MIE 1628 Big Data Science – Fall 2020
Final Project
Due Date: TBA

1 About the Data

1.1 Predict the Quality Rating of Stack Overflow Questions

60,000 Stack Overflow questions have been collected from 2016-2020 and classified into three categories:

1. HQ: High-quality posts with 30+ score and without a single edit.
2. LQ_EDIT: Low-quality posts with a negative score and with multiple community edits. However, they remain open after the edits.
3. LQ_CLOSE: Low-quality posts that were closed by the community without a single edit.

Moreover,

1. Question body is in HTML format.
2. All dates are in UTC format.

1.1.1 Features and Target

- Id: The id of the post
- Title: The title of Stack Overflow question
- Body: Question Body in HTML
- Tags: Question Tags
- CreationDate: Creation Date in UTC format
- Target: Quality rating = **Target**

1.2 Identify Potentially Hazardous Asteroids

The dataset contains different physical parameters and measurements for over 900,000 asteroids. Nowadays Machine Learning is solving so many problems in Astronomy and Astrophysics fields. This Dataset is officially maintained by Jet Propulsion Laboratory of California Institute of Technology which is an organization under NASA. In this Dataset all kinds of Data related to Asteroid is included.

1.2.1 Features and Target

- SPK-ID: Object primary SPK-ID
- Object ID: Object internal database ID
- Object fullname: Object full name/designation
- pdes: Object primary designation
- name: Object IAU name
- NEO: Near-Earth Object (NEO) flag
- PHA: Potentially Hazardous Asteroid (PHA) flag = **Target**
- H: Absolute magnitude parameter
- Diameter: object diameter (from equivalent sphere) km Unit
- Albedo: Geometric albedo
- Diameter_sigma: 1-sigma uncertainty in object diameter km Unit
- Orbit_id: Orbit solution ID
- Epoch: Epoch of osculation in modified Julian day form
- Equinox: Equinox of reference frame
- e: Eccentricity
- a: Semi-major axis au Unit
- q: perihelion distance au Unit
- i: inclination; angle with respect to x-y ecliptic plane
- tp: Time of perihelion passage TDB Unit
- moid_ld: Earth Minimum Orbit Intersection Distance au Unit

1.3 Improve the algorithm that classifies drugs based on their biological activity

This project is an ongoing Kaggle competition. You can find the details about the competition here - https://www.kaggle.com/c/lish-moa/overview

The Connectivity Map, a project within the Broad Institute of MIT and Harvard, the Laboratory for Innovation Science at Harvard (LISH), and the NIH Common Funds Library of Integrated Network-Based Cellular Signatures (LINCS), present this challenge with the goal of advancing drug development through improvements to MoA prediction algorithms.

The prize money is $30,000.

2 Learning Objectives

2.1 Data Cleaning

There may be missing values in the dataset, handle the missing values however you see fit and justify your approach. Provide some insight as to why you think the values are missing and how it might affect your overall analysis. For text data, there may be special characters present. Similarly, handle them as you see fit and justify your approach in the report.

2.2 Feature Engineering

The original features in the data should be used to create additional, innovative features. The goal here is for you to create additional features from existing ones which would boost the final model performance. Please support your new features with plots and/or statistical analysis.

2.3 Exploratory Data Analysis

Present at least 4 graphs (may include plots from 2.2, but new unique plots would be a plus) which may represent trends in data and explain how those trends might be helpful for your machine learning algorithm. All the graphs must be appropriately labelled (axes and titles).

2.4 Feature Selection

Analyze and visualize the importance of your final features and then from the analysis, select either manually or through feature selection algorithms the features chosen for your model. Using the feature selection technique and its justification would be a plus.

2.5 Model Implementation and tuning

Build at least 3 machine learning models to predict the target. Choose appropriate evaluation metrics to evaluate your models and provide graphs and statistical data to support your model. Draw conclusions on the strengths and weaknesses of each model and select a winner model based on this.

Tune the hyperparameters of the winner model and draw a comparison of tuned and untuned model using appropriate evaluation metric. Use k-fold cross-validation to get a better sense of the model's performance.

2.6 Model Testing and Discussion

One week before the project submission deadline, a new dataset will be released. Apply your winner model on this data and analyze whether your model is overfitting or not.

3 Note:

1. The project must be done using PySpark and/or Scala. Python should not be used apart from data visualization.
2. No other tool or software besides those mentioned in 1. can be used to modify the data files. For instance, using Microsoft Excel to clean the data is not allowed

4 Submission Format

4.1 Report

We will not specify a specific format for the report, but your report should include everything discussed in Section 2.

4.2 Group Presentation

In addition to presenting your work, each team member will be subjected to questions regarding the code and on-the-fly questions regarding big data and machine learning (only from material taught throughout the course).

4.3 Code Notebook

The code must be well organized and must have appropriate comments and explanations as mentioned in the learning objective.