**Group Project – Developing a predictive machine learning model (classifier) and deploying it as API**

**Due Date:** Group presentation in Week 12.

1) **Purpose:** The purpose of this project is to:

1. Get & Prepare the data:
2. Load and explore the dataset referenced in section 4 in this document using techniques learnt during this course.
3. Visualize the data and describe it thoroughly, identify correlations..etc.
4. Clean, transform categorical data and model the dataset using the techniques learnt throughout the course in preparation for building a predictive model.
5. Build a supervised predictive model based using a suitable classification algorithms in python and utilizing scikit-learn, pandas, numpy…etc. To provide predictions as specified in project specification, section 3 of this report.
6. Validate / score and evaluate the models and choose the best model.
7. Build an API for the model using Python Flask framework.
8. Build a simple front end to access the API and pass new feature values to the prediction model.

2) **Guidelines & Instructions:** Be sure to read the following general instructions carefully:

* General:
* This Project is to be completed in groups of 5 students.
* Read the textbook, lecture slides, class examples, and additional references provided here. You are free to research and use more materials and tools to implement a good solution, just make sure to reference it in your solution and your report.
* Presentation:
* Each group will have to present and demonstrate their solution in Week 12 or as agreed with their professor.
* During the presentation, each group needs to demonstrate the solution i.e. live code execution and maybe as needed illustrate part of the code. In addition, the group needs to present all assumptions, constraints, key findings in addition to the list of technologies used clearly.
* Submission:
* All code developed in python and any other language should be part of your submission.
* The should submission should be accompanied with either a report prepared as a Microsoft document or a power point presentation explaining the project and detailing all the assumptions, constraints applied. (Details in section 3).
* Name the project submission: “Bycycle\_theft\_Group\_Group#\_section\_section#COMP309Project” where Group# is your assigned group# and section# is your section number.
* The submission should be a zipped file containing the code and the written report.

**3) Project Specifications & deliverables**

Both the police department and the “general public” would make use of a software product that can give them an idea about the likelihood of bicycle theft. For the police department it would assist them in taking better measures of anti-theft around certain neighborhoods. For the public individuals, it would help them assess the need for additional precautions such as locks.

Based on the dataset described in section four of this document, which is actual data collected over the period of four years by the Toronto police department. You need to build a predictive service that based on certain features would provide a classification of either the bike is likely to be returned or not.

Please arrange to provide the following deliverables for your project.

1. Data exploration: a complete review and analysis of the dataset including:
   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.
   2. Statistical assessments including means, averages, correlations
   3. Missing data evaluations – use pandas, NumPy and any other python packages
   4. Graphs and visualizations – use pandas, matplotlib, seaborn, NumPy and any other python packages, you also can use power BI desktop.
2. Data modelling:
   1. Data transformations – includes handling missing data, categorical data management, data normalization and standardizations as needed.
   2. Feature selection – use pandas and sci-kit learn.
   3. Train, Test data splitting – use NumPy, sci-kit learn.
   4. Managing imbalanced classes if needed. Check here for info: https://elitedatascience.com/imbalanced-classes
3. Predictive model building
   1. Use logistic regression and decision trees as a minimum– use scikit learn
4. Model scoring and evaluation
   1. Present results as scores, confusion matrices and ROC - use sci-kit learn
   2. Select and recommend the best performing model
5. Deploying the model
   1. Using flask framework arrange to turn your selected machine-learning model into an API.
   2. Using pickle module arrange for Serialization & Deserialization of your model.
   3. Build a client to test your model API service. Use the test data, which was not previously used to train the module. You can use simple Jinja HTML templates with or without Java script, REACT or any other technology but at minimum use POSTMAN Client API.
6. Prepare a report explaining your project and detailing all the assumptions, constraints you applied should have the following sections:
   * Executive summary (to be written once nearing the end of project work, should describe the problem/solution and key findings)
   * Overview of your solution(to be written once nearing the end of project work)
   * Data exploration and findings (dataset field descriptions, graphs, visualizations, tools and libraries used….etc.)
   * Feature selection (tools and techniques used, results of different combinations…etc.)
   * Data modeling (data cleaning strategy, results of data cleaning, data wrangling techniques, assumptions and constraints)
   * Model building (train/ test data, sampling, algorithms tested, results: confusion matrixes ...etc.)

**4) Data Set**

This dataset contains actual Bicycle Thefts occurrences from 2014-2018 in the city of Toronto.

