文本

描述已自动生成

**Summary**

White Rock is a renowned asset management company that hopes to enhance digital capabilities facing a rapidly changing market, especially for the sales and marketing department to better understand customer churn behaviors. Our project plans to take advantage of a public dataset for machine learning to identify which factors are significantly associated with customer churn and key factors leading to an increase in purity between retained and churned customers.

In order to predict customer churn, we did machine learning and tried to find important features which are correlated to the customer churn. We first did data preparation on our dataset, which included data cleaning, normalization and sampling. Data preparation used stratified sampling with a 70%-30% train-test split with the relevant one-hot encoding, Z-score normalization and SMOTENC oversampling to balance the trainset. After that, feature engineering was performed. Logistic regression, decision tree and mutual information were chosen for feature engineering and profiling. During feature engineering, estimatedSalary was considered not significant while age is the most significant feature. Then, the reduced dataset with the best feature groups was fed into three different machine models for prediction. Models chosen were neural network, random forest and XG boost. Our best model, random forest, achieved 63% recall, 0.840 auc and 82% accuracy in testing.

Based on the prediction model, we got the variable importance outcome as the key factors of customer churn, in which we find age, isactivemember, Germany, gender, hasacard are the main cause of churn. Integrated with White Rock’s business circumstance, we gave reasonable analysis and solutions on the churn situation.

Customer churn can impose huge impacts on White Rock business as well as on the entire industry. The immediate and direct revenue loss would come from a loss of recurrent revenues, potential upsell opportunities, expected ROI due to disruption of ongoing investment and negative network effect. Subsequent costs following customer acquisition and retention activities can be more considerable. Additional costs may incur for launching campaigns and depending on the choice of prediction models, different groups are targeted, which sometimes may lead to zero revenues and even unnecessary costs. Compromise on investment strategy has to be made to cater for high liquidity demand. To outperform in the highly competitive marketplace, oftentimes revenues are sacrificed for customer growth. Additionally, awareness to apply machine learning and big data technology is raised to predict and reduce churn in the most cost-effective way.

Our model and analysis can bring several added value to asset management companies like White Rock. Above all, it can identify important factors influencing customer churn in a data-driven precise. The prediction of customer churn can give the company more time to take measures in advance. When applying the model to operation, the real-time analysis can better arrange the relationship between customer and company.

Measures to reduce the cost of customer churn can be divided into two aspects. On one side, White Rock should raise customer the retention rate to prevent loss. As regards the cost of customer churn already happened, it should either retain back lost customers or attract new customers and deal with negative brand implications as soon as possible.

**Contents**

[1. Motivation 1](#_Toc52810050)

[1.1 Background 1](#_Toc52810051)

[1.1.1 Internal corporate perspective: The meaning of digital market transformation 1](#_Toc52810052)

[1.1.2 Current contextual and global perspective: Coexisting challenges and opportunities 1](#_Toc52810053)

[1.2 Problem Statement 1](#_Toc52810054)

[1.2.1 Predict whether the customer wants to churn by machine learning model 1](#_Toc52810055)

[1.2.2 Decide the key factors that lead to client churn based on prediction model 2](#_Toc52810056)

[1.2.3 Measure the loss deal to client churn and offer solutions based on digital techniques 2](#_Toc52810057)

[1.2.4 Analyze impacts of client churn and offer solutions based on digital techniques 2](#_Toc52810058)

[2. Design 2](#_Toc52810059)

[2.1 Overview 3](#_Toc52810060)

[2.2 Dataset 3](#_Toc52810061)

[2.2.1 Source 3](#_Toc52810062)

[2.2.2 Overview 3](#_Toc52810063)

[2.2.3 Feasibility 3](#_Toc52810064)

[2.2.4 Visualization 3](#_Toc52810065)

[3. Experiment 4](#_Toc52810066)

[3.1 Data preparation 4](#_Toc52810067)

[3.1.1 Data Cleaning 4](#_Toc52810068)

[3.1.2 Data Normalization 5](#_Toc52810069)

[3.1.3 Data Sampling 5](#_Toc52810070)

[3.1.4 Dummy Variable 5](#_Toc52810071)

[3.2 Feature Engineering 6](#_Toc52810072)

[3.2.1 Logistic Regression 6](#_Toc52810073)

[3.2.2 Decision Tree 6](#_Toc52810074)

[3.2.3 Mutual Information 7](#_Toc52810075)

[3.2.4 Feature Selection and Profiling 8](#_Toc52810076)

[3.3 Model Training 8](#_Toc52810077)

