Wine Quality Prediction Using Regression

Abdul Hadi Saeed
Politecnico di Torino
Student id: 290480
S290480@studenti.polito.it

***Abstract*—The purpose of this given problem to predict the quality of the wine in this given evaluation by obtaining its analysis from the Wine Quality dataset. The proposed solution consists of Ridge Regression, which we have applied on the description column present in dataset. These descriptions are the reviews provided by the customers of wine in different countries. Two datasets are available, one is the development dataset and the other for the evaluation of the results of the wine quality review. The proposed solution is predicting the wine quality.**

I. Problem Overview

The given problem was to predict the Quality of wine on the given datasets. The dataset consists of various parameters such as Country, Designation, province and the most significant column was Description and Quality. The purpose was to predict the quality of wine by using Regression but for regression we need numerical data and the data on which we have to train the model was consist of Categorical data. We need to apply some techniques to change the data from categorical to numerical.

The Two types of datasets were given:

- A dataset for development having quality column
- An evaluation which doesn't have quality column

For predicting the quality of wine for the evaluation set, I will use the development dataset to construct a regression model because it has quality column on which we can train and test the model.

The development dataset has two columns on which it is important to concentrate. One is the 'quality' feature because it is our target feature, and the 'description' is another essential feature of the dataset because the description includes the customer's review of wine. The quality score is correlated with the wine review.

To infer the quality of the wine on description column which are the reviews provided by wine consumers, a regression model is needed.

First, the given dataset had to be pre-processed, then a model would be selected for regression and then hyperparameters would need to be tuned.

II. Proposed Approach

A. *Preprocessing*

In the given dataset, the two columns "Description" and "Quality" were important for training the model and also for prediction. Our targeted variable was "Quality" which was only present in one dataset (Development) not on another dataset (Evaluation), That's why we have drop the "quality" column for Preprocessing purpose. I have selected the Description column.

The description column was categorical which means that we cannot apply Regression model techniques until and unless we change it into non-categorical data. First the description column was composed of reviews along with different countries. Which means every word in a sentence has some meaning in description which is describing the quality of different wines.

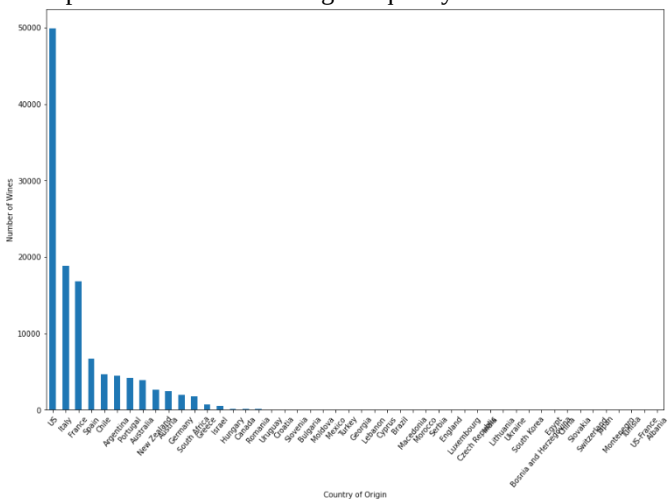

Figure1: Bar graph

In the given data set, It is seen that Among all countries producing wine, US has more than 50,000 types of wine in the wine review dataset, twice as much as the next one in the rank: Italy - the country famous for its wine. France also produces a lot of quality wine, having nearly 20,000 wines open to review.

Let's now take a look at the plot of all 44 countries by its highest rated wine, using the same plotting technique as above:

