STATS507 PROJECT

Predict the default probability
For Online Microlending platforms

Member: Yifeng He, Wang Xiang, Yanze Liu

CONTENTS

PART 01 **Project Overview**

PART 02 **Dataset Description**

PART 03 **Statistical Methods**

PART 04 **Conclusion**

PART 01

Project Overview

Current Situation

+

=

Time-consuming and high cost

Standard is not clear

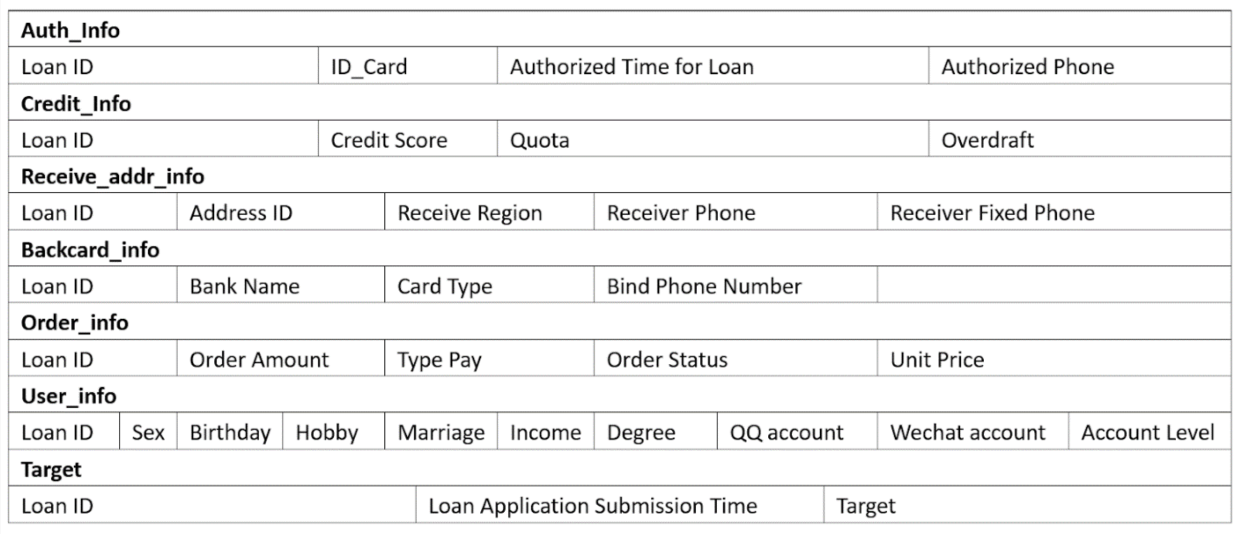
Misclassification

→ **Using programming and statistical methods, based on the user’s past financial records**

PART 02

Dataset Description

Dataset details



Auth_Info									
Loan ID	ID_Card	Authorized Time for Loan	Authorized Phone						
Credit_Info									
Loan ID	Credit Score	Quota	Overdraft						
Receive_addr_info									
Loan ID	Address ID	Receive Region	Receiver Phone	Receiver Fixed Phone					
Backcard_info									
Loan ID	Bank Name	Card Type	Bind Phone Number						
Order_info									
Loan ID	Order Amount	Type Pay	Order Status	Unit Price					
User_info									
Loan ID	Sex	Birthday	Hobby	Marriage	Income	Degree	QQ account	Wechat account	Account Level
Target									
Loan ID	Loan Application Submission Time	Target							