[3.3.1 Random forest 8](#_Toc52810078)

[3.3.2 XGBoost 8](#_Toc52810079)

[3.3.3 Artificial Neural Network (ANN) 8](#_Toc52810080)

[3.4 Result Analysis 8](#_Toc52810081)

[4.Discussion 10](#_Toc52810082)

[4.1 Cause of churn 10](#_Toc52810083)

[4.1.1 Age 10](#_Toc52810084)

[4.1.2 Is active member 10](#_Toc52810085)

[4.1.3 Germany 11](#_Toc52810086)

[4.1.4 Gender 11](#_Toc52810087)

[4.2 Costs of churn and corresponding impacts 11](#_Toc52810088)

[4.2.1 Costs of churn & impacts: Company-wise 11](#_Toc52810089)

[4.2.2 Costs of churn & impacts: Industry-wise 13](#_Toc52810090)

[4.3 Added values bring to the White Rock 14](#_Toc52810091)

[4.3.1 Identify important factors influencing customer churn in a data-driven precise 14](#_Toc52810092)

[4.3.2 Predict customer churn and take measures in advance 14](#_Toc52810093)

[4.3.3 Real-time analysis about operational performance in customer relationship management 14](#_Toc52810094)

[4.4 Measures to reduce bad influence and cost of customer churn 14](#_Toc52810095)

[4.4.1 Raise customer retention rate 15](#_Toc52810096)

[4.4.2 Decrease cost of customer churn already happened 15](#_Toc52810097)

[5. Conclusion 15](#_Toc52810098)

[6. Reference 16](#_Toc52810099)

**Table Directory**

[Table 1 Hyperparameters of Random Forest 7](file:////Users/tangmanhong/Desktop/AN6003_Group_Project_Report.docx#_Toc52668080)

[Table 2 Hyperparameters of XGBoost 8](#_Toc52668081)

**Figure Directory**

[Figure 1 Flow Chart 3](#_Toc52668107)

[Figure 2 Distribution of Churn 3](#_Toc52668108)

[Figure 3 Churn versus Other Variables 4](#_Toc52668109)

[Figure 4 Result of Logistic Regression 6](file:////Users/tangmanhong/Desktop/AN6003_Group_Project_Report.docx#_Toc52668110)

[Figure 5 Result of Decision Tree 6](file:////Users/tangmanhong/Desktop/AN6003_Group_Project_Report.docx#_Toc52668111)

[Figure 6 Mutual Information Result 7](file:////Users/tangmanhong/Desktop/AN6003_Group_Project_Report.docx#_Toc52668112)

[Figure 7 ANN Testing Result 8](file:////Users/tangmanhong/Desktop/AN6003_Group_Project_Report.docx#_Toc52668113)

[Figure 8 Random Forest Testing Result 8](file:////Users/tangmanhong/Desktop/AN6003_Group_Project_Report.docx#_Toc52668114)

[Figure 9 XGBoost Testing Result 9](file:////Users/tangmanhong/Desktop/AN6003_Group_Project_Report.docx#_Toc52668115)

# 1. Motivation

## 1.1 Background

White Rock, a renowned global asset management company with over 300 years of servicing experience in managing clients’ assets through various investment products, aims to deliver the best service to its clients to further guarantee sales excellence in a fast-changing and digitalized world. White Rock has to take advantage of cutting-edge analytics to facilitate accurate commercial decision-making for the benefit of employees and clients by better understanding its clients’ needs, its internal operations and streamlining operations for better operational efficiency.

### 1.1.1 Internal corporate perspective: The meaning of digital market transformation

As the market gain is the fundamental means for a company to sustain its operations and thrive, a prompt and relevant digitalization of sales and market distribution strategy is of great urgency that should be incorporated into the digital blueprint of White Rock as many of its competitors will also be embarking on their own development of digital transformation. Furthermore, facing a rapidly changing market, White Rock must remain competitive by capitalizing on the recent “data revolution”, harnessing greater insights from relevant information hidden in the proliferation of vast amounts of data available to companies today. This will enable White Rock to make more informed and better decisions that can contribute to increasing revenue and streamlining operations.

A good digital and analytics strategy has great potential to empower every aspect of White Rock’s sales and marketing department’s business process from customer segmentation, demand analysis to marketing deliverables and sales strategy. Data analysis and analytics can help segment White Rock's clients into groups, understand what key factors cause client churn and reveal the key driving factors of demand of customers. Such insights aim to deliver correct messages to the sales team and get them ready to engage with identified target clients having a clearer strategy and plan, hence reducing the cost of inefficient communication and redundant service deliveries.

