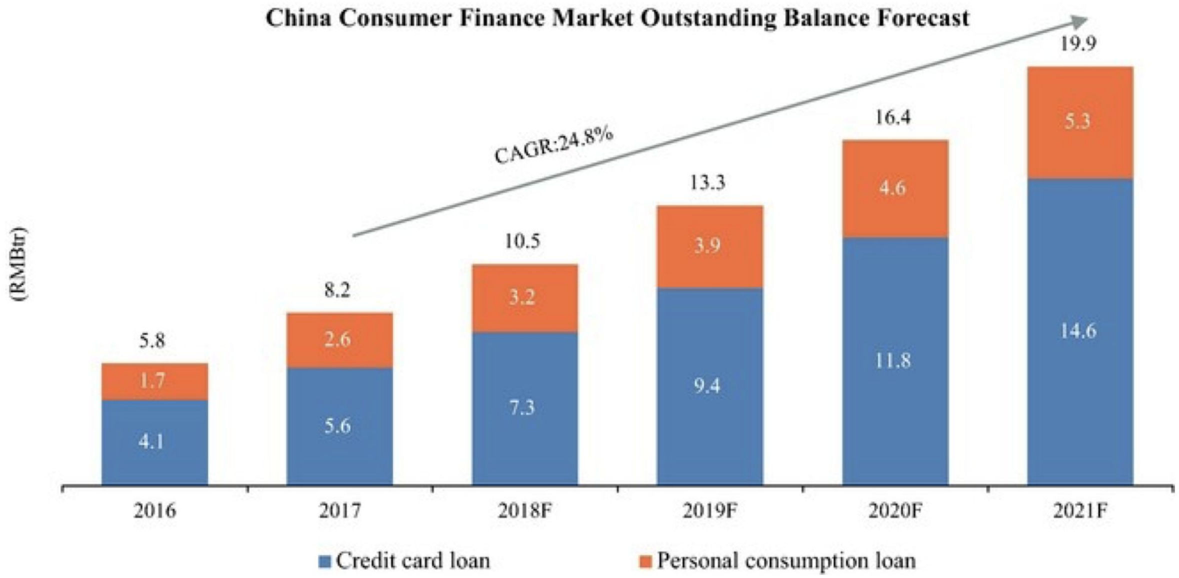
Client Presentation

Team A

小赢科技

China's Consumer Finance Market Industry Trend



- ❏ Broad market
- ❏ High growth rate (CAGR 24.8%)
- ❏ Large potential users
- ❏ Strict Regulation

Data source: Oliver Wyman Report

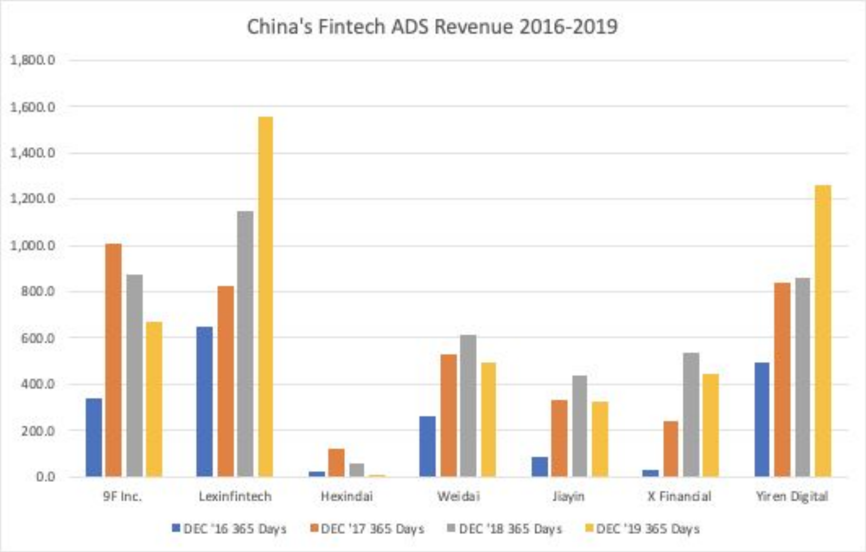
China’s Internet Finance Policy

China said all existing peer-to-peer (P2P) lending platforms ***must become small loan providers within two years****, a notice seen by Reuters on Wednesday showed, the latest official edict aimed at curbing the once-booming industry.*

