New York University Shanghai

Problem Set 2

BUSF-SHU 210: Business Analytics (Spring 2019)

1. Forecasting Auto Sales (Adapted from the exercise on pages 398-400 of *Analytics Edge*)

In this problem, we will try to predict monthly sales of an Auto Brand.

The file `Auto.csv` contains data for the problem. Each observation is a month, from January 2010 to February 2014. For each month, we have the following variables:

- $Month$ = the month of the year for the observation (1 = January, 2 = February, 3 = March, ...).
- $Year$ = the year of the observation.
- $AutoSales$ = the number of units of the Auto sold in the United States in the given month.
- $Unemployment$ = the estimated unemployment percentage in the United States in the given month.
- $Queries$ = a (normalized) approximation of the number of Google searches for "Auto" in the given month.
- CPI_energy = the monthly consumer price index (CPI) for energy for the given month.
- CPI_all = the consumer price index (CPI) for all products for the given month; this is a measure of the magnitude of the prices paid by consumer households for goods and services (e.g., food, clothing, electricity, etc.).

Load the data set into R and split the data set 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. You may want to use the function `subset()` (use the function `?subset` to figure out the usage of `subset()`).

(a) Build a linear regression model to predict monthly Auto sales using Unemployment, CPI_all, CPI_energy and Queries as the independent variables. Use all of the training set data to do this. Please show a screen shot of your linear regression model using the "summary" function. Clearly state the significance, the sign, and the magnitude of the association between the dependent variable and each independent variable.

(b) We would now like to improve the model by incorporating seasonality. Seasonality refers to the fact that demand is often cyclical/periodic in time. For example, demand for warm outerwear (like jackets and coats) is higher in fall/autumn and winter than in spring and summer.

In our problem, since our data includes the month of the year in which the units were sold, it is feasible for us to incorporate monthly seasonality. From a modeling point of view, it may be

reasonable that the month plays an effect in how many Auto units are sold.

To incorporate the seasonal effect due to the month, build a new linear regression model that predicts monthly Auto sales using Month as well as Unemployment, CPI_all, CPI_energy and Queries. Do not modify the training and testing data frames before building the model. Based on the model estimation results, how do you evaluate the new model compared with the original one?

(c) In the new model, given two monthly periods that are otherwise identical in Unemployment, CPI_all, CPI_energy and Queries, what is the absolute difference in predicted Auto sales given that one period is in January and one is in March? Consider again the new model, given two monthly periods that are otherwise identical in Unemployment, CPI_all, CPI_energy and Queries, what is the absolute difference in predicted Auto sales given that one period is in January and one is in May? Is there anything you feel uncomfortable about this finding?

(d) Alternatively, we consider Month as a factor variable, instead of a numeric variable. Then, we can use the binary variable technique introduced in Problem 2 of Homework 1 to build a linear regression model. Why do you think we should use the factor variable instead of the numeric variable to represent month? To convert a numeric variable into a factor variable, you may use the function `as.factor()`. To apply this function, you may type:
`Auto_train$MonthF=as.factor(Auto_train$Month)` and
`Auto_test$MonthF=as.factor(Auto_test$Month)` in the Jupyter Notebook or *R* console. In this way, you will not overwrite the original numeric variable Month.

(e) Re-run the regression with the Month variable modeled as a factor variable. (Create a new variable that models the Month as a factor. From the new regression results, what seasonality pattern have you observed?

(f) Another peculiar observation about the regression results (with month as a factor variable) is that the signs of the Queries variable and the CPI_energy variable. Why their signs are counter-intuitive? Please try to give an explanation for such phenomenon and find a way to address this issue. You may need to remove some independent variables and re-build the linear regression model.

(g) Use out-of-sample test to evaluate all your models built to estimate the sales of Auto. Report the out-of-sample R^2 of each model and discuss which model you would like recommend to this Auto Brand for their sales forecasting.

2. Election Forecasting

In this problem, you will use polling data from the months leading up to a presidential election to predict the winner by logistic regression. The file `polling.csv` contains the polling data for United States Presidential Election in 2004, 2008 and 2012. The variables are listed as follows:

- *State*: Name of state
- $Year$: Election year (2004, 2008, 2012)
- *Rasmussen* and $SurveyUSA$: Voters who said they were likely to vote Republican % - voters who said they were likely to vote Democrat %, from two major polling data resources, Rasmussen and SurveyUSA.
- $DiffCount$: Number of polls that predicted a Republican winner in the state - number of polls that predicted a Democratic winner
- *PropR*: The proportion of all polls that predicted a Republican winner
- *Republican*: Whether a Republican actually won that state in that particular election year (1/0)

Please solve the following questions.

(a) Read the data set `polling.csv` into R. Then, split the data into a training set, consisting of all the observations in 2004 and 2008, and a testing set consisting of observations in 2012.

Based on the training data set, let the baseline model be that we predict the outcome of 2012 election in each state will be the same as the outcome of 2008 election. Please Evaluate the false positive rate, the false negative rate, and the accuracy of the baseline model.

(b) A more credible baseline model would be to follow one of the polls and make a prediction. In our case, we will take the variable *Rasmussen* to make the prediction. Specifically, if the variable *Rasmussen* is positive, then the new baseline model predicts Republican will win; If negative, it predicts Democrat will win. And if the variable equals zero, the model would randomly predict which party will win.

To determine the sign of the variable, you can use the function `sign()`. Type `?sign` in R to see how to use the function.

Using the table function, we can compare the new baseline prediction (from the sign of *Rasmussen*) and the actual results for the testing set. Please take a screen shot of the confusion matrix and compute the overall accuracy. Take the cases in which the model does not know which to select as wrong predictions. Does the new baseline model outperform the original one in overall accuracy?

(c) As we start to think about building logistic regression models, we need to consider the possibility of multicollinearity within the independent variables. To some extent, some of the variables here are measuring the same thing. Compute the correlation among all variables

except for $State$ and $Year$. What do you observe? So, which independent variables would you recommend to include in the logistic regression model?

(d) Build a logistic regression model using the independent variables you recommend in part (c). Provide a screen shot of the summary of the model, and interpret the estimated coefficients of the independent variables.

(e) Build a classifier based on the logistic regression model built in part (d) (we set the threshold at t =0.5, since we do not have a preference between false positives and false negatives). Show the confusion matrix and compute the overall accuracy for the testing data set. Does your model perform better (on the testing set) than the two baseline models in terms of overall accuracy?

(f) Draw the out-of-sample ROC curve and compute the associated AUC for the logistic regression model built in part (d). How would you interpret the out-of-sample AUC in the contest of this problem?