### 1.1.2 Current contextual and global perspective: Coexisting challenges and opportunities

Digitization and data analytics initiatives explored by companies prior to COVID-19 have only intensified and accelerated as a result of COVID-19. The slowdown in business activities due to economic lull and supply chain disruptions as a result of COVID-19 offers a “downtime” that White Rock should make use of to enhance its digitization and data analytics capabilities before the market picks up again. Hence, now is a good time for White Rock to channel effort and resources into exploring and developing its own digital and analytics capabilities across relevant departments as a form of promising investment in itself.

## 1.2 Problem Statement

### 1.2.1 Predict whether the customer wants to churn by machine learning model

What is the suitable model to describe the churn behaviors for White Rock? How do we design the input and output for the models, adjust the parameters to get the best performance? Based on the prediction model, what kind of indicators should we use to evaluate the churn behaviors? We would select several machine learning methods and ask what is a good threshold when measuring the efficacy of the model. This will have to be in sync with information about the sales and marketing department’s overheads and key performance indicators (KPIs).

### 1.2.2 Decide the key factors that lead to client churn based on the prediction model

In order to reduce customer churn, increase retention rate and acquisition rate, we have to find out what kind of cause is significantly associated with customer churn and key factors leading to an increase in purity between retained and churned customers. In addition, the goal of the company's operations is to achieve optimal profitability. There is only a limited amount of time and resources that the sales and marketing department has and they have to correctly identify and focus on the key client type they need to target in their upcoming sales and marketing campaign.

### 1.2.3 Measure the loss deal to client churn and offer solutions based on digital techniques

Customer churn is a painful reality that White Rock has to deal with. After understanding what causes customer churn, we would give a more specific measure to help evaluate the cost of churn. Not only would help White Rock understand what is the cost of churn, but also what we can help to change the situation and gain back profit.

### 1.2.4 Analyze impacts of client churn and offer solutions based on digital techniques

Considering the status of White Rock as a globally well-known company, we would understand the effect of client churn for White Rock as well as the asset industry, through which we hope to give a comprehensive vision of interaction in the asset industry and how White Rock acts an important role in it.

**1.3 Project Objectives**

There are several key objectives this paper aiming at:

* Make up standardized data visualization analysis, so as to improve work efficiency and enhance digitalization for the company
* Set up a standardized model process, design specification for input and output data, reducing the repetitive manual work of data collection, data processing and data analyst.
* Find out the key factors to decrease client churn rate, so as to improve sales performance.
* Detect the loss and impact of customer churn for White Rock and asset industry, offer professional solutions.

# 2. Design

## 2.1 Overview

In order to solve White Rock’s business problem, our project will take advantage of a public dataset for machine learning to identify which factors are significantly associated with customer churn and key factors leading to an increase in purity between retained and churned customers. With data analysis, we will classify customers based on whether they exit or not and explore the mechanism beyond the results under business scenario and the costs of customer churn on the company and the entire asset management industry.

图示

描述已自动生成

Figure 1 Flow Chart

## 2.2 Dataset

### 2.2.1 Source

The dataset link:<https://www.kaggle.com/shrutimechlearn/churn-modelling>.

### 2.2.2 Overview

Churn\_Modelling is a public dataset consisting of details of 10000 bank's customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he continues to be a customer.

### 2.2.3 Feasibility

The key reason why we choose to use the Churn\_Modelling dataset to present our idea to solve this problem is its high applicability under this business scenario. Both banks and asset management companies belong to the financial industry, and Banks also provide some asset management services. Therefore, through the data analysis of bank customers, we can obtain some insights and inspiration applicable to the asset management company.

### 2.2.4 Visualization

#### 2.2.4.1 Distribution of Churn

图表, 饼图

描述已自动生成

Figure 2 Distribution of Churn

Since customer churn is the dependent variable in our project, we pay close attention to the distribution of customer churn in our data sample. As shown in the figure, the number of exited customers is one-quarter of the number of retained customers, which requires us to balance the data in the process of data cleaning and preparation.

#### 图表, 雷达图 描述已自动生成2.2.4.2 Churn versus Other Variables

Figure 3 Churn versus Other Variables

From the joint distribution of Churn and other variables, we can see that elder, female and people with more balance exits more. In the mid tenure level, there is less exit. Besides, EstimatedSalary and Creditscore have similar distribution for exited customers and retained customers.