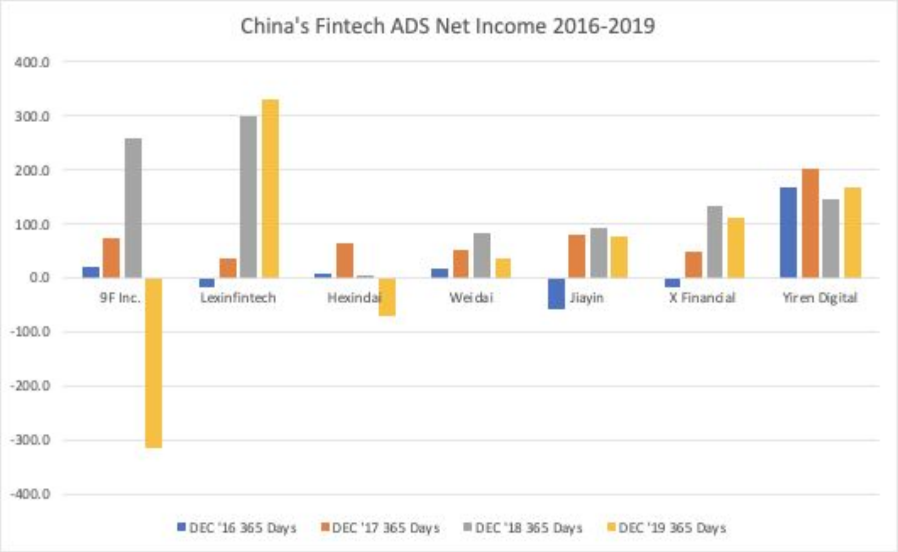
--Reuters, Nov. 28th 2019

Source: https://www.reuters.com/article/us-china-p2p-idUSKBN1Y2039

Most of China’s Fintech ADS Had Worse Financial Performance in 2019

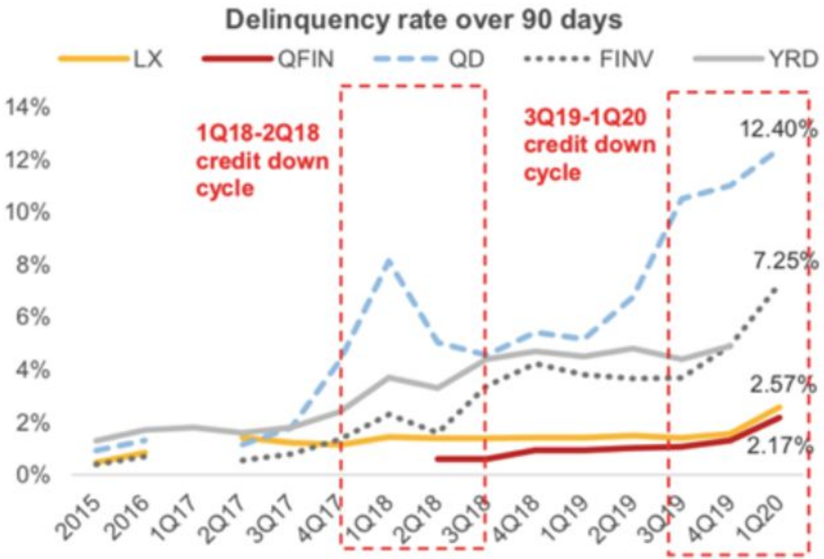


Source: Factset. All figures in millions of U.S. Dollar.

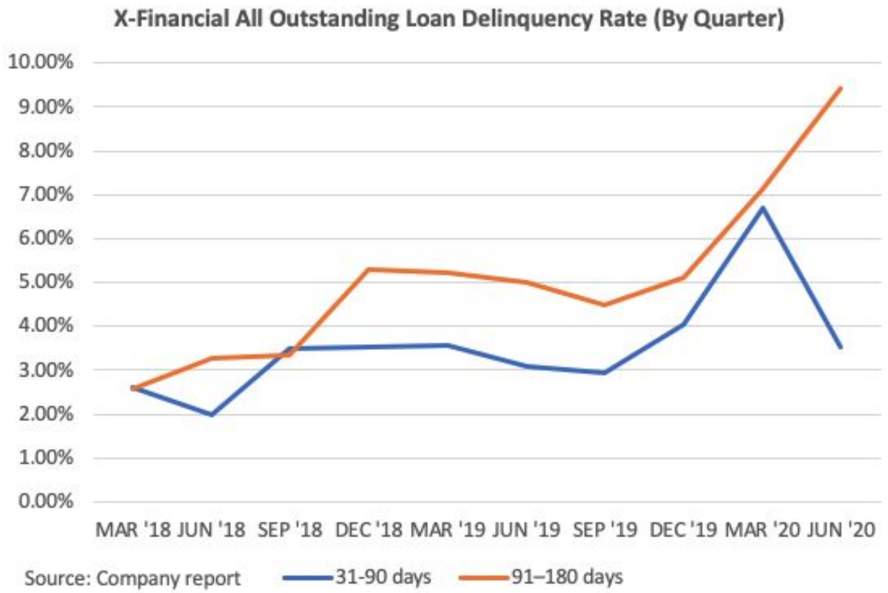


Source: Factset. All figures in millions of U.S. Dollar.

