University of Virginia
DS 5110: Big Data Systems
Prof Tashman

Final Project Instructions

Last updated: June 7, 2022

Instructions

Students will form teams of roughly 4 students with the goal of completing a substantial course project. This document outlines the necessary components of the final project.

Deliverables

The final group project has three deliverables of equal value (each worth 10% of overall grade):
I. Paper
II. Code
III. Presentation

These are detailed below, along with general components of the project.

Components

1. Research question

This is the motivation for the project. Examples:

a. We will predict which businesses will go bankrupt within 12 months
b. We will predict the rate of return of Apple stock tomorrow
c. We will construct a model to rank-order the risk of hospital patients with a given infection

A project proposal needs to be submitted by the group.
Since the project will be done over several months, it should be substantive, and it should be relevant. Imagine the project on your resume, you are interviewing at Google, and they ask you about it. Are you proud because it was a kickass project, or do you want to run and hide?

2. Data

There are several free sources of online data including the ones below. If you need suggestions, I'm happy to discuss. Larger datasets with more opportunity to show your data ninja skills are better. For example, finding a clean, 100 row dataset in csv format online is not ideal. This is an opportunity to be creative and show off! ***Please limit the size of the dataset to 5GB.***

Physionet
Physiological signals including ECGs
https://physionet.org/

NYC OpenData
https://opendata.cityofnewyork.us/

Kaggle
https://www.kaggle.com

StatLib---Datasets Archive
http://lib.stat.cmu.edu/datasets/

Large Global Address Dataset (language classification, for example)
https://openaddresses.io/

UCI Machine Learning Repository
http://archive.ics.uci.edu/ml/index.php
Datasets for Data Science and Data Mining
https://www.kdnuggets.com/datasets/index.html

Federal Reserve Economic Data
https://fred.stlouisfed.org/

Repository of Big Datasets
http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free

3. Code Development

To implement the modeling and analysis, it will be necessary to develop code.
The code should be in Pyspark. In specific circumstances, python code may be admissible, but you must discuss this with the instructor first. If small parts of the code are in Python (e.g., reading into a pandas dataframe), this is fine.
No other coding languages are admissible.

The code needs to be clearly written and documented in a Jupyter notebook (ipynb format). If you have all of the code in a single cell, this is fine. ***Please clearly describe what the code does at the top of each file. Additionally, place the code's "task" in the filename.***

For full credit (10 PTS), the code needs to include these sections in a clean, commented, and comprehensive manner:

1. data import and preprocessing (2 PTS)
preprocessing include such tasks as imputing, binning, filtering, outlier treatment, feature engineering, text processing

2. data splitting / sampling (1 PT)
sampling may not be needed, but splitting is a must

3. exploratory data analysis, with at least 2 graphs (2 PTS)

4. model construction, with at least 3 models (3 PTS)
ideally the models are constructed using pipelines

5. model evaluation (2 PTS)

this should include computation of relevant metrics, and a comparison between models

4. Modeling using Machine Learning

A complete project will consider at least 2 models:

1. A benchmark model, which is relatively simple. This could be a regression model with a small number of features (possibly a single feature). This provides a basis for comparison and a sanity check.
2. A more sophisticated model, which could be one of the models covered in class. The best model found in your experiments is called the **champion** model.

The model construction process should follow the best practices covered in class, including:

a. Data preprocessing. The required steps will depend on the model, and could include:

   i. dummy variable construction
   ii. feature scaling
   iii. handling missing values and outliers
   iv. handling semi-structured / unstructured data
   v. dimensionality reduction (e.g., PCA)

b. Data splitting (train/validation/test sets, for example). The test set should be left out for evaluation purposes. It should NOT be used in training.
c. K-fold cross validation of hyperparameters

5. Model Evaluation

For all appropriate models (benchmark, champion, and other relevant models), the following should be conducted:

a. Evaluate relevant metrics

For regression, this would include
   i. R-squared (for single factor)
   ii. Adjusted R-squared (for multifactor)

For classification, this would include:

   i. accuracy
   ii. precision, recall, F1 score
   ii. confusion matrix
   iv. area under ROC curve (AUROC)

Depending on the application, additional evaluation could make sense such as lift charts

b. Sensitivity analysis (optional)

Sensitivity analysis measures the effect of changing the model inputs or parameters. For example, if the model uses a hyperparameter C, how does AUROC change when feature X is increased/decreased by one standard deviation. The hope is that sensitivity is low.

6. Project Presentation

In the final live session of the course, each team will give a group presentation to the class.
The instructor will mention the time limit of the presentation, as it depends on the number of groups presenting. It is generally 7-10 minutes per group.

I encourage each member to present a portion of the project. One of the exciting things about being a data scientist is that they can drive major change at organizations. As a consequence, they can be called upon to communicate with executives. Strong communication skills (to a technical and non-technical audience) is critical.

Components of the presentation should include:
i. Executive summary: discuss the research question and what you have found
ii. Data summary: explain the target (response) variable, the predictors, sampling, etc.
iii. Variable transformations and preprocessing
iv. Models constructed (or planned to be constructed)
v. Model performance
vi. Conclusions and future research

A presentation earning full points will be strong in:
i. content
ii. organization / aesthetics
iii. delivery

7. Project Writeup

The project writeup should include the sections below. It could make sense to divide the section writing among teammates; in that case, give the paper a final review for consistency. **The paper should be no more than 6 pages, single-spaced.** Additional tables and figures can be included in an appendix that does not count against the page limit.

Sections:

i. Abstract
Although the abstract appears first, it should be written last. This includes a quick introduction, an overview of what was done, and a summary of findings.
ii. Data and Methods
iii. Results
iv. Conclusions
The conclusions section can include future work, if there was more time.

8. Team and Teammate Evaluation

Each team member needs to make a substantial contribution, and needs to be accountable.
If a teammate issue cannot be resolved within the group, please notify the instructor. Students not contributing meaningfully to the project will not receive an A in the course.

9. Data and Modeling Recommendations

- if your dataset is > 5GB, consider taking a meaningful sample to get within size requirement.
- early on, drop fields and records that are not needed.
- categorical variables with many levels can often be bucketed effectively. Conduct EDA to understand how best to bucket.
- be sure to scale features in regression models
- when testing if the pipeline works properly, try on sample of data to save runtime
- when building tree models like GBT, start small to insure things work. This means small number of trees, shallow depth (e.g., 3). GBT and RF can take a long time to train, and are more likely to use up memory.
- when the labels are imbalanced, first split the data into train/test. Next, can change the label proportions in the train set ONLY.

10. Recommended and Non-Recommended Project Types

These projects tend to go well:

- supervised learning with text and/or quantitative data; classification, regression
- recommendation algorithms
- sentiment analysis

We advise against these types of projects:

- computer vision, as the required data is massive, hyperparameter space can be large, and there isn't a variety of data
- time series modeling, as this requires specific knowledge to do well and isn't a lot of core Spark support. You can try this but user beware.

11. Final Notes and Advice

1. If any issues come up during the course of the project, please reach out so we can address them
2. I encourage you to take on a challenging project, but if it cannot be completed in full and on time, select something simpler
3. Practice the presentation beforehand
4. Collaborate with others
5. Meet face-to-face with your teammates
6. Have fun!