STAT 561 - Homework 3

Instructions

1. *Due Monday, April 21st at the 11:59pm. Any submission after that (24hrs after) will be graded out of 50%.*

2. *Your submission should include a* ***pdf*** *from your generated R markdown, and your R markdown file.*

3. *Make sure to highlight the part of the output that have the information you will be providing as answers.*

4. This work should be done as a group. Only one person in the group should submit the work with the names of all the people in the group on the document

---

Question 1

1. **Student performance data**. The Student Performance Data set is synthetically created dataset which includes 10,000 student profiles. The variables in the data are;

• Hours Studied: The total number of hours spent studying by each student.

• Previous Scores: The scores obtained by students in previous tests.

• Extracurricular Activities: Whether the student participates in extracurricular activities (Yes or No).

• Sleep Hours: The average number of hours of sleep the student had per day.

• Sample Question Papers Practiced: The number of sample question papers the student practiced.

• Performance Index: A measure of the overall performance of each student. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

Before you analyze your data, explore your data set to understand the nature of the variables in the data set. Do not provide any results from your exploration.

a) Formulate a question that you can use this data to answer. You would use multiple linear regression to answer this question. State clearly what your predictor variables would be and what your target variable would be.

In predictive modeling, one of the primary foci is on the model's performance with new data. This is crucial because evaluating error metrics on new data provides insights into the model's ability to generalize to unseen situations. The Root Mean Square Error (RMSE) serves as a measure of the model's average prediction error. For instance, an RMSE value of 6.5 indicates that the model's predictions are, on average, about 6 to 7 units off from the actual values. It is essential to select models that demonstrate strong performance on new data. The process involves:

Fitting the model on the training sample to learn the patterns in the data. Making predictions on a new, unseen sample to test the model's learned patterns. Validating the model's performance on this new sample to ensure its predictions are accurate and reliable.

b) Split your data into a training (70%) and test set (30% test).

c) Train a multiple linear regression model using the training set and a 10-fold cross validation. Use the train function from the caret package, specifying method = "lm" for linear regression. How would you interpret any 2 of the regression coefficients in the context of the student performance data? Provide the fitted equation.

Example code:

```
library(caret)
# Fit lm model using 10-fold CV: model
model <- train(
target~...., # replace target with the target variable
data=...,
method = "lm",
trControl = trainControl(
method = "cv",
number = 10,
verboseIter = TRUE
)
)


summary(model)
```

d) Evaluate the performance of your regression model on the test set. Use metrics such as R-squared and Root Mean Squared Error (RMSE) to assess how well the model predicts the target variable. Comment on this.

Example code:

```
# Predict on test set
prediction = predict(model,test)

# Compute errors
errors=prediction-test$target replace target with the target variable

# Calculate RMSE
sqrt(mean(errors^2))
```

e) Investigate the residuals from your regression model in c) to check model if model assumptions are satisfied. What do these diagnostics tell you about the model's assumptions and its suitability for the data?

f) Based on the insights gained from your analysis, propose interventions that could potentially improve student performance. Are there any violations of assumptions? Suggest remedies for these violations.

Question 2

In this study, we are going to determine whether individuals can get a loan or not based on the following predictor variables:

- Loan ID: A unique loan ID.
- Gender: Either male or female.
- Married: Weather Married(yes) or Not Married(No).
- Dependents: Number of persons depending on the client.
- Education: Applicant Education(Graduate or Undergraduate).
- Self Employed: Self-employed (Yes/No).
- Applicant Income: Applicant income.
- Co-applicant Income: Co-applicant income.
- Loan Amount: Loan amount in thousands.
- Loan Amount Term: Terms of the loan in months.
- Credit History: Credit history meets guidelines.
- Property Area: Applicants are living either Urban, Semi-Urban or Rural.

Target variable: Loan Status: Loan approved (Y/N).

(a) Split the data set into a training set and a test set (80% training, 30% test).

(b) Train the logistic regression model using the training set with a 10-fold cross-validation to optimize model parameters (Use the caret package).

(c) Evaluate the model on the test set using appropriate metrics ; Accuracy, Sensitivity, Specificity, and AUC (Area Under the ROC Curve). Analyze the confusion matrix to understand the model's performance in predicting an individual's loan status. Interpret each of these metrics in terms of the data and the model.