China's Fintech ADS Loan Delinquency Rates Were Up During 2019-2020 Fiscal Year



Source:UBS



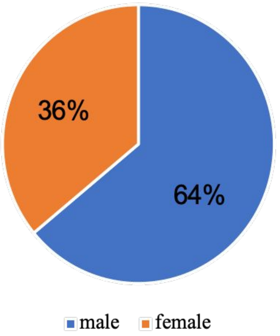
Source: Company report

User Portrait: Find growth points

Users

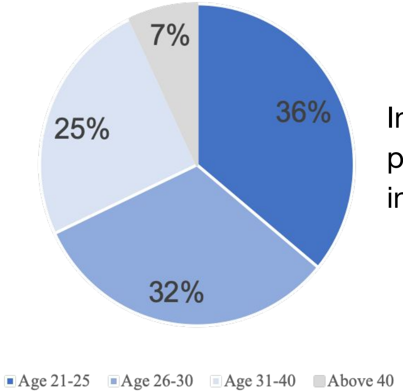
More than **47** million

Gender Distribution



Design Products specialized in women

Age Distribution



Increase the penetration rate in aged group

Source: company report

Research Methodology & Findings

Methodology: Annual Report, Research Report, Interview, Observations

Recent Findings：

- ❏ Our target customers are mainly male and young people.
- ❏ Revenue and profits continue to decline.
- ❏ The management team has changed recently. (CFO)
- ❏ X-Financial has changed the conversion ratio of ADS (1:2 vs 1:6)

Insights：

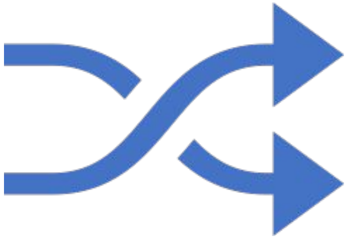
- ❏ Increase the penetration rate in aged group and female.
- ❏ The new ADS conversion rate may increase the stock price in order to prevent delisting.

Problem Statement

- ❑ To protect itself from borrower defaults and consequent losses, X Financial should be able to assess credit risk associated with each borrower as precisely as possible.
- ❑ During Covid-19, the delinquency rate continues increasing, which definitely affects the current financial situation of X Financial. The company is facing the risk of delisting since its stock price is below $1 and still showing a downturn trend.
- ❑ In this situation, there is every reason for us to build feasible models to estimate credit risk of each applicants, allowing the company to take corresponding strategies.
- ❑ Moreover, the company would utilize our customer-based model to rank the risk level of future credit card applicants to better control the risks.

Objective: Reducing the default rate

Default Rate



Build feasible models
To classify clients

Data Description

Dataset Overview

01 Application

Data about loan details and applicants' identity characteristics

02 Account

Data about applicants' account (e.g. number of loans, family relationship)

03 Inquiry

Data about credit history of applicants. From the credit bureau and third-party credit center.

04 Internal score

Scores calculated according to the applicant's address, id and other information.

05 Mobile info

Data about the applicant's mobile phones

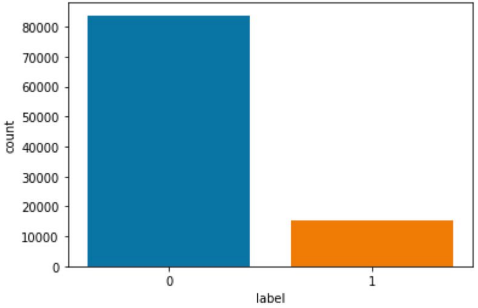
06 Pty3rd

Data provided by a third-party organization.

Exploratory Data Analysis

Target Variable

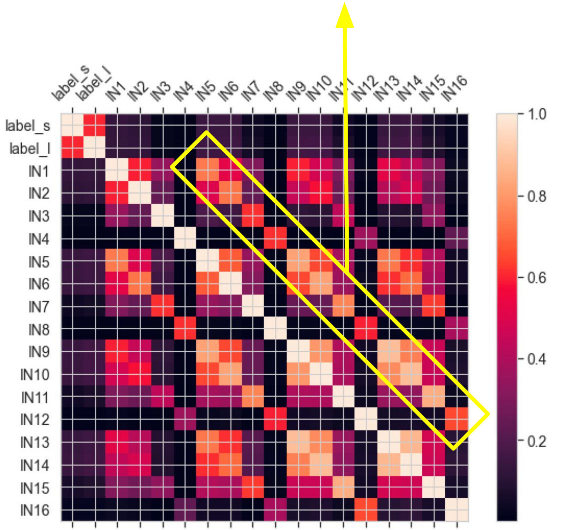
long term default label - imbalanced
(0-not default, 1-default)



