CHURN ANALYSIS IN SUBSCRIPTION-BASED BUSINESSES FROM PREDICTION TO INFERENCE

Final report and presentation

Aashi Parikh	13604590
Sammi Guan	13526948
Yangyang Jin	13647716
Weilin Sun	13462383

Table of Contents

1.0 Executive Summary

This report was commissioned by CFS to investigate and analyse their customer churn in the subscription business and based on the churn analysis of the subscription business, it makes recommendations on ways to reactivate lost users.

Analysis includes billing information for members, demographic information, service record information, and member engagement. The results of the analysis show that the churn rate of CFS members in 2015 accounted for 14.2%. The higher churn rate of members also directly led to the decline of the company's profitability. There is a large percentage of accounts with low balances, and these accounts also have a very high churn rate. Moreover, the age of churn customers is generally lower than that of non-churn customers, mainly around 30-35 years old. Most of them has low customer engagement (for example: FirstNet login, call frequency, etc.), and churners are mainly concentrated in customers who have been inactive for more than a year.

The purpose of this project is to help our clients solve their customer churn problem and provide corresponding solutions. To help them resolve this, it is recommended that:

- Proactively communicate with customers
- Define the most valuable customers to the company
- Increase customer engagement
- Define a roadmap for new customers
- Collect customer feedback

In addition, the analysis performed has some limitations, including:

- Unknown about the current economic situation of the company.
- There are data limitations. Results and forecasts are based on past behavioural data (2016), not present.

2.0 Business Problem

Customer retention is one of the main growth pillars of subscription-based business model. The superannuation fund market is highly competitive, and even within one product category, customers are free to choose from a multitude of fund markets. After a few bad experiences, or even once, customers may opt out. if there is a series of unhappy customer churn, both the loss of profits and the damage to the company's reputation can be huge.

Churn rate is a critical metric for companies with subscription-based business models. It can be calculated as follow formula:

$$ChurnRate = \frac{\textbf{(Churners at end of month)} + \textbf{(New members who churned BOM)}}{\textbf{(Total active members at the end on the month)}}$$

For example, 100 churned users at the end of the month and 14 new churned members at the beginning of the month, the total number of active members at the end of the month is 1000. Then the churn rate can be calculated (100+14)/1000 = 0.114 = 11.4%.

According to the CFS report, they showed three main problems:

1. **The cost of acquiring new customers and the rate of customer churn both are increasing at a rapid pace**
   For enterprises, if they want to promote and acquire new customers in a more cost-effective way, while maintaining old customers, they must dig deep into customer value. According to the survey, the cost of acquiring a new customer is 5 times that of retaining an old one. If we can seize the opportunity to serve old customers and gain their recognition, so as to retain more old customers, there is no doubt that this will be of great value to the company.

2. **High customer churn rates lead to a decline in company profitability**
   The higher the customer churn rate, the first impact is to reduce the profitability of the company. In fact, research by Bain & Company and Earl Sasser of Harvard Business School shows that increasing customer retention by 5% can lead to a 25% to 95% increase in profits.

3. **They know very little about member data**
   Understanding customer data plays a key role in a company's business growth. When companies can gain enough insight into their customers' data, they can make decisions about it to improve customer experience, increase customer loyalty and retention, and recurring revenue. Conversely, if the company doesn’t

know enough about its members' data, not only will the business problem not improve, but it may also snowball and become more difficult to deal with.

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. The Dowhy package is based on a Bayesian causal graph model for causal discovery of all possible methods and uses graph-based criteria to find possible explanations. In our team, we will use appropriate data analysis methods to explore this data and provide our client with the results they want.

3.0 Data Exploration

Our data set was derived from CFS member information records for June and December 2015, including age, account usage period, savings plan, billing information and service record information. The size of this dataset is relatively large, with nearly 270,000 members' record information and 88 behaviour attributes in a single dataset.

3.1 Attributes Description

The size of the dataset has been displayed below.

Dataset	# Attributes	Size (# recorded)
CFS_member_data	88	269522

3.2 Data Preprocessing

For the massive and high-dimensional data, how to effectively clean and pre-process the data set 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 normalised the data. The purpose of normalisation is to place the range of data within a specific cell range. Using the normalised 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.

3.3 Data Visualization

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.

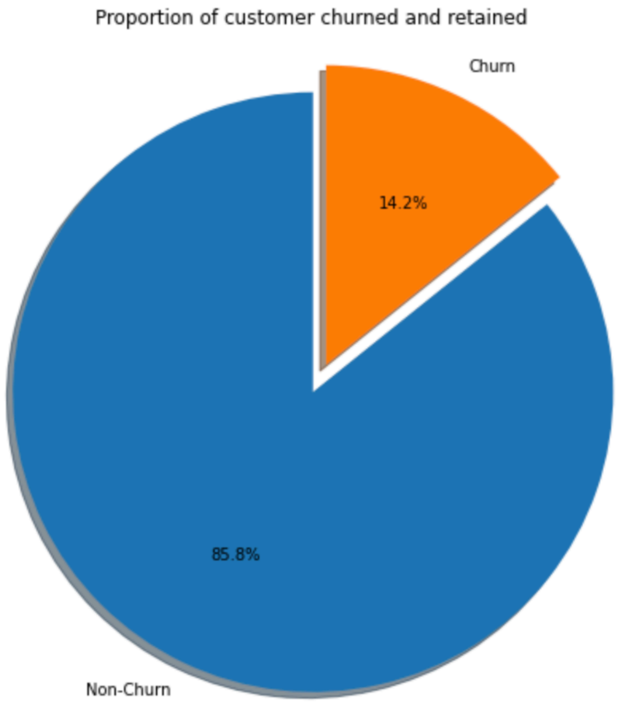


Figure 1 # Pie chart for Proportion of Customer Churned and Retained

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.

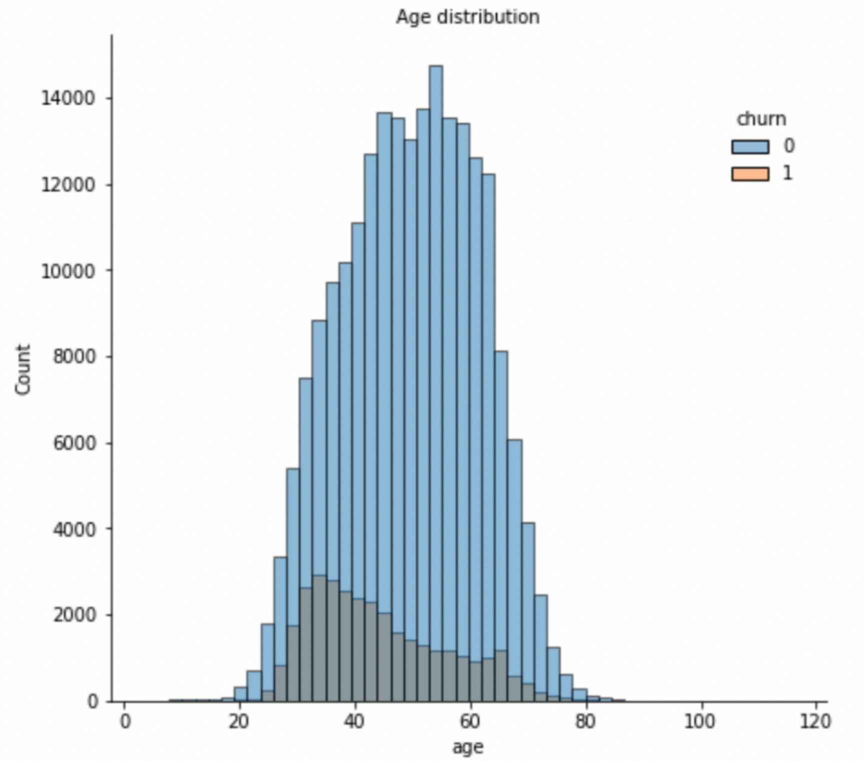


Figure 2 # Histogram for Age Distribution of Customers

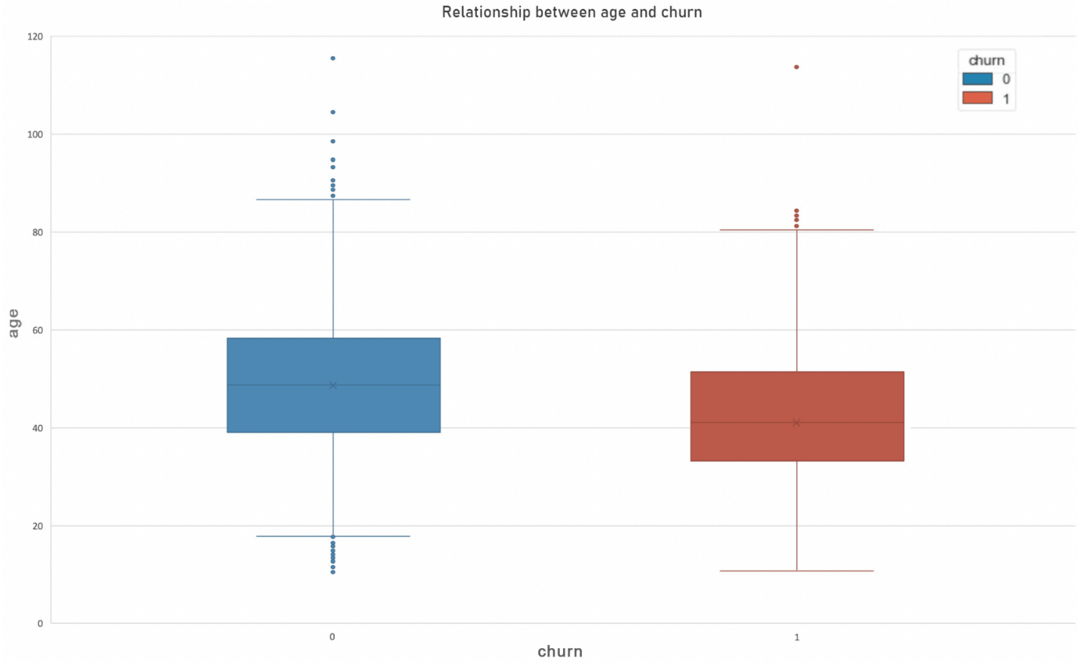


Figure 3 # Box plot for Age Distribution of Customers

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.

