Credit Risk Prediction in Loan Lending Scenario

Team: DataBears

Teammates: Wu Dingjun; Wang Hongjian; Zhu Xiangqian; Ding Zhiqin

Date: 2021.12.12

Background & Hypothesis

Lack of Credit Rate → Overdue Loan → Financial Risk

- **Problem:** The role of credit rate is highlighted due to the development of lending activities.
- **Our goals:** Evaluate the crisis of loan default, find critical factors and create a model to limit the bad lending.

Hypothesis:

- 1) One's behavior in the **past** can represent his or her behavior in the **future**
- 2) People with same characteristics will behave similarly

Basic

- 3) Credit risk is strongly correlated to the financial environment, loan types, borrowers' solvency and trustworthiness
- 4) people in better financial status are tend to repay on schedule

To be confirmed

Reference: Du Miaomiao. Southern Finance, 2008.

Related Work - common technologies

- **Linear regression** $Y = X\widehat{B} + e$

linear relationships between independent variables & categorical data

- **Logistic regression**

predict the credit rating with following formulation

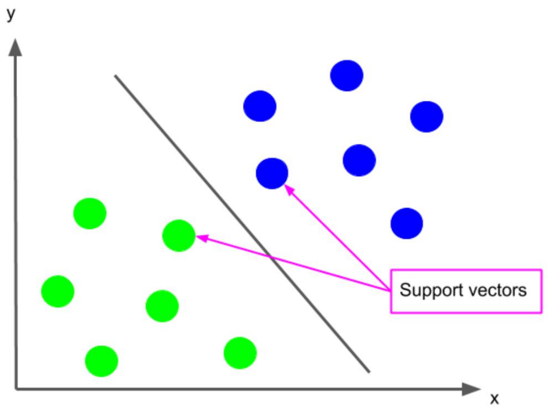
$$ln\frac{p}{(1-p)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

