Kaggle Competition Presentation

Bottle of Wine Price Prediction
by Group 12

AMMI

Group Members
1. Diene Madiou
2. Ogun Sewade Olaolu

Content

- Introduction
- Data Preparation
- Dealing with Null Values
- Feature Engineering
- Encoding
- Dealing with Text - “Description” column
- Model Prediction
- Conclusion

Introduction

- To predict the price of a bottle of wine based on a collection of over one hundred thousand reviews and other product features.
- Evaluation: Root Mean Squared Error (RMSE).

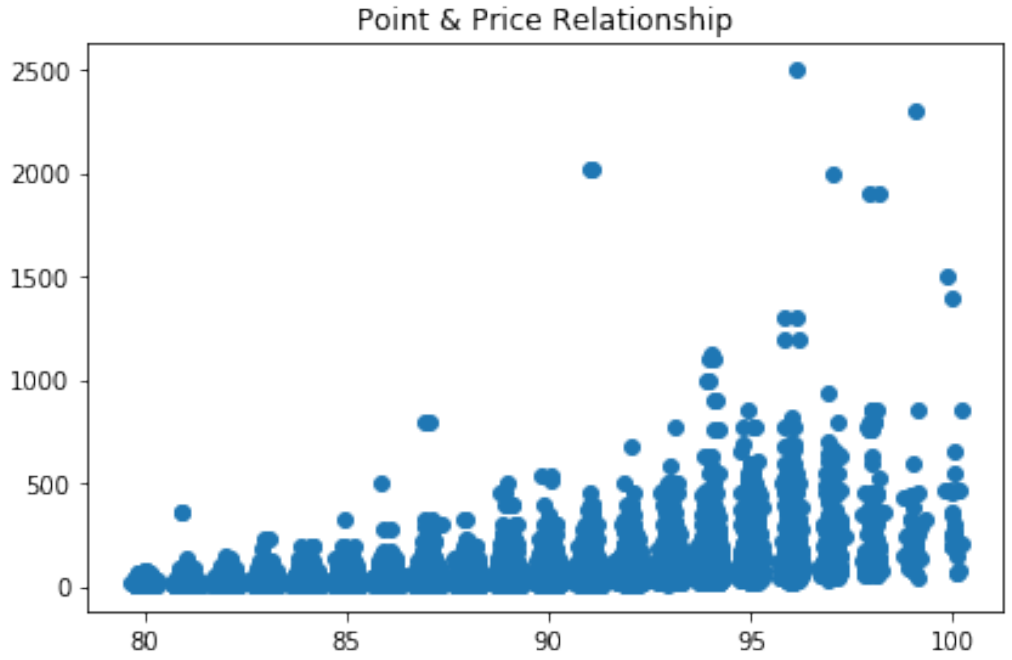
$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