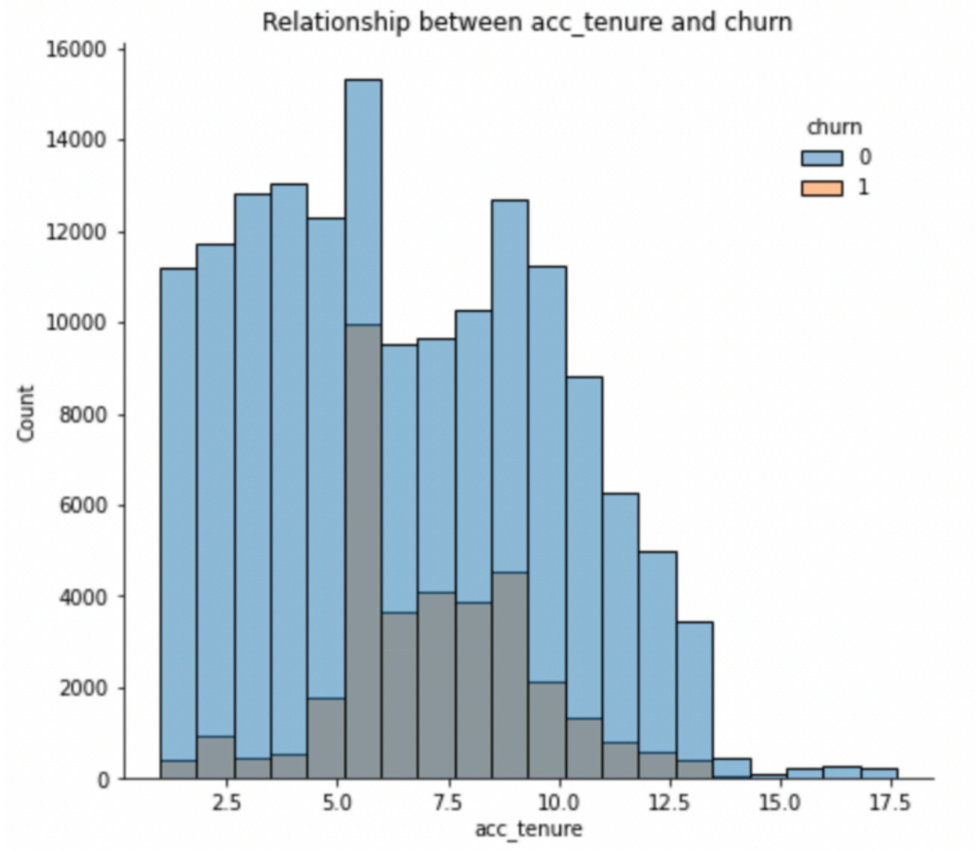


Figure 4 # Histogram for Relationship between Acc_tenure and Churn

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.

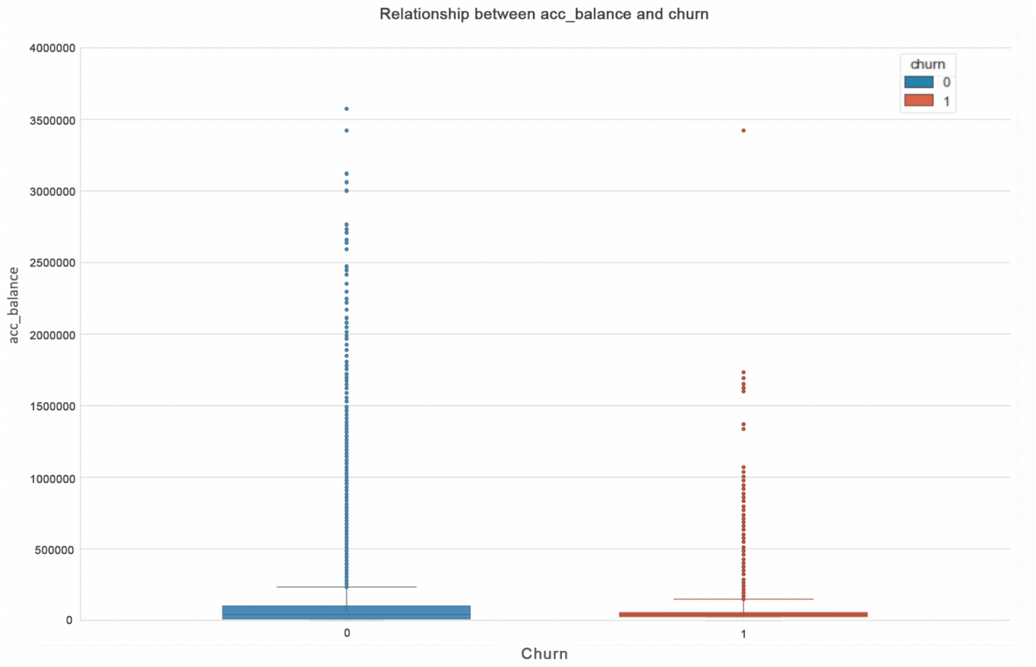


Figure 5 # Box plot for Relationship between Acc_tenure and Churn

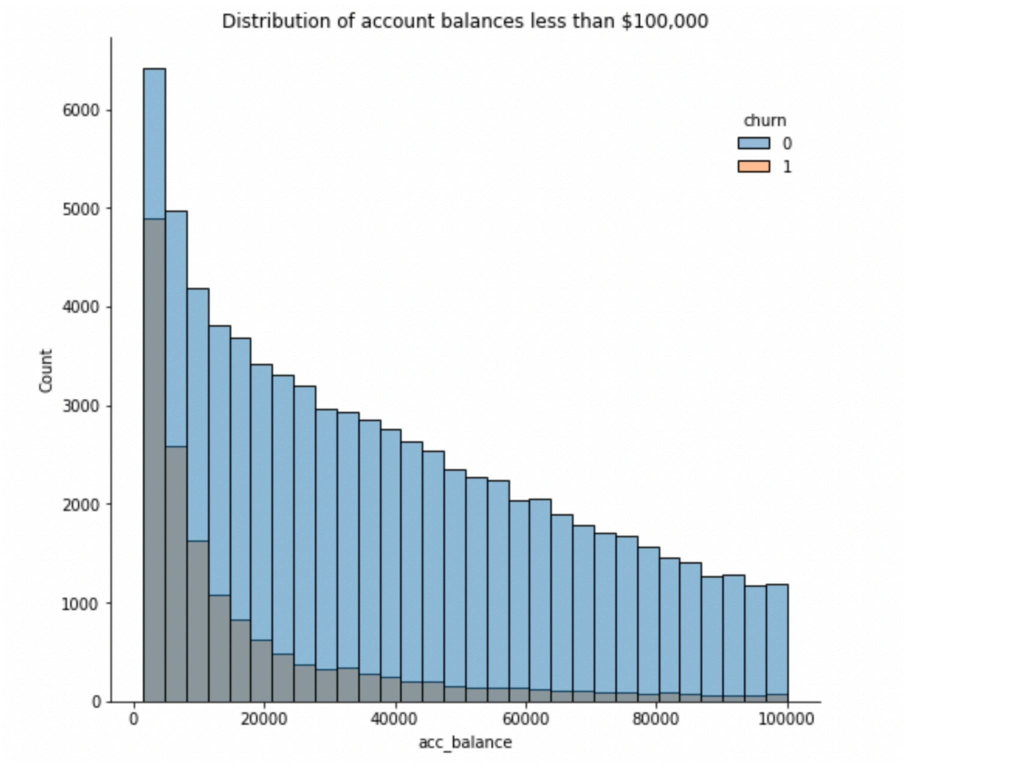


Figure 6 # Histogram for Distribution of account balance less than $ 100,000

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.

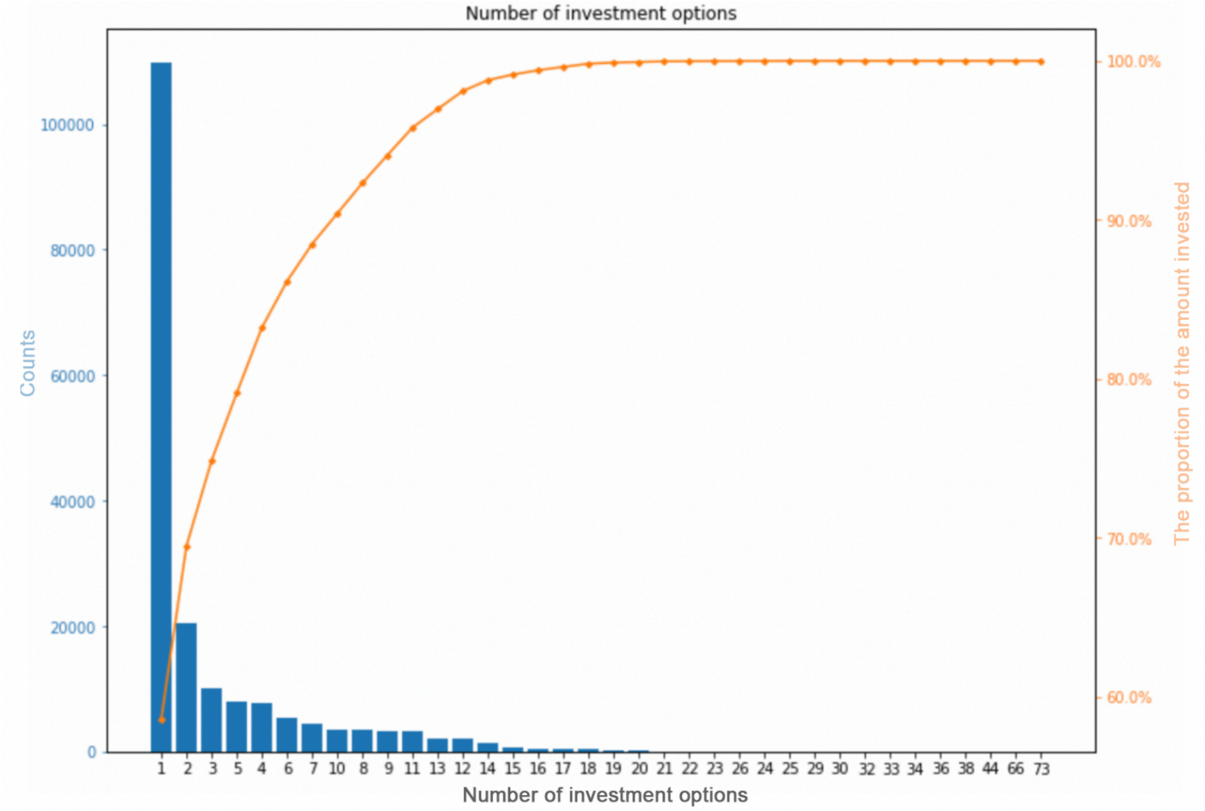


Figure 7 # Pareto for Number of Investment options

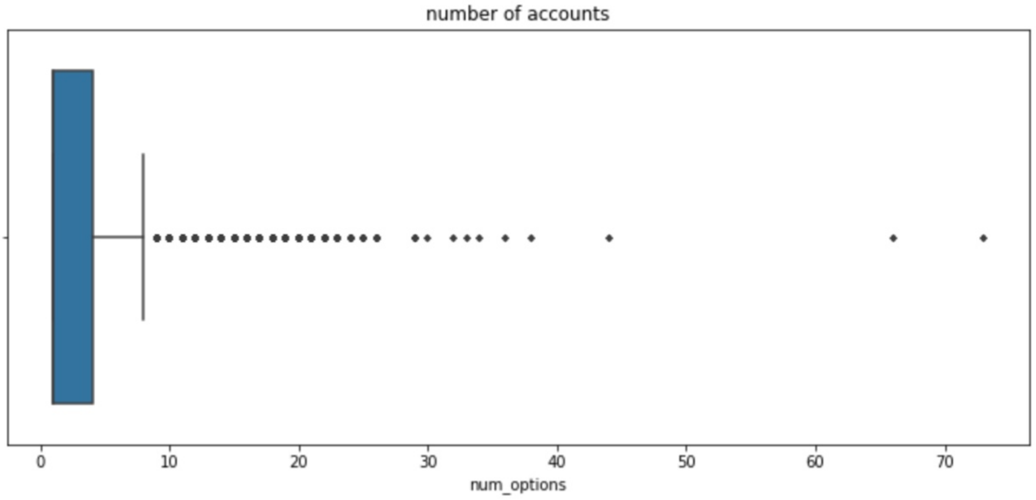


Figure 8 # Box plot for Number of Investment options

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.

