Red or White Wine Predictions

Capstone Project

LaShanni Butler

Flatiron School

9/17/19

Agenda

- Problem Statement
- Methodology
  - Exploratory data analysis (EDA)
  - Statistical analysis
  - Machine Learning models
  - Deep Learning model
- Conclusion
- Future work

Problem Statement

- *Can we predict red or white wine from a dataset?*

  - How we’ll do this:
    - Look at attributes (or features) of each type of wine
    - Look at low, medium, and high quality wines statistical interactions

Methodology

- Wine Dataset: Obtained from Univ. of California, Irvine (UCI)

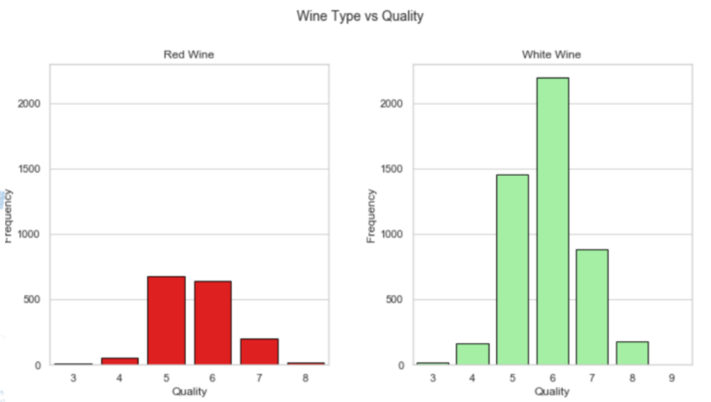
- Exploratory Data Analysis (EDA)
  - Analyze red and white wine data separately
  - Merge and analyze both datasets

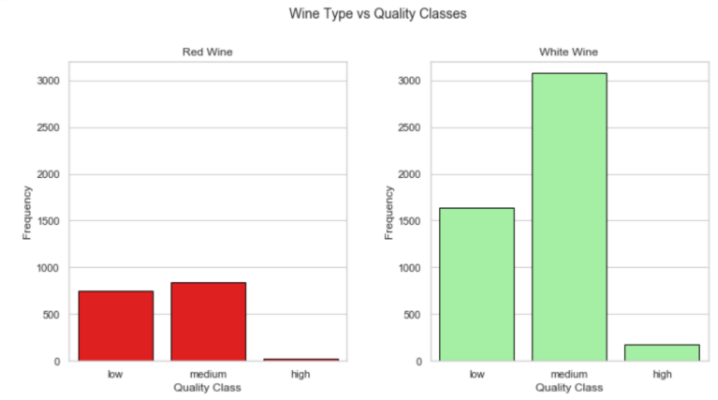
- Statistical analysis looking at wine quality

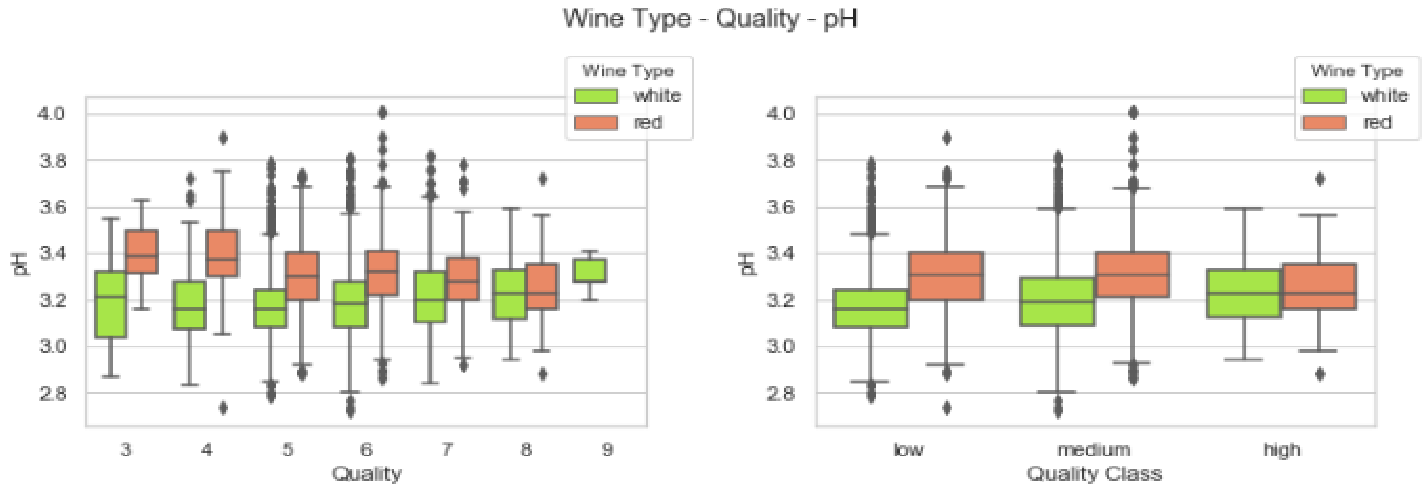
- Machine learning to predict red or white wine

