Texas Tech University Whitacre College of Engineering

CS 4331 – Data Mining

Spring 2020

Project

March 9, 2020

Full Name: ____ Class Section:______________________

Acknowledge your collaborators or source of solutions, if any. **Submission dates are given below.** This project is for getting familiar with *the Data Science Methodology*. For coding, use Python 3 or R. A subset of your answers will be graded. Include your .py or .r file with helpful comments included and no hard-coded file paths (only the file name should be present when reading or writing the file). All files should be submitted in a zip file and named CS4331_00?_eraiderlogin_Vn.zip where ? is replaced with your class section number, eraiderlogin is your eraiderlogin, and n is replaced with the version. You may submit up to 3 times and only the last version will be graded.

Let Dr. Mengel know by March 26, 2020, your team and a general description of the problem you would like to work on (be sure you can get the data). Team size may be up to 3 people if there is a demonstrated need for each team member. Team members are expected to fully participate in all phases which excludes one team member doing all of a task, such as data cleaning, modeling, or writing the report. Low team member participation will be handled on a case by case basis.

There are many tutorials on the web and elsewhere on data science and data sets. You may not do a tutorial (even if modified slightly or heavily) and turn it in as your work.

- Some sources of datasets are:
  - Kaggle - https://www.kaggle.com/datasets
  - DrivenData - https://www.drivendata.org/
  - Google Dataset Search - https://datasetsearch.research.google.com/
  - Microsoft Research Open Data - https://msropendata.com/
  - Awesome Public Datasets - https://github.com/awesomedata/awesome-public-datasets
  - LionBridge - https://lionbridge.ai/datasets/
  - FiveThirtyEight - https://data.fivethirtyeight.com/
  - Food Environment Atlas - https://www.ers.usda.gov/data-products/food-environment-atlas/
  - IES>NCES National Center for Education Statistics - https://nces.ed.gov/
  - World Bank Data Catalog - https://datacatalog.worldbank.org/
  - Harvard Dataverse - https://dataverse.harvard.edu/
  - DataHub.io - https://datahub.io/
- Sources of Country Data
  - EU Open Data - https://data.europa.eu/euodp/data/dataset
  - Data.Gov - https://www.data.gov/
  - New Zealand - https://catalogue.data.govt.nz/dataset
  - India - https://data.gov.in/
  - Northern Ireland - https://www.opendatani.gov.uk/
  - UK Data Service - https://www.ukdataservice.ac.uk/

Project I - 8 points, due April 16

For Project I, conduct the following 3 phases:

- Problem Understanding Phase
  - Enunciate clearly the problem objectives
  - Translate the objectives into a formulation of a problem that can be solved using data science.

- Data Preparation Phase
    - Prepare the data for analysis.
    - Identify outliers and determine what to do about them
    - Transform and standardize the data
    - Reclassify categorical variables
    - Bin numerical variables
    - Add an index field
- Exploratory Data Analysis Phase
    - Explore the data
    - Explore univariate relationships between predictors and the target variable
    - Explore multivariate relationships among the variables
    - Bin based on predictive value to enhance models
    - Derive new variables based on a combination of existing variables

Only one team member should submit your Python 3 or R scripts, your dataset(s), and an up to 2-page report on the 3 phases (1" margins, 10-point black font, single to double spacing). Be sure to highlight what each team member did in each phase. You will be graded on the thoroughness of each phase (it looks like you did 3 weeks of work), the quality of your work as taught in the course, the ability to discern the individual work of each team member, and on the novelty of your work (you can note similar efforts with the datasets and why your approach is novel).

Project II - 8 points, due April 30

For Project II, conduct the following 4 phases:

- Setup Phase
    - Cross-validation either twofold or n-fold or both - data partitions should be evaluated to ensure that they are indeed random
    - Balance the data
    - Establish baseline performance
- Modeling Phase
    - Apply state-of-the-art algorithms to uncover some significant profitable relationships lying hidden in the data (each team member can do a different modeling algorithm, for example)
    - Select and implement appropriate modeling algorithms (one per team member)
    - Make sure that the models outperform the baseline models
    - Fine-tune your model algorithms to optimize results
- Evaluation Phase
    - Assess how the models are doing at helping to solve the problem
    - Evaluate against the baseline performance measures from the Setup Phase
    - Determine if the models are solving the problem at hand and achieving the project objectives
    - Perform data-driven cost evaluation to model the actual costs involved
    - Determine which of a suite of models performs the best
- Model Deployment
    - Write up to a 3-page report on the Setup, Modeling, and Evaluation phases (1" margins, 10-point black font, single to double spacing)

Only one team member should submit your Python 3 or R scripts, your dataset(s), and Model Deployment Report. Be sure to highlight what each team member did in each phase. You will be graded on the thoroughness of each phase (it looks like you did 3 weeks of work), the quality of your work as taught in the course, , the ability to discern the individual work of each team member, and on the novelty of your work (you can note similar efforts with the datasets and why your approach is novel).

Source: Chantal D. Larose, Daniel T. Larose, Data Science Using Python and R

All assignments will be checked for academic misconduct (cheating, plagiarism, collusion, falsifying academic records, misrepresenting facts, violations of published professional ethics/standards, and any act or attempted act designed to give unfair academic advantage to oneself or another student) defined by "OP 34.12: Grading Procedures, Including Academic Integrity" of TTU.