**Section 1: Purpose**

You will be developing a predictive machine learning model (classifier) and deploying it as API.

The purpose of this project is to:

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

**Section 2: Guidelines and Instructions**

**Requirements**

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 neighbourhoods. 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 based on certain features that 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, and 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 can also use power BI desktop.

1. **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. For more information open the following [link](https://elitedatascience.com/imbalanced-classes).
2. **Predictive model building**
   1. Use logistic regression and decision trees as a minimum– use scikit learn.
3. **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.
4. **Deploying the model**
   1. Using a flask framework arrange to turn your selected machine-learning model into an API.
   2. Using the 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.
5. **Prepare a report**

Explain your project and detail all the assumptions and constraints you applied in the following sections:

* 1. Executive summary (to be written once nearing the end of the project work, should describe the problem/solution and key findings);
  2. Overview of your solution (to be written once nearing the end of the project work);
  3. Data exploration and findings (dataset field descriptions, graphs, visualizations, tools, and libraries used, etc.);
  4. Feature selection (tools and techniques used, results of different combinations, etc.);
  5. Data modeling (data cleaning strategy, results of data cleaning, data wrangling techniques, assumptions and constraint); and
  6. Model building (train/ test data, sampling, algorithms tested, results: confusion matrixes, etc.).

**Section 3: Data Set**

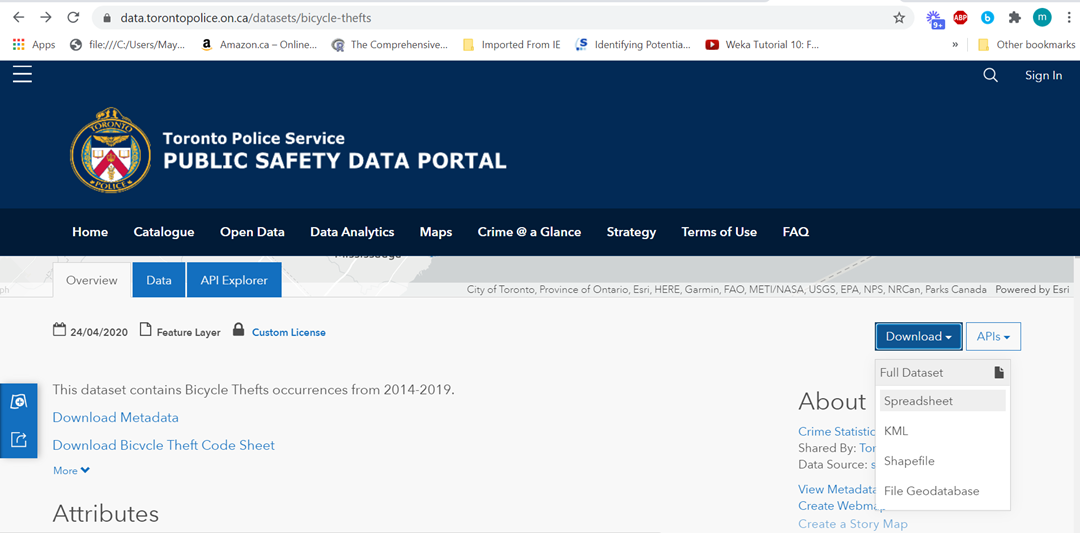
This dataset contains actual bicycle theft occurrences from 2014-2018 in the city of Toronto.

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 has 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.

* [Toronto Police Dataset](https://data.torontopolice.on.ca/datasets/bicycle-thefts)
* <https://data.torontopolice.on.ca/pages/bicycle-thefts>
* Use the download tab and select spreadsheet to download the dataset as a csv file, also download the Metadata file.



**Section 4: Reference Links**

1. [Toronto Police Dataset](https://data.torontopolice.on.ca/datasets/bicycle-thefts)
2. <https://scikit-learn.org/stable/>
3. <https://pandas.pydata.org/>
4. <https://numpy.org/>
5. <http://flask.pocoo.org/>
6. <https://www.getpostman.com/automated-testing>
7. <https://www.datacamp.com/community/tutorials/machine-learning-models-api-python>
8. <https://www.analyticsvidhya.com/blog/2017/09/machine-learning-models-as-apis-using-flask/>
9. <https://elitedatascience.com/imbalanced-classes>