Team A

Professor Wayne HuangT

5900 Capstone (X Financial)

November 23, 2019

X Financial Capstone Project: Results and Recommendations

RESULTS

Summary of Research Method

Our research method can be broken down into three steps. The first step we took was **data exploration and preparation**, in which we gained a basic understanding of the dataset, identified data problems and cleansed the data for future analysis. In this step, we learnt that there are 1,808 independent variables in 6 categories with the "Pty3rd" category taking up 81% of the total, which posted a great challenge as the variables in this category do not have descriptions. We also learnt from the dependent variable that 19% of the observations is "bad30." In terms of data processing, we performed various feature engineering techniques such as winsorizing outliers, binning continuous variables and encoding categorical variables.

The second step was **feature selection** where we tested various methods including univariate analysis, feature importance analysis, correlation analysis using heatmap and Boruta feature selection. The most effective approach was Boruta selection, which yielded 166 significant features. For logistic regression, we manually inspected the distribution of these 166 variables and performed binning to check their p-values, which further narrowed down the number of features to 26. The manual inspection and binning were very challenging in particular due to the sheer number of features selected by Boruta.

The final step was **model selection and evaluation**. We tested various methods including logistic regression and boosting methods. To find the model with the best performance, we used three metrics for assessment, which are K-S test result, AUC-ROC score and lift. We found that logistic regression generated better results with regards to lift and XGBoost regression showed the better result with regard to ROC score. For XGboost regression, we also found that the model would generate better ROC score without feature selection (i.e. using all variables) than with feature selection (i.e. using only Boruta selected variables), but would yield better lift with feature selection.

Model Performance

1. Logistic Regression

We built a Logistic Regression model with 26 selected variables. Firstly, we utilized results from Boruta feature selection with 166 variables, then we conducted pairwise correlation analysis to drop abundant variables to only include one variable from each high correlation group. Finally, we applied univariate analysis based on glm to narrow the number of features to just 26. The AUC-ROC score for this model is 0.6604 which is not as high as that for XGBoost models, however, the model lift reaches 177 as the highest. Furthermore, A/P scores (actual value to predicted value) are all relatively close to 100 compared to XGBoost models. The maximum K-S value is 32.03 which is better than all the K-S values from XGBoost models. Therefore, high model lifts and high K-S test statistics are indicators that our Logistic model has a good ranking power.

decile	count	actual	pred	pred_good	actual %	pred %	act_avg	model lift	A/P	KS_Stats
0	1999	668	674.0	1325.0	33.42	33.72	381.5	177.0	99.0	9.51
1	1998	585	553.0	1445.0	29.28	27.68	381.5	145.0	106.0	15.10
2	1998	504	496.0	1502.0	25.23	24.82	381.5	130.0	102.0	18.84
3	1999	437	450.0	1549.0	21.86	22.51	381.5	118.0	97.0	21.08
4	1998	423	410.0	1588.0	21.17	20.52	381.5	107.0	103.0	22.03
5	1998	385	370.0	1628.0	19.27	18.52	381.5	97.0	104.0	21.68
6	1999	304	326.0	1673.0	15.21	16.31	381.5	85.0	93.0	19.89
7	1998	252	269.0	1729.0	12.61	13.46	381.5	71.0	94.0	16.27
8	1998	171	182.0	1816.0	8.56	9.11	381.5	48.0	94.0	9.82
9	1998	86	78.0	1920.0	4.30	3.90	381.5	20.0	110.0	-0.00

Figure 2.1.1. Lift and K-S value of Logistic model with 26 selected variables

2. XGBoost

We built XGBoost models with all variables and with 166 variables selected by Boruta. The one we applied all the variables to has an AUC of **0.67405**, which is slightly higher than **0.672031**, the AUC score of the model with 166 selected variables. However, according to the gain charts (Figure 2.2.1 and Figure 2.2.2), the model with 166 selected variables has higher model lift and maximum K-S value than the other, which indicates a better ranking power. Since the gap between two AUC scores is small and we want to ensure the ranking power of our model, we chose the XGBoost model with 166 selected variables over the one with all variables.

