|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Homework overview |

The homework assigned in this class is a blend of machine learning fundamentals, numerical analysis, and research skills that is designed to reinforce the course learning objectives. Homeworks 4-6 and the take-home final will be iterations on a research and analysis paper & presentation that will benefit your organization.

A table of contents for the homework is shown below, and there are reference materials in the appendices.

[Homework #1 2](#_Toc137278740)

[Homework #2 3](#_Toc137278741)

[Homework #3 6](#_Toc137278742)

[Homework #4 8](#_Toc137278743)

[Homework #5 10](#_Toc137278744)

[Homework #6 12](#_Toc137278745)

[Final Project 14](#_Toc137278746)

[Appendix A – IEEE Citation Guide 16](#_Toc137278747)

[Appendix B – Guide to Formal Writing 17](#_Toc137278748)

[Appendix C – Bibliography Instructions 19](#_Toc137278749)

[Appendix D – Presentation Rubric 20](#_Toc137278750)

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Homework #1 |

This assignment will help me learn more about you, and also assist in further developing your research skills. In all paragraphs, strive to write in a formal writing style based on the guidance in Appendix B.

1. Write one paragraph introducing yourself, including your education, 5 years of work history, the base you work at, and your current role in your organization.
2. Write one paragraph describing what parts of the Data Analytics certificate have been the most challenging for you, and one paragraph discussing what parts have been the most enjoyable.
3. Write one paragraph describing what aspect of machine learning you’re most looking forward to learning about, and in the same paragraph list two ideas of how machine learning could benefit your organization. The two ideas won’t be judged on their feasibility, and could help your selection of a machine learning topic for HW 4-6 and the final project.
4. Add citations to this paragraph using the instructions below. “The CRISP-DM process contains the iterative steps of data preparation, data understanding, modeling, evaluation & deployment [1]. Several machine learning topics are detailed in the literature [2, 3]. Additionally, (insert a fact or opinion about machine learning) [4].”
   1. Create citations in the IEEE format using the Word bibliography tool. The IEEE format is shown in Appendix A, and instructions on using the Word tool are located in Appendix C. Using the Word bibliography tool, create a bibliography at the end of your document.
   2. Make sure the citations are inserted & auto-generated by the tool, not just typed in like this [1].
   3. Citation 1: IBM SPSS Modeler CRISP-DM Guide from DASC 500
   4. Citation 2, 3: Our two textbooks
   5. Citation 4: A webpage related to machine learning
5. Submit a Word document that contains your assignment.

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Homework #2 |

In this assignment you will thoroughly evaluate a classification problem.