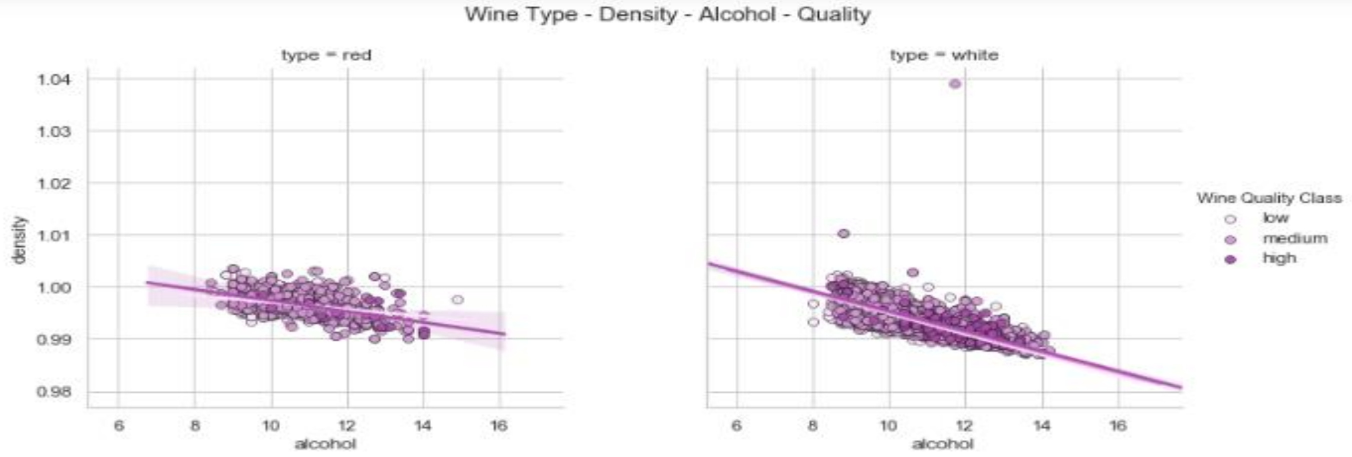
- Deep learning (“artificial brain”) to predict red or white wine

EDA


Wine Type vs Quality


Wine Type vs Quality Classes


Wine Type - Quality - pH


Wine Type - Density - Alcohol - Quality

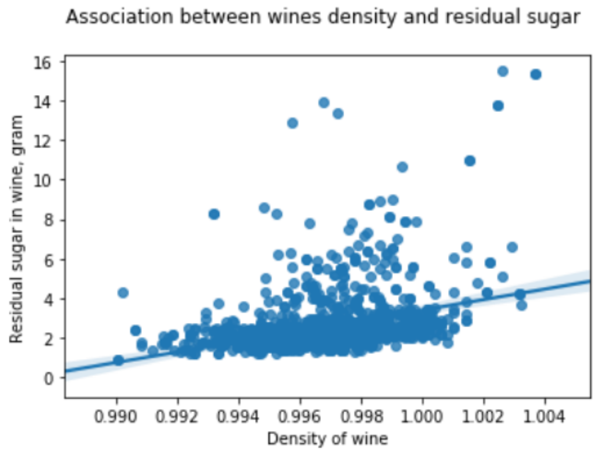
EDA

- Some key takeaways:
  - Alcohol concentration increases with the quality of wines.
  - There is no big difference in alcohol concentration between red and white wines in the same quality class.
  - Red wines are more dense than white wines. Additionally, red wines have a higher pH and sulphate concentration
  - Density has a relatively high negative correlation to alcohol (linear trend is decreasing from left to right).