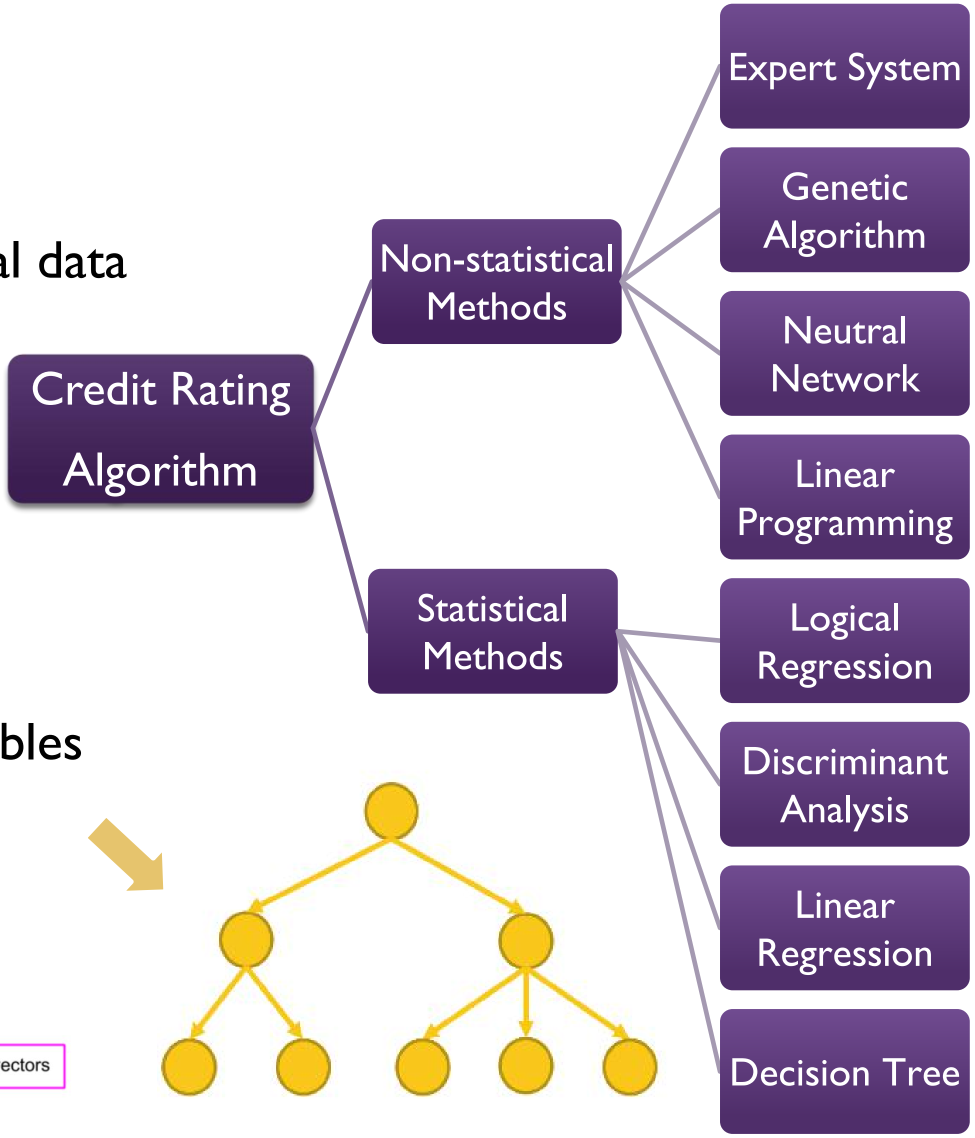
- **Decision Tree**

determine the classification rules by splitting the values of variables

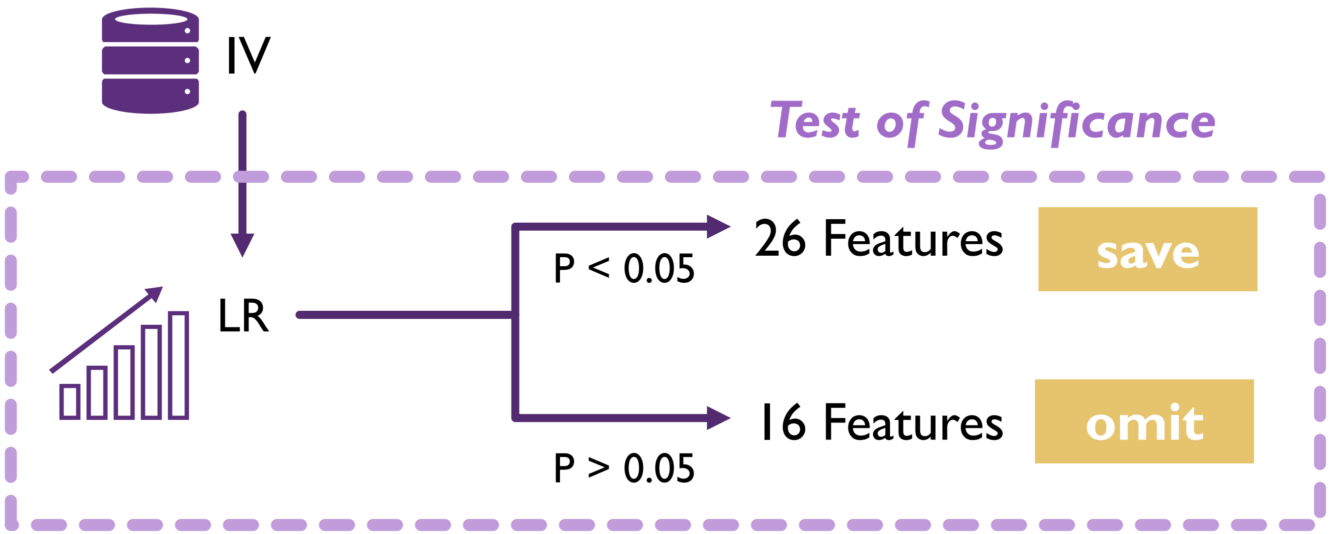
- **Support Vector Machine**

supervised learning model for classification and regression analysis





Determination of Factors



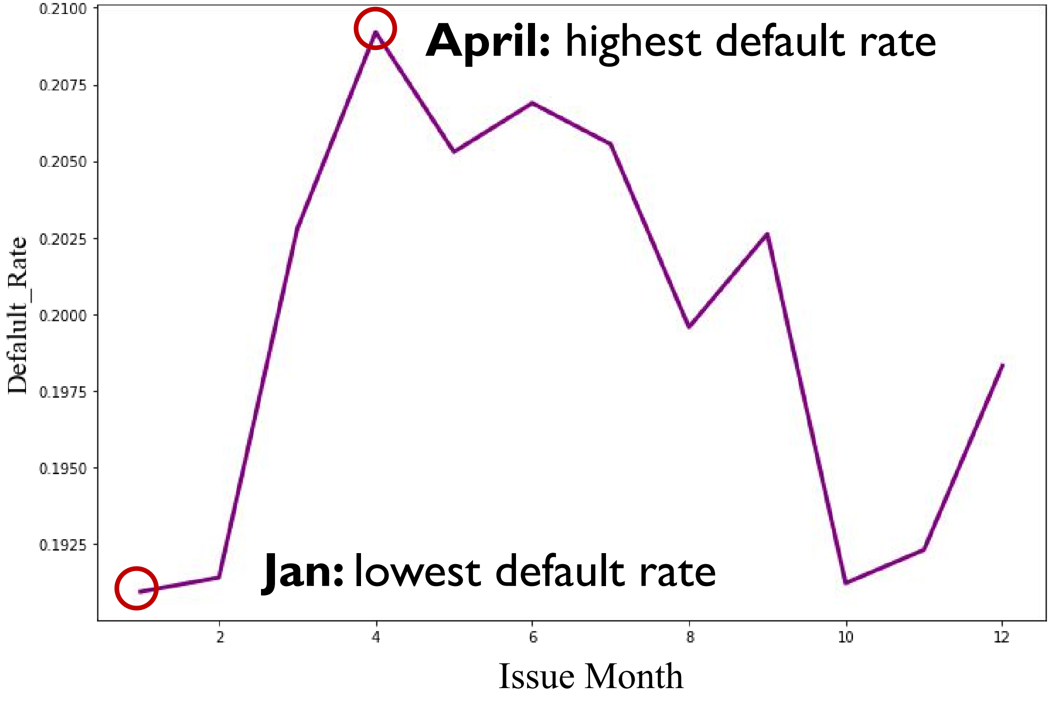
IV(Information Value) : IV reflects the predictive power of variables, mainly used for feature selection.

