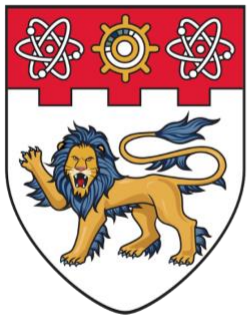


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC2406 - Analytics I: Visual & Predictive Techniques

Improving WhiteRock's Client Onboarding Through Analytics

Seminar Group: 7
Team: 4
Instructor: Sanli Pinar Darendeli

Name	Matriculation number
Brandon Chen Yun Xin	U1920186J
Janzy Chua Pei Lin	U1910545K
Jerry Lee Rui	U1910048J
Joelle Lee Shi Yau	U1810858J
Shwe Sin Oo	U1911031L

Contents

Executive Summary

This report explains the relationship between White rock and its intermediaries (banks) and provides an analysis regarding the customer onboarding process for the intermediaries, as well as a proposed model to speed up the onboarding process for the banks.

The current customer onboarding process that White Rock and banks adopt is inefficient due to limited resources and extensive manual work to analyse each customer's eligibility to onboard. With limited manpower to interact with large groups of customers, it is no doubt that customers can get frustrated at the long processing time. Hence, customers may choose to onboard with competitors and cause the banks to lose many potential customers. This translates to the loss of opportunity for White Rock to sell their products and generate revenue. Similarly, White Rock's asset management products are recommended less by the banks to less customers.

With limited resources, the banks should choose to channel more resources to onboard potential customers that are likely to bring in the most value to the banks. Hence, what the team proposes is to first segment the customers into 'high value' or 'low value' customers when they sign up for the onboarding process.

When customers register for the onboarding process, only general information of the customers are given. But only after onboarding will the banks know exactly how much value these customers bring in. Thus, the team will make use of a real dataset regarding details of bank customers that are already on board with a bank. The dataset contains both general information of the customers and their activities with the bank (Example, the number of bank products the customer purchased or whether they have a credit card with a bank etc.). The information on the customers' bank activities cannot be determined during the onboarding process but are indicators to justify the customers' values. Thus, in the dataset, the team segregated the customers based on their value using their bank activities, and used the general information to predict the customer value. Our model utilises logistic regression analysis to identify the general information that suggest a statistically significant impact on customer's values. Afterwards, the CART model is used to accurately classify customer's value based on each statistically significant variable.

Based on the model accuracy, CART model is a slightly better fit than the logistic regression model for predicting customer's values. However, there are many other factors that affect the accuracy as well as suitability of both models.

If WhiteRock uses the model to predict the value a customer may potentially bring before they are officially onboard, White Rock can easily identify the customers that they should secure quickly and thus allocate more resources to escalate the onboarding process. After onboarding them more quickly, it is easier for White Rock to make an effort and sell their investment products to these customers. This also gives White Rock a greater competitive advantage since almost every financial organization out there is facing the same inefficient onboarding process.

1 Introduction

White Rock is a 300-year-old company formed in London and now in 32 markets globally, including 8 in Asia. Primarily an asset manager, White Rock distributes investment products to not only institutional clients, but also intermediaries that are mainly banks all around the World. The intermediaries will promote and upsell White Rock's asset funds, contributing to a substantial proportion of White Rock's revenue.

1.1 Significance of the onboarding process for White Rock and Intermediary banks

White Rock has to onboard the new clients before promoting their investment products to them. Also, while some banks have their own asset management divisions that sell in-house asset management products, many intermediaries banks such as Morgan Stanley, Goldman Sachs and Deutsche adopt what is called the "open architecture model" (Marriage, 2015) where the majority of the asset management products they recommend to customers come from external asset management firms such as White Rock. Hence, intermediary banks can be effective channels for White Rock when these banks onboard new customers and recommend White Rock's asset management services or products to the customers.

As a result, the client onboarding process of both White Rock and these banks are of significant importance for White Rock to get new customers who will buy their investment products. Also, a generally pleasant onboarding experience creates a first good impression to customers, which may encourage customers to recommend others to come onboard as well. Therefore, WhiteRock and their intermediary banks will get more customers in the near future.

1.2 Inefficiencies in the current Onboarding Process

Currently, the onboarding process is done manually, suggesting that employees have to settle piles of paperwork, overnight mailings, and make redundant data requests just to analyse and review each customer's eligibility to onboard. To make matters worse, most financial institutions have "contacted the customers on average 10 times during the onboarding process and asked to submit between 5 and up to 100 documents", according to a study conducted by Forrester Consulting (Fenergo, 2015). The duration of the onboarding process is incredibly long as well and could take anywhere from 2 to 34 weeks.

This is further aggravated by the fact that many financial institutions still make use of legacy systems (Finextra, 2020) that have been running for over 30 years with approximately £2 trillion passing through them every day. Hence, their refusal to switch to modern systems could be because there is already so much money relying on these existing systems and changing them could be risky. As a result, many banks remain risk-averse and still prefer to use the good old "if it isn't broken, don't fix it' approach.

As a result, even for institutions that lean on technology, it is not sufficient to catch up with the incredibly fast paced nature of customer service today. With a large pool of customers waiting in

line and limited resources, the onboarding process has become highly inefficient and too time consuming.

1.3 Effects of inefficient customer onboarding process

The onboarding experience has a quantifiable impact on key variables such as customer loyalty, profitability and referrals by existing customers to new customers, reputation and brand equity. According to a study conducted by Forrester Consulting, “9 of the 13 interviewed organizations agreed or strongly agreed that the client's onboarding experience has a strong impact on the lifetime value of the client, **with 10 claiming that they have lost deals due to inefficient onboarding**.” (Fenergo, 2015). This shows how highly inefficient the onboarding process is for not just White Rock but other financial organizations or corporations in the world. This also proves that potential customers can easily come and go and that there is little effort that White Rock can do to retain customers after onboarding them so inefficiently.

Additionally, a customer would be **four times less likely to defect to a competitor** if the service-related problem is handled effectively (KPMG, 2016). This once again emphasizes that the customer onboarding process is the one-time opportunity to make a good first impression, which can lead to customer retention and active long-term customers.

