Amex Express-Credit Card Default Prediction

Mini Project 2 :Classification Models Comparison

By Clivia Kong

Business Context

American Express

Credit default prediction is central to managing risk in a consumer lending business.

Lending Decisions

Optimizing Lending Decisions, offering better customer experiences.

Risk Managements

Managing customer default risk

Approval Criteria

Improving application criteria

Business Question

Will the Amex Credit Card Customer be default or not?

Default will be regarded if customer not paying required amount in 120 days(Supervised Binary Classification)

Data Pipeline

Data Source

Datasets from Amex Kaggle Competition

EDA

Data Overview

Modeling

Preprocessing & Modeling

Comparison

Random Forest, Light Gradient Boosting, Logistic Regression & Gaussian Naïve Bayes

Summary

Improvements

EDA

Data Overview

Dataset: Profile features(189) for each customer at each statement date(5.5 million records)
Key features: Five types of features(Incl.11 Categorical features), Customer ID, Target

Delinquency Features

- Falling behind on required monthly payments to credit card companies.
- 96 features(9 categorical features)

D

S

Spend Features

- Customers' spending variables
- 22 features (1 time feature)

Payment Features

- Customers' payment behaviors.
- 3 features

P

B

Balance Features

- 40 features(2 categorical features)

R

Risk Features

- Variables regarding risk
- 28 features

EDA

Data Overview

Imbalanced Data

Target:

- Default:25.89%
- Not Default: 74.11%

01

Customer Records

Maximum 13 variables

Minimum 1 variables

02

Missing Values

- 67 features of total 189.
- 8 features have 90% missing values

03

Feature Correlation

29 features have over 90% correlation with each other

04

Preprocess

- Spend Features
- Delinquency Variables
- Balanced Features
- Risk Features
- Payment Features

- Categorical features

Label Encoder

- Fill Missing values

Forward fill by time-based Spend features

- Feature Selection

Drop columns with over 90% missing values

Drop columns with over 90% correlation

- Define Parameters Range
- Randomized Search
- Fit best score model parameters
- Comparison

Data Source

5.5 million records * 190 features

EDA

Multiple customer records

Preprocess

One customer record, 4.6 million records * 180 features

Modeling

Random Forest, Light Gradient Boosting & Logistic Regression

Summary

Model Results

- Random Forest
- Light GBM
- Logistic Regression
- Gaussian Naïve Bayes

Random Forest Model

Train & Test Result

Parameter Tuning Range

- Max_features： auto, sqrt, log2
- Max_depth:2 to 130
- Min_samples_leaf: 2 to 55
- Min_samples_split: 2 to 55
- Bootstrap: yes or no

Best Parameter

- Max_features： auto
- Max_depth:22
- Min_samples_leaf:16
- Min_samples_split: 14
- Bootstrap: no

Score Result (Test)

- Accuracy : 0.8930
- Precision : 0.7967
- Recall : 0.7837
- ROC AUC : 0.9533

Light Gradient Boosting Machine

Train & Test Result

Parameter Tuning Range

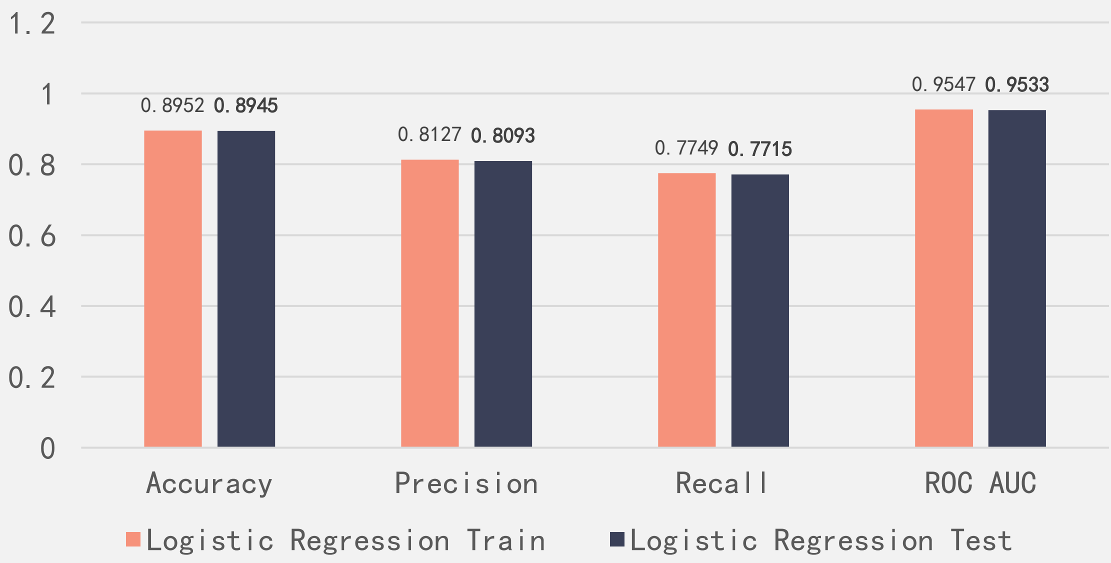
- Num_leaves: 100 to 500
- Max_depth:3 to 20
- Min_data_in_leaf: 10 to 1000
- Learning_rate: 0.01 to 0.31
- Lambda_l1: 0 to 100
- Lambda_l2: 0 to 100
- Bagging_fraction: 0.1 to 0.9