7 csv files

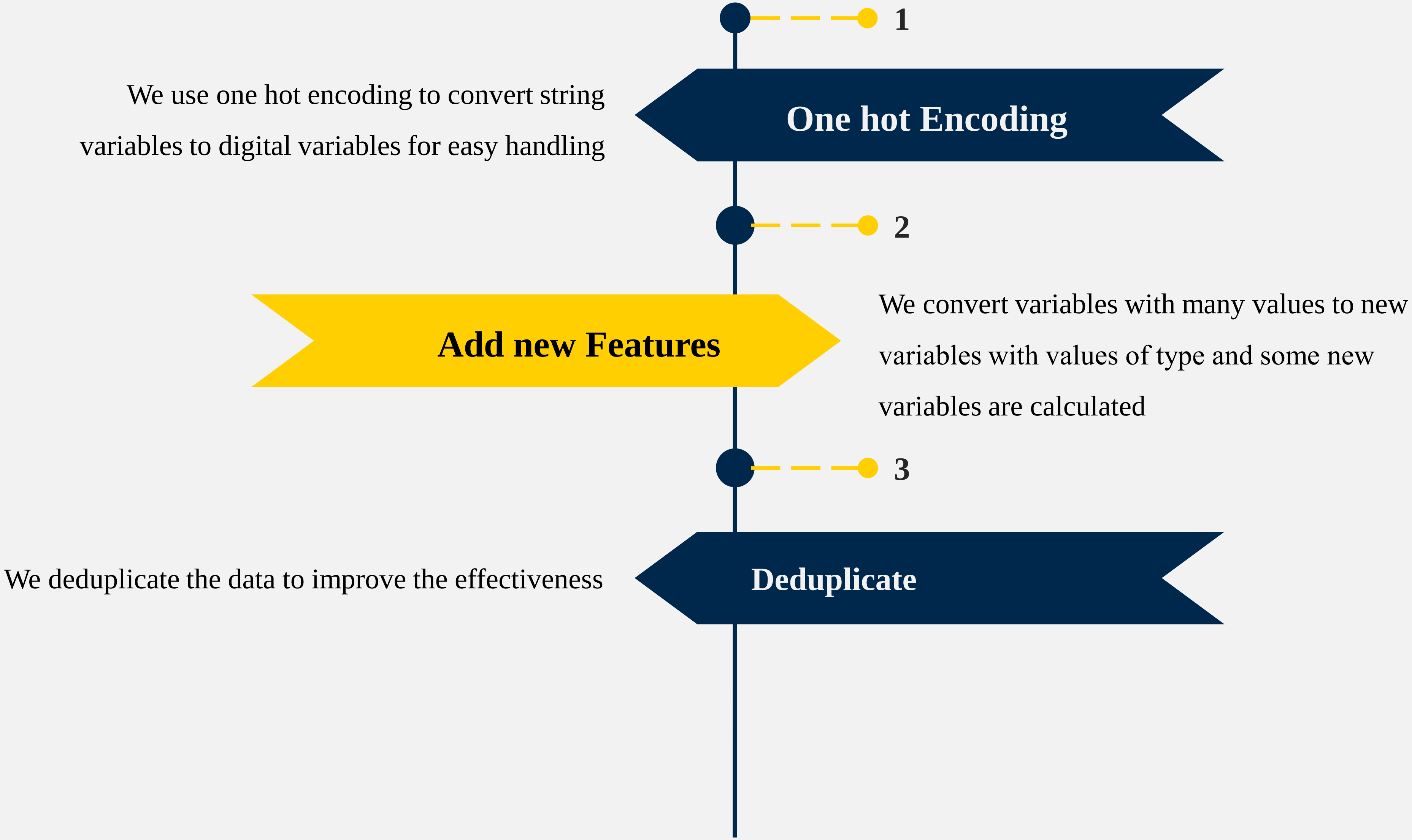
29 variables

120,000 observations

PART 03

Statistical Methods

Data cleaning & Feature Preprocessing



1

One hot Encoding

We use one hot encoding to convert string variables to digital variables for easy handling

2

Add new Features

We convert variables with many values to new variables with values of type and some new variables are calculated