Thus, if White Rock does not improve their highly inefficient customer onboarding process, White Rock and their intermediary banks will lose many potential customers to competitors during the onboarding process if the client develops an unfavorable impression towards the institution. If not, customers are more likely to be inactive or “one-time” customers of White Rock, resulting in higher churn rates as well. Word-of-mouth between customers also creates an undesirable domino effect. According to search by Medallia, 35% of bank customers share their negative experiences with a bank with family, friends and acquaintances (Medallia, 2018). This might lead to a loss of potential customers for White Rock and intermediaries.

1.4 Improving the Client Onboarding Processes

1.4.1 Customer Segmentation

Many financial organizations place a lot of emphasis on analyzing customers during the onboarding process because it is a great opportunity for them to evaluate many key factors such as potential customer loyalty and profit directly (Gundaniya, 2020). With the general information received regarding the customer during the initial onboarding process, one can filter and know which potential customers are financially capable to remain active and loyal with the bank. These are the customers that White Rock or other organisations do not wish to lose to competitors.

With limited resources and manpower to evaluate customers in the onboarding process, White Rock can take the first step to introduce **customer segmentation** in their onboarding process whereby instead of spending limited resources such as budget and manpower into targeting a broad scope of diverse customers, the banks could classify customers into different groups of

importance based on their economic value (Nguyen, 2020). After classification, White Rock will then allocate more resources to the high value customers. Thus, through customer segmentation, WhiteRock can better allocate resource to engage current and potential high value customers

1.4.2 Use of analytics in customer segmentation

A **customer value model** could help identify the most relevant factors that determine the value and priority level of a customer. These factors are unique to each organization and it is essential for White Rock and its intermediaries to carefully select them. However, both White Rock and the intermediary banks do not have the expertise nor the analytical tools to be able to quickly identify and segment customers accordingly. Thus, using analytical tools to predict the value of customers would significantly increase White Rock's potential customers and revenue.

Many papers have written both the conceptual and practical challenges associated with customer valuation (e.g. Berger and Nasr, 1998; Jain and Singh, 2002; Rust et al., 2004). However, only a few of them (Haenlein et al., 2007; Ekinci et al,, 2014) are customized to retail banking, hence, our approach will help White Rock and their intermediaries to predict the value of potential customers to allocate their resources more efficiently in client onboarding.

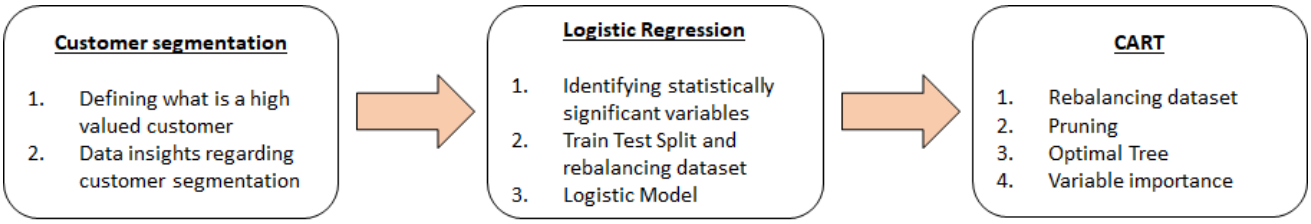
Through the model proposed at the later part of the report, we aim to use analytical tools to help White Rock identify which are the significant factors that contribute to customer value before predicting the value of each customer so that White Rock and their partnering banks can allocate the level of priority to customers based on their predicted value.

2 Approach

2.1 Overview of approach

Firstly, we will do customer segmentation and reclassify the customers into high and low value customers based on certain variables from the dataset. The remaining variables are then analysed with the help of domain knowledge and expert opinions to find out if they have any statistical significance with customer value.

Next, we would then build a model to predict the value of a customer based on information provided in the dataset using both the Logistic Regression and CART model. The diagram below gives the overview of the whole approach.



2.2 Dataset

Our dataset set belongs to one of the many intermediary banks of White Rock.

We will be extracting insights from this dataset as well as develop a predictive model to help determine the value of each customer. The dataset contains details of customers largely from 3 specific countries (i.e France, Germany and Spain). The dataset gives a brief overview of the customer's engagement with the bank by highlighting factors such as their active status and number of products purchased from the bank. The model that will be developed by using this dataset can then be put to use by White Rock with their own data set.

2.3 Customer Segmentation

In this model, we will be broadly segmenting our customers into 2 category levels accordingly. The value of the customer will be determined based on 3 variables respectively, number of products, existence of credit card and member status. These variables largely have an influence on the revenue of the bank over time.

Value	
High	Number of product > 1, Has Credit Card, Active Member
Low	All other combinations that do not meet criteria: Number of products = 1, No credit Card, Not Active Member Number of products = 2, No credit Card, Not Active Member Number of products = 3, No credit Card, Not Active Member Number of products = 4, No credit Card, Not Active Member Number of products = 1, Has credit Card, Not Active Member Number of products = 2, Has credit Card, Not Active Member Number of products = 3, Has credit Card, Not Active Member Number of products = 4, Has credit Card, Not Active Member Number of products = 1, No credit Card, Active Member Number of products = 2, No credit Card, Active Member Number of products = 3, No credit Card, Active Member Number of products = 4, No credit Card, Active Member Number of products = 1, Has credit Card, Active Member

1. Number of products

How many products and services customers purchase from a bank is a common and crucial (CSP, 2017) customer segmentation criteria as it analyses a customer's behavioural patterns as their relationship with the bank progresses. According to Pointillist, the company behind a widely used customer analytics software, it is important to analyse and be aware of how often and how much of a company's products customers are using as "usage behavior (Deasi, 2020) can be a strong predictive indicator of loyalty or churn and, therefore, customer value".

Also, when customers onboard with the bank, they would automatically have 1 baseline product which is usually in the form of a savings account. When customers have more than one product, it would suggest that the customer had purchased additional products from the bank. Therefore, such customers should be classified as high value since there is a higher likelihood that they will purchase the bank products.

2. Credit Card

Credit cards bear certain risks and are also used for investment by consumers. A credit card provides an open-ended revolving line of credit where a customer is allowed to borrow within a limit and pay back later. Failing to pay by the due date will lead to the borrower being charged with heavy interest. "The interests they (the banks) charge on unpaid balances are lucrative, approximately 24% per annum" (Dayani, 2020). According to the Washington Post, JPMorgan Chase has been reported to have earned a "record $36 billion profit" in 2018 wiith "credit card loans increasing 8 percent". Apart from interest, a bank may also earn its credit card revenue from other channels such as late fees, annual membership fees, merchant fees, and others. With many streams of revenue being generated from credit cards alone, a customer holding a credit card issued by the bank will no doubt be considered to be of higher value.

