Project Assignment

# Report Requirements

* Shall be a Microsoft Word Document.
* Spacing shall be 1.5 lines
* The teaching staff is aware that many data sets on the internet have sample notebooks available on Kaggle and other sites. 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.
* The project code and report are due at the same time and shall be submitted together. See submission instructions below.
* The maximum number of pages excluding the title page, table of contents, and appendix is 15 pages. Due to the large number of groups this semester, we intend to strictly enforce a grading penalty if the number of pages exceeds 15 (excluding title, TOC, and appendix) of 5 points off for every page over 15. We have a hard deadline to submit grades to the university and need to limit the scope of submissions in order to meet our grading timeline as set forth by the University. Sometimes it’s hard to decide what to cut from a report but oftentimes it makes for a better report because it forces the writers to really think about what adds the most value to the report.
* In general, the best reports present most data in terms of visualizations and tables followed by brief descriptions of the tables and visualizations. Be brief, targeted, and concise with words and liberal with tables and visualizations.
* Figures and tables shall be inline with the text. Don’t fill up the appendix with tables and figures and make the reader jump back and forth between the appendix and text. This will be penalized heavily.
* Sections defined in the report sections table below shall be present in the document. You are free to add sub sections but the major sections in the table must be included.
* Include technical references you used for your report in the appendix and link to your references from the main document.
* Sections shall be numbered. Major sections shall be numbered 1, 2, 3. Sub sections shall be numbered 1.1, 1.2, etc. I’m basically asking you to use the Microsoft Word built in heading number capability.
* High quality writing is expected with good grammar and punctuation. The document should be thoroughly proof-read and reviewed by the team prior to submission.
* Provide visualizations throughout the report which can be any form of plots, tables, graphics, etc.
* All projects shall include a data exploration component.
* All projects shall include an inference component. The inference component is defined as using trained models to gain insight about the problem you are trying to solve. Example: Comparing regression coefficients together to determine feature importance. Another example, using feature importance list obtained from a Random Forest model.
* All projects shall use multiple machine learning model types.
* All projects shall use spark as the main tool. **Projects that do not use spark as the main tool will not be accepted.** You may use tools outside of spark to support spark. A good example of using code outside of spark to support spark is using NLTK to label tweet data with positive and negative sentiment labels so you can train sentiment prediction models in spark. If in doubt, discuss with the professor. Essentially, your project should be 90 to 95% spark.
* All code shall run on Google Colab
* You are allowed to install packages you need on Colab. Installations must be performed via code written by you in your notebook.
* All plots, figures and tables shall have titles and axis labels
* All plots, figures and tables shall have (brief) text describing the figures and tables
* Putting code in the report is strongly discouraged.
* Do not waste space in your report describing how things work which we covered in class. For example, you don’t need to describe how random forest works, or how regularization works.
* Do describe how things work that were not covered in class. For example, if you decided to use some technique that is totally outside the scope of the class, it would be a good idea to provide some brief background information and reference information on this topic. Keep in mind that the grading team has a large workload and we don’t have hours of time to spend learning and researching new techniques while grading. You need to present a concise high-level overview that clearly explains the high-level points. If you don’t do a good job presenting it and we don’t understand it, it will show up in your grade.
* Remember that grading papers and code is a subjective process. We compare the level of effort and quality of your work to other project groups in the class before assigning final grades. If your work is significantly better or worse than other projects, you will be graded accordingly. For projects that are significantly better than other projects in the class I occasionally award more than 100 points.
* **Only** the major sections in the following table shall be included in the report. You are allowed to add sub sections.