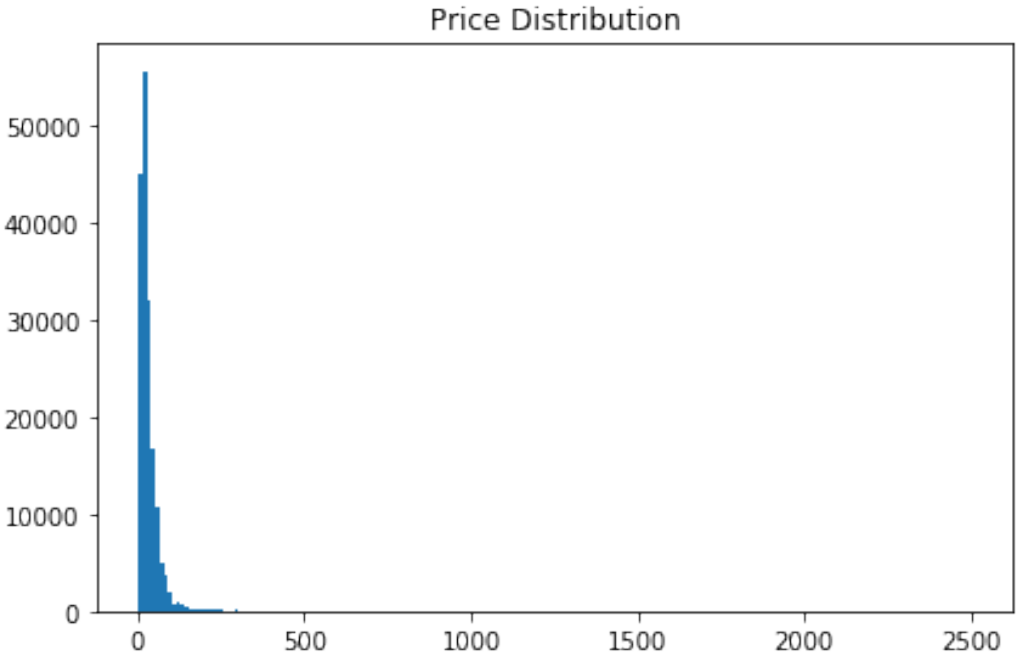
Data Preparation

	train_data (17500)	test_data (83210)
country	174953	83193
description	175000	83210
designation	122734	58386
points	175000	83210
price	175000	0
province	174953	83193
region_1	146466	69327
region_2	75394	35602
taster_name	65509	30970
taster_twitter_handle	62190	29369
title	82189	38786
variety	174999	83210
winery	175000	83210

```
RangeIndex: 175000 entries, 0 to 174999
Data columns (total 13 columns):
country                    174953 non-null object
description                175000 non-null object
designation                122734 non-null object
points                     175000 non-null float64
price                      175000 non-null float64
province                 6 174953 non-null object
region_1                   146466 non-null object
region_2                   75394 non-null object
taster_name                65509 non-null object
taster_twitter_handle      62190 non-null object
title                      82189 non-null object
variety                    174999 non-null object
winery                     175000 non-null object
```

Price Distribution





Dealing with Null Values

High number of null values in region_2, taster_name, taster_twitter_handle, title, designation, region_1

	% Null Values in Data Set
country	0.024786
description	0.000000
designation	29.855544
points	0.000000
province	0.024786
region_1	16.427327
region_1	16.427327
region_2	57.013284
taster_name	62.635452
taster_twitter_handle	64.540878
title	53.148600
variety	0.000000
winery	0.000000

Extracting Designation and Region from Title

	title	designation	region_1
	Mas Fi NV Brut Nature Reserva Sparkling (Cava)	Brut Nature Reserva	Cava
	Château Tayet 2014 Cuvée Prestige (Bordeaux S...	Cuvée Prestige	Bordeaux Supérieur
	NaN	Baccante	Sicilia
	Matrix 2013 Pinot Noir (Russian River Valley)	NaN	Russian River Valley
	NaN	Lellè Extra Dry	Prosecco
	Domaine d'Eole 2005 Réserve des Gardians Red (...	Réserve des Gardians	Coteaux d'Aix-en-Provence
	NaN	NaN	NaN
	NaN	Thauvenay	Sancerre
	NaN	Vin d'Eliza	Paso Robles
	Goat Bubbles 2011 Sierra Madre Vineyard Créman...	Sierra Madre Vineyard Crémant	Santa Maria Valley

Information of 'designation' and ‘region_1’ is contained in 'title'

Feature Engineering

- New Feature “Ratings” generated from Points (to reduce dimensionality)

