Review: Final Team Project

Introduction

The final project for this course will assess the development of an end-to-end statistical analysis. By the end of Module 2, you will be assigned to a group by your instructor. You and your teammate(s) will work closely to find a dataset, clean and prepare it for analysis, perform EDA, and produce a model. You will then create a final technical report discussing your analysis's results and your model's validity. You and your team must use GitHub as a code hosting platform to manage version control and collaboration during this project. Creating and adding a README Actions file to your GitHub repository is also necessary. It is *recommended*that you follow the [PEP 8Links to an external site.](https://pep8.org/) – Style Guide for your Python code in the final project.

Project Timeline

* Module 2 (by the end of Week 2): The course instructor will group students into teams of two to three members.
* Module 4 (by the end of Week 4): Each team will select and introduce a dataset. The team representative will submit the "Team Project Status Update Form."
* Module 7 (by the end of Week 7): Each team will submit the following deliverables for the course project in the final week:
  1. Technical Report: One PDF document with your final technical report. This should describe your preparation and analysis of the data, discuss the final model selection, and describe the statistics behind your final model selection.
  2. Team Presentation: One 8-10 minute video presentation by all team members. Do not exceed 10 minutes. Submit one mp4 file. This should be presented as a business presentation of your analysis to a non-technical audience. You will present your analysis and findings in an understandable way to any non-technical executive or business leader. This presentation should include one slide to showcase your collaborative efforts; you will create a presentation slide highlighting each team member's name and their contributions to the final project work and deliverables.

\*\*It is critical to note that no extensions will be given for any of the final projects' due dates for any reason, and final projects submitted after the final due date will not be graded.

Project Datasets

* You are responsible for finding a dataset for the final team project in this course. Several free data resources are provided by the [UCI Machine Learning RepositoryLinks to an external site.](https://archive.ics.uci.edu/datasets).

Requirements

* Divide the work equally between the team members for the following steps, and everyone needs to code and review the code.

For your technical report document:

* You should include the following sections in your report:
  + Introduction (what is/are the question(s) you want to answer)
  + Data Cleaning/Preparation
  + Exploratory Data Analysis
  + Model Selection
  + Model Analysis
  + Conclusion and Recommendations.
* Include an appendix with the output of your code from a technical notebook (i.e., Jupyter Notebook).

For your video presentation:

* Give an 8-10 minute presentation of your analysis aimed at a non-technical audience of business leaders.
* Each team member should participate equally in the presentation.

Project Deliverables and Submission Format

* Prepare and submit your Final Technical Report in PDF format on Canvas. You will use the naming convention Final-Project-Report-Team-Number.pdf (e.g., Final-Project-Report-Team-1.pdf)
* Prepare a recorded video presentation of your project using a screencasting tool, such as Screencast-O-Matic or Zoom, to record your screen and provide a voice narration.
  + Ensure that the sound quality of your video is good and each member presents an equal portion of the presentation.
  + Export the video file to an mp4 format.
  + You may use any recording software you wish. View [Recording Video Presentation and Submission Guidelines](https://sandiego.instructure.com/courses/14112/files/1936151?wrap=1) [Actions](https://sandiego.instructure.com/courses/14112/pages/review-final-team-project?module_item_id=539752)

 for AAI students.

* + You will use the naming convention Final-Project-Presentation-Team-Number.mp4 (e.g., Final-Project-Presentation-Team-1.mp4).
* Only ONE member of your team will need to submit these deliverables.
* You will submit the peer evaluation form individually in Assignment 7.1.

Groupwork Requirements

The final teamwork project is an essential component of our courses in the AAI program. This project is representative of the kind of collaborative project you may work on during your *career and real-world projects*. A significant portion of your final grade is drawn from your participation in this final group project, so you are strongly encouraged to work within your team and that each team member contributes *equally*to the final project deliverables. Team members should plan to have clear and ongoing communication with your team and engage with the project and its deliverables each week. Lack of participation and engagement with your team and your final project can result in a failing grade for the course.

