Lending Club Data

Shubham Chaudhary, Michelle Baginski, Kirun Haque

Problem Statement

- Find the interest rate for the customers that will be taking out a loan
- Find out if the customer’s loan will likely be Fully Paid off or Charged Off
- We used Regression and Classification to solve the problem statements

What is our dataset?

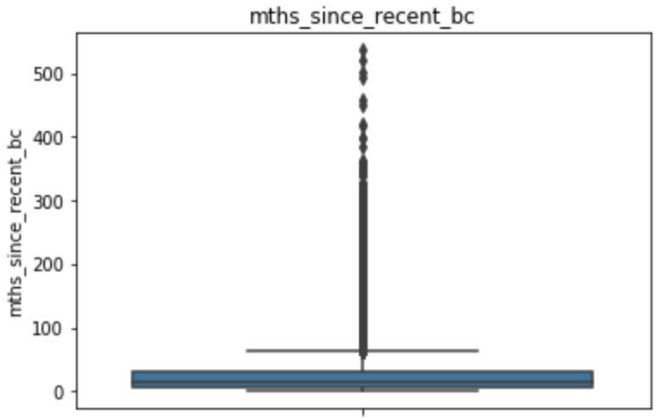
- Our data set contains loans from a 2 year span (2012-2013)
- It consists about 188,000 observations and 144 columns
  - It has variables like interest_rate, loan_status, and loant_amnt, etc.

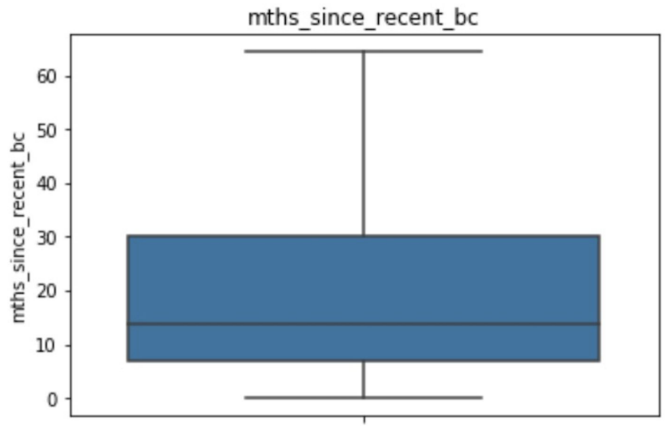
Data Preparation

- Removed variables with a single value
- Removed one of the variables with the correlation higher than 80%
- Variables that are missing with less than 20% of observations replaced with mean otherwise remove the variable
- Divided the data frame into 2 subsets
  - Numerical
  - Categorical

Data preparation

- Removed variables on the basis of p-value calculated using the linear regression method
- Used boxplots to analyze the dataset
  - What did we do to the outliers?





Data Preparation

- This is an imbalanced classification problem
- How did we make it balanced?
  - Used randomized undersampling to fix imbalanced dataset
    - Initially had 16% Charged Off and 84% of Fully Paid
    - Ended up with perfectly balanced dataset

Regression

- We performed 3 linear regression models with different variables
- First model with numerical data type variables and we got the R-Squared value to be 41%
- Second model with done with the complete dataset and we got the R-Squared value to be 99.4%
- The final model was done using specific variables where we got the accuracy to be 94.2%

Classification

- We made 3 classification models
  - Model 1 had 84.9% training accuracy, 77% test accuracy, 77% F1-Score,
  - Model 2 had 85.7% training accuracy, 77.7% test accuracy, 77% F1-Score
  - Model 3 had 88.21% training accuracy, 80% test accuracy, 80% F1-Score

Best Classification model

```
=== Confusion Matrix ===
[[6389 2567]
 [ 906 7941]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.88      0.71      0.79      8956
           1       0.76      0.90      0.82      8847

    accuracy                           0.80     17803
   macro avg       0.82      0.81      0.80     17803
weighted avg       0.82      0.80      0.80     17803



=== All AUC Scores ===
[0.88321193 0.876815   0.87533758 0.8775723  0.88519355 0.87789234
 0.88218542 0.88230857 0.89716215 0.88406711]


=== Mean AUC Score ===
Mean AUC Score - Random Forest:  0.8821745955069854
```

Results

- Our best linear regression model was able to define 94.3% of variance on the test dataset
- Our best classification model got 88.2% of the test accuracy with 80% F-1 score
- For regression, we selected 11 variables out of 144 variables to get the optimal accuracy
- For classification, we selected 14 variables out of 144 variables to get the optimal accuracy

CONCLUSION

- We concluded loan amount, loan status,home ownership, purpose of the loan, inquiries in the last 6 months, number of bankcard accounts, and number of accounts opened in the last 12 months were the most important variables to calculate interest rate and the defaulters