**Student: Seif Kungulio**

**Date: 03/16/2025**

**Subject: Project 3**

**Class: DSCI 512**

**Section: 01W**

**Instructor: Dr. Nengbing Tao**

**File Name: Project3\_Kungulio\_Seif.docx**

**PART I.**

1. Load the dataset mtcars.xlsx into memory and convert column **am** to a factor using factor() function.

A screenshot of a computer program

AI-generated content may be incorrect.

1. Split the data into training set and test set. The training set contains the first 35 observations, the test set containing the remaining observations.

A screenshot of a computer program

AI-generated content may be incorrect.

1. Build a logistic regression model with the response is **am** and the predictors are mpg, cyl, hp, and wt using glm() function.

A screenshot of a computer program

AI-generated content may be incorrect.

1. Compute the test error on the test data set using a confusion matrix. Is it a good model based on test error?

A screenshot of a computer program

AI-generated content may be incorrect.

The model, despite its 83.33% accuracy, is ineffective due to severe class imbalance. It correctly identifies all “automatic” cases but completely fails to predict “manual” vehicles, resulting in a sensitivity of 1.00 and specificity of 0.00. The Kappa score of 0 indicates no meaningful classification ability beyond random guessing. Additionally, the Balanced Accuracy of 0.50 and McNemar’s Test p-value of 1.000 confirm that the model lacks discrimination between classes. Overall, it overfits to “automatic” and fails to generalize, making it unreliable for classification.

**PART II.**

1. Build a linear model to forecast number of total rentals (count) using potential predictors, season, holiday, workingday, weather, atemp, and registered.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

The model effectively explains bike rental patterns, showing that seasonality, holidays, and working days have substantial effects. Weather conditions, particularly misty/cloudy and snowy conditions, decrease rentals, while temperature has a positive impact. Surprisingly, heavy rain does not show a significant impact, which may warrant further investigation. Finally, the number of registered users is the strongest predictor, reinforcing the idea that bike-sharing systems heavily rely on frequent users rather than occasional riders.

1. Perform best subset selection using bestglm() function based on BIC. What’s the best model based on BIC?

A computer screen shot of a program

AI-generated content may be incorrect.

This model suggests strong predictive performance, with seasonal, weather, holiday, and temperature-related factors all significantly influencing the outcome variable. The model’s high R – squared and significance levels indicate it is likely capturing the key drivers in the data effectively.

1. Compute the test error of the best model based on BIC using LOOCV.

A screenshot of a computer program

AI-generated content may be incorrect.

The test error is 35.09127

1. Calculate the test error of the best model based on BIC using 10-fold CV.

A screenshot of a computer program

AI-generated content may be incorrect.

The test error is 35.06165

**Note Qn3 & Qn4:**

An RMSE value of 35 indicates that, on average, the model's predictions deviate from the actual values by 35 units. In general, lower RMSE values signify better model performance, as they indicate smaller discrepancies between predicted and actual values.

1. Perform best subset selection using bestglm() function based on CV. What’s the best model based on CV?

A screenshot of a computer program

AI-generated content may be incorrect.

The model's high R-squared value and significant predictors indicate a good fit, with meaningful relationships between the predictors and the outcome. However, certain predictors, such as *workingdayYes* and *seasonFall*, show large effects that warrant careful interpretation to ensure they align with domain expectations and aren't driven by multicollinearity. Furthermore, the broad range of residuals suggests that while the model is generally accurate, there may be outliers or variations not fully captured by the current predictors.

1. Perform the backward stepwise selection using stepAIC() function. What’s the best model?

A screenshot of a computer program

AI-generated content may be incorrect.

The analysis used backward stepwise selection to identify the best predictive model for bike rentals, retaining all variables as their removal increased the Akaike Information Criterion (AIC).

A screenshot of a computer program

AI-generated content may be incorrect.

The final model, explaining 96.25% of rental variations, confirmed that season, holidays, working days, weather, temperature, and registered users significantly impact rentals. Rentals decrease in summer, fall, and winter, as well as on holidays and working days. Cloudy weather and light snow reduce rentals, while temperature positively influences usage. Registered users strongly drive demand, emphasizing their importance. These insights can help bike-sharing services optimize operations, adjust for seasonal trends, and encourage ridership through registration programs.