Customer Transaction Prediction

MSBA 6410
Chu Yun Hsiao, Yi Fang, Hao Chun Niu, Jiarui Hu, Shuyun Liu

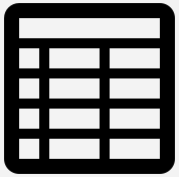
Table of Contents

01

Business Problem

Which customers will make a specific transaction in the future, irrespective of the amount of money transacted?

02
Our Dataset

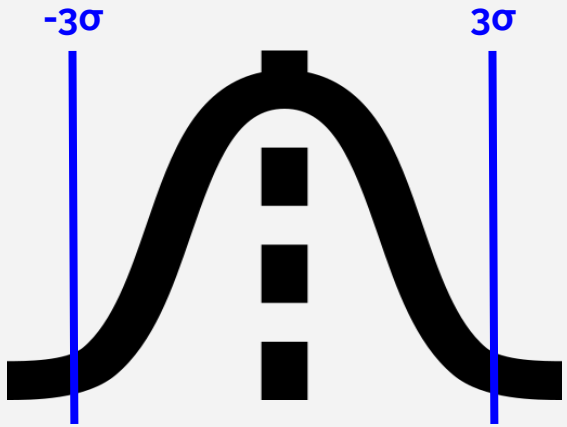


- 200 Anonymized numeric feature variables
- The binary target column
- An ID code column

03

Data Preprocessing

Outliers Issue



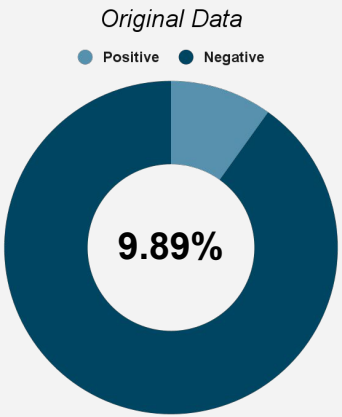
After excluding the outliers, 5.52% of data is removed.

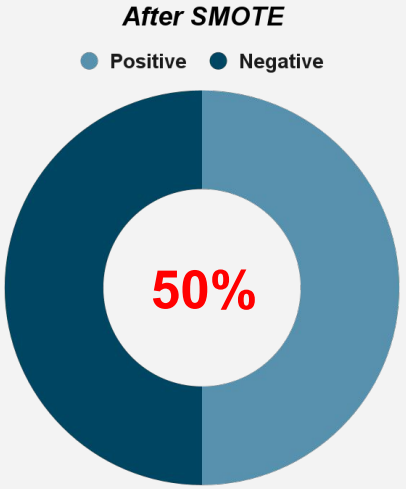
| Z-Score | > 3

03

Data Preprocessing

Imbalance Data Issue





04

Model Overview

- Logistic Regression with gridsearchCV
- XGBoost with gridsearchCV
- LightGBM with gridsearchCV

Logistic Regression with gridsearchCV

Why powerful?

<u>Easy to implement, interpret and efficient to train. Can be use as benchmark to measure performance</u>

Pros +

- Simplest machine learning algorithms
- Provide the importance of each feature

Cons −

- Over-fit when the datasets are on high dimensional
- Sensitive to outliers

XGBoost with gridsearchCV

Why powerful? **Learn from previous mistake through iteration**

Pros +

- General good result
- Less data preparation needed
- Provide insight on key factor

Cons −

- Result is more likely to be influenced by extreme value
- Can not transform categorical data in numerical form

LightGBM with gridsearchCV

Why powerful than XGBoost?

LightGBM grows vertically while XGBoost grows horizontally

Pros +

- Time efficient
- General good result
- Perform well with huge dataset

Cons −

- Not compatible with small dataset

05

Model Performance

- Logistic Regression with gridsearchCV
- XGBoost with gridsearchCV
- LightGBM with gridsearchCV

Logistic Regression with gridsearchCV

Hyperparameter Tuning

Penalty	l2
C	0.001
solver	lbfgs

Model Performance

Accuracy	Precision	Recall	F1-score
0.91	0.89	0.91	0.89

XGBoost with gridsearchCV

Hyperparameter Tuning

colsample_bytree	0.3
gamma	0.01
learning_rate	0.1
max_depth	3
n_estimators	200
objective	binary:logistic

Model Performance

Accuracy	Precision	Recall	F1-score
0.90	0.90	0.90	0.85

LightGBM with gridsearchCV

Hyperparameter Tuning

colsample_bytree	0.3
learning_rate	0.1
max_depth	3
n_estimators	200
objective	binary

Model Performance

Accuracy	Precision	Recall	F1-score
0.91	0.90	0.91	0.87

06

Conclusion & Business Implications

- **91%** of the customers can be predicted correctly for future transactions

Advantages of this analysis:

- More precise and targeted incentive plan
  - By Category, amount of money
- Fraud detection on irregular transactions
  - Comparing to predicted actions

Thank You!