**SUMMARIZED PROJECT MEETING MINUTES**

**(From Week 1 to Week 13)**

**GROUP 3 – DAB422-24W-001**

**Project Name: Crime Rate Analysis in Toronto, Canada**

**Main Attendees (Members of Group #3)**

1. Yen Nga Le

2. Srilakshmi Gummadidala

3. Tehsin Shaikh

4. Vinod Soloman Santhakumar

5. Prof. Abiodun Sodiq Shofoluwe

**Specific Activities and Outputs from prior weeks:**

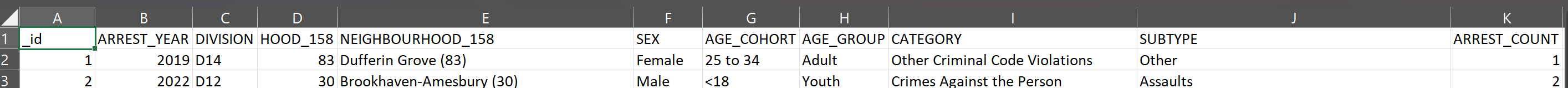
1. In **Week 1**, based on the result of Analysis on Capstone Project 1 with the topic as ***Crime Rate Analysis in Toronto, Canada***, all team members were introduced during the initial meeting (January 9, 2024) with the Professor. Sodiq. We provided a brief project overview, highlighted actions taken in the previous semester as well as outlined upcoming steps.
2. Carrying on **Week 2**, which commenced on January 16, 2024, our team engaged in an in-depth discussion with Prof. Sodiq concerning the project dataset. The focus was on identifying key variables crucial for gaining insights into crime and refining the project's scope. Additionally, we revisited dashboard graphs from the previous semester to provide context for our discussion. Prof. Sodiq offered valuable suggestions for our next steps, including the incorporation of factors affecting crime rates into the main dataset, particularly major crime indicators, to further explore demographic trends. Moreover, Prof. Sodiq provided a comprehensive overview of potential machine learning models suitable for implementation in subsequent phases, emphasizing the utilization of time series machine learning models. During this period, we established a GitHub profile for the Capstone Project and uploaded all project-related documents for both Capstone 1 and Capstone 2 to the repository.
3. Next, in **Week 3 & 4**, several relevant raw data sources which provide information of several essential factors affecting the main Crime Rate dataset are recognized via the below link:

<https://open.toronto.ca/catalogue/?owner_division=Toronto%20Police%20Services>

After thorough checking and preprocessing of datasets, a significant challenge arose due to irrelevant demographic characteristics in the original dataset during data collection and preprocessing steps. Additionally, some datasets lacked primary keys for merging into a single dataset for analysis. Consequently, the decision was made to focus solely on downloading one file named "Arrested and Charged Persons.csv" to gain insights into demographic factors.

Attempts were made to link this file with the main dataset using Power BI, SQL, or Excel. Team members initially tried to merge the datasets using Python programming, but encountered Many-to-Many combinations during data preprocessing. Following suggestions from Prof. Sodiq, the team explored merging the datasets using One-to-Many combinations through various tools such as Power BI, SQL, and Python. However, this approach proved to be less efficient than anticipated. Consequently, the decision was made to discontinue the approach.

[Police Annual Statistical Report - Arrested and Charged Persons - City of Toronto Open Data Portal](https://open.toronto.ca/dataset/police-annual-statistical-report-arrested-and-charged-persons/)



1. Moving on **Week 5, 6, & 7**, each team member continued working individually on the project's next steps, focusing on feature engineering and modeling methods About feature engineering steps, we were trying to identify which features are important for modeling after encoding all the non-numerical features to numerical and then attempting to evaluate which are best for modeling.

About modeling methods for predicting the trend, we tried applying for Time series forecasting methods as ARIMA and Clustering methods. Unfortunately, with ARIMA, the forecasting accuracy of these model are still low. The same low results for Logistic Regression accuracy and Random Forest Classifier accuracy. With Clustering model, the coding still gets error to run in python.

During this period, each team member reviewed and prepared their individual contributions to the capstone project, which were discussed during the mid-term interview with Prof. Sodiq on February 13, 2024.