| **Report Sections** | |
| --- | --- |
| **Item** | **Description** |
| Title Page | Include the name of your project, the team members who contributed to the project, and your assigned project group number (example: group7). |
| Table of Contents | A Microsoft Word auto generated table of contents for the entire report which includes the section numbers. |
| Abstract | * Provide a high-level overview of your project * Brief data set description * Link to data set * Provide a bulleted or numbered of interesting / surprising data exploration insights you found * Provide a bulleted or numbered list of the specific predictions you made * Provide a bulleted or numbered list of the specific inferences you made. * Provide a brief conclusion summary outlining the success / failure of your predictions and inference * 1 to 2 pages is advised for this section |
| Data Collection/ Cleaning | * Provide brief details about what you had to do to clean the data. For example, did the data have a lot of NA values. Did you drop whole columns of data or whole rows of data. How did you handle missing values? * 1 page is advised for this section |
| Data Exploration Insights Using Standard Statistical Techniques – Excluding Machine Learning | * Describe the results of your data exploration. * Provide some data exploration visualizations. * Tell me something interesting you learned about the data. * Maybe provide some statistical summaries * Maybe look at feature correlation * Example data exploration insight for a group using an airline flight delay dataset: We found that there was a higher frequency of delayed flights which originated at Chicago O’Hare airport as compared to other airports. We found evidence that these delays were caused mainly by … * Do not include model inference in this section. Machine learning model inference goes in it’s own section described below. * 2 to 3 pages are advised for this section |
| Methodology | * Provide a brief description of the methodology used to achieve the goals of your project. For example, if you are performing sentiment analysis on tweet data, describe the high-level methodology of your data science workflow to get your data ready to train models like stop word removal, word count, IDF, etc. * Describe feature engineering / data transformations that are common to all of the models described in the following sections? Example, you one hot encoded categorical variables that are used in all following models. Or you reduced dimensionality by using PCA and trained all the following models using PCA scores. * Describe the distribution of outcomes. Is your dataset imbalanced or balanced and what is the distribution of outcomes. Describe any sampling you may have done to balance your data set. * Describe your model scoring methodology. Example 1: If you are making predictions on an unbalanced data set, describe how you adjusted your scoring methodology to account for the unbalanced outcomes. If you are using a simple scoring metric then describe that. Example 2: For model predictions we are using AUC to score the model’s ability to accurately predict delayed flights. * If you are using tools outside of spark this is a good place to describe how you are using that technology. Do not say anything about seaborn, matplotlib, numpy, pandas, etc. Only talk about uncommon tools that we have not used in class to support your project. * Feel free to use block diagrams / tables / diagrams to describe your workflow. Note that I am not expecting you to present some off the shelf data science workflow; but rather, I want to know specifically what you needed to do on your project to get the ready for training your models. * 1 to 2 pages is advised for this section |
| Model Prediction | * A good technique for this section is to put each model in its own sub section with a descriptive title like Linear Regression Prediction of Flight Delay. A lot of the requested information can be efficiently conveyed with a table (the same general table can be used for each model). Or provide a combined summary table and text that describes the combined summary table. * For each summary table, include the following:   + Value being predicted.   + Model type: Random Forest, Linear Regression, neural network   + Scoring metric: F1, AUC, MSE, RMSE, etc.   + If cross validation was used. Number of folds if used   + Resulting Model Score * Provide brief text explanations for each table. Describe things like transformations or feature engineering that was not common to all models and not described in the methodology section. Describe things particular to this model that needed to be performed in order to meet your goals. * **There is no need to overly complicate things in this section. Don’t make things more complicated than they need to be.** * 2 to 3 pages are advised for this section |
| Model Inference | * Same rules as for prediction above: A good technique for this section is to put each model in its own sub section with a descriptive title like Linear Regression Inference of Flight Delay. A lot of the requested information can be efficiently conveyed with a table (the same general table can be used for each model). Or provide a combined summary table and text that describes the combined summary table. * For each summary table, include the following in some sorted order:   + Feature Name   + Feature metric being compared: Example: Regression coefficient, gini index, entropy, etc.   + Scoring metric: F1, AUC, MSE, RMSE, etc.   + If cross validation was used indicate that fact and the number of folds used to score your model. * For each table, provide a brief text explanation which highlights the following. Provide a bullet or numbered list for each of the following:   + The inference goal   + What you learned through your analysis   + Any special transformations / feature engineering you had to perform in order to correctly perform the inference. For example, transformed the data using a standard scaler so regression coefficients representing real data could be compared between each other.   + It is encouraged to stick with inference techniques covered in class.   + **There is no need to overly complicate things in this section. Don’t make things more complicated than they need to be.** * 2 to 3 pages are advised for this section |
| Conclusion | * Describe the results of the project by expanding upon the brief conclusion results outlined in the abstract. Summarize model comparison results in a prediction summary table. Summarize inference results in an inference summary table. Sort the summary tables. Use table column names which are similar to columns in previous sections. Maybe add sub sections for prediction and inference summaries. Provide results for every prediction and inference item presented in the abstract. * 1 to 2 pages is advised for this section |
| Appendix | * Include references you used. For example, if there are specific web sites or books you used, include them here and link to your references from the document. * You can include any additional information you consider important that might not be appropriate or too large to include in the main document. * Break the appendix up into sub sections as you see fit * The appendix is not the place to put lots of tables and figures that didn’t fit in the 15 pages you were allotted for the report text. |