# 3. Experiment

## 3.1 Data preparation

### 3.1.1 Data Cleaning

The data set mentioned in section 2.1 was used. The target variable is a binary categorical variable reflecting whether the customer has churned or remains to be a customer of the bank. The set does not contain any missing values or unknown value placeholder values. However, data pre-processing has to be done in order to prepare the data for feature engineering and machine learning. As the dataset is highly imbalanced, about 80% of the clients did not churn and 20% of the clients churned. We use stratified sampling and 70%-30% splitting to split the whole data set into the training set and testing set respectively with seed set. The proportion of clients who did not churn to clients is proportionally evenly distributed in both training and testing sets.

### 3.1.2 Data Normalization

Normalization has to be carried out for the continuous features as we intend to perform Synthetic Minority Over-sampling Technique (SMOTE) on the training set in order to balance the training set. More details on SMOTE are found in the next section. In addition, the logistic regression and neural network machine learning models require the continuous features to be normalized when their values are not measured on the same scale. The normalization is done separately for the training and testing sets using Z-score standardization. This is to ensure no information from the testing set is “leaked” into the training set that the machine learning models are going to learn from.

### 3.1.3 Data Sampling

After data cleaning and normalization, we have obtained an imbalanced dataset. In the training set, the amount of positive and negative samples is 1426 (20.37%) and 5574 (79.63%) respectively. In such a case, we can simply get a pretty high training accuracy by considering all our samples as negatives samples. Additionally, an imbalanced dataset would easily cause the machine learning algorithm to be more biased to the majority class during the training process, which consequently leads to the low sensitivity of the minority class.

Data under-sampling and over-sampling are therefore designed to mitigate the influence of an imbalanced dataset. In this experiment, over-sampling was performed. We applied SMOTE on the training dataset which creates synthetic positive samples according to the mid-way between two near neighbors. Since our training set contains mixed categorical and continuous value, we performed a special version of SMOTE which is also called SMOTE for Nominal and Continuous (SMOTENC) so that the most common category of nearest neighbors is assigned to the newly generated sample. The Imblearn package in python was used to perform SMOTENC on the training dataset. After calculation, we obtained a training set with the same amount of positive and negative samples (5574 for each). The new training set was then further used for the next step.

### 3.1.4 Dummy Variable

We perform two different types of dummy encoding. For the logistic regression and neural network models, we use “drop first” one-hot encoding to create dummy variables for the categorical variables. The base level of every categorical variable is absorbed into the intercept term for logistic regression and bias term of the input layer for the neural network. For the decision tree, random forest and XG boost models, we use the “typical” (no “drop first”) one-hot encoding to create dummy variables for the categorical variables. Unlike logistic regression and neural network, tree-based models do not have an intercept term or bias term for the inputs. Every level of every categorical variable has to be involved in the selection process that selects the maximum increase in purity at every split. Regardless of encoding, dummy variables belonging to the same category should always be considered together and not in isolation.

Our final dataset for logistic regression and neural network model construction contains 11 attributes with 6 continuous attributes and 5 dummy attributes with France being the base level of Geography, male being the base level of Gender, code 0 (no credit card) being the base level of HasCrCard and code 0 (non-active member) being the base level of IsActiveMember.

Our final dataset for decision tree, random forest and XG boost contains 15 attributes with 6 continuous attributes and 9 dummy attributes.

## 3.2 Feature Engineering

### 3.2.1 Logistic Regression

A logistic regression model is constructed by fitting all predictors available. The results 表格

描述已自动生成are shown below.

Figure 4 Result of Logistic Regression

The accuracy against the testing set of the logistic regression is not satisfactorily high. However, we can use the logistic regression model as a means to indicate the statistical significance of the association between a particular feature and the target variable. We observe that EstimatedSalary is a highly insignificant feature with an exceptionally high p-value of 0.315. All other features are statistically significant based on the p-values of the logistic regression model.

### 3.2.2 Decision Tree

图形用户界面, 应用程序

描述已自动生成A decision tree model is constructed by fitting all predictors available. The full tree is grown and then pruned by finding the optimal cost complexity parameter. The optimal cost complexity parameter is where the average cross-validation error using 10-fold cross-validation on the training set is minimum. The pruned tree is still quite large. An image of it can be found in tree.png. The results are shown below.

Figure 5 Result of Decision Tree

The accuracy against the testing set of the decision tree is better than logistic regression but still below 80%. We use the decision tree model as a means to identify the features that are “important” measured based on a relative importance scale normalized to a percentage. The importance measures every feature’s contribution to overall Gini gain. In other words, it measures the extent to which each feature contributes to purer splits of the target variable throughout the tree. It is observed that age contributed the most to overall Gini gain by a large margin as compared to all the other features.

