PES University, Bangalore

UE16CS322 - Data Analytics Session: Aug – Dec 2018

Week 5 – Assignment 4

Date of Submission: 21 September, 2018 **Max Marks:** 25

NOTE: *In your assignment books, write the question, R commands required to get the desired output and the obtained output. For questions that require you to plot graphs, you will have to also print the obtained output (graphs/charts) and attach it in your assignment book. The solutions to the assignment must be* ***hand-written****.*

TOPIC: Filling in missing values, Confusion matrix analysis

Betting on horse racing or horse betting commonly occurs at many horse races. Gamblers can stake money on a horse. Gambling on horses is prohibited at some tracks; Springdale Race Course, home of the nationally renowned Toronto-Dominion Bank (TD Bank) Carolina Cup and Colonial Cup Steeplechase in Camden, South Carolina, is known as one of the tracks where betting is illegal, due to a 1951 law. Where gambling is allowed, most tracks offer pari-mutuel betting where gamblers' money is pooled and shared proportionally among the winners once a deduction is made from the pool.

Source*: https://en.wikipedia.org/wiki/Betting_on_horse_racing*

Analysis of such data can aid in finding patterns, if any, and help them take informed decisions.

This dataset has been modified for this assignment. The original dataset can be found here as 'tips.csv'. The dataset 'tips_tailored.csv' has been modified a bit for the sake of this assignment.

Note: Please use the original tips.csv only for the questions where it is explicitly asked, for comparison purposes.

Question – 1 (8 points)

Missing values in a dataset is considered to be a trivial problem most of the time. What are the different types of missing values? When do they become a serious issue?

In this dataset, the *Odds* parameter has some missing values.
Fill these missing values with

1. The mean of the column
2. The median of the column
3. The mean value for that particular horse
4. The interpolation of the data points (mean)
5. Install the Mice package and study how the imputations happen. Is it possible for it to work with this dataset?

Get the root mean squared error of the each of the newly obtained columns (against the original column from 'tips').
Plot a line graph with all of the newly obtained (filled) columns, and also plot the original *Odds* column from the original dataset 'tips'.

State which of the above methods would be the best for filling in the missing values for this dataset.

Question – 2 (8 points)

A series of tests were conducted and a Naïve Bayes classifier was used to predict whether the bet was going to be Won or Lost. The predictions were added to the dataset in the column *Predicted Results*.

a). Obtain the confusion matrix of these predicted values.

Calculate the following metrics by hand:

1. Accuracy
2. Precision
3. Recall
4. Misclassification Rate
5. F1-score
6. F score with β=2 and ½. What is the significance of β?

b). Add another column to the dataset labelled 'Predictions1'. Make the entire column predict 'Lose'. Calculate the confusion matrix and the first 4 metrics mentioned above using the newly added Predictions column. What does this indicate?

c). Why is accuracy not enough for the evaluation of a classification model? Using this dataset, explain the intuition behind each of the metrics and hence, why they are important.

Question – 3 (4 points)

World Happiness Report gives us a near insight about the contentedness of people as opposed to economic production, social support, etc. Use the data for the Years: 2015, 2016 for the purpose of training and those of 2017 as testing. *Remove the country names that have not appeared 3 years in a row for the purpose of the assignment.*

Model 1: Build a linear model combining the data from 2015 and 2016, after appending year to each, and use the parameters Economy..GDP.per.Capita, Family, year, Health..Life.Expectancy. to predict Happiness.Score.

Model 2: Build a linear model combining the data from 2015 and 2016, after appending year to each, and use the parameters Economy..GDP.per.Capita, year, Health..Life.Expectancy. to predict Happiness.Score.

Using the testing data, now check the rms error for Happiness.Score as predicted by the two models and compare the same.

Question – 4 (5 points)

The dataset in SchoolData contains information to do with students in a school. We will use logistic regression to predict if these students will successfully be able to Pass the course given their attendance and other attributes.

Pre-processing of Data

Sample class imbalance is sometimes an issue with logistic regression. Clearly, the training dataset has a major imbalance, as the number of Students who have passed outnumber the students who have failed. Correct this imbalance by upsampling the records of those who have failed.

Building of model

Implement logistic regression model on the training data.The summary statistics of the model will clearly indicate the significance of each component. Extract only the significant components from the dataset and build another logistic regression model on the new data to predict the death of a character in the series. The Akaike Information Criteria is used as an evaluation tool for the model. State which model is better using this value.

Prediction using model

Use the two models built to predict the value of ‘Pass’ parameter in the TestData. Build the confusion matrix for both the predictions.