DAV 6150 Final Project Guidelines

***** You may work in small groups of no more than three (3) people to complete the Final Project *****

Throughout this course we explore a wide variety of machine learning algorithms, each of which has its own strengths and weaknesses. The algorithms we are exploring can be used to estimate/predict the values of either numeric or categorical response variables, with many algorithms being applicable to both. An essential component of data science and machine learning work is knowing when and how to apply a given type of algorithm when presented with a new challenge. For your **Final Project** of this course you are tasked with finding a data set of your own choosing and creating a suite of machine learning models whose performance will be compared and contrasted. Your suite of models **MUST** be comprised of at least **3 *distinct* types of machine learning models** discussed within the course learning materials (e.g., a K-means model, a decision tree, a random forest, a neural network, a regression model, a Naïve Bayes model, an XG Boost model, an SVM model, etc.), **at least one of which MUST be either an XG Boost or Neural Network model**. The data you select and the models you elect to construct **must** serve to answer one or more formal research questions that you define for purposes of framing your **Final Project** work.

You will construct and evaluate your three distinct machine learning models independently of one another, and identify which of those you believe is the "best" performing model.

You will then **construct an additional ensemble model** following the guidelines outlined below to determine whether the ensemble approach outperforms each of the three individual models you initially created.

This **Final Project** represents **20% of your final grade for the course**. Your **Final Project** is comprised of three separate deliverables:

1. A formal Final Project Proposal;
2. Your Final Project writeup + Python code (in the form of a Jupyter Notebook);
3. A "live" presentation of your work during our final Live Session (Module 15).

A summary of the schedule and scoring for these deliverables is provided below.

Deliverables Schedule

Deliverable	***Date***	***Points***
Proposal	Module 8	**30**
Final Project	Module 14-15	**170**
Final Project Presentation	During Final Live Session (Module 14-15)	**50**

Policy on Collaboration

You may work in teams of up to three people to complete the **Final Project**. Each project team member is responsible for understanding and being able to explain *all* of the submitted project work + Python code. Remember that you can take work that you find elsewhere as a base to build on, but you need to acknowledge the source, so that your grade is based upon what you actually contribute, ***not*** on what you start with.

Proposal Guidelines

Your first deliverable for this Project (30 Points) is the **Final Project Proposal**. The Proposal for the Final Project will be submitted in the form of a formal research proposal document. . Furthermore, you need to ensure that the Project you are proposing will satisfy all of the requirements specified in the **Final Project Checklist** (see below). Your proposal must include each of the sections outlined below and must be submitted in the form of a Jupyter Notebook.

Introduction (8 Points)

This section should provide some context for the basis of the research questions you plan to answer without actually describing the research questions themselves. For example, if your research questions are focused on a health related issue, you might provide a brief summary of how many people are affected by that issue each year either regionally, nationally or globally, including any infection rate or mortality statistics you were able to gather. Basically, in the Introduction you are trying to make the reader understand why the research questions you are going to propose are relevant and should be of interest to them.

Research Questions (8 Points)

Provide a single succinct sentence describing each of your research questions. Then provide a paragraph or two explaining how the results of your research might be used/implemented in the "real world".

Data to be Used (4 Points)

Clearly identify the sources of your data and explain the methods you will use to collect the data from those sources, e.g., "Data will need to be collected from this source via scraping of a web page..", etc.

Approach (10 Points)

Explain how you plan to manage the data you are collecting: e.g., will you be storing it within some sort of database, etc.? Explain what types of statistical analysis you plan to utilize to help answer your research questions. Explain any graphics you plan to generate to help answer your research questions. Identify the three different types of models you plan to construct and how you will compare their performance. Explain your plan for combining the output of the four required individual “weak learner” models into a single ensemble model, etc. The reader should come away with a clear understanding of how you plan to proceed with your work. Keep in mind that since this is a proposal, you need to be able to convince the reader that your proposed project is: a) realistic; and b) feasible within the time allotted for the project. Also, be sure to clearly articulate the roles and responsibilities of each team member for your Project work.

Your Final Project Proposal Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors.