	Count	Percent
0	83,839	84.62%
1	15,233	15.38%

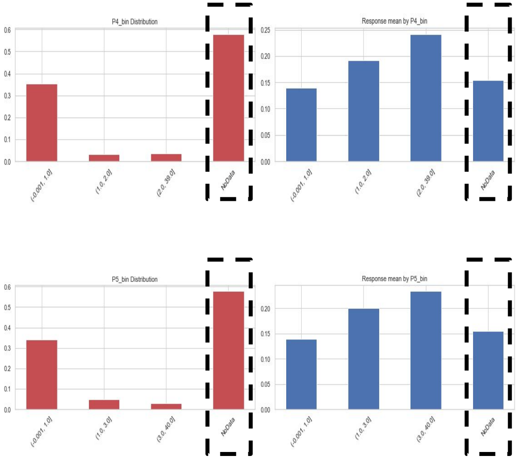
Correlations

Some variables are highly correlated.
For example, IN1&IN2, IN1&IN5, IN5&IN9



Distribution

Specific data bins have high default rate.
NoData group contains information.



Data Cleaning

- ❏ Merge & drop duplicate rows
- ❏ Drop columns with 100% NAs
- ❏ Create notnull features to represent notnull records
- ❏ Impute NAs with mode
  (also tried 0, mean, median)

Feature Engineering

- ❏ Bin continuous variables
- ❏ Encode categorical variables
  One-hot & Label Encoding
- ❏ Create interaction features
- ❏ Remove identical columns

99,072 rows, 343 features

244	18	66	15
Basic Variables	Notnull Variables	Binned Features	Interaction Features

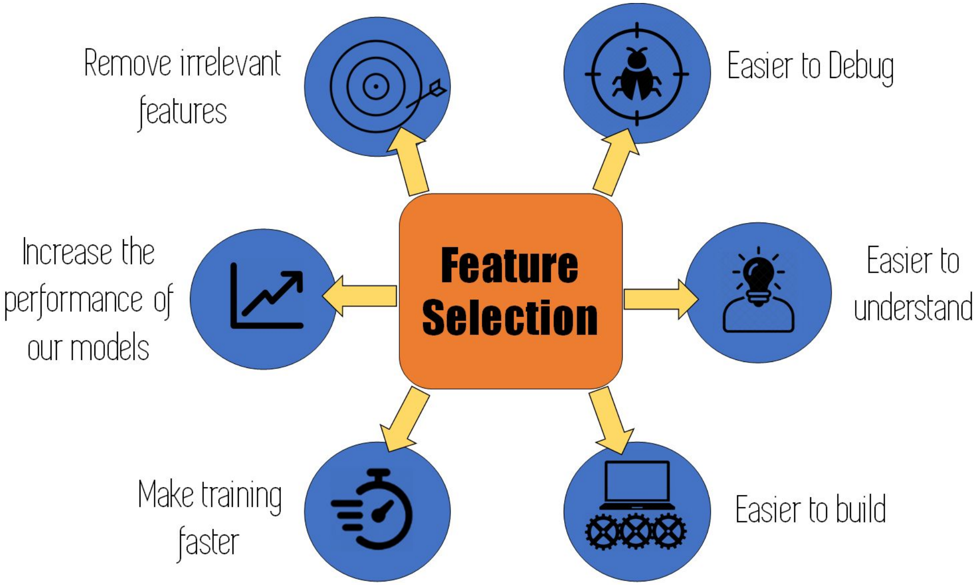
Feature Selection

Why Feature Selection

Sometimes, less is better!

- ❑ Train faster
- ❑ Reduce the risk of overfitting
- ❑ Original features + New features

➡ 68 selected features (Lasso)



Feature Selection

Version	Handling missing values	# of Features selected Lasso model
V3_1	No imputation	57
V3_2	Imputation with zero	73
V3_3	Imputation with mean	72
V3_4	Imputation with median	63
V3_5	Imputation with mode	60
V4_1	Imputation with mode (treat missing differently depending on features types)	69
V4_2	Imputation with mode (imputing first and binning)	64
V4_3	Imputation with mode (including polynomial features)	68
V5_xy	Imputation with mode	68

Modeling

Basic Models

-Basic Models include logistic regression, GBM, RF, AdaBoost and Extremely Randomized Tree

-Tune Models on basic universe and selected features

-GridSearch (cv=2) inside Stratified 5-fold cv

-Use the best models to run on oot dataset

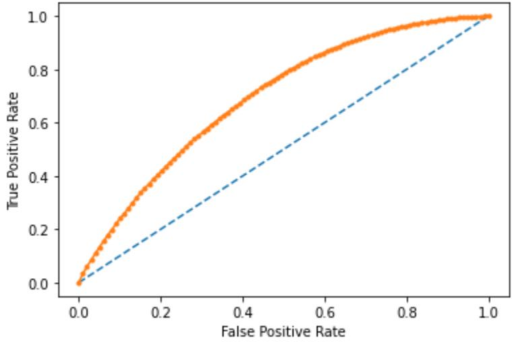
Model Performances

Dataset	Basic Universe	Selected Features	out-of-time
Metrics	ROC AUC/PR AUC	ROC AUC/PR AUC	ROC AUC/PR AUC
Gradient Boosting	0.696/0.266	0.694/0.265	0.621/0.080
Random Forest	0.686/0.256	0.687/0.258	0.617/0.082

