Project Assignment

# Report Requirements

* The project code and report are due at the same time and shall be submitted together. See submission instructions below.
* Recommended 10 to 12 pages maximum excluding table of contents, title page, and appendix. You can go over a little (like 1 or 2 pages) but reports that are significantly over (defined by me while I am grading) will be penalized. Its also fine to be a little under 10 pages as long as you meet the report requirements.
* Spacing shall be 1.5 lines
* Shall be a Microsoft Word Document.
* 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.
* All projects shall use multiple machine learning models.
* All projects shall use spark as the main tool. **Projects that do not use spark as the main tool shall receive a maximum grade of C-.** 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 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. The professor compares 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 * Provide a list of the specific predictions you made * Provide a list of the specific inferences you made. In this context, inference is using parameters of a trained model to gain insight about the problems you’re trying to solve. * Provide a brief conclusion summary outlining the success / failure of your predictions and inference * Provide a list of other goals your project may have. * The abstract should be no more than 1 to 2 pages. |
| Data Collection/ Cleaning / Exploration | * An expanded dataset description from what you provided in the project proposal. * If you collected the data by scraping a web site, describe that here. * Describe the results of your data exploration. * Provide some data exploration visualizations. * Tell me something interesting you learned about the data. * Provide brief details about what you had to do to get the data ready for training your models. For example, did the data have a lot of NA values. Did you end up dropping columns of data. Did you do feature engineering. Did you use dimensionality reduction. |
| 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 work flow to go from post data wrangled text from the previous section to trained models. Block diagrams work well here. Be brief with words and liberal with diagrams and figures. Some brief text along with a block diagram or 2 should be fine. |
| Models | * For each model:   + Describe what the model is predicting.   + Indicate the model type: Random Forest, Linear Regression, etc.   + Briefly describe any special data transformations required for this model. Example: Performed PCA transformation and used the first 20 components which described 80% of the data variance.   + Briefly describe how you scored the models (Example, MSE, AUC, F1, etc). Describe any special scoring goals and how you achieved the goals. Example, you were trying to minimize false negatives.   + Briefly describe any inference you performed using the model * A good technique for this section is to put each model in its own sub section. Briefly describe the items above in words. Follow up with one or more tables that summarize the prediction and inference results. The goal is to provide brief text and summarize your results in easy to read tables. |
| Conclusion | Describe the results of the project by expanding upon the brief conclusion results outlined in the abstract. Summarize model comparison results. Summarize inference results. Provide results for every prediction and inference item presented in the abstract. A good methodology for this section is to be brief with words and liberal with summary tables. |

# Report Grading (100 pts)

| **Report Grading Rubric**  **(All categories weighted equally)** | |
| --- | --- |
| **Item** | **Description** |
| Writing Quality / Following Instructions | * I use following instructions as one of the main distinguishing factors between reports that get an A and reports that do not get less than an A. * Project report file name was not changed. * Read and follow the project requirements. * **Projects that do not use spark as the primary data science tool will earn a maximum of a C- grade.** |
| Abstract | * Clearly defined project overview. * Prediction list * Inference list * Other list * Predictions and inference are targeted and specific. Don’t say we are going to predict things and stuff. * A concise summary of results related to the specific inference, prediction, and other lists. |
| Data Collection / Cleaning / Exploration | * See requirements above |
| Methodology | * See requirements above |
| Models | * See requirements above |
| Conclusion | * All predictions included * All inference included * All other included * See requirements above |
| 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 |

# Code Requirements

* All data files shall be contained inside a directory named **data\_groupN** in your github repo submission where the groupN part of the directory name is your assigned group name you used to clone your repo. For example, if you are in group 12, your directory name would be **data\_group12**. This directory should be at the top level of your git repo.
* Your project must load data from Google drive. You must follow the data loading example provided in the Colab Data Loading Example notebook distributed with the assignment. We expect your data to be located in a very specific directory named **/content/drive/My Drive/ist718\_data/data\_groupN**. The example notebook includes all needed detail. We are requiring this naming format so we do not have to modify your code to run it during grading.
* 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 a different section, etc.
* All code shall run on Colab without error.
* Models shall be trained using grid search.
* Data wrangling, and exploration can be performed using Pandas. All machine learning, inference, and analysis shall use spark. In short, the main data science work-flow shall be performed using spark.

# Code Grading (50 pts)

| **Project Code Grading Rubric** | | |
| --- | --- | --- |
| **Item** | **Description** | **Weight %** |
| Following Instructions | * Grid search present * Data located in a “data\_groupN” directory where **groupN** is your group number. * Data directory located at top level of git repo. * Data is loaded from a directory name using the following general format: **/content/drive/My Drive/ist718\_data/data\_groupN** * **The following instructions grade will be applied as all or nothing. If you follow all instructions, you get full credit. If you don’t follow all instructions, you get 0 points.** | 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 receive an overall maximum grade of C-.** | 80 |

# Submission:

* You will submit by adding all code and data files to your group github repo following the instructions outlined above.
* You shall use the supplied empty project report word file. You shall not change the name of the project report file. You will not add a new project report file to the repo. We will be looking for the specific project report file name during grading.