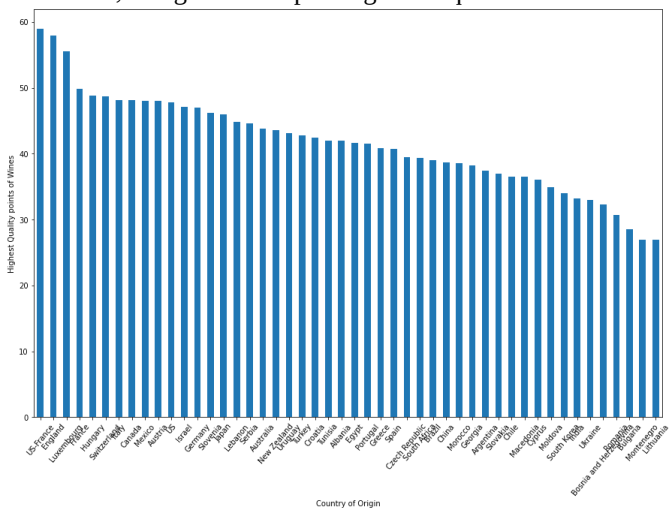

Figure2: Bar graph

For changing type from categorical to numerical, I applied preprocessing techniques which are following:

1) OneHot Encoding:

I have Encode categorical features as a one-hot numeric array. First, I used fit function of OneHot encoding and then transform the data into binary values and retrieve the data in parse matrix form.

2) TfidVectorizer

After OneHot Encoding, I have used the TfidVectorizer to preprocess my description column because TfidVectorizer Transforms text to feature vectors that can be used as input to estimator. vocabulary_ Is a dictionary that converts each token (word) to feature index in the matrix, each unique token gets a feature index. ... In each vector the numbers (weights) represent features tf-idf score and we need tf-idf score for predicting the quality of wines.

I have made the function of pre_processor( ) in which I am passing every sentence of description column. The pre_processor( ) function first cleaning the description columns from all the punctuations, numbers and alphanumerical values. After cleaning the description data the pre_processor( ) function is making all the sentences into lowercase.
After this, the function will perform Tokenization means the description column data will going to be split into words.
Next this function will remove all the Stop-Words, such as “is”, “am”, “are” and etc.

Removing stop words are not enough to proceed our description data for predicting wine quality because in our dataset the repeated words are also their which will affect Tf-idf score. So, function is performing stemming technique which will count all the repeated words as one. I.E: “play”, “played” this will be count as one word only.

Figure3: Words Cloud

After doing all this, In function of TfidfVectorizer( ) I have passed the pre_processor function as a parameter. I have passed another parameter of “Max-feature” this will only select top words which have higher tf-idf score.

Now, HotEncoding and TfidVectorize data is ready to merge, for merging we used hstack( ) our both vector data are merged row wise because we have the same number of row in both vectors but different number of columns.

B. Model selection

After deep Analysis, I have selected the Ridge Regression Model over Linear Regression model for predicting the quality of wine on given datasets. Because the number of independent variables are many in given data set, as well as we are not sure which of the independent variables influences dependent variable. In this kind of scenario, ridge regression plays a vital role than linear regression. Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable.

C. Hyperparameters tuning

We have used Ridge Regression, so for getting better result we change the values of some parameters to obtain the better result. We tune “Alpha” value in Ridge( ) function. We started putting values such as 0.06, 0.1 but when we are putting less values the result of Ridge Training and Testing score were less, so I decided to put higher values such as 0.8, 0.9 and I got very less score for Ridge Training and Testing because increasing “alpha” means pushing the Ridge Regression to be more robust against overfitting, but might be getting larger training error.

After changing the alpha values again and again Model got the right value which was alpha=0.201 in which score of Ridge Training and Testing was good.

Furthermore, I have applied the different values of "max_feature" in TfidfVectorizer( ) function such as top 100 or top 200. After doing tuning on max_feature parameter by applying different values, Model found the right value of max_feature which was 300 by changing the value into 300, our Ridge Training and Testing score got improved.

III. Results

After applying Ridge Regression model on our given data sets, the model predicted the values having good Ridge Training and Testing Score. For achieving this score, I applied some configuration such as pre-processing on data and then found out better score of Ridge regression.

```
Ridge_train_score
```

```
0.9492647504786756
```

```
Ridge_test_score
```

```
0.8137486601155097
```

References

[1] *nltk.tokenize package — NLTK 3.5 documentation*. (2016). Nltk. https://www.nltk.org/api/nltk.tokenize.html

[2] *sklearn.preprocessing.OneHotEncoder — scikit-learn 0.24.1 documentation*. (2018). OneHot Encoding. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[3] Q. (2019, January 3). *Ridge Regression for Better Usage - Towards Data Science*. Medium. https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db