DS 6021: Introduction to Predictive Modeling

Final Project Instructions and Rubric

Overview

This project is your opportunity to demonstrate a synthesis of skills: posing interesting research questions, finding reputable data, exploring the data through summaries and visualizations, building predictive and unsupervised learning models, and presenting results professionally.

Learning Objectives

1) **Define specific research questions** that can be answered with the data.

2) **Find a reputable data source.** The dataset should have at least 4 numeric variables and at least 4 categorical variables. Choose something your group is interested in — do not be afraid to pick unconventional topics. You may also combine multiple datasets that are related to an overall theme. Clearly document your data source(s) and explain why you chose them. You are encouraged to use modern data engineering techniques such as APIs or web scraping.

3) **Apply data engineering and cleaning steps.** Depending on the dataset, this might include renaming columns, removing irrelevant or duplicated rows/columns, mutating variables, creating new features, or subsetting the data to investigate specific questions.

4) **Visualize and summarize** the pertinent variables in the data. Your analysis should include a robust set of descriptive statistics and informative visualizations.

5) **Use predictive models and unsupervised learning** methods covered in the course (and beyond, if relevant) to investigate your research questions. Include each of the major models from class where appropriate: Linear Regression/GLMs, KNN, K-means clustering, PCA, and MLP.

6) **Present your work** to the class in a 6–8 slide presentation. The first slide must include the project title, group number, and list of group members in alphabetical order of last names. The second slide should summarize key visualizations and variable summaries. Subsequent slides should focus on your research questions and how you address them using predictive and unsupervised learning models. The presentation must include a demo of your app. Presentations should last 5 minutes with an additional 2 minutes for Q&A. Practice to ensure you finish on time. Use visuals effectively — avoid cluttered slides and emphasize clarity over quantity.

Project Deliverables

Turn in the following via Canvas, under Assignments.

1. Presentation slides saved as a PDF or HTML.
2. Original dataset(s) in CSV or Excel format. Include both raw and cleaned versions if applicable.
3. Compiled code: R Script, R Markdown file, HTML file, or Python notebook. All code should be well-commented and reproducible.
4. A published app (built in R Shiny or Python Dash). The app must be deployed publicly (e.g., shinyapps.io, Heroku, etc.) so it can be accessed for evaluation.

Examples of Apps:

- NFL Play Modeling Dashboard: Predicting Yards Gained per Play
- Freddie Mac Loan Default Modeling Dashboard
- City Weather Clustering with K-Means

Evaluation (100 points total)

The project and presentation are worth 100 points. The scores for each group will be based on the following rubric.

1. Introduction & Dataset Summary (5 pts)

- **0–2:** Minimal introduction; dataset unclear or poorly described.
- **3–4:** Adequate introduction; dataset described but missing important context.
- **5:** Clear and professional introduction; dataset well described (source, variables, rationale).

2. Data Engineering & Preparation (10 pts)

- **0–3:** Little or no cleaning or preprocessing shown.
- **4–7:** Some cleaning steps applied; limited justification.
- **8–10:** Thorough data preparation (cleaning, transformations, feature engineering) clearly tied to research questions.

3. Visualization & Exploratory Data Analysis (10 pts)

- **0–3:** Trivial questions or poorly designed charts.
- **4–7:** Charts answer reasonable questions but lack depth or clarity.
- **8–10:** Insightful questions explored with well-designed, clearly justified charts.

4. Predictive Modeling (30 pts)

- **0–10:** Models misapplied or unjustified; no evaluation.
- **11–20:** Models somewhat appropriate; some assessment of assumptions/performance.
- **21–30:** Models well justified by research questions; assumptions checked; performance clearly evaluated.

5. Unsupervised Learning (15 pts)

- **0–5:** No attempt or trivial use of clustering/PCA.
- **6–10:** Method applied but weak justification or limited interpretation.
- **11–15:** Strong use of clustering and/or PCA, justified by data and tied to research questions.

6. App/Dashboard & Presentation (15 pts)

- **0–5:** App missing, presentation unorganized or incomplete.
- **6–10:** App works but limited; presentation covers main points but uneven.
- **11–15:** Interactive, polished app; presentation professional, within time, with clear communication.

7. Conclusions & Insight (5 pts)

- **0–2:** Findings unjustified or irrelevant.
- **3–4:** Findings somewhat tied to questions but limited depth.
- **5:** Findings are coherent, justified, and clearly connected to research questions.

8. Ambition & Creativity (10 pts)

- **0–3:** Minimum effort; bare-bones scope.
- **4–7:** Reasonable effort; moderate scope.
- **8–10:** Ambitious scope; creative use of methods or data.

Total Points: 100