Classifying Borrowers who are Likely to Declare Default

Kyle Kim, Department of Statistics, University of Virginia

Introduction

- The main source of banks' profits is the net interest margin.
- However, loans are risky as borrowers may default on the loans.
- The 2008 Financial Crisis resulted from massive chains of defaults, started by Bear Stearns and followed by Lehman Brothers and many other institutions.
- The annual market size of auto loans is around $700 Billion, smaller than that of mortgages, but large enough to cause economic turmoil if massive default occurs.

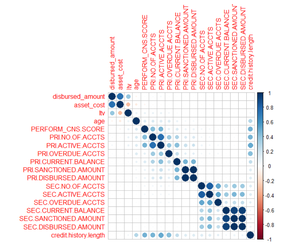
Objective

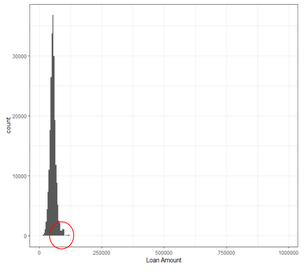
1. Identify key variables that are significant in predicting the probability of declaring default.
2. Compare different models that predict default and identify the best performing model.
3. Improve the best-performing model and discover relationship between variables and declare a default.

Data & EDA

•Downloaded from Kaggle but the original source is not released.
•225,000 observations and 41 variables.
•The main 3 categories of variables are borrowers' demographic, loan information, and Credit Bureau data including credit history.
•The data is slightly unbalanced. About 78.3% of observations did not declare a default. However, such unbalance is the norm in the industry and thus not corrected.

ACTIVE.ACCTS: Number of active loans
Loan Amount: Amount of the vehicle loan
DISBURSED.AMOUNT: Amount of loan prior to the current vehicle loan
Employment.Type: Type of Employment [Salaried or Self Employed]
Ltv: Loan to Value of the Asset
NO.OF.ACCTS: Total number of loans at the time of disbursement
OVERDUE.ACCTS: Number of Default accounts at the time of disbursement.
PERFORM_CNS.SCORE: Credit Bureau Score
SANCTIONED.AMOUNT: Amount of loan available to a borrower
*All numeric variables are standardized.

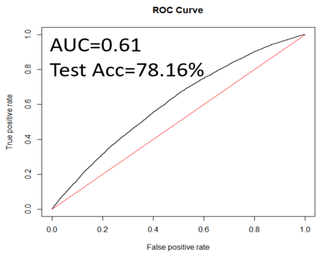


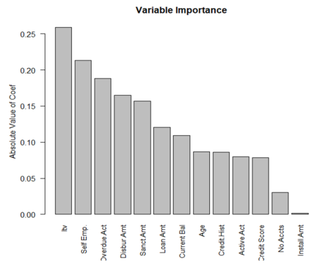


- Many categorical variables that describe borrowers' demographic are removed due to extreme unbalance (95:5).
- Multicollinearity issue fixed by adding variables based on the data dictionary.
  Ex: Primary and secondary loan data are added together as they represent borrowers' liabilities.
- Skewness: Many numeric variables have some skewness. Such behaviors are common in the financial data.

Logistic Regression

- The initial analysis started with logistic regression, mainly for its interpretability.
- Linearity couldn't be verified due to the presence of discrete variables .





Likelihood Ratio Test P-val = 0
Based on the absolute value of coefficient estimates, the top 3 most dominant variables are:
1. Ltv 2. self employment 3. overdue accounts.
   With a logistic model with ltv, employment type and overdue accounts, test accuracy slightly improved to 78.19%. Likelihood Ratio Tests p-val = 0.

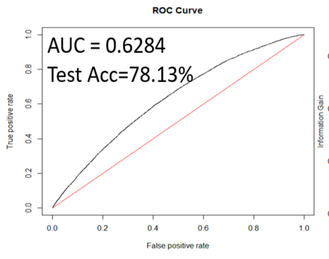
Model Improvement: Higher Order Terms & Higher Probability Cutoff
The standard probability cutoff for classifying an observation is 50% .

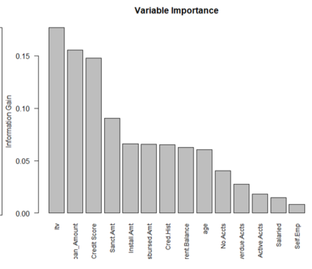
- Can we improve accuracy if the model only classifies borrowers as default at a value higher than 50% since defaults are not common in real life?
- What if Ltv is more emphasized by generating higher order values like ltv^2?

Accuracy, if classified as default when probability > 80%:	78.20%
Accuracy, with ltv^2:	78.16%
Accuracy, with ltv^2 and, at probability > 80%:	78.16%

No significant improvements are made.

XGboost





- Xgboost is selected because its performance is well known, has a short computation time, and can control overfitting with regularization parameters.
- AUC and Test Accuracy barely changed despite the increase in model complexity with Xgboost.
- Tuned parameters: Eta, max_depth, colsample_bytree,subsample

Based on the information gain, the top 3 most dominant variables are:
1. Ltv 2. Loan Amount 3. Credit Score

Model Improvement: Higher Order Terms & Higher Probability Cutoff

Accuracy, if classified as default when probability > 80%:	78.21%
Accuracy, with ltv^2:	78.16%
Accuracy, with ltv^2 and, at probability > 80%:	78.21%

No significant improvements are made.

Other models reviewed:
1. LDA 2. QDA 3. K-Means Clustering
- LDA had similar performance as logistic regression and Xgboost.
- QDA slightly underperformed.
- Not included because of similar performance and violation of multivariate normality assumption.

Conclusion

Bias and Variance Tradeoff:
- Logistic Regression & Xgboost have similar performance on the test data. This indicates bias exists because Xgboost has a lower bias but higher variance and vice versa for Logistic Regression.
- Since the performances are similar, linearity between the likelihood of declaring default and variables(logistic assumption) might exist. If linearity did not exist, Xgboost should have performed better than the Logistic Regression. Therefore, probabilistic results from the logistic regression can be accepted.

Conclusion:
- Accuracy ceiling is 78% & AUC = 0.6283 is the highest.
- With current data, Ltv is the most important variable that explains the probability of declaring default.
- False Negative Rate remained the same throughout changes in the threshold values.
- Model improvement with variable selection or adding higher orders did not make any improvements.

Since the data is about loans, loss-given default (LTD) is evaluated. LGD is the total amount banks lose due to borrowers' defaults. Few assumptions are made for calculation.

Logistic Regression:	$-552 Mil or 22.4% of total loan
Xgboost:	$-523 Mil or 21.2% of total loan

Based on LGD, Xgboost performed better than the logistic regression model.

Discussion & Future Work

Discussion:
- I think bias comes from the fact that defaults are not declared overnight. It occurs as borrowers' financial condition worsens along with the economic cycle, which is not covered in this dataset.

Future Work:
- Collect more variables to reduce bias.
- When the bias is controlled at a reasonable level, this project can be developed into a multiclass classification project that measures borrowers' risk like Moody's credit rating.
- Investigating the relationship between Ltv and default can deepen understanding of the vehicle loan industry.

Reference
1. *Pesantez-Narvaez, Jessica, et al. Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. 2019.*

I would like to acknowledge Dr. Li and Dr. Yu for their guidance.