increases the taste of wine becomes somewhat sour.So,that wine can be treated as bad wine.

Datasets and Inputs

I have taken this dataset from kaggle site and references for these are like what features can be involved in the wine making.The following features are

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol
- quality

The datset consists of 1599 rows and 12 columns.In this no missing values are present in any column.All these are positive values and no negative values are involved in this dataset.Only physiochemical inputs are involved in this and no sensory inputs(like wine brand, wine price etc.) are involved.

Solution Statement

This dataset can be viewed as classification or regression task.By seeing these classes they are not orderd and are imbalanced. Most of the wine quality are normal wines, not poor or excellent.More than one feature can decide the wine quality.We can use our model to train those data and predict the wine quality.So, I will use classification supervised model for this.I do use different classification models and caluculate those accuracy or F1score and take the best out of it.If I won't get high accuracy levels then I will use some optimization methods like Grid Search to tune the parameters for getting the better results.The different classification algorithms are RandomForestClassifiers,SVM and SGDC will use for this.And I will compare those algorithms find the best model.

Benchmark Model

I will use logistic regression model as my benchmark model.Accuracy and F1 score will be taken as reference to compare the results with other models.If it alone gives the best metric scores than other models then I will consider this as my model.

Evalution Metrics

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

Accuracy =(TP+TN)/(TP+TN+FP+FN)

In the Numerator, are our correct predictions (True positives and True Negatives)(Marked as red in the fig above) and in the denominator, are the kind of all predictions made by the algorithm(Right as well as wrong ones).

Precision is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer. The predicted positives (People predicted as cancerous are TP and FP) and the people actually having a cancer are TP.

Precision = TP/(TP+FP)

Recall is a measure that tells us what proportion of patients that actually had cancer was diagnosed by the algorithm as having cancer. The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP. (Note: FN is included because the Person actually had a cancer even though the model predicted otherwise).

$$\text{Recall} = TP/(TP+FN)$$

Another score is F1 score as an evalution metric .Between the original test values and predicted values.This score will be calculated for every model.After tuning the parameters crossvalidationscore will be calculated.

Project Design

At first we have to read the csv file into the dataframe with the help of pandas library and after that I will explore the data.I will check whether the columns have any null values or missing values.If they are I will preprocess the data.After preprocessing and I will describe the data.And made some visualizations on the features to make conclusions from it.As the quality values are mostly normal wines so,I will make this as binary classification as good or bad.For this will transform the quality values like equal or more than 7 quality will be treated as good otherwise bad. Then I will split my data into 80% training and 20% testing with some random state.After I will perform the logistic regression model and find the accuracy score model.Then different models are trained and tested using this data.For increasing the accuracy I will tune the parameters by using the Gridsearchcv method.The different algorithms include ensemble methods like RandomForestClassifier,Stochatic Gradient Descent Classifier and SVM.The I will compare those models by makin curves based on accuracy.

References

- https://archive.ics.uci.edu/ml/datasets/Wine+Quality
- https://www.sciencedirect.com/science/article/abs/pii/S0950329307000493
- https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance