Machine Learning Capstone Project

Title: Using Supervised learning to find whether the quality of wine is good or bad.

June 28th,2018

Sonia Agarwal

Proposal

Domain Background

Wine is an alcoholic beverage produced through the partial or total fermentation of grapes. Other fruits and plants like berries, apples, cherries, dandelions, and palm can also be fermented. It is one of the most consumed drink by the people. The Wine we are going to study here is White wine .It can be made with either white or red grapes. According to the technical point of view, a good quality wine consists of factors meeting specific criteria and set according to considerations of scientific, chemical and technical factors, which can always be verified analytically. The quality of wine is mainly affected by the ingredients used for making it. So, we use these ingredients to predict the quality of the wine. The major motivation for exploring this project is to develop the skill and understanding how to work out on real time datasets and how Machine-Learning is potentially making the lives of people easier.

Reference1: https://en.wikipedia.org/wiki/Wine_chemistry

Reference2: https://winefolly.com/tutorial/how-is-white-wine-made/

Reference3: https://en.wikipedia.org/wiki/Wine

Problem Statement

By using the dataset, the task is to determine of the quality of wine. Since the task is to figure out whether the quality of the wine is good or bad based on the chemical composition in manufacturing the wine. And hence we have to tackle with a binary classification problem that has two possible outcomes (‘0’ for bad quality and ‘1’ for good quality). By using machine learning techniques we can determine quality of wine. Several steps are involved in the project like Data

exploration, Data processing and finally testing various algorithms and techniques.

Datasets and Inputs

Link for dataset is given below.
Link:https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv

The dataset has 4898 instances and 11 attributes and a target variable. The 11 attributes are as follows:

- Fixed acidity -It is the level of fixed acidity
- Volatile acidity -it is the level of volatile acidity
- Citric acid -amount of citric acid
- Residual sugar -amount of residual sugar
- Chlorides -amount of chlorides
- Free sulfur dioxide -amount of free sulphur dioxide
- Total sulfur dioxide -amount of total sulphur dioxide
- Density -density of the wine
- pH –pH of the wine
- Alcohol-alcohol content in wine
- sulphates- amount of sulphates in the wine

output variable is(based on sensory data)

- Quality (score between 0 and 10)

By considering all the above features, we can predict the quality of the wine to be good or bad

Solution statement

The theme of the project is to determine the wine quality .Hence to achieve the potential quality I will apply classification model of supervised learning by passing the data to the model and predict whether the quality of wine is good or bad . I will use the above listed features to check for null values and outliers etc. And in order to predict the target variable ,I will convert all the values of it

to 1 and 0 to make it suitable for classification task. Then I will use a bench mark model on the data and then check the performance using the fbeta_score. Then I will apply several diverse classification models and evaluate the performances of the specified models and pick the model that generates the best fbeta_score. The best model that is determined to be generating the better performance is potentially optimized using the GRIDSEARCH CV technique and then check performance using the metric fbeta_score.

Benchmark model

I will use Logistic regression is used as benchmark model. Fbeta_score of benchmark model is used as reference and other model will be judge to perform better if their fbeta score will be greater than Logistic regression model.

Evaluation Metrics

In this project I will use the evaluation metric of fbeta_score. It measures the effectiveness of retrieval with respect to a user who attaches beta times as much importance to recall as precision.

$F\beta=(1+\beta 2).precision.recall/(\beta 2.precision)+recall$

Project Design

Data Exploration:

- Initially I will read csvfile into pandas data frame
- we can find number of rows and columns in it. For better understanding, we can plot scatter matrix of all features .

- Then I will plot the histograms of all the features to find skewness.We can plot heatmap as it is easy to identify the correlation between features

Data Preprocessing:

- We can check for missing values in my data set. If any missing values are found, I will replace with Nan.
- Then I will check for outliers and remove them to get cleaned data.
- We can separate target variable and store separately.

Data Standardization:

- We can apply scaling on my data for data standardization. I will use minmax scaling. And all the resulting values will be in the range of 0 and 1. Now, I will split my data into training and testing data.

Data modeling:

- We split the data into training data and testing data.
- We will check my benchmark model which is Logistic regression.
- based on that we will find the F beta score and accuracy scores.
- Then we will test with different algorithms like SVC, Decision Trees, KNN classifier, Adaboost Classifier and obtain their respective fbeta scores.
- And then find the optimal model among them which gives the highest accuracy .
- The model is refined by the application of GRIDSEARCHCV which helps in tuning the parameters optimal for the model

References

http://sebastianraschka.com/Articles/2014_about_feature_scaling.html

http://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html