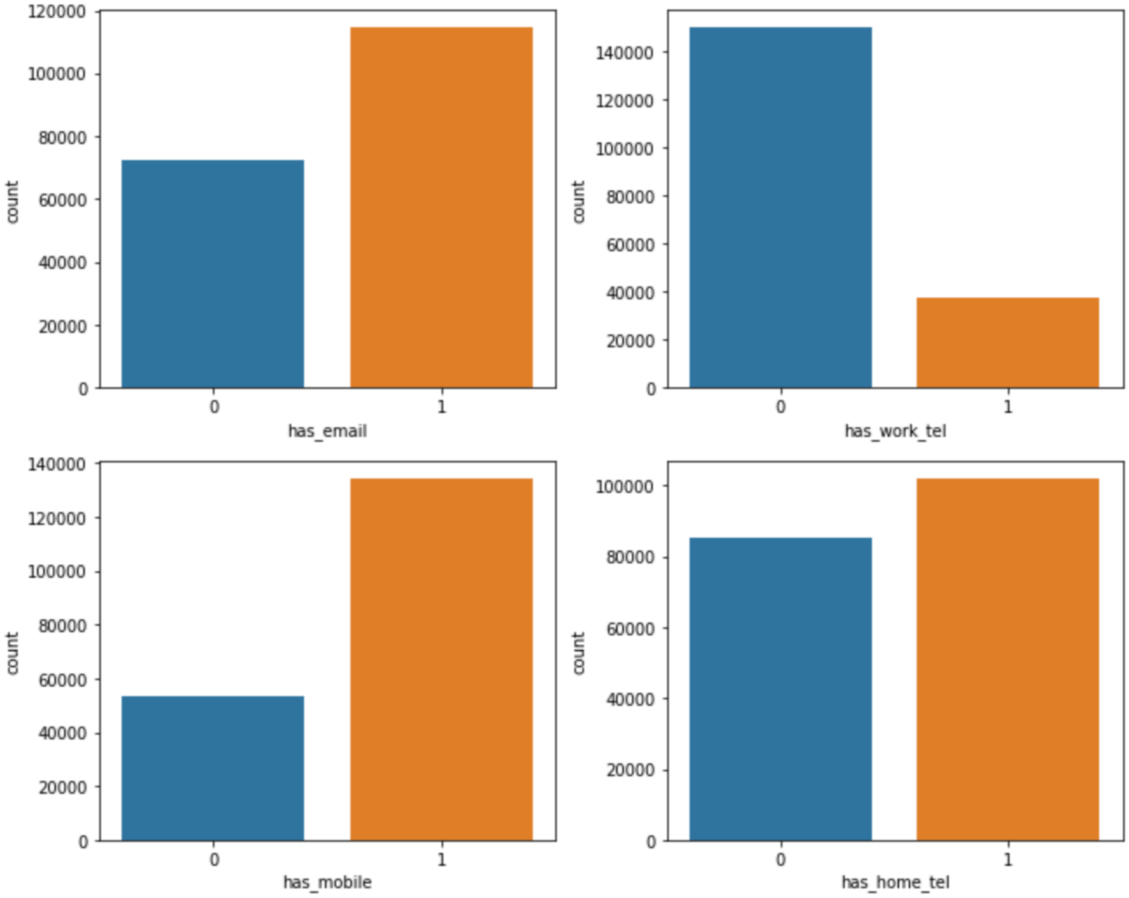


Figure 9 # Bar charts for has_email, has_work_tel, has_modile, and has_home_tel

From the histogram below (Refer to Figure 10), it can be found that there are many members' account balance changes showing 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).

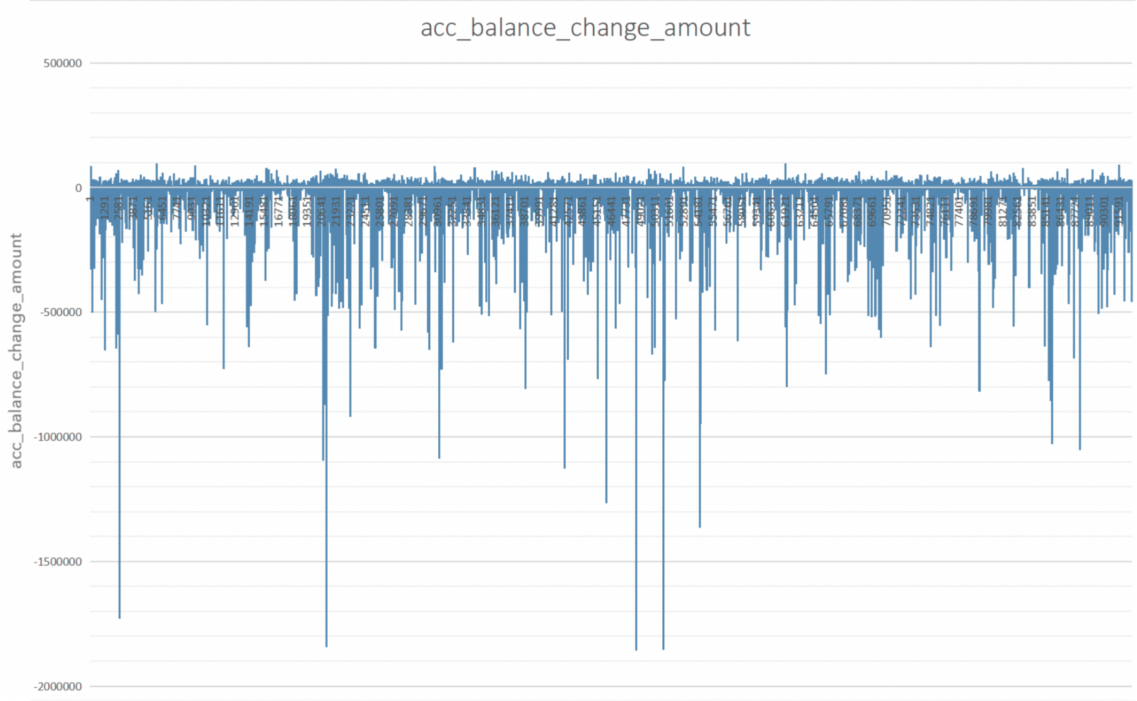


Figure 10 # Histogram for acc_balance_change_amount

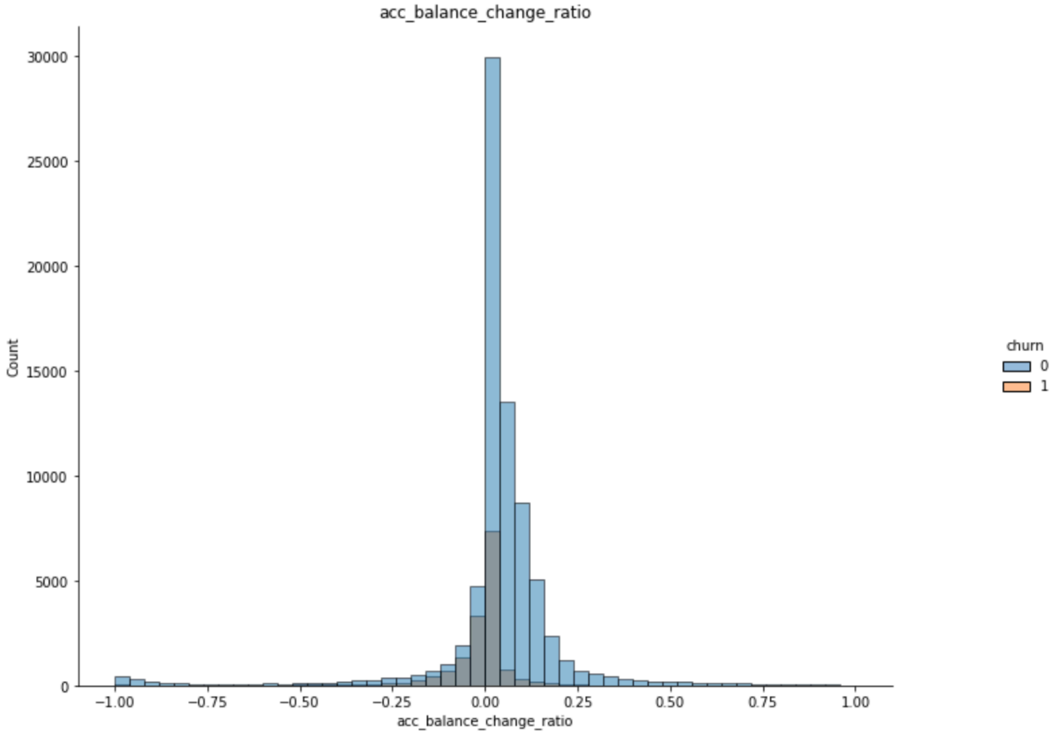


Figure 11 # Histogram for acc_balance_change_ratio

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.

