Wine Quality Project

Anshul Bhardwaj

Problem Statement

- To build a Machine Learning Model which will classify a wine in High quality or Low Quality using various input features.

Clients / Intended Audience

- This model can be used by anyone who wants to what makes a good wine, which ingredients affect the wine quality the most.
- This model can be used by anyone who wants to find out quality of wine given the input variables.

Dataset

- The dataset used for this project is taken from UCI Machine Learning repository.
- https://archive.ics.uci.edu/ml/datasets/wine+quality
- The dataset contains many features such as Fixed acidity, volatile acidity, pH, alcohol, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total Sulfur dioxide, Density and Sulphates

Data Cleaning & Data Wrangling

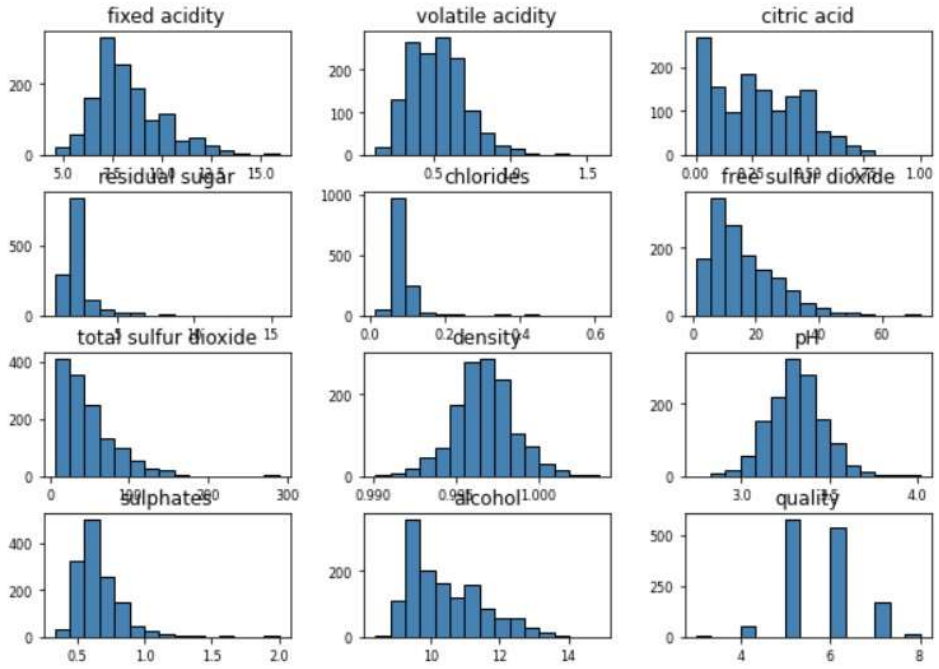
- Duplicate rows removed
  Original Dataset had 240 duplicate rows. They were removed using the following method:

```
df.drop_duplicates(keep = 'first', inplace = True)
```

