Credit Risk Modeling: Lending Club Dataset

Sijo Valayakkad Manikandan

westcreek™

Problem & Assumptions

Problem Definition:
Lending Club is a peer-peer lending company. In this data challenge, I am trying to assess the risk of loans based on historical data available.
Dataset:
Around 2.26 million loans and 145 features → 0.92 million current loans
Approach: Since we do not know the outcome of the Current loans, include such loans in the scoring/unseen data. Use the rest of the data for training and validation purposes. I have converted the problem to a binary classification problem by grouping performed as below

Bad Loan

- Charged Of'
- Does not meet the credit policy. Status:Charged Off
- Default

Good Loan

- Fully Paid,
- 'Late (31-120 days)
- In Grace Period
- Late (16-30 days)
- Does not meet the credit policy. Status:Fully Paid

→

Binary target:
Bad_loan(yes or no)

Exploratory Analysis & Feature Engineering

Figure 1: Distributions of loan, funded, and investor amounts

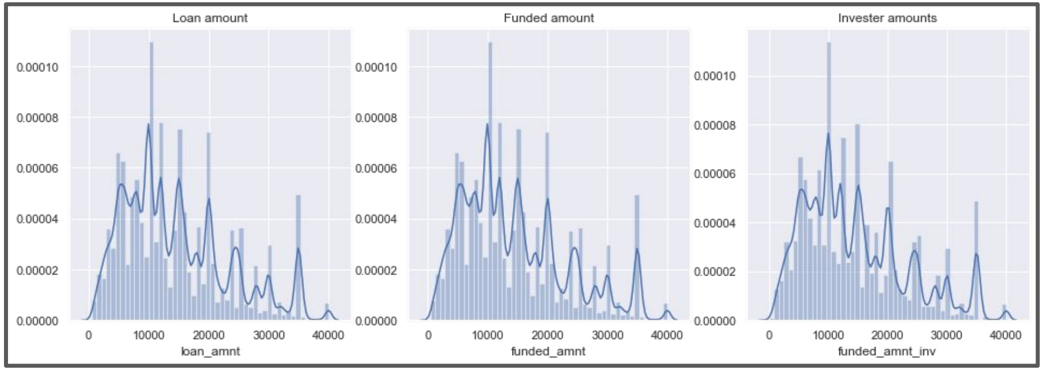


Figure 2: Grade vs bad loan proportion

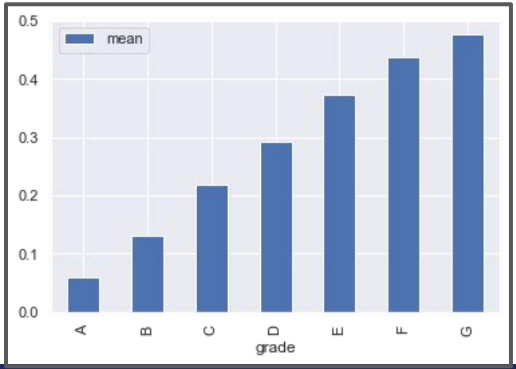
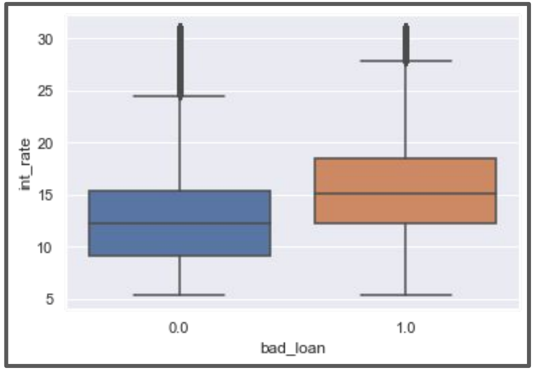


Figure 3: Interest Rate vs bad loan proportion



Initial Observations:

- High imbalance in the outcomes.
- Since I converted it into a binary problems and since I removed the current loans, around 19.5% of the loans (out of 1340973) are bad loans
- The distribution of funded amount, loan and investor amount follows similar distribution (good for business)
- High correlation among certain features
- Lending club assigned grade captures risk to a good extent (grade G & F has over 40% bad loan rate
- Certain features are a result of bad loan! For instance, debt_settlement_flag or collection_recovery_fee etc.
- Generally, higher interest rates tends to have more bad loans

Feature Selection and Engineering

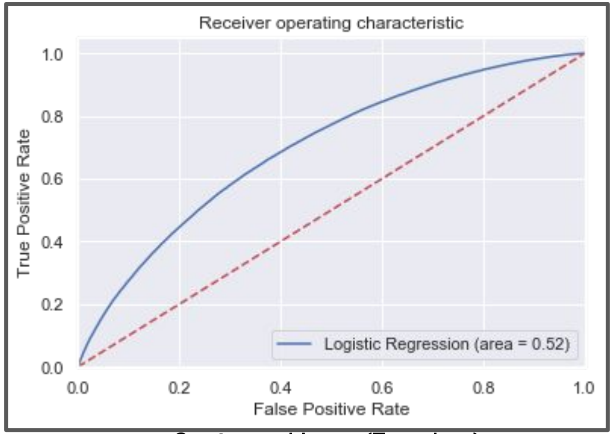
- **Using Variable Understanding & Exploratory Analysis:** Certain features are the result of bad or good loan. For instance, collection related features - collection recovery fees, recovery fees etc.
- **Zero Variance Removal:** Columns with only 1 unique value were removed
- **High missing values:** Variables with more than 50% of the observations missing were ignored - If time permits, we might be able to do smart imputation techniques. For now, any imputation would add a lot of bias. So removed those observations
- **Heavily correlated features:** Features with more than 0.95 correlation. For instance,
- **Model based feature selection using LASSO:** Built an initial Logistic Regression model using L1 regularization (used Cross validation to find the reg. parameter) and removed the coefficients with zero coefficient value

