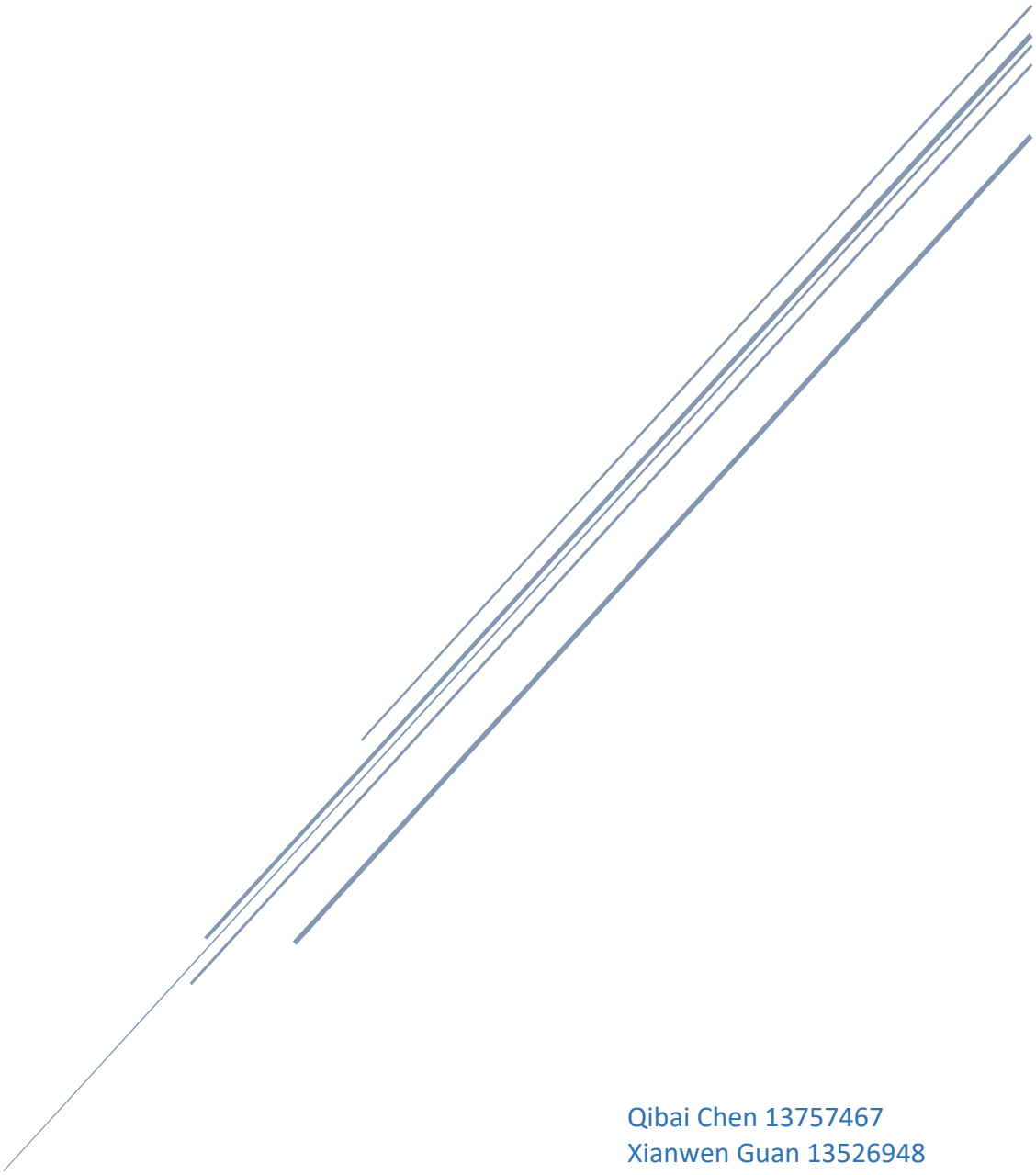
CHURN ANALYSIS IN SUBSCRIPTION-BASED BUSINESSES FROM PREDICTION TO INFERENCE

Mid-project update



Qibai Chen 13757467
Xianwen Guan 13526948
Weilin Sun 13462383
Yangyang Jin 13647716

Table of Contents

1.0 Business Problem

As mentioned in the previous project proposal, our client is an local Australian superannuation fund called Colonial First State (CFS). They showed that the cost of acquiring new customers and the rate of customer churn both are increasing at a rapid pace. The high churn of customer rate also leads to a decline in the company's profitability. Furthermore, they stated that they know very little about members data. Consider these aspects, CFS hopes that we can help them get the reasons for customer churn and find out countermeasures to reduce customer churn as much as possible.

This project will build a robust churn propensity model that can score each customer based on his probability of churn over the next six months. Several machine learning techniques for churn prediction have produced verifiable results from interpretable models, such as boosting, non-parametric, and logistic regression. However, this approach is only valid when the size of the customer database is very small and the sample size varies, not for larger datasets. Therefore, we propose deep learning (DL) algorithms to deal with massive financial data since feature transformation in deep learning can use historical data to weight features differently.

We also propose causal Bayesian networks to predict cause probabilities that lead to customer churn. We employed Bayesian causal graphs to encode assumptions and determine dependency levels between features. In our team, we will use appropriate data analysis methods to explore this data and provide our client with the results they want.

2.0 Data exploration

Our dataset comes from CFS member information records in June and December 2015, including age, account tenure, savings plan, billing information and service record information. As mentioned earlier, the volume of this dataset is relatively large, with a single dataset having nearly 270,000 member record information and 88 behavioural attributes. Therefore, how to perform data cleaning and data preprocessing on this dataset is very important.

Firstly, we aggregated membership data from June and December as our observation dataset. At the same time, we de-duplicate the integrated data, and only save the unique data unit. This step aims to save data storage space and improve write performance, thereby improving the model accuracy. Additionally, we performed binary transformation (One-Hot Encoding) on categorical data. The benefit of binary transformations is that it can

make our training data easier to use and more expressive. We also normalized the data. The purpose of normalization is to place the range of data within a specific cell range. Using the normalized data for data analysis can eliminate the dimension and simplify the model, thus speeding up the calculation.

To satisfy model performance, we also applied the two main inclusion criteria of account tenure and balance. Firstly, we only retained data on customers older than six months. Secondly, we removed account balances below $1,500 to improve forecasts, since predicting churn probabilities for inactive accounts is of low value to superannuation funds.

Finally, regarding the definition of churn, we define a customer as a "churner" if they close their account within the subsequent 6-month time window. Therefore, we use binary results for each customer [0 or 1], where 1 means the account is closed and 0 means that it was not closed in the subsequent 6-month time window.

3.0 Initial finding

This section contains the analysis of the data exploration so far, and some of the conclusions reached.

As shown from the pie chart below (refer to Figure 1), churn customers accounted for 14.2% of the total, and the remaining 85.8% belonged to non-churn customers. It also a severely imbalanced dataset.

