Project Proposal Assignment

Contents

[Assignment Goals 1](#_Toc96372623)

[Overall Requirements 1](#_Toc96372624)

[Project Proposal Section Requirements 2](#_Toc96372625)

[Grading Rubric 3](#_Toc96372626)

[Submission 3](#_Toc96372627)

# Assignment Goals

A preliminary summary of the project in a Microsoft Word document format. Note that by the time you write your proposal, you should have worked with the data set you plan to use for your project. If you change your data set, your proposal grade will be changed to a 0. If you have to change your data set, it shows that you didn’t really work with the data and think about your project before you wrote your proposal. For example, you originally decided to do image recognition but after you submit your proposal, you find that the images in the data set you chose are too large for your compute environment. In this case, you can change your data set but you will get a 0 grade on the proposal. Minor changes, additions, deletions, changes in direction are allowed. The goal here is to make sure that students work with the data before making a final selection.

# Overall Requirements

* The teaching staff is aware that many data sets on the internet have sample notebooks available on Kaggle. I always instruct the class faculty assistant to review Kaggle notebooks which might be available for each project team and compare to what the project teams submitted. We expect students to form their own business / scientific questions, do their own visualizations and data exploration, and write their own code. Notebooks that are found to largely copy what was already done in a Kaggle notebook will be heavily penalized up to receiving a zero grade for the project.
* Shall be a maximum of 3 pages
* Shall not contain python code.
* Line spacing shall be 1.5 lines.
* All plots and figures shall have titles, axis labels, and a brief description.
* The project team shall not be greater than 4 members
* Shall be a Microsoft Word Document
* High quality writing is expected with good grammar and punctuation. The document should be thoroughly proof-read and reviewed by all team members prior to submission.
* All machine learning, inference, analysis, data wrangling, and data exploration shall use spark. Visualizations can be performed using any python visualization package. In short, the data science workflow shall be performed using spark. For example, don’t pick a project that requires convolutional neural networks because convolutional neural networks are not included in spark. Don’t pick a project where the focus of your project depends on packages outside of spark. Essentially, your projects should be focused on topics learned in class using tools used in class. An example of an acceptable use of a package outside of spark is using an NLTK package to label a data set with positive or negative sentiment for the purposes of training a natural language processing model in spark. In this example, the NLTK package is being used to support your work in spark and is not the focus of your project. In summary, **the focus of your project shall be to apply spark to solve a business / scientific problem**. If you have a doubt, consult with the teaching staff. Projects that do not use spark as the primary tool will not be accepted and receive a zero grade.
* All projects shall run on Colab. If your project requires computing resources outside of Colab then it is not an appropriate project, will not be accepted, and receive a zero grade.
* The assignment shall be submitted by one (and only one) member of the team.
* The data set shall be readily available for public access and shall not require the graders to sign a non-disclosure agreement. Basically, we need to be able to download the data and the project code onto our own computers with absolutely no restrictions whatsoever.
* Note that for the final project, all data wrangling / cleaning must be performed in Spark so it is best to start out using spark during the proposal phase.
* Stock market projects are highly discouraged – especially projects that predict stock market prices. If you really want to do a stock market project please consult with the teaching staff.

# Project Proposal Section Requirements

The project proposal is a road map on how you will execute your project and **shall include the following sections** (and **only** the following sections) **in the order given** in the bulleted list below. Each bolded item in the list below is considered a required element of your project proposal.

* **Title page** including project title, group number, and all group member names.
* **Objective**: A concise, well written high-level objective of the project. Briefly describe what business or scientific problem you are trying to solve.
* **Data Set Description**:
  + Overview / Description
  + Number of rows and cols
  + Sample predictors (does not need to be an exhaustive list)
  + A link to the dataset
  + Anything interesting or surprising about the data
* **Preliminary Data Exploration:** Explore the data, provide brief summary statistics and visualizations. Tell me something interesting and or surprising about the data. If you need to reduce the size of your data to fit on Colab, now is the time to do that. The work that you do in this section of the proposal can be directly used in your project report.
* **Proposed Data Exploration Insights**

All projects shall include a bulleted or numbered list of proposed specific data exploration insights. These insights are things that can be determined by examining the data through statistical analysis or other means.

* + Example data exploration insight for a group using an airline flight delay dataset: We plan on examining which destinations are most likely to have delays by calculating average delay and frequency of delay for each destination and creating bar plots to summarize the average and frequency data.
* **Proposed Predictions**

All projects shall include a prediction component.

* + A bulleted or numbered list of specific proposed predictions you are trying to make. The list of items should be large enough to keep everyone on your team busy. Your predictions need to be as specific as possible. Don’t make general statements like we plan to make predictions which will aid business leaders; but rather, be very specific about the predictions. For example, say “We plan to build models to predict the monthly dollar sales to aid business leaders in making decisions.”
  + Try to think about interesting things to make predictions on. For example, if you are making Uber / Lyft fare price predictions, don’t just say you will try to predict the fare price; but rather, think about other things on which you can make predictions. For example, if there is a price multiplier column, try to predict the price multiplier and then use that model for inference on what predictors are most important to price changes. Or, think about how you might stratify the data to make predictions along common source / destination routes for the purpose of making a model focused on a more specific area of the data. The more specific models can then be used for more specific inference.
* **Proposed Model Inference Insights**

All projects shall include a model inference component. Model inference insights are things you intend to explore by using trained model parameters to learn something useful about the problem you are trying to solve.

* + Provide a bulleted or numbered list of specific model inference insights you would like to achieve in your project. Don’t say you plan to examine model parameters to gain insight; but rather, be specific about the insights you plan to explore and how those insights add value to the project.
  + For each insight, indicate exactly how your trained model will be used for inference.
  + Example poorly written inference item for a data set using Uber / Lyft data from an actual proposal turned in by a student group: Does fare vary with weather conditions.
  + Example well written inference item: Using our linear regression model which predicts fare, we intend to compare how each of the predictor variables contribute to the overall predicted fare. In addition, we will explore which predictor values have a positive effect on fare and which predictor values have a negative effect on fare.
  + Another example well written inference item: Using models which predict fares based on data stratified by start and end locations, we plan to examine how each of the predictor variables contribute to the resulting predicted fare. We will then compare the most important predictors across stratified start and end locations and see if different start and end locations have common or differing most important predictors.
* **Non Spark Packages:** Include an exhaustive list of packages outside of spark. You do not need to include numpy, pandas, matplotlib, and graphing packages in your list. For each package, state how the package relates to your project. Make sure that you understand that packages outside of spark, numpy, and pandas are only allowed to support work done in spark. Your project must be a spark focused project.

# Grading Rubric

* 10% Following instructions. If you do not follow instructions, you will likely not get an A grade. **For example, if multiple team members make submissions, don’t include the required sections, add new sections that are not in the list, submit in a format other than Microsoft Word, you will not get an A grade on your proposal.**
* 10% Writing quality: High quality writing is expected.
* 20% Project Objective:
* 20% Data set description
* 10% Preliminary Data Exploration
* 10 % Proposed Data Exploration Insights
* 10% Prediction
* 10% Inference

# Submission

* Submit your proposal by adding your project proposal to the github repo. It does not matter if you overwrite this document or add a new document to the repo.
* Only one submission is expected.
* All group members will receive the same grade for the submission.