**Practice Problem: Classification**

A manager at the bank is disturbed with more and more customers leaving their credit card services. They would really appreciate if one could predict for them who is gonna get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction

CLIENTNUM: client ID

Attrition\_Flag: 1 – churned customer, 0 – existing customer

Customer\_Age: age of a customer (numeric)

Gender: gender of a customer (F, M)

Dependent\_count: the number of dependents in the household (numeric)

Education\_Level: High School, Graduate, Uneducated, Unknown, College, Post-Graduate, Doctorate

Marital\_Status: Married, Single, Unknown, Divorced

Income\_Category: $60K - $80K, Less than $40K, $80K - $120K, $40K - $60K, $120K +, Unknown

Card\_Category: Blue, Gold, Silver, Platinum

Months\_on\_book: numeric

Months\_Inactive\_12\_mon: months a customer has been inactive

Credit\_Limit: numeric

Total\_Revolving\_Bal: numeric

Avg\_Utilization\_Ratio: how often a credit card is used

In using the above variables, please find the answers to the following questions. **Please use Attrition\_Flag as Y variable and all others as X variables**.

For a cut-off of .5 using Linear Classification (using Stepwise regression):

1. What is the R-square of the model? What does this number suggest?

*在fit model的時候不要按go，把所有parameters勾選起來再run model*

A picture containing table

Description automatically generated

1. What is the logOdds equation?

Table

Description automatically generated

1. What is the p-Value of the model and variables? What does this suggest?

Table

Description automatically generated

The p-value for the certain variables (see above) are above 0.05 which suggests that those variables are not statistically significant in predicting the churn rates.

1. What variable has the largest impact on the churn rate?

*Effect summary chart*

The major predictor is the number of months a user is inactive followed by total revolving balance and average credit card utilization rate

Table

Description automatically generated

1. What is the level of accuracy you see in your model?

*看到accuracy就要手算了*

About 83%

*為什麼這邊不是用交互驗證*

Application, table, Excel

Description automatically generated

1. What is the significance of the threshold? Does increasing the threshold to .95 give us a more accurate prediction?

As we increase a threshold, we get relatively the same accuracy level. However, the number of false negative and false positive predictions rise significantly.

Application, table, calendar

Description automatically generated

For a cut-off of .5 using Decision Tree Classification use **validation portion 0.4:**

Graphical user interface, application

Description automatically generated

1. What is the r square value for the training and validation data set?

Qr code

Description automatically generated

1. How many splits were made before the R square of the testing data started to drop?

Chart, line chart

Description automatically generated

9 splits

1. Why there were 9 splits?

The model found the optimal number of splits that prevent overfitting. The R-squared for the validation potion of the dataset starts dropping after 9 splits.

1. Which top 3 variables are important in making a decision? Which is the most important one?

Total Revolving Balance, Months Inactive, **Months on book (→箭頭小選單columns contribution)**

1. According to your model, which are the features of people who have a high likelihood of churning?

*看tree 藍色最多的*

who stayed more than 2 months inactive and have revolving balance of between 159 and 461

1. What is the probability a person who has total revolving balance greater than 600 and being inactive more than 2 months?

0.0961

看tree

Diagram, schematic

Description automatically generated

1. What is the level of accuracy you see in your model?

Table, calendar

Description automatically generated