CSIS 3290 Fundamental of Machine Learning
Project 03

Due date: Dec 06, 17:00

NOTE:

Instructor may ask the project group or individual members questions about the actual work and contribution after submission. The questions may be very specific (E.g., the lines of code written by a member and the explanation of the codes.) It may be done via email or in-person after the class without advanced notice. Failing to answer the questions satisfactorily will result in mark deduction. Please ensure all group members have roughly even contribution in the project, especially the coding part, and have full understanding on the whole project.

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 on Dataset:

1. The dataset must not be anyone we have used in the lab/assignment/tests.
2. The dataset needs to be recently published and not order than 5 years.
3. The dataset must not be a popular dataset in Kaggle, better never used in Kaggle.
4. The dataset cannot be too small (at least more than 1000) and too large, making sure it can be loaded in my computer.

Note on Methods:

1. The method used in the project must not be a Deep Learning approach.
2. Your program needs to finish running in 2 minutes.

Project Submission Requirements

(Note: A report in MS Word file is required.)
A zip file containing all materials in the submission. The zip file must have the following structure:

Name/structure	Details
Project.zip	Submitted zip file
↳ 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

A. Requirement

1. **Jupyter Notebook File(s) –** inside the main folder
   Put all your Jupyter Notebook file(s) in the main folder of your project. If you create python module(s) for your functions, please use some meaningful names 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 need to have a title page containing your name, student ID and the topic of your project. The document should include the following.

   - **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.

- **Member Contribution**

Each group needs to submit a peer evaluation matrix. Each cell should be a number between 1 and 4, which reflects how a member thinks the contribution by another member. The evaluation is opened to open to all members of your group (i.e., Every one can see how others grade on you), so that each member knows how to enhance their contribution in the project.

(Hint: You may refer to this link to see how to create a table in a Jupyter Markdown cell.)

Evaluator \ Evaluatee	Member 1	Member 2	Member 3	Member 4
Member 1				
Member 2				
Member 3				
Member 4				

Here is the rubric on how to evaluate your team members:

1 Point: No or very little contribution to the project; cannot deliver artifacts or largely miss the agreed deadline; showing no or very little passion in development.

2 Points: Little contribution to the project with no negative effect to the group; sometimes cannot deliver artifacts or miss the agreed deadline; mainly follow other members' idea and instructions.

3 Points: Fairly large and positive contribution to the project; can handle most of the assigned tasks and deliver artifacts on time;

4 Points: Large and positive contribution to the project; can help members to tackle problems; pro-active and passionate in the development.

B. Jupyter Notebook Requirement

Jupyter notebook that produces error message will receive a ZERO mark. Make sure your re-run the whole notebook before submission.

- **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.
  - 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.
  - Create necessary markdown cell to divide the code into sections and provide short description of the code.
  - 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 (can be included in the pipeline)
    - Your machine learning pipeline implementation(s)
    - Report, error (and plot of) metrics, and results analysis
    - Out-of-sample prediction

C. Presentation

Each group needs to create a video presentation which is not more than 20 minutes. [You may use any software to record the video. I have tried ActivePresenter (http://atomisystems.com/download/) and it works pretty well. There is a tutorial in http://atomisystems.com/activepresenter/tutorials/ on how to record and export the video. **Remember to save the presentation once after recording to avoid from losing your effort.] Upload your video to YouTube or any cloud storage. Include the link in the report, so that I can watch and grade your presentation.

D. Project Grading Criteria

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

Criteria		Grading
The project was submitted, named properly with all assets included in their corresponding folders to the Blackboard.		1 point
Report	Good English and report structure/composition	2 points
	The term project report shows that the student exhibits expertise in implementing machine learning techniques for data science. It 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.	8 points
Jupyter Notebook	Have necessary text content as specified in the requirement	2 points
	Good and clean code structure with comments as necessary	2 points
	The code implementation includes all the required information. It clearly shows that the students understand the problem at hand and implement the necessary dataset observation, transformation, exploration, and machine learning pipeline modeling and analysis to provide the solution of the problem.	20 points
	Code produces error message(s)	−23 points
Presentation	Excellent presentation with clear voice, sufficient eye contact, necessary demo/code explanation.	5 points