Team A

Professor Wayne Huang

5900 Capstone (X Financial)

28 September, 2019

X Financial Capstone Project Research Proposal

INTRODUCTION

X Financial, as a leading fintech company in China, focused on closing the gap between huge loan need from individuals and small businesses with those unmet demand from banks. Those individuals or business owners might not be able to get approved by traditional banks due to the lack of long term credit history due to the unsound credit system in China. It built its competitive advantage with a big data driven risk control system WinSAFE, which incorporate big data from various sources such as mobile info, credit bureau, call detail, etc. to evaluate borrowers' overall creditworthiness. In addition, its partnership with ZhongAn Online P&C Insurance Co., Ltd offered X Financial with extensive credit information as well as risk control technologies, which enhanced investor confidence and allowed the company to survive during the collapse of P2P (Peer-to-peer lending) industry in 2018. However, there are still barriers for the company to overcome in order to attract more customers and diversified funding base to support its sustainable growth.

FOCUS AND STRATEGY

Aligning with the company's mission "to utilize internet technology to build the leading personal finance company in China", X Financial currently focuses its strategy on offering more

diversified products, increasing brand awareness, acquiring new customers as well as retaining current users. A strong risk management and technology foundation will definitely strengthen investors' confidence and therefore help the company scale its funding sources.

CHALLENGE

The major challenge X-Financial faces is the heavy financial burden from unpaid customer loan and the increasing threats from competitors. Therefore, it needs a new and better credit score model with high ranking power to reject the loan applicants with highest risks. With a credit score model of better performance, X-Financial will be able to make decisions on credit application more effectively, segment the applicants based on their expected risks and adjust the approved amount and interest rate according to the results of our model, and thus control their risk exposure.

RESEARCH QUESTION AND HYPOTHESES

Given a dataset including applicants' demographic information, credit information, call data and additional information provided by third-party organizations, our study aims to develop a credit score model that determines whether an applicant can be classified as good or bad, and ranks their creditworthiness. The primary research question is that ***in terms of accuracy, stability and interpretability, what kind of big data algorithm has the best performance of identifying defaults***?

For research hypotheses, the variables filtered out from the feature selection process will each become a hypothesis of the analysis by itself. For instance, if there are 25 variables selected

to be used in the model, then there will be 25 hypotheses. However, if multicollinearity is detected during data preparation, we may group linearly related variables into categories. Further, since the goal of this analysis is to create an enhanced credit scoring model with robust ranking capability to reject loan applicants with the highest risks based on their various information, it would be rational to set the null and alternative hypothesis as the following:

Ho = an independent variable does not have strong predictive power over the dependent variable "bad30"

Ha = an independent variable has strong predictive power over the dependent variable "bad30"

To give an example, if YR_AGE is selected as a feature that will be used in our model, then the null hypothesis would state that YR_AGE is a strong factor in predicting the default risk of a borrower, whereas the alternative hypothesis would state it is not.

METHODOLOGY FOR RESEARCH

Credit risk modeling, which helps identify an individual's credit risk, is a prevalent research topic in the financial industry, which is of concern to a variety of stakeholders: institutions, consumers and regulators. For our project, the main purpose is to construct a credit score model with good performance and high accuracy to suggest rejections to the loan applicants with highest risks and propose acceptances to those with lower risks. A good model is based on both knowledge of the industry and understanding of models. Thus, our project requires industry research on business aspects (e.g. the loan application process, customer profiles and etc.) and modeling aspects (e.g. feature engineering methods, modeling methods, statistical

methods and etc.). Extracting information from secondary data is the main methodology for conducting the research. The secondary data include training data the client offered to us, resources on industry and models, and etc.. The first part of our research is to do industry research to gain a thorough understanding of business, which can help us gain more insights on data cleaning and model building. We propose that we can get the information from our client, major players in loan industry and online resources. The second part of our research is to construct our model, which requires steps of feature engineering, model building, parameter tuning and model selection. For this part, we plan to learn from industry benchmarks and academic resources in this field. The final part of our research is to test our model's performance and accuracy. For a more detailed explanation on our research methodology, this research proposal lists methodological steps as belows:

METHODOLOGICAL STEPS

Step 1 Formulate research problem

With the help of the client's clarification on variables and models and extensive, we need to understand the business problem thoroughly and rephrased it from an analytical point of view. We will use the predictive model to help the client reject the loan applicants with the highest risk and use the ranking power of the model to help them determine the approved amount and interest rate based on the different risks of the loan applicants.

Step 2 Data preparation and exploration

First, we will process data cleaning by checking and figuring out the data that are not complete, accurate or relevant. Then we will modify, replace or delete the dirty data parts based

on their characteristics. Then we will explore the statistical features of the data, including mean, median, total, variance etc. Based on the distribution of data points, we can consider discretization normalized, logarithmic or exponential distribution conversion for further analysis.

Step 3 Feature selection

In order to avoid overfitting and reducing and also decrease computational cost, we should do feature selection before modeling. We will adopt four feature selection techniques.