decile	count	actual	pred	pred_good	actual %	pred %	act_avg	model lift	A/P	KS_Stats
0	1999	715.0	652.0	1347.0	35.77	32.62	381.5	171.0	110.0	8.68
1	1998	580.0	545.0	1453.0	29.03	27.28	381.5	143.0	106.0	13.92
2	1998	502.0	495.0	1503.0	25.13	24.77	381.5	130.0	101.0	17.54
3	1999	432.0	455.0	1544.0	21.61	22.76	381.5	119.0	95.0	19.86
4	1998	423.0	417.0	1581.0	21.17	20.87	381.5	109.0	101.0	20.96
5	1998	399.0	378.0	1620.0	19.97	18.92	381.5	99.0	106.0	20.80
6	1999	290.0	333.0	1666.0	14.51	16.66	381.5	87.0	87.0	19.18
7	1998	237.0	276.0	1722.0	11.86	13.81	381.5	72.0	86.0	15.73
8	1998	156.0	191.0	1807.0	7.81	9.56	381.5	50.0	82.0	9.53
9	1998	81.0	88.0	1910.0	4.05	4.40	381.5	23.0	92.0	0.00

Figure 2.2.1. Lift and K-S value of XGBoost model with 166 selected variables

decile	count	actual	pred	pred_good	actual %	pred %	act_avg	model lift	A/P	KS_Stats
0	1999	706.0	642.0	1357.0	35.32	32.12	381.5	168.0	110.0	8.40
1	1998	592.0	538.0	1460.0	29.63	26.93	381.5	141.0	110.0	13.44
2	1998	527.0	490.0	1508.0	26.38	24.52	381.5	128.0	108.0	16.93
3	1999	430.0	452.0	1547.0	21.51	22.61	381.5	118.0	95.0	19.19
4	1998	432.0	415.0	1583.0	21.62	20.77	381.5	109.0	104.0	20.25
5	1998	350.0	379.0	1619.0	17.52	18.97	381.5	99.0	92.0	20.15
6	1999	319.0	335.0	1664.0	15.96	16.76	381.5	88.0	95.0	18.62
7	1998	224.0	280.0	1718.0	11.21	14.01	381.5	73.0	80.0	15.31
8	1998	152.0	199.0	1799.0	7.61	9.96	381.5	52.0	76.0	9.39
9	1998	83.0	92.0	1906.0	4.15	4.60	381.5	24.0	90.0	-0.00

Figure 2.2.2. Lift and K-S value of XGBoost model with all variables

Feature Importance

1. Logistic Regression

In the final logistic model we have 26 variables and 15 of them are third party variables. If we remove all the third party variables from the model, the roc score drops from 0.6604 to 0.6515, and the maximum lift drops from 177 to 169. In other words, the importance of third party variables is not very significant in the model.

In the final model, variables AP004, TD010, TD013, AP003 have the highest absolute values of coefficients, which suggests that they play an important role in the model.

2. XGBoost

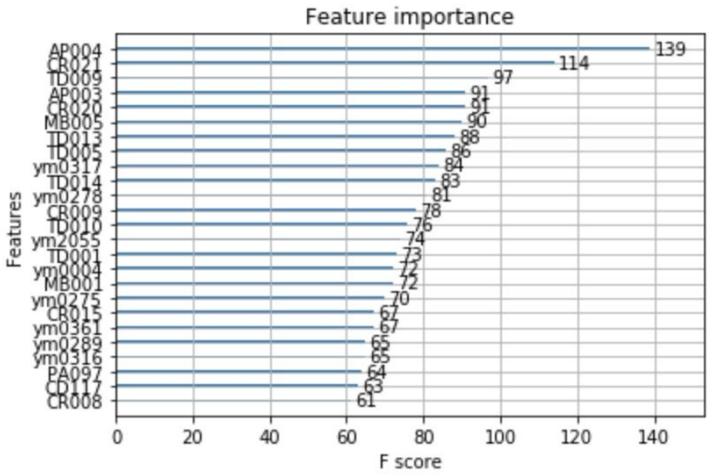


Figure 3.2.1. Feature Importance of XGBoost model selected variables

Based on the XGBoost Feature Importance Bar Chart plotted above, we can get insights on feature selection and result interpretation. Feature importance generated by XGBoost is calculated for a single decision tree by the number that each feature split point increases the performance measure, weighted by the number of observations the node takes (Brownlee, 2019). Generally, the more an attribute is used to make key decisions with decision trees, the higher its relative importance.