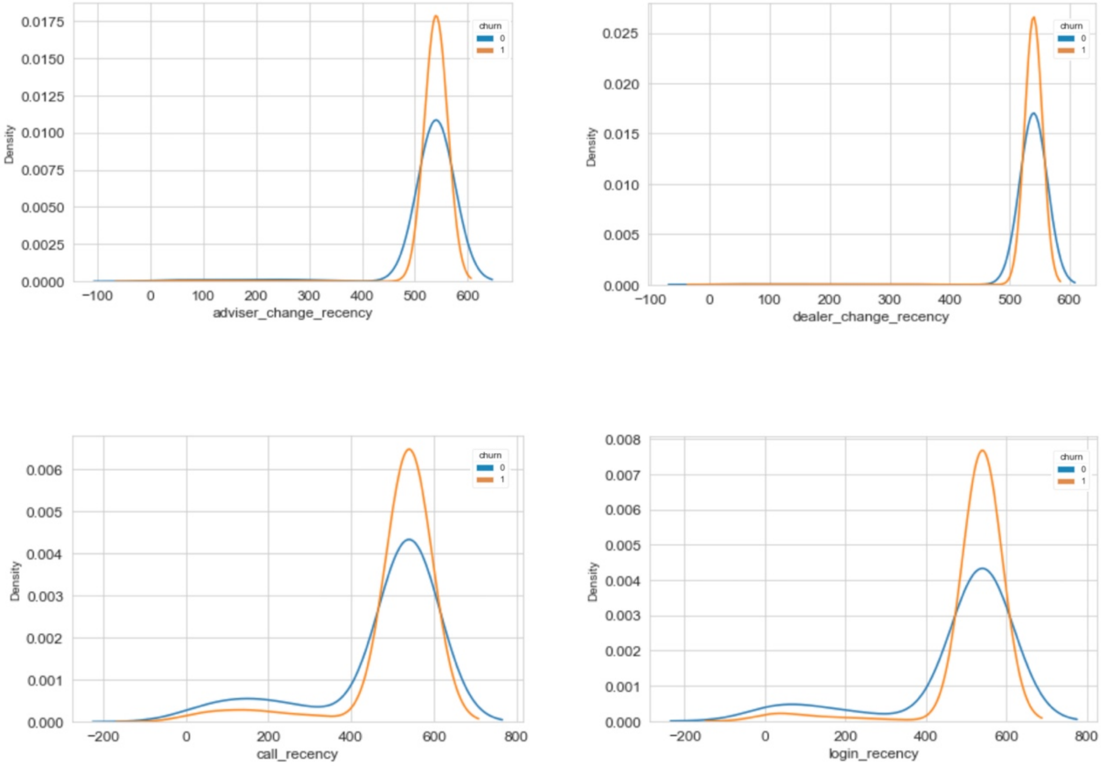
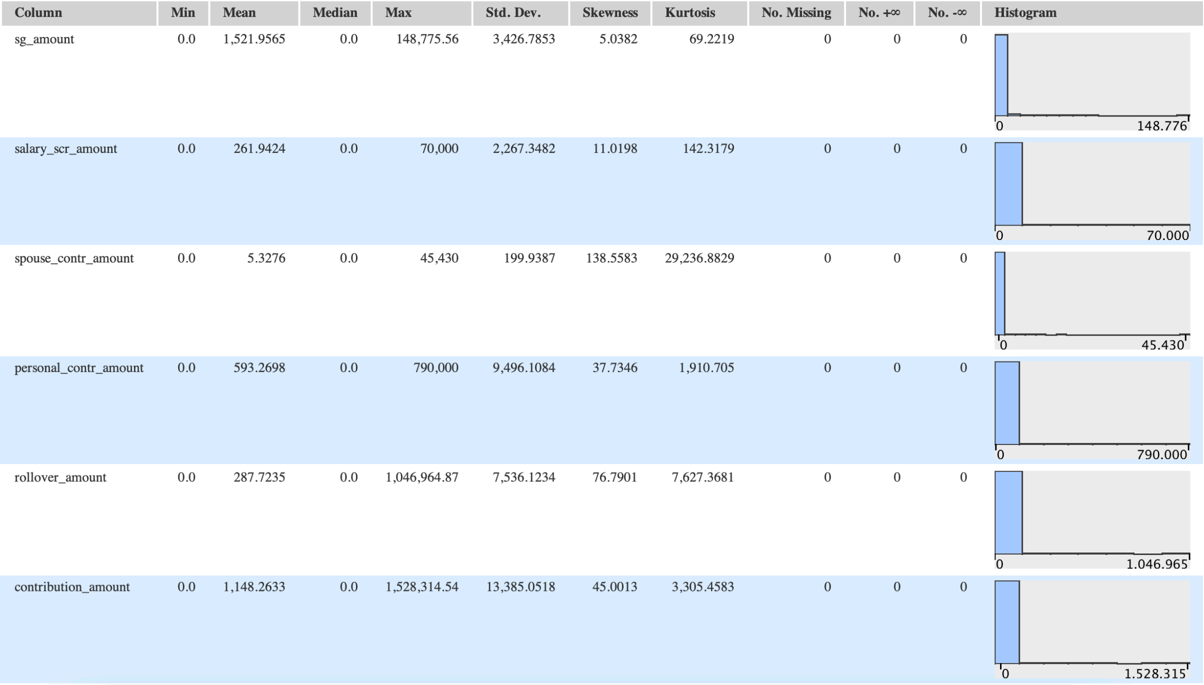


Figure 12 # Curve graph for adviser_change_recency, dealer_change_recency, call _recency, and login_recency

As shown from the statistics below (Refer to Figure 13), some contributions, such as SG contributions, personal contributions and spouse contributions, are generally low. Moreover, in comparing 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	0 148.776
salary_scr_amount	0.0	261.9424	0.0	70,000	2,267.3482	11.0198	142.3179	0	0	0	0 70.000
spouse_contr_amount	0.0	5.3276	0.0	45,430	199.9387	138.5583	29,236.8829	0	0	0	0 45.430
personal_contr_amount	0.0	593.2698	0.0	790,000	9,496.1084	37.7346	1,910.705	0	0	0	0 790.000
rollover_amount	0.0	287.7235	0.0	1,046,964.87	7,536.1234	76.7901	7,627.3681	0	0	0	0 1.046.965
contribution_amount	0.0	1,148.2633	0.0	1,528,314.54	13,385.0518	45.0013	3,305.4583	0	0	0	0 1.528.315

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

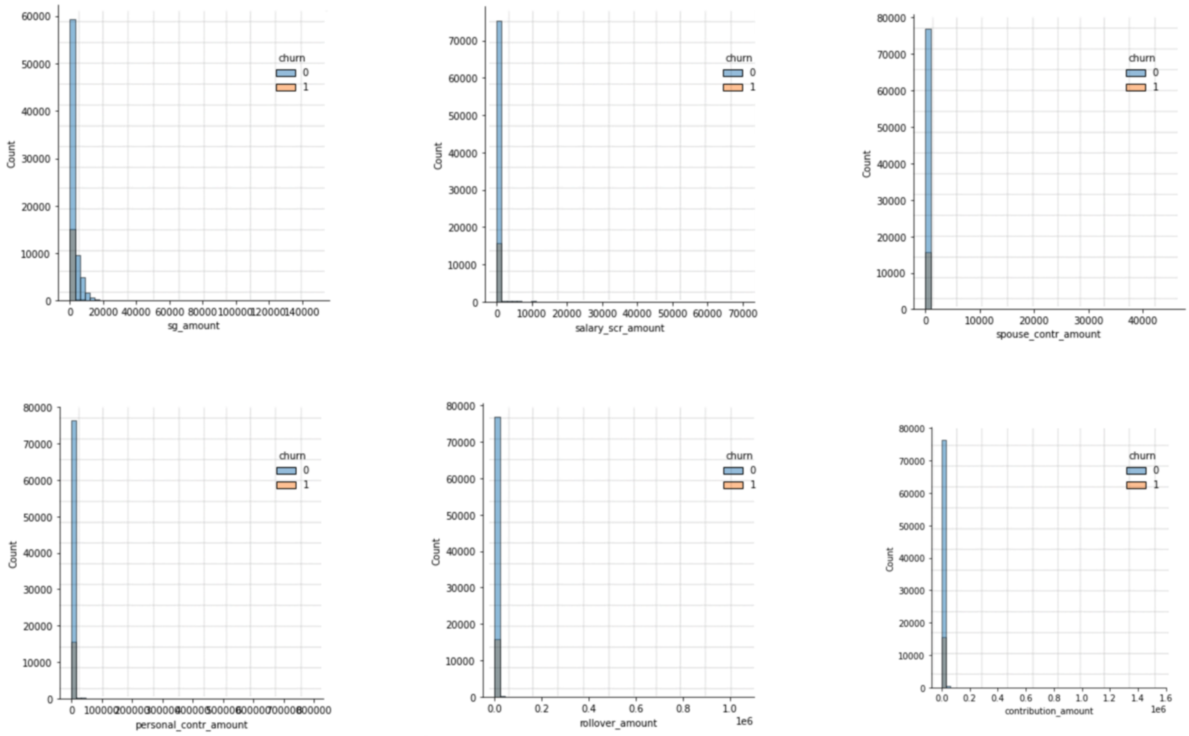


Figure 14 # Histogram for sg_amount, personal_contr_amount, rollover_amount, contribution_amount

In this heatmap below (Refer to Figure 15), it can be found that many attributes 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.

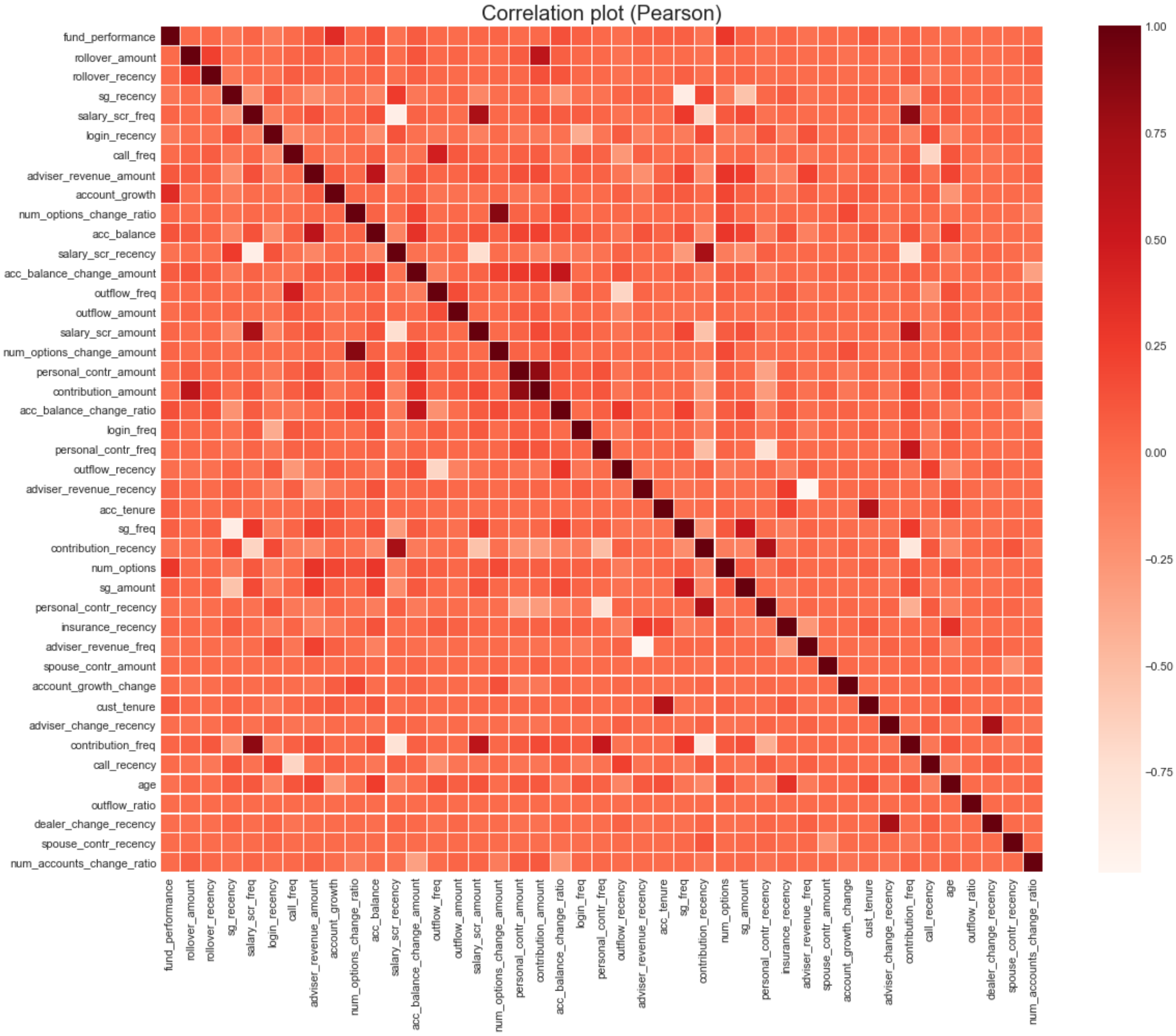
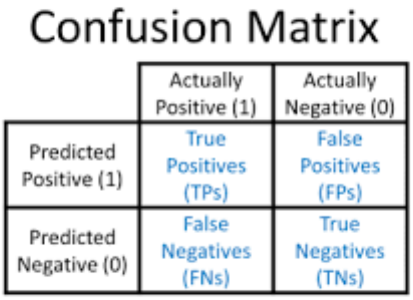


Figure 15 # Heatmap of Correlation for each attribute

4.0 Modelling

According to the project, we will first determine which evaluation metrics we will use to evaluate our model in this stage. Several essential measures for this issue area are sensitivity, specificity, Precision, F1 score, Geometric mean and Mathew correlation coefficient, ROC curve, and lastly, AUC score will be used to determined the performance of the models.