- Range of Points is 79 to 101

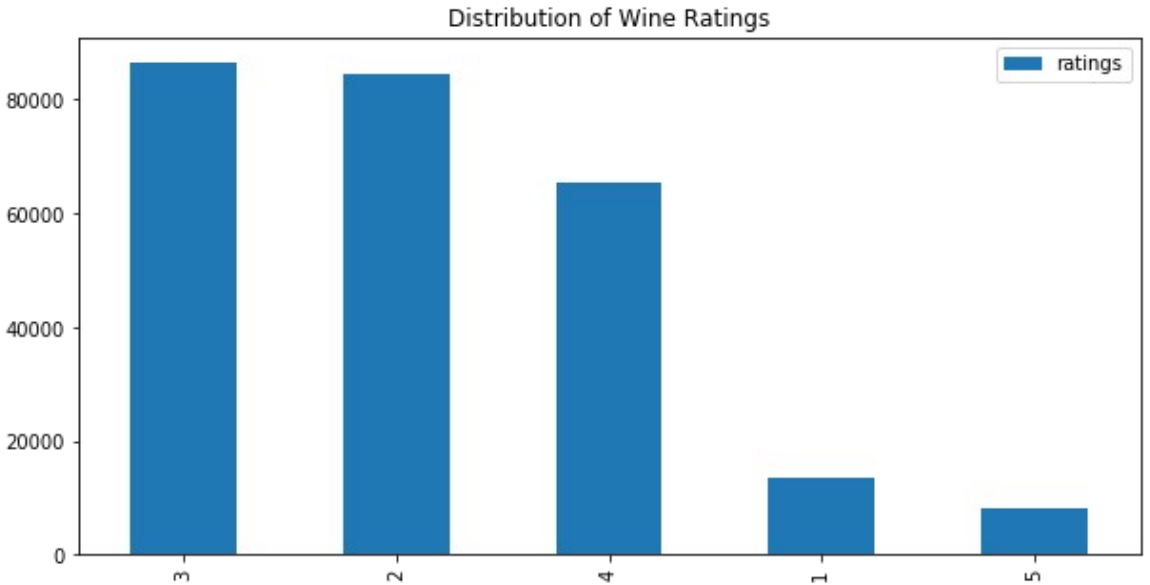
  1 -> Points 80 to 82 (Acceptable wines)

  2 -> Points 83 to 86 (Good wines)

  3 -> Points 87 to 89 (Very Good wines)

  4 -> Points 90 to 93 (Excellent wines)

  5 -> Points 94 to 101 (Superb wines)



Source: https://towardsdatascience.com/predicting-wine-quality-using-text-reviews-8bddaeb5285d

Feature Engineering (Contd.)

- New column ‘country_region’ was extracted from ‘countries’

Country/Territory	
Oman	Western Asia
Qatar	Western Asia
Saudi Arabia	Western Asia
State of Palestine	Western Asia
Syrian Arab Republic	Western Asia
	...
Nicaragua	Central America
Panama	Central America
Bermuda	Northern America
Canada	Northern America
Greenland	Northern America

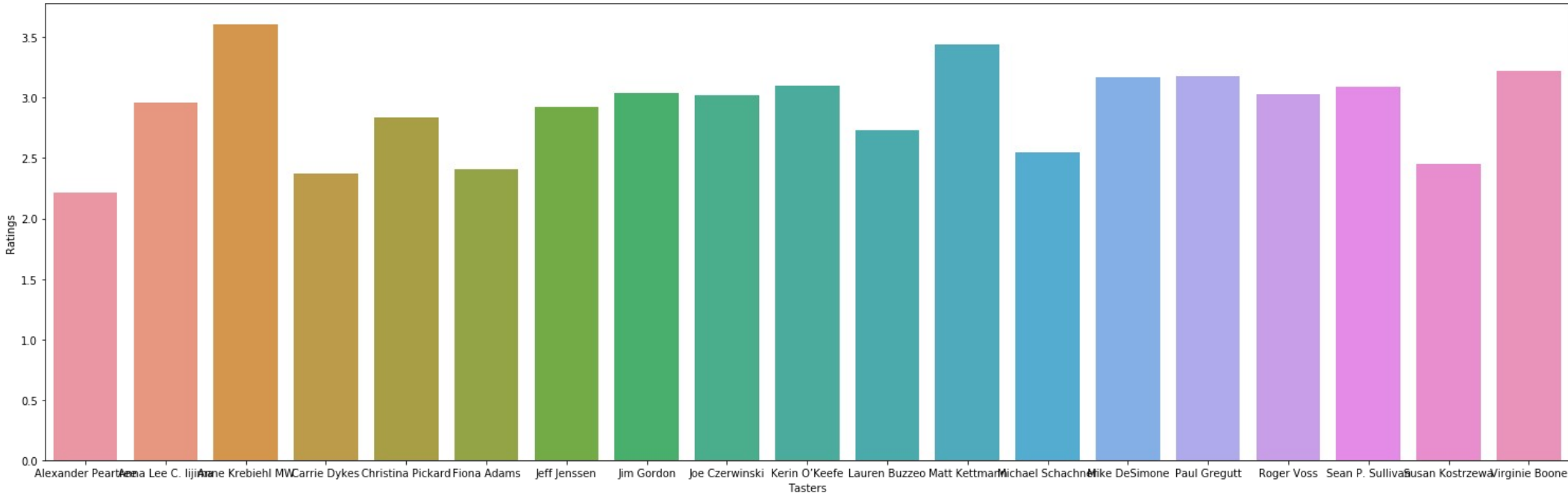
Northern America	116852
Southern Europe	61194
Western Europe	42337
South America	19807
Australia and New Zealand	11636
Southern Africa	3530
Western Asia	1482
Eastern Europe	1023
Central America	133
Northern Europe	85
Northern Africa	40
Southern Asia	17
Eastern Asia	10