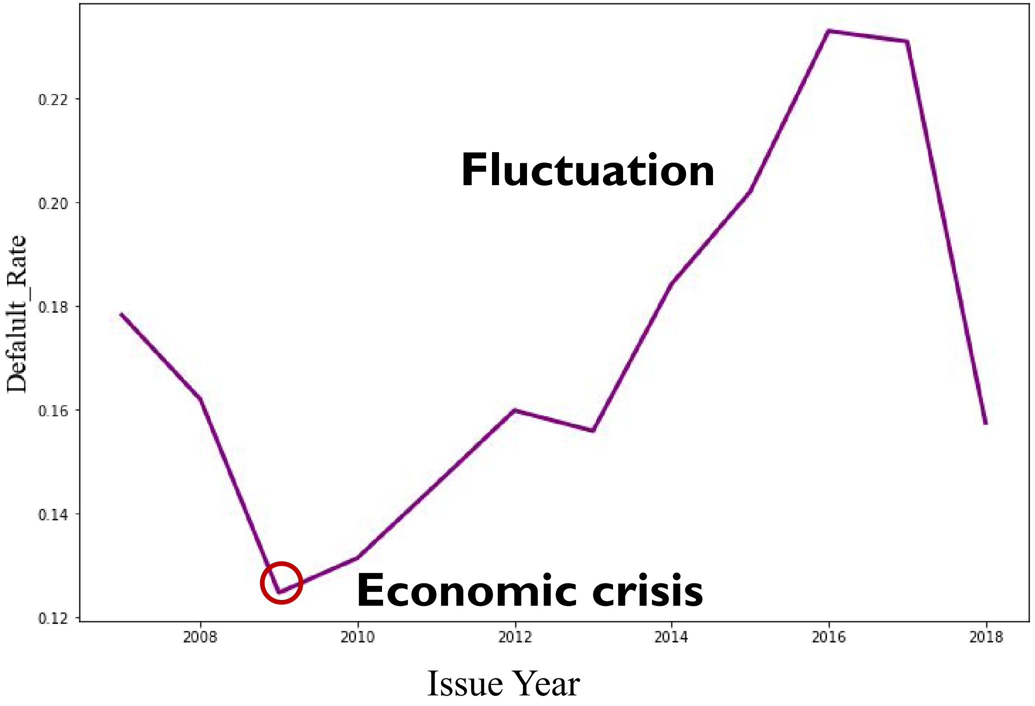
Reference: https://www.jianshu.com/p/cc4724a373f8

Category	Detailed Factors	Importance
Loan	Installment	9
	subGrade	1
	interestRate	2
	title	10
Solvency	term	3
	dti	4
	verificationStatus	5
	loanAmnt	8
	homeOwnership	12
	issueDate	6
	annualIncome	13
	revolUtil	14
Anonymity	n14	7
	n9	11
	n1	15

