Loan Default Prediction

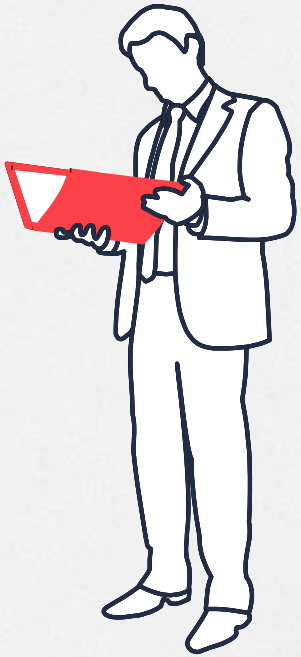
Group 12: Shiyi Hua, Qiuyi Lu, Vaishnavi Kodaganti

How Does Lending Club Work?

BORROWER

Risk

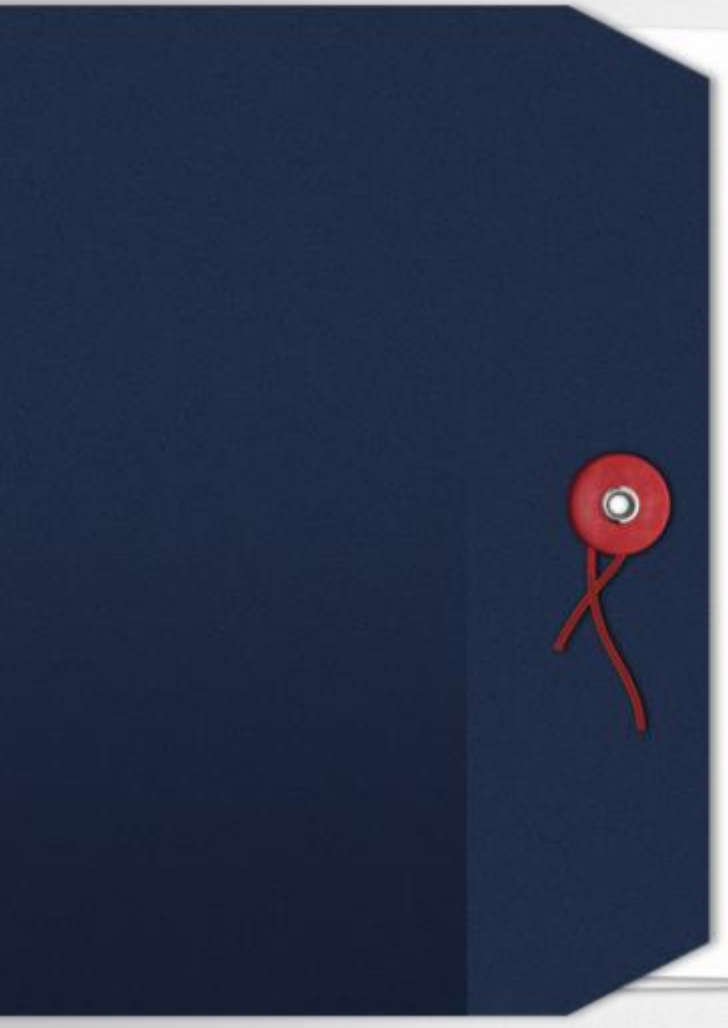
Lending club



INVESTOR

Return

Source: Kaggle



01. Discover business problem

02. Build model to solve problem

03. Result and Conclusion

01