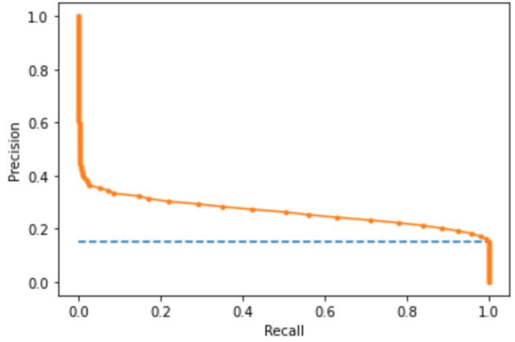
Performance drop for out-of-time dataset, possibly because out-of-time dataset (6%) has lower default rate than the original dataset (15%)

Model Performances

GBM on selected features:

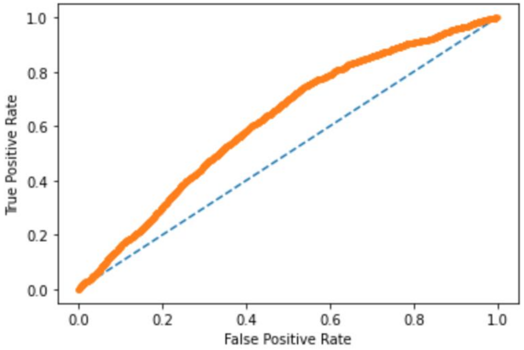


ROC AUC=0.694

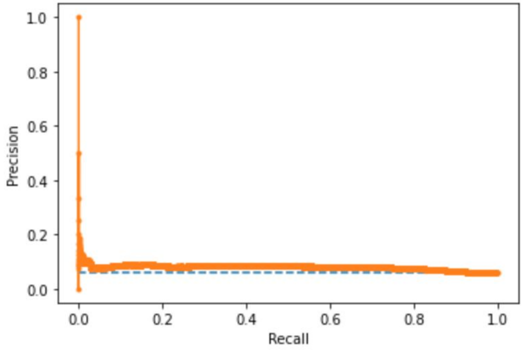


PR AUC=0.265

GBM on out-of-time dataset:



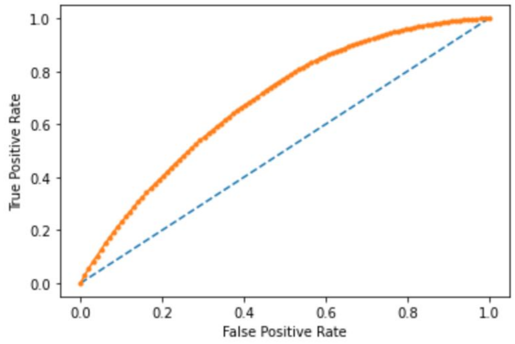
ROC AUC=0.621



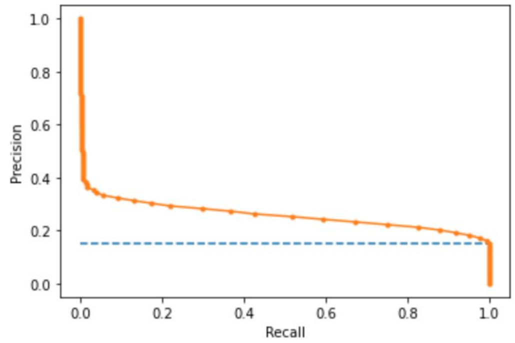
PR AUC=0.080

Model Performances

RF on selected features:

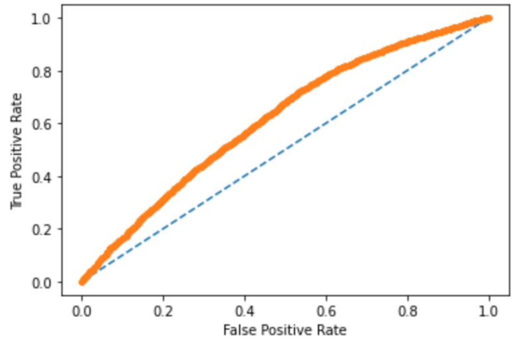


ROC AUC=0.687

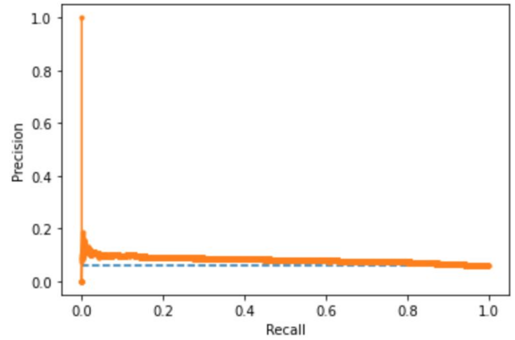


PR AUC=0.258

RF on out-of-time dataset:



