Term Project Final Report

Section A: Team 1
Authors: Sherry Tao, Winston Zhong, Cecilia Li, Koustubh Pareek, Henry Wu

Business Understanding

Credit cards, with over 176 million users in the U.S., have become an increasingly popular payment method and banks' revenue stream. Since acquiring a new customer is up to 25 times as expensive as retaining an existing one (Gallo, 2014), our team endeavours to investigate factors related to customers' churn rates and help banks predict those who are at risk of attrition. By performing data visualisation, including plotting and clustering, we hope to gain a better understanding of some characteristics of current cardholders. We also aim to train and test several predictive models, to predict whether a customer will churn and attempt to pick up some key variables that are most related to it. The bank can then explore looking into changing some variables that are related to churn, for example raising the credit limit, and assess the cost and benefits to achieve their business goals, which may be to improve the overall profit per customer.

Data Understanding and Preparation

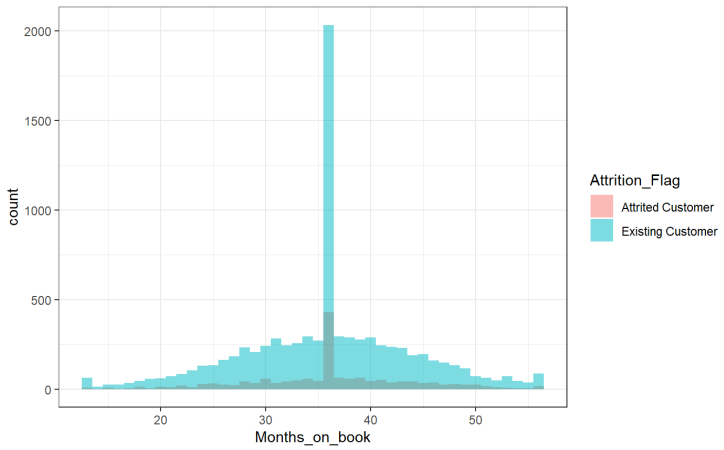
Our dataset is from Kaggle and contains over 10,000 rows and 23 columns of customer data, including an independent variable, Attrition_Flag, and 19 potential dependent variables that are can be used in our data mining process. There are no NAs or duplicate rows in our data.

After inserting our data into R, we dropped the columns that are irrelevant to the data mining process. The columns dropped are the Client Number, which is the unique identifier for each customer, and the last two columns which are not customer related variables. We created a target variable column called Churn using the Attrition_Flag column and calculated

the overall Churn rate to be 16.07%. Our cleaned data contains 10,127 rows and 21 columns. Of the 21 columns, 5 were non-numeric and we converted them to factors to support the data mining process.

Exploratory Data Analysis

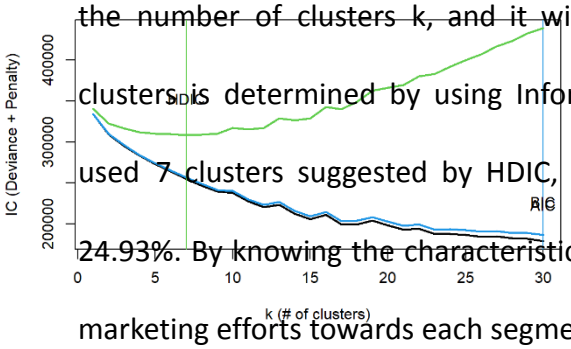
Our team wanted to know if there are any obvious differences in customer characteristics between the existing and attritted customer groups. We used visualisations to explore our data, specifically histograms of existing and attritted customers against Customer Age, Dependent_count, Months_on_book, Total_Relationship_Count, Months_Inactive_12_mon and Contacts_Count_12_mon to see if there are differences between the two groups. While visually they do not seem different, one interesting thing that we found was that there is a large spike in customers at the 36-month mark in the Months_on_book graph, suggesting that many customers applied for a credit card with the bank at that time.

