Module 6 Assignment: Predictive Model Exercise

1. ***In addition to the spreadsheet/code or programming output you submit, include a separate written document of 250-500 words that summarizes your process.***

I performed the following steps to create a logistic regression model for the given data:

* 1. Data Ingestion - I read the data using python – pandas as a dataframe (table) and analysed each variable given in the data set at the beginning by finding the statistics like mean, median, etc.; for the numerical values and count of values for the categorical variables to get a better understanding of the data being dealt with. I also checked for the presence of any null values row wise and column wise to see if the data needed to be cleaned on that perspective.
  2. Data Cleaning – Although the enrolment date column is not properly formatted, I decided to convert it into the format ‘YYYY-MM-DD’ to further break it down into columns month and year so that I could check if these parameters are closely related to churn or not. I didn’t want to lose out on too much data that was provided and try to garner as much information as I could. I also filtered the data by age to include 16+ and below 79 as there are rows with age 0 and 99+ which wouldn’t help the model and their removal doesn’t affect much (out of 24843, 2261 rows were filtered out).
  3. Data Visualization – To further clarify my categorical data and my numerical data, I visualized each variable using a histogram. I also did check the correlation between the numerical parameters however, their relationships were very weak to be considered.
  4. Data Pre-processing - Created dummy variables of the data as it allows easy interpretation and calculation. I further dropped the enrolment date column as I had the monthly and yearly information and the date was no longer useful. I created the down payment percentage column and checked it for null values. If present, I replaced them with the average value.
  5. Model Building (Logistic Regression) – I split the data into two tables: one with all features and the other without default which is the output of the prediction as Boolean values. These are then split into data for training and testing the model on an 80/20 percent basis.
  6. Confusion Matrix – On creating a confusion matrix, I found the below results:
     + True Negative - 3864
     + False Positive - 111
     + False Negative - 350
     + True Positive – 192

Which shows a good accuracy for the model. I also created an ROC curve to check the model’s accuracy and it’s about 91.2% accurate.

* 1. Features Selection – I identified the most important features of the model for future use.

1. ***Discuss why certain variables you expected to be significant are/are not and any other unexpected insights.***

* It was interesting to find that the number of defaulters is less than the ones not defaulting.
* While I accurately identified payment types Credit Card and Debit Card to be important features, I found it interesting that use bins were more important than payment types cheque and cash which were found to be highly related to churn during my exploratory data analysis.
* While I expected the additional columns year and month that I created to not have much significance, year 2021 was the third most important feature (probably due to not having information for the entire year) and months was more significant than down payment and price.
* Similarly, while I predicted age to be in the top 3 important features, it was in the bottom 3.
* Clearly, payment type and use should be considered for the model building more.