An Exploration of wine reviews and the Big Data of Wines

General Assembly Data Science Immersive - final Capstone Project

Project overview

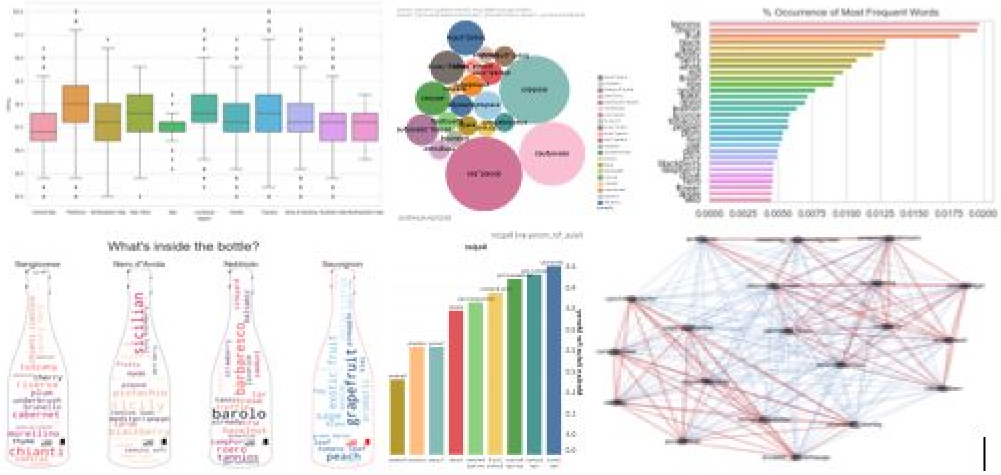
This repository contains the documentation and code for my Capstone project **An Exploration of wine reviews and the Big Data of Wines,** undertaken as part of General Assembly's Data Science Immersive course between November 2020 and February 2021.

The file Capstone_Github.pdf hosts a short presentation of the project for a non-technical audience. It covers goals, data, approach, basic description of model, findings, risks/limitations, impact and next steps.

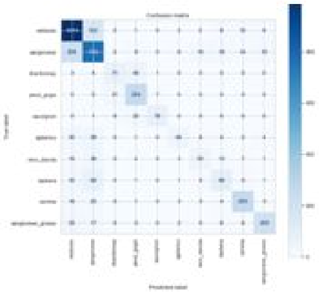
The notebook Scraping_capstone contains the Python code I used for scraping from Wine Enthusiast.

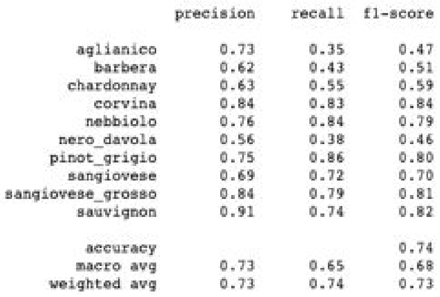
The notebook Capstone_cleaned contains the Python code of the project, and it's divided in 2 parts.

Part 1: Data cleaning, EDA and Preliminary Analysis – A quantitative description and visualization of the data.

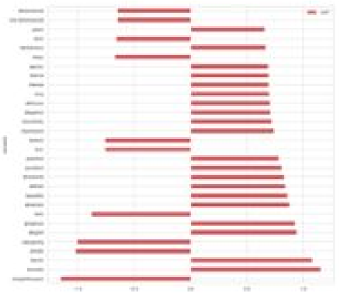


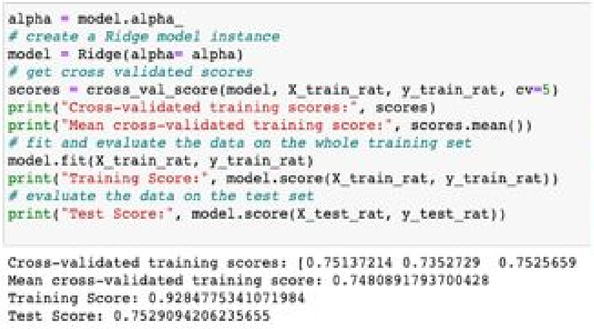
Part 2: Modelling & Results – Details of the models and approach with concisely commented code. Evaluation of the model performance and a discussion of the results. This includes regression models and classification models.