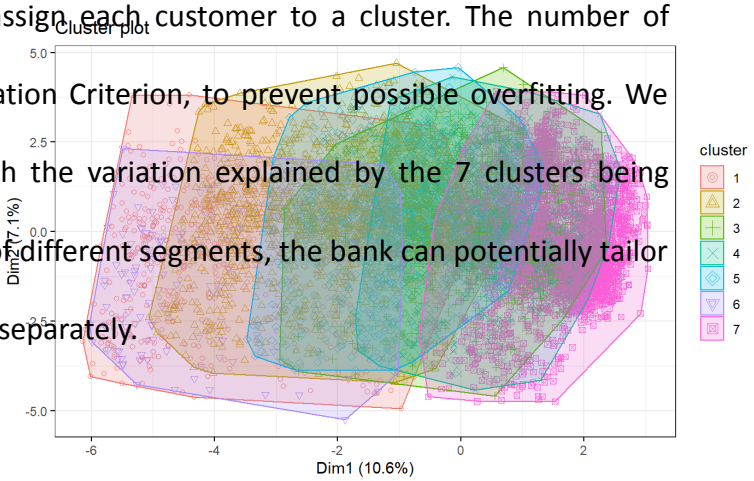


<u>K-means</u>

Our team used K-means clustering to explore our data to improve our understanding of the customers' characteristics. This unsupervised machine learning method segments customers such that those within the cluster are similar but they are dissimilar to those in other clusters. It is a practical way to segment customers, since we just have to input our data and

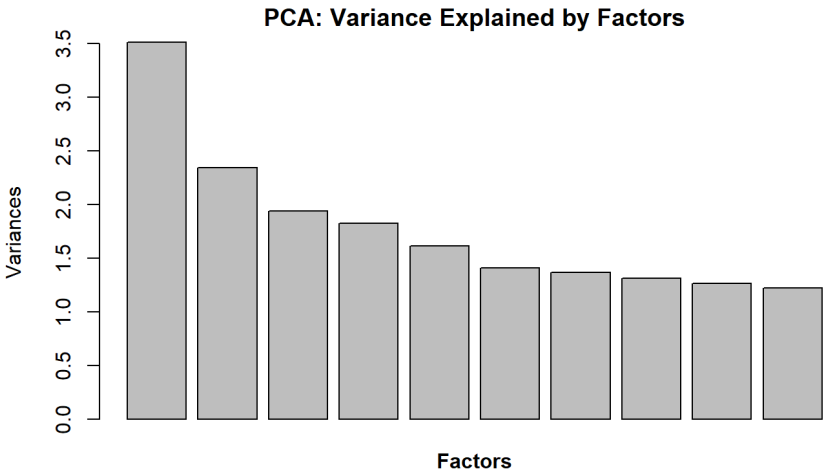
the number of clusters k, and it will assign each customer to a cluster. The number of clusters is determined by using Information Criterion, to prevent possible overfitting. We used 7 clusters suggested by HDIC, with the variation explained by the 7 clusters being 24.93%. By knowing the characteristics of different segments, the bank can potentially tailor marketing efforts towards each segment separately.





Principal Component Analysis (PCA)

Our team also used PCA, another unsupervised machine learning technique that extracts and creates a set of latent features from a high-dimensional dataset to capture as much information, or variance, as possible in the data. Latent features cannot be measured directly and is a combination of variables, and we can see the correlation between original variables and principal components. The variation explained by the top 4 latent features is 29.1%.



	cumulative.variance.percent <dbl>
Dim.1	10.6
Dim.2	17.7
Dim.3	23.6
Dim.4	29.1

Limitations of K-means and PCA

Although both K-means and PCA are simple to use and adopt, they each have their limitations. For K-means, we may not be able to select the most optimal number of clusters, despite using Information Criterion. In our case, AIC and BIC did not manage to find a suitable k less than 30. K-means also have trouble clustering data where clusters are of varying sizes and densities. Furthermore, the variance explained by the clusters may be small, which is the case for our dataset.