Source of data: http://statisticstimes.com/geography/countries-by-continents.php

Feature Engineering (Contd)

- Column ‘has_twitter_handle’ (Boolean) was created to know if the taster has an handle or not.
- Year of production extracted from ‘title’ as a standalone feature.
- Number of words (n_words) and Number of Characters (n_chars) extracted from ‘description’
- Polynomial interactions were done on the columns after encoding.

Exploring Tasters



Not much bias in their ratings

Encoding

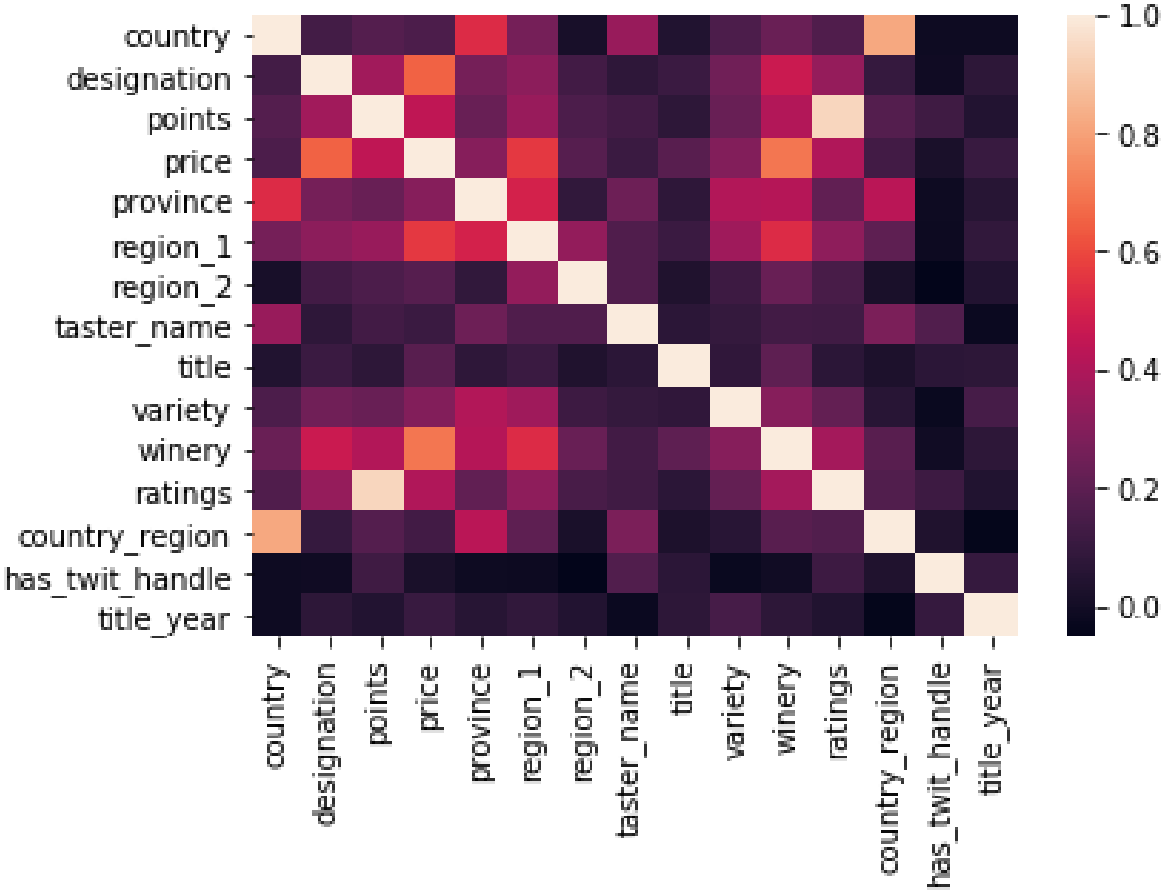
- Target Encoding.

