Machine Learning in Reviewing Loan Applications

Yanchen Chen

Objective

Investigate financial meaningful and statistical significant variables that are determinant in loan default likelihood prediction using logistic regression, decision tree, and XGBoost

Expectation

Optimizing predictive reliability

using low computational cost model with public financial data

Baseline Model

Our Baseline

Logistics regression using public data

Industry	Our Model
High Accuracy Computational Expensive Professional Experience required	Easy Implementation Reasonable Computational Cost Feasible Interpretation Accessible Data

Statistical Prespective

Model	Logistic (Baseline)	Logistic (Threshold)	Decision Tree	XGBoost
Advantage		Determining Probability	Meaningful Interpretation	More Efficiency
Recall	0.6403	0.9658	0.7819	0.8279
Accuracy	0.6454	0.6003	0.6542	0.6061

Business Prespective

Target Client

- Conservative Client: Model with high recall
- Aggressive Client: Model with high accuracy = higher coverage of clients

Reasonable Computational Cost Model

- Do not require advance coding skill
- Do not require professional understanding of data science
- Provide critical financial meaningful variables
    - Interest Rate
    - Annual Income
    - Loan Issue Date
    - Revolving Line Utilization Rate
    - Monthly Total Debt to Total Debt Obligation Ratio (DTI)
    - Total Credit Revolving Balance & Period Between Earliest Credit Line Date

Data Processing

- Variable Overview

Dependent Variable	Description
loan_status	Current status of the loan

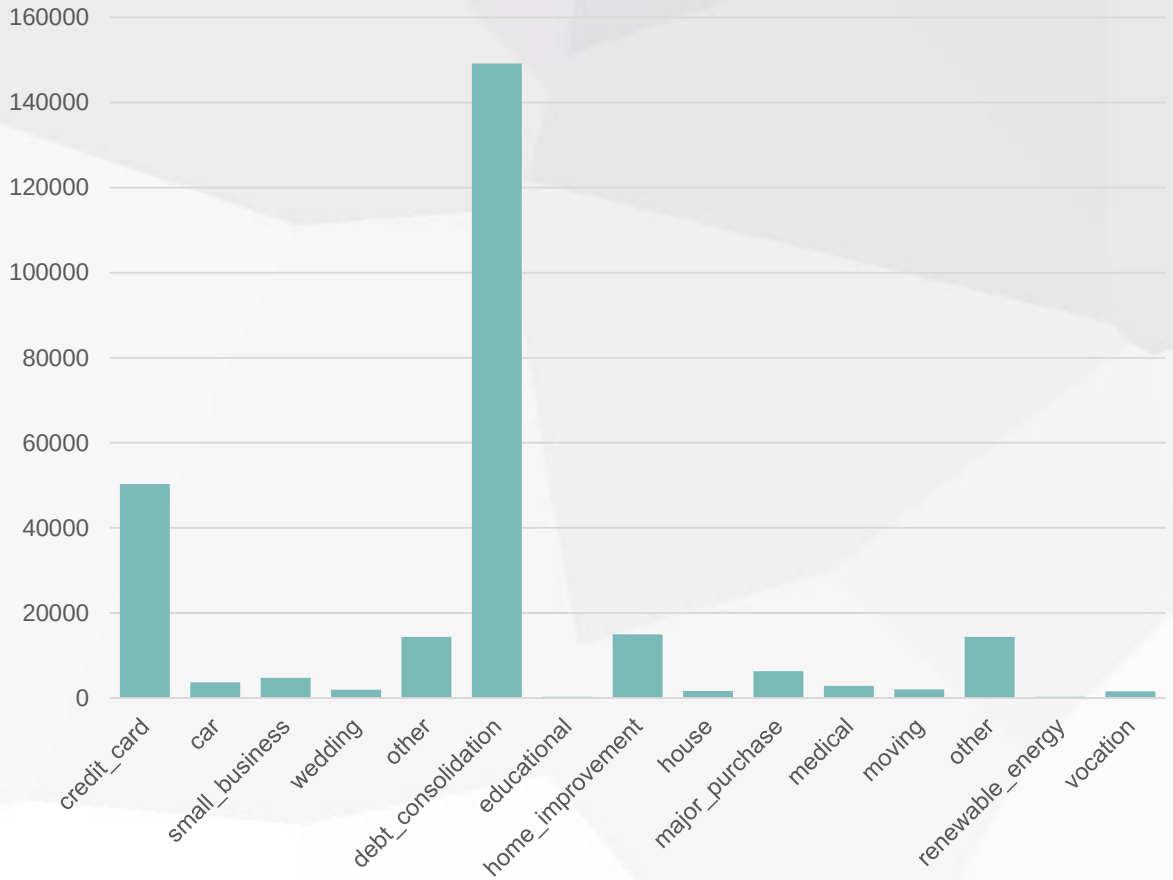
Integer predictor	Description
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
emp_length	Employment length in years
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
loan_amnt	Amount of the loan applied by the borrower
open_acc	The number of open credit lines in the borrower's credit file
pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance
term	The number of payments on the loan
total_acc	The total number of credit lines currently in the borrower's credit file
credit_month	The time interval between the loan was funded and the borrower's earliest reported credit line was opened(round)
desc_len	the length of loan description provided by the borrower(in words)