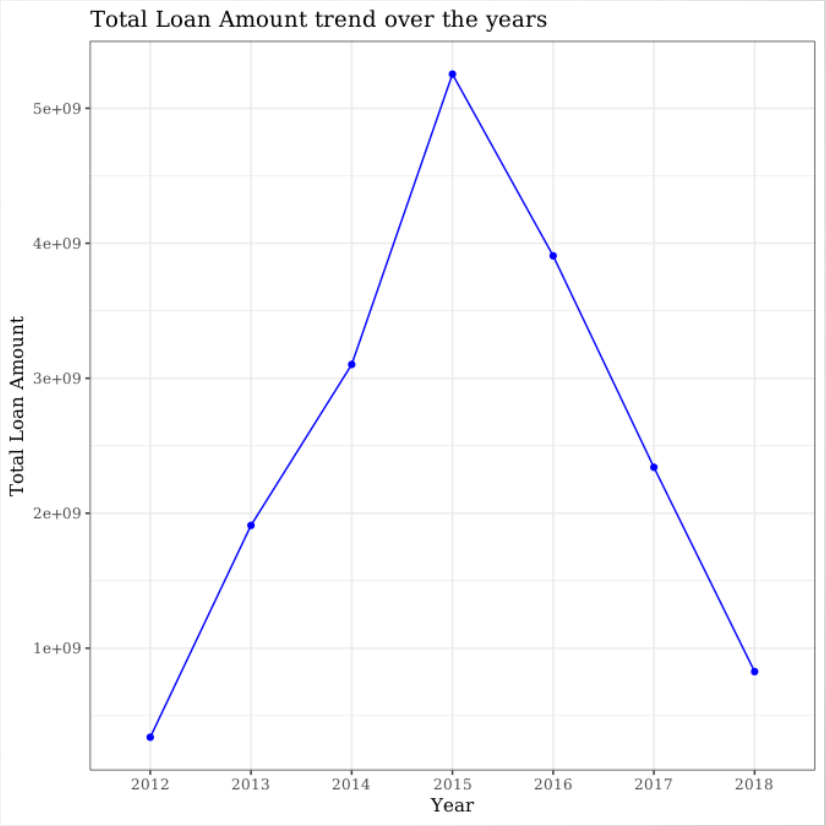
Discover the business problem

Data Exploratory Analysis

Business Probelm

Loan Amount

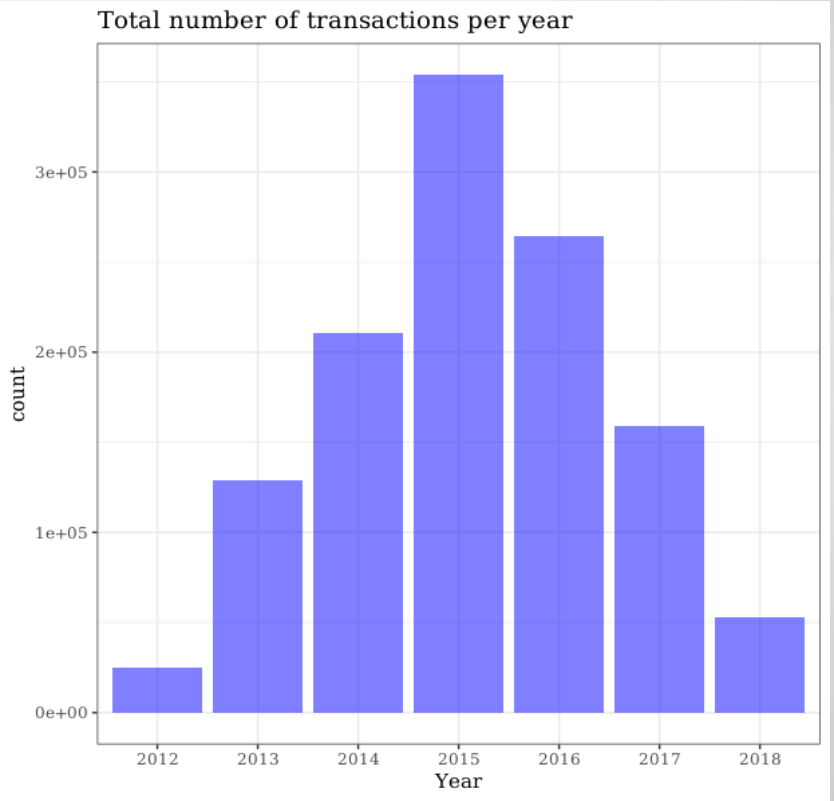
How has the business performed over the years ?



Business Probelm

Number of Transaction

Are the total number of transactions being handled on a growing or declining trend ?