  A Very Efficuent Preprocessing Scheme for High-Cardinality Categorical Attributes.

- One-Hot Encoding

Link: https://dl.acm.org/citation.cfm?id=507538

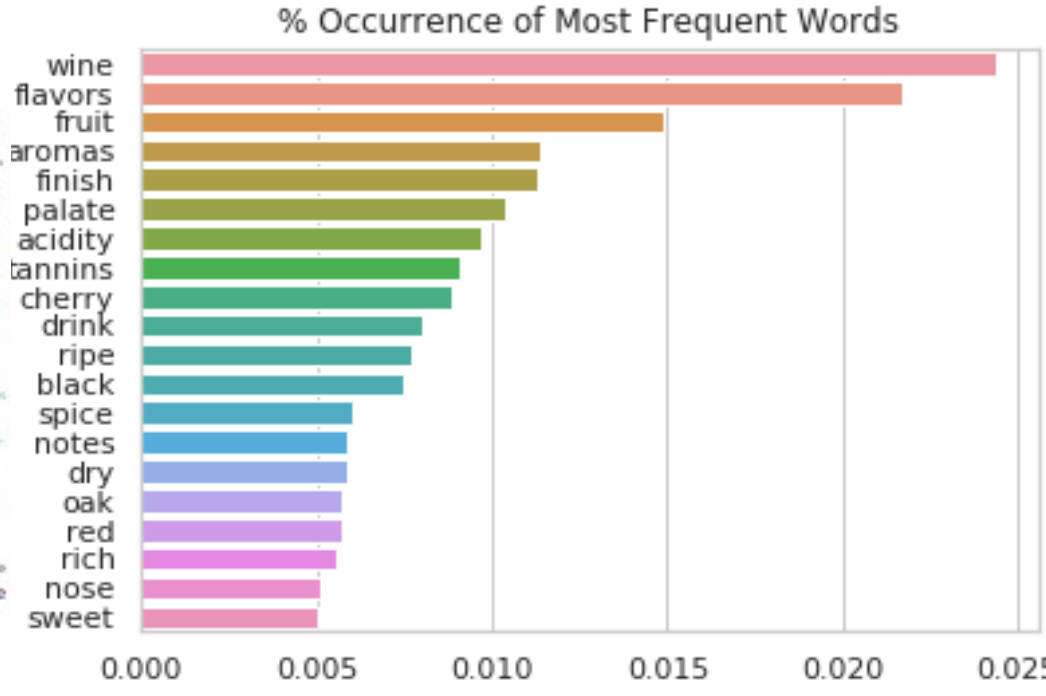
Feature Correlation



High correlation between points and ratings, *country and country_region*

Dealing with Text Column ‘Description’

- Word Cloud



TF-IDF applied on description column to generate a matrix of words.