Factor predictor	Description
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
home_ownership	The borrower's home ownership status. (Rent, Own, Mortgage, Other)
verification_status	Indicates if income was verified or not. (Landing Club Verified, Verified Income Source, Not Verified)
purpose	A category provided by the borrower for the loan request.

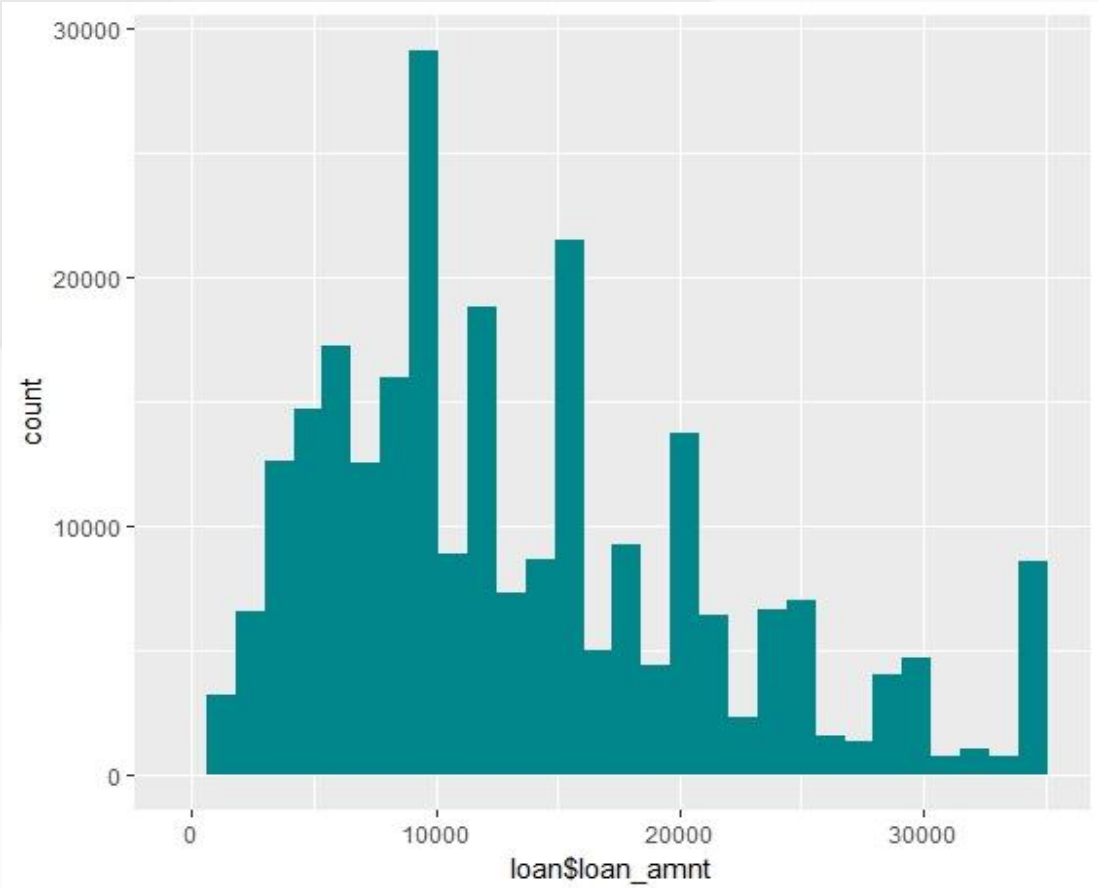
Numerical predictor	Description
annual_inc	The self-reported annual income provided by the borrower during registration
dti	Monthly debt payments on the debt obligations, divided by monthly income.
Interest rate	Interest Rate on the loan
revol_util	The amount of credit the borrower is using relative to all available revolving credit
tot_coll_amt	Total collection amounts ever owed
Total current balance	Total current balance of all accounts
mths_since_last_delin	The number of months since the borrower's last delinquency
mth_since_last_derog	Months since most recent 90-day or worse rating
mth_since_last_record	The number of months since the last public record

