Bootcamp Kaggle competition

Group 5
Emmanuel Owusu Ahenkan
Volviane Saphir MFOGO
Alex Sananka

African Masters in Machine Intelligence (AMMI- Ghana)

October 24, 2019, Accra

Outlines

1. Introduction
2. Explanatory Data Analysis
3. Feature Engineering
4. Algorithms Used
5. Conclusion

Outline

1. Introduction
2. Explanatory Data Analysis
3. Feature Engineering
4. Algorithms Used
5. Conclusion

AIMS | African Institute for Mathematical Sciences GHANA

Introduction

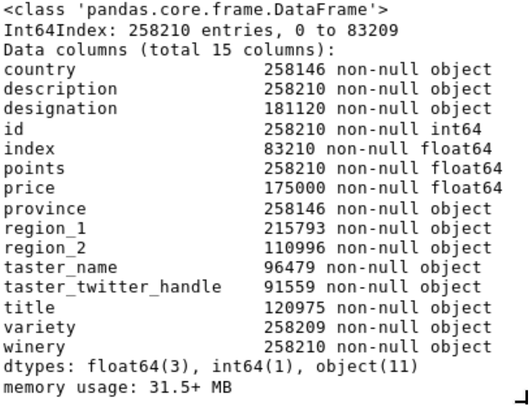
- Predicting the price of wine based on a collection of reviews and other product features.
- We use the Random Forest Regressor and the XGBoost Algorithm to predict the prices of wine.

Outline

1. Introduction
2. Explanatory Data Analysis
3. Feature Engineering
4. Algorithms Used
5. Conclusion

Explanatory Data Analysis

Let's start looking the information about our data.

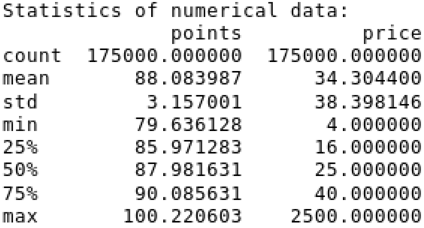


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 258210 entries, 0 to 83209
Data columns (total 15 columns):
country                  258146 non-null object
description              258210 non-null object
designation              181120 non-null object
id                       258210 non-null int64
index                    83210 non-null float64
points                   258210 non-null float64
price                    175000 non-null float64
province                 258146 non-null object
region_1                 215793 non-null object
region_2                 110996 non-null object
taster_name              96479 non-null object
taster_twitter_handle    91559 non-null object
title                    120975 non-null object
variety                  258209 non-null object
winery                   258210 non-null object
dtypes: float64(3), int64(1), object(11)
memory usage: 31.5+ MB
```

- We can see that the total number of columns is 15
- We can also see the number of non-null value of each columns and their type.

Explanatory Data Analysis

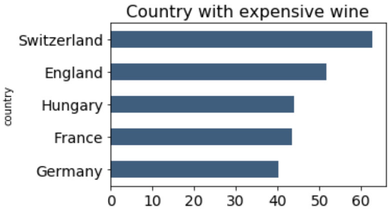
Now let's look at the distribution of Points and Prices

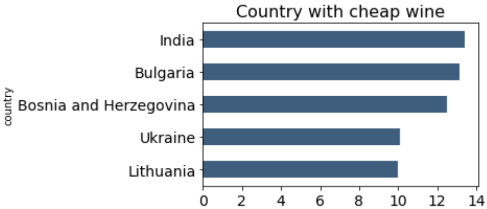


```
Statistics of numerical data:
              points          price
count  175000.000000  175000.000000
mean       88.083987      34.304400
std         3.157001      38.398146
min        79.636128       4.000000
25%        85.971283      16.000000
50%        87.981631      25.000000
75%        90.085631      40.000000
max       100.220603    2500.000000
```

- The values of points are distributed between 80 and 100
- Mean price of 34.3 and average points is 88

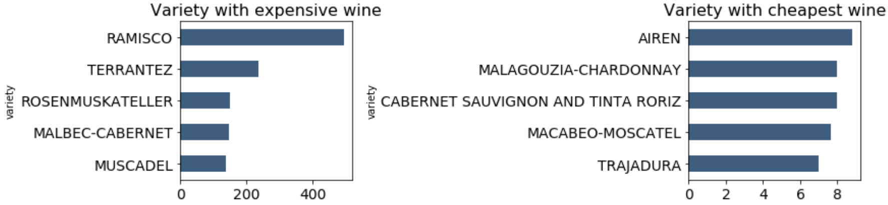
Explanatory Data Analysis





- Most expensive wine is from Switzerland
- Cheapest wine is from India

Explanatory Data Analysis



- Most expensive variety is Ramisco.
- Cheapest variety is Airen.

Outline

1 Introduction

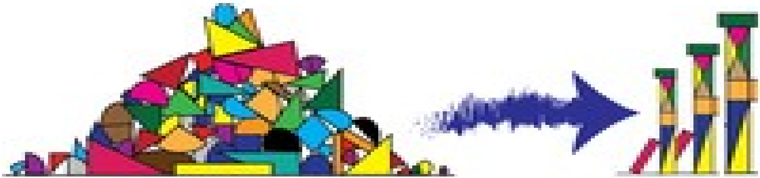
2 Explanatory Data Analysis

3 Feature Engineering

4 Algorithms Used

5 Conclusion

Feature Engineering



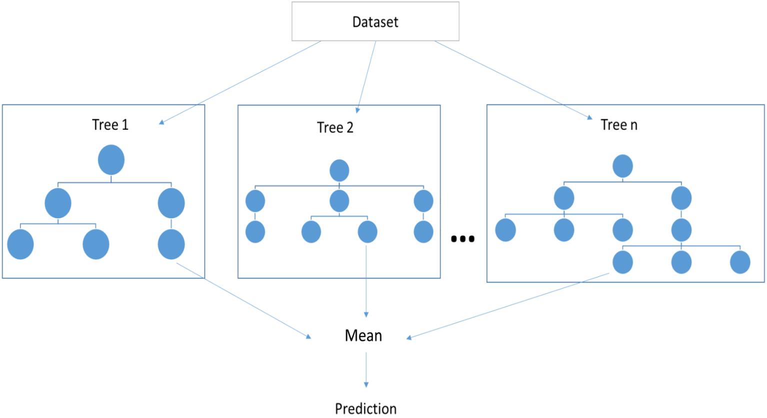
- Label encoding
- One Hot Encoding
- TF-IDF

Outline

1. Introduction
2. Explanatory Data Analysis
3. Feature Engineering
4. Algorithms Used
5. Conclusion

Random Forest Regression

Random forest is a Supervised Learning algorithm.



XGBoost

- Supervised machine learning algorithm
- Predicts target variable by combining estimates of simpler, weaker models.
- Incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models mis-classified.

Results

Model	RMSE (Validation)	RMSE (Test)
Random Forest Regressor	21.6	
XGBoost Regressor	23.8	
XGB + Random Forest	22.7	20.43

- 10-Fold cross validation applied in both models.
- Random Forest performs better than XGBoost.
- Combining the models gives even better results.

AIMS African Institute for Mathematical Sciences GHANA

Outline

1 Introduction

2 Explanatory Data Analysis

3 Feature Engineering

4 Algorithms Used

5 Conclusion

Conclusion

Two models were used with cross validation. Their predictions were averaged and RMSE of 20.4 was achieved.

References

- http://webdropin.com/wordpress99/answering-wine-related-questions-with-data/
- https://www.udemy.com/course/feature-engineering-for-machine-learning/

THANK YOU
FOR YOUR
ATTENTION