A confusion matrix is a table that is frequently used to describe the performance of a classification model (or "classifier") on a set of test data with known true values (Narkhede, 2018). The following table (Refer to Figure 16) is an example of the confusion matrix.



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 16 # Confusion Matrix

Condition positive (P)

- the amount of data points with true positive outcomes

Condition negative (N)

- the amount of data points with true negative outcomes

True positive (TP)

- A test result that accurately identifies the existence of a disease or feature.

True negative (TN)

- A test result that accurately identifies the absence of a disease or feature.

False positive (FP)

- A test result that incorrectly suggests the presence of a condition or attribute.

False negative (FN)

- A test result that incorrectly suggests the absence of a specific condition or feature.

This study will use several metrics, included accuracy, error rate, precision, Sensitivity, Specificity, ROC, F1 score, and Geometric Mean. The formula and description of each measurement will be displayed below.

Accuracy: Overall effectiveness of a classifier.

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN}$$

Error rate: Classification error.

$$\text{ERR} = \frac{FP+FN}{TP+TN+FP+FN}$$

Precision: Class agreement of the data labels with the positive labels given by the classifier.

$$\text{PRC} = \frac{TP}{TP+FP}$$

Sensitivity: Effectiveness of a classifier to identify positive labels.

$$\text{SNS} = \frac{TP}{TP+FN}$$

Specificity: How effectively a classifier identifies negative labels.

$$\text{SPC} = \frac{TN}{TN+FP}$$

ROC: Combined metric based on the Receiver Operating Characteristic (ROC) space.

$$\text{ROC} = \frac{\sqrt{SNS^2+SPC^2}}{\sqrt{2}}$$

F_1 score: Combination of precision (PRC) and sensitivity (SNS) in a single metric.

$$F_1 = 2\frac{PRC \cdot SNS}{PRC+SNS}$$

Geometric Mean: Combination of sensitivity (SNS) and specificity (SPC) in a single metric.

$$\text{GM} = \sqrt{SNS \cdot SPC}$$

AUC: AUC is area under the ROC curve which can also present the degree or level of separability of the curve. It indicates how well the model can discriminate between classes. The higher the AUC, the better the model predicts 0 classes as 0 and 1 courses as 1.

In this case, the value of precision, F_1 score, and AUC score will be the key metrics to determine the quality of a model.

Based on the dataset, the SMOTE (synthetic minority over-sampling technique) will be used to build the model, an integrated synthetic data algorithm for solving the Imbalanced class problem by combining Over-sampling minority classes and Under-sampling majority classes to synthesise data (Korstanje, 2021). Eleven modelling methods have been employed on the project, which are Naive Bayes, Logistic Regression, Decision Tree, RandomForest, AdaBoost, ExtraTrees, GradientBoosting, XGboost, Stack Ensemble (Hard Voting), Stack Ensemble (Soft Voting), and ANN Ensemble Classifier.

The voting ensemble, also known as the majority voting ensemble, was also employed in the project; it is a machine learning model that combines predictions from multiple distinct models, which is an approach that may be used to improve model performance and outperform any one model in the ensemble (Nair, 2021). The predictions from many models are combined in a voting ensemble. It can be used for regression or classification. In the case of regression, this entails taking the average of the models' predictions. When it comes to categorisation, each label's predictions are tallied together, and the label with the most votes is picked. In this case, two types of voting will be included: Hard Voting and Soft Voting.

- Hard Voting - Use models to predict the class with the largest sum.
  Each individual classifier votes for a class in hard voting (also known as majority voting), and the majority wins. In statistical terms, the ensemble's expected target label is the mode of the distribution of individually predicted labels.

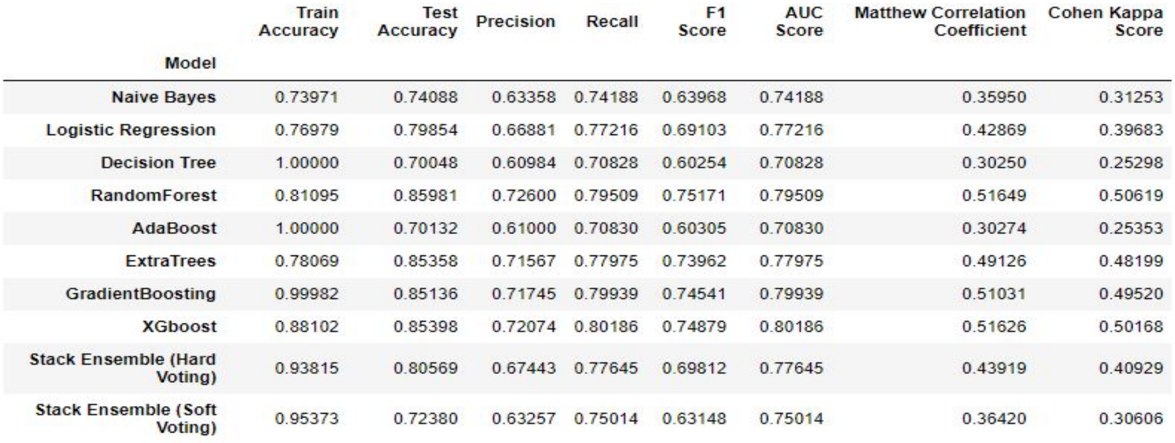
- Soft Voting - Use models to predict the class with the highest summed probability.
  Each individual classifier in soft voting delivers a probability value that a single data point belongs to a specified target class. The predictions are weighted by the significance of the classifier and totalled. The target label with the highest sum of weighted probability then receives the vote.

In addition, the Weighted Average Ensemble and Max Voting have also been used on the modelling.

Model averaging is an ensemble learning strategy in which each ensemble member contributes an equal amount to the final prediction (Brownlee, 2018). The ensemble prediction in regression is obtained as the average of the member forecasts. In the case of class label prediction, the prediction is calculated as the mode of the member predictions. The accuracy score and Cohen Kappa score of Weighted Average Ensemble are 0.806 and 0.409.

Meanwhile, each base model predicts and votes for each sample in max-voting. The final prediction class includes only the sample class with the most votes, resulting in accuracy and Cohen Kappa scores of 0.724 and 0.306, respectively.

Besides, the result of all used modelling methods is displayed in the following table and charts (Refer to Figure 17 & 18).



Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC Score	Matthew Correlation Coefficient	Cohen Kappa Score
Naive Bayes	0.73971	0.74088	0.63358	0.74188	0.63968	0.74188	0.35950	0.31253
Logistic Regression	0.76979	0.79854	0.66881	0.77216	0.69103	0.77216	0.42869	0.39683
Decision Tree	1.00000	0.70048	0.60984	0.70828	0.60254	0.70828	0.30250	0.25298
RandomForest	0.81095	0.85981	0.72600	0.79509	0.75171	0.79509	0.51649	0.50619
AdaBoost	1.00000	0.70132	0.61000	0.70830	0.60305	0.70830	0.30274	0.25353
ExtraTrees	0.78069	0.85358	0.71567	0.77975	0.73962	0.77975	0.49126	0.48199
GradientBoosting	0.99982	0.85136	0.71745	0.79939	0.74541	0.79939	0.51031	0.49520
XGboost	0.88102	0.85398	0.72074	0.80186	0.74879	0.80186	0.51626	0.50168
Stack Ensemble (Hard Voting)	0.93815	0.80569	0.67443	0.77645	0.69812	0.77645	0.43919	0.40929
Stack Ensemble (Soft Voting)	0.95373	0.72380	0.63257	0.75014	0.63148	0.75014	0.36420	0.30606

Figure 17 # the test results of each model

According to the evidence above, it can be summarised as the Random Forest, XGboost, and GradientBoosting have the better result among all the methods.

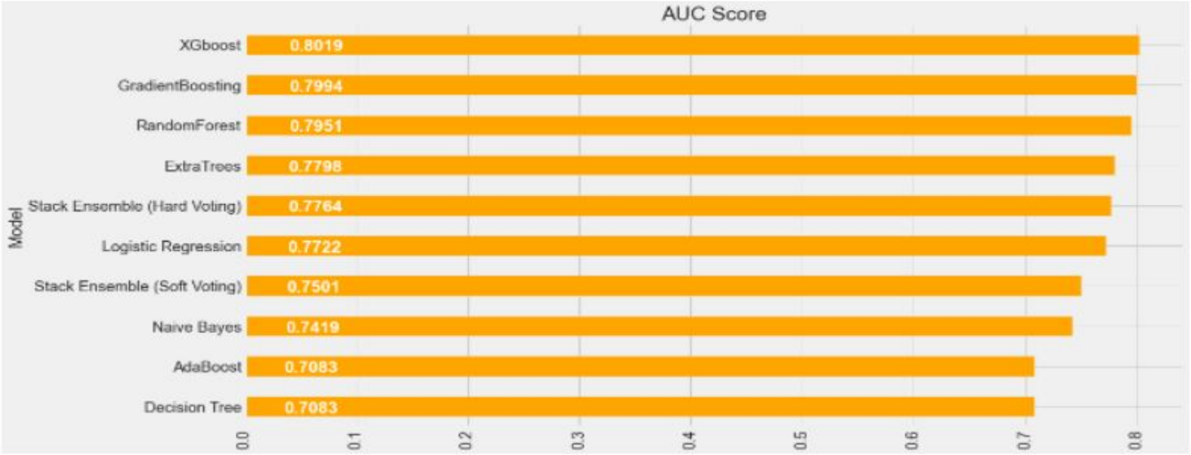
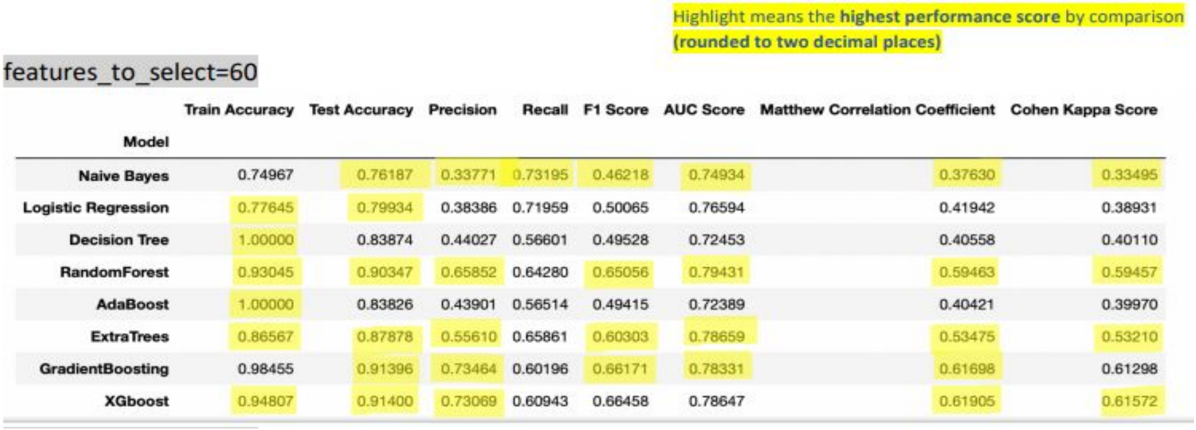


Figure 18 # the test results of each model

Furthermore, Recursive Feature Elimination (RFE) has been adopted is a feature selection algorithm with a wrapper that minimises model complexity by picking significant characteristics and discarding weaker ones (Kelley, 2022). RFE ranks feature based on the model's "coefficients" or "feature significance." It then eliminates a small number of features for each cycle, erasing any existing dependencies and collinearity in the model. RFE may pick from the training dataset those more or more significant characteristics to predict the target variable, lowering data complexity by deleting unnecessary features and enhancing model efficiency. The following table (Refer to Figure 19) shows the result of selecting different numbers of features.



