Assignment 1

Will Doyle

2025-01-07

Assignment: Analyzing and Modeling Postsecondary Credits

In this assignment we'll be pushing further out into the future. I want you to predict the number of postsecondary credits someone earned by 2018, based on the individual's entering characteristics as a ninth grader in 2009.

This assignment will assess your ability to: - Present and analyze a dependent variable. - Understand and apply regularization techniques. - Effectively communicate results through coding and visualization.

1. **Exploring the Dependent Variable**
   - Load the dataset and clean the data. The `hsls_extract2.csv` dataset is on the brightspace page, under datasets. Summarize `x5postern` using descriptive statistics (mean, median, standard deviation). Visualize its distribution using a histogram and a density plot. (hint: summarize, geom_histogram).
2. **Categorical Variable Analysis**
   - Identify any categorical independent variables in the dataset. Create a bar plot to explore the mean values of `x5postern` across different levels of one selected categorical variable. (hint: group_by, summarize, geom_col)
3. **Continuous Variable Analysis**
   - Select one continuous independent variable. Create a scatterplot with `x5postern` on the y-axis and the selected independent variable on the x-axis. Add a trendline to visualize the relationship. (hint: `geom_point`)
4. **Missing Data Handling**
   - Identify missing values for `x5postern` and other variables. Make sure to handle missing data! Remember that for most variables, all negative values indicate missing data. This is true for `x5postern.`
5. **Linear Regression Model**
   - Fit a standard linear regression model to predict `x5postern` using all available predictors. Report the coefficients, R^2, and RMSE of the model. Interpret the top three predictors based on their coefficients.
6. **Feature Engineering**
   - Modify the dataset by creating at least one new feature (e.g., an interaction term or a transformed variable (hint: `step_log`, `step_poly`)). Refit the linear regression model with this new feature. Compare the RMSE and R^2 values with the previous model.
7. **Lasso Regression Setup**
   - Set up a Lasso regression model using `tidymodels`. Use a range of penalties to evaluate how the model simplifies variable selection. Provide a summary of the variables retained at different penalty levels.
8. **Lasso Regression Evaluation**
   - Identify the penalty value among those you worked with that resulted in the lowest rmse.
9. **Visualizing Regularization**
   - Create two plots:
     - RMSE versus penalty values (on a logarithmic scale).
     - Coefficient estimates versus penalty values for the top two predictors.

- Interpret these plots and discuss how the penalty affects model performance and variable selection.

Submission Instructions:

- Include all code, outputs, and visualizations in an organized RMarkdown or similar document.
- Provide brief interpretations for each step, focusing on how the results contribute to understanding `x5postern`.