We have plotted 25 most important features based on our XGBoost model with best accuracy. Those features include 19 variables with descriptions and 6 third-party variables listed as below:

1	AP004: LOAN_TERM	14	ym2055
2	CR021: HIGHEST_CREDIT_CARD_LINE	15	TD001 :TD_CNT_QUERY_LAST_7Day_P2P
3	TD009: TD_CNT_QUERY_LAST_3MON_P2P	16	ym004
4	AP003: CODE_EDUCATION	17	MB001: CNT_CONTACT
5	CR020: HIGHEST_MONTHLY_LOAN_AMT	18	ym0275
6	MB005: YR_PHONE_ACTIVE	19	CR015:

			MONTH_CREDIT_CARD_MOB_MAX
7	TD013: TD_CNT_QUERY_LAST_6MON_P2P	20	ym0361
8	TD005: TD_CNT_QUERY_LAST_1MON_P2P	21	ym0289
9	ym0317	22	ym0316
10	TD014: TD_CNT_QUERY_LAST_6MON_SMALL_LOAN	23	PA097 : LEN_COLLECTION_OR_HIGH_RISK_CALLS_LAST_3MON
11	ym0278	24	CD117 : CNT_DISTINCT_OUTBOUND_CALLS_LAST_1MON
12	CR009: AMT_LOAN_TOTAL	25	CR008 : CNT_OFFLINE_QUERY_TIME_LAST_1MON
13	TD010: TD_CNT_QUERY_LAST_3MON_SMALL_LOAN		

According to the plot, the feature with the highest feature importance score is loan term; most features ranking top are those with descriptions; besides, most features are related to loan applicants behavioral data. For future model construction, we can further work on improving model accuracy by feature selection based on feature importance scores generated in scikit-learn.

Analytical Findings about Important Features

After univariate analysis of variables selected by Boruta, we have selected 26 golden features, which are significantly related to Bad30 and have distinct differences among different ranges based on logistic regression outcomes. They are listed as below:

11 Variables with Descriptions:	**15 Third-party Variables:**

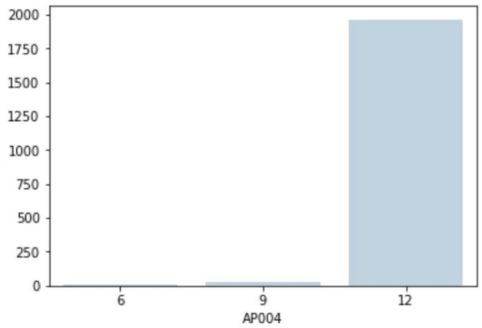
1	AP002: CODE_GENDER	1	YM0004
2	AP003: CODE_EDUCATION	2	YM0036
3	AP004: LOAN_TERM	3	YM0051
4	AP006: OS_TYPE	4	YM0052
5	MB005: YR_PHONE_ACTIVE	5	YM0153
6	CR008: CNT_OFFLINE_QUERY_TIME_LAST_1MON	6	YM0155
7	TD001: TD_CNT_QUERY_LAST_7Days_P2P	7	YM0294
8	TD010: TD_CNT_QUERY_LAST_3MON_SMALL_LOAN	8	YM0302
9	TD013: TD_CNT_QUERY_LAST_6MON_P2P	9	YM0317
10	TD016: TD_CNT_QUERY_LAST_6MON_OTHER	10	YM0329
11	CD164: CNT_SERVICE_OUTBOUND_CALLS_LAST_3MON	11	YM0361
		12	YM0366
		13	YM0403
		14	YM0922
		15	YM2055

Those golden features are predictors which can increase our models' ranking power and help generate business insights for X Financial. Based on our analysis, we have those business insights on customer analytics for X Financial from 3 dimensions.

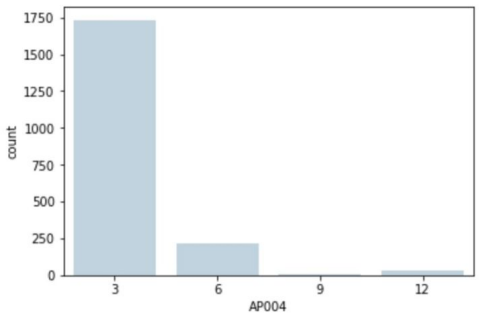
1. Loan Term

One important feature of loan application is the loan term. We produced diagrams of Loan Term Distribution of highest risk group (index0) and lowest risk group (index9). As we can see, over 95% of loan applicants in the highest risk group apply for a 12-month loan term, while most of those in the lowest risk group apply for a 3-month loan term. Thus, our insight is that loan applicants in highest risk group tend to apply longer term of loans. The reasons behind this phenomenon may be that longer-term loan applicants may have issues in personal finances.