- **One-Hot Encoding of Categorical Variables**
- **Features created:**
  - issue_earliest_diff - number of days between loan issue date and first credit line created
  - Converted values Y, N to 1 or 0.
- **Missing value imputation with median** (can be improved) - used median to avoid the outlier effects

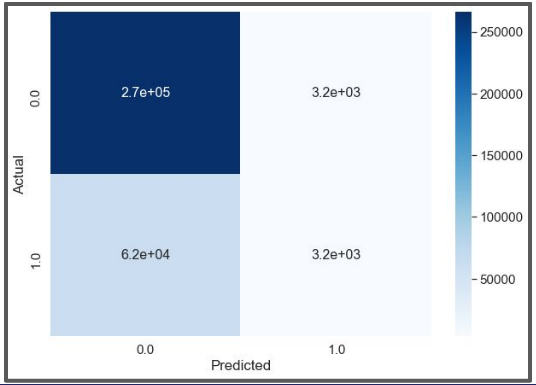
Base Model

- **Train - test split:** 80 - 20 stratified split to maintain the class proportion in both train and test
- **Base model:** Logistic Regression
- **Cross validation:** 5-fold
- **Hyperparameter tuning results:** C or 1/lambda = 0.001 or Lamba = 1000 with L1 regularization
- **Reducing False Negatives or Improving recall as the goal. Metric = Recall**

	precision	recall	f1-score	support
0.0	0.81	0.99	0.89	269632
1.0	0.50	0.05	0.09	65612
accuracy			0.80	335244
macro avg	0.66	0.52	0.49	335244
weighted avg	0.75	0.80	0.73	335244



Confusion Matrix (Test data)



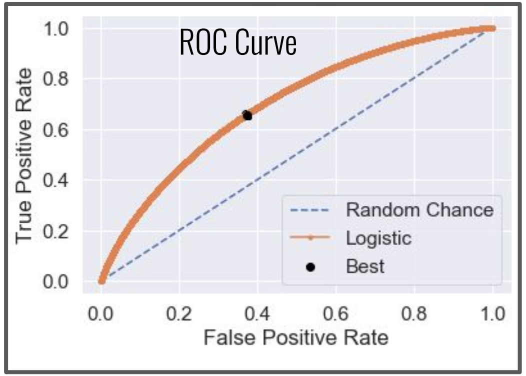
- Poor recall of the bad loans
- AUC of 0.52
- False Neg. Rate: 18.6% False Pos. Rate: 0.9%

Final Model results

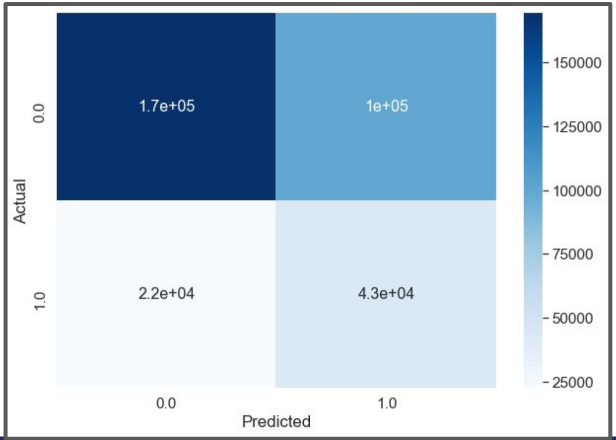
Actions taken:

- **Feature Selection:** Remove the coefficients with zero coefficient values
- **Threshold Selection:** Find the best point in the ROC curve and find the top-left most point (ROC curve on the right)

	precision	recall	f1-score	support
0.0	0.88	0.63	0.73	269632
1.0	0.30	0.66	0.41	65612
accuracy			0.63	335244
macro avg	0.59	0.64	0.57	335244
weighted avg	0.77	0.63	0.67	335244



Confusion Matrix (Test data)



- AUC of 0.643 False Neg. Rate 6.7%, False Positive Rate: 29.8%
- 23.6% improvement in AUC
- 64% improvement in False Neg. Rate
- Will be able to capture 66% of the bad loans

Insights and future work

Insights from the model (a few) :

- 1 unit increase in the interest rate increases the odds of bad loan by around 8%
- Earlier the credit line was issued, the lesser the odds of bad loan
- Initial Status of the loan -> If it is whole, then there is a higher odds of it being a bad loan by around 10%

Future Work:

- PCA to reduce dimensions since there are a lot of correlated variables and it would be useful to interpret the linear combination of those variables as well as use the Principal components in the model
- Outlier Detection
- Better Missing value imputation
- Better Feature Engineering:
    - For instance, address state can be grouped in different regions or pick highly vulnerable regions
    - Get seasonal effects (month of issual of loan)
    - Get zip code related info from open source datasets
- Ensemble multiple models: can build other models such as Random Forest, MLP classifier, Naive Bayes, KNN and ensemble the results
- Try undersampling, oversampling and SMOTE to resample the dataset to account for imbalanced class
- Perform the same feature transformations on the scoring data and predict. Use the prediction to take actions.

Thank you