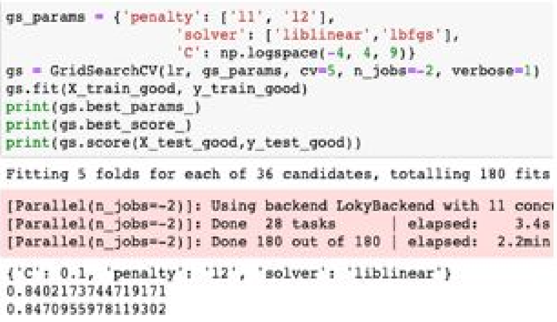
```
                   precision    recall  f1-score

         aglianico      0.73      0.35      0.47
           barbera      0.62      0.43      0.51
        chardonnay      0.63      0.55      0.59
           corvina      0.84      0.83      0.84
          nebbiolo      0.76      0.84      0.79
        nero_davola      0.56      0.38      0.46
       pinot_grigio      0.75      0.86      0.80
        sangiovese      0.69      0.72      0.70
 sangiovese_grosso      0.84      0.79      0.81
         sauvignon      0.91      0.74      0.82

          accuracy                          0.74
         macro avg      0.73      0.65      0.68
      weighted avg      0.73      0.74      0.73
```





```
alpha = model.alpha_
# create a Ridge model instance
model = Ridge(alpha= alpha)
# get cross validated scores
scores = cross_val_score(model, X_train_rat, y_train_rat, cv=5)
print("Cross-validated training scores:", scores)
print("Mean cross-validated training score:", scores.mean())
# fit and evaluate the data on the whole training set
model.fit(X_train_rat, y_train_rat)
print("Training Score:", model.score(X_train_rat, y_train_rat))
# evaluate the data on the test set
print("Test Score:", model.score(X_test_rat, y_test_rat))
```

```
Cross-validated training scores: [0.75137214 0.7352729  0.7525659
Mean cross-validated training score: 0.7480891793700428
Training Score: 0.9284775341071984
Test Score: 0.7529094206235655
```



```
gs_params = {'penalty': ['l1', 'l2'],
             'solver': ['liblinear','lbfgs'],
             'C': np.logspace(-4, 4, 9)}
gs = GridSearchCV(lr, gs_params, cv=5, n_jobs=-2, verbose=1)
gs.fit(X_train_good, y_train_good)
print(gs.best_params_)
print(gs.best_score_)
print(gs.score(X_test_good,y_test_good))
```

```
Fitting 5 folds for each of 36 candidates, totalling 180 fits

[Parallel(n_jobs=-2)]: Using backend LokyBackend with 11 conc
[Parallel(n_jobs=-2)]: Done  28 tasks      | elapsed:    3.4s
[Parallel(n_jobs=-2)]: Done 180 out of 180 | elapsed:  2.2min

{'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
0.8402173744719171
0.8470955978119302
```

Core tools used throughout the notebook are given below:

- NumPy, Pandas, Matplotlib, Seaborn, Selenium, NetworkX, NLTK, Statsmodel, CountVectorizer, TfidfVectorizer, Tableau.
  - Models: Svm ,Tf-idf, Regularization with Lasso, Ridge, ElasticNet, Linear/Logistic Regression, kNN, Decision Trees, Crossvalidation, Grid-Search, Naive Bayes

Executive Summary

The project was born from the idea of simplifying the world of wines with Big Data.
The main points I touched upon are:

1) Demystifying grape varieties creating a simple dictionary of descriptors (mainly aromas and flavours) from the wine reviews.
2) Identifying common descriptors among grape varieties that would expose which grape varieties share the highest number of descriptors and lead the way to a wine recommender.
3) Understand what are the strongest predictors of grape varieties, wine ratings and value for money/ This has been done not just looking at wine reviews but also at Appellation, Winery and other predictors.

Through my analysis of 41000 Italian wine reviews scraped from the Wine Enthusiast's website, I aim to identify flavour characteristics which are predictive of grape varieties such as Sauvignon Blanc or Sangiovese. I would like to use these insights to create a dictionary of flavours for each grape variety, map the common traits of different grapes and represent them into a flavours map and eventually build a wine recommender.
In the project, I answered the following questions:

- Can we predict grape varieties from the text reviews & what words are the best predictors?
- How are the grape varieties connected in terms of common flavour characteristics? Can we create a map of flavours?
- Can we predict above or below median rating from the text reviews and what words are the best predictors?
- Can we predict ratings from text reviews?
- Can we predict value for money from the text reviews?
- Can we predict if the wine will be good, very good, superb or excellent from the text reviews?
- Can we predict Rating from the winery?
- Can we predict above/below median rating from the appellation?

Key Learnings and future potential developments

Findings:

The NLP analysis, as pointed out by the network chart, was capable of perfectly breaking down the grape varieties into white and red and and within these 2 groups into full body and medium/light body grape varieties, southern and northern, used in blends or not.

We could predict grape varieties from the text reviews with high accuracy (0.74 with log regression using Countvectorizer) and the most important coefficients are consistent with the grape varieties. Furthermore, we could predict an above or below median rating from the text reviews with an 84% R^2 and exact ratings with a 0.75% score on the CV and Test set using a Ridge Normalization. In this project RandomForest, Decision Trees and kNN performed worse than Ridge.

We also predicted Value for Money from the reviews with a 0.54 score on the CV,Test set using a Ridge normalization. We predicted 'Final Score' with a 0.63 score on the CV Set and we predicted above/below median Rating from the Winery and Appellation with a 0.7 CV score.

For more details, please look at the notebook or the presentation on GitHub.

Potential Improvements:

The Dataset is biased towards red wines and the most famous wine regions; therefore, it would be beneficial to extend the analysis to a bigger wine reviews sample, for instance a dataset scraped from Vivino for wines from different regions.

Because of time constraints, the number of models was limited and it would be a good idea to extend the models to Neural networks as well.

Furthermore, some more clustering and network analysis would also add value to the project as the common traits among grape varieties became one of the most interesting findings of this project.