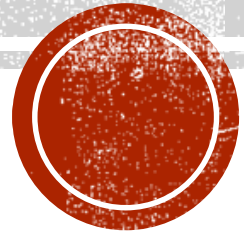
LOAN REPAYMENT ANALYSIS

For Lending Club and Investors



Ting Sit

July 2019

LENDING CLUB BUSINESS MODEL

- Unsecured personal loans ($1K - $40K)
- Investor fund loan request, obtain return through interests
- Lend Club revenue = origination fees (borrowers) + service fees (investor)

OBJECTIVE

- Develop models that help lending club to screen out bad loan's requests
- Benefit:
  - To Lending Club: Increase investors confidence on investing in Lending Club's loans → more fees → more revenues
  - To Investors: Minimize chances of loan default, which would be equivalent to investment loss (since loans are not FDIC insured)

Model Evaluation requirements:

- Balanced performance among precision and recall.
- Metrics: F1 score and misclassification rate

MODEL CONSIDERATION

- Classification problem: default (target = 0) and paid off (target =1)
- 2 models: Logistic Regression and Random Forest
- Features: numeric variables
  - years of credit
  - debt to income ratio (DTI),
  - total credit lines

RESULTS

Model	F1 score (cross validated)	misclassification rate
Baseline	0.917	.15
Logistic Regression	0.670	.46
Random Forest	0.883	.21

- Random Forest has higher f1 score which meet the evaluation criteria of balanced performance
  - Recommend Random Forest as the winning model.

CONCLUSION

1. Random Forest has balanced performance (high F1 score).
2. The model enables Lending Club to filter out some bad loan requests which reduce risk of profit loss to investors due to loan defaults. However, the model also reject loan requests that would pay off, which become lost revenue/investment opportunity for Lending Club and investors. Lending Club would need to weigh between the two trade offs and define the threshold tolerance on the potential misclassification.
3. From Data Exploratory Analysis, loan purpose is likely to impact loan paid off rate. A lot of applicants work in the military.

NEXT STEPS

- Introduce categorical features into the model to seek further improvement on F1 score.
- Cross-validation should be done in the form of Time Series Nested Cross-Validation to avoid contamination of time component on predicting results.
- Access 1-2 more classification models to have a broader measurement for model performance. I would consider Extreme Gradient Boosting and Support Vector Machines, for example.
- Features engineering which includes clustering to capture potential patterns that were not captured by looking at one feature alone.
- Run correlation test across features to access potential collinearity.