# **A screenshot of a video game Description automatically generated**

# Group Project 2

### **Submitted by: (Group 3)**

### Gayathri Srikanti: 301181090

### Parth Sharma : 301250690

### Marco Paulo Garcia: 301236921

### Mahsa Dorri: 301257460

### Fawad Khan:

### Submitted to: Professor Viji

### ICET Department,

### School of Engineering Technology and Applied Science

### Progress Campus, Block A

### Centennial College

### 

### Discipline:   Data Warehousing and Data Mining!

### 

### Due Date:    December 13, 2023

## Part 1: Executive summary

This project delves into the analysis of Bicycle Theft data, culminating in the development and deployment of two predictive models: the logistic regression and decision tree models. Equipped with foundational knowledge from class lectures, lab works, and course notes, the team navigated through multifaceted challenges to achieve success.

The project commenced with comprehensive data exploration, laying the groundwork for subsequent stages. However, challenges arose during the data modeling phase, particularly due to the team's lack of prior experience in data warehousing. Understanding the purpose of data modeling proved to be pivotal, and it was Lab 10 that provided the necessary insight into its relevance in constructing the predictive model.

The team faced initial hurdles comprehending the link between data modeling and the predictive model's construction. With the invaluable guidance of Professor Viji Angamuthu and newfound knowledge gained from Lab 10, the team navigated through these challenges, aligning their efforts with the project's objectives.

Despite encountering obstacles, the project culminated in the successful deployment of two predictive models. Through perseverance, utilization of available resources, and guidance from the professor, the team overcame their initial struggles to deliver a robust solution.

## Part 2: Overview of the Machine Learning Solution for Predicting Bicycle Theft Outcomes

### Introduction

This project represents a comprehensive effort to tackle the issue of bicycle thefts using machine learning. Collaboratively undertaken by a team of dedicated students, the project's aim was to build a predictive model that could accurately forecast the likelihood of a bicycle being returned if stolen. This tool was intended to serve both the Toronto Police Department and the general public, aiding in preemptive measures against bicycle thefts and enhancing community safety.

### Data Preparation and Exploration

**Dataset Utilization:** The project utilized a detailed dataset provided by the Toronto Police Department, encompassing various attributes related to bicycle theft incidents over four years.

**Data Cleaning and Analysis:** Using Python libraries like Pandas and Numpy, the team performed extensive data cleaning, transformation, and analysis. Key tasks included handling missing data, managing categorical data, and standardizing data for modeling.

**Visualization:** Power BI was employed to create insightful visualizations. This helped in identifying trends, patterns, and correlations within the data, laying a foundation for informed feature selection and model building.

### Model Development and Evaluation

**Algorithm Selection:** The team chose Logistic Regression and Decision Trees as primary algorithms, considering their efficacy in classification problems.

**Feature Engineering:** Based on the data exploration phase, important features influencing the likelihood of bicycle thefts were identified and used for model training.

**Model Training and Testing:** Models were trained and tested on the dataset, with a keen focus on handling class imbalances and ensuring robustness.

**Evaluation Metrics:** Models were evaluated using accuracy, confusion matrices, and ROC curves to determine their effectiveness in making predictions.

### Deployment and Accessibility

**API Development:** The best-performing model was turned into an API using the Flask framework. This involved serialization for model persistence and creating endpoints for real-time predictions.

**Front-End Interface:** A simple but functional front end was developed, enabling users to input features and receive predictions. The interface was designed to be user-friendly, ensuring ease of use for both the police department and the public.

### Conclusion

The project successfully culminated in a practical and accessible tool that leverages data analytics for social good. The predictive model, backed by rigorous data analysis and machine learning techniques, stands as a testament to the potential of technology in enhancing urban safety and community well-being.

## Part 3: Data exploration

1. Load and describe data elements (columns), provide descriptions & types, ranges, and values of elements as appropriate - use pandas, numpy and any other Python packages.

