Project 1: OPAN 6602 - Machine Learning I Fall 2024/Mod 2

The Project 1 is worth 30% of your final course grade. It's due November 18, 2024.

Your assignment is to analyze the data set contained in the Bike Share data set (Capital Bike Sharing data by hour.csv) and deliver a report, as described below. In your report, be sure to support your responses with statistical evidence from the data and your model.

Objective

Imagine that you are a consultant submitting a report to your client (Capital Bike Share). Your objective is to help your client predict demand for their total rentals of their product. More important than just a raw prediction, you need to interpret the results so that the client understands the drivers of demand and so that you can make actionable recommendations.

Data Set

For this analysis, you will use the Bike Sharing Systems Data set. The Bike Share data set (Capital Bike Sharing data by hour.csv) features data on bike rental demand for the bike sharing program in Washington, D.C. Using the bike sharing system, people rent a bike from one location and return it the same place or a different location on a need basis (mostly casual users) or membership (mostly regular users). This process is controlled by a network of automated kiosks across the city.

The data set shows hourly rental data for two years (2011 and 2012). The data set is for the first 19 days of each month. Variables include:

- instant: record index
- dteday: date
- season: four categories > 1=Winter, 2=Spring, 3=Summer, 4=Fall
- yr: year (0=2011; 1=2012)
- mnth: month (1 to 12)
- hr: hour (0-23)
- holiday: whether the day is a holiday (1) or not (0)
- weekday: day of the week
- workingday: if the day is neither a weekend nor holiday (1), otherwise (0)
- weathersit: Four categories of weather
  - 1 = Clear, few clouds, partly cloudy, cloudy)
  - 2 = Mist + cloudy, mist + broken clouds, mist + few clouds, mist
  - 3 = Light snow and rain + thunderstorm + scattered clouds, light rain + scattered clouds
  - 4 = Heavy rain + ice pallets + thunderstorm + mist, snow + fog
- temp: Hourly temperature in Celsius (normalized); normalized formula: (t-tmin)/(tmax-tmin)
- atemp: "Feels like" temperature in Celsius (normalized); normalized formula above
- hum: Relative humidity (normalized. The values are dividing to 100 (max))
- windspeed: Wind speed (normalized. The values are divided to 67 (max))
- registered: Number of registered users
- casual: number of casual users
- count: number of total rentals (registered + casual)

Deliverables

1. Report that includes visuals, analysis, and recommendations.

2. Code appendix of your R code.
3. Optional appendices for charts, tables, etc.

Main body and citations should be 8 pages or less. Report should be written in 12 point font with 1.5 line spacing.

Make sure your code is well-organized and well-commented. You won't be graded on your code, but I may use it to verify your results.

You may include up to 5 pages of appendices for charts, tables, etc.

Analysis to Perform & Suggested Report Outline

You do not need to structure your report as outlined below as it is a suggestion. However, pay close attention to the questions and recommended analysis.

1. Introduction

- Briefly describe the business problem and context.
- Define the dependent variable and independent variables (i.e., outcome and predictors).

2. Exploratory Data Analysis

- Provide summary statistics (mean, median, standard deviation, etc.).
- Include in the report univariate summaries of the whole data set, not just the training set. **Do this only after you have completed your analysis. Do not look at the test set and let it influence your analysis.**
- Identify any correlations between the independent variables and the dependent variable.
- Check for potential multicollinearity and explain why this might be a problem in regression.
- Include visualizations (scatter plots, correlation heat maps, etc.) in the main body only if they support your main point.

Set up

- Split your data into test and training sets.
- Split your training set into training and validation sets or use k-fold CV for validation.
- Get univariate summaries of all variables in the data set.
- Produce pairs plots of the data.

Questions to consider

- Are there variables that R has read in as numeric but should be categorical? (i.e., Variables that should be character or factor variables.) Could this affect your analysis? Re-code such variables as necessary.
- Which variables are outcome variables? Be sure to exclude any outcomes from the right hand side of the regression model.
- Are there seasonal differences? How do the time variables relate to each other? Is it worth reporting summary stats by one or more of those time variables?

3. Regression Model

- Build a Multiple Linear Regression model using a combination of your judgment and automated model selection procedures to select your final model.
- Provide justification of variables and transformations included and excluded.
- Report the following outputs from the model:
  - Coefficients for each independent variable.
  - R-squared and adjusted R-squared values.
  - RMSE values.
  - Significance (p-values) for each independent variable.
- Use visuals like residual plots to assess model fit.

Questions/Analyses to Consider

- Are there polynomial or interactions that are useful?
- If you include polynomial or interaction terms, how does that affect the marginal effects of the model?
- If you include polynomial terms, are there optimal points that might come in play? e.g., Does the polynomial hit a maximum or minimum value that could affect recommendations?
- How do your evaluation metrics compare across the training, validation, and test runs?
- Bonus for bragging rights: Do you get different results if you model total rentals, vs. registered and casual rentals separately?

4. Interpretation of Results

- Explain your key findings.
- Which variable(s) contribute the most to model fit? Consider whether partial R-squared, coefficient values, or both are the appropriate measure. If using partial R-squared and polynomial terms, consider the difference in interpretation between removing both the original variable and polynomial term to get partial R-squared or only removing one at a time.
- How do changes in independent variables affect the outcome variable (marginal effects).
- Discuss potential business implications of the model.

5. Limitations and Assumptions

- Discuss any linear regression assumptions that were violated and how you corrected for them.
- Discuss any limitations of the model, such as the potential for omitted variables or biases in the data.

6. Recommendations and Conclusions

- Provide actionable business recommendations based on the model results.
- Using the ethics framework from week 1, discuss any ethical issues that may arise from this project.

Evaluation Rubric

Rubric: 150 points.

1. Report Structure and Clarity 10
2. Problem Statement and Approach 15
3. Multiple Regression Model 40

4. Model Diagnostic Tests and Subset Selection 45
5. Managerial Insights and Recommendations 25
6. Writing Quality 10
7. Appendix 5