1. Data: The “LRS\_Pre\_Assessment\_trimmed\_rank.csv” physical fitness testing dataset is provided for use in this exercise. The data was collected from an active-duty squadron and the samples were deidentified at the point of collection.
2. The independent variables include numeric and categorical data related to demographics, mental health surveys, fitness participation surveys, injury history surveys, physical performance measures, and body composition assessments.
3. The dependent variable is whether or not the member passed their fitness test, and is titled *APFT\_1\_is\_pass*. For this label, pass = 1, and fail = 0.
4. Modeling: In a marked-up Jupyter notebook (\*.ipynb):
   1. Break your code into logical chunks, using multiple “text” and “code” sections, similar to the examples given in class.
   2. Drop the “flight” column and one-hot-encode the “rank” & “gender” columns with df\_ohe = pd.get\_dummies(df, drop\_first=True). *drop\_first* is needed so the columns are linearly independent.
   3. Create a Data Understanding table using .describe() and include 3 Data Understanding visualizations such as a scatterplot, histogram, pairplot or correlation matrix.
   4. Split df\_ohe into 70% train & 30% test
   5. Model #1: Train a statsmodels sm.Logit algorithm on the train dataset.
   6. Create a copy of the dataframe using df\_xform = df\_ohe.copy(deep=True)
   7. On df\_xform, perform a box-cox transform on the Age and ORS\_total features, and a yeo-johnson transform on the PTSD\_score feature. Refer to the week 3 “DASC 522 guide to feature standardization & transformation” for assistance.
   8. On df\_xform, standardize the following features using StandardScaler():
      1. Sleep\_score
      2. survey\_Burnout
      3. survey\_FitSat
      4. BMI
      5. BodyFatPerc
      6. MusclePerc
   9. Display df\_xform’s statistics using df\_xform.describe(), and verify the transformed features have mean~0 & standard deviation~1
   10. Model #2: split df\_xform into train/test datasets, and train a statsmodels sm.Logit algorithm on the train dataset.
   11. Model #3: using the training dataset from df\_xform, perform feature selection using p-value selection.
       1. Use the P>|z| values from the model summary, not the LLR p-value.
       2. Remove features one at a time until all remaining features have p < 0.05, which is 95% confidence they are statistically significant.
5. Metrics & overfitting
   1. Calculate the accuracy (on the test dataset) for Model 1/2/3, and use that to determine the “best” variation. The test dataset is critical to detecting overfitting on the model created from the train dataset, so use the model created from the train dataset. It will make me sad if you make a new model from the test dataset.
6. For the “best” model, using the train dataset, create a ROC curve plot and calculate the accuracy, odds ratio, AUC, classification report and confusion matrix.
   * 1. While the course example uses the LogisticRegression algorithm, use sm.Logit to calculate AUC (above) & ROC (below). This will require some investigation as the prediction methods differ.
7. For the “best” model, using the test dataset, create a ROC curve plot and calculate the accuracy, AUC, classification report and confusion matrix. The odds ratio does not need to be recalculated.
8. Discussion
9. Include a text block related to Business/Mission Understanding:
   * 1. Review the “Binary Classification metric summary” file from the week 2 Canvas page.
     2. Mention what the majority class is (either passing or failing the test), the % of datapoints in the majority class and whether or not the dataset is balanced.
     3. Discuss the penalty (if any) associated with a False Negative and False Positive. Mention if the penalty is equal or unequal.
     4. Discuss the metrics (accuracy/f1/etc) that would be most appropriate for this problem based on the balance & penalties
10. Include a text block relating to the best model:
    * 1. Discuss the most important 2-3 input variables, based on their z or p test scores.
      2. Include a discussion of the “most appropriate” performance metrics from part 2.f. Based on the train & test dataset performance, mention if the model has overfit the data
      3. Write in a formal writing style based on the Appendix B guidance, with the exception that references and citations are not required.
11. Upload your .ipynb python file to Canvas as your homework submission.

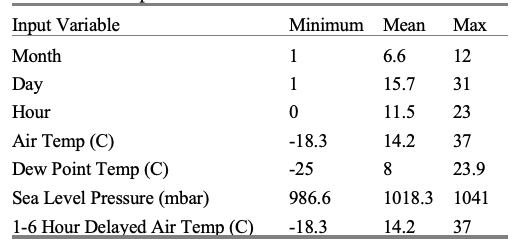
|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Homework #3 |

In this assignment you will perform manual & automated feature selection on data accessed from a database. Much of the world’s data resides in databases, and extracting this information for data analytics is an important skill. This will also reinforce your DASC 501 Database skills. In this assignment you’ll notice there is not a request to standardize, normalize, transform or split the dataset. While these are best practices for all projects, they were omitted in order to focus on databases and feature selection.

A video tutorial and sample code that will be useful in this homework are posted on the Syllabus tab of Canvas:

* Tutorial video (mp4)
* Loading database tutorial python file
* Sample database file
* Feature selection examples python file

The dataset for this assignment is from a prior Data Analytics student capstone project, from the 14th Weather Squadron. It contains 43,487 surface weather observations from 5 years of data at a single location, with the features shown in the table below. The last row is 6 features, for a 1 hour delay, 2 hour delay etc. The goal of the capstone project was to perform a regression analysis to predict the current temperature “Air Temp (C)” by using the previous six hours of temperature and other factors. The application is for quality control prior to entering the temperatures into an authoritative database. If the model prediction differed significantly from the new temperature it could be an error, and could be flagged for verification.



