Northeastern

Case Study 3 - One Dataset Many Users

Part 3 - Summary

INFO6105 37364 Data Science Engineering Methods and Tools

March 22, 2019

Project Group 2

Krina Devang Thakkar

Hardik Bhupendra Soni

Yash Thakur Lekhwani

Yashashri Shiral

Ritesh Manek

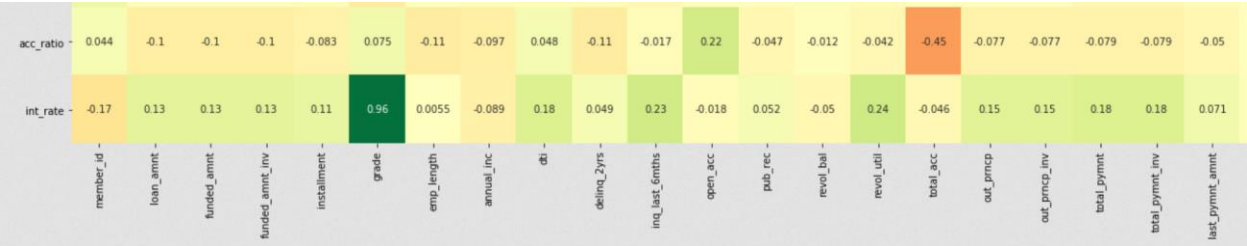
Design Specifications

1. Read the dataset and load it into dataframe. Further, split the dataset into Train and Test data in the ratio of 3 : 1.
2. Assign the Dependent and Independent features mentioned in the next section to the Train and Test data.
3. Create a function to predict the interest rate using each of the algorithms and calculate MAPE, Accuracy, RMS, MSE for them.
4. Call each of the algorithms, fit the variables in the model and predict the interest rate and accuracy for each of them.

Dependent and Independent Variables

Since we are helping Tola take better and accurate risks, we would want to predict the interest rates in the loans that he should invest. Thus, Interest Rate being our Dependent variable. Since Tola being a risk taker investor, we decided to use features that would increase the interest rate.

To identify the features that would contribute most to the quality of the resulting model we performed feature selection using **Correlation Matrix** on the cleaned and preprocessed data set. This helped us understand the relationship between multiple variables in the dataset. The following HeatMap showcases the same:



Have a look at the last row i.e. int_rate, see how the interest rate is correlated with other features, grade is the highly correlated with interest rate followed by debt to income ratio while some of the others are least related one. The following are the features selected using the Correlation Matrix:

- loan_amnt
- funded_amnt
- funded_amnt_inv
- installment
- grade
- dti
- inq_last_6mths
- revol_util
- out_prncp
- out_prncp_inv

- total_pymnt
- total_pymnt_inv
- purpose_debt_consolidation
- verification_status_Verified
- term_ 60 months

MAPE

The following are the MAPE score for each of the models:

Algorithm	Accuracy	MAPE
Linear Regression	Train: **94.73%** Test: **92.18%**	Train: **7.33** Test: **7.49**
Random Forest	Train: **99.91%** Test: **99:49%**	Train: **0.40** Test: **1.03**
Neural Network	Train: **99.46%** Test: **99.47%**	Train: **1.78** Test: **1.78**

Based on the MAPE score for each of the models, Random Forest and Neural Network has the highest and closest accuracy and with Random Forest having the lower MAPE score. Thus, **Random Forest** is the best model to predict the interest for Tola.

Model Performance Change with 5-Fold Cross Validation

Algorithms	Accuracy
Linear Regression	93.1%
Random Forest	99.02%
Neural Network	~99%

Using 5-fold cross validation, there is very less performance change in Random Forest and Neural Network, whereas linear regression has the highest difference.