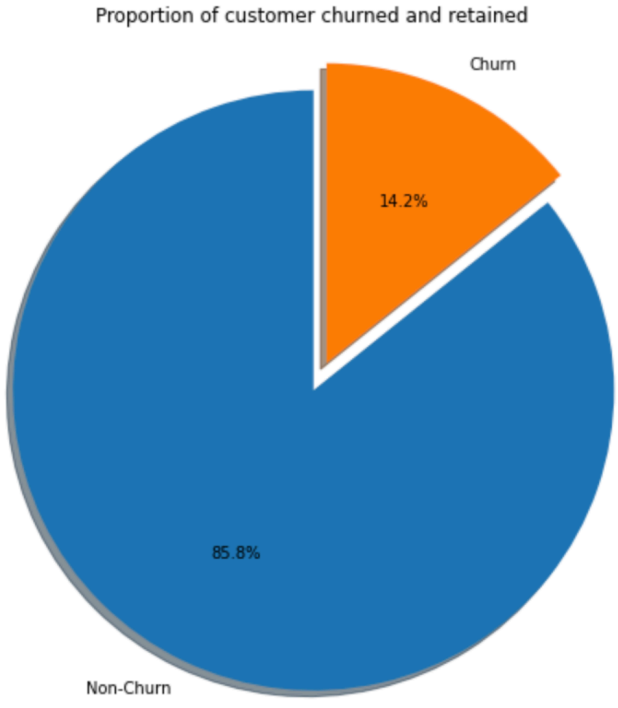


Figure 1 # Pie chart for Proportion of Customer Churned and Retained

In terms of age distribution, it can be found that the people who use the superannuation fund range from 20 to 80 years old, and are mainly concentrated in the middle-aged group of 40 to 60 years old (Refer to Figure 2). In addition, it can be found from the box plot that there are some outliers in the age data (Refer to Figure 3), and from the comparison of age and customer churn, the age of churn customers is generally lower than that of non-churn customers.

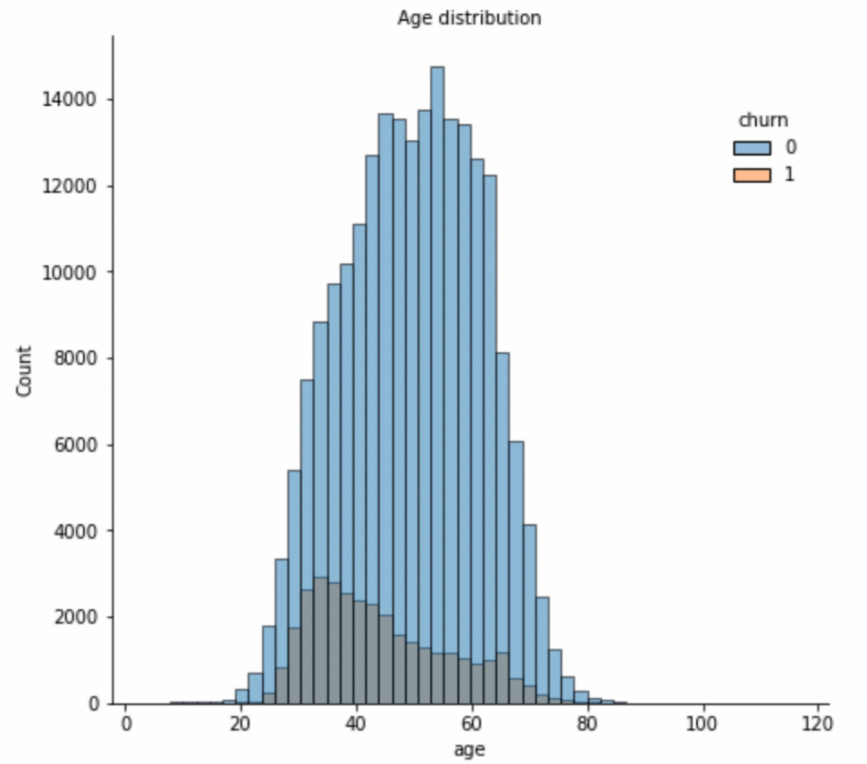


Figure 2 # Histogram for Age Distribution of Customers

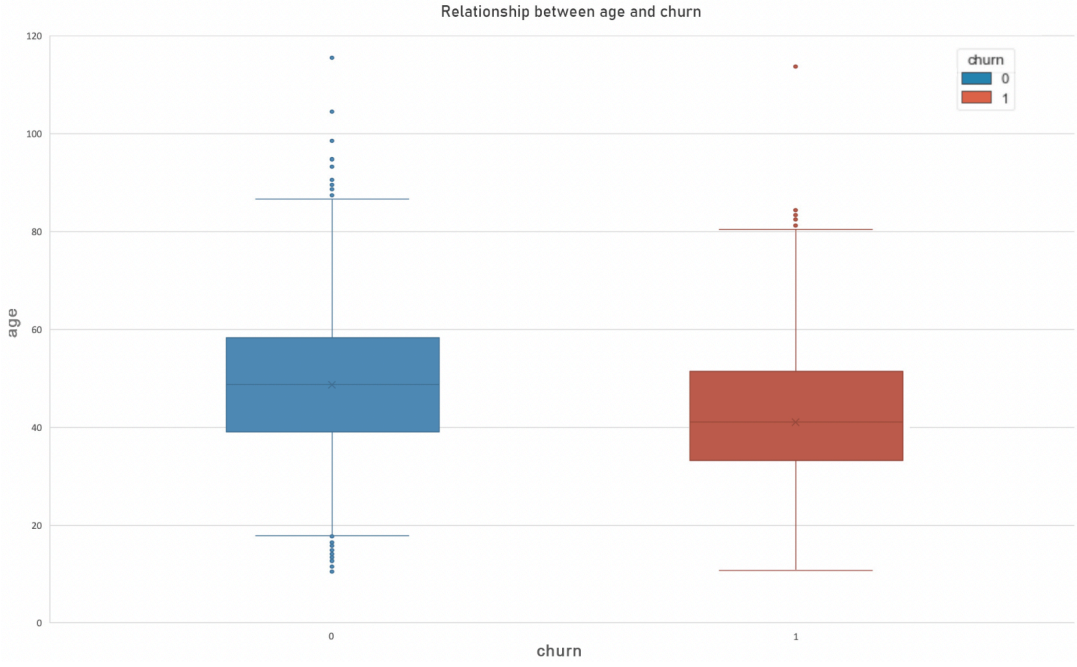


Figure 3 # Box plot for Age Distribution of Customers

As mentioned before, we only keep data for more than six months account tenure. Since, the churn rate is usually low five months after account opening. However, as shown from the Histogram (Refer to Figure 4), more customers began to feel dissatisfied and opted out starting from the sixth month, especially the sixth month.