As for PCA, it tends to find linear correlations between variables, which is sometimes undesirable. It could fail in cases where the mean and covariance are not enough to define datasets. Furthermore, similar to K-means, the variance explained by the 4 latent features may be small and is the case for our dataset as well.

Modelling

The goal of our models is to make accurate predictions as to whether a customer will churn based on their characteristics. We would also like to explore the significant variables that the company can influence to reduce attrition. Since we are solving a classification problem, we utilized two practical approaches: Statistical Procedure Based Approach and Machine Learning-Based Approach. Six models were built and tested: Logistic Regression (LR), LR with Lasso, LR with Post Lasso, Classification Tree, Extreme Gradient Boosting (XGBoost), and Neural Network.

Logistic Regression

Our team ran a logistic regression with all variables, and many are significant variables as indicated by their p-values.

```
Call:
glm(formula = Churn ~ ., family = "binomial", data = data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1363  -0.3630  -0.1703  -0.0676   3.5363

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                7.080e+00  5.333e-01  13.275  < 2e-16 ***
Customer_Age              -4.077e-03  8.652e-03  -0.471 0.637446
Dependent_count            1.382e-01  3.346e-02   4.129 3.64e-05 ***
Months_on_book            -9.596e-03  8.592e-03  -1.117 0.264092
Total_Relationship_Count  -4.538e-01  3.087e-02 -14.699  < 2e-16 ***
Months_Inactive_12_mon     4.996e-01  4.220e-02  11.838  < 2e-16 ***
Contacts_Count_12_mon      4.791e-01  4.089e-02  11.717  < 2e-16 ***
Credit_Limit              -2.101e-05  7.698e-06  -2.729 0.006353 **
Total_Revolving_Bal       -9.789e-04  8.176e-05 -11.974  < 2e-16 ***
```

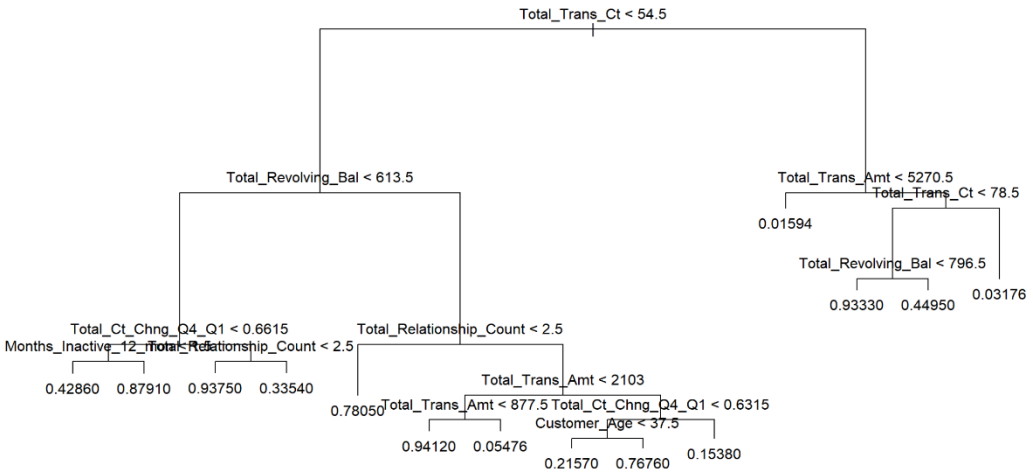
The model gives us six significant factors that have an impact on the churn rate: number of dependents (Dependent_Count), number of Bank Products held by the customer (Total_Relationship_Count), number of months inactive in the last twelve months (Months_Inactive_12_mon), number of contacts in the last twelve months (Contacts_count_12_mon), credit limit, and total revolving balance on the card. Three of the variables have positive coefficients, which are number of dependents, number of months inactive, and number of contacts in the last twelve months. This means that as the number of dependents increase by 1 unit, the log odds of churn increase by 0.138.

Logistic Regression with Lasso and Post Lasso

Our team used Lasso and Post-Lasso to regularise the Logistic Regression using the theoretical choice of λ. The regularisation aims to achieve a better Out of Sample prediction.

