East West University
Department of Computer Science and Engineering

CSE303: Project Description [Spring 2021]
Course Instructor: Dr. Mohammad Rezwanul Huq

Objectives:

The following qualities will be judged through the project of this course.

1. Understanding and analyzing the data set in terms of data distribution and correlations.
2. Understanding and applying theory and practices to implement machine learning models such as Logistic Regression, Support Vector machines, etc.
3. Ability to work in a team.
4. Ability to demonstrate and present the major concepts behind your ML models.
5. Overall aptitude level in this course.

Dataset Description:

The dataset is the same for all teams. It can be downloaded from the following link.

https://drive.google.com/drive/folders/1-HJ8w60hJwuQHiHEcU48kgIDGffGQyII?usp=sharing

The dataset has 30 features. All of the feature columns, cat0 - cat18, are categorical, and the feature columns cont0 - cont10 are continuous.

The complete dataset folder contains the followings:

1. train.csv – the training data with the target column (either 0 or 1). Each row has a unique identifier. Therefore, 32 columns in total are available in the dataset, and there are 300,000 rows.

2. test.csv - the test set; you will be predicting the target for each row in this file (the probability of the binary target). There is no target column in this dataset.

3. sample_submission.csv – A sample submission file contains the id of the row from the test dataset and the predicted value of the target variable. You must submit this file for each of your developed model separately and must be named as group-id_model_number_result.csv (group1_model1_result.csv)

Project Description:

Given the dataset, you need to develop a supervised machine learning model for classification. You may develop models based on the topics covered in the class, such as Logistic Regression and Support Vector Machine. You may try with changing different parameters to obtain better training accuracy. There are two baseline models that you must implement first. These are the followings:

1. Logistic Regression with 'liblinear' solver and C = 1.
[sklearn.linear_model.LogisticRegression]

~~2. Support Vector Classifier with 'linear' kernel and C = 1.~~
~~[sklearn.svm.SVC]~~
2. Linear Support Vector Classifier with C = 1
[sklearn.svm.LinearSVC]

Besides these models, you must include at least two other models (with different parameters), which outperform these model performance for the training data set.

Remember the followings while doing your project:

- Data preprocessing is important. (checking NULL, duplicate)
- Encoding categorical data is required.
- Dimension reduction could be beneficial.
- Evaluate your model's performance in terms of accuracy, precision, recall, f1-score, ROC Area Under the Curve (AUC) score.
- Visualize and compare the performance of different models using appropriate charts, figures, and so on.

Project Report:

You must write and submit a report following the template. You may extend the report by adding additional sections if required. The report must be submitted in both doc/docx and pdf format. You also have to submit the originality report with your report. See the tutorial video (https://www.youtube.com/watch?v=Xrrei9jeib4) on how to provide an "Originality Report" in Google Classroom is attached herewith. For students, the relevant part starts from 0:55 to 2:20 in the video. Failure to submit the originality report will result in a deduction of one-third marks automatically.

Project Deliverables:

1. All codes (py files).
2. Sample submission files, one for each model.
3. The project report both in doc/docx and pdf format.

4. The originality report.

Put everything inside a zip file. The name of the zip file should be CSE303_group1.zip.