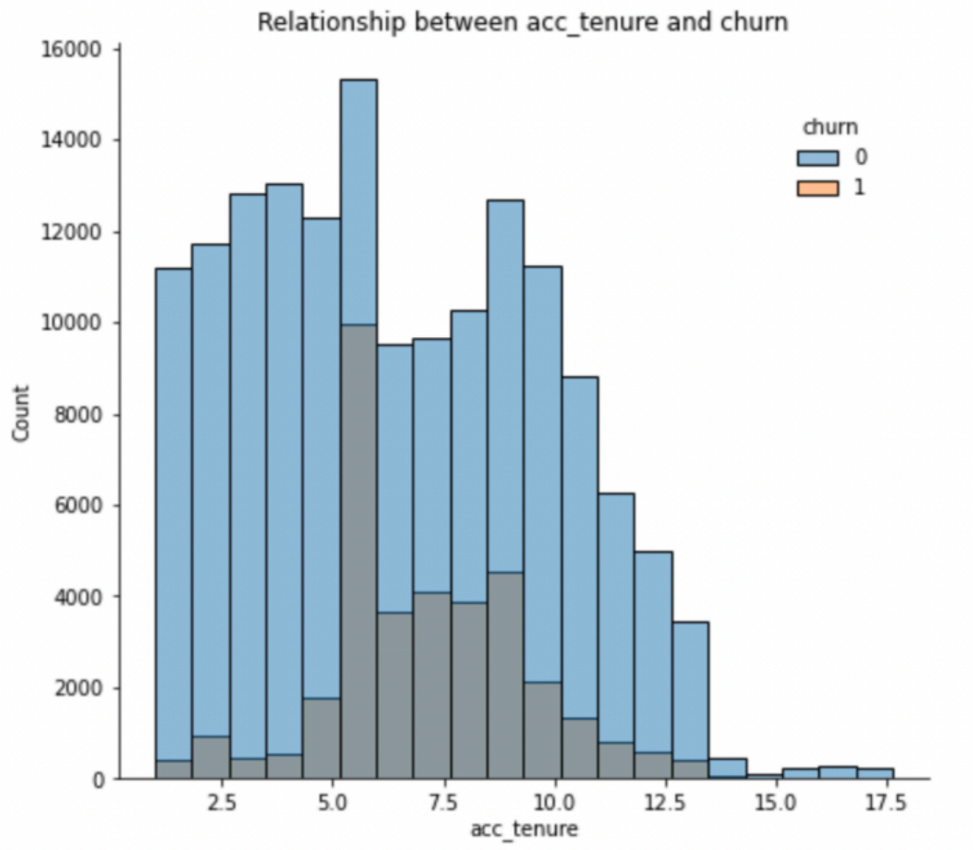


Figure 4 # Histogram for Relationship between Acc_tenure and Churn

From the box plot below (Refer to Figure 5), it can be found that customers' account balances are mainly distributed within $100,000, and the churn rate is also the largest among this part of customers. In addition, in the distribution of account balances within $100,000 (Refer to Figure 6), a considerable proportion of account balances are low, and the churn rate of these low-balance customers is very high. By contrast, customers with higher account balances have lower churn rates.

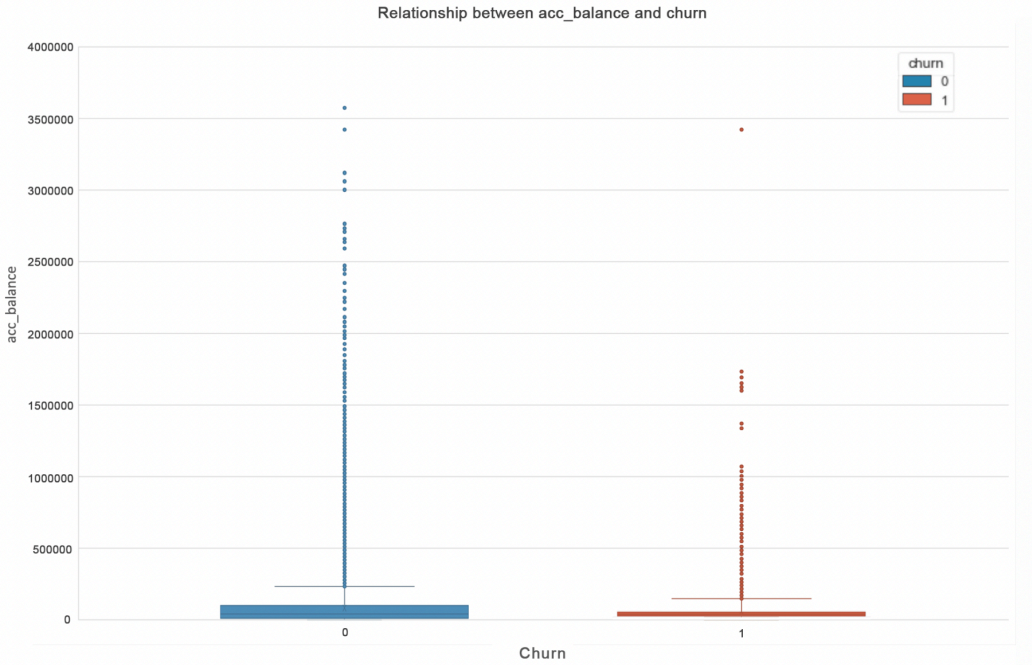


Figure 5 # Box plot for Relationship between Acc_tenure and Churn

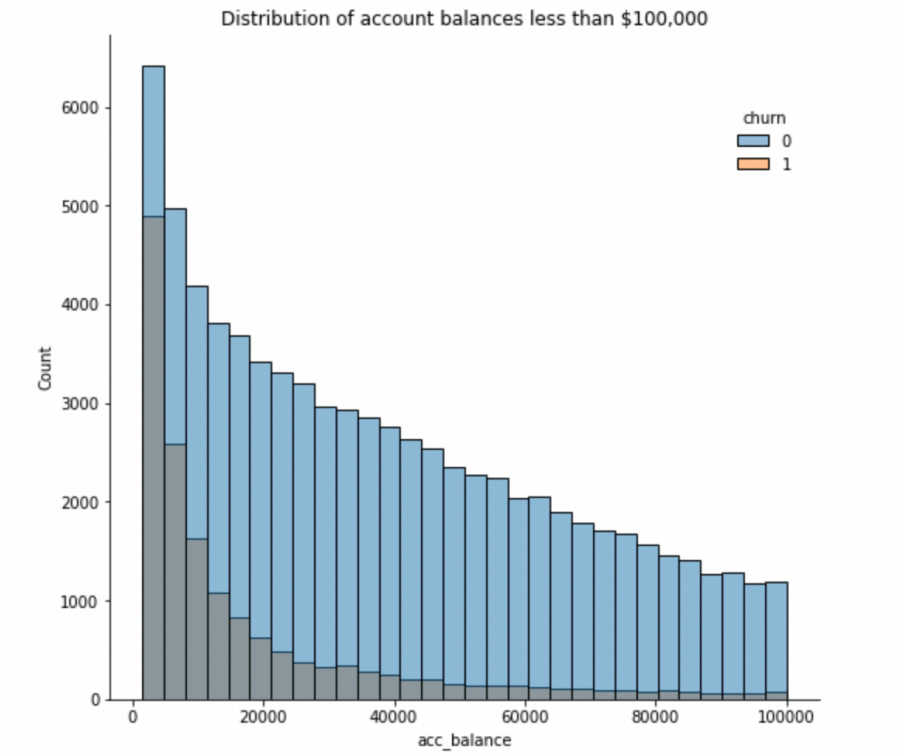


Figure 6 # Histogram for Distribution of account balance less than $ 100,000

From the Pareto chart below (Refer to Figure 7), it can be found that the customers who chose only one investment are the most, accounting for about 60% of the total, and those with less than ten investments account for 90% of the total. However, as shown from the box chart below (Refer to Figure 8), some investments with more than ten items appear to be outliers.

