SUMMARY

-Shreyas B

Model used:

Random Forests Classifier

Random forests Classifier is a supervised learning algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.

Random forest uses Gini importance to calculate the importance of each feature. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables. This score will help us to choose the most important features and drop the least important ones for model building.

As we are presented with a very large dataset, we have to understand the weightage of different features. Random Forest models implement a level of differentiation because each tree will split based on different features. This level of differentiation provides a greater ensemble to aggregate over, ergo producing a more accurate predictor.

Features extracted:

After analysing the training data, these features were extracted:

'country': No change -> Label Encoder

'revlist': Derived from review_title using a user-defined function reducetext( )

'review_polarity': Derived from review_description using sentimental analysis

'points': No change

'price': No change

'province': No change -> Label Encoder

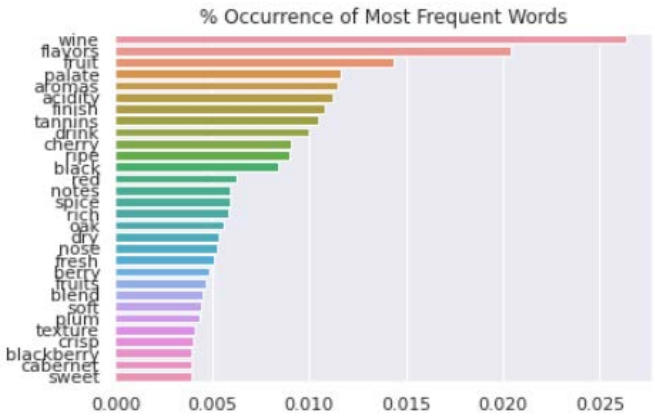
Model accuracy in train:

0.8388548057259714 (with n_estimators:1000)

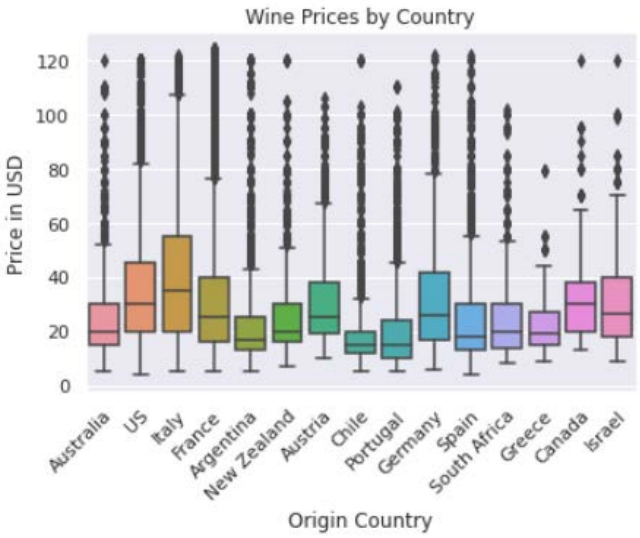
Important Visualizations:

Training data:

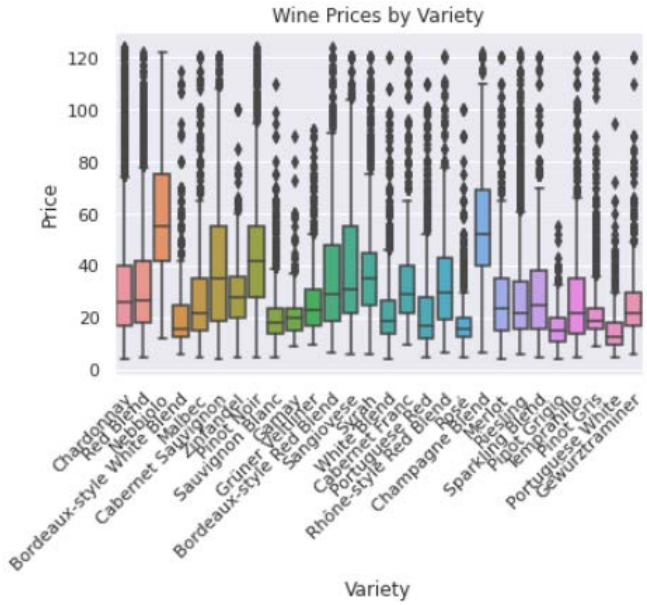
Frequency of word occurrence



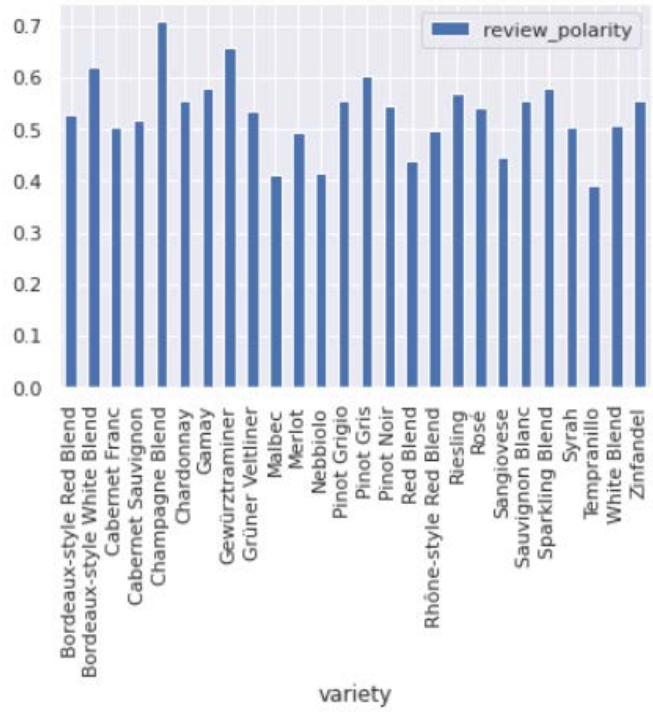
Wine Prices w.r.t country



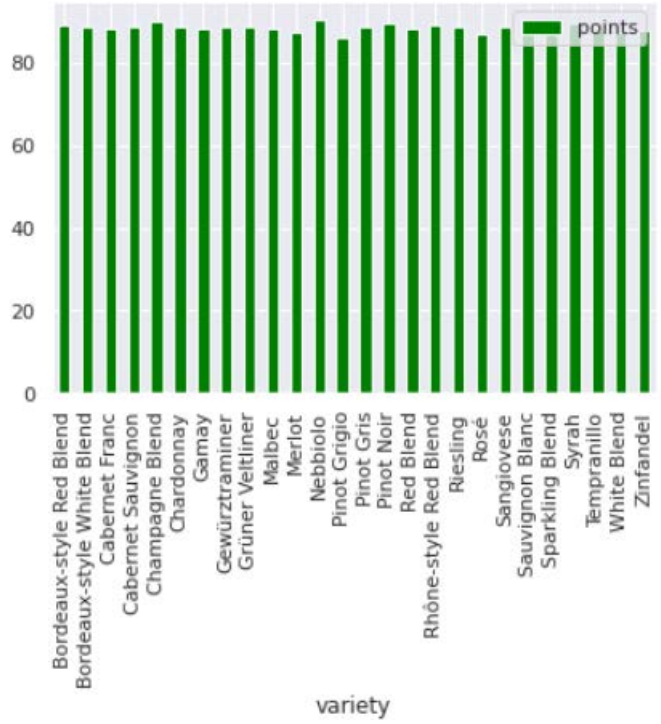
Wine Prices w.r.t Variety



Review_description Sentiment Analysis w.r.t Variety

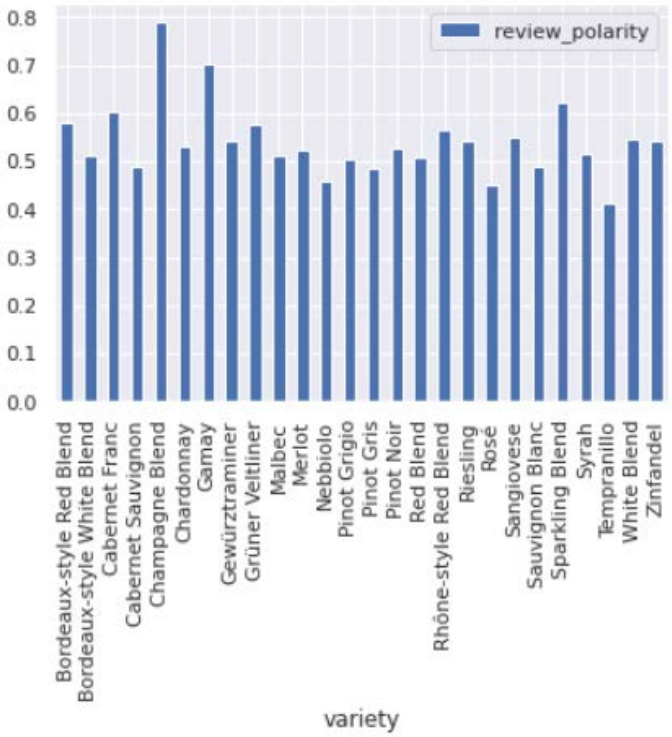


Points vs Variety

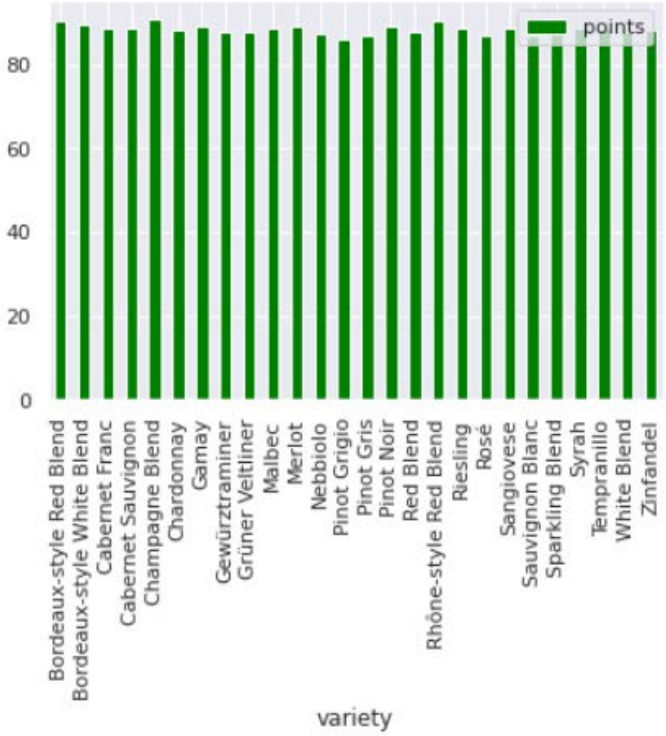


Testing Data:

Review_description Sentiment Analysis w.r.t Variety



Points vs Variety



Wine Prices w.r.t Variety