### 3.2.3 Mutual Information

文本

描述已自动生成The Mutual Information (MI) of two random variables is a calculation of the mutual dependence between them. In other words, it measures how much the uncertainty of one variable is reduced by knowing another variable. In this experiment, we applied MI to measure the dependency between each variable and target. The image below shows the MI of each feature.

Figure 6 Mutual Information Result

MI is zero if two variables are totally independent. We therefore set the threshold as 0.01, meaning that attributes whose MI is less than 0.01 will be removed by the algorithm. We can note that age achieves the highest MI while EstimatedSalary has the lowest.

### 3.2.4 Feature Selection and Profiling

We find that across the regression, decision tree and mutual information models, EstimatedSalary appears to be a rather unimportant feature in distinguishing between whether a customer churns or remains a customer of the bank. We will drop this feature in our feature selection. On the other hand, Age appears to be a rather important feature in distinguishing between whether a customer churns or remains a customer of the bank. The older a customer is, the more likely he/she will churn.

## 3.3 Model Training

### 3.3.1 Random forest

A randomized grid search is used to find the best hyperparameter combination of random forest. Grid search is run 100 times randomly out of multiple groups. The one with the best area under the roc curve (auc) is selected as the final hyperparameter for the random forest. The final hyperparameter of Random Forest is shown below:

Table 1 Hyperparameters of Random Forest

### 3.3.2 XGBoost

The training of XGBoost used the same strategy as Random Forest, as explained in section 1.3.1. The final hyperparameter of XGBoost is displayed below:

Table 2 Hyperparameters of XGBoost

### 3.3.3 Artificial Neural Network (ANN)

The architecture of ANN is a three-layer ANN with one hidden layer. The first and second layers have dimensionalities of the output space of 12 and 6 respectively, with ReLU activation. A sigmoid layer is used in the final layer. The optimizer of the network is Nadam with binary cross-entropy as the loss function. Nadam is also called Nesterov-accelerated Adaptive Moment Estimation which combines Adam with Nesterov momentum.

## 3.4 Result Analysis

The testing result of the three models are shown below:

Figure 7 ANN Testing Result

图表

描述已自动生成

Figure 4: ANN testing result

图表

描述已自动生成

Figure 8 Random Forest Testing Result

图表

描述已自动生成We can note that random forest classifier achieves the highest accuracy and auc while ANN has the highest recall. Since our goal is to train the classification model that can recognize customer has churned, we need to focus on recall and auc. Therefore, we can conclude that ANN and random forest are better than XGBoost. Additionally, the random forest is better than ANN because of a higher auc, stable accuracy and precision.

Figure 9 XGBoost Testing Result

# 4.Discussion

## 4.1 Cause of churn

To build up the customer churn prediction model help White Rock better understand what are the key factors that make customers leave the company and better understand the requirement of customers. Analyzing the cause of churn, we also suggest the appropriate actions that have to be carried out to reduce customer churn.

### 4.1.1 Age

Age is a powerful discriminator on customer churn. Considering a customer enters an older age range, the odds of a churn rate increase by the factor . There might be several explanations for this phenomenon. Firstly, when people get older, they are more or less refuse to take a risk because they are the breadwinner of the entire family, so they switched to other kinds of conservative investments like traditional bank products. Secondly, older customers have more experience in buying asset products that they know how to make the best choice to gain profits, thus if they are not fully satisfied with the current financial outcomes, they might end their relations with the asset company. For a younger demographic, they are not so experienced or have the same wealth of choice than older people given they have a lower estimated salary. Thirdly, unlike older customers who may consider pensions, inheritance, taxes, which means they will emphasize on finding the best deal, it seems to be unnecessary for younger customers to switch to another company because they don’t have that various kinds of asset and the outcome of asset management are similar across the industry for them. Finally, this is attributable to fact that the older generation might value customer service and real-life interaction more than younger customers, even though today’s trend is as less real-life interaction as possible. White Rock is no exception heading towards with digitization, meaning that would cause a loss in the older customers if they switch to companies that better offer offline services and traditional values today. Based on the churn model, and consider the further digitization way that plans to go, White Rock might need to weigh its impact on losing older customers and consider preferential policies to engage with those people.