Data Analysis

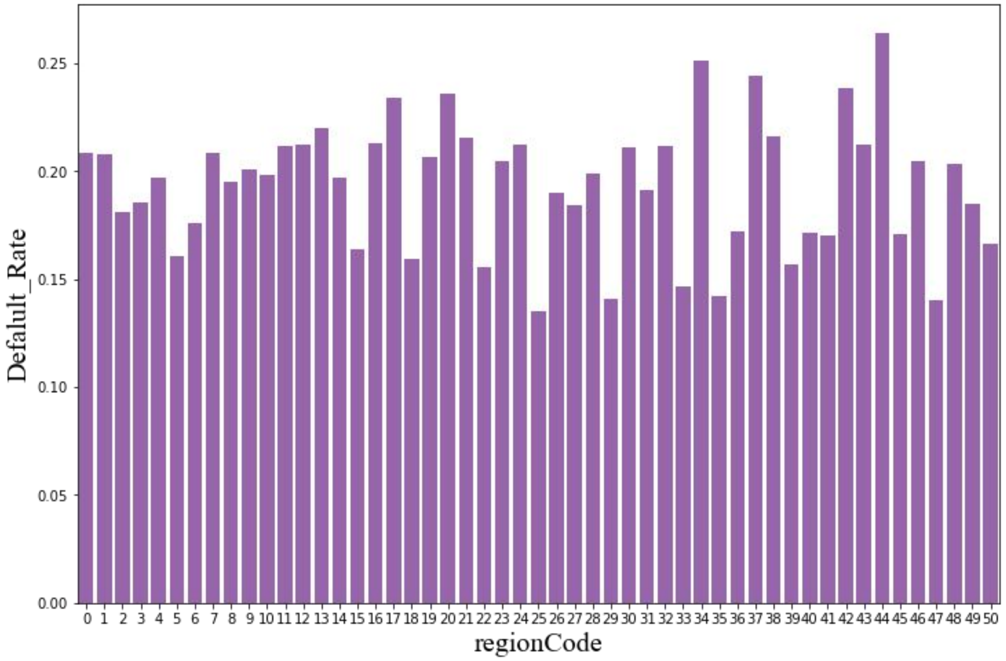
Time Period



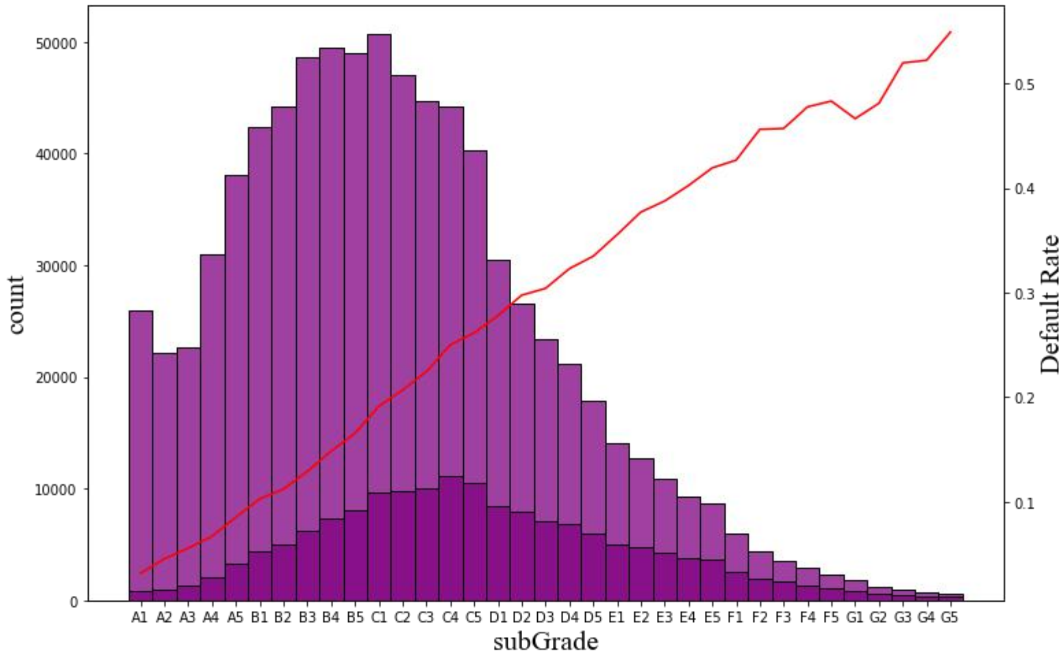


Data Analysis

Continents

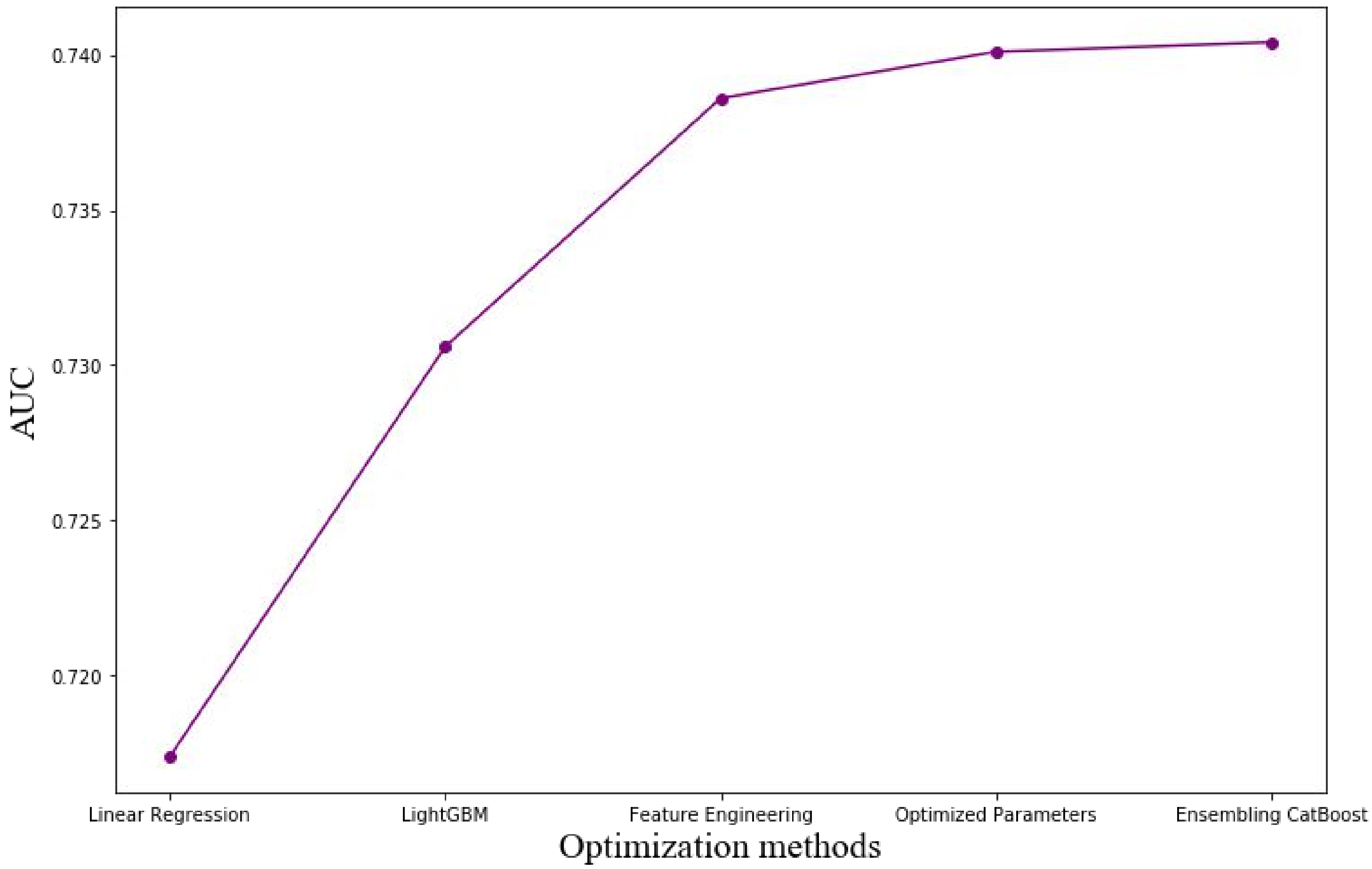


Loan Category



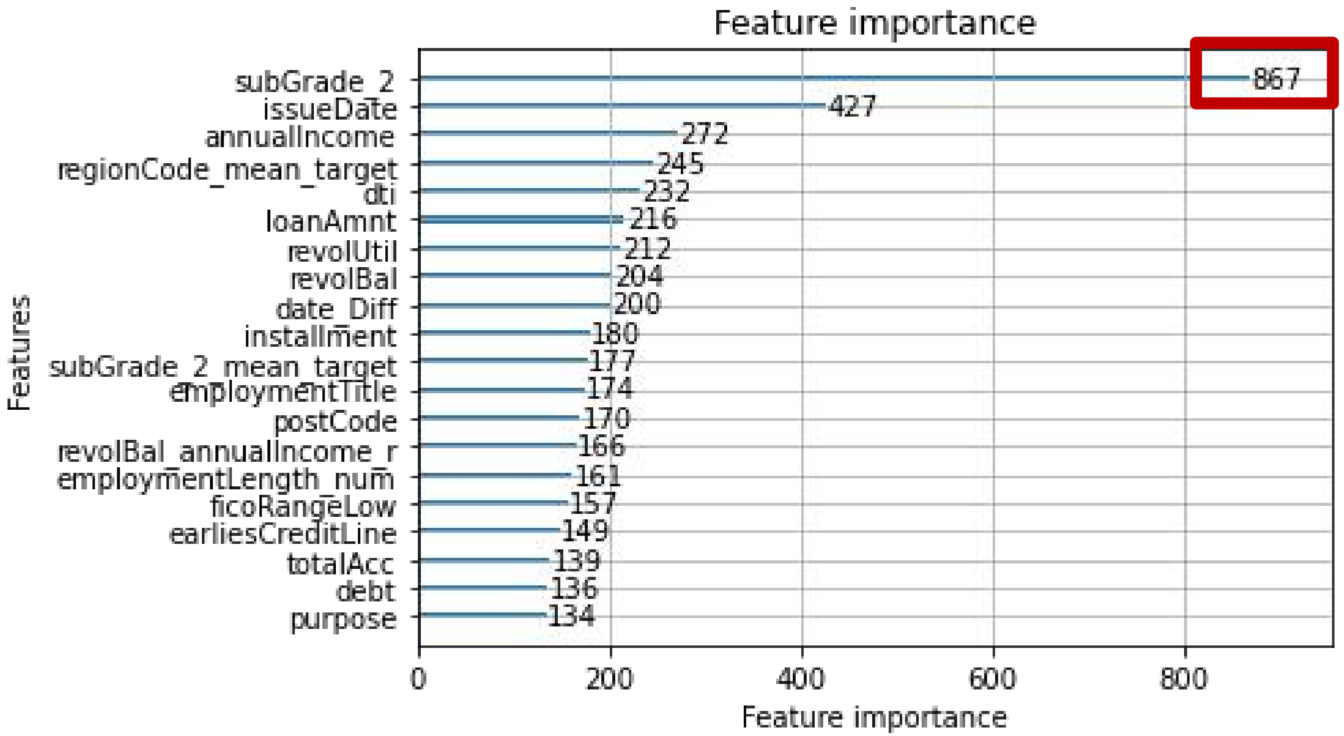
* Most continents have a similar default rate
* Continent 44 achieve the highest rate > 0.25

Performances of Model