3. Active Member

Inactive customers carry a higher risk of being unprofitable, because they no longer generate any revenue, while the client relationship may still lead to costs such as regular mailing of account statements (Haenlein et al., 2007). A client may no longer be active, but still own an account. One possible reason for this could be that in the absence of regular account maintenance charges, where this ownership does not involve any costs. Hence, there is only limited motivation for the customer to formally end the business relationship with his/her bank. It can therefore be assumed that customer value and profitability are heavily influenced by the activity level of the customer.

High Value Customers

In a nutshell, customers of 'high' value must fulfill all 3 criterias mentioned above. Even if customers met some of the criterias but not all, they are not considered as a high value customer in this dataset. This is largely due to the limited resources to attend to the large pool of customers. Hence, the most ideal way is to specifically pick out the customers who can bring the most benefits to the banks through the 3 criterias so that the resources allocated to these high value customers are justified.

For example, a customer might have more than one product and credit card but remain inactive to the bank. This can indicate that the customer has no interest to invest with the bank and the bank should then channel more resources to others who prove that they should get more resources from the bank.

2.4 Independent Variables

Our model will be using these 4 different independent variables based on published insights that might have an influence on customer value in banking.

1. Age

According to Campbell and Frei (2004), age can be assumed to influence profitability by its impact on consumption patterns in retail banking. It is important to take into consideration that financial needs and investment purposes change as a customer becomes more mature. Hence, banks must take into account such changes and adapt accordingly so that trust is formed. According to a study called Segmentation of Bank Customers by Age (Stanley, Fords, Richards, 1985), surveying users of 31 retail banks in Midwestern US states indicates that promoting specific products to specific groups proved greater chance for a successful sale. In particular, middle-aged customers tend to be more profitable than younger ones because they tend to maintain higher balances and are more likely to have mortgages. Hence, knowing the age distribution among its clients will help banks determine the type of products best suited for them. Baumann, Elliott and Burton (2012) report that the older the customer, the greater the intention to remain with their bank for a longer time. Loyal customers are more interested in the products and services of their own banks such as when considering investments in the financial market.

2. Gender

De Matos, Henrique, and De Rosa (2013) suggest that women are more likely to remain loyal to their bank when compared to men, which can be attributed to the fact that men are more willing to take risks than are women, and socially, men are expected to behave in this way, in agreement with the social role theory. This can be supported by Deloitte's findings (Deloitte, 2019) that millennial men are significantly more likely to switch banks (22%) than millennial women (13%), leading to a significant difference in loyalty. This serves as an important factor for banks as loyal customers not only increase the value of the business, but they also enable it to maintain costs lower than those associated with attracting new customers (Barroso Castro and Martín Armario, 1999).

3. Credit Rating

Credit rating can influence the amount of products the customer can buy from the bank. It indicates how good or bad of a risk one might be to the banks as a customer. The higher the number, the better the credit score. It is important to gain insight into the creditworthiness of individuals and assess one's financial soundness. Credit scoring can help financial institutions grow their portfolios by pre-qualifying customers for new products or cross sell products after

assessing their risk based on credit rating. This is essential as cross-selling has become a strategic priority for many banks in recent years. The incremental cost of selling to current customers is generally much lower than to new customers (Groenfeldt, 2012).

4. Estimated Salary

Income will have an influence on consumer purchasing behaviour and financial decisions. Hilgert, Hogarth, and Beverly (2003) report respondents with lower incomes were less likely to pay their bills on time than those with higher incomes. In addition, Aizcorbe, Kennickell, and Moore (2003) found that families with lower incomes are less likely to save. Consumers with high income will be more likely to own products and services provided by the bank and financial institutions.

5. Geography

Segmenting customers by demographics is a common practice among banks and geography is "particularly useful for identifying potential threats or opportunities as it is closely linked to other forms of demographic segmentation such as age or socio-economic status. (i.e. by 2043, the US will have a majority-minority population)". This helps banks make important decisions on whether to target the customers in that particular area or take their operations and services elsewhere.

3 Data Cleaning

Before any analysis takes place, there is a need to check if there are any missing data. Since, every column and row are filled in, the dataset has no missing values.

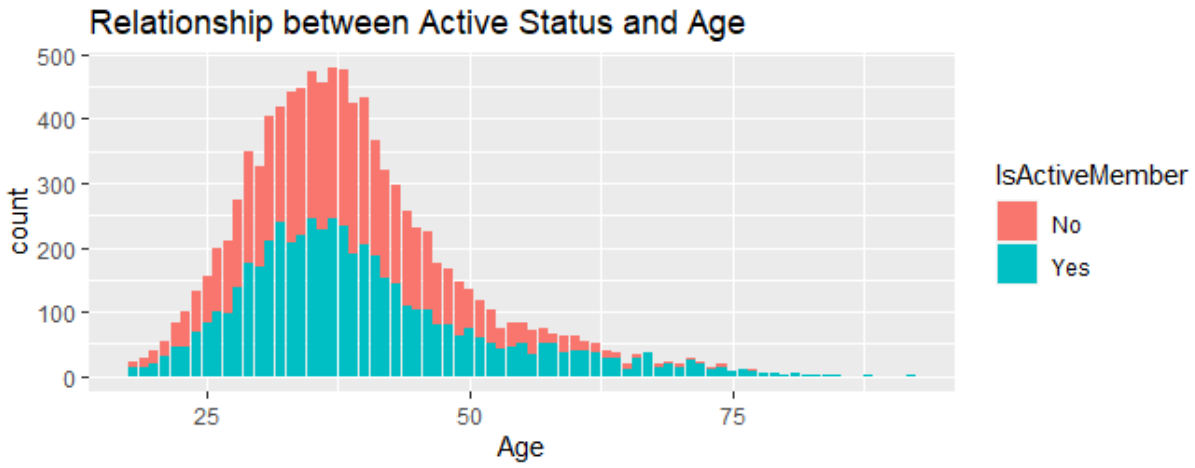
Next, a new column - CustomerValue, is added to classify the customers to 2 categories - "High and Low". As previously mentioned, a high valued customer is defined as a customer who has more than 1 product with the bank, has credit card with the bank, and is considered an active member with the bank. Hence, under the CustomerValue column, 'High' means that the customer has fulfilled the 3 criterias to be classified as a high valued customer with the bank while "Low" means that the customer is a regular customer.

