BUS 336: Data Analytics & Visualization

Gohram Baloch

Assignment 2

Linear Regression

Total Marks: 60

Due date: 11:59 PM June 16, 2024

Instructions:

Please read the instructions carefully as your assignment will be graded based on the assumption that the instructions are followed.

1. This assignment is based on the data file `Honda_Sales.csv`, which has been uploaded to Canvas. The assignment focuses on linear Regression and should be completed individually.
2. You will need to submit:
   a. A single PDF file on Crowdmark. While answering each question, you report must include lines of R code used and the resulting output (e.g., tables, graphs).
   b. The complete .R file and tableau file (.twb) along with csv files used, to be uploaded on Canvas.
3. Late submissions will not be accepted.

DO NOT PANIC! The assignment might be a bit trickier, but if you start early and put in the effort, you'll be able to handle it just fine. Remember, you are doing great in the course. Keep up the good work!

Forecasting Honda Civic Sales in Canada

The Honda Civic is a popular compact car model that has been manufactured and sold in Canada by Honda Canada Inc. since 1972. As a car manufacturer, it is important for Honda to accurately predict monthly sales in order to make informed decisions about production and marketing. In a given period, if the company produces more units than how many consumers will buy, the company will not earn money on the unsold units and will incur additional costs due to having to store those units in inventory before they can be sold. If it produces fewer units than how many consumers will buy, the company will earn less than it potentially could have earned. Being able to predict consumer sales, therefore, is of first order importance to the company.

In this case study, we will use linear regression to predict monthly sales of the Honda Civic in Canada using economic indicators and Google search queries.

Data:

The data set used for this case study includes monthly sales figures for the Honda Civic in Canada from 2010 to 2014, as well as economic indicators such as unemployment rate, Consumer Price Index (CPI) for all goods and services, CPI for energy, and Google search queries for the Honda Civic.

Table 1: Data Codebook

Variable	Description
`Month`	the month of the year for the observation (1 = January, 2 = February, 3 = March, ...).
`Year`	the year of the observation.
`Sales`	the number of units of the Honda Civic sold in the United States in the given month.
`Unemployment`	the estimated unemployment percentage in Canada in the given month.
`Queries`	the number of Google searches for “Honda Civic” in the given month.
`CPI_energy`	the monthly consumer price index (CPI) for energy for the given month.
`CPI_all`	the monthly consumer price index (CPI) for all products; this is a measure of the magnitude of the prices paid by consumer households for goods and services (e.g., food, clothing, electricity, etc.).

Part 1: Data Exploration [15 points]

a) Load the data and split it into training and testing sets as follows: place all observations for 2012 and earlier in the training set, and all observations for 2013 and 2014 into the testing set. **[3 points]**

   i) What percentage of the actual data are in the training and testing datasets? **[2 points]**

   ii) Save your training data frame you used in part (a) as a csv file to plot the monthly sales from 2010 to 2012 in Tableau. Make sure your plot labels are clear and meaningful. **[5 points]**

   iii) List down insights you are able to derive through visual inspection of the graph. **[5 points]**

Part 2: Building the Model (35 points)

Let us now build a linear regression that could help Honda to predict future monthly sales in R. NOTE: we always use the training data for model building

a) Build a linear regression model using `Unemployment`, `CPI_All`, `CPI_Energy`, and `Queries` as the independent variables. Call it `model1` **[2 points]**

   i) What is the model's R^2? How many variables are significant assuming 90% confidence interval? **[2 points]**

b) Build a new model by including two new variables `Year` and `Month` in `model1` Call it `model2`. **[1 points]**

   i. What is the model's R^2? How many variables are significant assuming 90% confidence interval? **[2 points]**

   ii. Does `model2` performs better than `model1`? Justify your claim. **[2 points]**

c) You may be experiencing an uneasy feeling that there is something not quite right in how we have modeled the effect of the calendar month on the monthly sales of Honda. If so, you are right. In particular, we added `Month` as a variable, but `Month` is an ordinary numeric variable. In fact, we must convert Month to a factor variable before adding it to the model. What is the best explanation for why we must do this? **[2 points]**

   i) Create a new variable, call it `Month_Factor` that models the `Month` as a factor (using the `as.factor` function) instead of overwriting the current `Month` variable. We'll still use the numeric version of Month later in the problem. **[1 points]**

   ii) Build a new model, call it `model3` by replacing variable `Month` with `Month_Factor` in `model2`. **[1 points]**

   iii) What is the model's R^2? **[1 points]**

   iv) Does `model3` performs better compared to `model1` and `model2`? Justify your claim. **[2 points]**

d) You may observe that there is still room for improvement by removing insignificant variables in `model3`. Apply Step-wise Elimination approach using step() function to `model3` to build your `bestmodel`. **[1 point]**

   i) What is the model's R^2? Which variables are removed via elimination approach? **[2 points]**

   ii) Write down the regression equation for the `bestmodel`. **[4 points]**

   iii) What is the coefficient of the `Year` variable? What is the interpretation of this coefficient? **[2 points]**

   iv) Using the `bestmodel`, make predictions on the training dataset using `predict` function. Call it `predictTrain` **[2 points]**

   v) Calculate RMSE, MAE, and MAPE for the training dataset. **[3 points]**

   vi) Save your vector `predictTrain` as a csv file. Make sure to set 'row.names = TRUE' in 'write.csv' function while saving the csv file. In your Tableau workbook, go to Connections > add > Load this file to your Tableau Workbook. Under Files (left pane) > drag your `predictTrain` csv file to Canvas (Top right) to create relationship with existing table with actual sales. Under data grid (bottom pane), create relationship by selecting `ID` and `F1`, respectively. Now plot a dual line plot (one for actual sales and other for predicted sales) to compare actual vs predicted sales. Don't forget to synchronize your axis! **[5 points]**

Part 3: Model's Out-of-Sample Performance (10 points)

In part II, we were able to identify the `bestmodel`. In this section, we will see how the model performs on unseen data in the testing dataset.

(a) Using the `bestmodel`, make predictions on the testing dataset. Call it `predictTest` **[1 points]**

(b) Repeat Parts 1(a)(ii) and 2(d)(vi) to plot dual line plot for testing data as opposed to training data **[4 points]**.

(c) Calculate R^2, RMSE, MAE, and MAPE for the testing dataset. **[2 points]**

(d) Does the model perform better or worse on the testing dataset compared to the training dataset? Justify your claim by comparing the performance measures for both datasets. **[3 points]**