Define Default

Loan Status

Label = 1

Charged Off

Late

Default

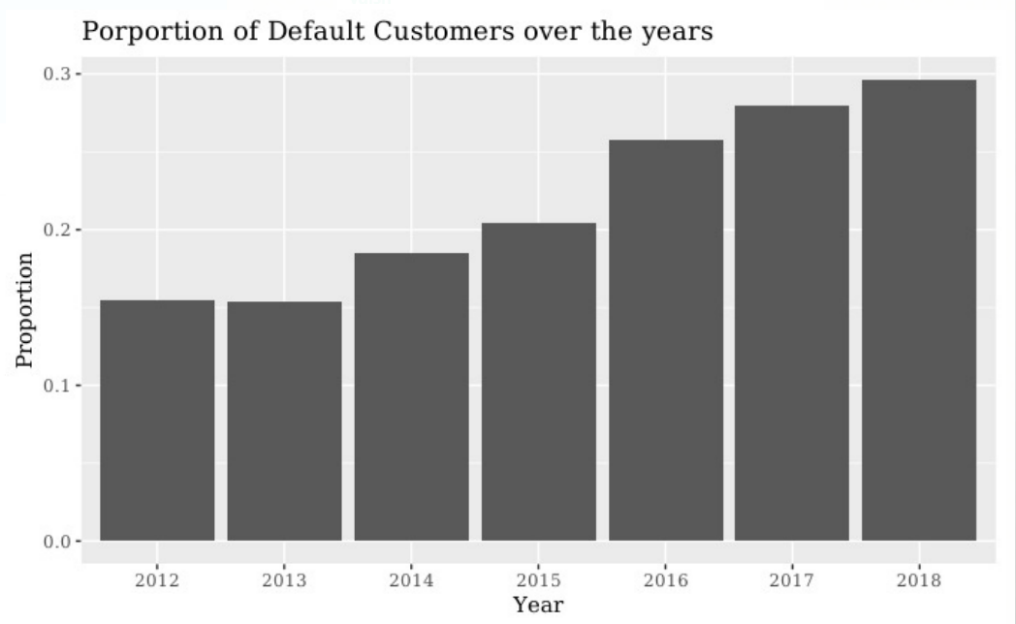
Label = 0

Fully Paid ✔

Business Problem

Default Rate

Has the Annual Default Rate of borrowers increased /decreased over the years?

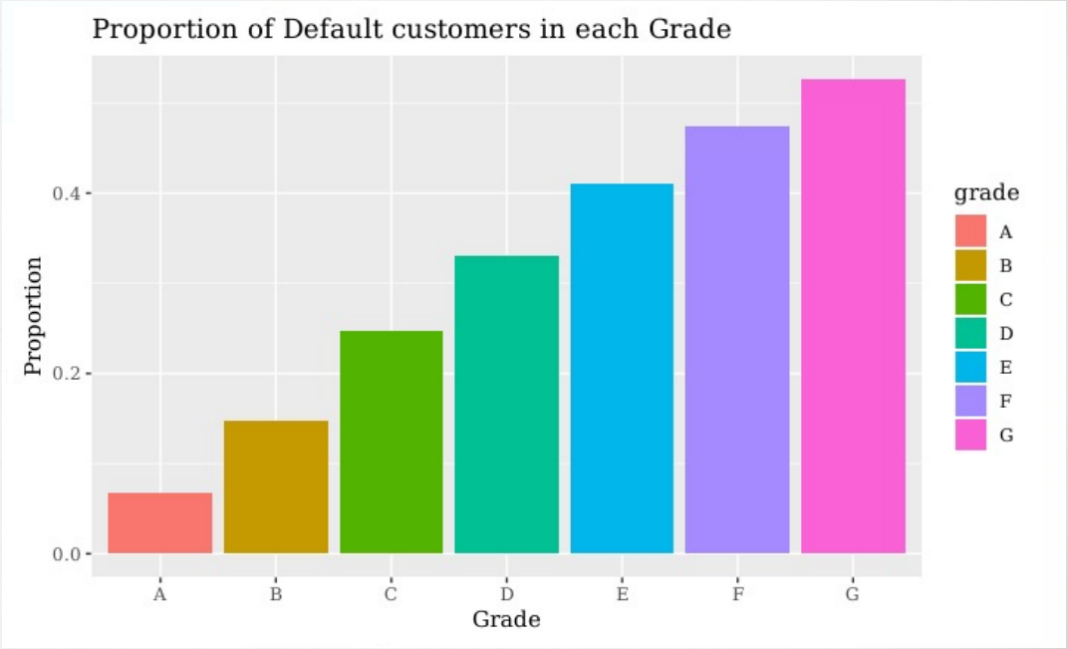


Customer Portrait of Defaulted Records

Customer Portrait

Are borrowers belonging to grade A loans less likely to default?