ROC AUC=0.617



PR AUC=0.082

Models with Resampling

-GBM and RF with oversampling and undersampling

-Tune Models on basic universe and selected features

-GridSearch (cv=2) inside Stratified 5-fold cv

-Use the best models to run on oot dataset

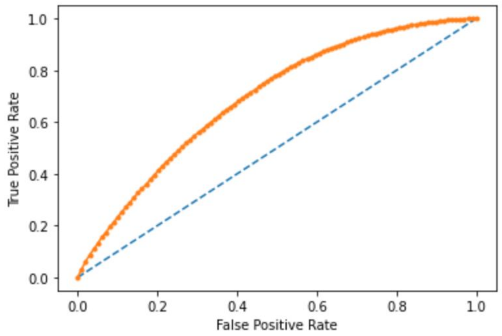
Model Performances

Dataset	Basic Universe	Selected Features	out-of-time
Metrics	ROC AUC/PR AUC	ROC AUC/PR AUC	ROC AUC/PR AUC
GBM oversampling	0.695/0.265	0.691/0.263	0.635/0.091
GBM undersampling	0.695/0.264	0.693/0.263	0.632/0.088
RF oversampling	0.687/0.258	0.682/0.248	0.624/0.085
RF undersampling	0.687/0.257	0.688/0.258	0.629/0.087

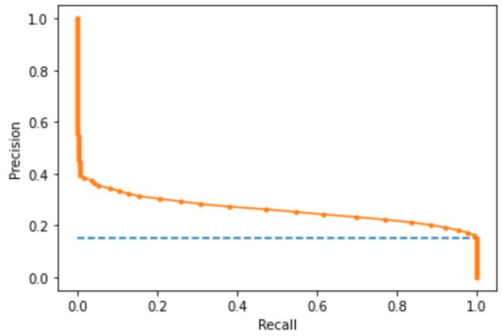
Though performance on basic universe and selected features don't improve, performance improved on out-of-time dataset.

Model Performances

GBM with oversampling on selected features:

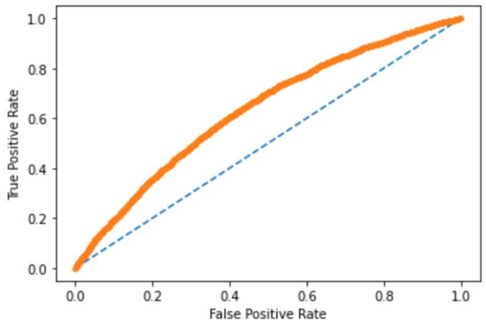


ROC AUC=0.691

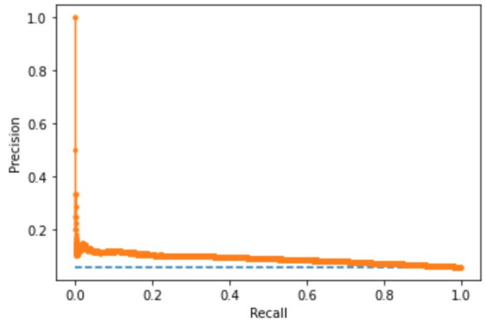


PR AUC=0.263

GBM with oversampling on out-of-time dataset:



ROC AUC=0.635



PR AUC=0.091

Ensemble Models

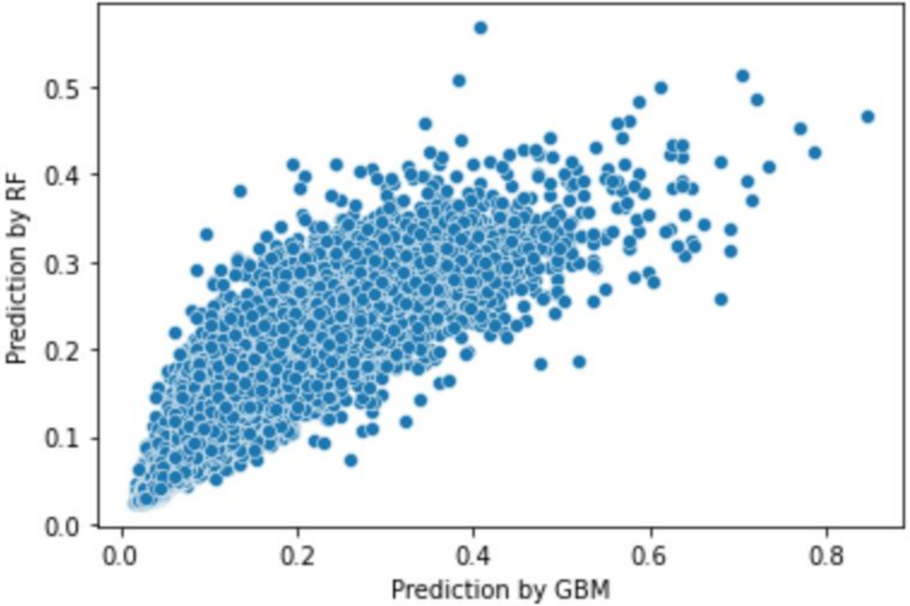
-Ensemble the basic models with best performances

-Models include:

1. Voting Model based on GBM and RF
2. Stacking Model based on GBM and RF

Validate Stacking Model

Scatterplot of Predictions by GBM and RF on Selected Features



There are differences between predictions by GBM and RF, but the differences are not very significant.

