**University of Technology Sydney**

**36106:** [**Machine Learning Algorithms and Applications**](https://canvas.uts.edu.au/courses/26202)

**Assignment 3 – Final Report**

**Report:**

**Student Name and ID:**

Kin Hang Chan - 14072995

Aibarna Basnet -

Rohit Sharma - 24590960

**Date: 26/05/2023**

**Business Understanding**

The bank has been collecting transactional data from customers for 3 years. Bank transactions have been recorded in multiple csv files. Combining with the customer database, the data will form a strong base for analysis.

The bank’s newly established data scientist team has been engaged in this confidential project. The project objective is to utilise the data collected in order to bring direct value to the business or the end customers. After thorough review of the data on record, the team concludes the following business use cases could be adopted.

1. To predict customer’s total spending expenditure for the upcoming month
2. To identify fraud transactions based on
3. To cluster our customers with similar behaviours and preferences in groups based on their age and the expenditure.

This report will explain the business problems identified, how these applications will provide a solution, and how they will benefit our customers and our business.

**Data Understanding and Data Preparation**

The dataset being used comprises transactional data collected from the bank’s computer system and online banking platform over a four-year period, spanning from 2018 to 2022. It contains a vast amount of information, including 4,260,904 rows of records and 10 features. Additionally, customer data from a separate dataset, consisting of 1,000 rows and 15 features, has been merged, resulting in a comprehensive dataset with detailed transaction and customer information.

The dataset contains features such as Credit Card number, Customer Account number, Customer Transaction number, Transaction time, Category of purchase, Transaction amount, Fraud Indicator, SSN number etc.

It is seen that the given dataset's features are only marginally useful for machine learning tasks, highlighting the necessity for feature engineering approaches to extract more useful data.

During dataset preparation, several techniques were performed. The ‘dob’ column was converted into datetime format, and the age was calculated based on the current date. The ‘amt’ column and the calculated ‘age’ were appropriately converted to their respective data types. Categorical variables like 'category', 'gender', 'zip', 'trans\_wkday', 'trans\_hr', 'amt\_bin', 'city\_pop\_bin', 'dist\_bin', and 'age\_bin' were encoded to numeric values for modelling purposes. Additionally, data binning was performed on 'amt', 'city\_pop', 'dist', and 'age' to create binned variables for Classification purpose. Certain columns such as 'cc\_num\_x', 'acct\_num', 'merchant', 'job', and unnecessary columns related to transaction time were dropped from the dataset to simplify and avoid collinearity. Also, a new DataFrame called "avg\_monthly\_spending" was made to determine the average monthly spending by category. Overall, the code applied data cleaning, feature engineering, binning, encoding, and column dropping techniques to prepare the dataset for further analysis and modelling.

Furthermore, it is clear that the dataset is very unbalanced, with the majority of samples (99.88%) representing non-fraudulent transactions and the minority of samples (0.012%) representing fraudulent transactions. This extreme class disparity makes it difficult to create efficient algorithms that can reliably detect fraudulent behaviour.

Several cities stand out for having significant transaction volumes among the 726 cities in the dataset. Notably, the top cities with much larger transaction volumes than others include Brooklyn, Denver, Austin, Houston, and Dallas. With regard to the 'shopping\_pos' category, Brooklyn has the most transactions, indicating a significant customer presence. The category 'shopping\_pos' has the most transactions out of the 14 available transaction types, demonstrating a strong preference for retail purchases. Contrarily, the analysis shows that customers typically spend less on transactions involving travel, identifying a potential area for additional research and focused marketing efforts.

Interesting insights are revealed by further research of gender-based transaction patterns. According to the research, men spend a significant amount of money on gas and other transportation-related expenses, underscoring the significance of this category in their spending patterns. Females, on the other hand, show a stronger propensity to engage in food shopping activities, suggesting their attention to home necessities. Seasonal patterns in transaction activity also offer important insights. It is noteworthy that the data shows that the summer season saw the most transactions, mostly fuelled by purchases in the Food & Dining category. This finding implies that people have a stronger a tendency to go out to eat or engage in food-related activities in the summer. On the other hand, Autumn shows the lowest number of transactions, suggesting that overall consumer spending may slow down during this season.

**Modelling**

**Evaluation and Deployment**

**References**

**Appendix**