EDA - Variable Overview

Loan Purpose



Loan Amount



Data Understanding - Dataset Quality

- Completeness
  - 887,379 lines of loan records and 74 variables
- Validity & Accuracy
  - From the lending company called Lending Club
- Availability & Reliability
  - Online open source with reliable reputation

Data Cleaning

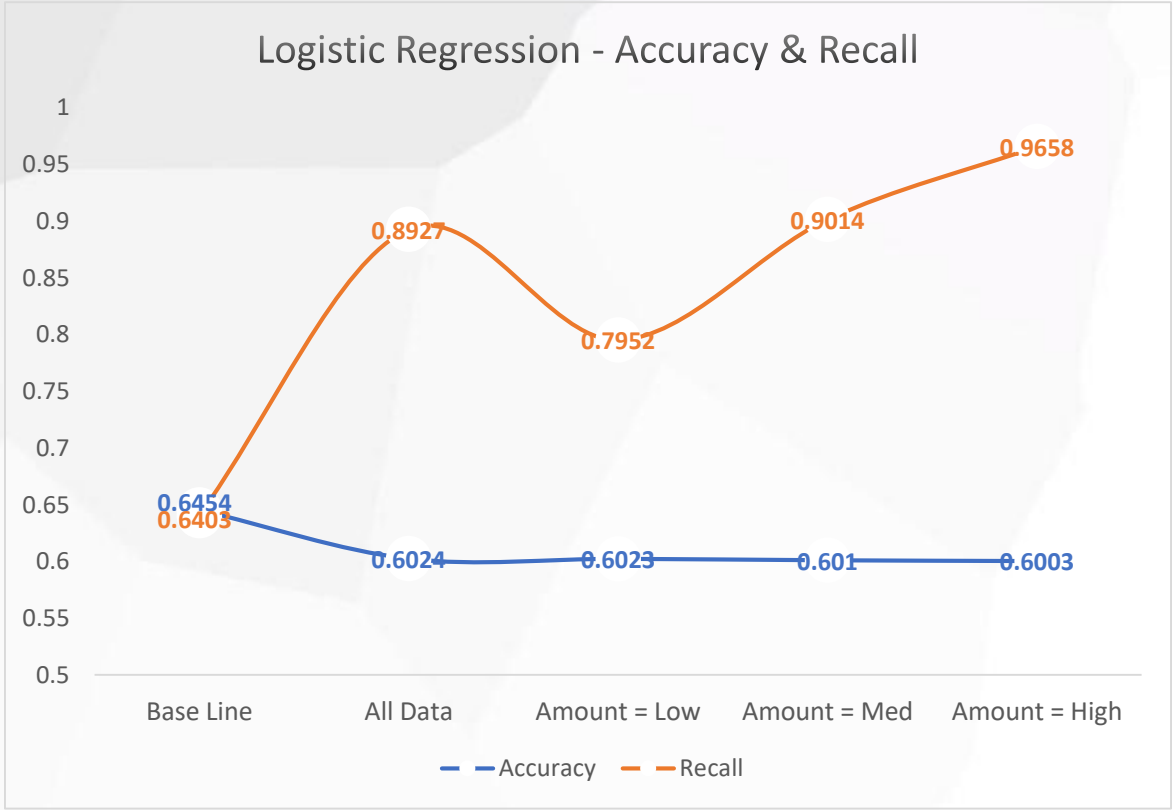
- Data preparation Tech & Other Key Task

- Remove irrelevant variables according to financial definition
- Response (Loan status)
  - Default & Charge off = 1
  - Fully Paid = 0
- NA value - reassign as Infinity, mean, specified value
- Outlier - none
- Character - transform into factor levels
- Date - transform into time
- Divided into three subset - low, medium, high loan amount

Logistic regression - threshold & loan amount

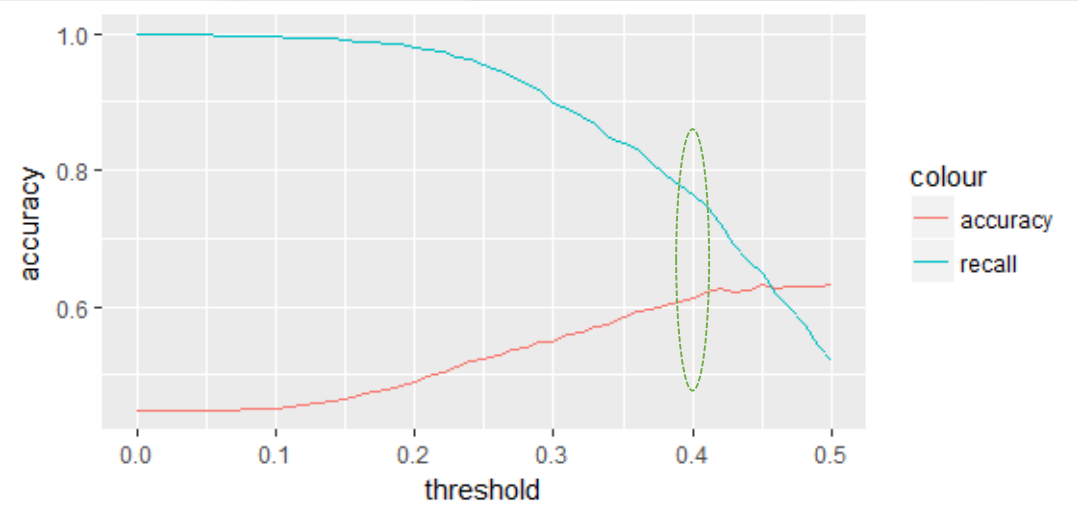
	Base Line	All Loan Amount	Low Loan Amount	Med Loan Amount	High Loan Amount
Accuracy	0.6454	0.6024	0.6023	0.6010	0.6003
Recall	0.6403	0.8927	0.7952	0.9014	0.9658
Threshold	0.5	0.34	0.38	0.34	0.29

