Name: \_\_Sri Vamshi Polela\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

CSDA 5430 Predictive Analytics

Week 2 Assignment

1. To make numerical predictions, regression trees are implemented by adding regression into decision trees. How does a regression tree differ from a classification tree? What are the common features between regression trees and classification trees? Use Toyota car case in question 5 to support your discussion.

**Differences:**

* Regression trees and classification trees are both two types of decision trees, but they address different types of problems.

1. **Prediction:**

* Classification trees are designed for predicting the categorical target variable. Here the output is a class label i.e., “YES/NO”.
* Regression trees are used for predicting the continuous or numerical target variables. The output is a numeric or real number.

1. **Impurity measures:**

* Impurity measure for nodes in classification is calculated by two methods: Gini Index & Entropy-based measure.
* In regression trees, the impurity measure is calculated by the sum of the squared deviations from the mean of the terminal node. This is equivalent to the sum of the squared errors because the mean of the terminal node is exactly the prediction. **(SDR is the criteria.)**

1. **Evaluating Performance:**

* For Classification trees, Confusion matrix, Accuracy, Sensitivity, F1-Score etc are used.
* Whereas for regression trees, R-Squared, Mean Square Error etc are used.

**Common Features:**

* Both are represented in tree structure, but the values are different in nodes.
* Both types of trees are built on decision nodes, where features are used to split the dataset recursively until certain criteria are met.

**Toyota car case:**

* "In the Toyota car dataset, a classification tree can be constructed to categorize cars based on a specific criterion related to their mileage (KM). For example, if the mileage is less than the average, the classification tree may classify those cars as 'Good' and others as 'Bad.' The output of this classification tree is a binary decision indicating whether a car is deemed good or bad based on the defined criteria.
* A regression tree model can be utilized to predict the price of used cars based on relevant features such as mileage and year of manufacture. Unlike classification trees, which predict discrete classes, a regression tree predicts a continuous numeric output. In this case, the output of the regression tree would be a predicted price for each car. This predictive model can be valuable for estimating the resale value of Toyota cars based on specific attributes. **(Overall good. Both are supervised learning with outcome to predict. At the terminal node, classification tree performs majority vote for predicted class assignment whereas regression tree calculates average as predicted outcome value.)**

**9/10 points**

1. Given the following sample data, calculate mean, variance, standard deviation of each variable (*Age* and *Income*). In addition, calculate covariance and correlation between the two variables. Assuming *Income* is the outcome variable and *Age* is the predictor, create a scatter plot between the two. Does the scatter plot show a linear relationship? What does the correlation between the two variables tell us?

|  |  |  |
| --- | --- | --- |
| Record | Age | Income ($) |
| 1 | 25 | 49,000 |
| 2 | 56 | 156,000 |
| 3 | 65 | 99,000 |
| 4 | 32 | 192,000 |
| 5 | 41 | 39,000 |
| 6 | 49 | 57,000 |

**Solution:** To calculate the mean, variance, & Standard deviation I used the built-in R functions:

To use the below functions, load the data as data frame.

**Load the data:**



**Mean, Variance & Standard Deviation:**

A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated

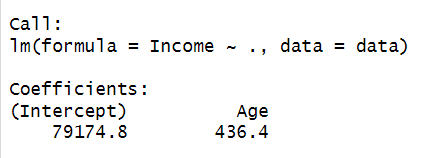
**Covariance and Correlation:**

A computer screen shot of a number

Description automatically generated with medium confidence

**Scatter Plot between Age and Income:**

**Linear regression Model:**

**A graph with a line and dots

Description automatically generated**

* Yes, there is a positive linear relationship between Age and Income. By looking at the coefficients, there is an estimated change in Income for a one-unit increase in Age is 436.4. On average, for every additional year of Age, the estimated Income increases by 436.4.
* The correlation coefficient of 0.104 between Age and Income suggests a very weak positive linear relationship. This indicates that changes in Age are not strongly associated with proportional changes in Income. **(Okay. Correlation of 0.1 indicates a very weak or no correlation. Plot shows the same.)**

1. For the dataset in question 2, assuming a liner relationship exists between the outcome (*Income*) and the predictor (*Age*), calculate the estimated slope and intercept using the formular for OLS estimation. Then, fit a liner regression model on the data. What is the estimated model expression? For the first observation, what are the actual value, fitted value, and residual of the estimation?

**Solution:**

**Ordinary Least Squares Estimation:**

* Minimized sum of the squared errors
  + The sum of squared vertical distance from each dot on the scatter plot to the regression line.

