ML-Assignment Summary

Submitted By: Srushti Sangawar

Date: 08-05-2020

Step 1: Defining the Problem

- Main objective: Predicting the variety of wine as accurately as possible
- Target Feature: Variety
- Type of Problem: Classification problem with 28 types of variety.
- Trying to predict using the review_descriptions of the particular variety

Step 2: Choosing a measure of Success

- Classification problems use evaluation metrics as precision, accuracy and recall. Therefore Classification Report will be used

Step 3: Preparing the data

- Dealing with missing values: There were 35 missing values for country. Since 35<<82657 I chose to delete those particular rows. I also decided to delete the particular rows for missing data for price (5569) which would not affect much. There were many missing values for region_1, region_2, designation and user name which were filled with -1 which I assumed won't be required for the model as much.
- Handling categorical data: The variety column was encoded
- Selecting Meaningful Features: Review_Description was selected to be a meaningful feature.
- Finding Duplicate Values: There were 4730 duplicate titles and description.
  The description was cleaned before it was fed to the model using NLTK library, regex library and CountVectorizer to convert it into a matrix since model understands only numerical values. Later using TFID transformer TF-IDF features were generated.

Step 4: Fitting and predicting

- Two different Classifiers were used to check which gives the better accuracy: RandomForestClassifier and KNN.

Step 5: Accuracy of the Model

- With RandomForestClassifier:

21	0.71	0.22	0.34	511
22	0.76	0.60	0.67	1078
23	0.86	0.42	0.57	478
24	0.78	0.43	0.55	932
25	0.58	0.21	0.31	392
26	0.85	0.37	0.52	492
27	0.89	0.44	0.59	635
accuracy			0.63	21699
macro avg	0.74	0.45	0.52	21699
weighted avg	0.67	0.63	0.60	21699

- With KNN:

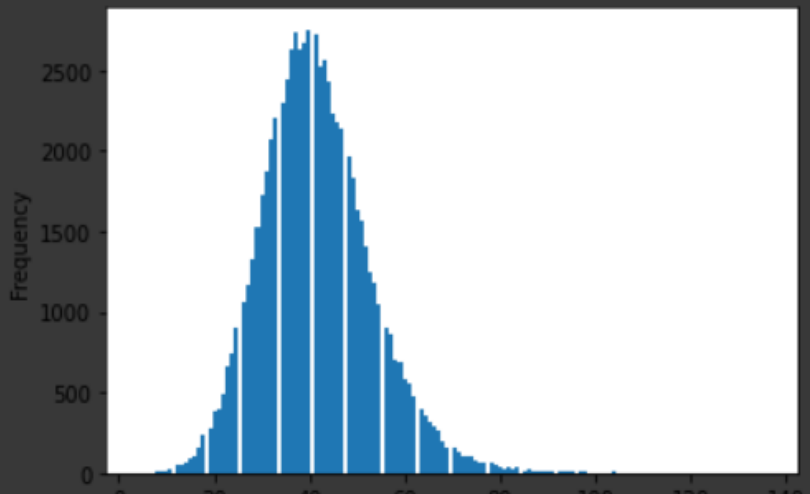
19	0.74	0.68	0.71	1106
20	0.58	0.52	0.55	741
21	0.52	0.29	0.37	526
22	0.66	0.52	0.58	1113
23	0.72	0.49	0.58	479
24	0.52	0.26	0.34	953
25	0.43	0.23	0.30	430
26	0.75	0.42	0.54	472
27	0.59	0.33	0.43	632
accuracy			0.51	21699
macro avg	0.51	0.45	0.47	21699
weighted avg	0.54	0.51	0.51	21699