- Value-add
  - By choosing different thresholds, recall increases up to 0.32.
  - By dividing data by loan amount, accuracy/recall increases.

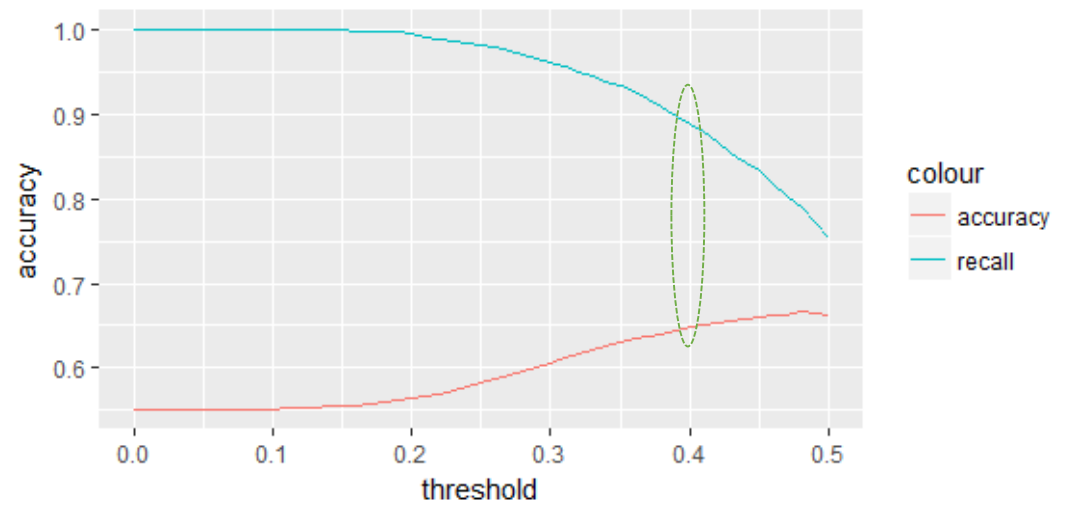


Logistic regression - Accuracy & recall trade off

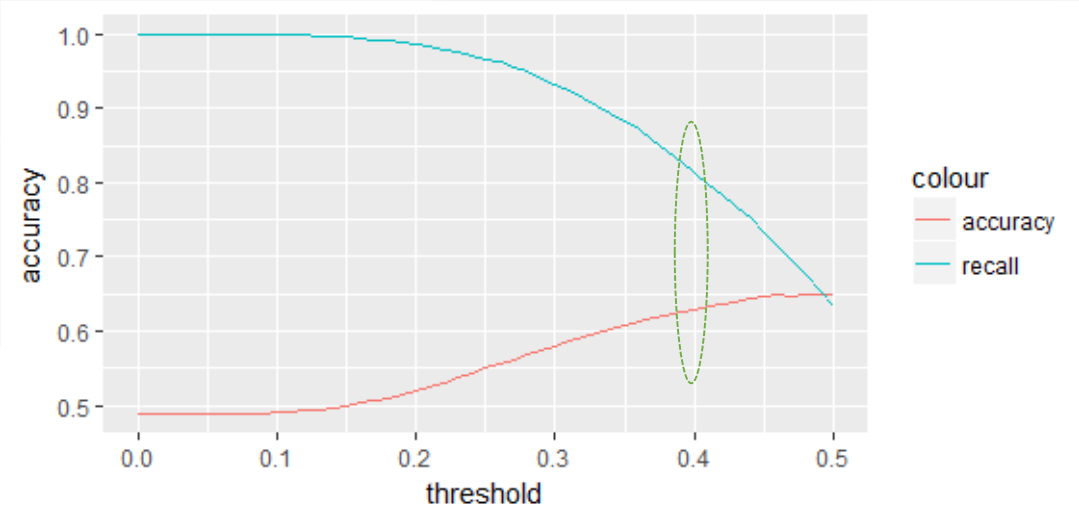
Loan Amount = Low



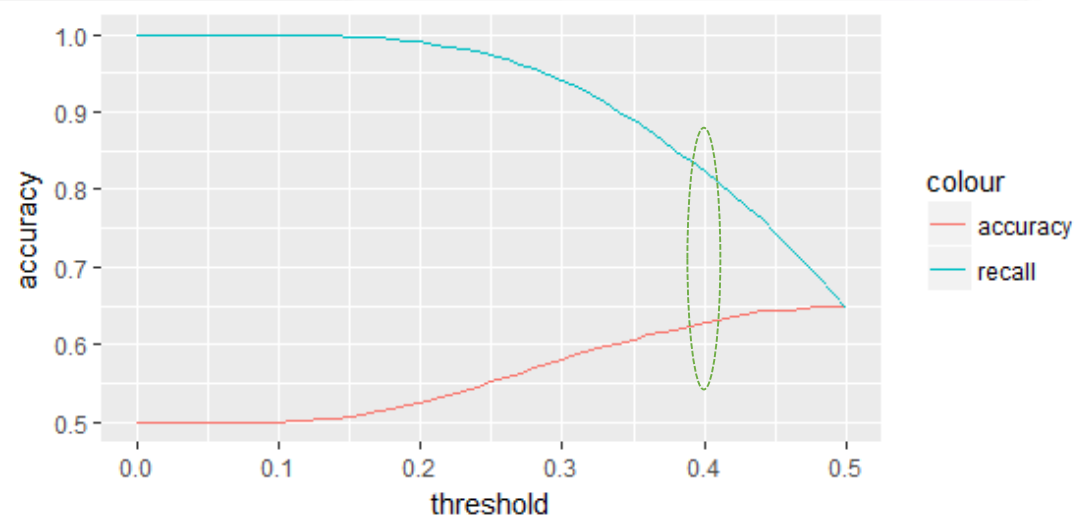
Loan Amount = High



Loan Amount = Med