### 4.1.2 Is active member

From the logistics regression result, we can see that an active member is a critical variable to decide an exited customer. An active member means a customer is “active” to be involved in the company’s services, following news of the latest products, using channels whether it’s online, phone calls, or offline ‘VIP’ service. For non-active members, the odds of probability to leave White Rock would increase by . If we come to describe the portrait of active and inactive customers, we may find inactive customers has higher balance although lower credit scores, which means they might not likely to be active in asset activities although has the financial capability. This could be a potential breaking point for White Rock to reduce churn rate if pay more efforts to build a reliable relationship with non-active

customers by channel promotion and better service. Considering White Rock’s high proportion of active members, which is 51% of the total amount, our recommendation is to utilize special services or discounts to encourage active customers to recommend non-active customers to join the asset activity.

### 4.1.3 Germany

Another important discovery we get from the churn model is that customers from Germany are more likely to churn compared to customers from France and Spain. A customer in Germany would have a higher churn probability of . If looking into the characteristics of German, French and Spanish, the average age and active member amount do not have much difference among the group, whereas German has significantly higher balance and estimated salary than French and Spanish. When we did a further market survey, we find that Germany actually doesn’t have an investment culture and does not like debt. As the "European Savings Champion", Germans are notoriously bad at and have no interest in investing, even young people. According to a market survey in 2015, however, only about 7% of Germans save money to increase wealth. The main purposes of their deposits are " pension", "emergency" and "buy commodities." All in all, the profile of consumers in Germany has a clear differentiation with respect to the other two countries. Considering this problem, White Rock might hope to narrow the team members in Germany and switch more effort to retaining customers in Spain and France markets.

### 4.1.4 Gender

Gender is another indicator that weighs high to decide the churn likelihood. In the logistic Regression Model, a female customer has odds of churn rate higher than male customers. Regardless of age, active member status, card status, and even geography, women are 10% more likely to churn. Although women and men have a similar balance on average and estimated salary, an explanation may arise from the statistical fact, women are more conservative than men because it risks on asset investment activities. In this case, White Rock might focus more on male customers and arrange sound investment to cater to the needs of women customers.

From the model result, we found that credit score and estimated salary has a marginal effect to predict churn behaviors, thus influence to a lesser extent the probability of churn.

When comes back to churn, actually there are avoidable and unavoidable reasons for companies to lose customers, no exception for White Rock. Some customer churns are inevitable. The churn rate of 20% now could need White Rock to pay more effort on avoidable churn, that is what we will make the value. Our objective aims to help White Rock detect avoidable churns and find what needs to take action. Based on the understanding of our model, for example, White Rock could find out elders needs more support in purchasing the product; channel promotion should be added to active members; Germany does not have an investment culture; focus on male customers and then dive deep into an area to pay effort.

## 4.2 Costs of churn and corresponding impacts

Customer churn always leads to loss of sales, a decrease of market share, disruption of further investments, lower Return of Investment (ROI), and additional payouts to either acquire new customers or retain lost customers.

### 4.2.1 Costs of churn & impacts: Company-wise

#### 4.2.1.1 Loss of recurrent revenue

The most direct cost of customer churn should be the loss of recurrent business revenue. If the annual average asset management fee for one customer is $10,000, losing 10 customers will lose $100,000 in revenue per year.

#### 4.2.1.2 Loss of expansion opportunity revenue

In addition to annual recurring revenue, there’s also a potential loss for renewal income or upsell deals. Data from the website of ClientSuccess shows the probability of upselling to an existing customer is around 65%, while the probability of selling products to a new customer is only 13%. That is to say, once customers decide to terminate their subscription, White Rock will face a 65% chance of losing a new upselling or cross-selling of other products. So it is wise for the management of White Rock to keep in touch with their existing customer as frequently as possible to avoid customer churn and is encouraged to do upselling or cross-selling to existing customers.

#### 4.2.1.3 Disruption of ongoing investments

Once customers decide to terminate the subscription and turn to other asset management companies, all assets that White Rock has been managing for customers will be withdrawn as per the agreed date and time. But these assets might have been further invested by White Rock and the investment term is not yet reached, leading to the consequences that the expected revenues will not be obtained with loss of return on investment.

#### 4.2.1.4 Additional costs for retention activities

Generally, customer retention has been shown to be highly profitable to companies because attracting new clients costs five to six times more than retaining existing customers. However, additional costs will be burnt for launching customer retention campaigns, such as offering an incentive (a discount or another promotional offer) to encourage customers with the highest likelihood of churning to extend their subscription or to keep their account active.

If we go a step further to evaluate the individual “treatment effect” towards below 4 different types of customers, we will find there may exist other retention costs, based on the model an asset management company chooses. A treatment may concern any action taken towards a customer, such as a discount offered to retain a customer.

