MGT 6203 Group Project Proposal Template

TEAM INFORMATION (1 point)

Team #: 036. **Team Members:** Tania Isabelle De Pena, Alejandro Jose De Jesus, Leslie Francis, Vaisag Radhakrishnan

1. Tania: I'm Tania De Pena;(GT ID: tpen7). I have a bachelor's in industrial Engineering. I'm currently working as a Business Intelligence Manager at Banco Popular Dominicano. Over the past 5 years I have work on the financial sector on different roles such as Business Intelligent analyst and Data Scientist.

2. Alejandro: my name is Alejandro De Jesus(GT ID: ajesus6). My bachelor's degree is in Electronics and Communications engineering. I have experience in web development and data analytics, and I am currently working as a data scientist at Qik Banco Digital Dominicano. Throughout the years, I have worked on projects related to data quality, data extraction and processing pipelines, data visualization for upper management and credit scoring model building, monitoring and adaptation.

3. Leslie: My name is Leslie Francis(GT ID: lfrancis35). I have a bachelor's and master's degree in computer science. Most of my experience has been in data engineering and some years in software engineering. I currently work as a BI Engineer for Amazon.

4. Vaisag: I am Vaisag Radhakrishnan(GTID : vradhakr8). I am a SQL developer. I did my undergraduate in Electronics. Over the past 9 years, I have worked in IT. As part of my CSE 6242 project, we examined soccer play-by-play data to predict a team's performance.

OBJECTIVE/PROBLEM (5 points)

Project Title: Fraud Detection and Prevention

Companies providing financial services across the world are working hard to stop fraudulent transactions from happening. To identify and stop potential financial loss for both the businesses and their clients, high accuracy fraud detection is essential.

Problem Statement:

In this project, we are using e-commerce transaction data to detect fraudulent transactions. The aim is to find out whether a transaction is fraud or not. Identifying this will help banks and payment providers alert their customers in case of suspicious activities, so they can block these transactions.

Primary Research Question: Can a fraudulent transaction be reliably identified using consumer behavior and profiling data? Some of the supplementary research questions are below

1. What are the key factors that indicate that a transaction may be fraudulent? Such as - the weekday where we have more fraudulent transactions, the hours (Morning, evening, night) where we have more fraudulent transaction, the source that has more fraudulent transactions, the country of origin etc.
2. Is it possible to forecast fraud?
3. Other descriptive analytics questions like the below will also be addressed by the team based on the dataset – are men more likely to commit fraud than women? Are younger people more likely to commit fraud than middle aged or older people?

Business Justification:

Data from the Federal Trade Commission shows that consumers claimed $8.8 billion in losses due to fraud in 2022 alone. This is a 30% increase over the reported losses from 2021, which were in turn a 70% increase from 2020. When not dealt with proactively, fraud can affect companies' bottom lines by pulling resources from the core businesses, damaging brand reputation and customer experience, and squandering profits. In some cases, it can lead to customer churn and damage the fidelity and trust of the customers.

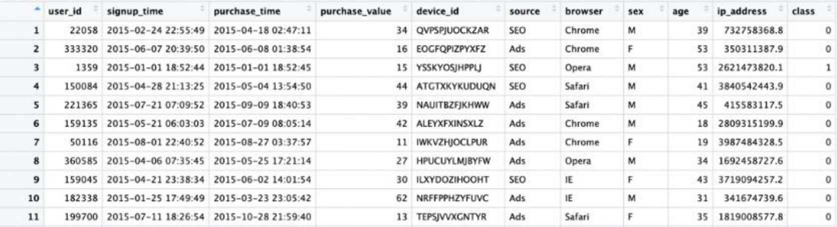
DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.): The data is source from this Kaggle repository.

Data Description:

- **Fraud Data table:** This table contains e-commerce transactions carried out by customers
- **Ip Address to Country table:** This table includes a mapping from ip address to corresponding source country from which the transaction occurred.

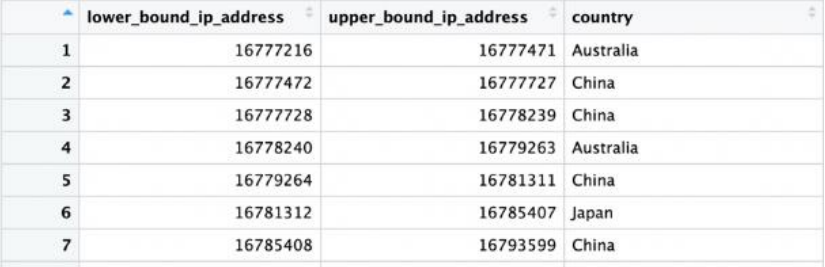
Fraud Data table – Dependent variable: Class – response variable. Class = 1 represents a fraudulent transaction. Independent variables: Signup time and Purchase time – This shows the time and date when user signed up and completed the transactions, Purchase value – the amount associated with the transaction, Source – source from which the user reached the website for carrying out the purchase, Sex – Male/Female, Age of the customer, etc. See Figure 1 for a screenshot of the first few rows including a few features. The hypothesis is that the variables source, time between a signup and purchase, and age play an important role in predicting the fraud. The team is planning on creating additional indicator variables to represent young, middle, and old age groups, high and low value transactions etc.