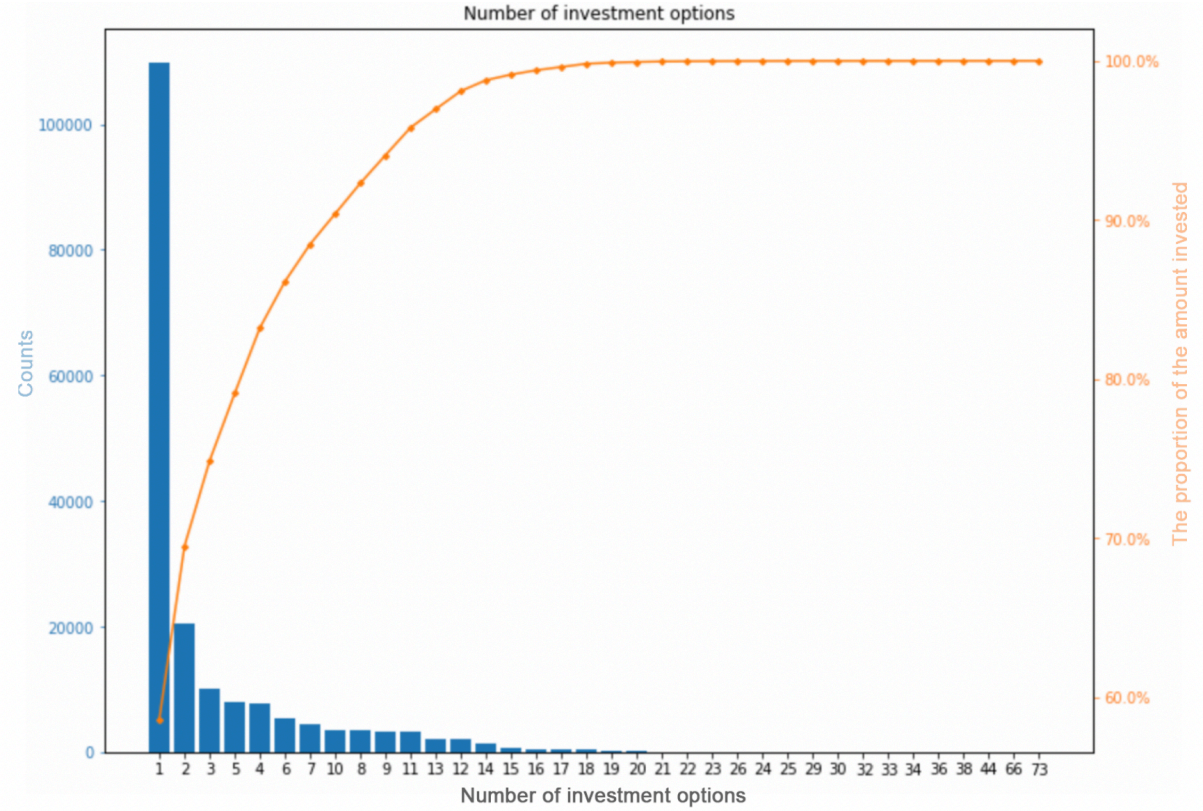


Figure 7 # Pareto for Number of Investment options

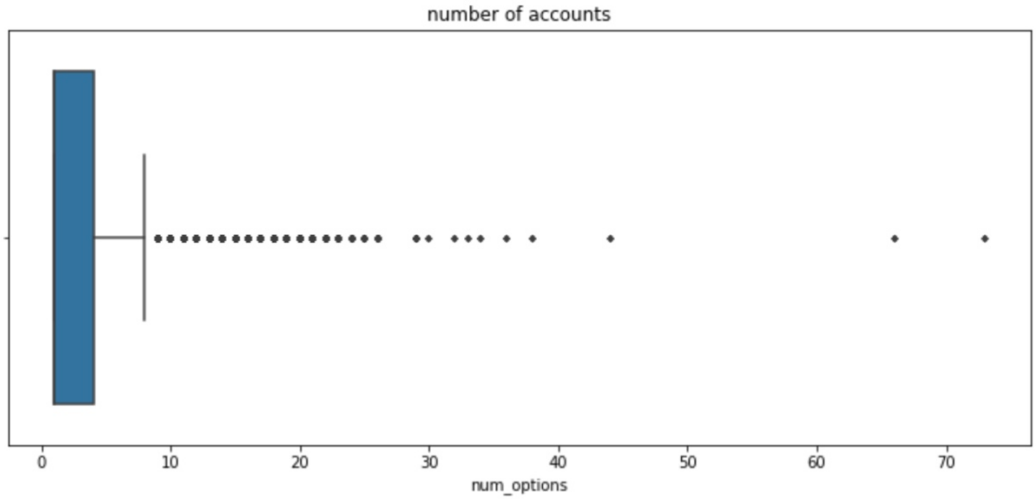


Figure 8 # Box plot for Number of Investment options

As shown from the bar chart below (Refer to Figure 9), most customers choose mobile phones as their personal contact information, followed by email. However, the number of customers who record work phones was significantly lower than the others, with about three- quarters not choosing to record their work phones.

