Project_manaskm2

Manas Kumar Mukherjee

11/23/2019

1. [10 Points, half a page] Project description and summary. This part should summerise your goal, your approach, and your results.

Overview

In this project, I've used the 'Wine Review' data from Kaggle. In this dataset, there are the following thirteen features related to the wine reviews.

country, description, designation, points, price, province, region_1, region_2, taster_name, taster_twitter_handle, title, variety, winery

Goals:

1. Build two statistical models on the *wine review* data and predict the points rated by WineEnthusiast on a scale of 1 - 100.
2. Recommending wineries based on the customer's price and taste preferences.

Approach:

Here is the steps followed to achieve the first goal.

- At first, an extensive exploraty data analsyis is done on the entire dataset.
- Next, applied different types of feature engineeriFng techniques on individual features. Example-
    - All features are merged and evaluated using various machine learning algorithm.
    - Best two models are choosen based on the low RMSE.
- Split the data into train and test set with 70-30 split.
- Built base models using two suitable algorithms called - *Lasso* and *XGBoost* and recorded the performance. Here, rmse is used for evaluation.
- Tuned parameters to optimzed the model.

To recommend the wineries, two different procedures are followed.

1. In the first approach, top 5 wineries which sell wines which have a *fruity* flavor, made of *pinot noir* grape, and with a price less than 20 dollars are recommended.
2. In the second approach, top 5 wineries are recommended based on the input from the customer. Here if he/she tells the name of the preferred wine, then 5 wineres which produce that kind of wine are recommended.

Result:

In part1, the *Lasso* and *XGBoost* based regression models produced the following RMSE values.

Lasso RMSE: 1.61 XGBoost RMSE: 1.7

In the second part, the recommendation is based on the cusotmer's specic wine choice. Ex -If we assume the end-customer has liked a wine with title *Rascal 2014 Pinot Noir (Oregon)* then the recommended wineries are

- Huia
- Willamette Valley Vineyards
- Acrobat
- Underwood
- Pali

2. [10 Points, within 1 page] Data processing. Describe how you process the data so that it can be analyzed by a regression model. This includes, but not limited to, processing text data.

The following feature engineeing processes are applied on the various input features.

Handling missing values:

- Price columns had a large(~8K) number of missing values which are removed from the dataset.
- Only one record with missing variety column is also removed.
- 59 records with missing country information are removed.
- 10K records were duplicate records are also removed.