Here are the tasks for this assignment

1. In a marked-up Python .ipynb file, create code to extract data from two tables (X\_table & y\_table) in the “weather\_data.db” SQL database file. Load them into two dataframes df\_X and dy\_y. The tutorial covers this procedure.
   1. Use the df.join() class method to merge the two dataframes on the “Record\_ID” column, into a dataframe called df\_new. Then, drop the “Record\_ID” column.
   2. Split df\_new into X & y dataframes, where y only contains the AIRTEMPERATURE feature.
2. In this section you will perform data understanding, and create 4 families of models. You may reuse any of the code in the “Load database tutorial” and “Week 4A feature…” files included in the homework.
3. Create a Data Understanding table & two python Data Understanding visualizations such as a scatterplot, histogram, pairplot or correlation matrix.
4. Full model: Model all variables using ols regression, and display a model summary.
   1. p-test selection: Use the p-test results to manually select the best variables, similar to part 2.k of Homework 2. Display a model summary of the best p-test model.
   2. AIC selection: Use the AIC analysis code from the “Analysis – AIC & BIC” section of the “4 Week 4A feature…” file to create a set of models.
      1. Use the “AIC Analysis” code to find the best model using AIC. This takes about 3 minutes if you use Colab.
      2. Display a model summary of the best AIC model.
   3. RFE selection: Use the “RFE iteration” code from the “Week 4A feature…” file to find the minimum set of features that achieves score = 0.98. This is equivalent to R2 = 0.98.
      1. Display a model summary of the best RFE model.
      2. In a code comment, mention which features to not contribute to improving model performance
5. In your Python file, also:
6. Enhance readability by occasionally adding text blocks above your code blocks, similar to the course examples.
7. Include a 200+ word evaluation of the modeling efforts at the end of the file. Be sure to discuss the differences in the variables selected by the different methods.
8. For part 3.b, write in a formal writing style based on the guidance in Appendix B. References are not required.
9. Submit
10. Upload your .ipynb file to Canvas as your homework submission.

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Homework #4 |

In this assignment you will start your capstone project that will be iterated upon for the remainder of the course. You will perform a classical modeling analysis on a problem of interest, and compare its performance to a neural network approach.

1. Select a topic of interest to your organization that has data available, or data that can be obtained a moderate amount of effort. The topic dataset should have 3 or more variables, and should be labeled. Talk to the instructor if you are having difficulty identifying a dataset. If your prior certificate assignments are related to your organization, you may build upon that topic.
2. Using the Word or LaTeX journal template uploaded to the Syllabus section of Canvas, add a descriptive title and write 1000 words of CRISP-DM Business Understanding (BU) and Data Understanding (DU) on your topic.
   1. 1000 words meets the objective, and be sure not to exceed 1500 words for homework #4. The split between BU & DU is up to you.
   2. Write using a formal writing style based on the Appendix B guidance.
   3. Your target audience are machine learning colleagues that are familiar with statistics and machine learning, but not experts in your career field. For example, you don’t need to explain regression.
3. In your Business Understanding section include:
   1. Problem motivation
   2. Mission understanding, and background needed to understand the problem.
   3. Your research question. This could be similar to “Using available data, how well can classical machine learning and neural network algorithms predict (describe output)?
   4. Your success criteria, which could include performance or the ability to generalize.
   5. A description of the process used to acquire your data, such as the originating office, survey or experiment.
4. In your Data Understanding section, include relevant Data Understanding calculations (descriptive statistics) and create visualizations such as, histograms or scatterplots.
   1. Include the most insightful 3 python visualizations in your paper, along with an explanation.
   2. Include a table with the number of observations, features and distribution of your input and output variables.
   3. If you are performing classification, address the number of classes, and if the class representations in the data are balanced, or unequally represented.
5. Include 6 references that are 2014 or newer with these attributes:
6. Recommended split: 2-3 references on topic background, and 3-4 references on machine learning analyses of the topic area.
7. References must be unclassified, and may include CUI material if they are Distro A,B,C,D or E.
   * 1. If your paper is CUI, add a password to the Word file and email the password separately to me.
     2. If a reference is CUI, annotate that in the bibliography after the title
   1. Select from journal articles, course readings, conference proceedings, books, DTIC, and quality webpages. Use webpages as a last resort if information is unavailable elsewhere.
   2. Graphics must have a reference as well - if a graphic would be suitable, try to source it from a DoD report or journal as they do not have copyright restrictions.
   3. Manage your references using a bibliography tool – instructions for the Word tool are given in Appendix C. Select the IEEE format.
   4. Add a bibliography to your paper using a bibliography tool (Word, Mendeley, Zotero etc).
   5. Update the sentence just before the bibliography to mention which reference tool you are using
8. Upload the following to canvas in a single zip file named *(your\_last\_name)\_HW4*. If the files are uploaded separately, Canvas renames them when I download, which can cause Colab datafile loading issues for me.
   1. Word or pdf paper
   2. Python .ipynb file
   3. Dataset
   4. PDF version of all references. Save a pdf version of any web pages, but PDFs of books or textbooks are not required.