Model	Description
Linear Regression	Baseline; standard normalization
LightGBM	Faster training speed; GBDT model; good of tabular data
Feature Engineering	Create more features
Optimized Parameters	Hyperparameter optimization by grid search
Ensembling CatBoost	Ensembling with CatBoost

Factor Correlation Analysis



Ranks	Pre-determination	Model selected
1	subGrade	subGrade
2	interestRate	issueDate
3	term	annualIncome
4	dti	regionCode
5	verificationStatus	dti
6	issueDate	loanAmnt

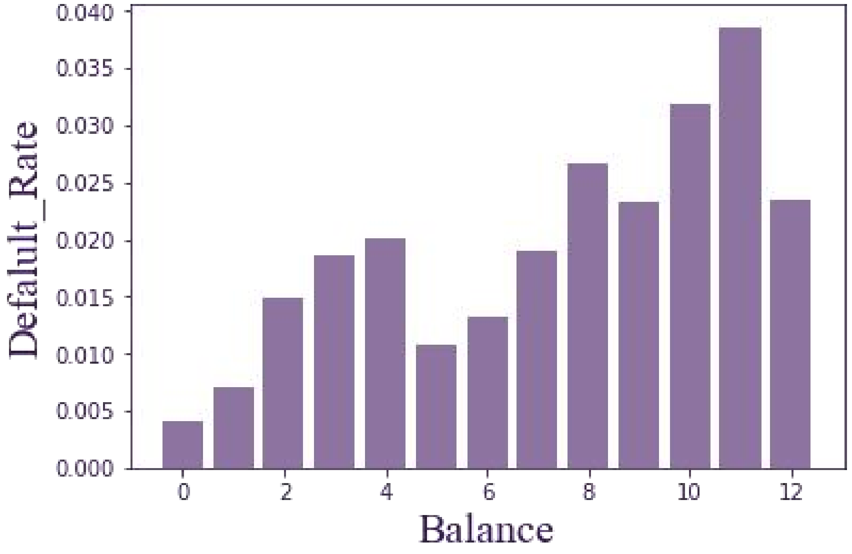
Feature importance
It reflects total gains of splits which use the feature in our decision tree model.

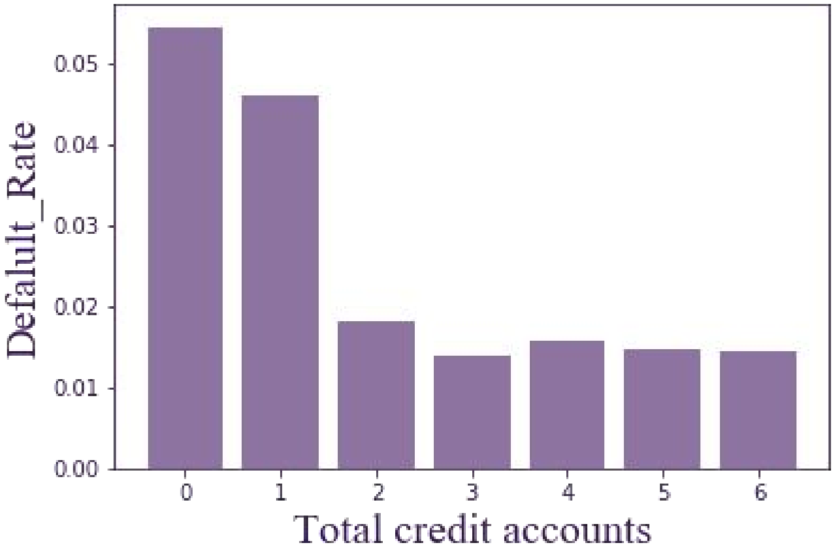
Additional Evidence: dataset Ⅰ - later dataset of ***lending club***