Highlight means the **highest performance score** by comparison **(rounded to two decimal places)**

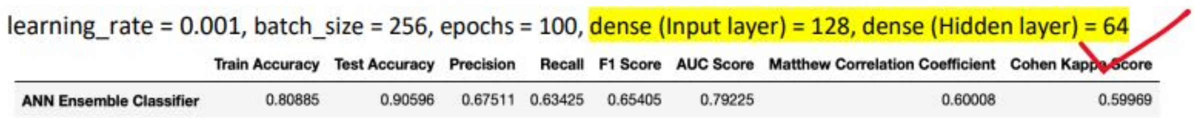
features_to_select=60

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC Score	Matthew Correlation Coefficient	Cohen Kappa Score
Naive Bayes	0.74967	0.76187	0.33771	0.73195	0.46218	0.74934	0.37630	0.33495
Logistic Regression	0.77645	0.79934	0.38386	0.71959	0.50065	0.76594	0.41942	0.38931
Decision Tree	1.00000	0.83874	0.44027	0.56601	0.49528	0.72453	0.40558	0.40110
RandomForest	0.93045	0.90347	0.65852	0.64280	0.65056	0.79431	0.59463	0.59457
AdaBoost	1.00000	0.83826	0.43901	0.56514	0.49415	0.72389	0.40421	0.39970
ExtraTrees	0.86567	0.87878	0.55610	0.65861	0.60303	0.78659	0.53475	0.53210
GradientBoosting	0.98455	0.91396	0.73464	0.60196	0.66171	0.78331	0.61698	0.61298
XGboost	0.94807	0.91400	0.73069	0.60943	0.66458	0.78647	0.61905	0.61572

Figure 19 # the results of each model's when features_to_select = 60

Based on the above tables with different features selected, it can be found that when the feature selection is about 60, the models will have a better performance on both F1 and AUC scores. Therefore, we adjusted the feature selection to 60, which will be explained in detail next section.

Additionally, we also add Artificial neural networks (ANNs) as one of the methods to construct the model. ANN is trained by analysing samples with a known "input" and "output," creating probability-weighted connections between the two stored inside the net's data structure (Techopedia, 2022). A neural network is typically trained from a given example by calculating the difference between the network's processed output (often a prediction) and target output. Unfortunately, this distinction is a mistake. The network then updates its weighted associations using this error value and a learning strategy. With each change, the neural network will create increasingly comparable output to the goal output. Following a sufficient number of these modifications, the training might be ended depending on specific criteria (Deepanshi, 2021). By testing the different parameters on learning_rate, bach_size, epochs, and dense of input and hidden layer, the best result (Refer to Figure 20) on precision, F1 Score, and AUC score of the ANN model have been shown below.

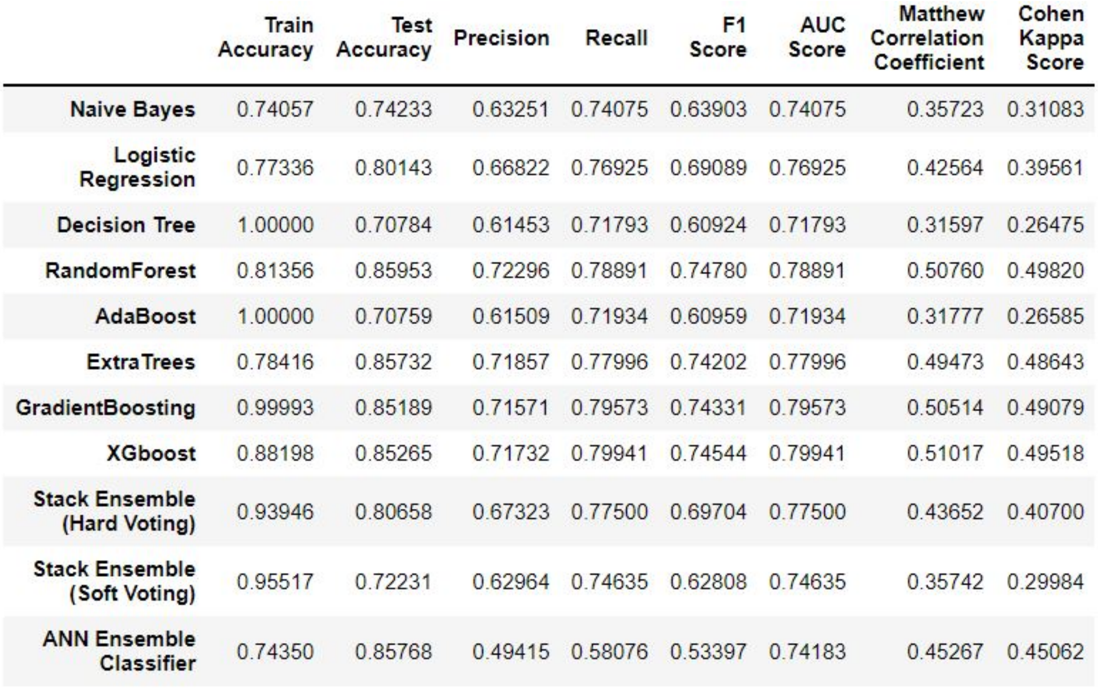


learning_rate = 0.001, batch_size = 256, epochs = 100, dense (Input layer) = 128, dense (Hidden layer) = 64

	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC Score	Matthew Correlation Coefficient	Cohen Kappa Score
ANN Ensemble Classifier	0.80885	0.90596	0.67511	0.63425	0.65405	0.79225	0.60008	0.59969

Figure 20 # the result of ANN Ensemble Classifier

In this case, a table (Refer to Figure 21) that contains each model's overall result has been published below.



	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC Score	Matthew Correlation Coefficient	Cohen Kappa Score
Naive Bayes	0.74057	0.74233	0.63251	0.74075	0.63903	0.74075	0.35723	0.31083
Logistic Regression	0.77336	0.80143	0.66822	0.76925	0.69089	0.76925	0.42564	0.39561
Decision Tree	1.00000	0.70784	0.61453	0.71793	0.60924	0.71793	0.31597	0.26475
RandomForest	0.81356	0.85953	0.72296	0.78891	0.74780	0.78891	0.50760	0.49820
AdaBoost	1.00000	0.70759	0.61509	0.71934	0.60959	0.71934	0.31777	0.26585
ExtraTrees	0.78416	0.85732	0.71857	0.77996	0.74202	0.77996	0.49473	0.48643
GradientBoosting	0.99993	0.85189	0.71571	0.79573	0.74331	0.79573	0.50514	0.49079
XGboost	0.88198	0.85265	0.71732	0.79941	0.74544	0.79941	0.51017	0.49518
Stack Ensemble (Hard Voting)	0.93946	0.80658	0.67323	0.77500	0.69704	0.77500	0.43652	0.40700
Stack Ensemble (Soft Voting)	0.95517	0.72231	0.62964	0.74635	0.62808	0.74635	0.35742	0.29984
ANN Ensemble Classifier	0.74350	0.85768	0.49415	0.58076	0.53397	0.74183	0.45267	0.45062

Figure 21 # the final test results of each model

To examined the results, three of the key metrics have been focused which are Precision, F1 score, and AUC score. In terms of Precision and F1 Score, Random Forest outperformed the other models, as seen in the table above. Although XGboost is a great model in the AUC Score assessment table, with a highest score among the models, Random Forest is more appropriate for the project after a comprehensive evaluation.

5.0 Experiment Analysis

5.1 Churn Prediction Results

Moreover, the Random Forest Variable Importance has been used to enhance the performance of the model. The mean reduction in impurity (or gini significance) mechanism is the default approach for computing variable importance: The improvement in the split-criterion at each split in each tree is the important measure ascribed to the splitting variable, and it is accumulated independently for each variable across all the trees in the forest (Lewinson, 2019).

In this case, SHAP (SHapley Additive exPlanations) has also been used to develop the prediction results of the model. SHAP is based on the Shapley value which compute the average marginal contribution of a feature x to a model score (Gopinath, 2021). It is uniqueness can be listed as follows:

- Completeness/Efficiency: the sum of each player feature's contribution must equal the entire model score minus the average payoff model score of the comparison group.
- Dummy: In the case of an ML model, a non-used input feature must be ascribed a contribution of zero.
- Symmetry: Characteristics contribute equally to the final model score.
- Monotonicity: If feature A consistently contributes more to the model score than feature B, the Shapley value of feature A must reflect this and be greater than the Shapley value of feature B.
- Linearity: Assume the game model score is decided by the sum of two intermediate values, each obtained from the input characteristics. Then, the contribution assigned to each feature must equal the sum of the feature's contributions to the intermediate values.

Then, SHAP can be explained as a visualisation tool that can be used to visually describe the output of a machine learning model which can also be used to explain any model's prediction by calculating the contribution of each feature to the prediction (Verma, 2021). The SHAP diagram could demonstrate how much each predictor contributes to the target variable, either positively or negatively. This is similar to the variable importance plot, except it may display the positive or negative association between each variable and the target (Kuo, 2019). The SHAP chart illustrated the following data:

- Variable importance: Variables are sorted in descending order.
- Impact: The horizontal placement indicates whether the influence of that number is related with a greater or lower prediction.

- Original cost: The colour indicates whether that variable is high (in red) or low (in blue) for that observation.

According to the SHAP values below (Refers to Figure 22), the most influential features which will be pushing the prediction higher such as age, account tenure, and the days since the last day of SG contribution. On the other hand, some features will be dropping the prediction lower which include the customers' account balance and its change amount.