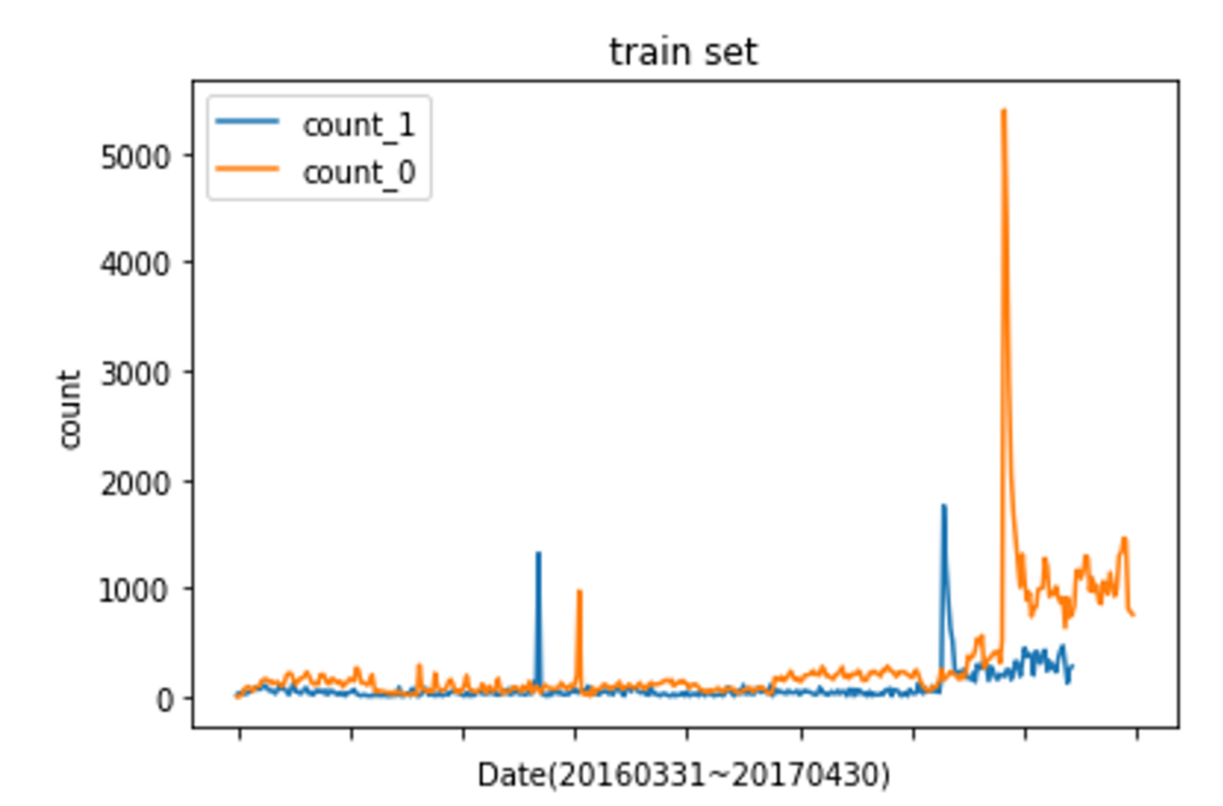
3

Deduplicate

We deduplicate the data to improve the effectiveness

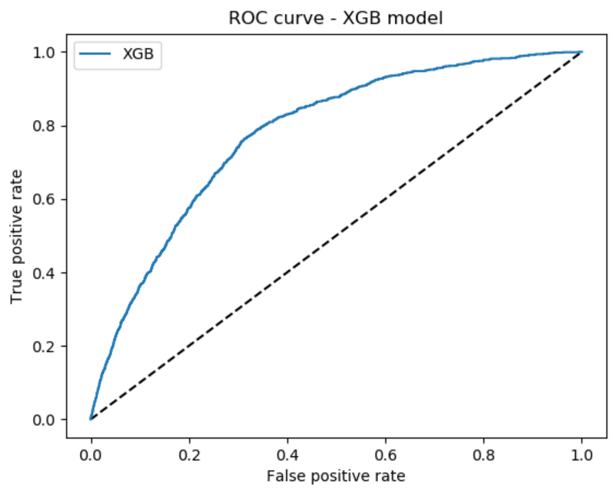
Understand the features

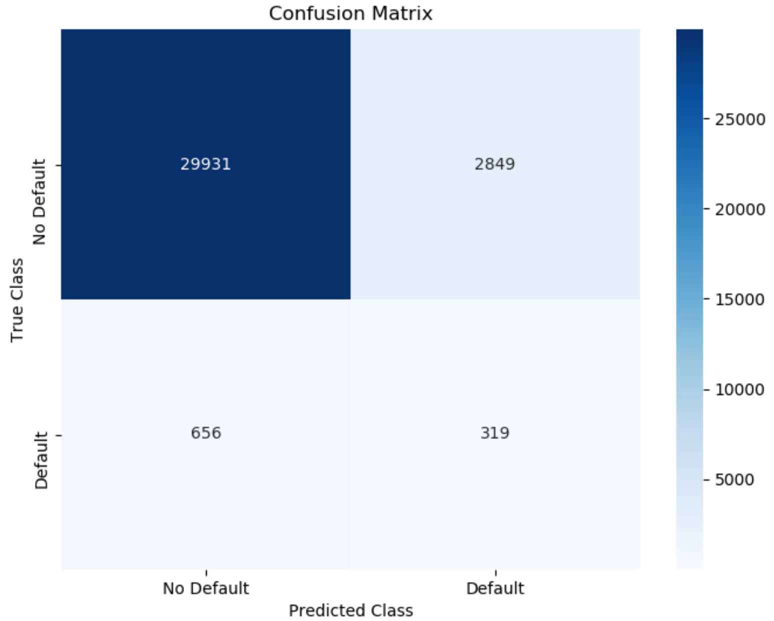
● Several Spikes



Model Training and Result Evaluation

XGBoost


ROC curve - XGB model


Confusion Matrix

Accuracy score:

89.62%