# Report Grading (100 pts)

| **Report Grading Rubric**  **(All categories weighted equally)** | |
| --- | --- |
| **Item** | **Description** |
| Writing Quality | * The document is expected to be well written and edited. |
| Following Instructions | * I use following instructions as one of the main distinguishing factors between reports that get an A and reports that get less than an A. * Read and follow the project requirements. * **Projects that do not use spark as the primary data science tool will not be accepted.** |
| Abstract | * Clearly defined project overview. * Well defined prediction list * Well defined inference list * A concise summary of results related to the specific inference, prediction, and other lists. |
| Data Collection / Cleaning | * Adequate description of data wrangling like how NA and null values were handled. |
| Data Exploration Insights | * Well defined insight conclusions which are supported by data, tables, figures, and plots. * Relevant statistical summaries with insightful descriptions |
| Methodology | * Well defined and understandable descriptions of how the data was manipulated to get ready for training. * Common transformations / feature engineering accurately and concisely described. |
| Predictions | * Required data is thoughtfully presented in an easy to read and understandable format including the required tables. |
| Inference | * Required data is thoughtfully presented in an easy to read and understandable format including the required tables. |
| Conclusion | * All predictions included * All inference included * Summary tables are included * Valid conclusions are made based on the results of the previous sections. |

# Code Requirements

* You are required to download data files from either your group’s github repo, AWS, or directly from the internet.
* The code is expected to be well organized and well commented so that it’s easy to follow. Try to break the code up into logical sections that perform major tasks. For example, perform all data wrangling in one section, develop machine learning models in their own sub sections, inference in different sub sections, etc.
* All code shall run on Colab without error.
* Models shall be trained using grid search where applicable. Linear / logistic regression models do not necessarily have to be trained with grid search if it is not needed for your application.
* All data wrangling / cleaning, machine learning, inference, and analysis shall use spark. In short, the main data science work-flow shall be performed using spark. Plots and visualizations can be performed outside of spark.

# Code Grading (50 pts)

| **Project Code Grading Rubric** | | |
| --- | --- | --- |
| **Item** | **Description** | **Weight %** |
| Following Instructions | * Grid search present for models other than regression or for regression where needed. * Data located in a “data\_files” directory. * data\_files Directory is located at top level of git repo. | 20 |
| Subjective Grading | * Code effectively split up into modules: * Effective use of comments. Is the code easy to follow. * Does the code correlate with the information in the report * Does the code run without error on Colab * **Projects that do not use spark as the main data science tool will not be accepted.** | 80 |

# Submission:

* Data files are submitted in your github repo in a directory named data\_files.
* All code is submitted in your github repo in a directory named “code”.
* Your project report is submitted as a Microsoft Word file at the top-level directory of your github repo. The name of the report file shall be named project\_report\_groupN.docx where the ‘N’ indicates your group number. Example: project\_report\_group1.docx.