**Type 1**: Customers who would never churn.

**Type 2**: Customers to churn regardless of the campaign used.

**Type 3**: Customers who do not churn only because they have been exposed to a retention campaign.

**Type 4**: “Sleeping dogs”. Customers who would churn only because they were exposed to a retention campaign.

Ideally, only retention investment on Type 3 customers will render a high ROI and is regarded as a successful retention activity. Type 3 is our target group. Nevertheless, in the real business world, it is quite common to see Type 1, 2 and 4 customers, on whom launching campaigns brings different consequences to AM companies:

* A high recall, low precision model predicts customers who initially don’t intend to quit as “potentially lost customers” and guides the company to offer a retention package and spend unnecessary costs. These customers can belong to Type 1 or Type 4. Targeting Type 1 does not generate additional returns yet in fact leads to additional costs, while targeting Type 4 customers can have an adverse reaction, generating no additional revenues but leading to unexpected loss of churn.
* A low recall, high precision model fails to identify “potentially lost customers” and provides nothing for retention. These customers can belong to Type 2 or Type 3. Luckily missing Type 2 from the campaign is nothing severe, while missing Type 3 customers can cause a huge loss.

To compare actual loss or trade-off between the above 2 models, While Rock needs in-depth analysis with transactional sales data based on groups.

#### 4.2.1.5 Additional costs for acquisition activities

Customer acquisition campaign, apart from the retention plan, though sometimes required, demands a higher cost of serving new customers, developing a sense of brand loyalty over time. New referrals through positive word-of-mouth will only be generated after building up trust and reliable customer relationships in a longer time, while potential risk also exists that dissatisfied new customers might spread negative word-of-mouth, leading to loss of potential sales opportunities.

#### 4.2.1.6 Negative network effect and impact on brand leadership

According to the White House Office of Customer Affairs, if a customer is dissatisfied with a company’s product or services, he will tell around 9 to 15 people around him about his uncomfortable experience with the company. The negative influence is quite huge by multiplying this number by the number of customer churn and it will lead to a bad brand implication.

The brand reputation is always the most key consideration for a new customer to choose their asset management company. Because customers will only put their money into the organization if they really trust it. And some researchers even regard brand reputation as the most valuable asset for a company. So the cost of customer churn in the field of company brand can be far more severe than some physical cost directly related to revenue loss.

#### 4.2.1.7 Change in investment strategies & product design choices

Liquidity, i.e. flexibility provided to customers to cash out their assets from the existing subscription at any time per their request, is one of the decisive factors that customers value when choosing asset managers or products.

On the one hand, commit such high liquidity sets traps for asset management companies themselves as it offers enough adjustable space for customers to switch from one company to another, which causes future sales loss of the company. On the other hand, to fulfill the commitment and further attract more customers, asset managers have to compromise their investment strategies to alter the product design choices to short-duration investments either high-risk products that may yield higher revenues in short term such as stocks, or low revenues products usually featured with lower risks.

### 4.2.2 Costs of churn & impacts: Industry-wise

#### 4.2.2.1 Rising competition in the Industry

Since the financial crisis, there has been an increase in competition in the industry, more firms and more advisors are competing for the same clients and the same assets. Having access to basically the same products, tools, and models, asset managers find it difficult to prove that their services are differentiated and their advice outperforms others on yielding higher ROI under a more transparent environment in the asset management industry nowadays.

To compete against customers’ existing offers from competitors, part of the profits is considered to be surrendered to win more deals.

#### 4.2.2.2 Increasing demand for big data and analytics

Following the above analysis of costs of churn from different aspects, we understand that customer churn does impose a great impact on future revenues and corporate management strategy, hence **the true business objective here is to reduce customer churn.**

As volumes of consumer data continue to grow exponentially, most AM companies have been using big data and simple analytics to deliver key business insights and to speed up the process of due diligence - hours of human work are thus reduced to seconds, freeing asset managers to focus on downstream tasks.

More advanced analytics, such as customer churn prediction models and descriptive and predictive analytics that combine internal and external, structured and unstructured data, is designed to create more complete and insightful client profiles, facilitate an accurate segmentation of the customer base, target the customers who are most likely to churn, identify factors that lead to customer churn and advise measures to reduce churn in the most cost-effective way.