Yes.

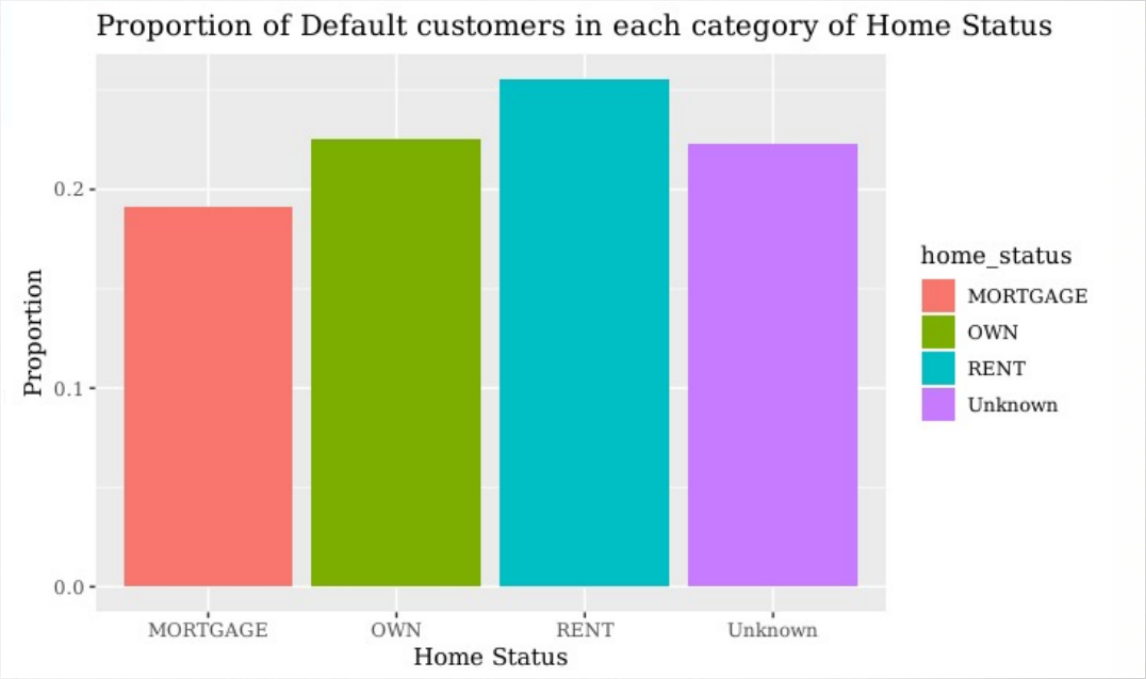


Discover the business problem

Customer Portrait

Do borrowers with Mortgage have the highest default rate?

No.

