**Practice Problem: Classification with Logistic Regression and Decision Trees**

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 going to 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 0.5 using Linear Classification (using Logistic regression):

1. What is the R-square of the model? What does this number suggest?
2. What is the logOdds equation?
3. What is the p-Value of the model and variables? What does this suggest?
4. What variable has the largest impact on the churn rate?
5. What is the level of accuracy you see in your model?
6. What is the significance of the threshold? Does increasing the threshold to .95 give us a more accurate prediction?

For a cut-off of 0.5 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?
2. How many splits were made before the R square of the testing data started to drop?
3. Why there were 9 splits?

1. Which top 3 variables are important in making a decision? Which is the most important one?
2. According to your model, which are the features of people who have a high likelihood of churning?
3. What is the probability a person who has total revolving balance greater than 600 and being inactive more than 2 months?
4. What is the level of accuracy you see in your model?