## 4.3 Added values bring to the White Rock

### 4.3.1 Identify important factors influencing customer churn in a data-driven precise

From this dataset, we test the features of age, gender, demographic features, personal account information, etc. The result and related explanations are covered in part 3 and part 4.1. From these results, White Rock can target specific groups of customers and use relevant knowledge to develop tailored asset management products for these groups.

### 4.3.2 Predict customer churn and take measures in advance

Our customer churn model can be applied to make classification according to the likelihood of churn in future times and companies can then target customers with higher churn risk. Specific promotional campaigns can also be applied to these groups of customers.

### 4.3.3 Real-time analysis about operational performance in customer relationship management

During operation, these models will be trained with live data about customer information and refresh regularly to keep in touch with the latest customer churn occasion and predict existing customers promptly to make sure White Rock can react to potential customer churn in advance.

## 4.4 Measures to reduce bad influence and cost of customer churn

### 4.4.1 Raise customer retention rate

To increase customer retention rate is the root measure to cut cost for customer churn. First of all, asset management companies can create barriers that discourage customers to change their choices. Furthermore, they can develop loyalty programs and encourage renewal activities. According to the analysis result of customer churn data, if a customer is an active member of the asset management company, the less likelihood the customer will cease the current contract and turn to another asset management company. If White Rock can create and develop more loyalty programs as well as upselling and cross-selling activity for customers, the relationship between customers and the company can be much tighter. Finally, using machine learning models to identify customers with a high likelihood to cease subscriptions can be a provisional measure. The collection and analysis of customer data can support the creation of a churn risk analysis that helps identify customers who should be focused on.

### 4.4.2 Decrease cost of customer churn already happened

Based on the relevant cost of customer churn mentioned in 4.2.1, once customers decide to cease their contract, the company should take measures and campaigns to either retain back lost customers or attract new customers to minimize the loss of recurrent revenue and disruption of ongoing investment.

For the influence on brand and reputation, the company should immediately ask for feedback from leaving customers as regards the reason why they choose to leave. If they leave because of poor services or dissatisfied performance of the company, relevant staff should deal with the complaints as soon as possible to avoid the spread of bad impressions to more people.

# 5. Conclusion

Our study set a background in the asset management industry and use data provided by White Rock to investigate customer churn. The whole process includes background understanding, data exploration and visualization, modeling, result analysis and business insight through our work.

Based on our business understanding of the problem White Rock faces, that is to find out the factors which are significantly associated with customer churn and whether predict a customer will exit or not, we utilize the public dataset of the bank industry which is similar to asset management for machine learning. We get an overview of the potential factors and useful information with the exploration and understanding of data and then do the data cleaning in preparation for the feature selection.

In model training, we choose random forest, XGBoost and artificial neural network to predict the likelihood of churn in for customers. The result shows that random forest classifier achieves the highest accuracy and auc while ANN has the highest recall. Since our goal is to train the classification model that can recognize customer has churned, our main focuses are recall and auc. Therefore, we can jump to the conclusion that ANN and random forest are better than XGBoost. Additionally, the random forest is better than ANN because of a higher auc, stable accuracy and precision.

For the business insight part, we analyze the key factors that make customers leave the company and better understand the requirement of customers. Then we list the cost of customer churn from both company-wise and industry-wise to make asset management companies more aware of the potential harm from customer churn. At last, we delivered the benefits of our model and propose actions to reduce the cost for customer churn.

In future research, we can put the real data of customer churn from asset management companies and include more customer information to our model to get a better insight into customer churn.

# 6. Reference

[1]Karvana, K.G., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. 2019 International Workshop on Big Data and Information Security (IWBIS), 33-38.

[2] Li, Y., & Wang, B. (2018). A Study on Customer Churn of Commercial Banks Based on Learning from Label Proportions. 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 1241-1247.

[3]Farooqi, R., & Iqbal, N. (2017). Effectiveness of Data mining in Banking Industry: An empirical study. International Journal of Advanced Research in Computer Science, 8, 827-830.

[4]Kumar, D.A., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. Int. J. Data Anal. Tech. Strateg., 1, 4-28.

[5]Kaur, M., Singh, K., & Sharma, N. (2013). Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers.

[6]Yu, F., Liu, D., Kumar, S., Jebara, T., & Chang, S. (2013). \(\propto\)SVM for Learning with Label Proportions. ArXiv, abs/1306.0886.

[7] loris Devriendt , Jeroen Berrevoets, Wouter Verbekeloris Devriendt , Jeroen Berrevoets, Wouter Verbeke (2019). Why you should stop predicting customer churn and start using uplift models. Information Sciences, Volume 525, July 2020, Pages 54-66.