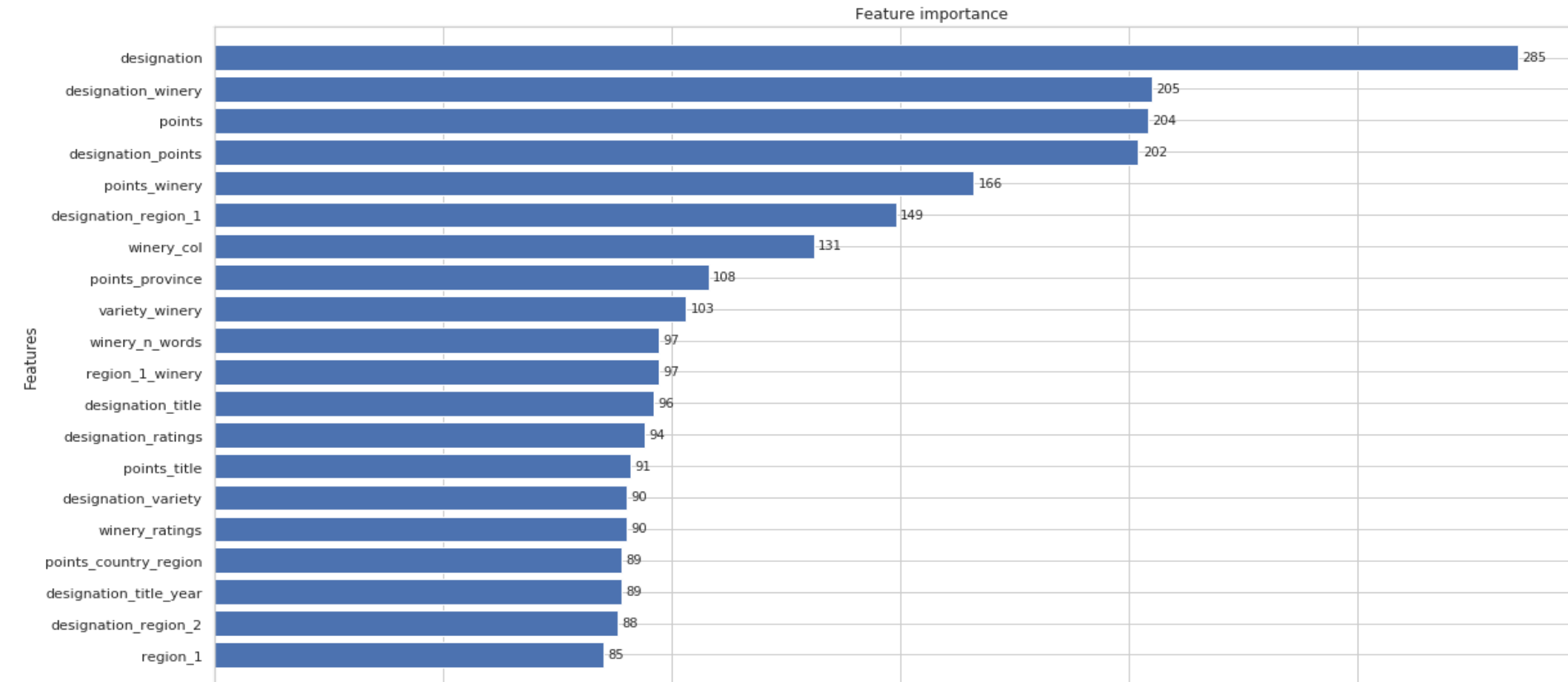
Model Prediction - XGBoosting

K-Fold Cross Validation – Sample

```
n_estimators:  72

Model Report
R2 Score : 0.9483
RMSE Score (Train): 17.454611
    train-rmse-mean  train-rmse-std  test-rmse-mean  test-rmse-std
67         8.653176        0.220763       16.375265       2.412200
68         8.630584        0.222806       16.381915       2.402022
69         8.601453        0.220496       16.376520       2.405377
70         8.568110        0.222205       16.375224       2.404391
71         8.540594        0.223437       16.368812       2.402139
```

Model Prediction – Feature Importance



Conclusion

- XGBoost was the main winner for my predictive modelling (17.05 & 17.21 Private). However, it was difficult avoiding overfitting.
- Other Things We tried

  Models: Random Forest, Neural Network, Ridge Regression

  Scaling features using MinMax Scaler and Log Scaling

  Dimensionality Reduction using TruncatedSVD/PCA. Lot of information lost

  Topic Extraction using Latent Dirichlet Allocation.

  Ensembling of Models.
- What we will try next time: Outlier Removal.

Thank You

Merci

E se