**COMP 247 Project**

**Group Project – Developing a predictive machine learning model (classifier) and deploy it as a web API for inference**

**1) Purpose**

The purpose of this project is to:

1. Pre-processing - Retrieve & prepare the data: Load and explore the dataset referenced in section 4 in this document using techniques learnt during this course.
2. Visualize the data and describe it thoroughly, identify correlations..etc.
3. Clean, transform categorical data and model the dataset using the techniques learnt throughout the course in preparation for building a predictive model.
4. Model building & fine tuningBuild a supervised predictive model based using a suitable classification algorithm(s) in python , utilizing scikit-learn, pandas, numpy…etc. To provide predictions as specified in project specification, section 3 of this report.
5. Validate / score and evaluate the models and choose the best model after carrying out hyper-parameter tuning.
6. Model deployment: Build an API for the model using Python Flask framework.
7. Deploy the model on local host.
8. Build a simple front end to access the API and pass new feature values to the prediction model for inference.

**2) Guidelines & Instructions**

Be sure to read all the guidelines and instructions:

General instructions:

- This Project is to be completed in groups of 5 students.

- Read the textbook, course lecture content, class examples, and additional references provided here. Each team 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 14 or as agreed with their professor.

- During the presentation, each group needs to demonstrate the solution i.e. live code execution and as needed illustrate part of the code. In addition, the group needs to present all key findings, assumptions, constraints, and the list of technologies used clearly.

- Submission:

- Two submissions are required as follows:

   part #1 : Due week #10, this will include deliverables #1 & 2 (Data exploration & Data modeling). worth 13%

   part #2: Due week #14 this will build on part #1 and will include a full report & demonstration of all the deliverables. worth 13%

- All code developed in python and any other language should be part of the submission.

- The submission should be accompanied with a report prepared as a Microsoft document or a pdf explaining the project findings and detailing all the assumptions, constraints applied. (Details in section 3).

- Name the project submission: “KSI\_Group\_Group#\_section\_section#COMP247Project” where Group# is the assigned group# and section# is the groups section number.

- The submission should be a zipped file containing the code and the written report.

Peer Evaluation

- Each team member will complete a peer evaluation form relating to other team members and this will count in the final grade.

- Grades are issued based on contribution to the project work.

**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 fatal collisions that involve loss of life. For the police department it would assist them in taking better measures of security and better planning for road conditions around certain neighborhoods. For the public individuals, it would help them assess the need for additional precautions at certain times and weather conditions and neighborhoods.

Based on the dataset described in point four below, which is actual data collected over the period of five 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 incident would result in fatality or not.**

Please arrange to provide the following deliverables for your project.

1. Data exploration: a complete review and analysis of the dataset including:

* Load and describe data elements (columns), provide descriptions & types, ranges and values of elements as appropriate. – use pandas, numpy and any other python packages.
* Statistical assessments including means, averages, correlations
* Missing data evaluations – use pandas, numpy and any other python packages
* Graphs and visualizations – use pandas, matplotlib, seaborn, numpy and any other python packages, you also can use power BI desktop.

2. Data modelling:

* Data transformations – includes handling missing data, categorical data management, data normalization and standardizations as needed.
* Feature selection – use pandas and sci-kit learn. (The group needs to justify each feature used and any data columns discarded)
* Train, Test data splitting – use numpy, sci-kit learn.
* Managing imbalanced classes if needed. Check here for info: <https://elitedatascience.com/imbalanced-classes>
* Use pipelines class to streamline all the pre-processing transformations.

3. Predictive model building

* Use logistic regression, decision trees, SVM, Random forest and neural networks algorithms as a minimum– use scikit learn
* Fine tune the models using Grid search and randomized grid search.

4. Model scoring and evaluation

* Present results as accuracy , precision, recall, F1 scores, confusion matrices and plot the ROC curves of the models - use sci-kit learn
* Select and recommend the best performing model

5. Deploying the model

* Using flask framework arrange to turn your selected machine-learning model into an analytics API.
* Using pickle module arrange for Serialization & Deserialization of your model.
* Deploy your model on local host.
* 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:

* Table of contents
* 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 includes all traffic collisions events where a person was either Killed or Seriously Injured (KSI) from 2006 – 2020 in the city of Toronto. (might change to 2021 depending on frequency of update)

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 the incident 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 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.