- Dataset had no missing or NaN values

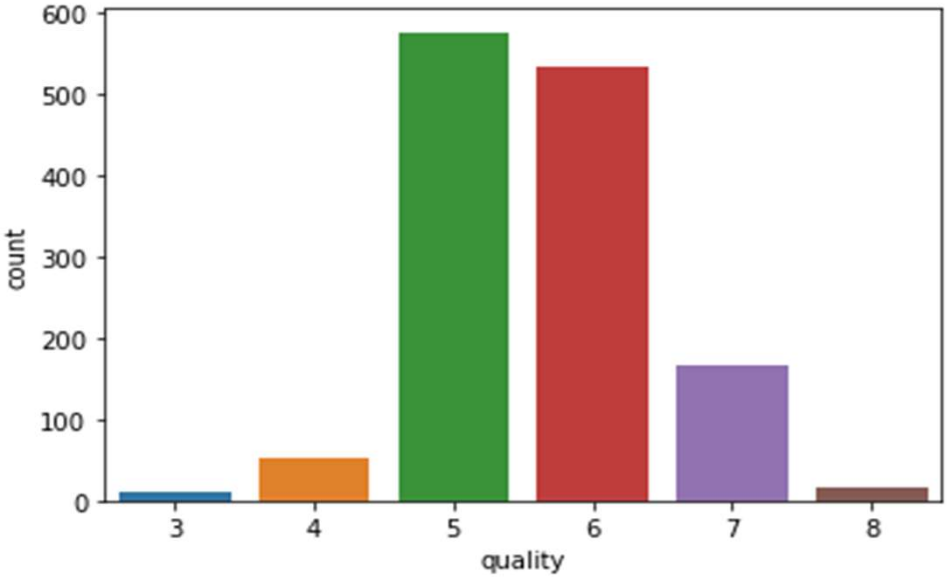
Exploratory Data Analysis

- Distribution of Features



- This image above depicts the distribution of various feature of Wine dataset.

Feature Engineering

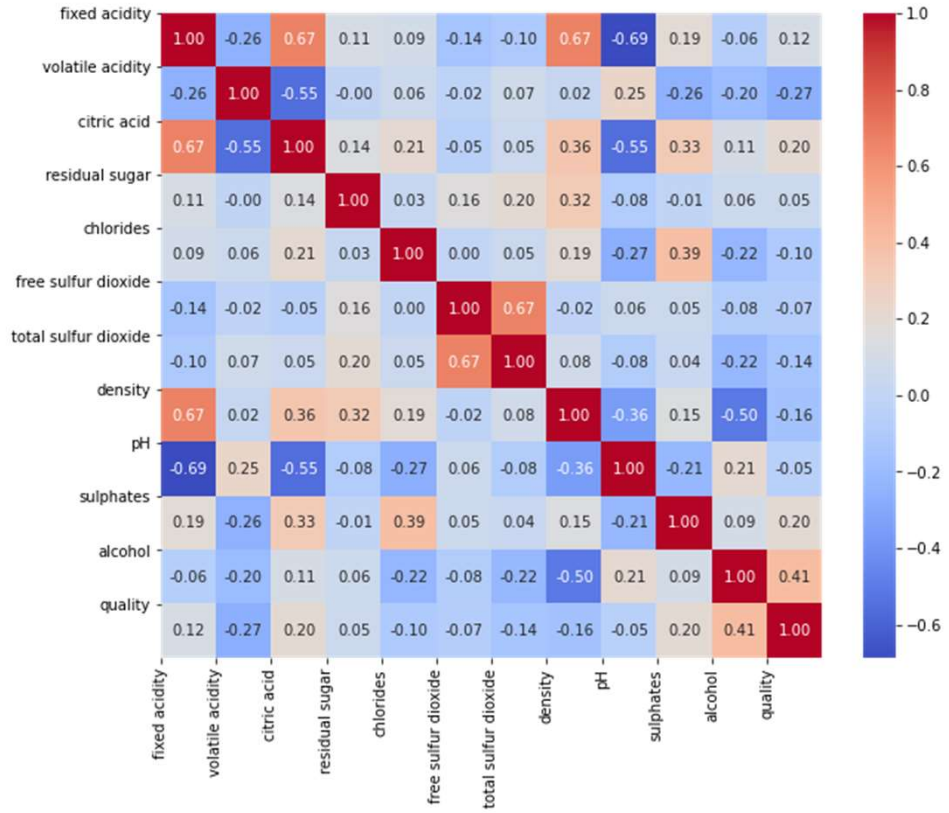


- This plot depicts destitution of wine quality.

- In the given Dataset the Target variable 'Quality' was in the range of from 3 to 8. As per project's requirement the Quality features was split into two labels 'Good' and 'Bad' using the following method.

```
bins = (2, 6.5, 8)
labels = ['bad', 'good']
df['quality'] = pd.cut(x = df['quality'], bins = bins, labels = labels)
```

Correlation Matrix of Variables



- The correlation matrix above depicts the correlation between various features.
- Independent Variable Alcohol has highest correlation with dependent variable quality.