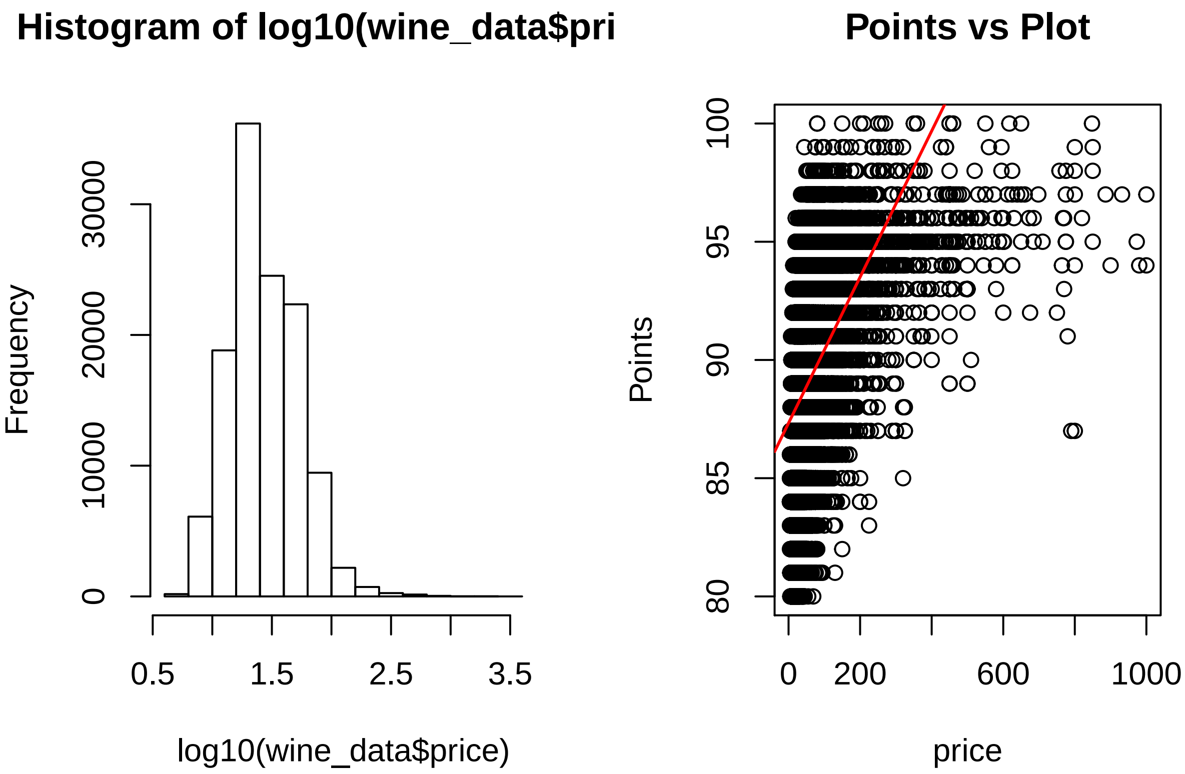
Feature engineering:

- Price values are distributed in 0 - 3000$ range but most wine prices are less than $200. To make the distribution less skewed, price feature is transformed to log10 scale.
- The text data of the description column is featurized using the TF-IDF technique.
- After doing some thorough exploratory data analysis the following 3 categorial variables were included in the model building process. Before using these categorical features(factor variables) in the model, these are transformed to dummy variable represetntaiton using one-hot encoding technique.
  - The original wine dataset has too many type of grapes(variety). To make the model more interpretable, only the top 50 most frequent varities are kept as it as and rest others are renamed as 'other_varity'.
  - Similar to the processing of the *variety* feature, only the top 25 most frequet countries are kept as it is and rest are categoried as 'other_country'.
- The vintage/year values are extracted from the *title* column and a new feature called *vintage_age* is created by subtracting each vintage year from 2020(baseline). PN: For all records where no year was mentioned in the title, a default baseline value(2020) is used.
- Another new feature called 'continent' is created by mapping each country to its continent. This is done by using a library called `countrycode`.

3. [10 Points, within 1 page] Descriptive statistics with tables/figures. Provide summaries of your data, which may motivate your particular choice of the regression model.

3.1 Importance of wine-price w.r.t the points.

The following plot sugguests that there is a strong relationship between the wine points and their prices.



After doing EDA and cross-validation, it is observed that the following 3 categorical variables(Variety, TasterName, and Country) are important to model the wine price data.

3.2 Importance of the categorical columns

3.2 Importance of the description column

The wine-description field is an important factor which are closely related to the wine price and point. To use this text feature in the regression model, the text is represented using the TF-IDF approach. From the train set itself, a vocabulary of ~29,000 words as features were created. To handle these large number of features and making sure that the final model interpretable, the *Lasso* linear regression method which does implicit *feature selection* is choosen. In addition to the *Lasso* model, the XGBoost method is slected because of its reputation and robustness in dealing with the missing data.

4. [35 Points, within 3 pages] Regression model analysis.

The *Lasso* and “XGBoost” regression models are used for this use cases. Both the algorithms are very popular and robust. Lasso also does implicit variable selection which is required to reduce the high number of features generated from the TF-IDF representation of the ‘description’

Here are the details of these two models.