This Data Frame has 15 features(columns) also 34290 records.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Columns name: ‘X', 'Y', 'OBJECTID', 'EVENT\_UNIQUE\_ID', 'PRIMARY\_OFFENCE', 'OCC\_DATE', 'OCC\_YEAR', 'OCC\_MONTH', 'OCC\_DOW', 'OCC\_DAY', 'OCC\_DOY', 'OCC\_HOUR', 'REPORT\_DATE', 'REPORT\_YEAR', 'REPORT\_MONTH', REPORT\_DOW', 'REPORT\_DAY', 'REPORT\_DOY', 'REPORT\_HOUR', 'DIVISION', 'LOCATION\_TYPE', 'PREMISES\_TYPE', 'BIKE\_MAKE', 'BIKE\_MODEL', 'BIKE\_TYPE', 'BIKE\_SPEED', 'BIKE\_COLOUR', 'BIKE\_COST', 'STATUS', 'HOOD\_158', 'NEIGHBOURHOOD\_158', 'HOOD\_140', 'NEIGHBOURHOOD\_140', 'LONG\_WGS84', 'LAT\_WGS84’

1. Statistical assessments include means, averages, and correlations.

We use medal\_data\_group3.describe() to get information.

A screenshot of a computer

Description automatically generated

1. Missing data evaluations – use pandas, NumPy, and any other Python packages.

fill any null values with 0, ‘MISSING’, and median.

A screenshot of a computer program

Description automatically generated

1. Graphs and visualizations – use pandas, matplotlib, seaborn, numpy, and any other Python packages. You can also use a power BI desktop.

We try to get histograms that give us some information which you can find after the code.

A computer screen shot of a computer code

Description automatically generated

* This histograms show in which years we had the highest rate of incidents.

A graph with numbers and a line

Description automatically generated

A graph of a bicycle theft

Description automatically generated

#### *In this histogram, we get that in 160/365 until 260/365 we have the most stolen bike. This means in summer and fall.*

#### *A graph with blue lines Description automatically generated*

#### *The other data that we need to know is in which season and month we have more incidents. (Seasonality Effect Diagram) these diagrams show in fall and summer stolen bike rate increased*

#### *A graph with colorful lines Description automatically generated*

#### *A blue line on a white background Description automatically generated*

* Also in the Pandemic year, we have the highest rate of incidents following diagram shows this.

A graph with blue lines

Description automatically generated

* First of all we can see some out liars in the cost diagrams that need to handle

A diagram of a scatter plot

Description automatically generated

* Here we can see the barchart of the same features

A graph of a bicycle cost

Description automatically generated

* This will be divided into two bar charts that cost less than 10000 which the lions share of the data frame

A graph of a bicycle cost

Description automatically generated

* and more than 10000 which hast the lowest numbers of incident so we need to consider some of them as a outliar

A graph of a bar graph

Description automatically generated

* After handling the outlier our bike cost is

A blue rectangular object with white lines

Description automatically generated with medium confidence

* In this diagrams which is the result of the codes we can see the numbers of stolen bike based on location and location and bike cost

A screenshot of a computer program

Description automatically generated

* This shows the apartment and the shelter are two high rates of stolen bikes in Toronto.

A graph of a number of blue and black bars

Description automatically generated with medium confidence

* This diagram shows based on location and cost the dealerships and the homeless shelter.