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Homework #5 |

In this assignment you will continue a project that will be iterated upon for the remainder of the course.

1. Address feedback from the prior homework.
2. Add 1000 words of CRISP-DM Data Preparation, Metrics & Modeling to your paper. After these additions your paper should not exceed 3000 words. Write using a formal writing style using the guidelines in Appendix B.
3. In your Data Preparation section, document the methods that you used, such as transformations or filtering. If applicable, include up to 2 before/after figures showing the results of your data preparation.
4. In your Metrics section, justify the choice of metrics that will be used to compare classical & NN model performance.
5. You will conduct 3 modeling efforts – a trivial model (baseline), classical analysis (classical regression or classical classification), and a neural network analysis. Document how you performed the modeling in the Methods section, and the model performance in the Analysis and Results section.
6. Trivial model (more description in week 8 bonus material):
   1. Trivial models are an important comparison for the reader so they know the value that your modeling has added. They are different for regression & classification, and are considered trivial because they take very little effort to create
   2. For regression, the trivial model is the mean of the labels. Calculate the model performance metrics, such as R2 or MSE, that result for a model that always predicts the label mean.
   3. For classification, there are 2 trivial models. Calculate the model performance metrics, such as AUC, accuracy or f1, that result from these models:
      1. The first model always predicts the majority class.
      2. The second model predicts randomly.
7. Classical modeling approach:
   1. Justify your selection of 3 classical modeling variations, and describe the differences. Variations could include different algorithms, adding or removing input variables, or transformations.
   2. Discuss the contributions of input variables to your model (z or p test).
   3. Document performance metrics for each modeling variation:
      1. If regression: R2, MSE and residual analysis. Residual analysis could include a plot of the label vs. the predicted label.
      2. If classification: Accuracy, precision, recall and area under the ROC curve (AUC).
   4. Create a table that summarizes part 7c for your 3 variations. Also include the performance of your trivial model.
8. Neural network:
   1. Justify your selection of optimization algorithm, loss function and regularization technique, based on the type of problem you are modeling (see concept map on slide 8 of *6B Optimization.pptx*).
   2. Document how you monitor to prevent overfitting, such as a train/validate split, or cross-validation. Justify your selection, based on your number of datapoints and number of input variables.
   3. Perform a hyperparameter sweep on neurons per layer, number of layers and 2 other hyperparameters. Select the model with the best performance.
   4. Investigate how 1 regularization technique affects your modeling, and document the results.
   5. Select a “best” model and create a figure that shows the NN did not overfit while training (train/test metric vs epochs).
   6. Create a table (similar to part 7d) that summarizes the performance of part 8c/8d NN 3 variations.
9. Add references to your paper so you have a total of 10 or more. The references should have the same attributes as homework #4.
10. Upload the following to canvas in a single zip file named *(your\_last\_name)\_HW5*:
    1. Word or pdf paper
    2. Python file
    3. Dataset
    4. PDF version of any new references. PDFs of books or textbooks are not required.

