Data Science Final Project

Your final assignment will be an end-to-end data science project using a dataset of your choice. This project will follow the format of a research paper/manuscript. 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.
- The ability to identify, load, and prepare a data set for a given problem.
- The ability to analyze a data set including the ability to understand which data attributes (dimensions) affect the outcome.
- The ability to perform basic data analysis and statistical inference.
- The ability to perform supervised learning of prediction models.
- The ability to perform unsupervised learning.
- The ability to perform data visualization and report generation.
- The ability to assess the quality of predictions and inferences.
- The ability to apply methods to real world data sets.

Project Proposal

You must write a project proposal (~1 page), 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 members. You will work in the small groups that you have been assigned to.

Sample datasets have been provided at the end of this document. You may use of one of these sources or propose the use of 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.

> **Research Question** – what problem(s) you are trying to solve, what questions you are trying to answer.
>
> **Hypothesis** – A well-formed hypothesis (statement) that you will be testing. This should follow the form discussed in class. Be as specific as possible.
>
> **Dataset** – 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. A naïve reader should be able given enough information to understand the work done on the dataset.

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. It is expected that you will use an established machine learning library – you are not expected to implement an algorithm from scratch.

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 Presentation

Your group will create and deliver a 5-10-minute presentation on your project during finals week of the quarter. In your presentation you should identify and explain the following:

- Pertinent background information – You should provide the audience with the background necessary to understand the research question, and where the data set came from and how it was created.
- The hypothesis – You should clearly state the hypothesis that you will be testing.
- The experimental design – You should provide a clear description of the experimental design, including methods for data pre-processing, data visualization, any machine learning models applied to the dataset, statistical tests, and result metrics.
- Results – You should give a clear explanation of the results found from this experimental setup. Including expected results and any unexpected results or observations gained from the processing of the dataset.
- Discussion – With results in hand, you should reflect on the process that you went through. This should include implications that your results have on the broader topic and any future studies you would propose based off the study findings.

Project Submission

You should submit all files associated with your project through Canvas by Saturday night in the finals week. A single zip file should be submitted. A single submission may be made for all group members – All group members names should be included on the title page of the presentation. Required files include:

- Presentation slide deck (pptx or pdf) of the presentation that will be given in week 10
- A pdf of the Jupyter notebook containing the entire project from beginning to end. (any classes or methods written should be defined within the notebook)

Grading

Grading will be based on the provided presentation grading rubric. Please see syllabus for class grade weight.

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 number of downloads 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.php