A graph showing a number of bars

Description automatically generated with medium confidence

* We can get the correlation matrix

A screenshot of a graph

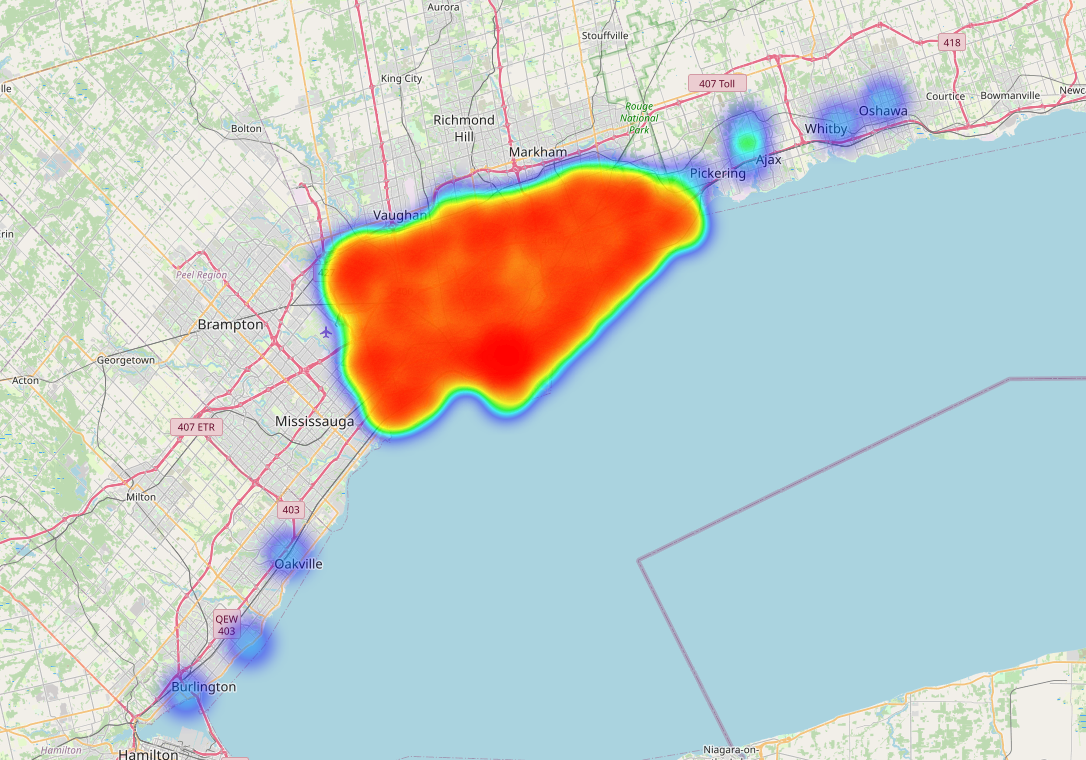
Description automatically generated

* **Geospatial Visualization of Bike Thefts in Toronto:** we use Geopandas and Shapely to create a geospatial visualization of bike thefts in Toronto. We convert our DataFrame into a GeoDataFrame using the longitude and latitude columns, setting the coordinate reference system to EPSG 4326. The map is centered on Canada, with the locations of bike thefts in Toronto plotted as red points.

A map of canada with red dot

Description automatically generated

* **Heatmap of Bike Thefts in Toronto:** We create a heatmap to visualize the concentration of bike thefts in Toronto. We use Folium to create an interactive map, with the heatmap layer representing the density of theft locations. The map is centered around Toronto with an appropriate zoom level, allowing for a detailed view of theft hotspots.



## Part 4: Feature Selection in Predictive Modeling of Bicycle Theft Outcomes

### Tools and Techniques Used

1. **Python Libraries:** The primary tools for feature selection were Python libraries like Pandas, NumPy, and Scikit-learn. These libraries provided functionalities for data manipulation, statistical analysis, and feature importance evaluation.
2. **Visualization Tools:** Power BI was used for visual data exploration, helping to identify potential features based on patterns and trends in the data.
3. **Feature Importance Analysis:** The feature\_importance.py script likely used techniques such as decision tree-based feature importance or model coefficient analysis (in the case of logistic regression) to identify the most relevant features.
4. **Correlation Analysis:** Statistical methods to analyze correlations between different variables were used to understand the relationships and potential collinearity among features.

### Feature Selection Process

* **Initial Data Review:** The process began with an initial examination of the dataset to understand the nature of each variable.
* **Data Cleaning and Transformation:** Before selecting features, the data was cleaned, and categorical variables were transformed appropriately to ensure they could be effectively used in the modeling process.
* **Exploratory Data Analysis (EDA):** This involved visually and statistically examining the data to identify potential predictors of bicycle theft outcomes.
* **Iterative Testing:** Different combinations of features were tested in the models to assess their impact on model performance. This iterative approach helped in refining the feature set.

### Results of Different Combinations

* **Decision Tree Model:** The feature importance extracted from the decision tree model provided insights into which variables most strongly influenced the predictions. Features like location, time of theft, and bike characteristics (like color, type, cost) might have been identified as significant.
* **Logistic Regression Model:** Coefficients from the logistic regression model offered a different perspective on feature relevance. This might have highlighted different sets of features or confirmed the importance of certain variables identified by the decision tree model.
* **Optimized Feature Set:** The final set of features chosen for the models was likely a result of balancing predictive power and model simplicity. This set would have been the one that provided the best performance metrics during model evaluation.

### Conclusion

The careful selection of features was a key aspect of building effective predictive models for bicycle theft outcomes. The use of various tools and techniques ensured a thorough evaluation of potential predictors, leading to the identification of an optimized set of features. This not only improved model accuracy but also provided deeper insights into the factors influencing bicycle thefts.