Statistics

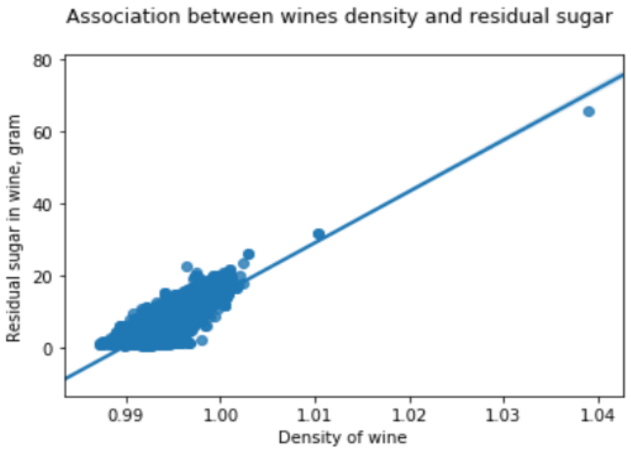
- Pearson's coefficient:

Red Wine


Association between wines density and residual sugar

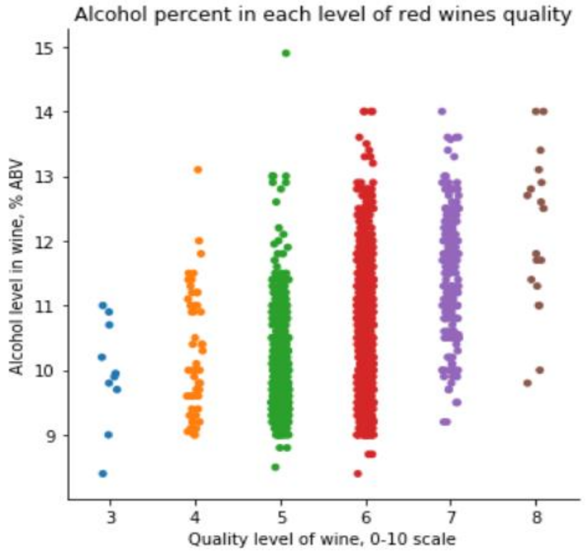
```
(0.3552833709833765, 9.013041728296711e-49)
```

White Wine

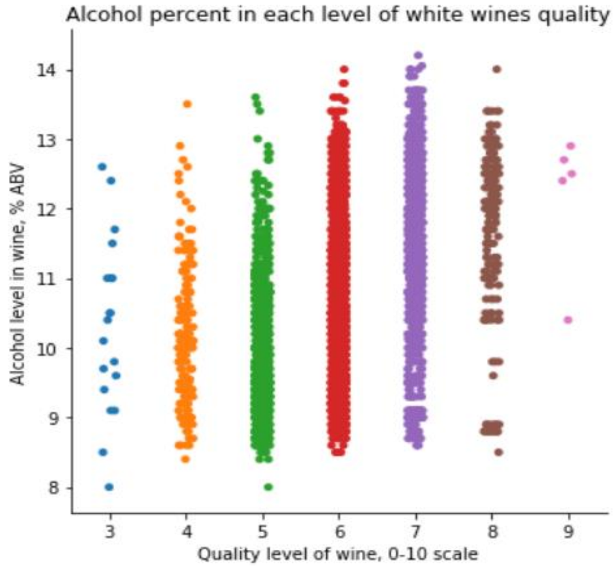

Association between wines density and residual sugar

```
(0.8389664549045837, 0.0)
```

Red Wine


Alcohol percent in each level of red wines quality

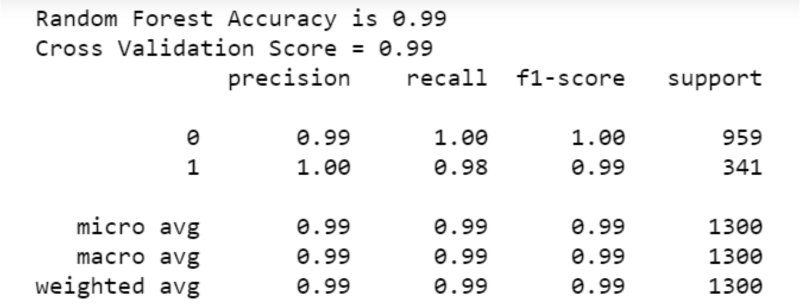
White Wine


Alcohol percent in each level of white wines quality

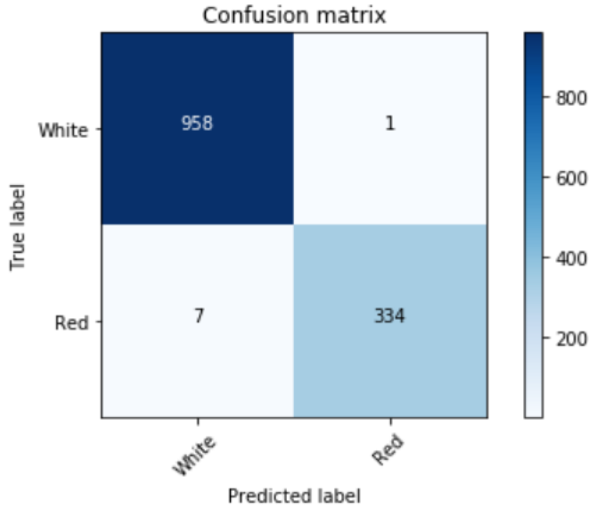
Machine Learning Models

- Machine learning models: K Nearest Neighbors, Logistic regression, Support Vector Machine (SVM), Decision Tree, Random Forest