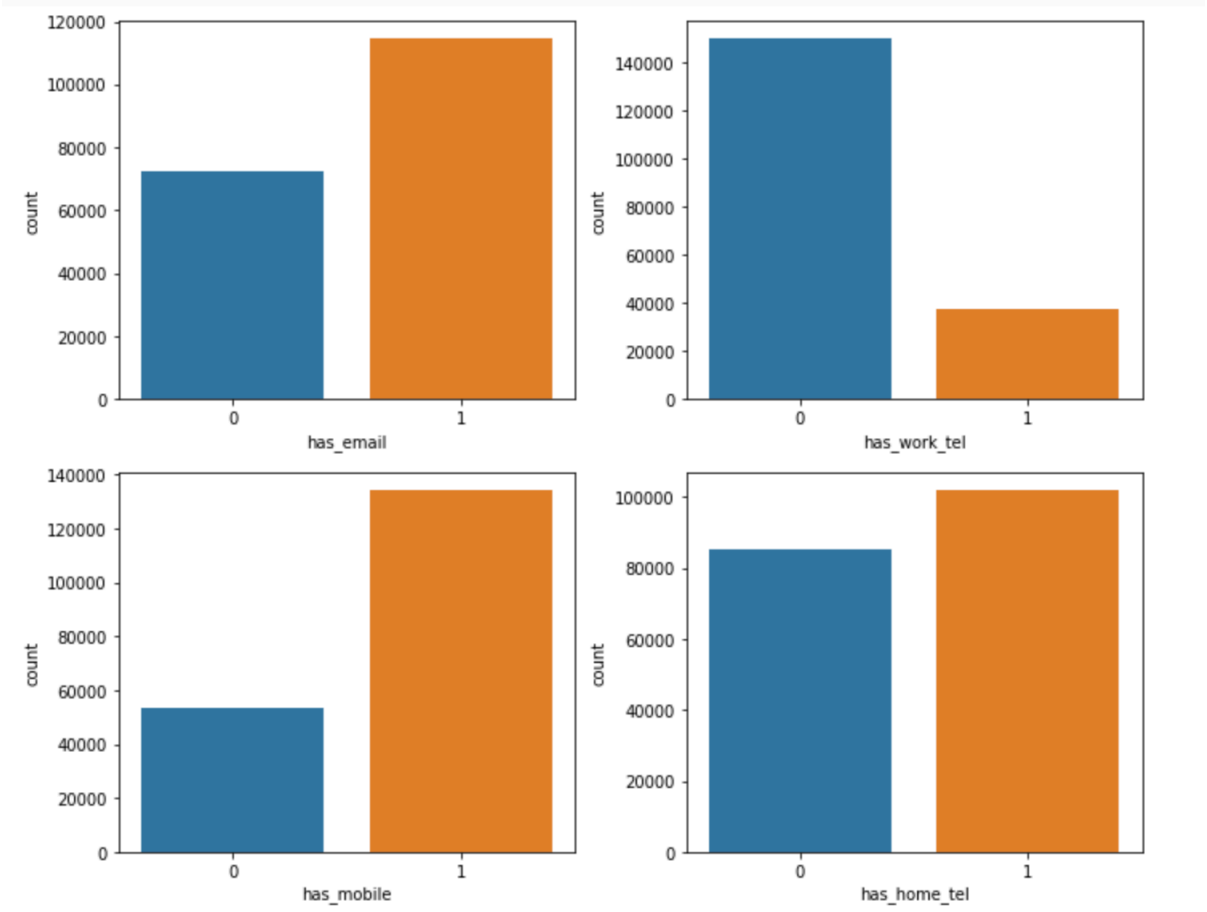


Figure 9 # Bar charts for has_email, has_work_tel, has_modile, and has_home_tel

From the histogram below (Refer to Figure 10), it can be found that there are many members' account balance changes showing a negative growth, and the amount is relatively large. At the same time, it can be found from the ratio of balance change that those who show negative growth in the balance change are usually also churners (Refer to Figure 11).