A screenshot of a computer program

Description automatically generated

A graph with blue and white bars

Description automatically generated

Based on the chart, PRIMARY\_OFFENCE seems to be the most important feature, with a score substantially higher than the others. This suggests that the primary offense is highly predictive of the outcome the model is trying to predict.

The next set of important features includes BIKE\_SPEED, BIKE\_COST, and BIKE\_MAKE, which might be factors related to the likelihood of a bike being stolen or recovered if that is the target variable of the model.

On the other end, features like OCC\_MONTH, REPORT\_MONTH, DIVISION, and NEIGHBOURHOOD identifiers have lower importance scores. While they do contribute to the model, their impact is smaller compared to the top features.

## Part 5: Data Modeling in Bicycle Theft Prediction Project.

### Data Cleaning Strategy

1. **Handling Missing Values:** The approach included:

* Dropping irrelevant columns such as 'X', 'Y', 'OBJECTID', 'EVENT\_UNIQUE\_ID', and others.
* Removing rows where the 'STATUS' column was 'UNKNOWN'.
* Imputing missing values in 'BIKE\_MAKE', 'BIKE\_SPEED', 'BIKE\_COST', and 'BIKE\_COLOUR' using methods like median, mean, or mode.

A screen shot of a computer code

Description automatically generated

1. Data Type Conversions: The project involved converting categorical variables to numeric types through label encoding and normalizing numerical columns with Min-Max scaling.
2. Normalization and Handling Class Imbalance

* Data Normalization: Applied to ensure a uniform scale across both categorical and numerical features.
* Class Imbalance with SMOTE: Employed Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset, enhancing the model's ability to generalize.

A screen shot of a computer program

Description automatically generated

1. Feature Selection

* Recursive Feature Elimination (RFE): This method, paired with a Logistic Regression estimator, was used to identify the top 5 most relevant features for the model.

A computer screen shot of a program

Description automatically generated

### Results of Data Cleaning

* **Enhanced Data Quality:** Removal of irrelevant features and imputation of missing values led to a more accurate and reliable dataset. Uniform data representation achieved through consistent data type conversions.
* **Dataset Size Reduction:** Streamlined dataset by excluding specific columns and rows, particularly those with 'UNKNOWN' status. Refined dataset focus, enhancing relevance for the logistic regression model.

### Assumptions and Constraints

#### Assumptions:

* **Data Completeness:** It was assumed that the dataset provided a comprehensive view of bicycle theft incidents over the specified period.
* **Predictive** **Relevance:** There was an underlying assumption that historical data and patterns would be predictive of future incidents.

#### Constraints:

* **Limited Historical Data:** The dataset's coverage was confined to a specific period, potentially affecting its applicability to other time frames.
* **Data Privacy and Security:** Maintaining the anonymity and security of potentially sensitive information within the dataset posed a significant constraint.
* **Computational Resources**: The extent of data processing and model complexity might have been influenced by limitations in computational resources.

### Conclusion

The data modeling phase proved pivotal in the successful execution of our project. Meticulous cleaning and transformation of the dataset ensured accuracy and relevancy, forming a solid foundation for the development of effective predictive models. This process, augmented by a strategic consideration of underlying assumptions and operational constraints, set the stage for insightful analysis and robust model building. The outcome was a harmonious blend of data integrity and analytical precision, essential for the predictive modeling of bicycle theft outcomes.

## Part 6: Model Building in Bicycle Theft Prediction Project

1. Balancing classes and splitting data for training.

**A screen shot of a computer code

Description automatically generated**

1. The ROC curve demonstrates the model's good predictive ability, significantly better than random, with an AUC of 0.75.

**A graph of a curve

Description automatically generated**

1. The accuracy of the model is 0.69. Class 0 has a precision of 0.68 and recall of 0.72, while class 1 has a precision of 0.71 and recall of 0.66. The confusion matrix indicates true positives, false positives, true negatives, and false negatives for the predictions.

**A screenshot of a computer

Description automatically generated**

1. The logistic regression model is then trained and evaluated, yielding an accuracy of 69%, with precision, recall, and f1-score close to 0.7 for both classes. The confusion matrix and ROC curve are generated, showing the model's performance, with an AUC of 0.75, indicating a good predictive ability.

**A black and white card with a hexagon and white text

Description automatically generated**