Description: This dataset represents loans made through the Lending Club.

Format: A data frame with 10,000 observations on **55 variables**.

Difference with our dataset: **later data**; personal information is not hidden.





IV Ranks	Features
1	emp_title
2	sub_grade
3	paid_total
4	interest_rate
5	state
6	balance
7	total_debit_limit
8	Installment
9	num_total_cc_accounts

OpenIntro

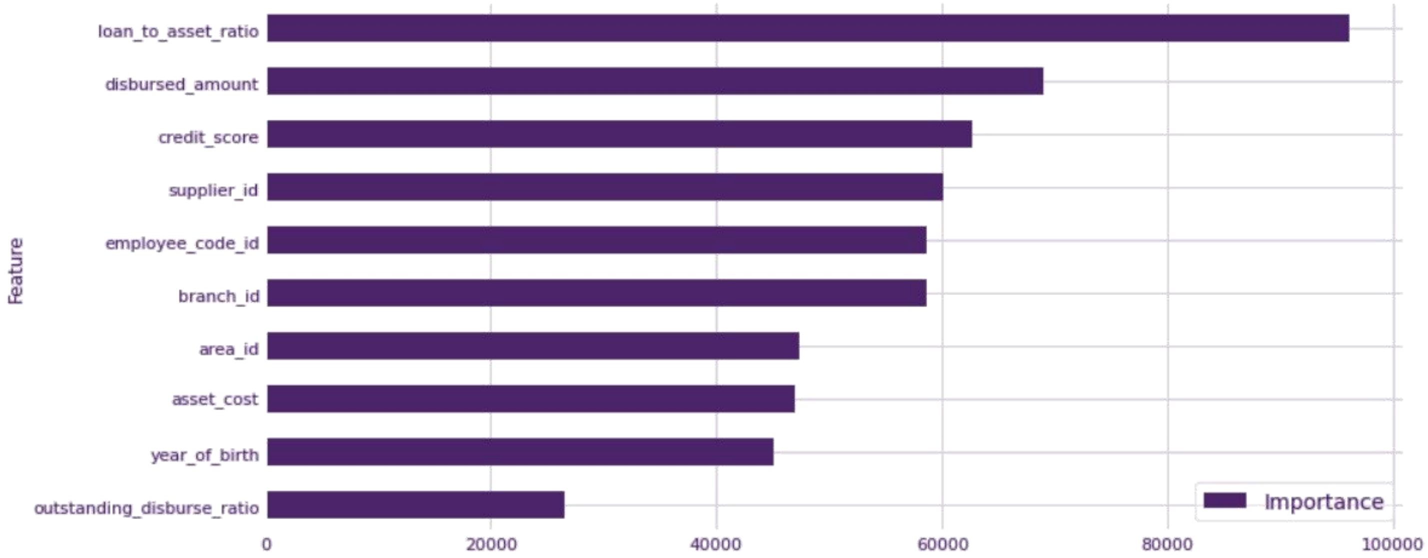
Access : https://www.openintro.org/data/index.php?data=loans_full_schema

Additional Evidence: dataset Ⅱ - dataset of ***car loan default forecast***

Description: This data represents the borrower's car loan information.

Format: A data frame with 150,000 observations on 52 variables.

Features of the dataset: specific loan purpose, more personal information.