Lastly, right before using the dataset in the model, the individual columns that are categorical are read as factors and the continuous variables are read as numerical. The table shows the list of categorical and continuous variables in the dataset.

Categorical Variable (factor)	Continuous variable (numerical)
Geography, Gender, NumOfProducts, HasCrCard, IsActiveMember, CustomerValue, Tenure	Age, Estimated Salary, CreditScore, Balance

4 Data Exploration and Insights

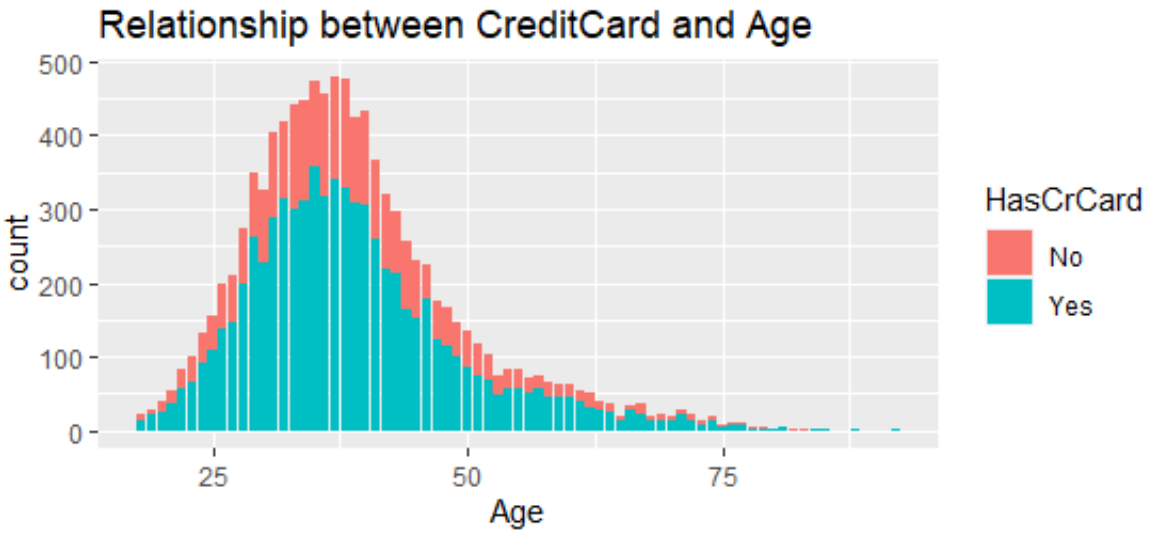
4.1 General Insights



The red area of the graph shows the number of customers who are not an active member of the bank across age, while the blue part of the graph shows the number of customers who are an active member of the bank.

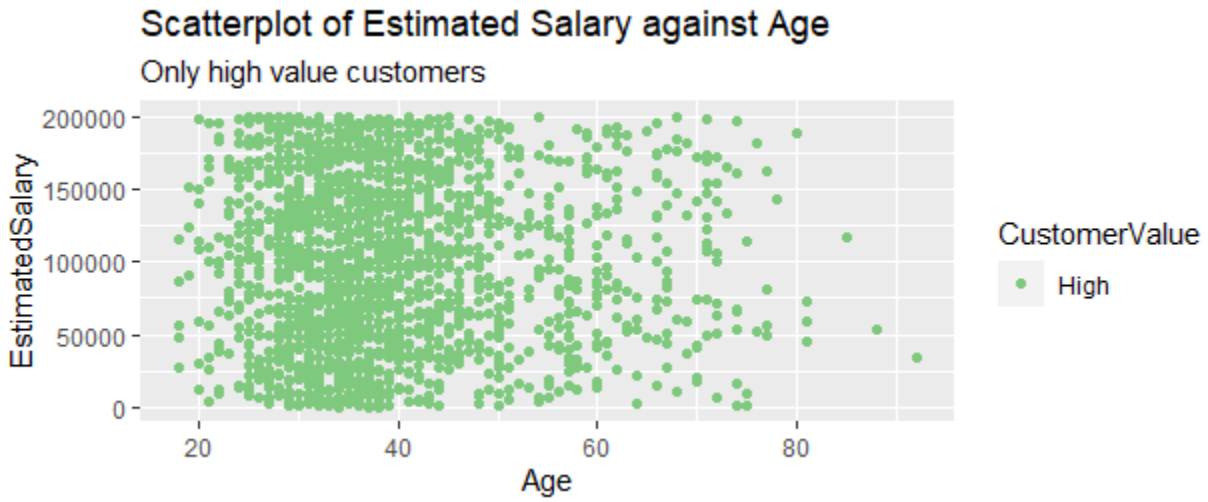
Most of the active members of the banks come from the age group of 30 to 40 years old. However, there are approximately as many non-active members as the number of active members. This suggests that banks could revise their strategies that are more tailored to this age group so that they can get more active members and sell more investment products to the existing clients. Also, White Rock can also take note that since most of the customers are from this age group, the products that get intermediary banks to sell should be more customised and suited for 30-40 years old.

Also, those who are older tend to use their accounts much less actively than those who are younger. One possible explanation of this trend is that people who are older might not receive wages into their accounts. Hence, we can expect less of the high value customers to come from the older age group.