- Out of all the input features the distribution of alcohol feature was different for Good and Bad quality. Other input feature showed not much difference in distribution for both Good and Bad Quality.
- I decided to further analyze alcohol feature and did some feature engineering on it.
- I decided to cut alcohol feature into three segments. Low, Medium and High respectively using the following method.

```
bins = [0,10,12,16]

labels2 = ['low', 'medium','high']

df['alcohol'] = pd.cut(x = df['alcohol'], bins = bins, labels = labels2)
```

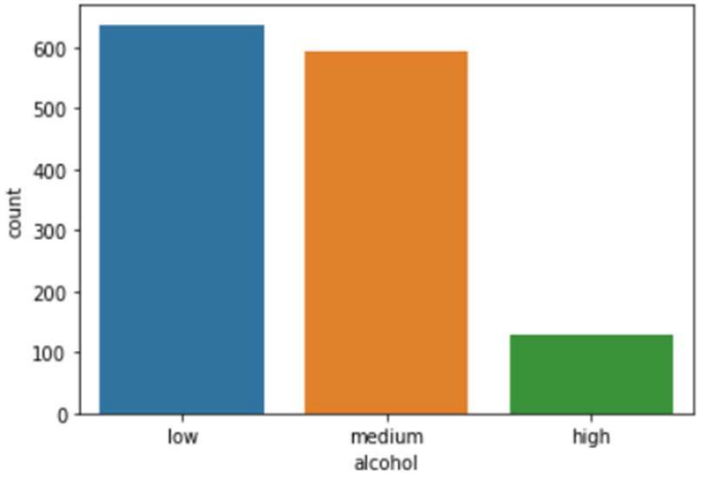


Image : Distribution of alcohol after feature engineering

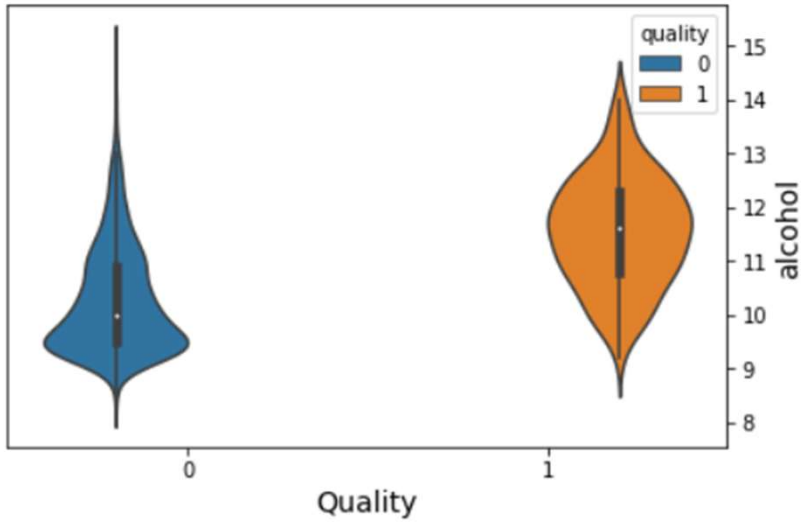


Image : alcohol amount distribution for Low and High quality wine.

Outlier Removal

&

Data Balancing

- I used Statistical method to remove outliers. Any row containing a value having z-score higher than 3 was removed.

```
from scipy import stats

z = np.abs(stats.zscore(df))

df = df[(z < 3).all(axis=1)]
```

- As the dataset was highly Unbalanced, it used SMOTE for balancing it.

```
oversample = SMOTE(sampling_strategy=0.5, random_state=42)

X_train, y_train = oversample.fit_resample(X_train0, y_train0)
```

Machine Learning Model Comparisons

- Three different Machine Learning Algorithms were used.

Model	Accuracy	ROC_AUC Score
Logistic Regression	89.49%	0.868
Random Forest Classification	90.8%	0.850
Xgboost Classifier	92.24%	0.887

- All models performed similar but Xgboost performed the best with ROC_AUC Score of .887.
- Alcohol feature has highest importance.

Future Improvements

- A better outlier removal method
- More feature Engineering