Recall:

32.72%

Precision:

10.07%

Valid AUC Score:

0.620

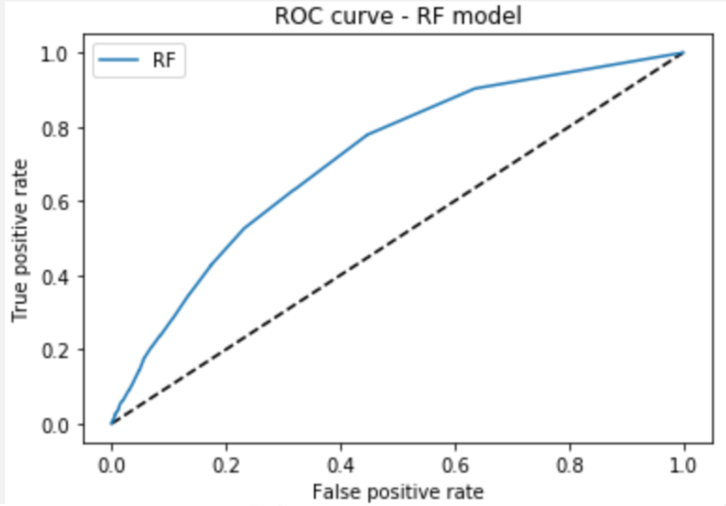
imbalance in

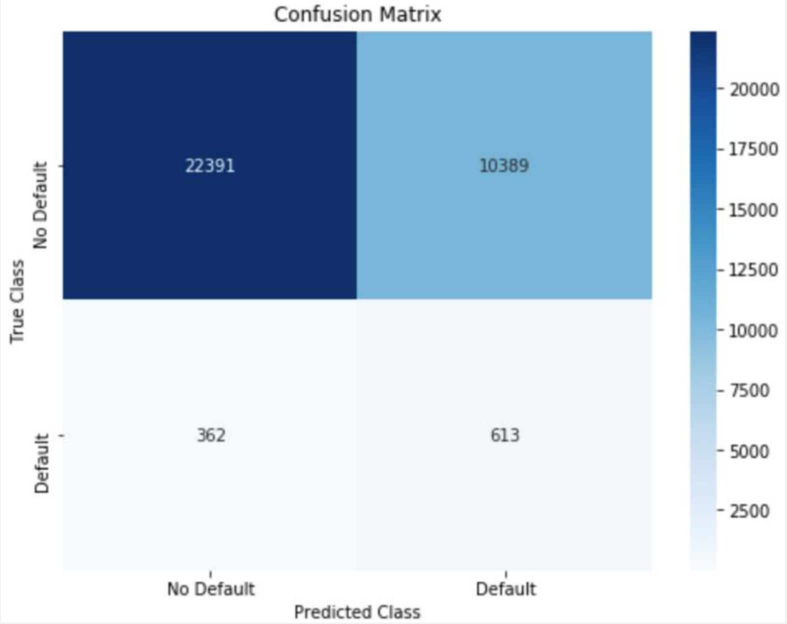
data

Model Training and Result Evaluation

“

Random Forest





	No Default	Default
No Default	22391	10389
Default	362	613

Accuracy score:

68.15%

Recall:

62.87%

Precision:

5.57%

”