Best Parameter

- Num_leaves: 150
- Max_depth:11
- Min_data_in_leaf: 160
- Learning_rate: 0.05
- Lambda_l1: 20
- Lambda_l2: 0
- Bagging_fraction: 0.9

Logistic Regression Model

Train & Test Result



Parameter Tuning Range

- Max_features： auto, sqrt, log2
- Max_depth:2 to 130
- Min_samples_leaf: 2 to 55
- Min_samples_split: 2 to 55
- Bootstrap: yes or no

Best Parameter

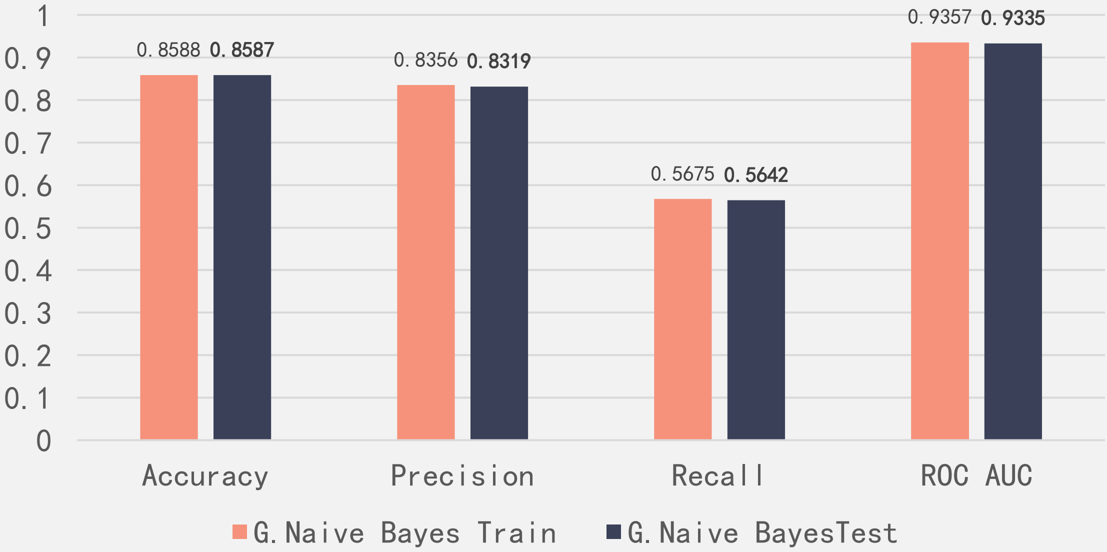
- Max_features： auto
- Max_depth:22
- Min_samples_leaf:16
- Min_samples_split: 14
- Bootstrap: no

Score Result (Test)