Upload your Jupyter Notebook within the provided Final Project Proposal Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_FinalProjectProposal**" (e.g., J_Smith_FinalProjectProposal). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member MUST submit their own copy of the team’s work within Canvas.***

Proposal Approval

Once you’ve submitted your proposal its content will be reviewed for purposes of determining whether or not what you have proposed is acceptable as a Final Project for this DAV 6150 course. If so, you will be

conditionally approved to start work on your Final Project. If not, you will receive detailed feedback regarding any issues that need to be addressed before you can receive approval for your Project. You will be able to re-submit your Proposal as many times as necessary to achieve the required approval. Once you receive the conditional approval you will have earned the full 30 points possible for the Proposal component of the Final Project (assuming you had originally submitted the Proposal no later than the due date specified in Module 8).

Your Second deliverable for this Project (170 Points) is your Final Project Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Abstract (10 Points):** Use 250 words or less to summarize your problem, methodology, and major outcomes.

2) **Introduction (10 Points):** Describe your project, including the scientific or business motivation for the research question you have chosen to answer. This section should summarize the content of your Final Project Proposal, so be sure to explain your research question, describe the source and content of the data set you have chosen to work with, and summarize your approach to meeting the requirements for the Project.

3) **Exploratory Data Analysis (30 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis, including any preliminary predictive inferences. This section should include any Python code used for the EDA.

4) **Data Preparation (15 Points)**: Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA, including any feature engineering techniques you have applied to the data set. This section should include any Python code used for Data Preparation.

5) **Prepped Data Review (10 Points)**: Explain + present your post-Data Prep EDA analysis. This section should include any Python code used for re-running your EDA on the variables adjusted during your Data Preparation work.

6) **Machine Learning Models (40 Points)**: Explain + present the work done to construct your three distinct machine learning models, including your feature selection / dimensionality reduction decisions and the process by which you selected the hyperparameters for your models. This section should include any Python code used for feature selection, dimensionality reduction, and model building.

7) **Model Selection (15 Points)**: Explain your model selection criteria. Identify your preferred model from the distinct models you selected for the Project. Compare / contrast its performance with that of your other models. Discuss why you've selected that specific model as your preferred model. Apply your preferred model to the testing subset and discuss your results. Did your preferred model perform as well as expected? Be sure include any Python code used as part of your model selection work and to frame your discussion within the context of the classification performance metrics you have derived from the models.

8) **Ensemble Model (30 Points)**: Construct an ensemble model comprised of appropriate "weak learners" relative to the type of response variable you are attempting to estimate:

    - If you are attempting to **estimate a floating point value**, your ensemble **MUST** be comprised of at least one of each of the following: 1) either a multiple regression or polynomial regression model; 2) a KNN regressor; 3) an SVM regressor; and 4) a decision tree regressor.

- If you are attempting to **estimate an integer value**, your ensemble **MUST** be comprised of at least one of each of the following: 1) A "count" regression model; 2) a KNN regressor; 3) an SVM regressor; 4) a decision tree regressor.
- If you are attempting to **estimate the value of a categorical variable**, your ensemble **MUST** be comprised of at least one of each of the following: 1) An appropriate logistic regression model; 2) a KNN classifier; 3) an SVM classifier; 4) a decision tree classifier.

Explain how your ensemble model combines the output of the individual component models to calculate a prediction. Compare the results of the ensemble model against those of the individual models you created in the **Machine Learning Models** section (outlined above): does the ensemble approach outperform all of the individual models? If so, explain why the ensemble approach is more effective for your data. If not, explain what might be the reason for the ensemble model's underperformance.

9) **Conclusions (10 Points):** Summarize your work and clearly state the conclusions of your research. Were you able to answer the research questions you originally posed in your Proposal? Comment on any potential future extensions of the work you've completed for the Project.

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload your Jupyter Notebook within the provided Final Project Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_FInalProject**" (e.g., J_Smith_FinalProject). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member MUST submit their own copy of the team's work within Canvas.***

Your third deliverable for this Project (50 Points) is an approximately 10 minute live presentation of your work. Your presentation should include a brief overview of your research questions and the data you selected to work with, your EDA work, a high-level explanation of your data preparation + feature selection process, a discussion of your three distinct individual types of models including the hyperparameter values you selected for each, a summary of your model selection process, an explanation of why you chose your preferred model, and comments on the performance of your preferred model when applied to the testing data set. Finally, explain your ensemble model and comment on its performance relative to that of the three distinct models you choose to construct.

Final Project Checklist

To receive full credit for the Final Project, you'll need to deliver on all of the items outlined in the checklist below. **Please read carefully through this checklist before you make your project proposal**. You are (within these checklist constraints) strongly urged to limit scope and make the necessary simplifying assumptions so that you can deliver your work on time.