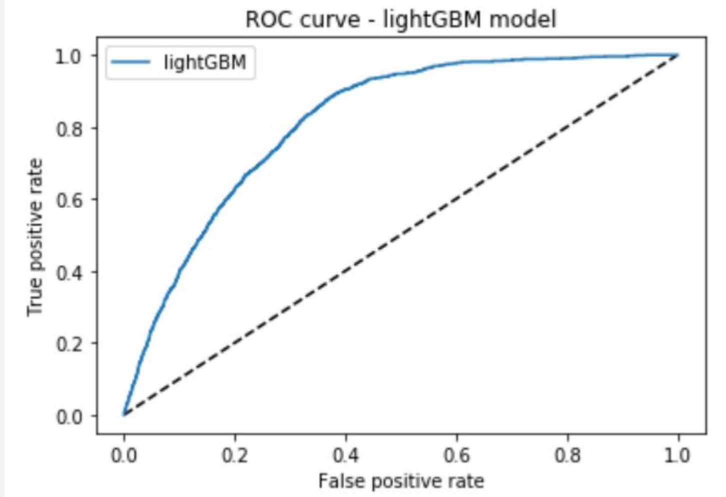
Valid AUC Score:

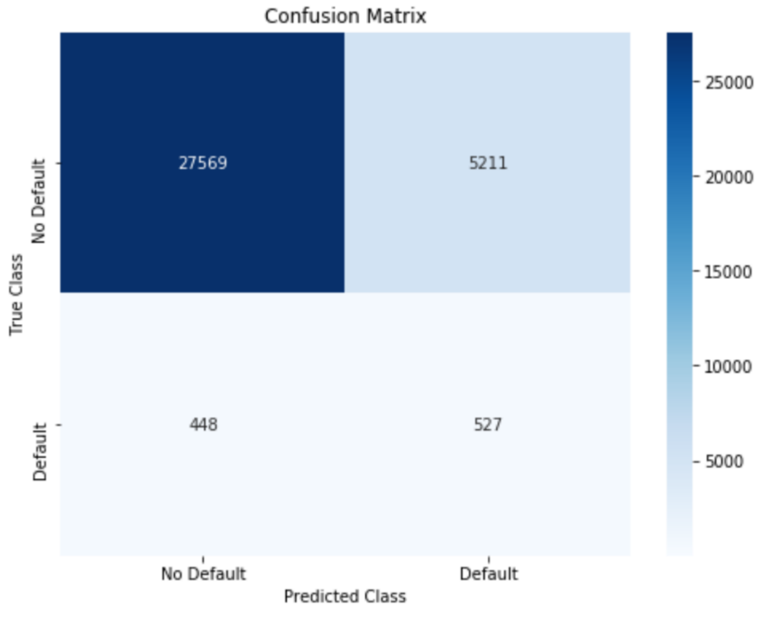
0.710

Model Training and Result Evaluation

“

Light GBM





Accuracy score:

83.24%

Recall:

54.05%

Precision:

9.18%

”