While these are the expectations for the project work, if you experience difficulty with project advancement or team dynamics, contact your instructor for assistance promptly. If you cannot perform the project as a team, contact your instructor to explore pursuing the final project independently.

View the assignment rubric on the Final Team Project page to understand how your work will be assessed.

**Dataset Sources**

**1. Kaggle (https://www.kaggle.com/datasets)**

* **Type of datasets**: Wide variety of datasets for classification, regression, and other machine learning tasks.
* **Best for**: Beginners and experienced data scientists alike.
* **Notable datasets**:
  + **Titanic dataset**: Classic dataset for binary classification (predicting survival).
  + **Housing price dataset**: For regression tasks (predicting house prices).

**2. UCI Machine Learning Repository (**[**https://archive.ics.uci.edu/ml/index.php**](https://archive.ics.uci.edu/ml/index.php)**)**

* **Type of datasets**: Large collection of datasets for both classification and regression.
* **Best for**: Academic and educational projects.
* **Notable datasets**:
  + **Iris dataset**: Popular classification dataset for predicting flower species.
  + **Wine quality dataset**: For regression tasks to predict wine quality based on chemical attributes.

**3. Google Dataset Search (https://datasetsearch.research.google.com/)**

* **Type of datasets**: A search engine for datasets across various domains, including statistics, machine learning, and social sciences.
* **Best for**: Finding specific datasets from a variety of sources.
* **Notable datasets**: Use search filters to find datasets for your classification or regression needs.

**4. StatLib (http://lib.stat.cmu.edu/datasets/)**

* **Type of datasets**: A curated collection of statistical datasets from Carnegie Mellon University.
* **Best for**: Simple statistical analysis and basic machine learning models.
* **Notable datasets**: Includes datasets on time series, multivariate analysis, and various topics.

**5. OpenML (**[**https://www.openml.org/**](https://www.openml.org/)**)**

* **Type of datasets**: Datasets for machine learning experiments, offering both classification and regression datasets.
* **Best for**: Building, evaluating, and comparing machine learning models.
* **Notable datasets**:
  + **Adult income dataset**: Commonly used for classification (predicting whether a person's income is above or below a threshold).
  + **Air quality dataset**: For regression tasks (predicting air pollution levels).

**6. Data.gov (**[**https://www.data.gov/**](https://www.data.gov/)**)**

* **Type of datasets**: US government open data, including health, finance, and environment.
* **Best for**: Public datasets that span multiple industries and fields.
* **Notable datasets**: Search based on the subject, you can find both classification and regression datasets.

**7. Awesome Public Datasets GitHub Repository (**[**https://github.com/awesomedata/awesome-public-datasets**](https://github.com/awesomedata/awesome-public-datasets)**)**

* **Type of datasets**: A community-maintained list of publicly available datasets from various fields.
* **Best for**: Exploring various domains, including economics, healthcare, and social sciences.

**8. Datasets from R (https://vincentarelbundock.github.io/Rdatasets/)**

* **Type of datasets**: Popular datasets that can be easily loaded into R but can be exported for use in Python or other languages.
* **Best for**: Simple and clean datasets for statistical modeling.
* **Notable datasets**:
  + **mtcars dataset**: For regression (predicting car performance).
  + **Diamonds dataset**: For regression (predicting diamond prices).

**9. FiveThirtyEight Datasets (https://data.fivethirtyeight.com/)**

* **Type of datasets**: Datasets from FiveThirtyEight's journalism covering sports, politics, economics, and more.
* **Best for**: Real-world datasets with an emphasis on storytelling.
* **Notable datasets**: Examples include political polling data and datasets on economic indicators.

**10. Academic Journal Supplementary Datasets**

* Many academic articles provide supplementary datasets used in the studies. These datasets are often good for educational or research purposes.
* **Where to look**: Google Scholar or journal repositories (many provide links to downloadable datasets).