1. During **Week 8 & 9**, the team continued investigating and deploying feature engineering and modeling methods, aiming to identify the most relevant codes and models for predicting trends in the dataset. Each team member constructed various models and extensively evaluated their performance to determine the most effective ones for analysis. Rigorous testing and comparison were conducted to ensure the reliability and effectiveness of the modeling approach. The team also gained an understanding of overfitting and implemented strategies to address it, such as selecting a set of important features to refine the modeling approach and ensure accurate predictions on unseen data.

The Interim Report was prepared and submitted by the Week 8 deadline, with contributions from all team members. It provided an in-depth analysis of Toronto's crime rates using data science techniques, detailing the entire process from problem definition and stakeholder identification to data collection, cleaning, exploratory data analysis (EDA), and data modeling. Machine learning algorithms like Random Forest Classifier, Logistic Regression, Neural Networks, and Time Series Analysis were employed for predicting future crime rates, highlighting their significance in revealing patterns and informing strategies for crime prevention and safety improvement.

During a meeting, the professor reviewed the team's feature importance assessments and modeling approaches. It was suggested to reconstruct the model using a select set of important features and to scale the feature and target variables to address potential generalization issues in the current models. Additionally, the team explored different deployment methods to test and implement the most suitable one for the project once the final model is selected.

1. After the March Break vacation, during **Week 11 & 12**, the team reviewed and implemented changes based on the professor's recommendations to refine the modeling approach, focusing on addressing overfitting and enhancing model performance. This involved adopting a different feature selection approach, selecting a subset of the top 10 features deemed most relevant for analysis. Lazy Predict was utilized to expedite model evaluation, while scaling techniques tailored to the LSTM model were implemented to mitigate overfitting issues.

The interim presentation was skillfully prepared and presented by Srilakshmi Gummadidala and Tehsin Shaikh due to the absence of two team members for personal reasons. We provided a comprehensive overview of the project, covering its inception, exploratory data analysis, feature importance identification, and modeling efforts. Our clear and concise communication effectively conveyed the progress and insights gleaned from our analysis.

Through rigorous testing and comparison, the team arrived at the consensus that the most suitable final model comprises Neural Networks and time series models for temporal analysis. This conclusion was drawn based on empirical evidence suggesting that other classification models were prone to overfitting and therefore deemed less reliable for the task at hand. As part of the ongoing efforts to enhance the analysis and interpretate the model better, a visualization graph illustrating the Neural Network model's performance was created.

Additionally, during our discussions, the professor suggested exploring various deployment options suitable for our project's needs. We were actively working on investigating these suggestions to identify the most appropriate deployment techniques tailored to our machine learning models. These suggestions would help us ensure the seamless integration and efficient operation of our models in practical applications.

1. In **Week 13**, we have already discussed about our Flask API deployment strategy and worked with the professor to resolve deployment errors. And then, we reviewed our approach, focused on improving stability, and discussed error handling best practices. This session helped us optimize our deployment workflow and better equip ourselves for challenges.

**Challenges/Disagreements:**

* Time is quite rushed to complete fully the project.
* After mid-semester, we had to face with the shortage of team members to conduct or run the code of project with a little bit stress. In details, Vinod Soloman Santhakumar, has been absent since the Mid Term Interview (Week 8) due to his personal health matters and Yen Nga Le had to undertake an emergency travel during 1 week (Week 12) to her home country, Vietnam.

**Planned Activities for coming weeks:**

* Continuing to Identify and address deployment errors and Optimize deployment workflow for stability
* Following a successful deployment, shift focus towards preparing comprehensive documentation, including both a presentation and a detailed report.

**Key Resources and Repository:**

* Main dataset link – Crime rate:

[Major Crime Indicators Open Data | Major Crime Indicators Open Data | Toronto Police Service Public Safety Data Portal](https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-open-data/explore)

* Additional dataset link - Demographic:

[Police Annual Statistical Report - Arrested and Charged Persons - City of Toronto Open Data Portal](https://open.toronto.ca/dataset/police-annual-statistical-report-arrested-and-charged-persons/)

* GitHub link:

<https://github.com/VinodSolomon/Crime-Rate-Analysis-Toronto>