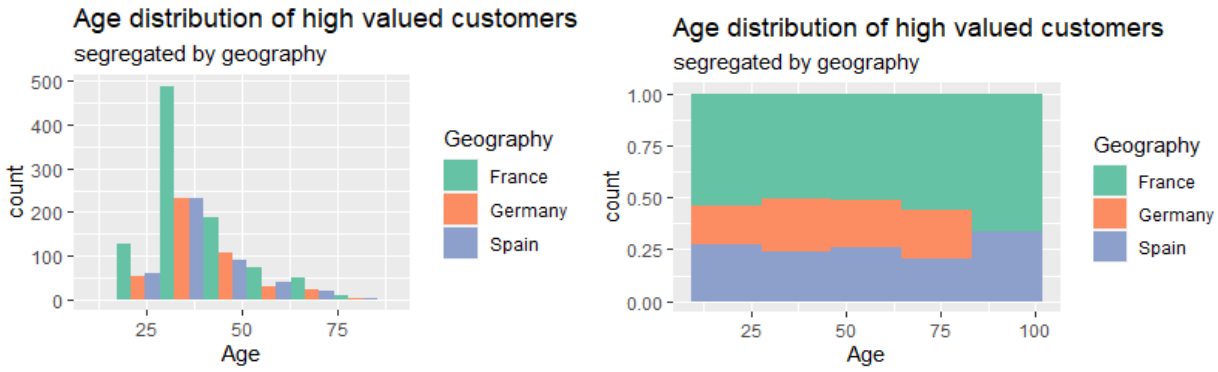
The red area graph represents the number of customers without a credit card while the blue part represents customers who have a credit card with the bank. The age group where most people own a credit card is between the age group of 30 to 40 years old. Not many of the younger generation (below 30 years old) seem to own a credit card according to our dataset. This is aligned with the survey conducted by Deloitte where younger consumers are less likely to take on less credit card debt compared to their predecessors. A possible reason is because the younger consumers have less financial capability to settle the debt a credit card may bring. Thus, we can expect that the high values customers in our dataset will less likely come from customers below 30 years old.

4.2 Characteristics of High valued Customers

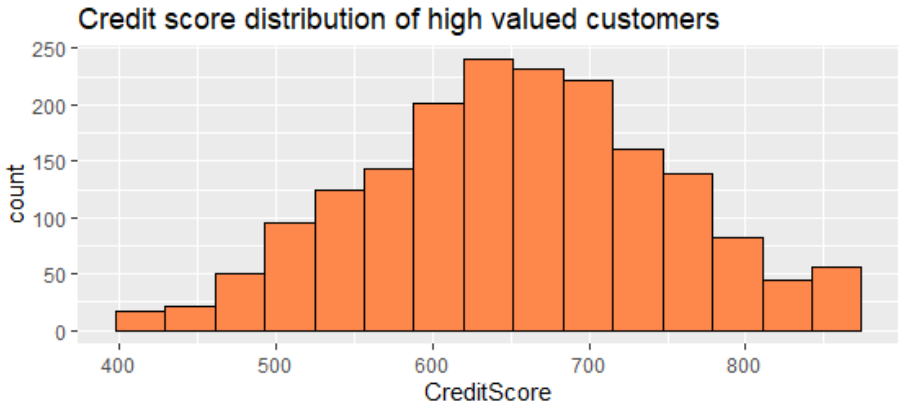


From the first 2 graphs explained above, we will expect most of the high valued customers to be between the age 30 and 40 years. From the scatter plot graph here, it shows the estimated salary against age for the high value customers only. As expected, most of the points are clustered between 30 to 40 years old since most of the high value customers are from that age group.

From the graph, we can tell that there is a very huge range of estimated salary at any point of age. From what we know, estimated salary can influence purchase habits and a generally higher salary will lead to customers choosing to invest more and supposedly bring more value to financial organisations.



From the graphs shown, most of the high valued customers across the 3 countries come from the age range 30 to 50 years old. At that age range, almost half comes from France and the remaining half is split between Germany and Spain. This may suggest that the country that customers reside in may play a part in determining whether they are more likely to be a high value customer. For the case of the dataset we use, a customer from France is more likely to be a high valued customer.



An acceptable credit score is generally above 600. Among the high valued customers, the majority have a credit score above 600. This suggests that most of the high valued customers are qualified for most of the bank products as their high credit rating brings assurance to banks that there will be less risk of default. With such assurance, banks could then cross-sell products of White Rock to high valued customers. However, a substantial number of high valued customers have a low credit score which is unusual. A low credit score suggests that customers are unable to pay up on time and have a higher risk of default. Because credit score is not one of the criteria that the group has established as high value, this unusual insight is discovered.

5 Logistic Regression

5.1 Purpose

An initial logistic regression analysis was performed to identify and validate the factors that might have a statistically significant impact on a customer's value, as well as understand the relationship between them. The independent variables chosen for the model were 'Age', 'EstimatedSalary', 'Gender', 'Geography' and 'CreditScore'. Based on domain knowledge, these factors have an influence on a customer's value.

5.2 Statistically Significant Variables

As seen from Figure 6.2.1, the logistic regression model obtained a *p-value* of 0.0246 for 'Age'. The *p-value* is an indicator of statistical significance of a variable and the smaller the *p-value*, the greater the significance. Since the *p-value* of 'Age' is lower than the benchmark of 0.05, 'Age' is concluded to have a statistically significant impact on a customer's value. Meanwhile, the other variables all have *p-values* above the benchmark which cannot conclude statistical significance of the variables.

5.3 Model Prediction

After identifying ‘Age’ as a statistically significant variable, another logistic regression model is built to predict a customer’s value. This logistic regression model would only fit in ‘Age’ as the independent variable.

5.3.1 Train Test Split

The initial dataset is first divided into a train set and a test set. The purpose of the train set is to train the model while the purpose of the test set is to validate the model’s accuracy on an unseen dataset so that predictions will not be biased towards the train set. The ratio chosen for the train set to test set is 70:30 which is in line with the general standards set across the field of machine learning.

5.3.2 Rebalancing the Dataset

As seen from the dataset, the proportion of customers is heavily skewed towards the ‘low’’ value. Based on the train set alone, the number of ‘high’ value customers (5718) is greater than the number of ‘low’ value customers (1282) by slightly over 400%. With too few samples from the minority category, the model might produce less than reliable estimates. Thus, the train set is rebalanced by undersampling the majority category to obtain an equal number of both ‘high’ and ‘low’ customers for training the model.