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Homework #6 |

1. Address feedback from homework #5 in your Word paper and modeling.
2. Add the topics mentioned in parts #3-6 below to your paper, so that your paper totals 3000-4000 words. Write using a formal writing style using the guidelines in Appendix B.
3. Model evaluation:
   1. Evaluate the results of your efforts using the criteria established in homework #4. Assess the value of this effort to your organization.
   2. Discuss the final model selected, and why it is was selected.
   3. Mention inferences that can be drawn from the classical modeling and from the data. These could include notable trends or features that are highly correlated with the label. Particularly insightful findings could be mentioned in your abstract and conclusion.
4. Model Application (a.k.a. CRISP-DM deployment):
   1. Discuss how your model could be used to further the mission of your organization, qualitatively or quantitatively.
   2. Qualitative applications - mention how you could apply inferences from classical modeling - here is an example from an AF Drug Testing Lab student:
      1. Inference: younger, less educated individuals who exhibit sensation seeking behavior and are open to experience tend to be at higher risk for THC use.
      2. Application: we recommend a personality survey for new employees, and focus drug prevention efforts on personnel that the model indicates are at highest risk for TCH use.
      3. If your inferences were not strong, it is OK to document that instead.
   3. Quantitative applications - how can you best use the prediction of the model to make a decision?
      1. Document an analysis (excel, Python or other) showing how a decision could be made.
      2. Be sure to mention the impact of that decision, such as we could save $500K, avoid 20% of space launch failures, or predeploy food aid to areas that have the highest risk for a food crisis.
      3. If your model performance was not strong, write this section based on what could be accomplished if model performance was improved. Also, offer ideas on how the model could be improved.
5. Abstract:
   1. In 150-200 words, describe and motivate the problem you are working on. Mention your approach and list your key findings or results, along with why they are important.
   2. Include 7 or more numbers, such as number of datapoints, modeling metrics, or the benefits of the model.
   3. Compare the performance of the best model to the trivial model.
   4. If someone only has time to read your abstract, they should be able to understand your work.
6. Conclusion:
   1. Summarize your method, modeling and results, and discuss how they answer your research question.
   2. Address the impact of known limitations, and provide suggestions for future work on your topic.
   3. Be sure not to add any new information.
7. Upload the following to canvas in a single zip file named *(your\_last\_name)\_HW6*:
   1. Updated Word or pdf paper
   2. Python file & dataset
   3. PDF version of any new references. PDFs of books or textbooks are not required.

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Final Project |

In the course “Final” you will demonstrate excellent verbal technical communication skills by presenting the project that you have iterated upon during the course.

1. Address feedback from homework #6 in your paper.
2. Create presentation from your paper that is less than 20 minutes in duration, addressing the most important points of the work that you’ve performed. There is no limit on the slide count.
   1. Your target audience are machine learning colleagues that are familiar with statistics and machine learning, but not experts in your career field.
   2. Use the slide template that is provided, and you will be assessed using the presentation rubric in Appendix D.
   3. Here are two good practices for all presentations:
      1. Complete the "gain attention & interest" part of the rubric on the title slide. Make the audience eager to listen to your presentation by mentioning a few important or interesting aspects of your work, which could include "If we can accurately predict (xyz) then we can (list mission impacts)."
      2. Don’t read more that 3-4 sequential words from any slide. If you do, this can frustrate the audience since people can read much faster that you can talk. I'd recommend having fewer words per slide, and then use those words to help you remember the full sentences that you want to say.
3. Record your presentation using one of these methods. The final recording should be <50 MB.
   1. Use the instructions at this [link](https://support.office.com/en-us/article/Record-a-slide-show-with-narration-and-slide-timings-0B9502C6-5F6C-40AE-B1E7-E47D8741161C#OfficeVersion=Windows) to record an audio track of your presentation within powerpoint, and upload that file.
   2. Use your phone or other device to record an audio track of your presentation, being sure to mention the current slide number as you advance slides. Upload both the powerpoint and audio track to Canvas.
4. Upload the following to canvas in a single zip file named *(your\_last\_name)\_final*:
   1. Powerpoint file
   2. Your audio file if you used option 3.b.
   3. Final Word or pdf paper
   4. Final Python file & dataset
   5. PDF version of all references, except for books/textbooks.