Classification Tree

Our team ran a classification tree, which is easy to interpret and visualise.



Machine Learning

Our team tried two machine learning methods, XGBoost and Neural Network. XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient-boosting framework. Neural Network is a type of prediction model comprised of an input layer, one or more hidden layers, and an output layer. For our Neural Network model, we chose 3 hidden layers with ReLU activation for each hidden layer and an output layer with sigmoid activation. The activation function can be changed to create different models, and regularisation and dropout can also be added to models which are more complex to prevent overfitting. Although the accuracy of both machine learning methods exceed the statistical approaches, they are uninterpretable, and it becomes difficult to form recommendations in terms of changing certain variables to influence customer behaviour.

Evaluation

We first split the data into the training and testing data, with 80% of the dataset set as training data and the remaining being testing data. The same training and testing data were used to train and test the 6 models, to ensure comparability between the models during evaluation. We used out of sample metrics for evaluation since our goal was to have accurate predictions of future customers.

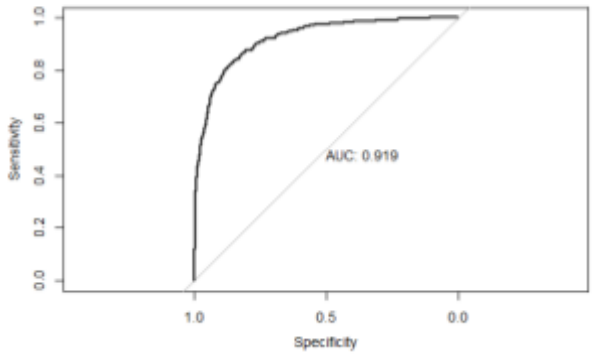
Our team decided to use the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) graph as our evaluation criteria for the 6 models. The ROC curve plots the True Positive Rate vs False Positive Rate as the threshold varies, and AUC is the probability that a random chosen positive instance will be ranked higher than a randomly chosen negative instance. AUC ranges from 0 to 1, with random guessing being 0.5, and the higher the AUC, the better the model is at predicting if a customer will churn or stay.

Logistic Regression, Lasso and Post-Lasso

The logistic regression yielded an AUC of 0.919, which is considerable for a relatively simple model. This may be due to the imbalance of the data, with only 16% of customers churning. Using Lasso did not improve the AUC and returned identical results as logistic regression and Post-Lasso even decreased the AUC to 0.918.

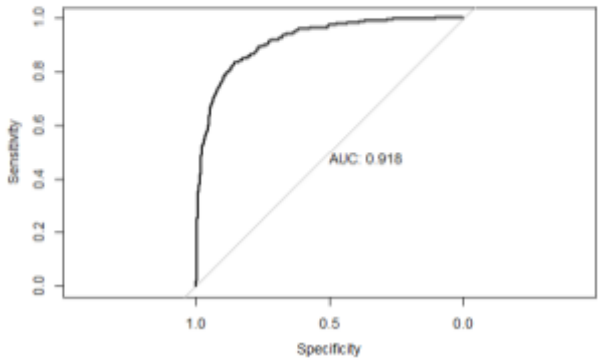
Logistic Regression

```
         6         14         18         20         22         28
0.11148214 0.00204744 0.07766357 0.10514043 0.96508314 0.06208084
```



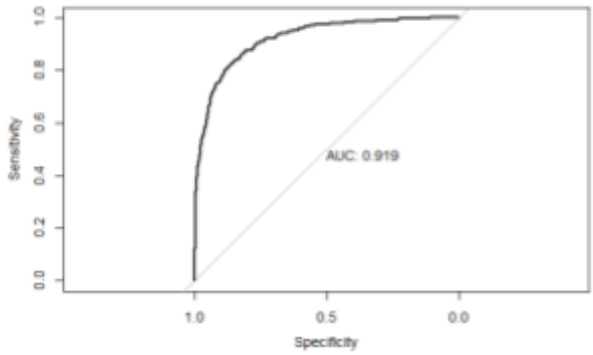
Logistic Regression with Lasso

```
                  s0
6  0.121418115
14 0.002189179
18 0.078591041
20 0.111289044
22 0.963313105
28 0.066425802
```