Model Performances

Dataset	Basic Universe	Selected Features	oot
Metrics	ROC AUC/PR AUC	ROC AUC/PR AUC	ROC AUC/PR AUC
Voting on GBM and RF	0.693/0.264	0.694/0.265	0.631/0.087
Stacking on GBM and RF	0.694/0.263	0.693/0.264	0.631/0.087

Performances didn’t improve on the 3 datasets.

Neural Networks

-Scale data with MinMaxScaler()

-**Structure:**

-5 hidden layers

-Number of neurons:[80,30,20,10,5]

-activation=’relu’

-optimizer=’adam’

-loss=’binary_crossentropy’

-metric= ‘accuracy’

-epoch=10

-batch_size=100

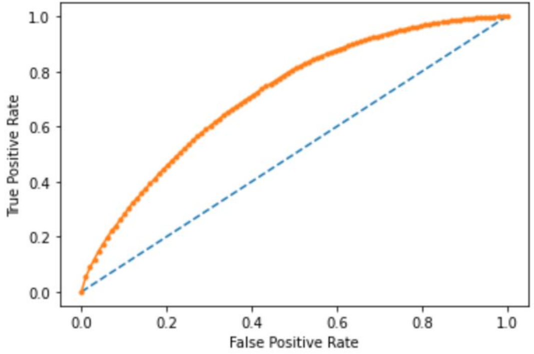
Model Performances

Dataset	Basic Universe	Selected Features	out-of-time
Metrics	ROC AUC/PR AUC	ROC AUC/PR AUC	ROC AUC/PR AUC
NN	0.723/0.315	0.716/0.308	0.572/0.074

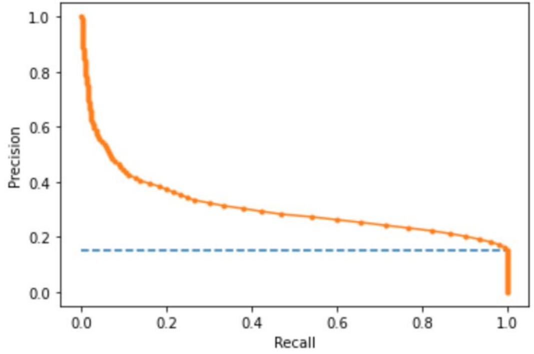
Though performance on basic universe and selected features improve with neural network, performance on out-of-time dataset drops more.

Model Performances

Neural Networks on selected features:

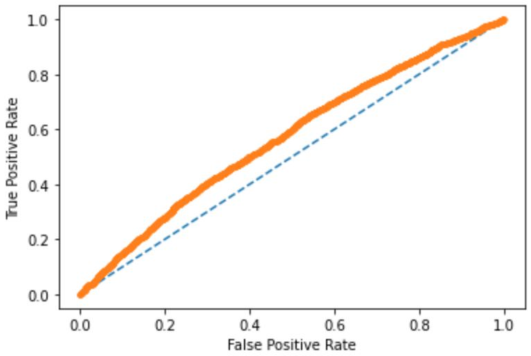


ROC AUC=0.716

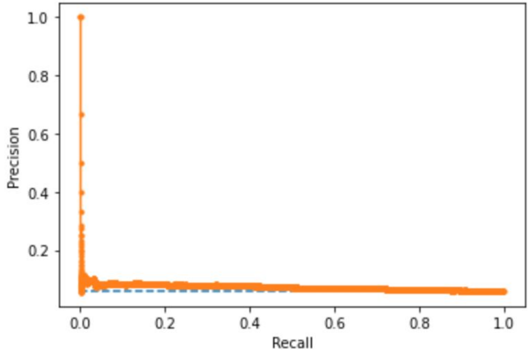


PR AUC=0.308

Neural Networks on out-of-time dataset:



ROC AUC=0.572



PR AUC=0.074

Model Performance by Demographic Groups

-Group by gender (male and female)

-Group by age (20-26, 26-30, 31-51)

-Use the group from selected features to tune and train the model and apply it to the group in oot

Dataset	**male_oot**	**female_oot**
Performance	0.617/0.084	0.643/0.083

Dataset	**20_26_oot**	**27_30_oot**	**31_51_oot**
Performance	0.589/0.096	0.641/0.087	0.606/0.07

The model has relatively better prediction power for female clients and clients between 27 and 30 years old.