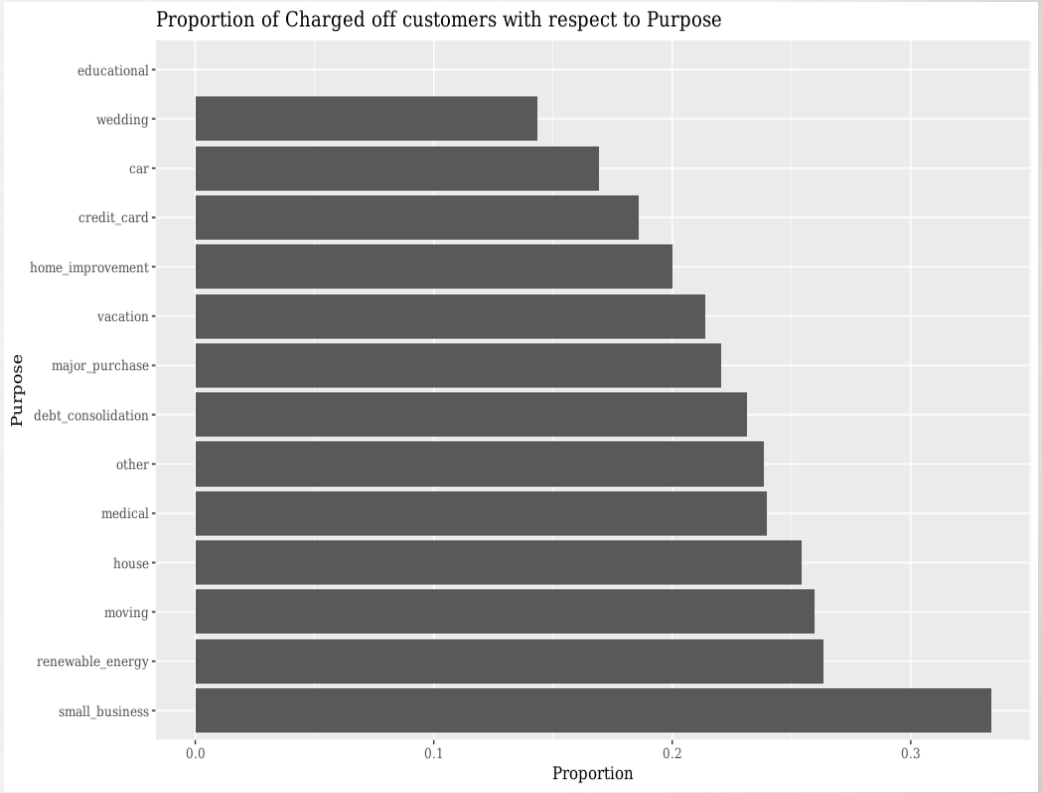


Discover the business problem

Customer Portrait

Is purpose for borrowers a factor that influence the default?

Yes.