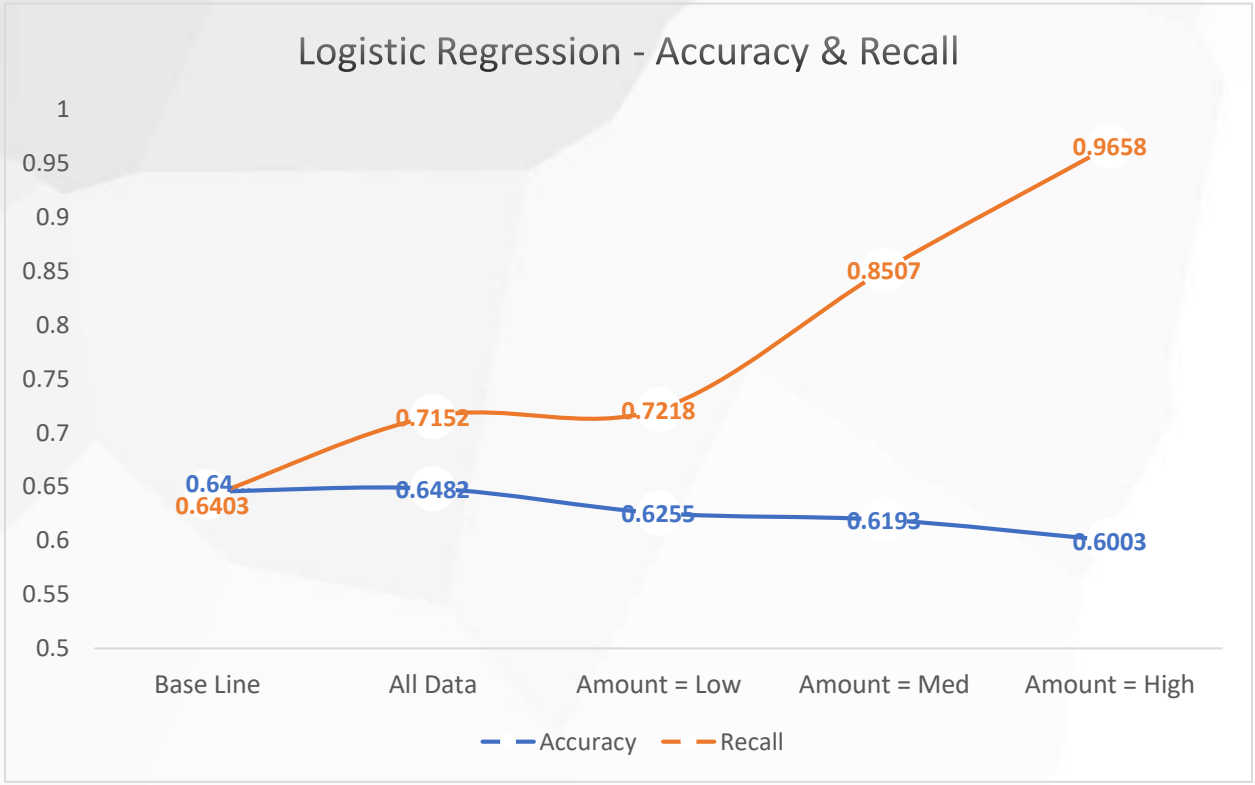
All Data



Logistic regression - loan amount

	Base Line	All Loan Amount	Low Loan Amount	Med Loan Amount	High Loan Amount
Accuracy	0.6454	0.6482	0.6255	0.6193	0.6003
Recall	0.6403	0.7152	0.7218	0.8507	0.9658
Threshold	0.5	0.46	0.42	0.38	0.29

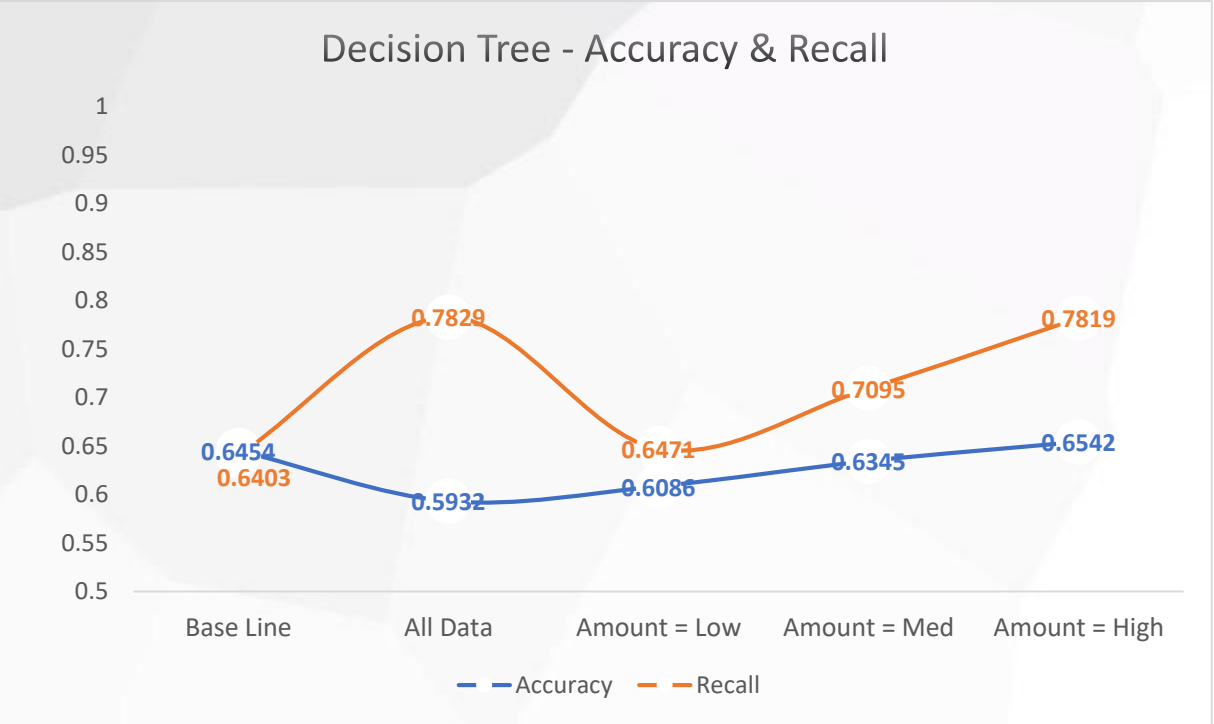
- Extra advantage
  - Flexible choices can be made according to different risk attitudes / strategies



Classification Tree

	Base Line	All Loan Amount	Low Loan Amount	Med Loan Amount	High Loan Amount
Accuracy	0.6454	0.5932	0.6086	0.6345	0.6542
Recall	0.6403	0.7829	0.6471	0.7095	0.7819

- Value-add
  - Increase recall by 0.14 only sacrificing accuracy by 0.05
  - Easy to visualize graphically
  - Closer to human decision-making