[KSI dataset](https://data.torontopolice.on.ca/search?collection=Dataset&q=traffic)

Use the download tab and select spreadsheet to download the dataset as a csv file, also download the pdf guide the describes the metadata and navigate to the specific section that describes the KSI dataset. You can find the guide at this link [Open data documentation](https://data.torontopolice.on.ca/search?collection=document)

**5) Evaluation of software solution:**

|  |  |
| --- | --- |
| **Evaluation Component** | **Percentage** |
| **Data exploration** | **20%** |
| **Data modelling** | **20%** |
| **Predictive model building** | **10%** |
| **Model scoring and evaluation** | **5%** |
| **Model deployment** | **20%** |
| **Report phase #1, Report phase #2** | **5%, 5%** |
| **Presentation phase #1, Presentation phase #2** | **5%, 5%** |
| **Innovation & creativity phase #1, phase # 2** | **2.5% , 2.5%** |
| **Total** | **100%** |

**6) Rubric:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Evaluation Component | Not acceptable  Score =  0 | Below average  Score 1 -5.9 points | Average  Score  6-6.9 points | Competent  Score  7-8.9 points | Excellent  Score  9-10 points |
| Data exploration  Score multiply by 1 | Data exploration completely missing or what is submitted is below 30% with no relationship analysis. | Only 50%-60% of dataset attributes have been explored or exploration not complete on # of missing values, only a few relationships are captured. | Only 60%-70% of dataset attributes have been explored or exploration not complete on # of missing values not all relationships are captured. | Most dataset attributes columns have been explored and a complete description of each attribute value meaning has been reported in addition to exploring some relationships between attributes. | All dataset attributes columns have been explored and a complete description of each attribute value meaning has been reported in addition to exploring all relationships between attributes. |
| Data modelling  Score multiply by 1.5 | Used the data attributes as is with no cleaning or manipulation. | Identified a data model with no justification.  Minimum data cleanings. | Identified a data model with 50-70 %   Justification.  Applied moderate data cleaning. | Identified a data model with 80 %   Justification.  Applied data cleaning. | Identified a comprehensive data model.  Data model clearly states which attributes should be dropped and why also which attributes should be transformed with a logical justification. |
| Predictive model building  Score multiply by 1.5 | Did not build any to build the predictive models. | Used one of the specified minimum algorithms effectively to build the predictive models with mistakes. No hyper-parameter tuning. | Used one of the specified minimum algorithms effectively to build the predictive models without mistakes. Minimum hyper-parameter tuning. | Used specified minimum algorithms effectively to build the predictive models with minor mistakes.  Minimum hyper-parameter tuning. | Used specified minimum algorithms effectively to build the predictive models without any mistakes. Carried out hyperparameter tuning effectively. |
| Model scoring and evaluation  Score multiply by 1.5 | No model evaluations  conducted | Some metrics are generated for each model with no comparisons / conclusions presented. | Some metrics are generated for each model with minimum comparisons presented with partial conclusions. | All metrics are generated for each model and a comprehensive comparison presented with partial conclusions. | All metrics are generated for each model and a comprehensive comparison presented with clear conclusions. |
| Model deployment  Score multiply by 2.5 | No demo  No web server code developed | Some code with many errors, not working. | Submitted and presented a working analytics web app with some issues in message exchange Code is not modular nor documented. | Submitted and presented a working analytics web app showing messages exchanged, through a frontend or simulator.  Code is modular and partially documented. | Submitted and presented a working analytics web app showing messages exchanged, through a frontend or simulator.  Code is modular and well documented. |
| Report  Score multiply by 1 | No Report submitted. | Has no clear introduction or has an irrelevant introduction; gives reader no focus or outline of the report. | Has introduction relevant to the topic and gives a clear outline; is not organized; has many repetitions the reader would get confused. | Has introduction relevant to the topic and gives a clear outline; is mostly organized; provides adequate “road map” for the reader. | Has a clear introduction & outline that catches reader’s interest; maintains focus throughout; summarizes main points |
| Presentation  Score multiply by 1 | Team not ready for presentation. | Some team members do not participate.  All read off the slides.  Voice too low or too high.  Demonstrates little grasp of information; has undeveloped or unclear answers to questions | Some team members participate in the delivery;  Some team members read off the slides.  In response to questions the team shows ease in answering questions but does not elaborate. | All team members participate equally in the delivery;  Some team members read off the slides.  In response to questions the team shows ease in answering questions but does not elaborate. | All team members participate equally in the delivery;  Team projects enthusiasm, interest, and confidence; uses body language effectively.  In response to questions demonstrates full knowledge of topic. |
| Innovation & creativity  Score multiply by 0.5 | No suggestion or implementation of new algorithms | Team suggests new non-relevant techniques/models/algorithms without implementation | Team suggests new relevant techniques/models/algorithms without implementation. | Team researches and demos new relevant techniques/models/algorithms with some unclarity. Or introduces a front-end with a few mistakes. | Team researches and demos new relevant techniques/models/algorithms efficiently. Or builds a front end. |