02

Build model to solve problem

Logistic Regression &Deep Learning

Feature Selection & Data Cleaning

2.24m rows × 145 features → 1.19m rows × 23 features

Log Transformation

- total accounts

Drop NA

- employment length
- % trades never default

Cut Numeric into Intervals

- annual income
- % trades never default

Categorization

- # default within past 2 yrs
- # charge-offs within 12 months

Baseline Model:Logistic Regression

Train Model

- Training size: 50%

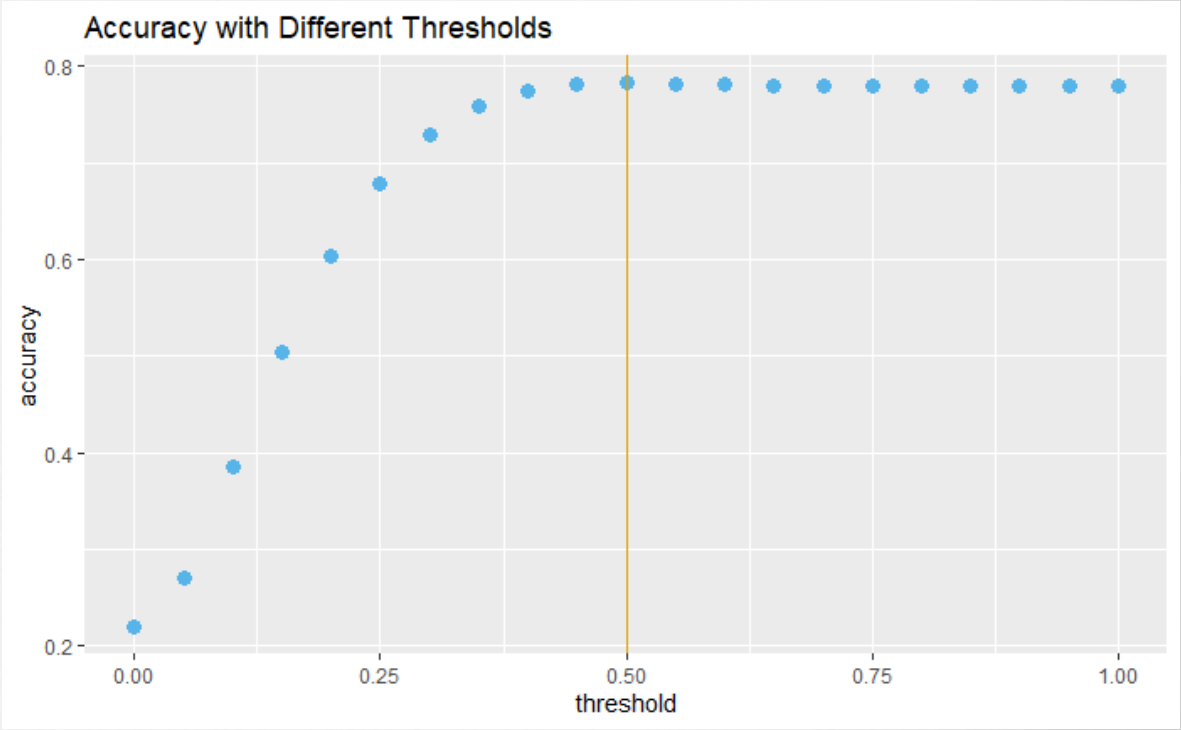
Accuracy: 78%

Detection rate: 7%

Threshold: 0.5

- Reason: imbalanced dataset (20% are label 1)

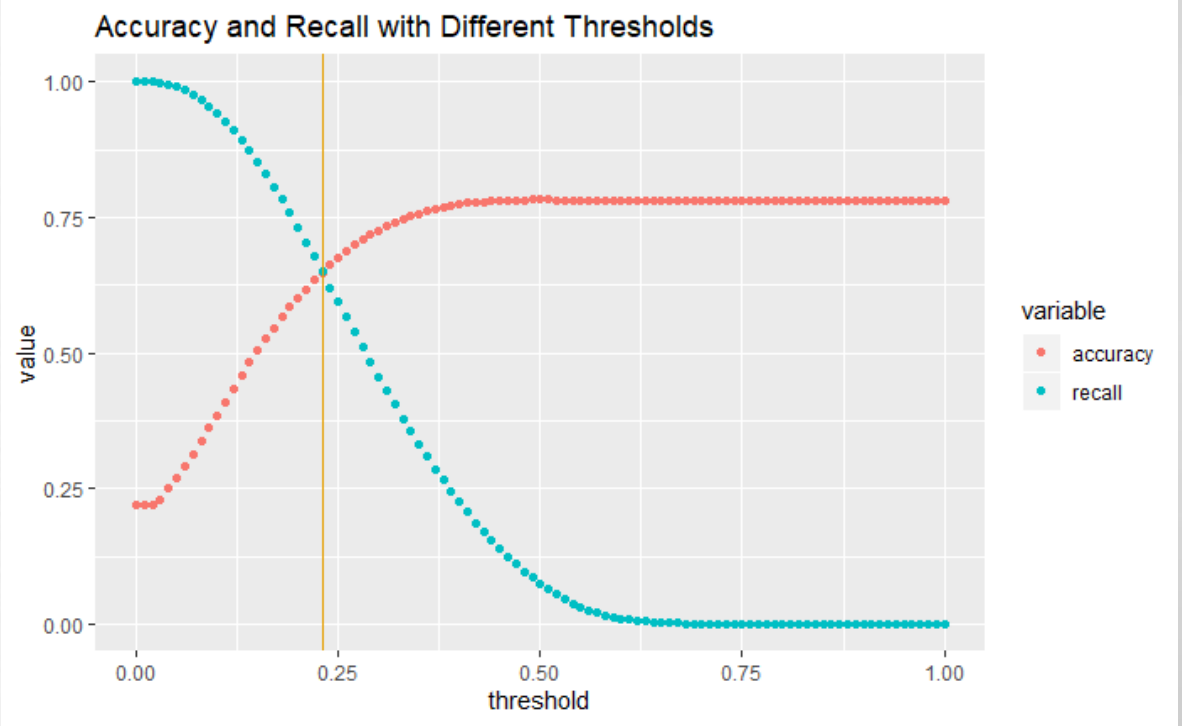
→ improved in model 2



Baseline Model

Tune Model

- Validation size: 25%
- Threshold: 0.23
- Accuracy: 65%
- Detection rate: 65%



Baseline Model

Test Model

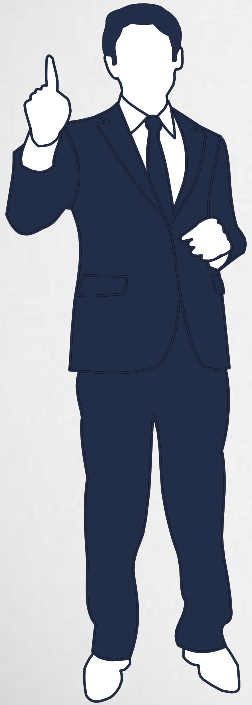
- Test size: 25%

Among 100 truly defaulted people, we detected 65.