**Recommended Business-Oriented Datasets for Simple Projects:**

* **Customer Segmentation Dataset** (Kaggle): Use a classification model to group customers based on behavior or demographics.
* **Bank Marketing Dataset** (UCI): Build a binary classification model to predict whether a customer will subscribe to a term deposit.
* **E-Commerce Data** (Kaggle): Perform regression or classification to predict customer purchasing habits or product success.
* **Online Retail Data** (UCI): Build models predicting sales volume or customer purchase behavior for a retail store.

**1. Customer Churn Prediction (Classification)**

* **Problem**: Predict whether a customer will stop using a service (churn) based on features like customer demographics, usage patterns, or service complaints.
* **Techniques**: Logistic Regression, Random Forest, Support Vector Machines (SVM)
* **Dataset**: Telco Customer Churn Dataset

**2. House Price Prediction (Regression)**

* **Problem**: Predict the selling price of a house based on features such as the size, number of bedrooms, and neighborhood.
* **Techniques**: Linear Regression, Ridge Regression, Decision Trees
* **Dataset**: Ames Housing Dataset

**3. Loan Default Prediction (Classification)**

* **Problem**: Predict whether a loan applicant is likely to default on a loan based on features like income, credit score, and loan amount.
* **Techniques**: Logistic Regression, Decision Trees, Gradient Boosting
* **Dataset**: Lending Club Loan Data

**4. Medical Diagnosis Prediction (Classification)**

* **Problem**: Predict whether a patient has a certain medical condition (e.g., diabetes, heart disease) based on medical records.
* **Techniques**: Logistic Regression, K-Nearest Neighbors (KNN), Random Forest
* **Dataset**: Pima Indians Diabetes Dataset or Heart Disease UCI

**5. Car Price Prediction (Regression)**

* **Problem**: Predict the price of a used car based on its features such as age, mileage, brand, and fuel type.
* **Techniques**: Linear Regression, Random Forest, Gradient Boosting
* **Dataset**: Car Price Dataset

**6. Employee Attrition Prediction (Classification)**

* **Problem**: Predict whether an employee will leave a company based on factors like salary, years of experience, job satisfaction, etc.
* **Techniques**: Logistic Regression, Random Forest, Gradient Boosting
* **Dataset**: IBM HR Analytics Employee Attrition & Performance

**7. Stock Price Prediction (Regression)**

* **Problem**: Predict stock prices based on historical stock data, including closing price, trading volume, etc.
* **Techniques**: Linear Regression, Time Series Analysis, Lasso Regression
* **Dataset**: Stock Prices Dataset

**8. Credit Card Fraud Detection (Classification)**

* **Problem**: Predict whether a transaction is fraudulent based on features such as transaction amount, location, and time of day.
* **Techniques**: Logistic Regression, SVM, Random Forest
* **Dataset**: Credit Card Fraud Detection Dataset

**9. Energy Consumption Prediction (Regression)**

* **Problem**: Predict energy consumption in households based on weather, household size, and time of day.
* **Techniques**: Linear Regression, Random Forest, Time Series Analysis
* **Dataset**: [Household Electric Power Consumption Dataset](https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption)

**10. Sentiment Analysis of Product Reviews (Classification)**

* **Problem**: Classify the sentiment of product reviews (positive or negative) based on review text using Natural Language Processing (NLP) techniques.
* **Techniques**: Logistic Regression, Naive Bayes, SVM
* **Dataset**: Amazon Product Reviews Dataset

**11. Admission Prediction for University (Regression)**

* **Problem**: Predict whether a student will get admission to a university based on GRE score, GPA, and other features.
* **Techniques**: Linear Regression, Decision Trees
* **Dataset**: Graduate Admission Dataset

**12. Wine Quality Prediction (Classification/Regression)**

* **Problem**: Predict the quality of wine based on its chemical properties like acidity, sugar content, pH, etc.
* **Techniques**: Logistic Regression, Random Forest, SVM
* **Dataset**: Wine Quality Dataset

**13. Airbnb Price Prediction (Regression)**

* **Problem**: Predict the price of an Airbnb listing based on location, size, amenities, and other features.
* **Techniques**: Linear Regression, Decision Trees, Random Forest
* **Dataset**: Airbnb Listings Dataset