4.1 Regression model using Lasso

4.1 Data Processsing

4.2 Handling categorical variables

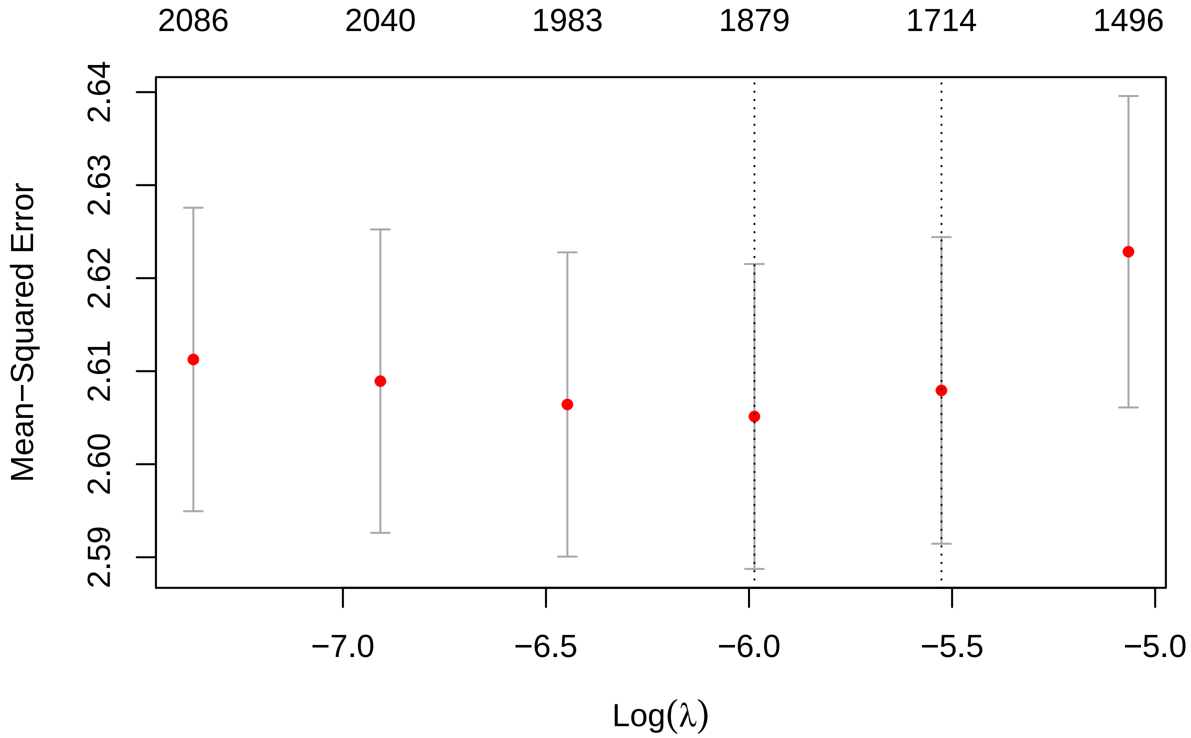
4.3 Create final train and test set

4.4 Build the best lasso model using cross-validation

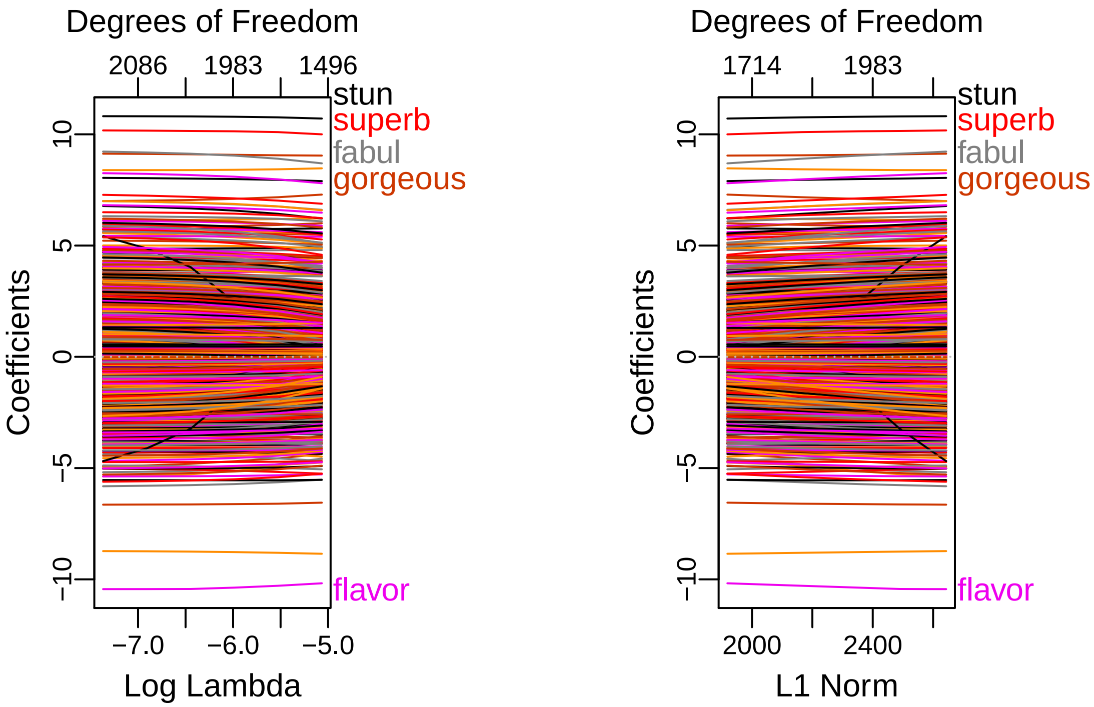
4.5 Prediction using the lasso model

```
## [1] "Test RMSE : 1.6042"
```

4.6 Plot(MSE vs Lambda) of the fitted lasso model.



4.7 Lasso model interpretation



The above plot shows that with the increase lambda values(L1 regularization), the more and more features' co-efficients become 0. Here the *lambda* shrinks the less important feature coefficients to zero, thus it implicitely does feature selection which is one of the reason behind considering this model.

```
## (Intercept)      flavor      superb        stun    gorgeous
##   82.714855   10.605129   10.591467   10.230525    9.007839
```