Besides, longer-term loans are also more expensive and require higher interests, which may contribute to the increasing default rate.



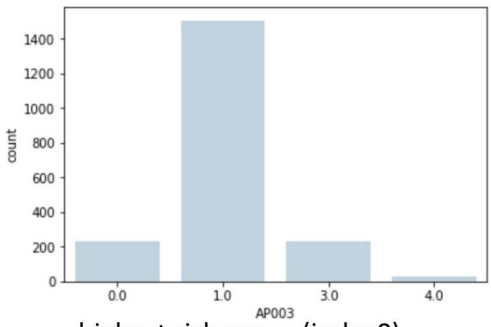
highest risk group (index0)



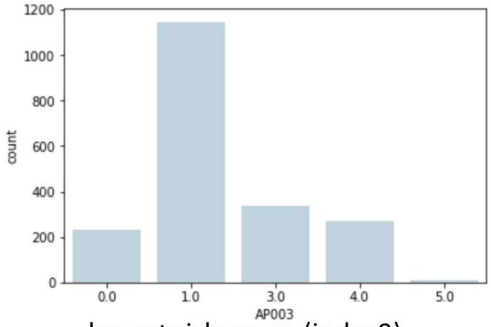
Lowest risk group (index9)

2. Education

Education is also an important feature which influences the default rate. Comparing education distribution of highest risk group (index0) with lowest risk group (index9), we can see that loan Applicants in highest risk group (index0) tend to be less educated. Since education levels are coded by numbers ranging from 0 to 6. We infer that 0 denotes the lowest education level while 6 is the highest education level. After we computed the average education level while we treat them as numbers, the average education level of all loan applicants is 1.575289. Comparing with the average education level of the whole group, the highest risk group is less educated (Index 0 average: 1.163082), and the lowest risk group is more educated (Index 9 average: 1.649149). The difference between the average education level of the whole group and that of the highest risk group is large. Then we may confirm our conclusion that less-educated customers have higher default rate. However, education may not be a significant predictor for the best applicant group.



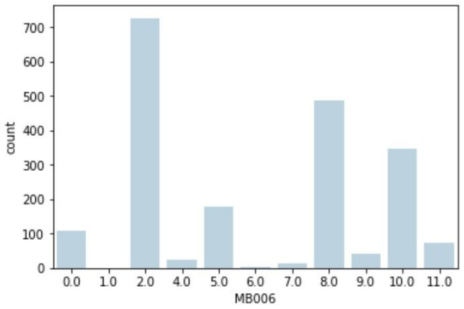
highest risk group (index0)



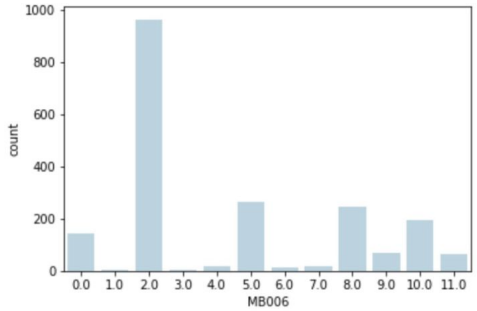
lowest risk group (index9)

3. Cell Phone Brands

Cell Phone Brands can be a good customer segmentation metric. By comparing the cell phone brands information of applicants in the highest risk group (index0) with those in the lowest risk group (index9), we find that applicants using Oppo and Vivo are riskier than others. The reason may be different cell phone brands present applicants' personal finance levels.



highest risk group (index0)



Lowest risk group (index9)

RECOMMENDATIONS

Model Selection Rationale

Financial institutions construct credit default models to determine the likelihood that a corporate or a person client may default on its credit obligations. Applying the credit predictive models will help financial institutions evaluate the customer's loan risk to determine the

corresponding approved amount, interest rate, risk limits, etc. Therefore, the model selection is a crucial step in credit risk prediction. In credit risk analysis practice, the most commonly used models are Logistic Regression, XGBoost.

Logistic regression is a typical classification algorithm that can be used to predict binary results given individual independent variables. It is particularly suitable for determining the independent variable risk of personal loan default and has unique advantages for final qualitative and quantitative research on credit default risk. The first significant advantage of the Logistic model is that it can solve the regression problem of discontinuous variables, especially for the use of the dependent variable as a categorical variable to obtain accurate results. Second, when there are both discrete and continuous variables in the variables, the prediction accuracy of logistic regression is usually higher. Third, Logistic regression can intuitively reflect the interpretability of each variable on default, and can help us give corresponding economic explanation.