- ☐ Proposal describes your motivation for your work.
- ☐ Proposal describes from where you plan to source your data.

- ☐ Project uses data ***that has not been provided*** with any DAV 6150 Assignment or Project. (**See the attached list of data sets that are provided to you for various DAV 6150 Assignments + Projects)**. Also, data ***must NOT be sourced from*** Kaggle.com or any Python library (e.g., scikit-learn).
- ☐ The data set **must include at least 4 potential explanatory variables** (anything less will be deemed too simplistic for the Final Project). If you are struggling to find individual data sets having at least 4 potential explanatory variables that are useful for answering your research question(s), consider combining one or more distinct data sets to meet this requirement.
- ☐ Your project has a recognizable and reproducible "data science workflow" (e.g., data gathering, EDA, data prep, feature selection/dimensionality reduction, etc.).
- ☐ Project includes statistical analysis and graphics that describe and/or validate your data (e.g., EDA).
- ☐ Project includes appropriate data preparation + data transformation operations as needed.
- ☐ Project uses appropriate feature engineering, feature selection and/or dimensionality reduction techniques where needed.
- ☐ Project must make us of at least 3 ***distinct*** types of machine learning models discussed within this course's learning materials (e.g., KNN, neural network, regression, clustering, Naïve Bayes, XG Boost, SVM, random forest, etc.). In other words, it is not acceptable to, for example, construct 3 slightly different logistic regression models for this Project. Each of the models you construct must be of a distinctive type and not simply a variation on a single type of model. However, you are welcome to construct multiple variations of the same type of model to help you identify the "best" model of a specific type (e.g, the "best" KNN or neural network, etc.). Furthermore, at least one of the three distinct types of models you use **MUST be either an XG Boost or Neural Network model.**
- ☐ Project must include valid cross validation + appropriate training/validation/testing approaches for all machine learning models.
- ☐ Project must compare + improve models via appropriate model evaluation techniques + identify "best" model to use relative to your selected data set.
- ☐ Project must include the use of at least one ensemble model comprised of the 4 distinct types of machine learning models specified in the section labeled **Part 8) Ensemble Models** shown above.
- ☐ Presentation: Was the presentation delivered in the allotted time (15 minutes)?
- ☐ Presentation: Did you show (at least) one challenge you encountered during your work and what you did when you encountered that challenge? If you didn't encounter any challenges, your Final Project was clearly too easy for you!
- ☐ Presentation: Did the audience come away with a clear understanding of your motivation for undertaking the project?
- ☐ Presentation: Did the audience come away with a clear understanding of at least one insight you gained or conclusion you reached or hypothesis you "confirmed" (rejected or failed to reject…)?
- ☐ Code and data: Have you delivered the submitted code and data where it is reproducible and self-contained within a Jupyter Notebook on GitHub? Can someone else reproduce your results with what you've delivered? You won't receive full credit if your Python code references data on your local machine!
- ☐ Code and data: Does all of the delivered code run without errors?
- ☐ Deadline management: Were your draft project proposal, project, and presentation delivered on time? Please turn in your work on time! You are of course welcome to deliver your work ahead of schedule!

List of Data Sets Provided for DAV 6150 Assignments + Projects

All of the following data sets are being provided for your use with various DAV 6150 Assignments or Projects. The Final Project requires you to use a data set that has **not** been provided to you for any DAV 6150 Assignment or Project. As such, you are ***NOT*** allowed to use any of these data sets for any aspect of your Final Project work. You are also ***NOT*** allowed to use any data set that is embedded within any Python library (e.g., any of the scikit-learn data sets). You are also ***NOT*** allowed to use any data set provided by the Kaggle.com website.

Automobile Characteristics: https://archive.ics.uci.edu/ml/datasets/Automobile

Wine Quality: https://archive.ics.uci.edu/ml/datasets/Wine+Quality. *NOTE*: There are **many** variations of this data set available on the web. No variations of this data set may be used for the Final Project.

Online News Article Sharing: https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

Insurance Company Data: https://www.kaggle.com/rluyck/insurance-company?select=Customer_data.csv

Online Web Browsing Metrics:
https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

Movie Reviews: http://www.cs.cornell.edu/people/pabo/movie-review-data/

NYSED 2018-2019 HS Graduation Metrics: https://data.nysed.gov/downloads.php