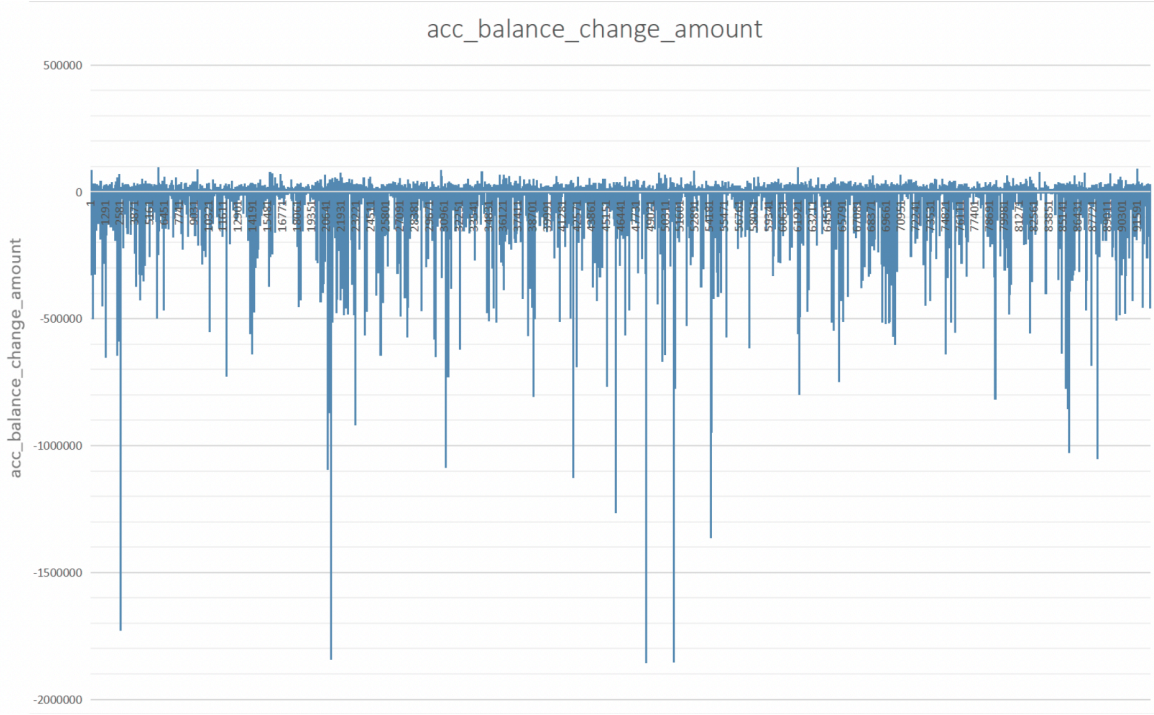


Figure 10 # Histogram for acc_balance_change_amount

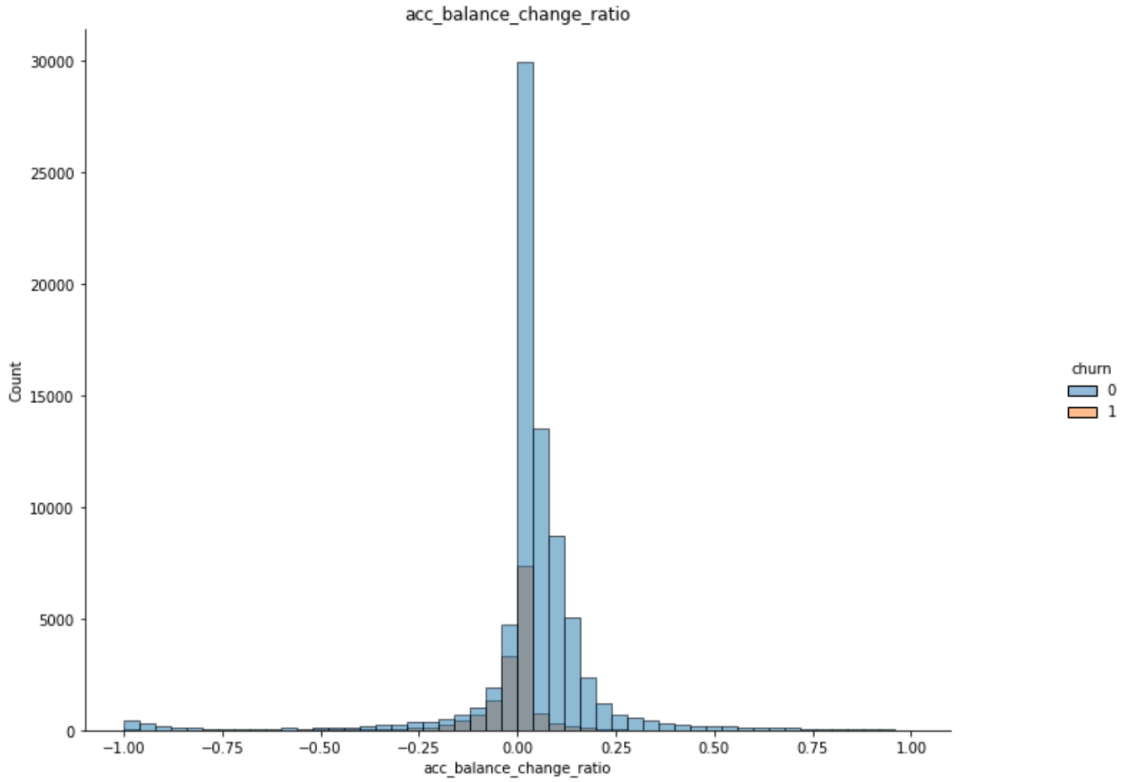


Figure 11 # Histogram for acc_balance_change_ratio

In the curve graph below (Refer to Figure 12), we can see the distribution of days since the last change of advisors, dealers, logins to FirstNet, and incoming calls. It can be found that the customer engagement is low in majority of population and churner is concentrated among customers who have been inactive for more than a year.

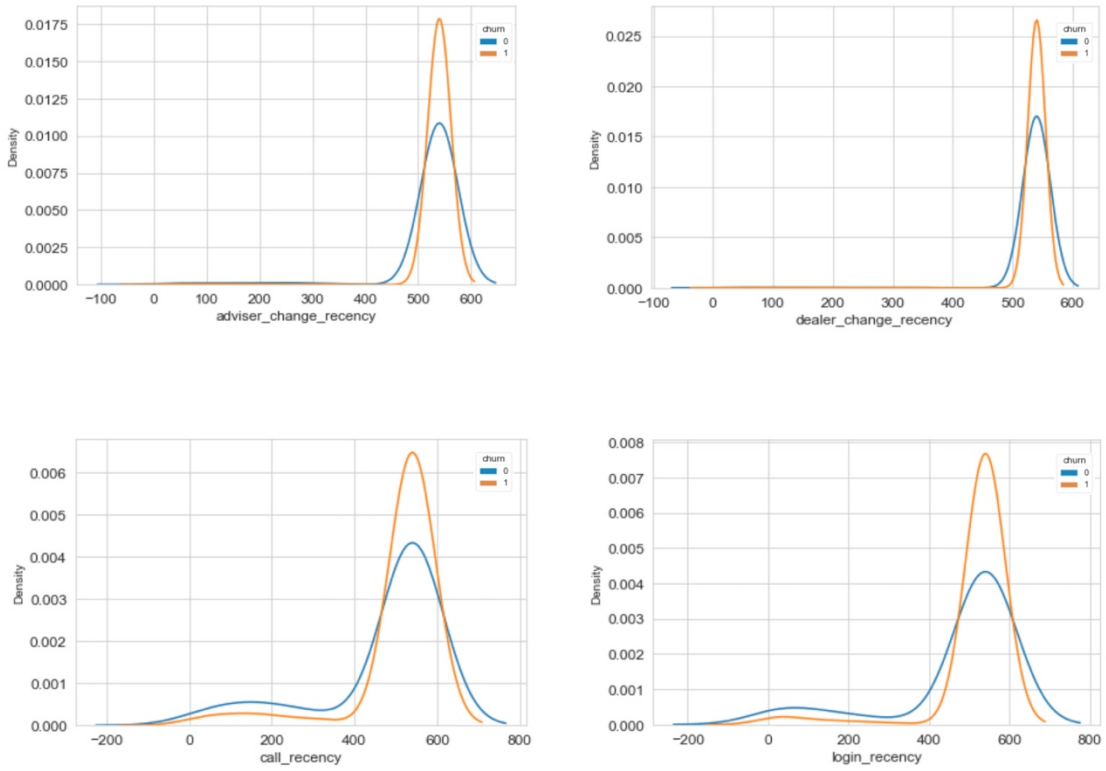


Figure 12 # Curve graph for adviser_change_recency, dealer_change_recency, call _recency, and login_recency

As shown from the statistics below (Refer to Figure 13), some contributions such as SG contributions, personal contributions and spouse contributions are generally low. And in the comparison of the different contributions and churn below (Refer to Figure 14), the churn customers are more distributed in the lower contribution amount.

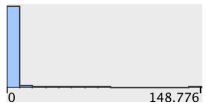
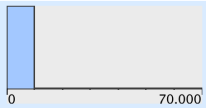
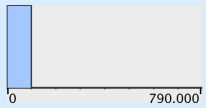
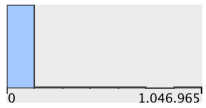
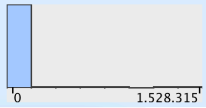
Column	Min	Mean	Median	Max	Std. Dev.	Skewness	Kurtosis	No. Missing	No. +∞	No. -∞	Histogram
sg_amount	0.0	1,521.9565	0.0	148,775.56	3,426.7853	5.0382	69.2219	0	0	0	
salary_scr_amount	0.0	261.9424	0.0	70,000	2,267.3482	11.0198	142.3179	0	0	0	
spouse_contr_amount	0.0	5.3276	0.0	45,430	199.9387	138.5583	29,236.8829	0	0	0	
personal_contr_amount	0.0	593.2698	0.0	790,000	9,496.1084	37.7346	1,910.705	0	0	0	
rollover_amount	0.0	287.7235	0.0	1,046,964.87	7,536.1234	76.7901	7,627.3681	0	0	0	
contribution_amount	0.0	1,148.2633	0.0	1,528,314.54	13,385.0518	45.0013	3,305.4583	0	0	0	

Figure 13 # Statistics for sg_amount, personal_contr_amount, rollover_amount, contribution_amount

