PSTAT 194

Prof: Tashman

Final Project

Last updated: Oct 18, 2017

**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.

**Components**

**1. Research question**

This is the motivation for the project. Examples:

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

Give this thought, and discuss within the group and with the instructor before starting the project. 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!

StatLib---Datasets Archive

http://lib.stat.cmu.edu/datasets/

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. 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:

1. 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)

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

**4. 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

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.

**5. Project Presentation**

In the final weeks of the course, each team will have the opportunity to present their project to the class. I encourage each member 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.

**5. 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.

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.

**Final Notes and Advice**

1. It is expected that each team member will make important contributions to the project

2. If any issues come up during the course of the project, please reach out so we can address them.

3. Practice the presentation beforehand

4. Collaborate with others.

5. Have fun!