5.3.3 Predicted Results

After balancing the train set, the logistic regression model is trained with ‘Age’ as the only independent variable. From the summary as shown in Figure 6.3.3, we can identify the logistic function and the coefficient of ‘Age’ as seen below. The *p-value* is also in line with our previous findings.

$$Z = 0.4772 - 0.0122*(Age)$$

The coefficient of ‘Age’ implies that if a customer’s age increases by an additional year, the odds of them being ‘high’ value will decrease by a factor of e^0.0122 = 1.012. Based on ‘Age’ alone, the model then uses the test set to predict the probabilities of ‘high’ value which are then compared to the threshold probability of 0.5. If the predicted probability exceeds 0.5, the customer’s value is predicted as ‘high’. The predicted customer values are then compared to the actual customer values in a confusion matrix as seen below.

Cust Value	**'High' (Predicted)**	**'Low' (Predicted)**
'High' (Actual)	324	226
'Low' (Actual)	1330	1120

As seen from the confusion matrix, the sensitivity (true positive rate) can be calculated to be 324/(324+226) = 0.589 while the specificity (true negative rate) is 1120/(1120+1330) = 0.457. This means that when the actual value was 'high', the model predicted 58% correctly whereas when the actual value was 'low', 45% was predicted correctly.

5.4 Evaluation & Analysis

The overall accuracy obtained for the model is 0.495 which translates to **49.5%**. While this might suggest a slight association between 'Age' and 'CustomerValue', in general, the relationship between these two factors cannot be distinctly explained by the logistic regression model. It does not make logical sense that the odds will always decrease even if a customer's age surpasses the human life expectancy as well. Thus, for this particular dataset, a logistic regression model is inconclusive.

Nevertheless, White Rock can still take this model into consideration for future datasets that they might obtain from banks across the World. Each country has unique characteristics which might influence the accuracy of the model.

6 Classification and Regression Tree (CART)

6.1 Purpose

With the inconclusiveness of the logistic regression model, a CART model was then used to obtain more accurate predictions of a customer's value. A CART model segregates a customer's value based on all input variables and identifies the variable that creates the best homogeneous sets of values. The independent variables input into the model were 'Age', 'EstimatedSalary', 'Gender', 'CreditScore' and 'Geography'.

6.2 Model Prediction

6.2.1 Rebalancing the Dataset

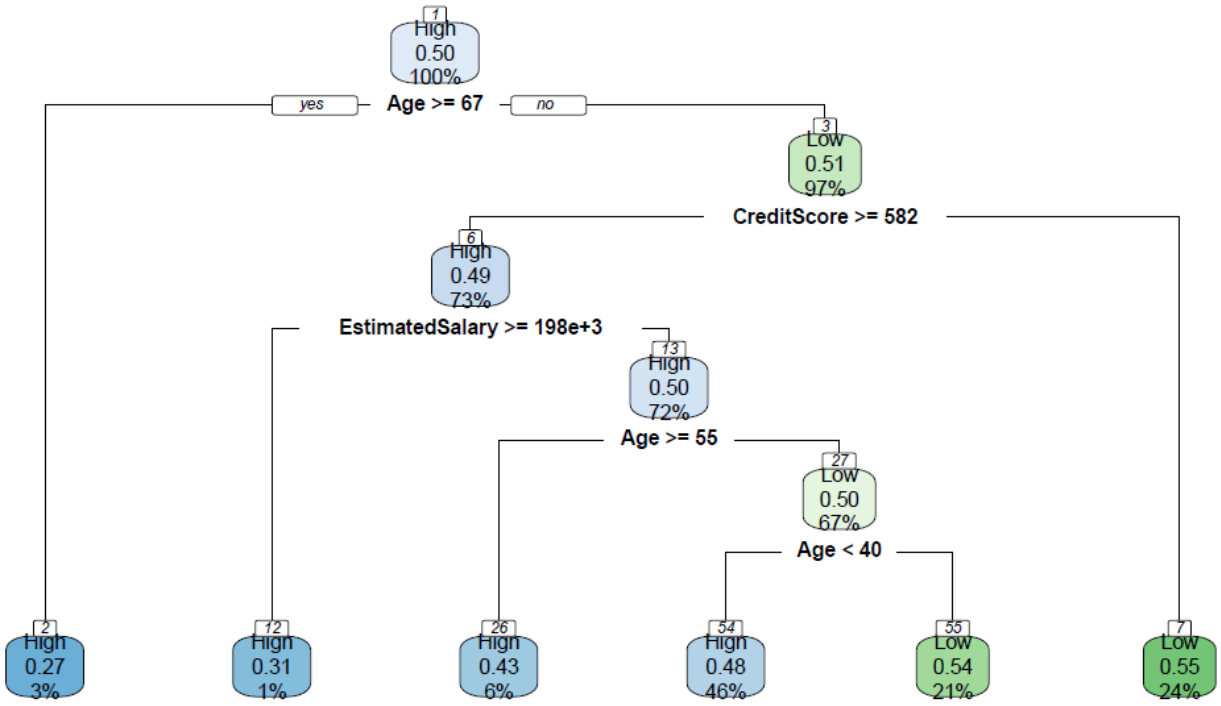
As mentioned in section 6.3.2, the proportion of customers is heavily skewed towards the 'low'' value. While a CART model generally handles skew better than logistic regression, the dataset still requires rebalancing since the main objective is to accurately predict the minority group ('high' value). Undersampling of the majority group ('low' value) is thus performed to obtain a more

'balanced' dataset such that more weightage is allocated to the minority group. The model will then make more correct predictions of the minority group instead of the majority group.

6.2.2 Pruning

The CART model is first allowed to grow till its maximum depth with the objective of reducing the misclassification error at each node. However, an increase in size also increases the likelihood of overfitting. Thus, pruning is performed to reduce the size of decision trees by removing sections of the tree. This reduces the complexity of the final classifier, and hence improves predictive accuracy by reducing likelihood of overfitting. The tree is pruned to the optimal point identified where the cross-validation error is within one standard error of the lowest cross validation error and the size of the tree is minimized. The optimal complexity parameter is identified to be 0.008456.

6.2.3 Optimal Tree



The optimal tree after pruning is shown as above. A condition for each split is presented under parent nodes. Observations that meet the condition (i.e., when the condition yields an answer of "yes") go to the child node on the left, otherwise go to the child node on the right.

For the first split, the decision rule is whether age is greater or equal to 67. 3% of the customers who satisfied the decision rule are placed in the left child node 2 and they have a 27% chance of being a high valued customer. 97% of the customers who do not satisfy the rule are placed in the right child node 3 and they have a 51% chance of being a low valued customer.