Among 100 people we predicted as default, only 35 are truly defaulted.

Baseline Model

What does the model tell us?



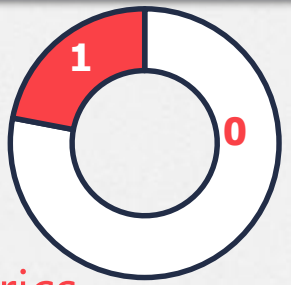
If a person ……

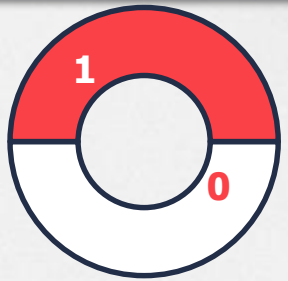
Borrows $**15,000** for **5years**
Has **2-year** work experience
verified annual income $**100,000**

…
His likelihood of default is **33.74% > 23%.**
Based on our model, we will NOT lend him any loan.

Model2: Logistic regression by one-hot encoding

- Balanced sample:





- Transform factor to numerics

[Morgage, Rent, Unkown]	**One-hot encoding**	[0,1,0]
Income	**Standardize**	(Income-mean)/sd
Credit grade (A,B,C…)	**Ordinal encoding**	[1,2,3…]

- Create a new feature: monthly income/monthly debt
- Tune the parameters (penalty='l1' to prevent overfitting)

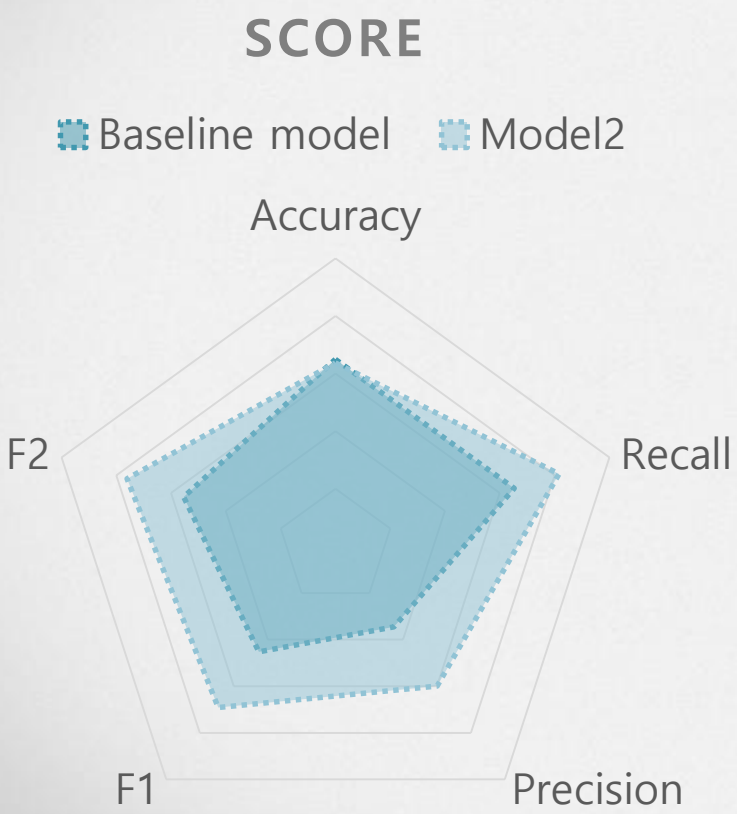
Model2 outperformed Baseline

Probability < **0.4**

Probability > **0.4**

among 100 defaulted people, we detected 81

Model2 outperformed Baseline



	Baseline model	Model2
Accuracy	0.65	0.64
Recall	0.65	0.82
Precision	0.35	0.60
F1	0.45	0.69
F2	0.55	0.76

* on test data

Deep Learning is Sexy ?