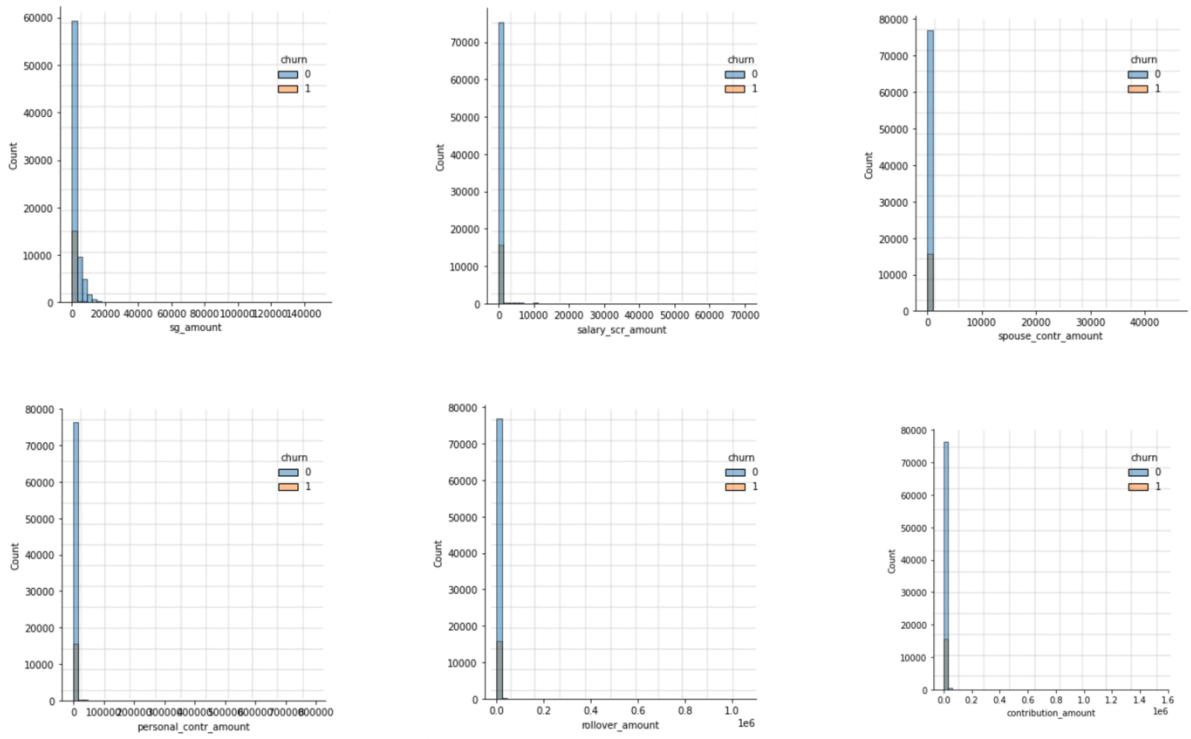


Figure 14 # Histogram for sg_amount, personal_contr_amount, rollover_amount, contribution_amount

In this heatmap below (Refer to Figure 15), it can be found that there are many attributes that show strong relationships, such as salary_scr_freq and salary_scr_amount, insurance_types and insurance_recency. At the same time, some attributes also show a strong negative correlation, such as salary_scr_freq and salary_scr_recency, adviser_change_freq and adviser_change_recency and so on.

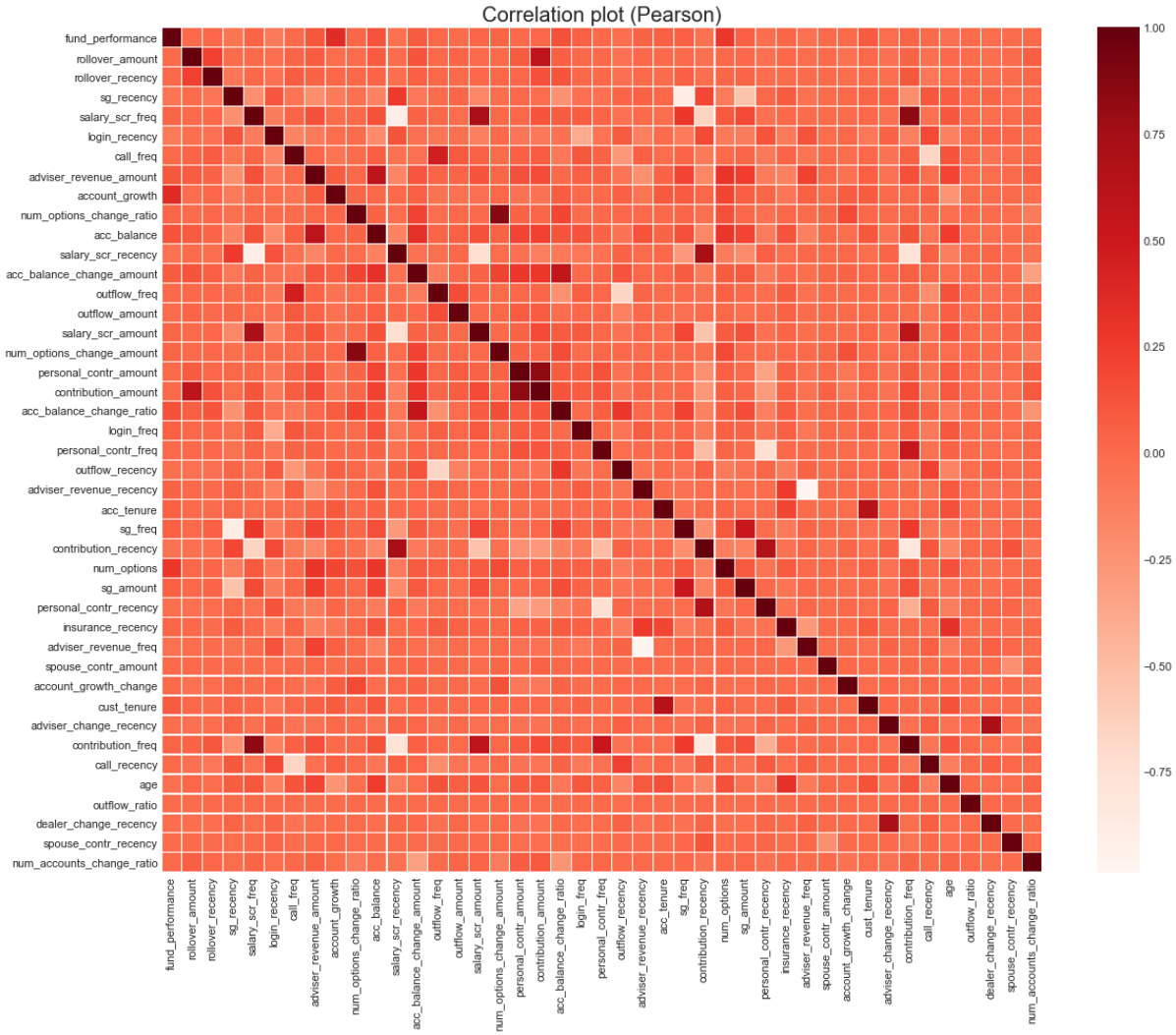


Figure 15 # Heatmap of Correlation for each attribute

We also used the random forest for feature selection. From the weight diagram of feature importance, we can see which feature have the most significant on customer churn, such as acc_balance_change_amount, acc_balance_change_ratio and acc_balance (see Figure 16).