**A screenshot of a computer screen

Description automatically generatedUsing the above formula in R:**

The estimated **slope** is 436.4.

**Intercept**: 79174.79

* **Linear Regression Model:** Using lm() function to apply linear regression model on the data.

**A white background with black text

Description automatically generated**

* **Estimated Model Expression:** Income = 79174.8 + 436.4 \*Age
* A screenshot of a computer code

  Description automatically generated**The actual value, fitted value, and residual of the estimation for first observation:**

Actual value: 49000

Fitted value: 90084.42

Residual value: -41084.42

**(All good)**

1. Given the following dataset on weather play hours and the decision tree split by one predictor *Outlook*, calculate the SDR for the split. Calculate predicted mean value of playing hours (outcome variable) for each terminal node of the split.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outlook | Temp | Humidity | Windy | Hours Played |
| Rainy | Hot | High | Weak | 26 |
| Rainy | Hot | High | Strong | 30 |
| Overcast | Hot | High | Weak | 48 |
| Sunny | Mild | High | Weak | 46 |
| Sunny | Cool | Normal | Weak | 62 |
| Sunny | Cool | Normal | Strong | 23 |
| Overcast | Cool | Normal | Strong | 43 |
| Rainy | Mild | High | Weak | 36 |
| Rainy | Cool | Normal | Weak | 38 |
| Sunny | Mild | Normal | Weak | 48 |
| Rainy | Mild | Normal | Strong | 48 |
| Overcast | Mild | High | Strong | 62 |
| Overcast | Hot | Normal | Weak | 44 |
| Sunny | Mild | High | Strong | 30 |

A diagram of weather forecast

Description automatically generated

**Solution:**

**Standard Deviation reduction:** The impurity or homogeneity of a node during the tree-building process. It is commonly used as a criterion for determining the best split at each decision node.

**Formula:**

Given total number of observations = 14

For the above split, the SDR is 0.832

And predicted mean values for each terminal node are:

* Left leaf node =49.25
* Right leaf node =38.7

A screenshot of a computer program

Description automatically generated

**(Look good)**

1. **The dataset *ToyotaCorolla.csv* contains data on used cars (Toyota Corolla) on sale during the late summer of 2004 in the Netherlands. It has 1436 records containing details on many attributes, including Price, Age, Kilometers, HP, and other specifications. The dealership would like to predict the price of a used Toyota Corolla car based on its specifications so the dealer will be able to predict the profit that the dealership will get for the used cars, assuming higher priced cars generate more profit.**

**In R, perform the following steps as discussed in Ch09 on the first 1,000 cars and the following variables: *Price, Age\_08\_04, KM, Fuel\_Type, HP, Met\_Color, Automatic, Doors, Quarterly\_Tax,* and *Weight*. A description of each variable is given at the end.**

* **Step 1: Collecting data (Discuss the business problem and how the data can support the analytics.)**

**Business Problem:**

* Dealership wants to predict the price of used cars to predict the profit that the dealership will get for the used cars.

**Data Understanding:**

* The dataset "ToyotaCorolla.csv" comprises information on 1436 used Toyota Corolla cars for sale in the Netherlands in late summer 2004.
* The business objective is to predict the selling price of these cars based on key specifications, aiding the dealership in estimating potential profits.
* The selected variables, including Price, Age\_08\_04, KM, Fuel\_Type, HP, Met\_Color, Automatic, Doors, Quarterly\_Tax, and Weight, have been chosen to capture relevant features influencing car pricing.
* The analysis involves leveraging these attributes to build a predictive model that supports informed pricing decisions, contributing to enhanced profitability for the dealership.

**(Okay. The goal is to predict used car price based on other predictor values. Outcome variable is Price. Not all predators are useful for the prediction.)**

* **Step 2: Exploring data and preparing data (Use only the first 1000 rows of data and specified variables.**

Check if the categorical data are factored correctly. Create a histogram of the outcome for the data. Use the correlation matrix and *pairs.panels()* function to explore relationships between features. Create training partition (60%) and testing partition (40%) with randomized partitioning. Set a seed so the results can be reproduced.)

* **Outcome variable: Price**
* The histogram plot illustrates that most car prices in the dataset fall within the range of 7000 to 13000. This concentration indicates a common price range for the Toyota Corolla cars, suggesting a prevalent pricing trend among the observed used vehicles.

