TO: 010620 Data Science Cohort
DATE: March 13, 2020
SUBJECT: Module 5 Project Guidelines

PROJECT GOAL

The goal of this project is to be able to utilize Classification modeling to answer questions that your company/stakeholder may be interested in. You will be tested on your ability to gather information from a real-world database and generate analytical insights that will be meaningful to the stakeholder.

Choosing your data

In this project, you are free to choose any dataset you like that would enable you to build a predictive model regarding information your company/stakeholder is interested in.

For example, if you were given a data set of credit card transactions, a driving question might be:
"Can I detect fraudulent transactions?"

If you're having trouble coming up with ideas, we recommend googling APIs for a subject of interest to you. You can also merge different datasets.

Stakeholders

Picking an audience at the beginning of your project helps you define the scope of the project. Once a stakeholder is picked, keep them in mind as you are conducting your analysis. When translating statistical models for a non-technical audience, be sure you are answering questions that are relevant to the stakeholder and being clear with the limitations of your findings.

Project Requirements:

Data Source

For this project, it is recommended that you gather your own data using API sources. You are required to obtain data **NOT** from any pre-cleaned data sources. Ensure that your dataset contains information that would help you build predictive models.
(Be aware of GitHub limitations with data size)

Modeling Requirements

The goal of this project is to have you complete a very common real-world predictive task in regard to Classification Modeling. However, real world problems often come with a significant degree of ambiguity, which requires you to use your knowledge of statistics and data science to think critically about and answer. You are required to use classification to build your model, however, you may try to answer more than one overarching question which would require you to use different modeling tools.

The following must be included:

- Assumptions (if any) of the model checked
- Model validation
- Model comparisons to the other models used
- Interpretation of model results
- Predictions

Visualization Requirements

EDA Visualizations

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. This is a key element of all data science projects. It is important to conduct EDA before doing any modeling so that you understand the characteristics of your data, to help with data cleaning, and to apply appropriate models.

The objectives of EDA are to:

- To find key business insights that modeling is not necessary for
- To uncover anomalies in your data to assist in data cleaning
- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions of any model you are going to use
- Support the selection of appropriate models and techniques

Model/Metric Visualizations

Classification is an area of data science that lends itself well to intuitive data visualizations. *Any findings worth mentioning in this problem are probably also worth visualizing*. Your notebook should make use of data visualizations as appropriate to make your findings obvious to any readers.

Also, remember that if a visualization is worth creating, then it's also worth taking the extra few minutes to make sure that it is easily understandable and well-formatted. When creating visualizations, make sure that they have:

- A title
- Clearly labeled X and Y axes, with appropriate scale for each
- A legend, when necessary
- No overlapping text that makes it hard to read
- An intelligent use of color--multiple lines should have different colors and/or symbols to make them easily differentiable to the eye (please, no rainbow color scheme), color should be used to represent something!
- An appropriate amount of information--avoid creating graphs that are "too busy"--for instance, don't create a line graph with 25 different lines on it

<u>Project Deliverables</u>

Your team is expected to use git as a collaborative tool for this project to manage version control and history. All documents must be contained in a git repository that you create. You should use the templates provided by instructors here.

1. **A README.md file** listing project members, goals, responsibilities, and a summary of the files in the repository. This summary should also include a guide to navigate your notebook.

2. **Multiple commits and at least one push every day**.

a. Must include short, descriptive commit messages.
b. Each project member should commit at least once.
c. Be sure to use branches to work individually and merge to master when complete.

3. **Master Notebook** - This notebook is targeted to a technical audience. The following should be present:
    a. You must **source & clean your data**. All boring stuff should be pushed to a .py file that is imported.
    b. You will work on a classification model, **compare different models** and compare their performances. Be sure that you **include justifications of these decisions** in your technical notebook.
    c. **Visualizations** to support each of your models built. *Make sure to check for any assumptions.
    d. **Clean and commented code** so an independent party can read your analysis and concur with your analytical choices.
    e. **Documentation** of where the data came from- API and any additional CSV sources.
    f. **Custom functions** should be stored in a .py file and imported whenever possible.
    g. Code should follow **PEP 8 standards.**

4. **Python files** - You should include .py files using the templates provided in your GitHub repo and the functions in them in your technical notebook. Example files may be:
    a. data_prep.py
    b. visualizations.py
    c. utils.py (for extraneous functions)

5. **Slidedeck -** You should include a PDF of your slide deck targeted at the non-technical audience in your repo. It should include:
    a. The purpose of your analysis and why it matters.
    b. A high-level overview of your data sources.
    c. Analysis of your test results.
    d. All pertinent visualizations from your analysis.
    e. Actionable insights based on the results of your classification results.
    f. Conclusions and possible future actions
    g. No more than 10 slides.
    h. No python screenshots

6. **Presentation -** Your presentation should:
    a. Be aimed at a non-technical audience
    b. Avoid technical jargon and explain results in a clear, actionable way for the audience.
    c. Contain between 5-10 professional quality slides.
    d. Include a high-level overview of your methodology and findings, such as including the fraud results for the example above.
    e. A brief explanation of what metrics you defined as "best" in order complete this project
    f. Take no more than 5 minutes to present

Project schedule:

03/13 Friday Afternoon - Project Assignment

- Start brainstorming potential questions
- Begin researching data sources of interest
- Schedule Monday check-in with coaches

03/16 Monday Morning - Send in a project proposal (via Google Forms):

- The problem you are trying to solve
- Data sources
- Preliminary EDA
- The types of model you plan on employing
- Plan for how the team will divide the work.

03/16 Monday Afternoon - Check in with coaches to review the agenda:

- Models used
- Model validation
- Have a version of master notebook completed

03/17 Tuesday Morning - Demo presentation with feedback from instructors based on agenda:

- Have a draft of deck completed
- Have a version of master notebook completed

03/18 Wednesday Afternoon - Presentations

- Afternoon project presentation to the class
- This will be done via Zoom

If any requirements are missing or if significant gaps in understanding are uncovered, be prepared to do one or all of the following:

- Perform additional data cleanup, visualization, and/or feature selection
- Submit an improved version
- Meet again for another project presentation

What won't happen:

- You won't be yelled at, belittled, or scolded
- You won't be put on the spot without support
- There's nothing you can do to instantly fail or blow it