Model Interpretation

-Use SHAP value to interpret the model with the best performance on out-of-time dataset (GBM with oversampling)

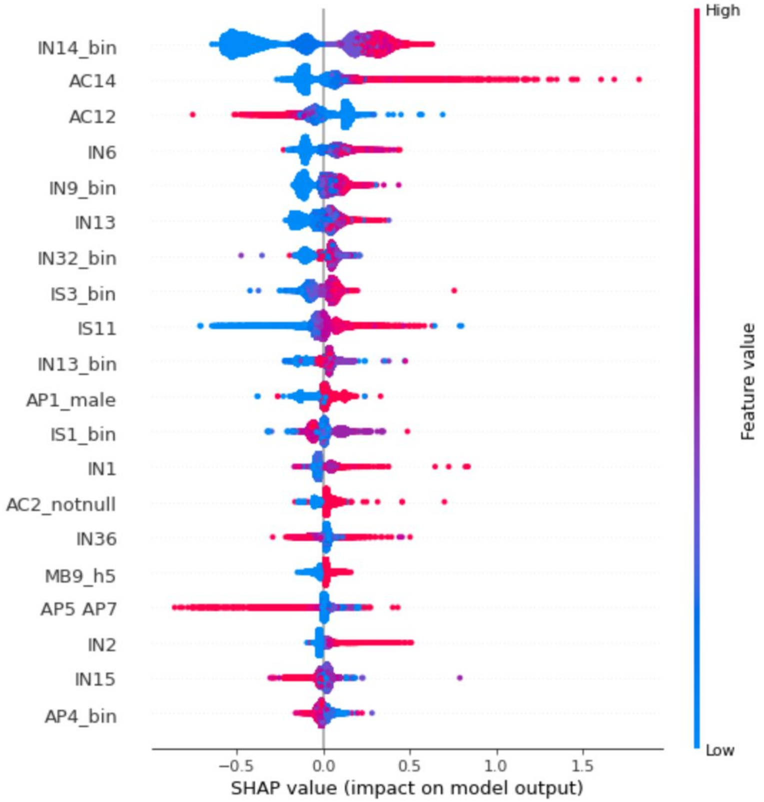
Why SHAP value?

-SHAP value can give the average of the marginal contributions across all variables permutations and help interpret any models.

-SHAP value can show the correlations between X and y

-SHAP value can help interpret a single observation

Feature Importance



-Features of top importance are from ***Inquiry, Account & Internal Score*** tables

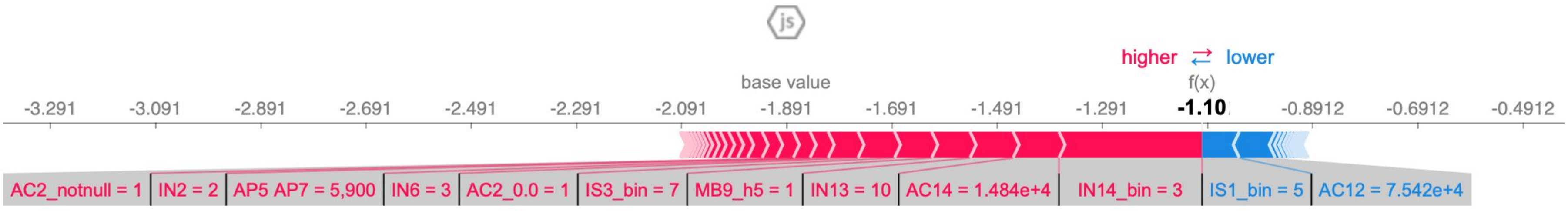
-Number of inquiry is positively correlated with default probability

-Loan balance (AC14) is positively correlated with default probability

-Male clients are more likely to default (AP1_male)

-Younger clients are more likely to default (AP4_bin)

Single Observation Interpretation



-Features in red contribute to higher default probability, while features in blue contribute to lower default probability

-Base value is the average prediction of all observations from the model (raw prediction)

-f(x) is the prediction for this single observation, it has higher default probability than average

-This client can improve their profile by lowering AC14 (loan balance)

Potential Obstacles

-Data Drift & Model Decay

E.g. New mobile phone types; Lower default rate after implementing the model

-> The model needs to be updated constantly

-Delayed Model Monitor

We can only get actual long term default status after years, and by the time the model might have been updated

-> hard to monitor model performance but can monitor input variables

Recommendations

-Build separate models for different demographic groups

-Adjust marketing target to clients with lower default probability (female, older clients)

-Provide credit improvement suggestions based on SHAP value to clients whose applications get disapproved

-Monitor input variable and update model constantly to avoid model decay

Thanks for listening!

Do you have any questions?