We are able to conclude that a customer's value is likely to be high when his age is greater or equal to 65. This agrees with our intuition that when a person is older, they naturally have more savings in their bank account.

The 2nd split occurs at the 3rd node where the decision rule is whether a customer's credit score is greater or equal to 582. 73% of the customers in node 3 who satisfied the rule are placed in the left child node 6 and they have 49% chance of being a high valued customer. 24% of the customers in node 3 who do not satisfy the rule are placed on the right node 7 and they have a 55% chance of being a low valued customer. It allows us to gather that a customer's value is likely to be high when his credit score is greater than 582. Credit Score is an important variable as it reflects the reliability of a customer. Therefore, a higher credit score is usually linked to a higher customer value since they will have a lower risk assessment.

Generally, this model result is easy to read and White Rock can make use of this model to quickly classify and identify a customer's value during the onboarding process.

6.2.4 Variable Importance

Variable importance assesses the magnitude of the impact of an independent variable on a customer's value. As seen from the table below, 'Age', 'Credit Score' and 'Estimated Salary' of a customer have the largest impact on their value. Amongst these factors, 'Age' has the highest importance in determining a customer's value.

	Age	**Credit Score**	**Estimated Salary**
Variable Importance	16.069062	3.662301	2.287246
Proportion	72%	16%	12%

6.2.5 Predicted Results

Using the CART model, the predicted values of customers are obtained. The predicted values are then compared to actual values of the dataset in a confusion matrix as seen below.

Cust Value	**'High' (Predicted)**	**'Low' (Predicted)**
'High' (Actual)	1081	751
'Low' (Actual)	940	892

As seen from the confusion matrix, the sensitivity (true positive rate) can be calculated to be 1081/(1081+751) = 0.60 while the specificity (true negative rate) is 892/(892+940) = 0.486. This means that when the actual value was ‘high’, the model predicted 60% correctly whereas when the actual value was ‘low’, 48.6% was predicted correctly.

6.3 Evaluation & Analysis

The overall accuracy obtained for the model is 0.538 which translates to **54%**. While the accuracy is not very high, there is a slight improvement from the logistic regression model which suggests a better fit at predicting a customer’s value from the chosen dataset. Ultimately, the poor accuracy might be due to various factors such as imperfect information based solely on the chosen dataset. The limitations of the models will be further discussed in the section below.

7 Model Comparison

Before diving in to compare the models we have developed, a comprehensive research on each of the two models was analysed (Refer to Appendix). It is important to take note that each of the models have their own strengths and weaknesses in specific areas. Therefore, we will need to take into account the kind of predictive performance that the banks may require: accuracy, interpretability, ranking, probability estimation.Thus, while there is no one model that is superior, one may be more suitable than the other depending on the output that we are looking for.

Model Type	**Logistics Regression**	**CART**
Analytical Description	"*Age*" is statistically significant. Model Accuracy at 49% with a sensitivity of **58%** and specificity of 45%.	"*Credit Score*", "*Estimated Salary*" and "*Age*" are important variables. Model Accuracy at 54% with a sensitivity of **60%** and specificity of 49%.
Advantages	Easily identifies variables that might have a linear association with the Customer's value.	Does not require manual train test split since 10-fold cross validation is automatically carried out. Does not require backwards elimination to identify significant variables. Able to provide better predictions for data which is not linearly separable such as ours. Variables are identified and ranked by importance.
Disadvantages	Unable to handle data which is non-linearly separable. Does not handle skewed dataset well which necessitates rebalancing of dataset.	Requires pruning to reduce risk of overfitting.

We will choose CART over logistic regression due to the higher accuracy and the variables based on literature reviews were included in the CART model.

8 Limitations of the Model

8.1 Segmentation of Customer

Currently our model looks at Number of products, credit cards and active members to determine that they are customers of “high value” to the bank and White Rock. This is a simple method of categorizing the value of the customer and there may be other factors that affect the value of the customers.

8.1.1 Transactions of the customer over time

The concept of customer value (CLV), which was defined more than 30 years ago by Kotler as ‘‘the present value of the future profit stream expected over a given time horizon of transacting with the customer’’ (Kotler, 1974, p. 24). The balance given in the data set provided was more likely to be captured at a point in time. Our segmentation of customer value can be further improved if we have the transaction history of the customer with the bank over the given time period the customer was with the bank reflecting the monetary value and frequency.

8.1.2 Type of purchases

Banks usually offer multiple products and services and consumers globally hold an average of 7.4 banking products across all their various relationships. However, based on the dataset given the maximum of products owned by customers is 4. Due to this limitation, it affects how we segment our customers as we consider customers holding more than 1 product as “High value”. Additionally, according to Fader et al. (2005b) information on past customer recency, frequency and monetary value can be used to determine customer value with high accuracy. With data collected on the recency, frequency and monetary value of the products bought by customers we would be able to increase the accuracy in determining customer value.

8.2 Model was built upon data from 3 countries of a bank (E.g. France, Germany and Spain)

Model will be restricted to predict largely in the 3 countries while White Rock currently has 32 markets globally, there will be a need to build the model based on data collected from banks of different countries. This is due to the different income distribution and banking preferences across countries. Additionally, the model was built using data from a bank. If White Rock is working with multiple banks in the world, it would be desirable to replicate the model using data from other banks.

8.3 Predictors of Customer Value

Predictors of customer value have been analysed and listed by various researchers in the literature. Although some of these studies have emphasised common predictors or variables, there is no agreement due to the diversity of the industries investigated. The common variables are lifestyle characteristics of the customer (Haenlein et al., 2007), macroeconomic environment

and competition determine the activity level in the industry and the customer's brand-switching behaviour, which lead to changes in customer value generated at the company level (e.g. Gupta *et al.*, 2006). Due to the data set given, we are not able to factor such variables into our model.

9 Conclusion & Future Directions

Our model can aid White Rock and financial intermediaries to allocate resources efficiently in the process of on-boarding by predicting the value of the potential customer. Inefficient client onboarding will have a tremendous impact on customer loyalty, profitability and referrals by existing customers to new customers, reputation and brand equity.

Customers of “high value” should be prioritized to prevent these customers from switching to competitors in the industry and dropping out during the on-boarding process.