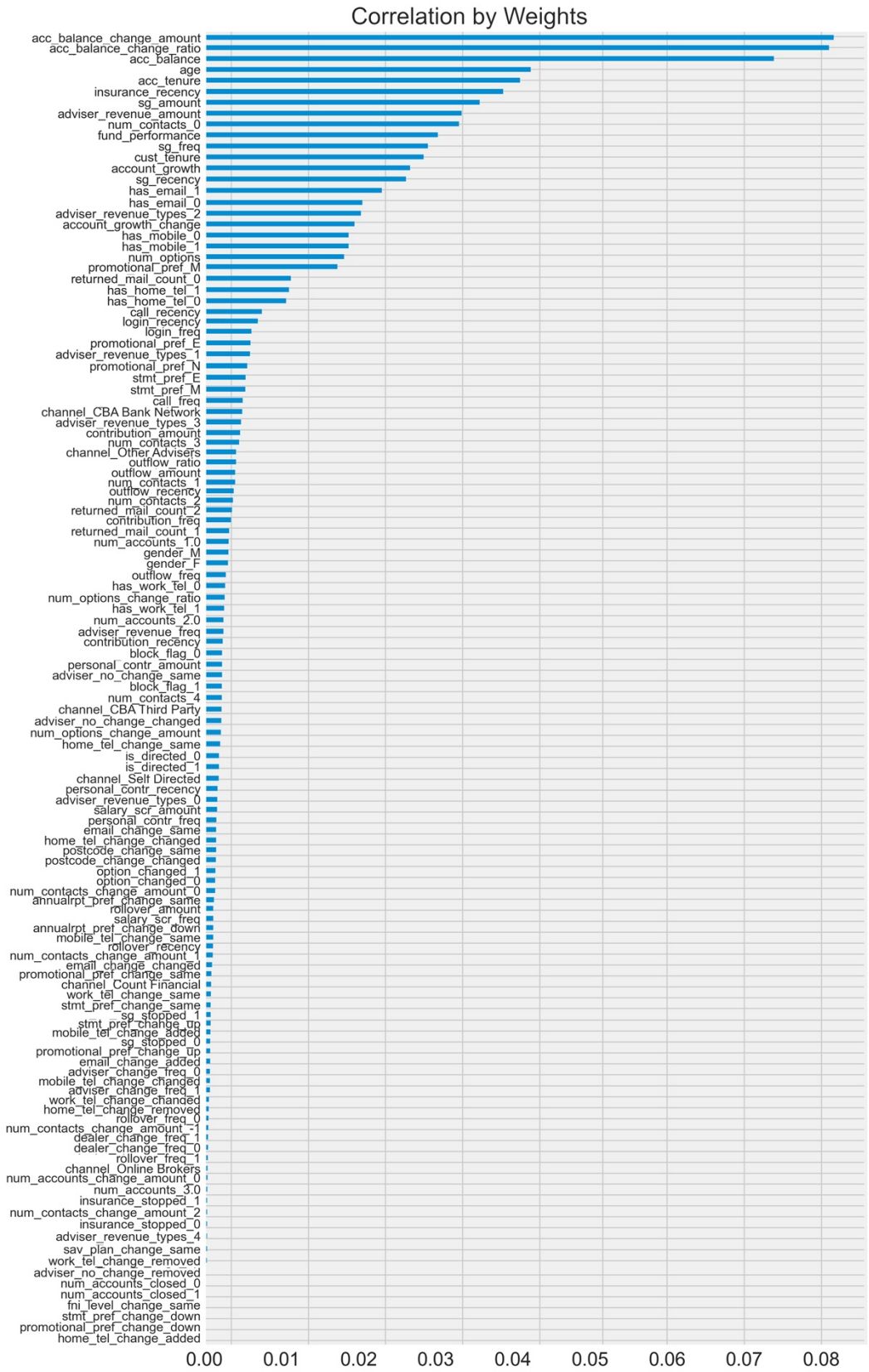


Figure 16 # Correlation by Weights

4.0 Challenges encountered

This section details some challenges we encountered so far and the identified risks. At the same time, steps have been taken to address these issues and mitigate risks as they arise.

The current challenges and solutions are as follows:

- Imbalanced Dataset

As mentioned before, there is a severe data imbalance problem in the dataset. In this case, we applied the SMOTE method for data training. It can analyze the minority class and add new samples to the dataset based on the minority class sample, thereby improving the prediction ability of the minority class

- Dataset Complexity

Due to the large volume and complexity of the dataset, how to perform data preprocessing and data cleaning is a challenge. In this project, through some data preprocessing such as One-Hot Encoding and normalization, we eliminated redundant data, accelerated the speed, and improved the model's performance.

- Model Performance Issues

In model training, how to further improve the model performance is also an inevitable problem. In this project, we try to use voting ensembles to improve the performance of the model. It combines predictions from multiple other models, and ideally it can achieve better performance than any single model used in the ensemble. In addition, we will also use feature selection to remove unimportant features, which can further improve model performance

Risks and mitigation:

• Misunderstanding customer requirements

Risk can be mitigated by proactively communicating with clients and acknowledging that updating our project schedule will help get their exact requirements and expectations.

• Lack of professional skills of team members

This risk can be mitigated by frequent group meetings to keep the project updated and ensure that the team leader assigns tasks based on the team members' proficiency.

• Absence of a meeting

This risk can be mitigated by having a separate communication with the team leader to ensure that the progress of the project can be followed.

• Conflicts of opinion among team members

By being patient with the opinions of other team members, expressing their differences, and summarizing problems to resolve conflicts of views

5.0 Updated project plan

At this stage, we have done data exploration and data preprocessing. We also built models like logistic regression, decision tree, random forest, and XGBoost and evaluated them.

Next we will continue with the causality analysis of churn. Essentially, our project time is sufficient because we move everything forward to prevent contingencies or problems that may arise.

The revised work plan is shown in the figure below (Refer to Figure 17). Our project plans and deadlines remain the same and still follow the CRISP-DM methodology, as we have done so far in this project.

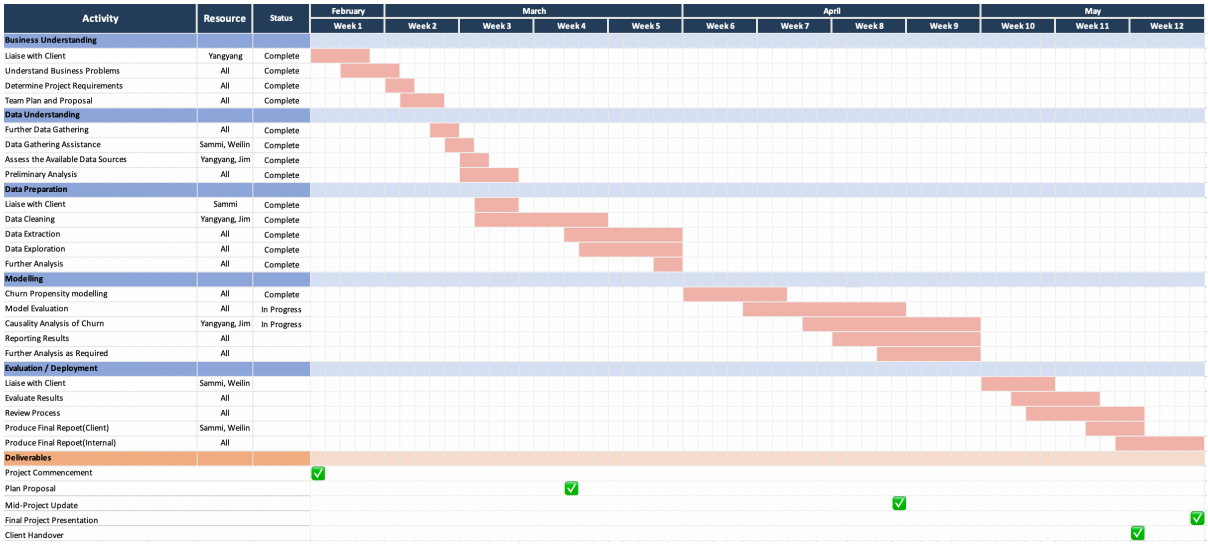


Figure 17 # Project Timeline