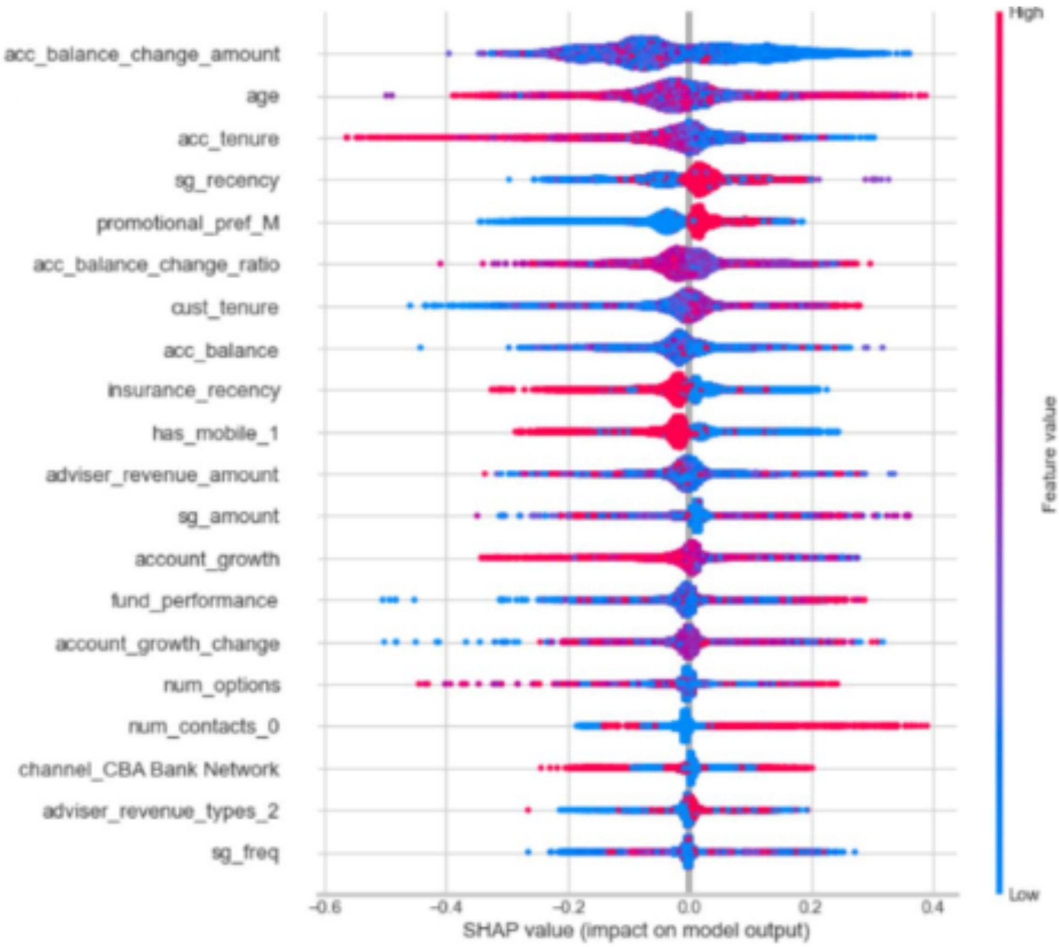


Figure 22 # the SHAP which will impact on model output

In order to construct a churned customer ranking based on a better comprehension of the dataset, we select all the active customers, about one forth of the total (68555/270000). In this case, two types of churners will be identified, and the criteria has been set up as follows:

High profitability churners are who those have higher value in the following features:

- "Contribution_frequency" > 6
- "Confidence (churned)" >= 0.8
- "Prediction(churn)" = churned

Low profitability churners are who those have lower value in the following features:

- “Contribution_frequency” <= 6
- “Confidence (churned)” < 0.8
- “Prediction(churn)” = churned

By following the scales above, 313 higher and 18930 lower profitability customers have been determine, key findings is shown below.

The two pie charts (Figure 23 & 24) below clearly show the gender distribution of higher and lower profitability customers, it is evident that the two charts present an opposite situation. For the customer who has high profitability, the distribution of the male (58%) is significantly more than the female (42%).

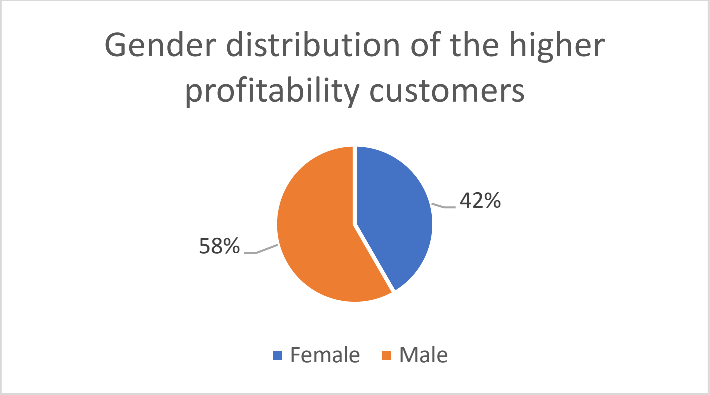


Figure 23 # Gender distribution of the higher profitability customers

However, the data presents a different way for the second chart, in which female takes about 57% of the total low profitability customers, and the male takes about 43%.

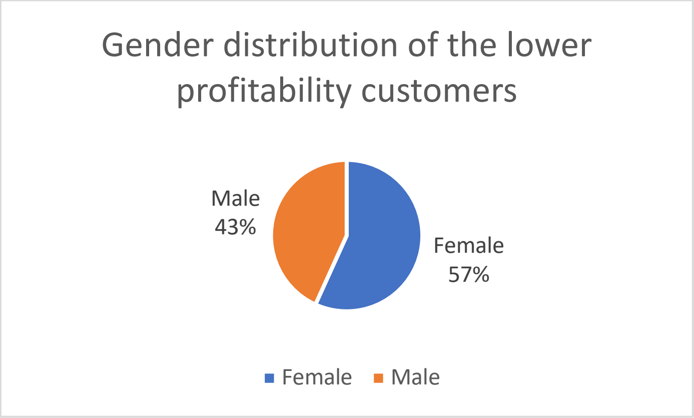


Figure 24 # Gender distribution of the lower profitability customers

To sum up, the female has a higher probability of being a lower profitability customer than a male.

Moreover, the following column charts (Refer to Figure 25 & 26) illustrated the age distribution of the higher and lower profitability customers. As we can see from the first chart, the most outstanding share of customers between the ages of 51 and 60, reaching 35 per cent of the total higher profitability customers. By contrast, only three customers who have higher profitability were under the age of 30. On the other hand, the distribution of each age group of the low profitability customers will be more even than the higher one, which is between 21% and 25%. However, the customers who are less than 30 years of age only take up 5% of the category's total.

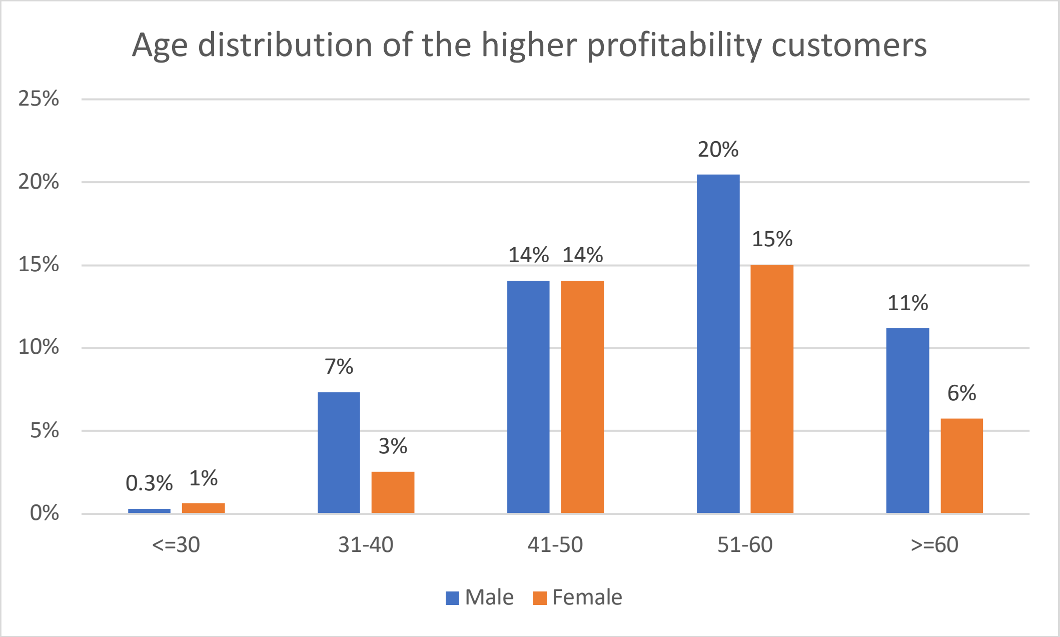


Figure 25 # Age distribution of the higher profitability customers

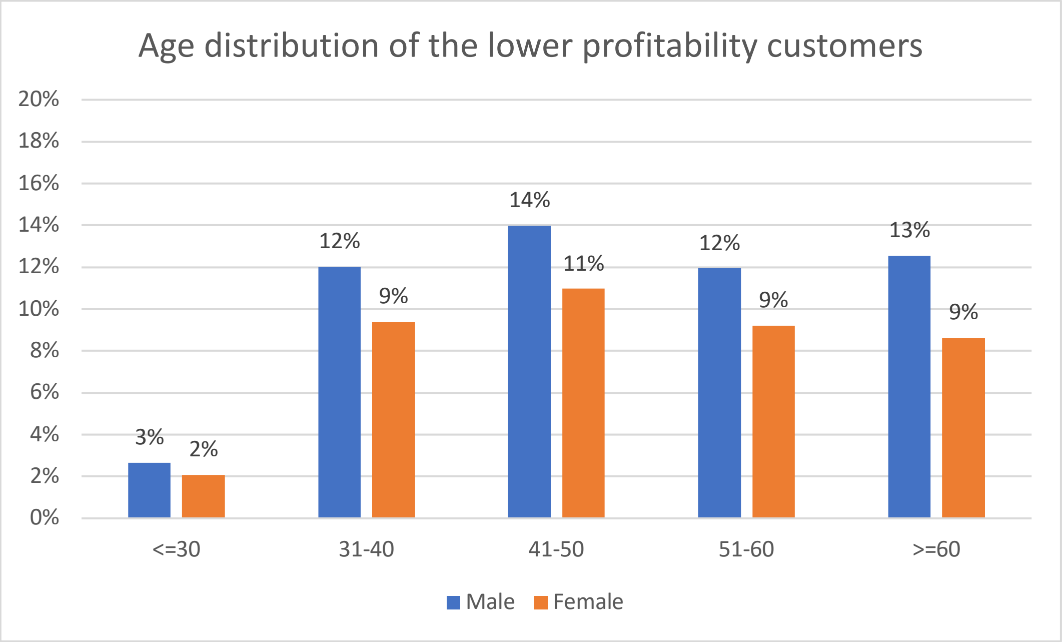


Figure 26 # Age distribution of the lower profitability customers

5.2 Causal Inference Analysis

Causal inference is known as the process of determining the independent, real consequence of a specific phenomena that is a component of a broader system which will study on the assumptions, designs, and estimation procedures that allow researchers to make causal inferences from data. The purpose of causal inference is to assess the causal influence of treatment (a choice or action) on the outcome (i.e., the result of treatment). In this case, the Casual Analysis Algorithm process is as follows:

- Identify the causal effect
- Assess the estimated effect on the outcome
- Measure causal effect estimation
- Refute the obtained estimate
- Our assumption is correct if the new estimate effect value is NOT changed

During the process of the Causal Analysis, the ‘DoWhy’ Python library as a necessary method to employed which aims to simplify the use of causal reasoning in machine learning applications (Rodriguez, 2020). In this case, the following four essential phases have been used to describe any causal inference problem in a workflow:

- Model: DoWhy uses a network of causal linkages to model each situation.
- Identify: DoWhy searches the input graph for all potential methods to identify a desired causal impact based on the graphical model. It employs graph-based criteria as well as do-calculus to identify possible expressions that might identify the causal impact.
- Estimate: DoWhy calculates the causal influence using statistical techniques such as matching or instrumental factors. The current version of DoWhy supports estimate methods that focus on estimating the treatment assignment, such as propensity-based-stratification or propensity-score-matching, as well as regression approaches that focus on estimating the response surface.
- Verify: DoWhy employs several robustness measures to confirm the validity of the causal effect.