1) Univariate Selection

Univariate Selection can select those features that have the strongest relationship with the output variable. We conducted univariate analysis to find out the correlation between independent variables and the target variable. The result shows that there are 27 variables have a p-value less than 0.05 in our analysis.

2) Feature Importance

Feature Importance can give us a score for each variable of our data. The higher the score is, the more important or relevant is the feature towards the output variable. After calculating the feature importance scores, we can reduce the feature set.

3) Correlation Matrix with Heatmap

Correlation Matrix with Heatmap can state how the features are related to each other and the target variable. We will plot heatmap of correlated features using the seaborn library to identify which features are most related to the target variable and check whether the correlation is positive or negative.

4) Boruta Feature Selection

Boruta Feature Selection is a feature selection method used in high-dimensional data. The Boruta algorithm is based on random forest and its goal is to select all feature sets related to the dependent variable, rather than selecting a feature set that minimizes the model cost function for a particular model.

Step 4 Model selection:

1) Logistic regression

Logistic regression is the most widely used method in credit risk research. Even though the benchmark provided uses this method and we might not be able to improve it a lot, we still would like to test out with different variables and acquire a deeper understanding of the data, which could be beneficial for future analysis.

2) Random forest

Random forest is a supervised learning approach that corrects for decision trees' habit of overfitting to the training set. We would also like to test this approach in our project.

3) Boosting methods (including XGBoost, Gradient Boosting Decision Tree)

Boosting is a supervised learning method that uses a set of weak learners to create a strong learner. Gradient Boosting Decision Tree uses Newton-Raphson method in every iteration to reduce residual values. XGBoost is considered to be a faster boosting method with better performances. Chen, Chen and Li (2017) applied GBDT and XGBoost with credit related data from Qian Hai credit information enterprise and it was proven to be practical. That is why we would also like to apply those methods in our project.

Step 5 Model evaluation & Answer to research question:

1) KS test

The the Kolmogorov–Smirnov test (KS) is mainly to evaluate how well our predictive model is able to discriminate between default and non-default users. It will compare the cumulative distribution of default and non-default users and plug the maximum difference into KS probability function to calculate the probability value. The lower the probability value is, the less similar the two distributions are. The higher or more close to 1 the probability value is, the more similar the two distributions are.

2) PIS test

The population stability index (PSI) measures the score distribution differences between the model development sample and the test sample, which is the most common indicator of model stability assessment. Generally, there is a common rule for reference: if PSI is less than 10%, it means that the model stability is high. If PSI is between 10% and 25%, then the current sample should be investigated further for reasons of this high PSI. If PSI is beyond 25%, we need to develop a new model on a more recent sample.

In conclusion, we would like to try out different feature selection methods to increase the interpretability as well as accuracy. With the numbers we get from KS and PIS tests, we will be able to find out the most accurate and stable model. The model with the best performance will be our answer to the research question.

DATA DESCRIPTION

1) General description

The dataset has 79939 observations and 1808 variables. To be more specific, there are 8 application variables, 7 of which are numeric, the rest one is datetime. There are 6 mobile info

variables, 2 of which are categorical, the rest are numeric. There are also 21 credit bureau variables, 16 credit center variables, 277 call detail variables and 1480 pty3rd party variables, all of which are numeric. The target variable is ‘bad30’.

2) Data quality issues

a. Third party data

Of 1480 third party variables, more than 600 of them largely consist of 0. To make it worse, we are not able to find out the explanation of these data. In other words, third party variables has low correlation with the target variable and we might not have enough rationale to add them to our model manually. We will continue exploring solutions to this problem in our future analysis.

b. Inaccurate data

The dataset has already been screened. It now has 19% default risk, which is higher than the real value. This means the final model we get might need to be adjusted before the company puts it in practice.

Work Cited

"Financial Reports." *X Financial-Welcome*, http://ir.xiaoyinggroup.com/index.php?s=120.

"X Financial." *SEC.gov*, https://www.sec.gov/Archives/edgar/data/1725033/000104746918006156/a2236689zf-1a.htm.

Pathak, Manish. "Feature Selection in R with the Boruta R Package.", Datacamp, https://www.datacamp.com/community/tutorials/feature-selection-R-boruta. Accessed 7 March, 2018

https://en.wikipedia.org/wiki/Random_forest

Yaofei Chen, Yijie Chen, Ming Li. 'Credit score prediction model based onXGBoost', 2017. *基于XGBoost的信用评分预测模型中国知网*, http://kns.cnki.net/kcms/detail/detail.aspx?filename=GTJX201712001008&dbcode=CPFD&dbname=CPFDTEMP&v=.

SEC.gov,https://www.sec.gov/Archives/edgar/data/1725033/000104746918006156/a2236689zf-1a.htm#targetText=Our mission is to utilize,personal finance company China.&targetText=Notes:,facilitated during the relevant period. http://ir.xiaoyinggroup.com/.