A graph of a graph showing the price of a product

Description automatically generated with medium confidence

**(It shows the skewed distribution.)**

**Using the correlation matrix and pairs.panel() :**

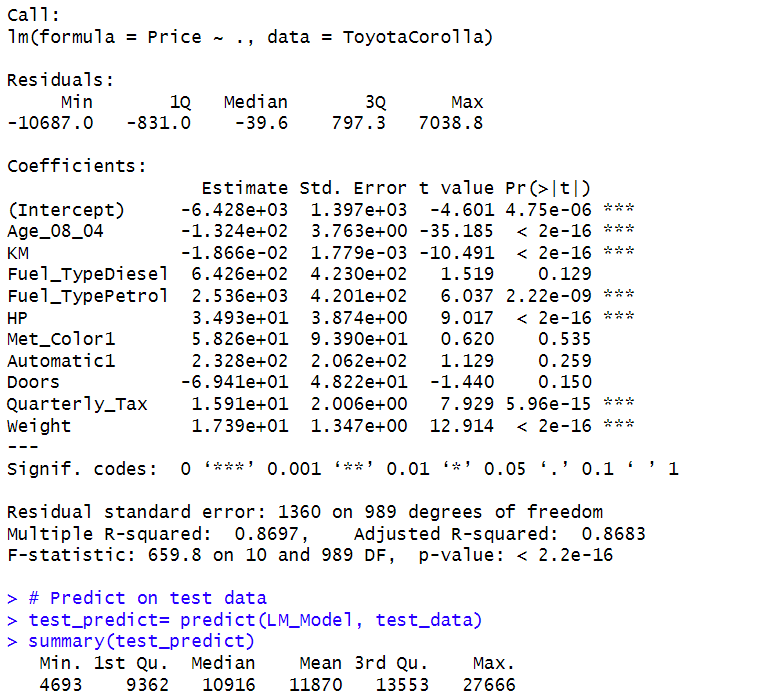
A diagram of a graph

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generated

* The correlation matrix indicates strong negative correlation between "Price" and "Age\_08\_04" (-0.87), suggesting that newer cars tend to have higher prices.
* Additionally, "Price" shows a moderate positive correlation with "Weight" (0.61), implying that heavier cars may have higher prices.
* On the other hand, there is a weak negative correlation between "Price" and "KM" (-0.60), indicating that higher mileage is associated with lower prices. These insights are further explored visually using the pairs.panels() function. **(Okay. It is unnecessary to factor on Met\_Color and Automatic, since they are already in dummy format.)**
* **Step 3: Training a model on the data (Use the function *lm()* for the linear regression algorithm. Remember to fit the model on the training data.)**

**Trained a model and fitted it on the training data**

****

**(The model looks good)**

* **Step 4: Evaluating model performance (Explore the model output and identify any insignificant predictors. Make predictions on the testing data with the model. Compare statistics (min, 1st Qu, mean, 3rd Qu, max, etc.) for predicted and observed values. Do they correlate with each other?)**

A screenshot of a computer

Description automatically generated

**Insignificant Predictors:**

* I found that Fuel\_typeDiesel, Met\_color, Automatic, Doors are insignificant.

* Yes, they are correlated.
* A correlation coefficient of 0.9276899 suggests a strong positive linear relationship between the predicted and observed values. This indicates that as predicted values increases, the observed values also tends to increase, and vice versa.

A close-up of a number

Description automatically generated

**(Comparison on other measures such as statistics?)**

* **Step 5: Improving model performance (Use *rpart::rpart()* function to fit a regression tree model on the same training data. Make predictions on the testing data with the regression tree model. Compare the performance between the regression tree model and the linear regression model. Is there any performance improvement based on your comparison?)**
* The Mean Absolute Error (MAE) is a measure of the average absolute differences between predicted and observed values. In this comparison:
* **Linear Regression Model MAE: 1066.996**
* **Regression Tree Model MAE: 1218.346**
* A lower MAE indicates better model performance. Therefore, based on the comparison, the Linear Regression Model outperforms the Regression Tree Model, as it has a lower MAE.

A computer code with blue text

Description automatically generated

**(How about other measure, such as correlation and prediction distribution as you did in previous step? The regression tree model uses much smaller number of predictors.)**

**18/20 points**

**Total 57/60 points**

**Description of Variables in Toyota Corolla Dataset**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Price | Offer price in euros |
| Age\_08\_04 | Age in months as of August 2004 |
| KM | Accumulated kilometers on odometer |
| Fuel\_Type | Fuel type (*Petrol*, *Diesel*, *CNG*) |
| HP | Horsepower |
| Met\_Color | Metallic color? (Yes = 1, No = 0) |
| Automatic | Automatic (Yes = 1, No = 0) |
| Doors | Number of doors |
| Quarterly\_Tax | Quarterly road tax in euros |
| Weight | Weight in kilograms |