DOUGLAS COLLEGE – Summer 2021

CSIS 3290 – 001 – Fundamental of Machine Learning Term Project

Your project will be considered as a **PLAGIARISM** if:

1. The project is similar to the work submitted by other students within the same term, other terms or other sections.
2. The project is similar to any work found in the Kaggle or any other websites.

Project resulted from plagiarism will receive a ZERO mark and the students associated with the work will be reported to the Dean office for academic dishonesty case.

Jupyter Notebook that produces error message will receive a ZERO mark.

Project Description:

Through out this course you have been learning and acquiring key skills to successfully implement machine learning algorithms to perform data science. In this Project, you will be **working with a dataset of your choice to reflect on your learning in this course**. Note: **you cannot use the dataset that we have used in the lab/assignment/tests.**

Project Submission Requirements

File/folder structure and naming convention

You need to create a folder named Project_ABcXXXXX with A signifies the first letter of your first name, Bc signifies the first two letters of your last name and XXXXX denotes the last five digits of your student ID. The project **must be submitted as a zip file**. Other type of compression (tar.gz, tar, bz2, rar) is not acceptable. Please **make sure to check whether your zip file can be unzipped and contains all the required files for the project to work properly**.

The zip file must have the following structure:

Name/structure	Details
Project_ABcXXXXX.zip	Submitted zip file
└→ Project_ABcXXXXX	The project's folder
└→ Report	The project's document folder containing your term report, and images (if any)
└→ Dataset	The project's dataset folder containing the dataset
└→ < jupyter notebook>	Collection(s) of your Jupyter Notebook file(s) for the project

Project's Components

1. **Jupyter Notebook File(s) –** inside the main folder
   Put all your Jupyter Notebook file(s) in the main folder of your project. Please use Project_ABcXXXXX as the naming convention of your Jupyter Notebook main file. If you create python module(s) for your functions, please use a meaningful name for them. Please look at the Jupyter Notebook requirements mentioned in the following page(s).

2. **Dataset File(s) –** inside the dataset folder
   Please provide the datasets that you use in your project. Any temporary csv file that you used while building the solution should be placed here as well. Please provide a short documentation in text file explaining the dataset (feature) and how it was obtained, i.e., links, etc.

3. **Term Project Report (5 to 10 pages) –** inside the documentation folder
   Prepare your report professionally. Your English writing skill and the structure/composition of the report is part of the marking. You do not need to have a title page, but make sure to write down your name, student ID and the topic of your project. The document should include the following.

Jupyter Notebook Requirement

Jupyter notebook that produces error message will receive a ZERO mark

- **Text Content**
  - Provide your name, student ID and a title at the top of the page.
  - Specify any major references used in creating your machine learning pipeline.
  - Any python modules used in the project that were not covered in the class should be briefly described. References for these modules should be provided.
  - Provide brief introduction of the problem you are trying to solve, a bit info about the dataset, and the summary of steps in your machine learning pipeline.
  - Make sure to use the correct markdown heading signifying the steps performed in your Jupyter notebook. Any code should be accompanied with some brief text content describing the steps/procedures being implemented in the corresponding cell.
  - Any visualization or plot must be accompanied with some text explaining your observation.

- **Code implementation**
  - Your implementation should be clean from any unnecessary code (whether it is commented or not), have a clear flow of logic and purpose, and include comments as necessary.
  - The code implementation should match the project report content. Depending on how you implement your machine learning pipeline, you should have part of the code that performs:
    - Data wrangling and transformation
    - EDA (interesting plots/charts of the data with observation is highly appreciated)
    - Feature engineering that includes transformation, selection, or scaling
    - Your machine learning pipeline implementation(s)
    - Report, error (and plot of) metrics, and results analysis
    - Out-of-sample prediction

Project Report Requirement

- **Introduction and discovery**
  - Introducing the business domain (a brief background about the target company/organization/dataset)
  - Framing the problem (what questions do you want to address in your analysis and why are they important)
  - Developing initial hypotheses

- **Data Preparation**
  - Data inventory – provide a brief introduction of the dataset(s), where you obtain it, and summary of the features you have
  - Data processing – provide the summary statistics, a brief peek of the data and briefly specify any data transformation done in your pre-processing

- **Model Planning and Implementation**
  - Proposed model(s) and justification – justification could be based on the structure of the data and literature review of past similar studies
  - Determine if the situation warrants a single model or a series of techniques as part of a larger workflow (e.g., you could begin by using cluster analysis and then apply regression techniques to each cluster identified). Or the data could be repurposed to do both multiple regression and classification, etc
  - How you made your project workflows more efficient (hint: use of pipelines)
  - Discuss how the chosen techniques facilitate testing of the hypotheses and provide insight on the modeling objectives.

- **Results Interpretation and Implications**
  - Show the results of your machine learning implementation. Provide some plots and/or summary tables of your result(s)
  - Assess if the results are statistically significant and valid. Question to consider when interpreting the results:
    - Does the model appear valid and accurate on the test data?
    - Does the model output/behavior make sense to the domain experts?
    - Do the parameter values make sense in the context of the domain?
    - Is the model sufficiently accurate to meet the goal?
    - Does the model avoid intolerable mistakes?
    - Are more data or inputs needed?
    - Is a different form of the model required to address the problem?
  - Communicate and document the key findings and major insights derived from the analysis

- **Out-of-sample Predictions**
  - Using your final model, perform predictions using new data (i.e., out-of-sample data) and comment on the results
  - Note: You will need to generate/obtain new data for out-of-sample predictions. This is different from test dataset, which is used for model testing. New data is trying to simulate how your model would perform if deployed in the production environment in the real world.

- **Concluding Remarks**
  - Summary of the analytics process went through, major findings and key business (managerial) implications

Project Grading Criteria

The project will be graded on a scale of 40 points.

Criteria		Grading
The project was submitted, named properly with all components included in their corresponding folders to the Blackboard.		1 point
Report	Good English and report structure/composition	2 points
	• The report includes all the required components as stated above. The content matches with the submitted Jupyter notebook implementation. • The term project report shows that the student exhibits expertise in implementing machine learning techniques for data science. • The report shows that the student can come up with a problem definition, collect necessary dataset(s), plan and implement machine learning pipeline model(s), and analyze the obtained test results convincingly. • The report is coherent and flow without any jump of logic. Justifications, assumptions, tables and charts are explained and commented.	8 points
Jupyter Notebook	Have necessary text content as specified in the requirement	3 points
	Good and clean code structure with comments as necessary	2 points
	• The code includes all the required steps and does not produce any error messages. • The code shows that the student understands the problem at hand and implement the necessary dataset observation, transformation, exploration, and machine learning pipeline modeling and analysis to provide solution to the problem. • The actions implemented in code have a clear purpose. The assumption and decision taken have sound reason. • The complexity of the problem and its code implementation should be higher than any work that the student has done within the course, i.e., midterm and project 1.	24 points
Penalty	The code produces error message(s)	–29 marks