Based on the table (Refer to Figure 27) of the result of the causal analysis, some some variables have a visible impact on churn. For instance, the annual report preference and the days since the last day of SG contribution will increase the probability of churn by about 15 percent. On the other hand, the statement preference and customer communication preference will influence the churn rate positively which will drop the churn probability by 14 and 8 per cent, respectively.

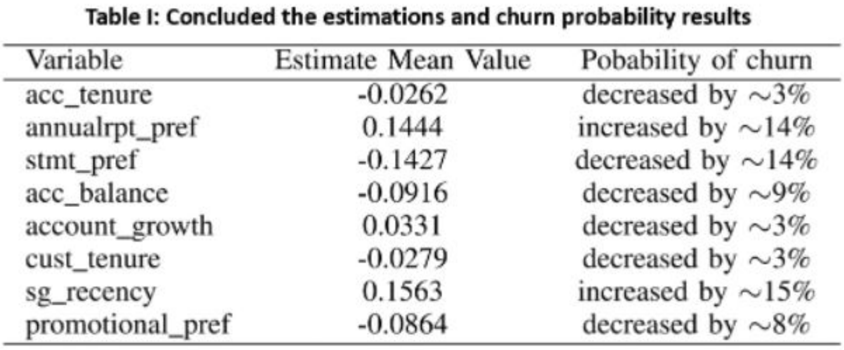


Table I: Concluded the estimations and churn probability results

Variable	Estimate Mean Value	Pobability of churn
acc_tenure	-0.0262	decreased by ~3%
annualrpt_pref	0.1444	increased by ~14%
stmt_pref	-0.1427	decreased by ~14%
acc_balance	-0.0916	decreased by ~9%
account_growth	0.0331	decreased by ~3%
cust_tenure	-0.0279	decreased by ~3%
sg_recency	0.1563	increased by ~15%
promotional_pref	-0.0864	decreased by ~8%

Figure 27 # the estimations and churn probability results of causal analysis

6.0 Recommendations

In response to the business problems faced by CFS, we recommend that CFS can implement the following measures:

- **Proactively communicate with customers**

Sometimes, existing customers may have lost sight of the value of a product or service as the initial novelty wears off. Maintaining effective communication with customers can remind customers of product value or recent feedbacks.

- **Define the most valuable customers to the company**

By scoring and categorising each member's churn rate for the next fiscal year. Divide the churned customers into high-profit members and low-profit members. For high-profit members, CFS can offer their best rewards, such as service discounts, regular customer events, and so on. For low-profit members, the company encourages them to increase their business use.

- **Increase customer engagement**

Increase customer engagement through appropriate strategies. Focusing on customer satisfaction and retention increases customer engagement and fosters long-term relationships. This may involve regular communication through product-led content, blogs, newsletters, and more to reinforce what they can get out of your product or service.

- **Define a roadmap for new customers**

Some people who start using a new product or service may not understand the product's features or value. If the customer cannot figure out the benefits or services of this product in the first place. They may lose interest very quickly. To ease the transition, a roadmap for setting up new customers can guide new customers through the features of a company's product or service, future value, and more.

- **Collect customer feedback**

  Companies can identify customers who are unhappy with their business and understand why they feel that way, leading to further improvements.

7.0 Deployment

Our project plans and deadlines follow the CRISP-DM methodology and we are presently in the deployment phase. This project aims to find out the cause of customer churn and countermeasures to reduce customer churn, aiming to increase profitability for our client (CFS). Therefore, instead of deploying data models into a working environment, we use appropriate data analysis methods and provide our client with the findings on the cause of customer churn.

The data visualisation provides a better understanding of CFS's customers. Furthermore, Causal Bayesian Networks provide prediction of cause probabilities that lead to customer churn, we also employed Bayesian causal graphs to encode assumptions and determine dependency levels between features. Based on the findings, our client will gain deep knowledge of the cause of customer churn. We also provided numerous recommendations that CFS can take into consideration to reduce customer churn. To ensure the deployment plan can run simultaneously, we provide regular monitoring and maintenance to fix any bugs and update the database in time.

Reflecting upon previous experience is essential in order to develop better communication skills, conflict resolution and improve future performance. Our team comprised four members, during this project, each member gained a different level of professional skills. The group also learned that good teamwork with effective communication brings the key success of the project, moreover, having the right people in the correct roles is also an important factor. Overall, project reflection enables such observations to help with performance and enhance discipline-specific knowledge.

8.0 Challenges Encountered

1. One of the main challenges encountered was the pre-processing of the dataset. Due to the large volume and complexity of the dataset, performing data pre-processing and data cleaning become a challenge. Without adequately preparing the dataset, it is difficult for further analysis. We used data pre-processing techniques like One-Hot Encoding and normalisation to remove redundant data, accelerated the speed, and improved the model’s performance.

2. Unbalanced observation causes a class imbalance problem in the dataset. In this case, we applied the SMOTE method for data training. It can analyse 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.

3. In model training, improving model performance has become an inevitable problem. In this project, we employed voting ensembles to improve the model's performance. 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 also use feature selection to remove inconsequential features, which enhances model performance even further.

4. Time-consuming feature engineering in a massive financial dataset with High-Dimensional Sparse Data. We used the intrinsic property of Artificial Neural Network (ANN) based on different weights of given historical data.

5. Improve prediction confidence in deep learning classification (the probability that a customer predicted as a churner is indeed churned). We propose a casualty analysis method with the use of DoWhy to predict the set of cause that lead to deliberate churn.

6. Provide a measure of uncertainty in the prediction of customer churn rate. We calculate magnitude of causes by exploiting counterfactual conditions prediction method.

   Magnitude of cause = Actual Probability - Counterfactual Probability

9.0 Conclusion

Customer churn simply refers to terminating a relationship with a business, gaining new customers, and retaining existing customers are two main business marketing strategies. However, customer acquisition typically costs way higher than customer retention. Hence, our company is entrusted to analysed possible customer churn causes for a superannuation fund company over the next six months.

Based on the data exploration and pre-processing, we transform the data into a more effective format for our further analysis, follow by data visualisation, the result clearly shown that the dataset is uneven distributed on a 14.2% churn customer compared to 85.8% retaining customer, and this has become a challenge for us, to solve the data imbalance problem, we applied SMOTE method for data training to improve the prediction ability of the minority class, unbalanced churned, and non-churned classes were levelled in pre-processing. A new propensity model was designed and integrated with the causal Bayesian networks to predict cause probabilities that lead to customer churn.

Throughout the modelling stage, eleven modelling methodologiess were employed and evaluation metrics on test data confirm that RandomForest, XGBoost and GradientBoosting have the best results based on the AUC F1 Score. Moreover, the Random Forest Variable Importance confirms that acc_balance_change_amount, acc_balance_change_ratio and acc_balance have the most significant impact on customer churn. Furthermore, ANNs have also been applied as one of the methods in constructing the model. Finally, it has been confirmed that GradientBoosting has the best performance among others based on precision, F1 and AUC Score.

Under experiment analysis, we focused on the impact of a different number of feature selections impact on model performance. We have tested four different numbers of feature selection: 60, 90, 120, 150. As a result, we decided to use the feature choices of 60 as the model's accuracy performs the highest, F1 and AUC Score are also acceptable under feature choice of 60. To maximise the value of AUC, F1 Score and Accuracy in ANN Classifier, we have made changes to some of the settings. By changing learning_rate for 0.00001, 0.001, 0.01, 0.1, it's confirmed that 0.001 is the best learning_rate compared to others. Identically, by changing batch_size and epochs, changing_dense, the accuracy, F1 and AUC Score has increased, and it's summarised that GradientBoost is the most appropriate to build the model. Furthermore, we have setup high and low profitability churners for casual analysis; as a result, the distribution of the males (58%) compared to females (42%) in high profitability churners. However, the result is completely opposite in low profitability churner of females (57%) compared to males 43%.

Causal analysis results confirmed variable representing recent SG contribution, annual report preference changed, account growth and balance amount were identified as confounding factors for customer churn with a high degree of belief. Furthermore, churn rate can be reduced by approximately 3% for customer with active account longer than one year, consistent with expert knowledge.

In the deployment stage, based on all the results and findings we gathered from our analysis, we suggested several methods that CFS can implement to reduce the customer churn rate. Overall, project reflection enables such observations to help with performance and enhance discipline-specific knowledge.

10.0 Reference List

- Brownlee, J. (2018, December 28). *How to Develop a Weighted Average Ensemble for Deep Learning Neural Networks*. MachineLearningMastery. https://machinelearningmastery.com/weighted-average-ensemble-for-deep-learning-neural-networks/

- Deepanshi. (2021, May 25). *Beginners Guide to Artificial Neural Network*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/beginners-guide-to-artificial-neural-network/

- Gopinath, D. (2021, October 26). The Shapley Value for ML Models. Towards Data Science. https://towardsdatascience.com/the-shapley-value-for-ml-models-f1100bff78d1

- Kelley, K. (2022, February 9). *Recursive Feature Elimination: What It Is and Why It Matters*. Simplilearn. https://www.simplilearn.com/recursive-feature-elimination-article

- Korstanje, J. (2021, August 30). *SMOTE*. Towards Data Science. https://towardsdatascience.com/smote-fdce2f605729

- Kuo, C. (2019, September 14). *Explain Your Model with the SHAP Values*. Towards Data Science. https://towardsdatascience.com/explain-your-model-with-the-shap-values-bc36aac4de3d

- Lewinson, E. (2019, February 12). *Explaining Feature Importance by example of a Random Forest*. Towards Data Science. https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e

- Nair, A. (2021, October 12). *Combine Your Machine Learning Models With Voting*. Towards Data Science. https://towardsdatascience.com/combine-your-machine-learning-models-with-voting-fa1b42790d84

- Narkhede, S. (2018, May 9). *Understanding Confusion Matrix*. Towards Data Science. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

- Rodriguez, J. (2020, August 28). *Microsoft's DoWhy is a Cool Framework for Causal Inference*. KDnuggests. https://www.kdnuggets.com/2020/08/microsoft-dowhy-framework-causal-inference.html

- Techopedia. (2022, April 30). *Artificial Neural Network (ANN).* https://www.techopedia.com/definition/5967/artificial-neural-network-ann

- Verma, Y. (2021, December 25). *A Complete Guide to SHAP – SHAPley Additive exPlanations for Practitioners.* Analyticsindiamag. https://analyticsindiamag.com/a-complete-guide-to-shap-shapley-additive-explanations-for-practitioners/