In the Lasso model, the TF_IDF representation of the *descripton* was used as features. The above result shows that positive adjuctives like superb, fabulous, amazing, delicious etc. were highly related to the points of a wine.

4.2 Regression model using XGBoost

The model parameters are selected after doing cross-validation. Mainly the following three parameters are tuned to get the optimal result.

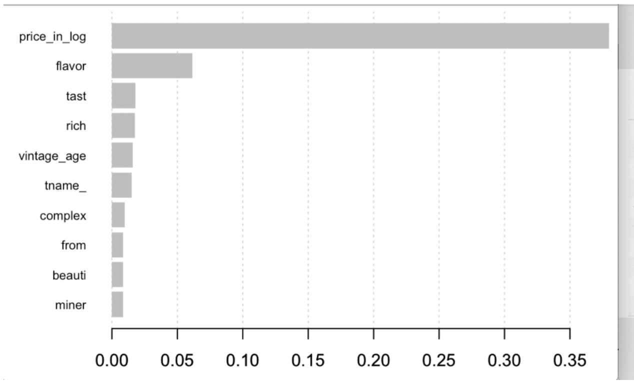
- max_depth is increased to 10 from 6.
- The model was fit multiple times usnig decaying eta. It got trained 4 times with eta-values 0.6, 0.25, and 0.1.
- subsample and colsample_bytree params tuned to control the number of featuers used while building trees. It improved training time as well as helped to avoid over-fitting.
- Alpha is used to control the L1 regularization(similar to the above Lasso model).