- Accuracy : 0.8930
- Precision : 0.7967
- Recall : 0.7837
- ROC AUC : 0.9533

Gaussian Naïve Bayes

Train & Test Result



Parameter Tuning Range

- Var_smoothing: 1.0e-100 to 1

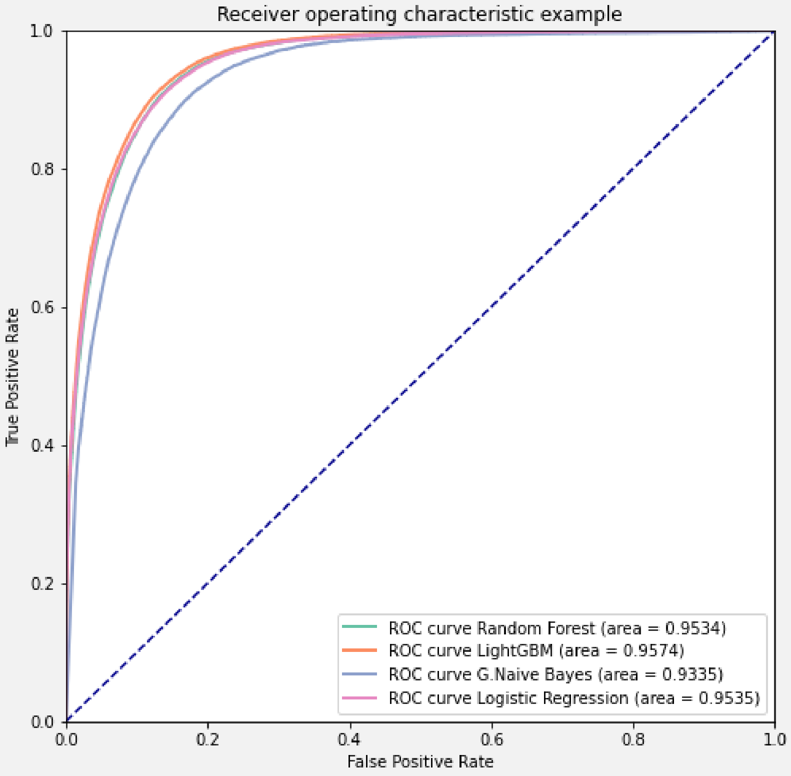
Best Parameter

- Var_smoothing: 8.497534359086438e-08

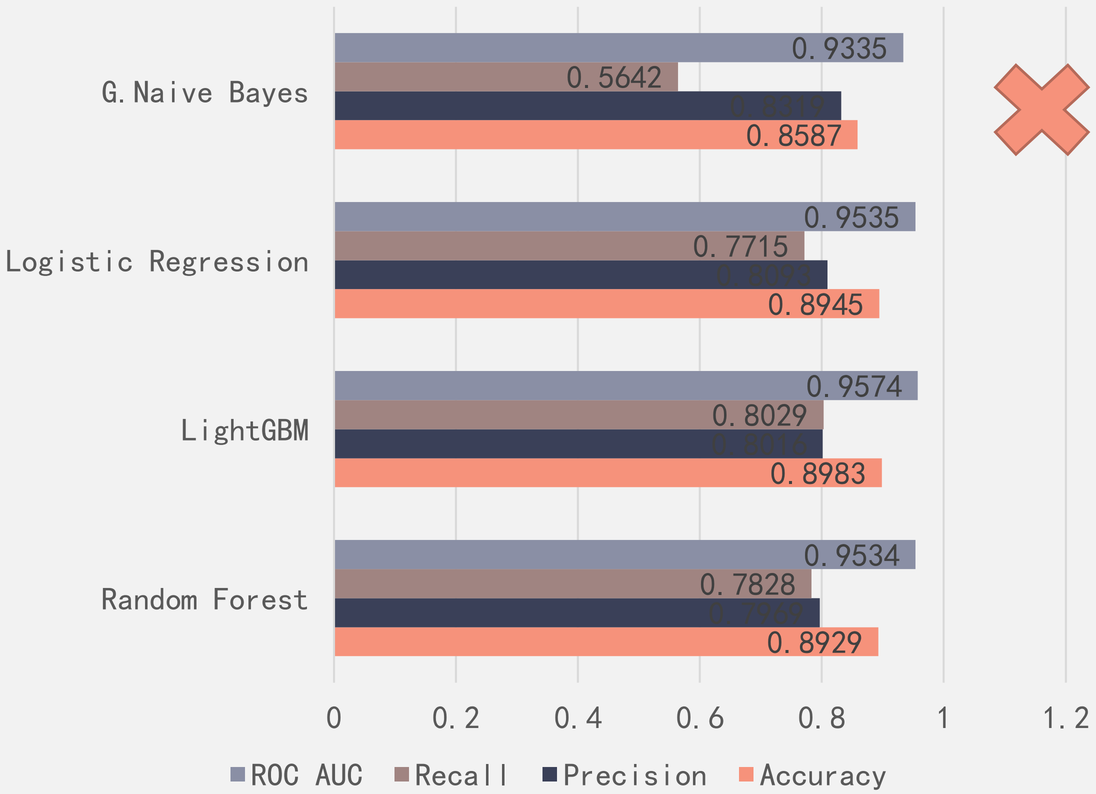
Score Result (Test)

- Accuracy : 0.8587
- Precision : 0.8319
- Recall : 0.5642
- ROC AUC : 0.9335

Model Comparison- Test Score Result



Recall: actual default customer% is correctly classified



Model Comparison-Cost Analysis

Assumptions

- Approval Membership Income: $100 per card
- Continuous Interest Income:$150 per card
- Fixed Costs: $30 000
- Variable Costs: $40 per card
- Average Default costs: $3000 per card

Model	Approvals	Fixed costs	Variable costs	Default costs	Lose customers	Total costs	Revenue	Profit
Formula	TN+FP	Fixed	(TN+FP)*$40	FN*$4000	FP*($100+150)	Fixed +Variable +Default	TN($100+$150) +FP*$100	Revenue-Costs
Random Forest	85,237	$30,000	$3,409,480	$19,221,000	$1,647,800	$24,308,280	$22,807,060	-$1,501,220
Light GBM	85,237	$30,000	$3,409,480	$17,439,000	$1,640,520	$22,519,000	$22,811,740	$292,740
Logistic Regression	86,616	$30,000	$3,464,640	$16,083,000	$1,887,200	$21,464,840	$23,039,280	$1,574,440

Improvements

- Feature engineering and EDA is important, decide the score range
- Parameter tuning slightly improve the score
- Grid search or Bayesian search may give more accurate parameters

Thank You

By Clivia Kong

References

- Kaggle Competition:
  - https://www.kaggle.com/competitions/amex-default-prediction
- Parameters Tuning:
  - https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/
  - https://lightgbm.readthedocs.io/en/v3.3.2/Parameters-Tuning.html
  - https://towardsdatascience.com/bayesian-optimization-for-hyperparameter-tuning-how-and-why-655b0ee0b399
- Coding:
  - https://www.kaggle.com/code/nicapotato/titanic-voting-pipeline-stack-and-guide/notebook