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Appendix A – IEEE Citation Guide |

A PDF file is available at the link below:

[IEEE Citation Guidelines](https://ieee-dataport.org/sites/default/files/analysis/27/IEEE%20Citation%20Guidelines.pdf)

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Appendix B – Guide to Formal Writing |

In formal writing, your goal is to communicate your main points in a clear, efficient and unambiguous manner. The reader should be able to understand your meaning on their first review of a sentence or paragraph.

Strive to match the writing style of the journal articles and reports that are assigned as course readings, and the topics below are a starting point:

1. Avoid casual language, imprecise language and slang:
   1. For casual language, while you might *say* “$15.2 billion worth of fuel”, in formal writing you would *write* “the fuel cost was $15.2 billion.” Also, “the prototype appears to be a solid candidate” could be rewritten as “the prototype has potential to meet the requirement”
   2. Examples of imprecise language are “done” and “some.” “Done” could be “performed” or “completed”, and in many cases you can replace “some” with a number.
2. Write from a third-person perspective, which is typical in academic writing. This means not using personal pronouns (words such as I, me or my), and avoiding referring to yourself or your reader. In this style, instead of writing “I used MSE to compare the outputs of…” you would write “MSE was used to compare the outputs of…”
3. The IEEE citation style is citation and then punctuation, like this example [1].
4. Only capitalize proper nouns. If it is important to emphasize a word, it is acceptable to use underlining or *italic font* instead of capitalizing it.
5. Always include a space between a number and its unit, and a leading zero before a number that is <1, like these examples: 5 kW and 0.6 kWh.
6. Avoid excessive significant digits, especially if there is any ambiguity in the parameters used in a calculation (i.e. 1.6 vs. 1.648343).
7. Hyphens are used to connect two adjectives that modify a noun, and an example is “well-known author”. Don’t use if the modifications were used after the noun, such as “the author was well known.”
8. Parentheses are great to define acronyms, but in other cases try to avoid them (pssst they make it seem like you are whispering something to the reader, which seems informal).
9. Semicolons are like cayenne pepper; they are best in low doses. Use them sparingly; they can be distracting.

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Appendix C – Bibliography Instructions |

Instructions for using Word’s bibliography tool:

* To add a reference and cite it
  1. Click where you’d like the reference to be inserted, typically before the punctuation at the end of a sentence
  2. Go to the References bar, make sure the IEEE style is selected & click “Insert Citation”
  3. Click “Add New Source”.
  4. Select the proper “Type of Source” and separate multiple author names with semicolons.
  5. Select the “Corporate Author” checkbox if an organization created the report. If you don’t, Department of Defense will appear as D. Defense
* To cite a previously added reference
  1. Click where you’d like the reference to be inserted, typically at the end of a sentence, but before the punctuation mark
  2. Go to the References bar and click “Citations”
  3. Double-click the citation that you’d like to add
  4. To update your bibliography after adding a reference, click on the reference list and select “Update Citations and Bibliography”

|  |  |
| --- | --- |
| C:\Users\tford\Google Drive\AFIT\SENG520 SP17 DL AFMC\Images\AFIT Logo 3.png | DASC 522 Machine Learning |
|  | Appendix D – Presentation Rubric |