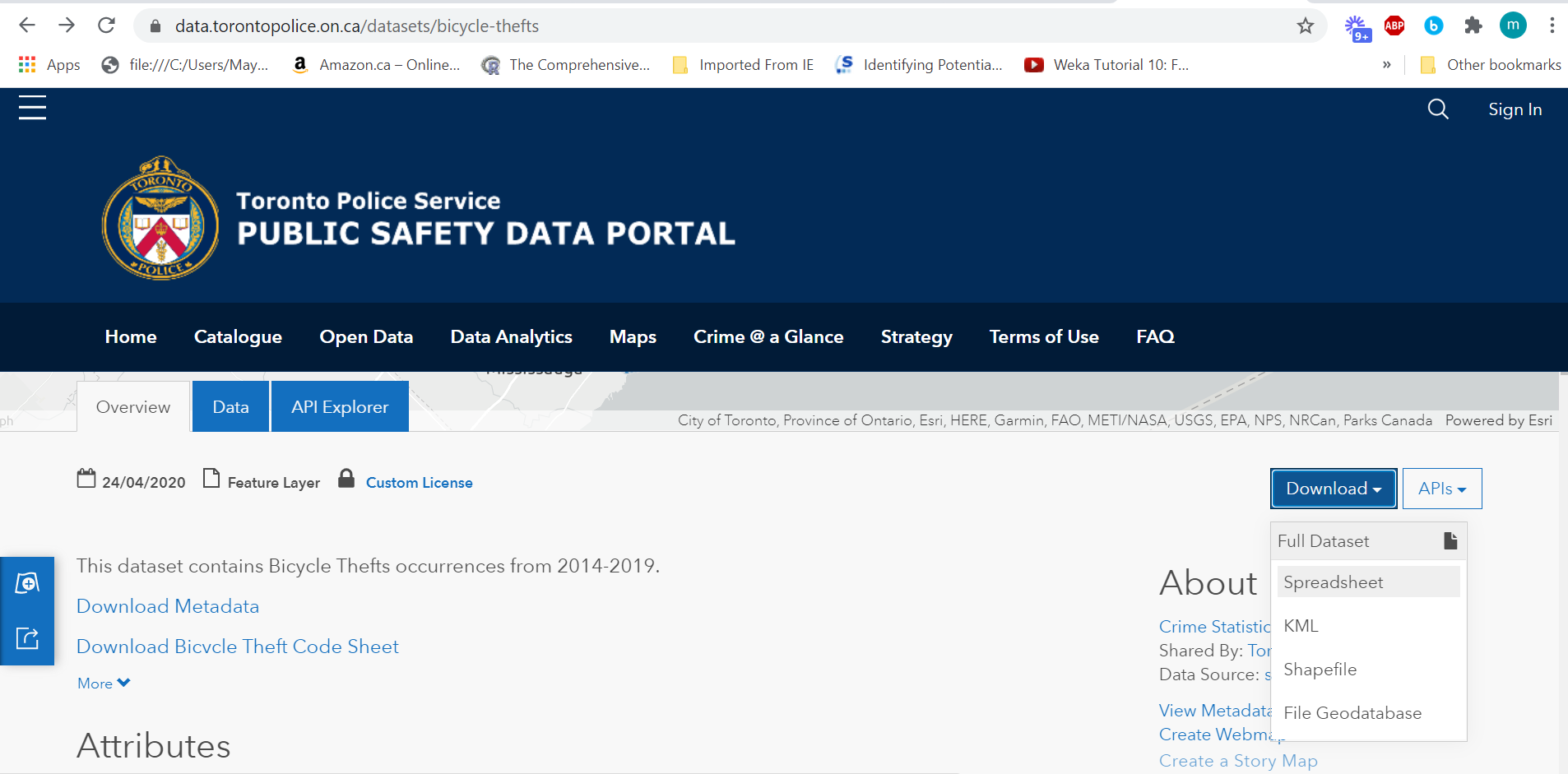
In accordance with the Municipal Freedom of Information and Protection of Privacy Act, the Toronto Police Service has taken the necessary measures to protect the privacy of individuals involved in the reported occurrences. No personal information related to any of the parties involved in the occurrence will be released as open data.

The location of crime occurrences have been deliberately offset to the nearest road intersection node to protect the privacy of parties involved in the occurrence. All location data must be considered as an approximate location of the occurrence and users are advised not to interpret any of these locations as related to a specific address or individual.

The reported crime dataset is intended to provide communities with information regarding public safety and awareness. The data supplied to the Toronto Police Service by the reporting parties is preliminary and may not have been fully verified.

<https://data.torontopolice.on.ca/datasets/bicycle-thefts>

Use the download tab and select spreadsheet to download the dataset as a csv file, also download the Metadata file.



**5) Evaluation of software solution:**

|  |  |
| --- | --- |
| **Evaluation Component** | **Percentage** |
| Data exploration | 10% |
| Data modelling | 15% |
| Predictive model building | 15% |
| Model scoring and evaluation | 10% |
| Model deployment | 25% |
| Report | 10% |
| Presentation | 10% |
| Innovation & creativity | 5% |
| **Total** | **100%** |

**6) References:**

<https://scikit-learn.org/stable/>

<https://pandas.pydata.org/>

<https://numpy.org/>

<http://flask.pocoo.org/>

<https://www.getpostman.com/automated-testing>

<https://www.datacamp.com/community/tutorials/machine-learning-models-api-python>

<https://www.analyticsvidhya.com/blog/2017/09/machine-learning-models-as-apis-using-flask/>

<http://data.torontopolice.on.ca/datasets/16f2b8a1c76547c69fec14b7f8541ffc_0>

https://elitedatascience.com/imbalanced-classes