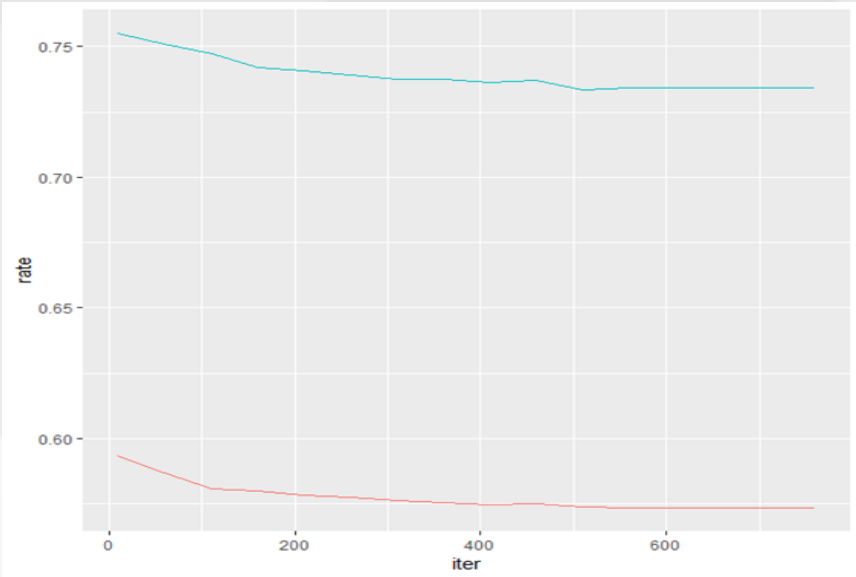


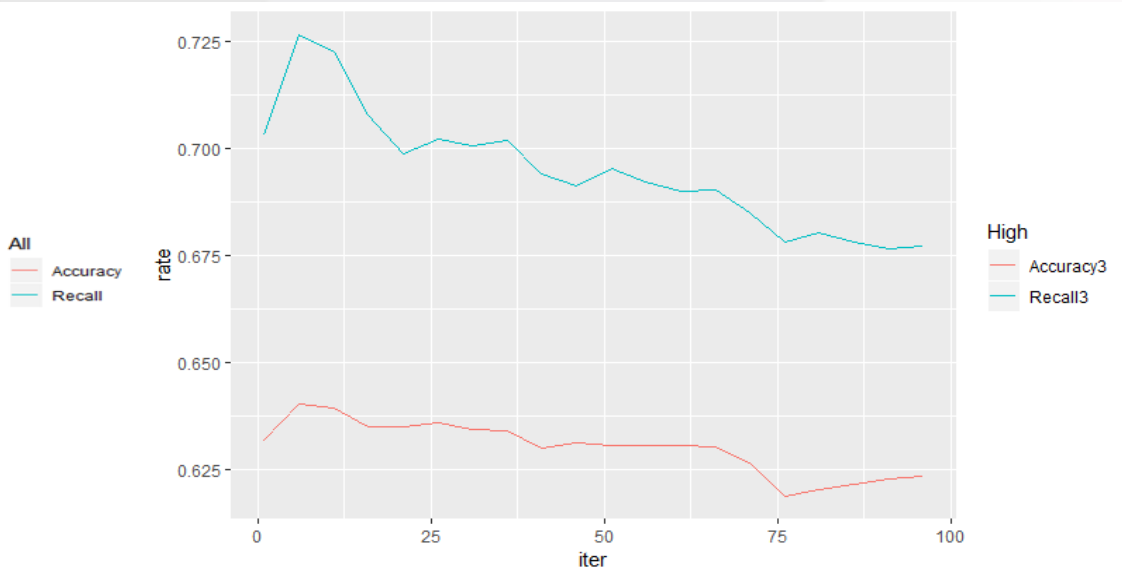
XGBoost

	All Loan Amount	Low Loan Amount	Med Loan Amount	High Loan Amount
Accuracy	0.5735	0.5934	0.6186	0.6234
Recall	0.7341	0.4815	0.6161	0.6771

- After splitting the dataset based on loan amount, accuracy increases as a trade off as recall
- Target client: risk conservative companies