Model3:Deep Learning Model

2 hidden layers
first: 15 nodes
second: 5 nodes

Result is not exciting

Accuracy: 0.6581 Detection: 0.663

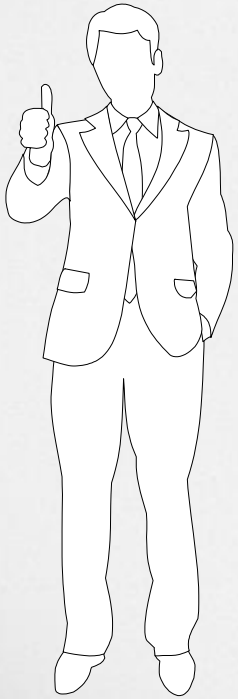
Not better than model 2

03

Result and Conclusion

Select Logistic Regression model as final model

Select model 2 as final model



What is the ranking of key features?

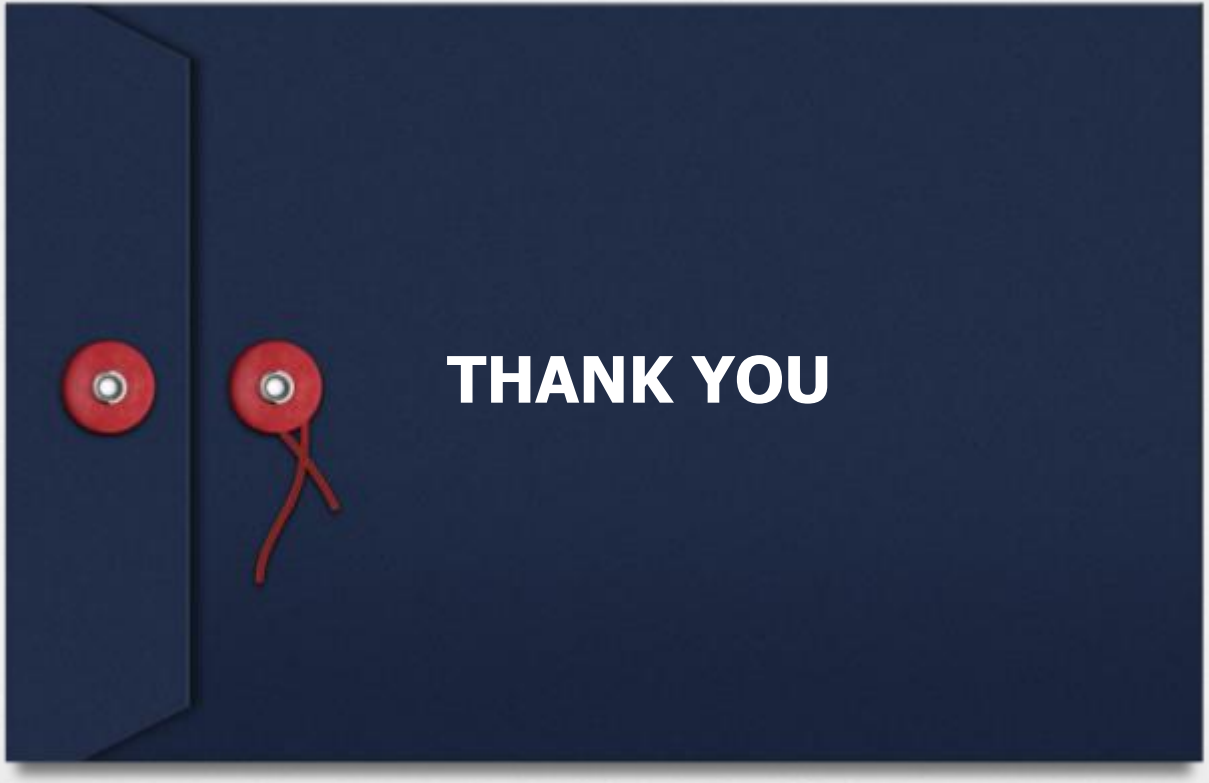
Interest rate
Debt-covered ratio
Sub-grade
Annual income
Total-account
Loan amount
Percentage never delinquent
Employ length
Term
(By Random Forest Classifier)

Conclusion

Build a logistic regression model to predict the default rate.
Although the accuracy is 60%, the detection rate (80%) is high.

We found the most important features that determine a person's default probability. Helpful to manually review applicants.

Collect more features that can measure the ability to pay.
(e.g. past spending)
Apply more advanced model to improve accuracy.(Xgboost)


THANK YOU