Project 3

For this project you can work alone or work with a partner. See wolfware for more details about letting me know if you are working with a partner.

You'll be tasked with the following:

- Finding a data set you can explore and fit supervised learning models with
- Conducting EDA through pandas-on-spark and/or spark SQL
- Fitting three different classes of models and choosing a best model
- Writing a narrative (via a notebook) with some explanations and discussions as you go through the above. (See below for details.)

Report Components

Introduction

You should discuss the goals of the notebook, introduce your data set, and give the source for your data set.

Supervised Learning Idea and Data Split

Give a discussion as to why we want to what we are generally trying to do with supervised learning where prediction is our goal. Discuss why we want to split our data into a training and test set.

You should also split the data into a training and test set.

EDA

You should have a narrative that goes through what you are trying to accomplish in the EDA, why you are looking at a particular graph or statistic, and how you interpret what you've made. The EDA should be done on the training data only. You should use pandas-on-spark or spark SQL data frames (but `matplotlib` is fine)

Part of the final's purpose is to see if you can judge what should and shouldn't be included in an EDA.

Modeling

Next, you should fit three different classes of models (they can be the ones we did in class or you can branch out). You can have a numeric response or a binary response.

With each model type you use, you should describe the general idea of the model/how it works. These discussion don't need to be super long, but they should be clear and hit on the most important points about how the model works.

You should use CV to choose among the candidate models for each model type.

- You should set up a pipeline in `pyspark` for each of your models
- At least one of the pipelines should include at least two transformations prior to the model fit (`estimator`)
- You can use the same set of transformations for multiple models (if appropriate)

If working with a partner, I'd recommend going through the process of data transformations and model fit together for one model type. Then each take another model type and modify for that case.

Lastly, you should evaluate the best models on the test set and state which overall model was deemed the best.

Note: If you are unable to get the modeling to work in pyspark, you can do everything via sklearn. However, you would then earn a maximum of 25 points in this section.

Rubric for Grading (total = 100 points)

Item	Points	Notes
Introduction	15	Worth either 0, 3, 6, . . . , 15
Supervised Learning and Data Split	15	Worth either 0, 3, 6, . . . , 15
EDA	30	Worth either 0, 5, 10, . . . , 30
Modeling	40	Worth either 0, 5, 10, . . . , 40

Notes on grading:

- For each item in the rubric, your grade will be lowered one level for each each error (syntax, logical, or other) in the code and for each required item that is missing or lacking a description.
- **You should use Good Programming Practices when coding (see wolfware). If you do not follow GPP you can lose up to 25 points on the project.**

The reports should include a narrative throughout, section headings, graphs outputted in appropriate places, etc. To be clear **be sure to include markdown text describing what you are doing, even when not explicitly asked for!** Points will be deducted from appropriate sections as appropriate.