- Confusion matrix: a summary of prediction results on a classification problem
  - Correct & incorrect predictions are summarized with count values by each class

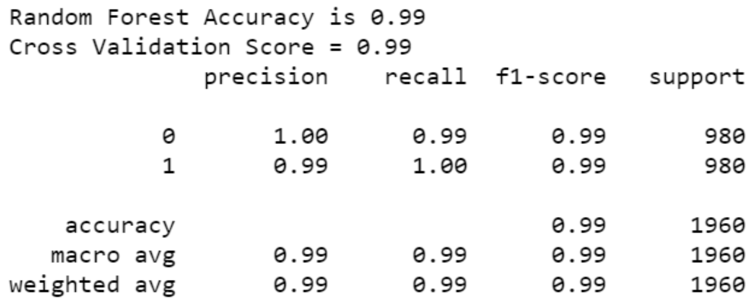


```
Random Forest Accuracy is 0.99
Cross Validation Score = 0.99
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       959
           1       1.00      0.98      0.99       341

   micro avg       0.99      0.99      0.99      1300
   macro avg       0.99      0.99      0.99      1300
weighted avg       0.99      0.99      0.99      1300
```

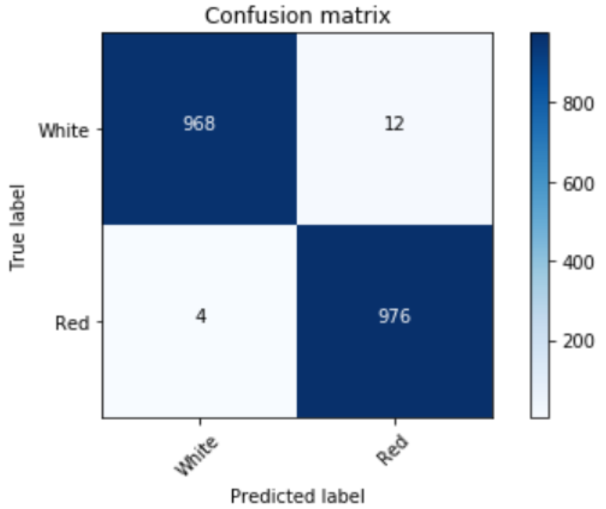


Machine Learning Models (Balanced dataset)

- Random Forest performed well, but the dataset was imbalanced
- Balancing the dataset, yields the same level of accuracy
- We can feel more confident with these results, since the dataset was balanced



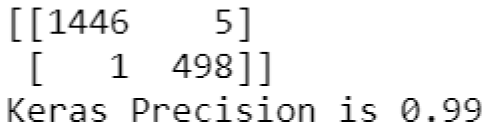
```
Random Forest Accuracy is 0.99
Cross Validation Score = 0.99
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       980
           1       0.99      1.00      0.99       980

    accuracy                           0.99      1960
   macro avg       0.99      0.99      0.99      1960
weighted avg       0.99      0.99      0.99      1960
```