**14. Cancer Detection (Classification)**

* **Problem**: Predict whether a tumor is benign or malignant based on features derived from medical images or diagnostics.
* **Techniques**: Logistic Regression, SVM, Random Forest
* **Dataset**: Breast Cancer Wisconsin Dataset

**15. Flight Delay Prediction (Classification/Regression)**

* **Problem**: Predict flight delays based on factors such as weather, time of day, and airport traffic.
* **Techniques**: Logistic Regression, Random Forest, Gradient Boosting
* **Dataset**: Flight Delay Dataset

Notes for Project (from Office Hours)

1. Groups will be assigned by this weekend

AAI-500 Group 9

Group Members

Carrie Little

Ian Lucas

Teachers & TAs

Leonid Shpaner

1. Groups to meet at least once a week to discuss progress

1st Meeting – In person, Copley Library 226 Group or Individual Study Room 10 AM Sat. 9/21

(week 3 day 5)

Decide next meetings, chat/zoom/in Person

Weekday

Weekend

Setup Presentation Walk-thru / Record

Decide with-without camera video

1. Recommendation to choose either regression or classification models for the project.

**Classification Model:**

* **Purpose**: A classification model is used to predict **categorical outcomes** or **discrete labels**.
* **Output**: The output is a class or label (e.g., "spam" vs. "not spam", "cat" vs. "dog").
* **Type of Task**: Classification tasks are aimed at sorting data points into one of several predefined categories.
* **Examples**:
  + **Binary classification**: Predicting if an email is spam (1) or not spam (0).
  + **Multiclass classification**: Predicting the type of fruit (e.g., apple, banana, or orange).
* **Common Algorithms**: Logistic regression, Decision trees, Random forests, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Neural Networks for classification.

**Regression Model:**

* **Purpose**: A regression model is used to predict **continuous numeric outcomes**.
* **Output**: The output is a continuous value (e.g., a real number like house prices, temperature, or salary).
* **Type of Task**: Regression tasks aim to find the relationships between variables and predict numeric quantities.
* **Examples**:
  + Predicting house prices based on square footage and location.
  + Predicting someone's income based on education level and years of experience.
* **Common Algorithms**: Linear regression, Polynomial regression, Ridge regression, Lasso regression, Neural Networks for regression.

1. Pick a use case first that you are trying to solve and then select data from UCI or any other source.

Define business type questions, what we are asking to solve

1. Don't overcomplicate what you are trying to solve for or predict. Anything with a possible Y/N outcome for an expected value is ideal.
2. Use Github to collaborate on code

Setup GitHub repository for project, branch for each group member

1. Use youtube to upload project final video. You can record your "10 min video" with your team using zoom and screen share. So it feels like a presentation. You can even talk through slides with your voice recording and not have video if you choose.

**Report Structure**

* Introduction
* Business Objective
  + Financial Planning Firm
* Research Questions
* Dataset Overview
* Data Cleaning/Preparation
* Missing Values
* Erroneous Data
* Data Consistency
  + Data Dictionary
* Exploratory Data Analysis
  + **Statistics/Describe**
  + **Numerical Attributes**
  + Categorical Attributes
* Model Selection
  + Classification
    - Purpose: A classification model is used to predict categorical outcomes or discrete labels.
    - Output: The output is a class or label (e.g., "spam" vs. "not spam", "cat" vs. "dog").
    - Type of Task: Classification tasks are aimed at sorting data points into one of several predefined categories.
  + **Logistic regression**, **Decision trees**, Random forests, Support Vector Machines (SVM),

**k-Nearest Neighbors (k-NN)**, Neural Networks for classification.

* + Regression
    - Purpose: A regression model is used to predict continuous numeric outcomes.
    - Output: The output is a continuous value (e.g., a real number like house prices, temperature, or salary).
    - **Linear regression**, Polynomial regression, Ridge regression, Lasso regression, Neural Networks for regression
* Model Analysis
  + - Standard Deviation/Error
    - P Values
* Conclusion and Recommendations.