**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/>

[KSI dataset](https://data.torontopolice.on.ca/datasets/TorontoPS::ksi/about)

[PUBLIC SAFETY DATA PORTAL: OPEN DATA DOCUMENTATION](https://ago-item-storage.s3.us-east-1.amazonaws.com/c0b17f1888544078bf650f3b8b04d35d/PSDP_Open_Data_Documentation.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEMb%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJGMEQCIA1YfDmhR6VtaBTbCOhwMBzSsTwBce1bePd1G23AWmrrAiAyHhceE1GTfVO2XcMQWbSZZf3uoZKcjGaeQL8A5qSYgCrcBAiO%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F8BEAAaDDYwNDc1ODEwMjY2NSIMIQZGPu6c9Ej7aeBeKrAEcHgQKG4roAZyVAloJwVxhiCK%2FikUQfI280kGdp48jqHC2TobFBh5PfpRPryD%2BYchNXLg%2BJ8OexcLuk6iinRj6bcJGZcvoMRSdBbRx3dKJJdGHYCz6VNT0%2BIADpePEN0792YhRQsjcvTNm%2FuKuqJcFxO60GSvAfPMtMLEMESwoZ3NWRcpzQk7DhjH7gaCNZYNNP6t%2FFpHoalNmNH8nMmzZmbeRy1MaEEUedGtbYRTdYFeewxCq3JjuRvBRl7HM%2B7wlwDlsu2l%2FYCvITEFKhAwvFgTdT4tOtMJHzisP5LeHlvFPWSmGhrEO9lJx7HkWk2bE8i4K7r4GZc5DR2UdEtr%2FIePoUElXVqZk8yRt9dA22mZr846%2FzCA3XsG4y9G%2BJv1fsZqSGZZp5j5ba%2BVDJiKatwv6bIGEXDaXnynrqFNwe%2F59kosLH7jecg1canrm0BB6fk0vbHPgH%2FJXRbNZfteSj6mdpx3wlYyePq0eSvhOZdhJdw7PRiQUOHwtIHzQ7XTvYBqlgSJlVsCQwJNfTbn%2BzH6x5Q0%2BC6w4yZ%2FkrfPRDM55qOOJgrfHfuGS4IWaoNN3sJCMCbyLW7nyFDQLHxXsRjGttMwebAEnKq0JjLDdowf1AzGhObpPwYkkcYaO9Nq0f8sBEcFZBIlWHa35dY83O3WBPRsktlFpMRlv4gQbRQTAYN9Vyf4LkrDJ%2FXzaS4nnvexTQ3wNqEnl7FHMuAl%2BiEHYZiyevCrfPw%2FCXeXPTIw8ZrPkwY6qgGavDRPQB8paix69lnZqSD6GVGCWiCipLSLZjqeg1fcY3w4U82btiRDgyC%2BYLcpx%2BMBG7vfYPjPEWmzpDXHwNokTSSuwsF8wFsSDR%2FZaRZY%2FPR8gOE5gbC%2Fa%2Bw6t345C%2Bky9N7hUC0BXAuzY%2Bg9R%2B%2BA69YF5Rm8eHv2FRr%2BBYyQ6rwS5eVcTDVJnDOO6Gr4gbBq7x1oTN9jMVtkgkgBTCOivQdwNAwwQN%2FwaQ%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20220505T144248Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAYZTTEKKEQLTJ47HU%2F20220505%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=bc3783b540bdee4bcb6d8e6924a096d4da4f271180c8aa32fde7d560aec2c824)

<https://elitedatascience.com/imbalanced-classes>