Learning outcomes:

1. Multiple linear regression models
2. Transformations using Tukey's approach
3. Understanding the interpretation for categorical variables
4. Understanding the multiple R^2 and Adj R^2
5. Addressing the multi collinearity using
   i. AIC
   ii. VIF

Activity

Problem Statement

A large child education toy company which sells edutainment tablets and gaming systems both online and in retail stores wanted to analyze the customer data. They have been operating for the last few years and have maintained all transactional information data. The given data 'CustomerData.csv' is a sample of customer level data extracted and processed for the analysis from various sets of transactional files.

The objectives of today's activity are to build a regression model to predict the customer revenue based on other factors and understand the influence of other attributes on revenue.

Steps:

1. Read the data 'CustomerData.csv' into R.
2. Understand the structure of the data and perform the required pre-processing steps
   a. Drop the attribute 'CustomerID'
   b. Convert 'City' into factor
3. Look at the summary of the data and the structure of the data
4. Split the data into train and test data sets
5. Build the linear regression model and interpret the results
6. Review residual plots and analyze the model summary
7. Evaluate error metrics evaluation on train data and test data

   library(DMwR)

8. Spend the time on understanding the summary of the results and perform experiments with multiple combinations of attributes (by dropping the attributes which are not significant).
9. Standardize the data (except the target variable) and split into train and test.

Note: It is always better **not** to standardize the target variable. Let us say, you built the model with standardized target variable. Now you want to use this model to apply on the test data to predict the outcome. What will be output? The output would be in standardized form. However, you need to again **un-standardize** the results to show it to the business users. Therefore, it is recommended not to standardize the target variable.

10. Build the model again using the standardized dataset and compare the summaries with the un-standardized data.

11. Understand the interpretation for the categorical variables. (for example, “FavoriteChannelOfTransaction”)

12. Draw the scatter plots between each of the independent variable and the target variable. Study these graphs and identify if we need to do transformations. Using Tukey’s transformation as the guide lines and come up with the transformations that you can apply.

13. Build the models after making the transformations and build the models and check if these models are better than the original model (i.e.; model built before any transformations)

14. Check for multi-collinearity and perform dimensionality reduction analysis

    a. VIF
    library(car)
    vif(model_name)
    Build model by dropping attributes with high collinearity (>10) obtained from VIF

    b. AIC
    library(MASS)
    stepAIC(model_name, direction = "both")
    Build model by using attributes obtained from AIC

15. Compute the error metrics on both Train and Test data