Logistic Regression with Post-Lasso

```
         6         14         18         20         22         28
0.184137520 0.002522187 0.053006201 0.135113712 0.971153316 0.080508925
```

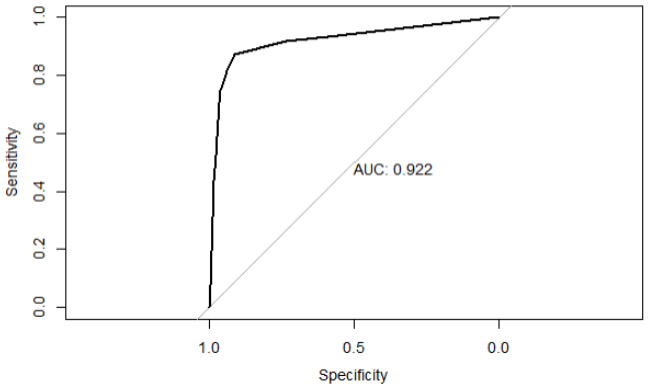


Classification Tree

The classification tree performed better than logistic regression and lasso, obtaining a higher AUC of 0.922.

Classification Tree

```
        6         14         18         20         22         28
0.05288462 0.05288462 0.05288462 0.05288462 0.83225806 0.05288462
```

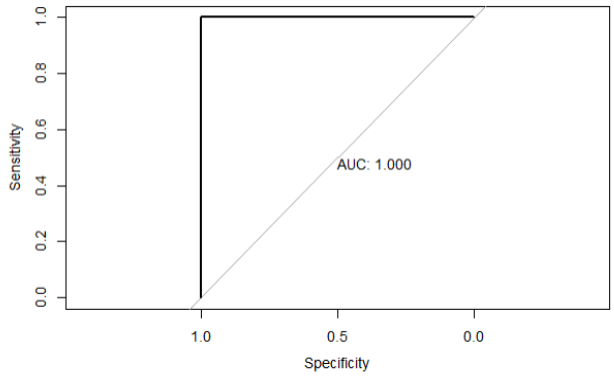


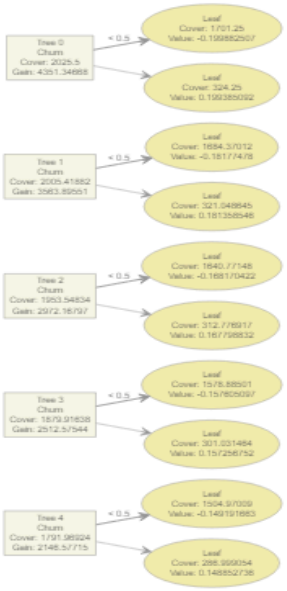
Machine Learning

Both machine learning methods yielded the highest AUC curves at 1, which means that they accurately predicted all 2025 customers in the testing data.