XGBoost is one of the most popular machine learning algorithms. It is a comprehensive optimization of the traditional Boosting algorithm. It allows users to customize the optimization goals and evaluation criteria, making the model highly flexible. At the same time, XGBoost also implements parallel processing, which has high efficiency and low time-consuming. These advantages are ideal for accurately predicting loan defaults. Compared to the traditional statistical models, XGBoost can capture more complex nonlinear relationships between variables and improve the prediction accuracy. However, like other machine learning algorithms, XGBoost belongs to the black box model and does not intuitively reflect the interpretability of

each variable for default. In addition, XGBoost has more parameters that are more complicated to be adjusted.

In summary, XGBoost model has high classification accuracy, but the results lack robustness and interpretability. Logistic regression model is slightly inferior to the machine learning model in classification accuracy, but it has good robustness and high interpretability. Therefore, we recommend to use both Logistic Regression and XGBoost and then evaluate them by using the metrics AUC_ROC and K-S Test.

AUC - ROC curve (receiver operating characteristic curve) is a performance measurement for binary classification models at various probability thresholds. This curve plots two parameters that are True Positive Rate and False Positive Rate. AUC stands for the two-dimensional area underneath the entire ROC curve. It measures the quality of the model's predictions no matter what classification threshold is chosen.The range of AUC is from 0.5 to 1. The closer AUC is getting to 1, the better the model is.

The KS test (Kolmogorov–Smirnov) can check if two independent distributions are similar or different. It compares the cumulative distribution of default and non-default users and plug the maximum difference into KS probability function to calculate the probability value. The larger the KS value, the greater the degree to which the model can distinguish between default and non-default users.

Model Deployment

As part of the CRISP-DM method, the model deployment is the final process of integrating machine learning models into existing production environment and solve prediction

or classification problems for real business. Only if the model is productionized, it begins to generate business values. There are differences between our training data with real data since we did feature engineering as well as feature selection before running the model. In addition, the size of datasets is different, and we need to consider the production cost and computational power when deploy our model. Here are some aspects that X-Financial may take into considerations when conducting model deployment.

1. Reproducibility

Although XGBoost performs slightly better than logistic regression in this case, when considering reproducibility, it might be much easier to implement logistic regression model as one can simply extract the golden features selected by univariate analysis. We may modulate the preprocessing and feature engineering code for reproduce.

2. Timeliness

The business manager should consider how often the model should be updated with new training data considering that the customer traits may shift from time to time.

3. Cost-effectiveness

When transferring the model from Jupyter Notebook with sandbox data to real system environment or cloud with large quantity of data, one may take into consideration of cost-effectiveness. For example, advanced machine learning models may burn out one-month budget on AWS within 30 minutes.

Business Recommendations

1. Spend wisely on third-party data

X-Financial should consider the credibility, explainability and reliability when purchasing third-party data. Based on our model, few third-party variables play a substantial role in predicting the target variable. X-Financial should consider the cost-effectiveness and purchase third-party data from a reliable source.

2. Request supplemental documents for high-risk customers

X-Financial should follow the patterns of high-risk borrowers in our analytical findings and request supplemental documents for high-risk group to mitigate risks. For instance, long-term loan borrowers and customers with lower education levels tend to have higher default rate. X-Financial could request supplemental documents like proof of income or certificate of deposit to ensure that these borrowers are less likely to default.

3. Attract and maintain customers with low-risk

From our data analysis and modeling, customers who are less likely to default share some general patterns including loan terms, cell phone brand, gender and education. X-Financial could increase marketing spending on this targeted group to attract more new customers. On the other hand, X-Financial can provide existing customers who have not defaulted on record with rebate or discount for their next loans to maintain its customer base.

Work Cited

Brownlee, J. (2019, August 21). Feature Importance and Feature Selection With XGBoost in Python. Retrieved from https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/.

Samiullah, Christopher. "How to Deploy Machine Learning Models." *Sitewide ATOM*, https://christophergs.github.io/machine learning/2019/03/17/how-to-deploy-machine-learning-models/.

Kealhofer, S (2003), 'Quantifying credit risk I: Default prediction.' Financial Analysts Journal, 59(1):30-44.