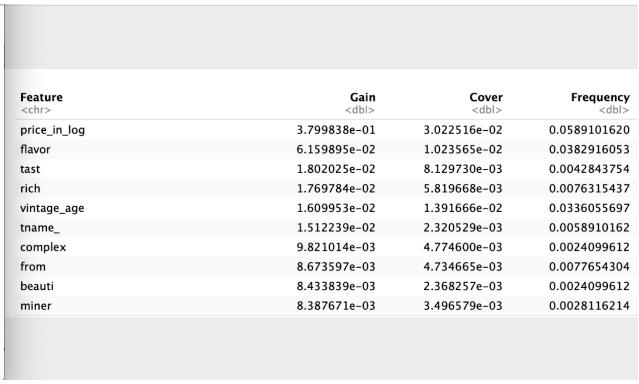
```
## [1] "Test RMSE : 1.7069"
```

Feature importance

Importance weightage

The feature importance table shows that the top 5 important fetures are *Price*, *flavor*, *taste*, *richness* and *vintage* of the wine. The result suggests that the *Price* is the single most significant factor related to the *Point* of a wine. As expected, the *vintage* of the wine also came up as one of the top 5 important attribute.





Feature <chr>	Gain <dbl>	Cover <dbl>	Frequency <dbl>
price_in_log	3.799838e-01	3.022516e-02	0.0589101620
flavor	6.159895e-02	1.023565e-02	0.0382916053
tast	1.802025e-02	8.129730e-03	0.0042843754
rich	1.769784e-02	5.819668e-03	0.0076315437
vintage_age	1.609953e-02	1.391666e-02	0.0336055697
tname_	1.512239e-02	2.320529e-03	0.0058910162
complex	9.821014e-03	4.774600e-03	0.0024099612
from	8.673597e-03	4.734665e-03	0.0077654304
beauti	8.433839e-03	2.368257e-03	0.0024099612
miner	8.387671e-03	3.496579e-03	0.0028116214

5. [20 Points, within 1 page] Recommend five different wineries for a customer that is interested in purchasing a pinot noir, with a price less than 20 dollars, and has a fruity taste.

The following filters are used to find the potential wines and wineries which sells *fruity pinot noir* with price less than $20. *Description* contains the word *fruit*, *Price* is less than $20 and *Variety* is "pinot noir". There are 522 wines which and 367 related wineries satisfy this critera.

```
## [1] 522  15
```

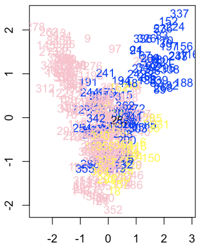


Figure 1: MDS-Scaling

5.1 Aanlysis of concatenated wine description per winery

The descriptions of all the wines per winery are concatenated and those are represented using the TF-IDF form.

To recommend wineries, we assumed that the customer would share his favorvite/preferred wine. The recommendation model will get the description of the customer selected wine, match it with the all the wines' descriptions(concatenated) per winery and sort it by similarity score.

In this cases, we have considered the end-customer has liked a wine with title *Rascal 2014 Pinot Noir (Oregon)*

```
##                            winery MeanScore
## 143                          Huia  91.00000
## 359 Willamette Valley Vineyards  90.00000
## 5                         Acrobat  87.66667
## 327                     Underwood  87.00000
## 233                          Pali  83.00000
```

The above list shows the list of 5 recommended wineries where the customer can find wines simliar to "Rascal 2014 Pinot Noir (Oregon)"

The following chart shows, Multidimensional-scaling(MDS using tf-idf of description) of all the 368 winries. It clearly shows that there are 3 different groups of wineries. The right hand side plot shows the same plot with colors based on the continents where these wineries belong. These two plots show similar kind of clustering which suggests that there are some differences in the wineries in Americas, Europe and Oceania. So whlie recommending wineries, the above recommendation model will first find a subset of the wineries which belong to the same continent as the customer, then it would find winneries based on the customer's wine preference and ultimaltely rank those wineries based on the average score/points of those wineries.