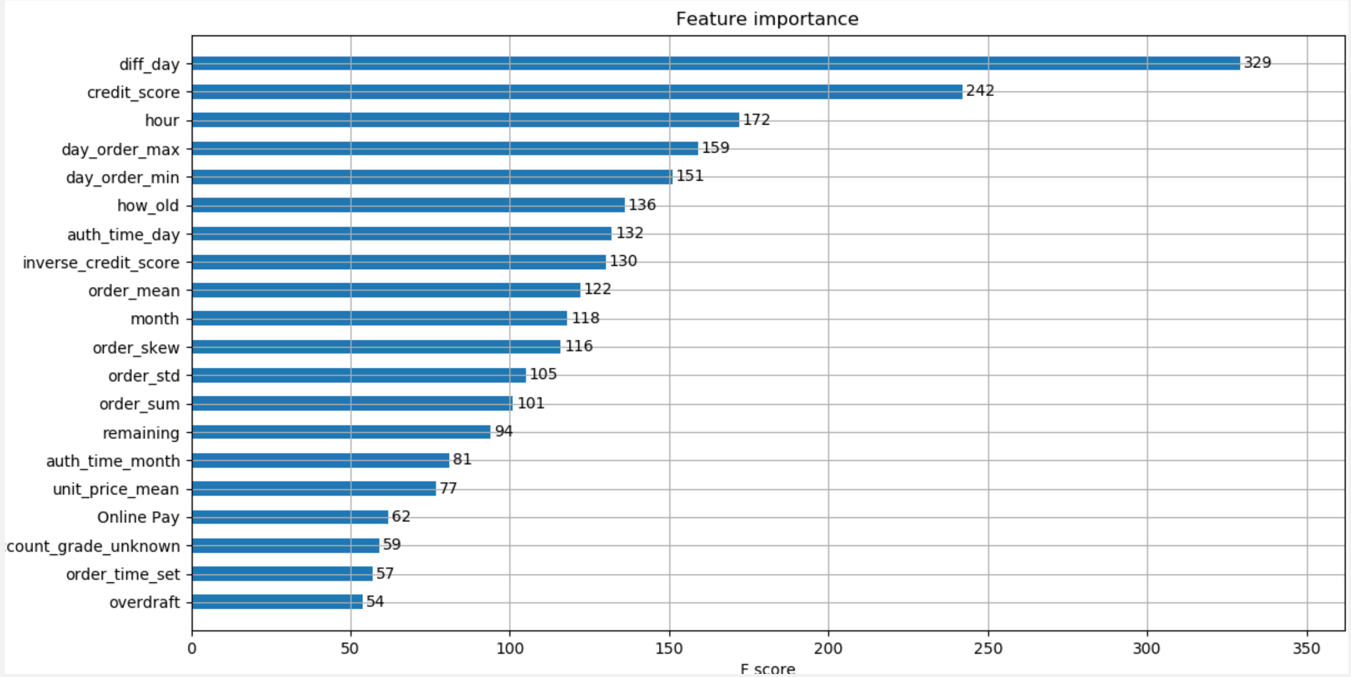
Valid AUC Score:

0.815

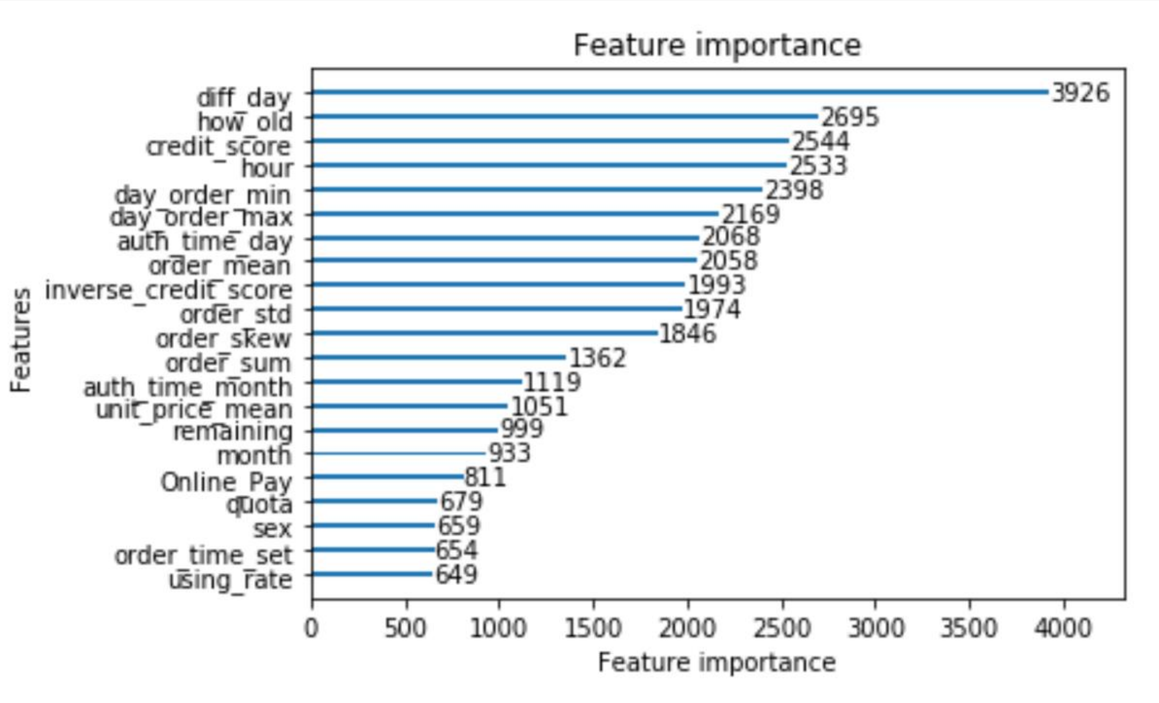
Feature Selection

- **Select 20 most important variables**

XGBoost



Light GBM



The results obtained by the two methods are similar

PART 04

Conclusion

Conclusion

Comparison

A XGBoost

B Random Forest

C Light GBM

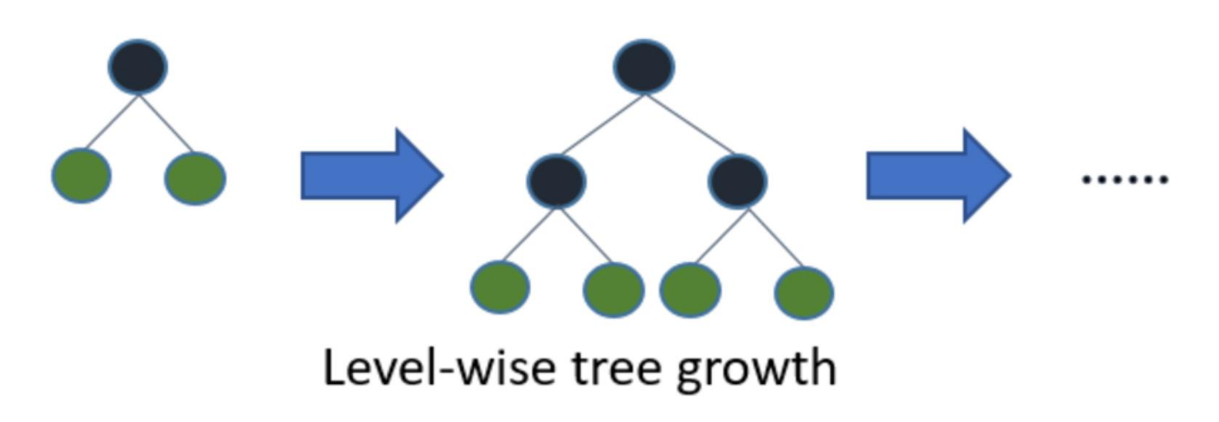
Knowledge Sharing

Light GBM

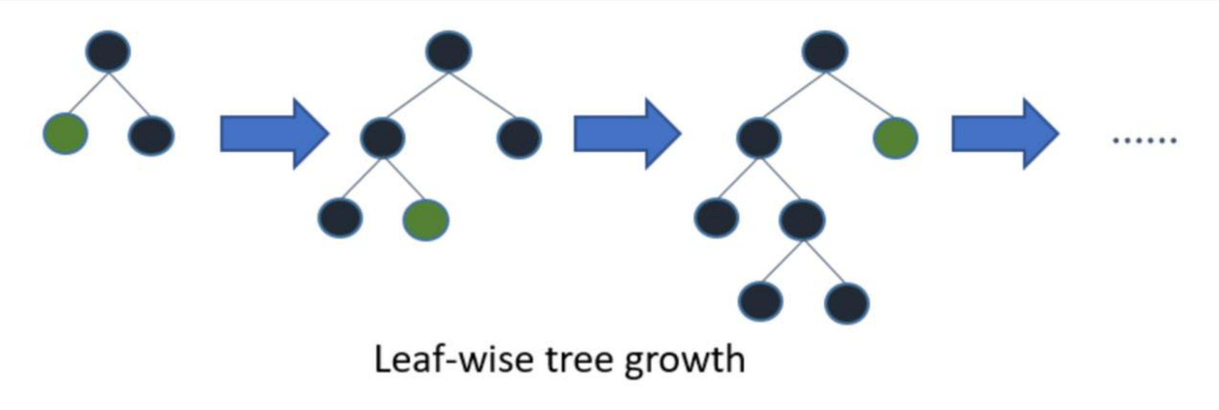
Light GBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithm [1]

What different from other boosting algorithms

For most boosting algorithms:



For light GBM:



Reference

1. Mandot, P. (2018, December 1). What is LightGBM, How to implement it? How to fine tune the parameters? Retrieved from https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

THANK YOU