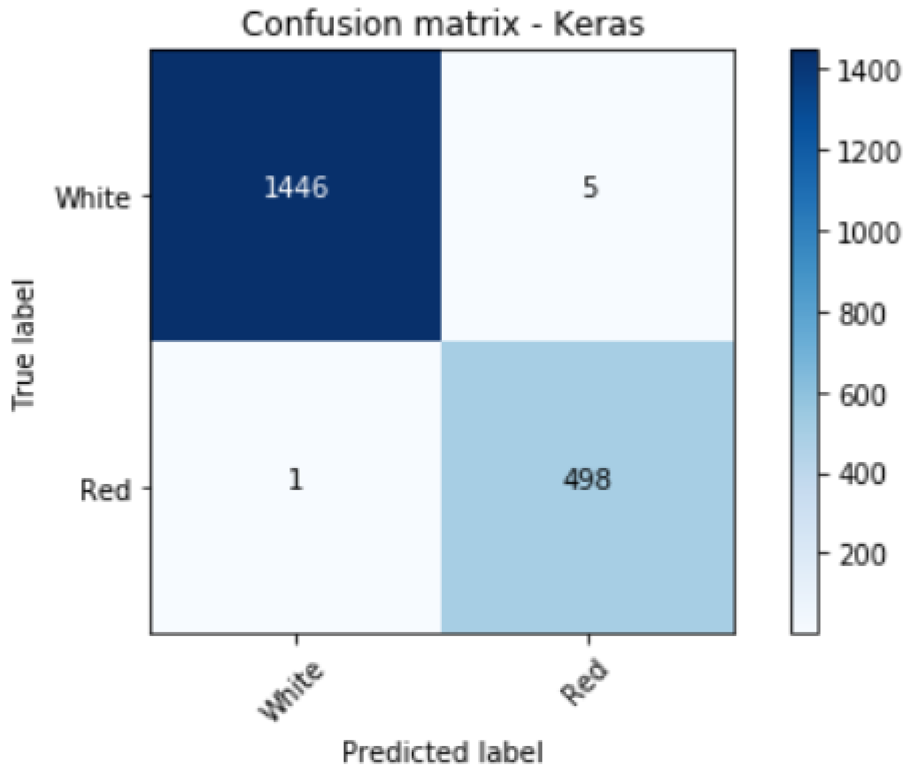
Confusion matrix

Deep learning Model

- Keras is a high-level neural networks focused on enabling fast experimentation

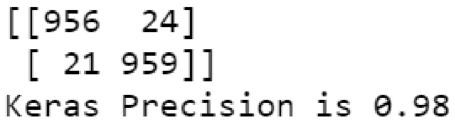


```
[[1446    5]
 [   1  498]]
Keras Precision is 0.99
```

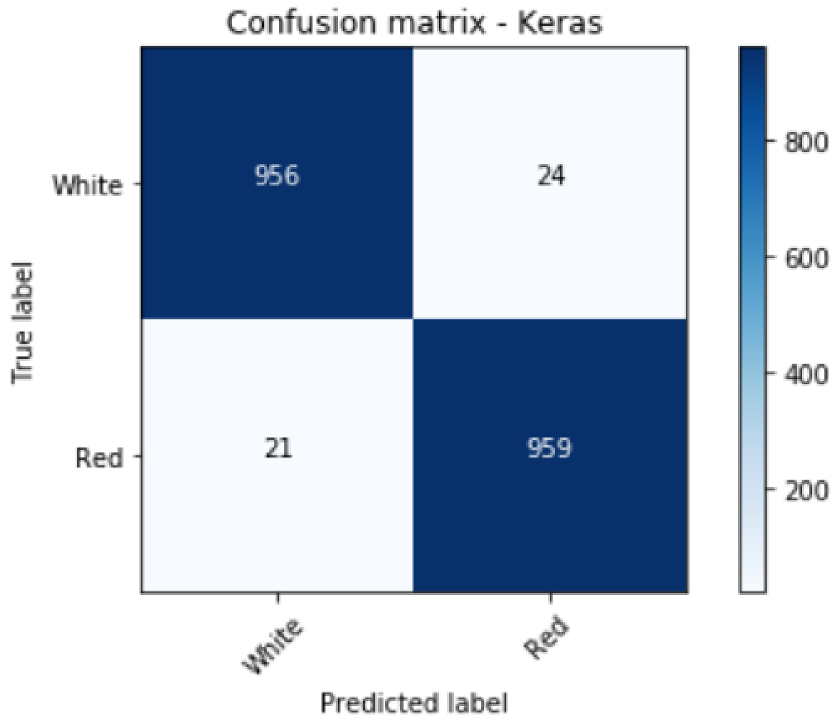


Deep learning Model (Balanced dataset)

- Again, the dataset yielded a high result, but was imbalanced
- Balancing the dataset yields a slightly lower lever of accuracy
- We can trust these results, since predictions were based off of a balanced dataset



```
[[956  24]
 [ 21 959]]
Keras Precision is 0.98
```



Conclusions

- EDA showed several strong relationships between the features and wine types
- Statistical analysis further shows some strong positive correlations
- Imbalanced datasets: Random Forest had 99% accuracy, Keras had a 100% accuracy in predicting wine type
- Balanced datasets: Random Forest had 99% accuracy, and Keras had 98% accuracy in predicting wine type
- Wine type classes were imbalanced, which I believe influenced high levels of modeling accuracy

Future work

- Improve parameters in Keras deep learning model to improve accuracy
- Possibly adjust training and testing set of data to see if that will yield higher accuracy
- Add more features to the dataset, making it more robust for testing
- Explore other deep learning models to determine level of accuracy

Thank You!

- Questions/Concerns/Comments?
- Video Walkthrough: https://drive.google.com/open?id=11FEHhyaleBIa5WBvfB4jbmY-TcI8DSNm