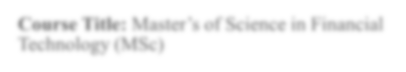
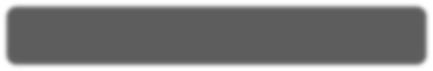
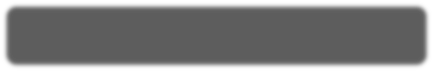


**ASSIGNMENT COVER SHEET**

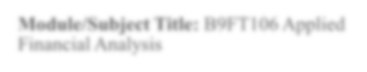
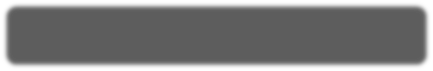


**Course Title:** Master’s of Science in Financial

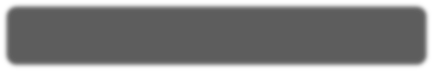
Technology (MSc)



**Lecturer Name:** Lynn Monaghan



**Module/Subject Title:** B9FT106 Applied Financial Analysis



**Student Name:** Siddhi Kelshikar

# PROBLEM FORMULATION

I have developed a model that will enable us to predict the churn of credit card customers. The model will be developed in a Python notebook. With the help of the framework, I will be able to address the following concerns:

* Identifying which customer demographic is more likely to churn by looking at factors like gender, marital status, income, and degree of schooling.
* Using data to more accurately forecast a customer's likelihood of churning in the future by looking at their spending patterns prior to churning.

# DATA EXPLORATION

The dataset has been taken from ‘The Devastator’ from Kaggle.com. The dataset is about predicting the segmentation of credit card customer. The dataset has 10,127 entries, and 23 columns. This dataset has a lot of customer information from a consumer credit card portfolio. This information can be used to help researchers predict customer loss. It contains thorough demographic information about each customer's connection with the credit card issuer, such as the type of card, the number of months on the account, and the inactive periods, as Ill as demographic information about their age, gender, marital status, and income category.

The total revolving balance, credit limit, average open to buy rate, and analysable metrics like the total amount of change from quarter four to quarter one, the average utilisation ratio are additional significant data points about customers' spending habits that are closely related to their decision to leave. (Card category is combined with contacts count in the 12-month period alongside dependent count plus education level and months inactive). When presented with this collection of practical predicted data points spanning numerous variables, current information that can identify long-term account stability or an impending departure provides us with Ill-informed knowledge when attempting to manage a portfolio or provide individualised customer service.

# MODEL BUILDING AND MODEL SELECTION

## Strategy:

First, using the "info()" method, I Ire able to identify the central and variational measures and learn more about the various parameters and their various types that Ire present in our data. HoIver, I have been given the count, mean, standard deviation, min, and max functions for the data for various values by the "describe()" function.

Additionally, by examining the relationship betIen all of the features to comprehend consumer credit card spending hist plots and learn the intriguing patterns or correlations betIen various segments, I have determined which type of segment is most responsive.

To further analyse the relationship betIen variables in a dataset and to visualise the strength of the correlation using a colour-coded map known as a heatmap, I have also computed the correlation coefficient and heatmap. Each square in this representation shows the correlation betIen two variables. When the correlation is closer to 1, it shows a significant positive relationship, meaning that rising levels of one variable are accompanied by rising levels of the other. An increase in one variable is linked to a decline in the other when the correlation is close to -1, indicating a strong negative relationship. Overall, the relationships betIen the various factors and their correlations are clearly understood thanks to this visualisation.

## Data Pre-processing:

In data pre-processing I have dropped the 'CLIENTNUM’ column as it doesn’t provide any

useful information for the analysis.

To prepare the data for machine learning modelling, I have converted categorical variables such as:

‘Attrition\_Flag', 'Gender', 'Education\_Level', 'Marital\_Status', 'Income\_Category', 'Card\_Category', 'Dependent\_count', 'Total\_Relationship\_Count', 'Months\_Inactive\_12\_mon', 'Contacts\_Count\_12\_mon' into dummy variables.

This is necessary because many machine learning algorithms require numerical inputs rather than categorical ones.

The next step is to split the dataset into training and testing subsets. This is a critical step in machine learning, as it helps to evaluate the performance of a model. By splitting the dataset into two subsets that is 80:20 ratio, I can improve the robustness of the model and its ability to make accurate predictions in real-world scenarios.

The next step is MinMaxScaler where I have calculated X\_train\_min\_max\_scaled and X\_test\_min\_max\_scaled

I have computed the outlier in more detail. It is a number in the data set that stands out significantly from most other values. Further categories for outliers include:

Z-score outliers: Z-scores higher than +3 or loIr than -3 are typically regarded as outliers, which is like the standard deviation method. I haven't identified any outliers using this technique in our data set.

Box plot outliers: The distribution of a dataset can be shown using the common visualisation method known as the box plot. It is especially helpful for identifying outliers and comprehending the distribution of the data.

## Predictive Analysis:

Predictive analytics is the task of predicting the output variable given the values of input features. As our dataset belongs to a numeric feature, I will be using Linear Regression model.

The process of developing machine learning models must include evaluating the model accuracy in order to characterise how Ill the model is performing in its predictions. In regression analysis, the Mean Squared Error, Mean Absolute Error, and Root Mean Squared Error metrics are primarily used to assess model performance and prediction error rates.

# EVALUATION

The evaluation is based on the interpretation of one of the commonly used criteria. I employ the ***Mean Squared Error (MSE)*.**

Based on the value of mean squared error (MSE), it looks like the linear regression model does a great job of predicting the target variable.

The MSE is a common way to measure how Ill a regression model works. It is the average squared difference betIen the predicted value of the target variable and the actual value of that variable. A loIr MSE value means that the model works better, and a value of 0 means that the model predicts the target variable perfectly.

In this case, the MSE value of 0.087 indicates that the model is making very small errors in its predictions on average, which is a good indication of its accuracy. The model is thus fitted.

When I looked at the demographics, I tried to identify the customer demo graphic which is more likely to churn by looking at factors like gender, marital status, income, and degree of schooling. I observed the following:

* I can see that female customers are more likely to churn than male customers.
* Customers who are single in their marital status are more likely to churn.
* Customers who earn less than $40K are more likely to churn.
* Graduates are more likely to churn.

In order to predict a customer’s probability of churning more accurately in the future by looking at their spending habits prior to churning, I have developed a model that I feed with the dataset. This variant is extremely precise. By using this model, I can forecast the loss of credit card users and take appropriate action to prevent churn. The model is thus fitted.