	user_id	signup_time	purchase_time	purchase_value	device_id	source	browser	sex	age	ip_address	class
1	22058	2015-02-24 22:55:49	2015-04-18 02:47:11	34	QVPSPJUOCKZAR	SEO	Chrome	M	39	732758368.8	0
2	333320	2015-06-07 20:39:50	2015-06-08 01:38:54	16	EOGFQPIZPYXFZ	Ads	Chrome	F	53	350311387.9	0
3	1359	2015-01-01 18:52:44	2015-01-01 18:52:45	15	YSSKYOSJHPPLJ	SEO	Opera	M	53	2621473820.1	1
4	150084	2015-04-28 21:13:25	2015-05-04 13:54:50	44	ATGTXKYKUDUQN	SEO	Safari	M	41	3840542443.9	0
5	221365	2015-07-21 07:09:52	2015-09-09 18:40:53	39	NAUITBZFJKHWW	Ads	Safari	M	45	415583117.5	0
6	159135	2015-05-21 06:03:03	2015-07-09 08:05:14	42	ALEYXFXINSXLZ	Ads	Chrome	M	18	2809315199.9	0
7	50116	2015-08-01 22:40:52	2015-08-27 03:37:57	11	IWKVZHJOCLPUR	Ads	Chrome	F	19	3987484328.5	0
8	360585	2015-04-06 07:35:45	2015-05-25 17:21:14	27	HPUCUYLMJBYFW	Ads	Opera	M	34	1692458727.6	0
9	159045	2015-04-21 23:38:34	2015-06-02 14:01:54	30	ILXYDOZIHOOHT	SEO	IE	F	43	3719094257.2	0
10	182338	2015-01-25 17:49:49	2015-03-23 23:05:42	62	NRFFPPHZYFUVC	Ads	IE	M	31	341674739.6	0
11	199700	2015-07-11 18:26:54	2015-10-28 21:59:40	13	TEPSJVVXGNTYR	Ads	Safari	F	35	1819008577.8	0

Figure 1: First few rows of the fraud data table

Ip Address to Country table – lower_bound_ip_address and upper_bound_ip_address– Range of ip addresses for the associated country. Country – country name in text format. Figure 2 shows the first few rows and columns of this table.



	lower_bound_ip_address	upper_bound_ip_address	country
1	16777216	16777471	Australia
2	16777472	16777727	China
3	16777728	16778239	China
4	16778240	16779263	Australia
5	16779264	16781311	China
6	16781312	16785407	Japan
7	16785408	16793599	China

Figure 2: First few rows of the ip address to country table

APPROACH/METHODOLOGY (8 points)

Before we start modeling, we need to perform exploratory data analysis to understand our data. Following the EDA, we will split the given training data into training, validation, and test data. This problem calls for a classification model to say whether the transaction is fraudulent or not (2 categories). We will try several models such as KNN, Random Forest, Logistic regression etc and decide on a model with the best lift. We must take into consideration that response variable is not balanced. Some models have parameters to account for this. We could also try oversampling or under sampling to reduce this imbalance. To compare models, we can use cross validation. Also, we would need to choose a performance metric to compare them. Once the model is chosen, we will use some cross validation technique like Monte Carlo cross validation to tune the hyperparameters.

Anticipated Conclusions/Hypothesis: The hypothesis is that young people are more likely to commit fraudulent transactions as they are more technologically savvy compared to older people. Additionally, as mentioned earlier, we will look at whether the source and how long the customer was a member of the store before the purchase had influence in the transaction being fraud.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

Early fraud detection not only assists companies in minimizing losses, but also safeguards clients from fraudsters. This study aims to identify frauds by analyzing the transaction's characteristics. With machine learning, a 100% accurate detection is not always feasible. Additionally, delaying a transaction can make customers unhappy because their purchase was unsuccessful. Due to this, if a suspicious transaction is successfully detected, one alternative will be to get in touch with the customer to confirm the transaction. If the transaction was incorrectly identified by the system, this will help the bank validate it directly with the customer.

With numerous options available online for purchase, customer satisfaction is of paramount importance for businesses these days. Protecting a customer from fraudsters and warning them about it will help build that relationship and trustworthiness between the business and the customer.

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

Phase	12-Mar	19-Mar	26-Mar	2-Apr	9-Apr	16-Apr
Exploratory Data Analysis	🟧	🟦				
Data Wrangling and Cleaning		🟦	🟧			
Modeling			🟦	🟧		
Feature Engineering and Tweaking the Model			🟦	🟦	🟦	
Reporting					🟦	🟧

Table 1: Shows the tentative project plan. All dates are End of week Dates. Orange represents deadlines. Details in Table 2.

Deadline 1	Project Proposal Document Due
Deadline 2	Plan Presentation Video Due
Deadline 3	Progress Report Due
Deadline 4	Final Report and Code Due

Table 2: Deadlines for the project

Appendix (any preliminary figures or charts that you would like to include):