Our current model can be further improved by capturing more data of the customer that can influence the value of the customer. Subsequently, the team would then revise our model and increase the accuracy of the model to predict the value of potential customers.

In conclusion, there is a need for White Rock and intermediaries to mitigate huge potential loss of customers in client-onboarding.

10 Appendices

General Comparison between CART and Logistic Regression

	CART	**Logistic regression**
Input data types	Works well with categorical data	Unable to handle pure categorical data thus there is a need to convert the data into numerical format.
Missing values	Missing values are handled automatically through surrogates	Does not handle missing values.
Is the data highly skewed?	Handles skewed data well if the tree is allowed to be grown fully without cutting off or pruning.	Skewed data can be handled by balancing the dataset.
Decision boundaries	Non-linear classifiers. They do not require data to be linearly separable. When generating decision boundaries, decision trees “bisect the space into smaller and smaller regions. However, when classes are not defined or separated very well, it can lead to overfitting.	Assumes that the data is linearly (or curvy linearly) separable in space. When generating decision boundaries, logistic regression fits a single line to divide the space exactly in two which could be limiting when dealing with higher dimensional data.
Decision making	Automatically handles decision making	A decision threshold has to be set

References

Barroso Castro, C. and Martín Armario, E. (1999), “Nivel de Servicio y Retención de Clientes: El Caso de la Banca en España”, Revista Española de Investigación de Marketing ESIC, Vol. 3 No. 1, pp. 9-33.

Baumann, Chris & Elliott, Greg & Burton, Suzan. (2012). Modeling customer satisfaction and loyalty: Survey data versus data mining. Journal of Services Marketing. 26. 148-157. 10.1108/08876041211223951.

Berger, Paul D. and Nasr, Nada I. (1998) Customer lifetime value: Marketing models and applications. Journal of Interactive Marketing 12(1), 17–30.

Campbell, Dennis and Frei, Frances (2004) The persistence of customer profitability: Empirical evidence and implications from a financial services firm. Journal of Service Research 7(2), 107–123.

Customer Segmentation: Where Banks Can Find Value. (2017, February 8). CSP. https://www.csp.com/customer-segmentation-value/#.X57CiogzY2x

Dayani, D. (2017, July 4). Here Is How Credit Cards Really Work, And How Banks And Credit Card Companies Make Money From Us. DollarsAndSense.Sg. https://dollarsandsense.sg/here-is-how-credit-card-really-works-and-how-banks-and-credit-cards-companies-make-money-from-us/

DeAsi, G. (2020, October 21). 10 Powerful Behavioral Segmentation Methods to Understand Your Customers. Pointillist. https://www.pointillist.com/blog/behavioral-segmentation/

Deloitteeditor. (2019, October 3). Gender Differences in Banking Behavior. Deloitte. https://deloitte.wsj.com/cmo/2019/03/26/gender-differences-in-banking-behavior/

Editorial Team. (2020, May 13). Are banks that run on legacy systems able to compete with their digital counterparts? Finextra Research. https://www.finextra.com/blogposting/18751/are-banks-that-run-on-legacy-systems-able-to-compete-with-their-digital-counterparts

Ekinci, Y., Uray, N. and Ülengin, F. (2014), "A customer lifetime value model for the banking industry: a guide to marketing actions", European Journal of Marketing, Vol. 48 No. 3/4, pp. 761-784. https://doi.org/10.1108/EJM-12-2011-0714

F. (2018a, June 28). New Research Study Measures the Time, Cost and Challenges of ClientOnboarding.https://www.prnewswire.com/news-releases/new-research-study-measures-the-time-cost-and-challenges-of-client-onboarding-290630341.html

Fader, Peter S., Hardie, Bruce G.S. and Lok Lee, Ka. (2005b) RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research 42(4), 415–430

Fernandez V. Maria and Barbon F. (2019) Credit Scoring in Financial Inclusion. https://www.cgap.org/sites/default/files/publications/2019_07_Technical_Guide_CreditScore.pdf

Greonfeldt Tom (2012, August 23). Tech and Geography Beat Too Big To Fail at Wells Fargo. https://www.forbes.com/sites/tomgroenfeldt/2012/08/23/tech-and-geography-beat-too-big-to-fail-at-wells-fargo/?sh=6ce295754902

Gundaniya, N. (2020, August 28). Future of customer onboarding in banks. Digital Finance Solutions, Ewallet Payment System, Wallet App Development. https://www.digipay.guru/blog/future-of-customer-onboarding-in-banks/

Haenlein, M., Kaplan, A. and Beeser, A. (2007), "A model to determine customer lifetime value in a retail banking context", European Management Journal, Vol. 25 No. 3, pp. 221-234.

Jain, Dipak and Singh, Siddhartha S. (2002) Customer lifetime value research in marketing: A review and future directions. Journal of Interactive Marketing 16(2), 34–46.

Kotler, Philip (1974) Marketing during periods of shortage Journal of Marketing 38(3), 20–29

KPMG (2016). Transforming client onboarding. https://assets.kpmg/content/dam/kpmg/pdf/2016/07/transforming-client-onboarding.pdf

Medallia. (2018). The Customer Experience Tipping Point. https://go.medallia.com/rs/669-VLQ-276/images/Medallia_Ipsos_The_Customer_Experience_Tipping_Point.pdf

Marriage, M. (2015, November 30). Banks embrace 'the age of asset management.' Financial Times. https://www.ft.com/content/40c39d3e-9440-11e5-b190-291e94b77c8f

Nguyen, T. A. (2019, June 3). Customer Segmentation: A Step by Step Guide for Growth-OpenViewLabs.OpenView.https://openviewpartners.com/blog/customer-segmentation.X52CeIgzY2x

Q. (2017, June 19). What Are The Advantages Of Logistic Regression Over Decision Trees? Forbes.

Segmentation of Bank Customers by Age | Emerald Insight. (1985, March 1). Emerald Insight. https://www.emerald.com/insight/content/doi/10.1108/eb010761/full/htmlhttps://www.forbes.com/sites/quora/2017/06/19/what-are-the-advantages-of-logistic-regression-over-decision-trees/?sh=629fc9ab2c35

Rust, Roland T., Lemon, Katherine N. and Zeithaml, Valarie A. (2004) Return on marketing: Using customer equity to focus marketing strategy. Journal of Marketing 68(1), 109–127.