XGBoost

```
[1] 0.2980450 0.2980450 0.2980450 0.2980450 0.7015398 0.2980450
```

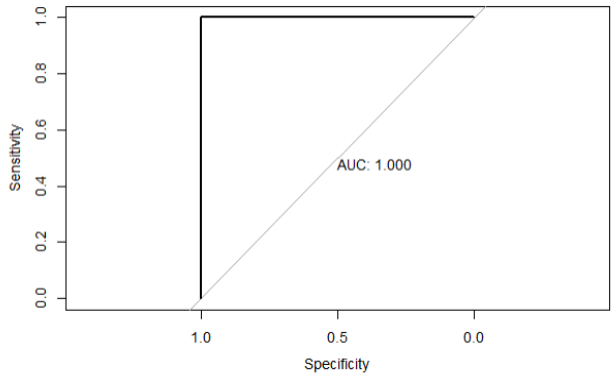


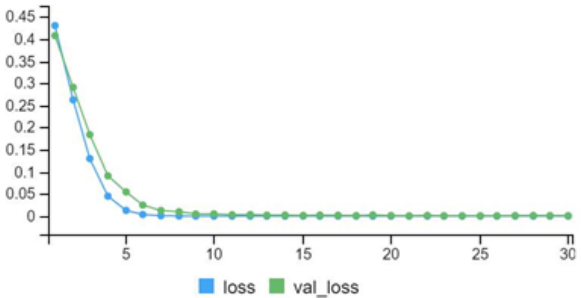


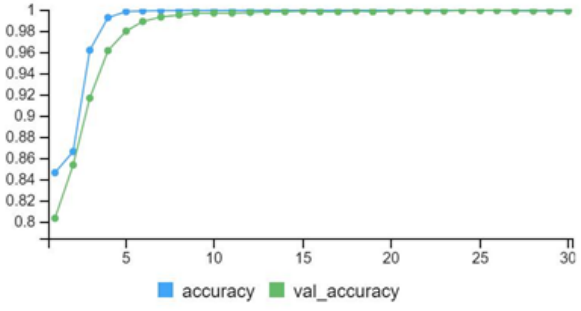
Neural Network

```{r}
head(pred.NN1)
```

```
             [,1]
[1,] 2.264654e-13
[2,] 1.202584e-12
[3,] 2.079260e-26
[4,] 4.121057e-10
[5,] 1.107943e-14
[6,] 1.016979e-11
```







Deployment

The models obtained through the data mining process can be used by the bank to make business decisions. The bank can combine the result of our models with a Cost-Benefit Matrix, thereby combining prediction with the business setting. For example, the bank can choose to increase the credit limit of customers who are at risk of attrition. A drawback on that would be that the customer might default, which is a potential cost. The calculation of expected profit would lead to a business decision on whether to increase the credit limit of each customer and by how much.

Another example of a business application is to offer cashbacks after the customer spends a certain amount with their credit card. In this case, for a customer with characteristic X

$$E[Profit \mid X,\ Cashback] = P(X, Cashback)E[V(X, D)|noChurn,\ X,\ Cashback]$$

And the goal is the maximise $E[Profit \mid X,\ Cashback]$.

<u>Limitations associated with using data mining models</u>

As with most data-driven approaches, one limitation of the data is that it may not be representative of the population. This is due to the low response rates of bank surveys, typically between 5% and 30% (Dube, 2020). As such, the models trained on this data may not capture all customer characteristics that are relevant for prediction, and hence be unable to accurately predict all customers.

Another limitation is that credit card may not be the reason that a customer churns from a bank. Other factors, such as portfolio performance and quality of service that are not included in the model may be the true cause of attrition (Reier, 1999). However, finding out these factors would require the banks to perform causal analysis instead of simply relying on a predictive model.

The third limitation is related to the goals of the bank. If the goal of the bank is to increase profits, there may be other factors that are better at achieving this goal such as offering more loans, or encouraging customers to have more deposits. The data mining model is extremely specific regarding the credit card data and may not be able to be used in solving other business goals.

Finally, there are some variables in the model that are difficult for the bank to control. Examples include Education_Level, Marital_Status, Income_Category, etc. Without the ability to control some variables, the bank is limited in the ways it can influence the customer to remain with the bank.

Appendix

References

Dube, D. (2020, May 27). *Why Retaining Customers For Banks Is As Important As Winning New Ones*. Retrieved from Forbes: https://www.forbes.com/sites/forbestechcouncil/2020/05/27/why-retaining-customers-for-banks-is-as-important-as-winning-new-ones/?sh=6c6e3d933f98

Gallo, A. (2014, October 29). *The Value of Keeping the Right Customers*. Retrieved from Harvard Business Review: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

Reier, S. (1999, October 28). *Customer Loyalty Is Harder to Maintain : Making Sure Service Keeps Clients Happy*. Retrieved from The New York Times: https://www.nytimes.com/1999/10/28/news/customer-loyalty-is-harder-to-maintain-making-sure-service-keeps.html