Conclusion: RandomForestClassifier works much better the KNN

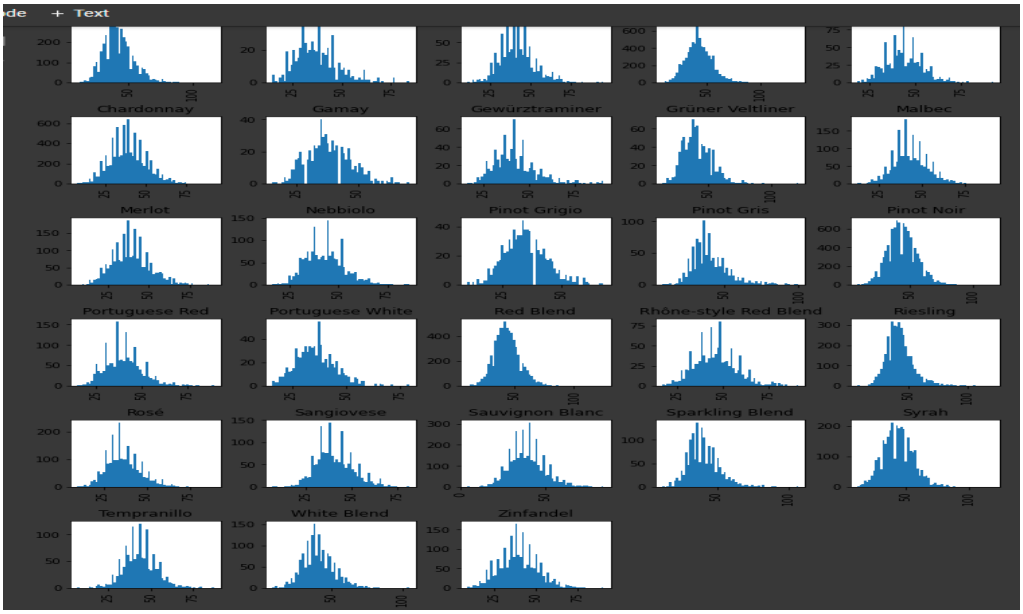
Some Data Visualisations:

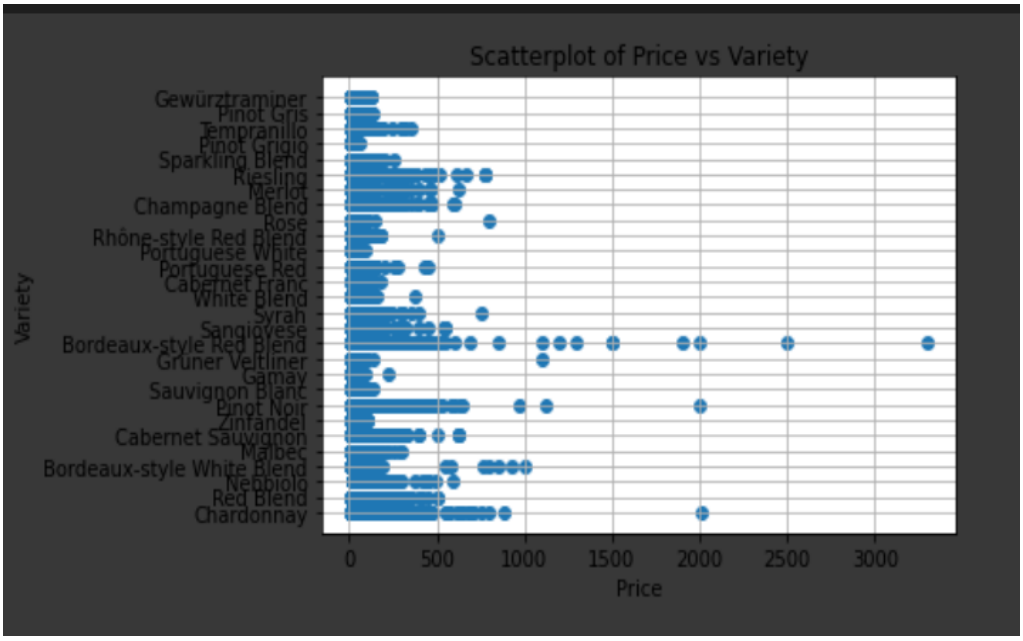
```
[ ]  #visualizing length of description
     data['description_lengths'].plot.hist(bins=150)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fed3a727828>
```



Description_lengths V/S Variety





Top 5 Actionable Insights:

1. Pinot Noir has the highest number of reviews(10587) with an average rating of 89.43 which suggest it's the most popular while Gamay has the least number of reviews(816) with an average rating of 88.05 which suggests it's the least popular. There price differ by a value of 26.47 only. It is advised to increase the stock for Pinot Noir.
   (It was possible that less number of reviews meant rarer and expensive wine but it's neither of it hence it's least popular)
2. US sell the maximum number of wines followed by France and least is sold by India. Since US has a higher stock of wines, it might be possible for the company to buy more wine at a cheaper rate than present.
3. Points show the least correlation. All the points are almost equally distributed among the varieties hence it's not a good idea for the company to rely on the points.
4. Cheaper wines are not bought frequently as usually expected but rather average priced wines are more in demand.
5. Most of the wines with prices more than 100 are scaled more than 90(out of 100) in points .They are lesser in demand due to their price but they are top rated. So it's advisable to keep them in stock.