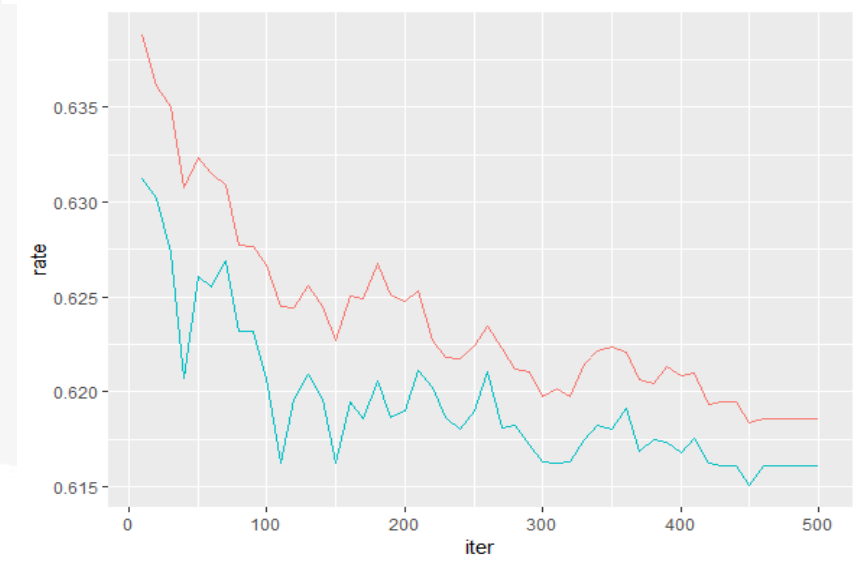
XGBoost

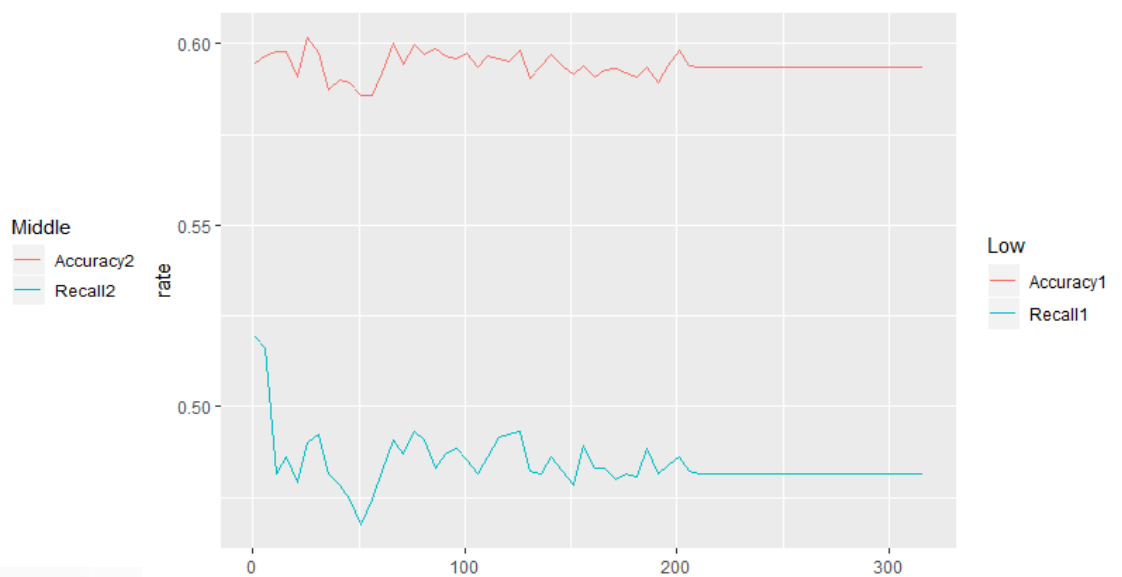

0.75
0.70
0.65
0.60
rate
0
200
400
600
iter


All
Accuracy
Recall

0.725
0.700
0.675
0.650
0.625
rate
0
25
50
75
100
iter

High
Accuracy3
Recall3


0.635
0.630
0.625
0.620
0.615
rate
0
100
200
300
400
500
iter


Middle
Accuracy2
Recall2

0.60
0.55
0.50
rate
0
100
200
300
iter

Low
Accuracy1
Recall1

Variable Selection

Business understanding

- Irrelevant variables
- Pre-approval variables only
- Financially intersect/overlap variables

Statistical meaning

- Majority with missing value(NA)
- Least importance

Workflow - Logistic Regression

The beginning

- Load training and out-of-sample datasets

Fit the logistic regression

- Use backwards stepwise for variable selection
- Set threshold value from 0 to 0.5, step by 0.01

Result

- Make prediction
- Calculate the accuracy rate and recall from the confusion matrix
- Plot accuracy and recall against different thresholds

Workflow - Decision Tree

The beginning

Load training and out-of-sample datasets

Grow the Tree

- Method as ‘class’
- Control with minsplit 20 and maxdepth 8
- Printcp display cp table
- Plotcp plot cross-validation results
- Select best cp value and grow the tree again

Result

- Make prediction using the new tree model
- Calculate the accuracy rate and recall from the confusion matrix

Workflow - XGBoost

The beginning

Load training and out-of-sample datasets

XGBoost Algorithm

- Create a for loop iteration from 1 to 400+
- Implement XGBboost and ConfusionMatrix function

Result

- Make prediction
- Calculate the accuracy rate and recall from the confusion matrix
- Comparison plot

THANKS!

Q&A