01 DTI (Loan to asset ratio), is important feature for default prediction.

02 Personal information (Age, job, etc.) may improve our model.

Access : http://challenge.xfyun.cn/topic/info?type=car-loan

Additional Evidence

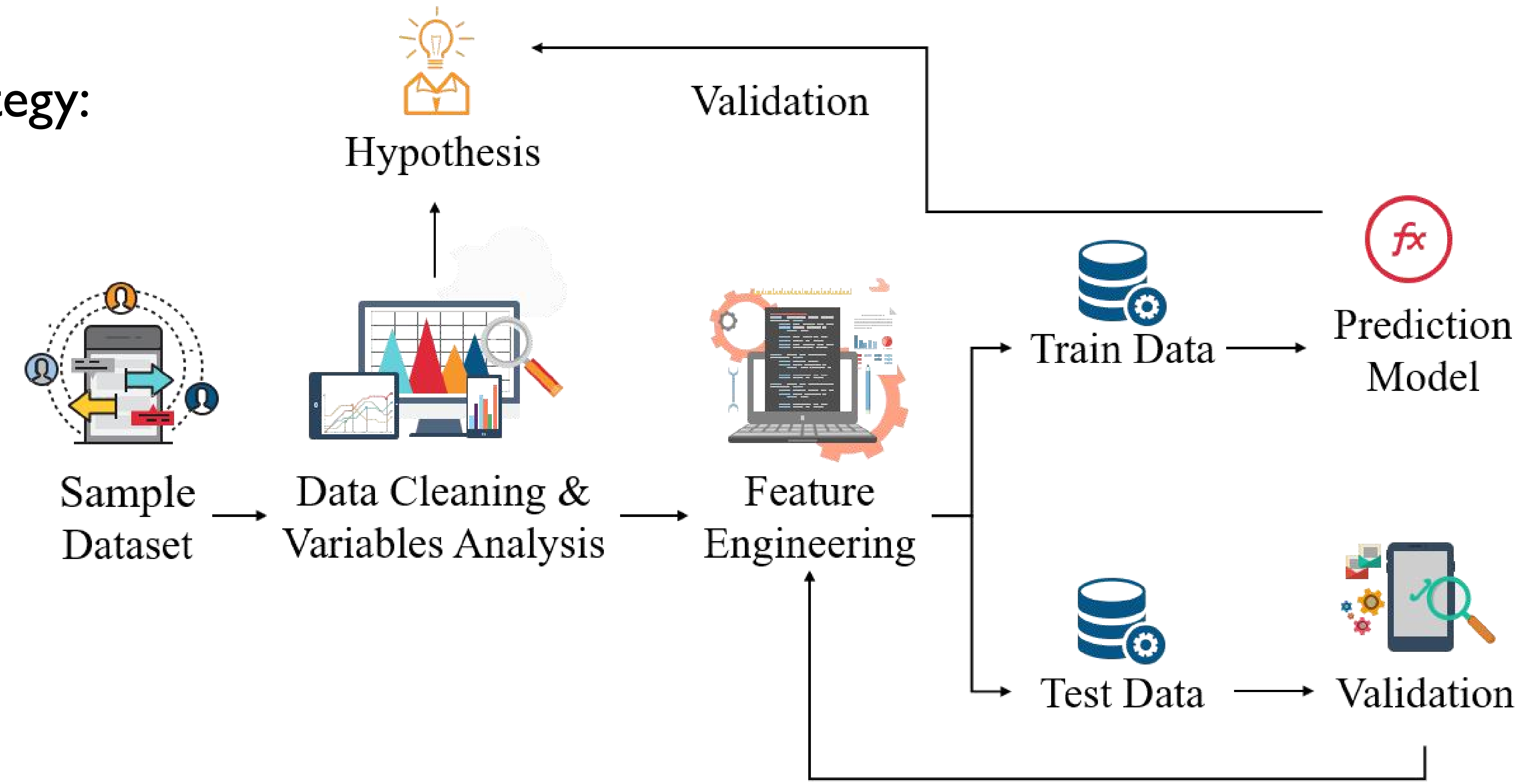
Historical research:

Topic	Year	Author	Main Results
Factors influencing the credit risk	1989	Steenackers, Goovaerts	Age, occupation and other personal characteristics, as well as employment information such as the length of working hours, whether to work in state organs
Peer-to-peer lending	2009	Lyer et al.	Descriptive information can alleviate information asymmetry to some extent and help investors identify customers' credit risk
The inner workings of Lending Club	2017	Bhatnagar Pujun et al.	Build models to predict the interest rate and if a loan will be approved
Credit Risk with Financial News	2020	Tam Tran-The	Proposes a predictive downgrade model using solely news data represented by neural network embeddings

Reference:
Steenackers A, 1989； Iyer R, 2009； Bhatnagar Pujun 2017; Tam, 2020.

Proposed Solution

Our strategy:



PROSPECTS

Limitation of Hidden information

- We can make more detailed analysis with more realistic data (geographic information, etc).
- But there are hidden dangers in privacy security.

Distorted data identification

- Based on empirical results and numerical analysis， Identify cases where personal information has been falsified.

Valuable indicates supplement

- Search for other valid metrics and improve current standards.

Future Works

■ Time value

The accuracy of the model may decrease due to concept drift. The latest data is required for adjustment.

■ Generalization

More datasets can be used to verify the efficiency of our algorithms. Different algorithms may be suitable for different datasets.