Data Science Project

Your final assignment will be an end-to-end data science project using a dataset of your choice. This will allow you to explore a practical data science problem you are interested in in greater detail.

Course and project outcomes:

- Understand the basic process of data science and exploratory data analysis including modes of inquiry (hypothesis driven, data driven, and methods driven).
- Identify, access, load, and prepare (clean) a data set for a given problem.
- Select, apply, and interpret appropriate visual and statistical methods to analyze distributions of individual variables and relationships between pairs of variables.
- Communicate findings through generated data visualizations and reports.
- Apply and interpret unsupervised learning models for exploratory data analysis.
- Generate appropriate supervised learning problem descriptions.
- Determine and apply appropriate experimental setup, evaluation metrics, and models for supervised learning problems.
- Engineer features for machine learning tasks.
- Perform and interpret feature selection to identify relationships between features and predicted variables.
- Apply methods to real world data sets.

Project Proposal

You must write a project proposal, and have it approved by the instructor. Your proposal must identify your dataset, the objectives for analyzing this dataset, the first ten records of the dataset demonstrating that you have been able to load the data, and team member (if any). You may work in teams of two, but this is not required.

I've provided sample datasets below which are all acceptable, but you are free to propose a different dataset.

Deliverables

You are expected to follow the data science workflow presented in class and include this workflow within a well-documented Jupyter Notebook. Each of the bolded headings should be addressed and clearly identified in your notebook.

Hypothesis – what problem(s) you are trying to solve, what questions are you trying to answer.

Data Set – explanation of the dataset. This should include an explanation of why the dataset was created, who created it, and an explanation of features, target variables, etc. It is important to provide appropriate references.

Data preprocessing – identification of required features, proper handling of missing values (removal, imputation, etc.), and supporting documentation for your decisions such as the number of missing values relative to the size of the dataset.

Data analysis and visualization – quantitative (statistical) and visual analysis of your data. You should apply quantitative measures of centrality, distribution, and correlation. In addition, you should provide visual analysis through plots that are appropriate for a better understanding of your data and how the data can be used to answer your question. Extra credit for novel insightful visualizations. Using the appropriate clustering algorithm for different measures can be very helpful for better understanding the relationships in your data.

Data modeling and prediction – development of a predictive model, regression or classification. Your model development should include experimentation with feature selection. The effect of different features on your target variable as well as associations between features should be identified. You should appropriately use train and test data when evaluating your models.

Results analysis – Summary of your interpretation of the results. Are the results statistically significant? I.e., did you prove or disprove your hypothesis. How could you improve your analysis? Lessons learned, and feedback on the assignment.

Project Submission

Your project submission should include a README.md, main Jupyter Notebook file, HTML version of your notebook file, and any additional images and supporting Python source. Your entire project should be zipped and submitted on Blackboard. The name of the submitted zip file must include names of team members and the title (abbreviated) of your project. If you use code from another project, it must be clearly identified. Plagiarism will not be tolerated.

Grading

Grading will be based on each of the **bolded** requirements for your notebook. This project counts as 20% of your grade.

Sample Datasets

Airbnb price prediction
https://www.kaggle.com/stevezhenghp/airbnb-price-prediction

PIMA Indians Diabetes Database
Predict whether or not the patients in the dataset have diabetes or not
https://www.kaggle.com/uciml/pima-indians-diabetes-database

TMDB movie success prediction database
https://www.kaggle.com/tmdb/tmdb-movie-metadata/home

S&P Stock Data
Predict future price (of course)
https://www.kaggle.com/camnugent/sandp500

Across the Bay 10K Race

- What do the different distributions of runners look like (i.e. geographic, age, gender)?
- What does the runner finish time distribution look like?
- How did actual times compare to the times runners estimated for themselves (based on bib number since I did not have data for the estimated times but bib number assignment was sequential according to estimated time)?
- Is estimated time a good measure for dividing runners into waves?
- How well did runners stick to their assigned waves?
- How could wave organization be improved for next year?

https://sites.google.com/site/atb10kbridgerace/home

Other ideas:

- Predicting the up votes / down votes of a reddit comment. Data will be pulled via the Reddit API
- Predicting the rating of a game on Steam
- Predicting which team will win a game in a match up
- https://www.kaggle.com/datasets
- https://archive.ics.uci.edu/ml/datasets.html