PES University, Bangalore
UE15CS322 - Data Analytics

Session: Aug – Dec 2017
Week 5 – Assignment 5

Date of Submission: **Max Marks:** 20

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: Logistic Regression and Linear Regression

"When you play the game of thrones, you win or you die."

Breathtaking. Incredible. Ominous. Spectacular. These are some of the milder adjectives that have been used to describe the epic saga. Simply put, Game of Thrones is the greatest show of all time. When season one aired, the folks most excited for it were the fans of the book series. However, it quickly caught on to the rest of the world, inspiring a whole new set of people to take up the books. Game of Thrones can no longer be categorised as 'just a TV show'. George R R Martin's creation has become a cultural phenomenon that incites a level of devotion, love and passion that is usually only seen by massive Hollywood franchises or sports teams. Every week, people stop their daily lives because they are absolutely addicted to find out what happens next in Westeros.

Excerpts from 'GoT: Addicted to the Game', The Hindu

The Grand Finale of Game of Thrones, Season 7 aired last month and fans of GoT will once again have to wait for what feels like an eternity for more! While the fans wait, let's get our hands dirty by exploring the Game of Thrones dataset. Some brilliant work has already been done on this dataset, available on A Song of Ice and Data.

Question – 1 (10 points)

The dataset in folder A5_Q1 contains information about various characters in the series. We will use logistic regression to predict if these characters will be alive at the end of the whole series.

Pre-processing of Data

a) The training dataset contains a large number of missing values. We will use book mentions, gender, if the character has dead relations, number of dead relations, nobility, popularity, marital status, age and information about alive spouse, mother, father and heir for the prediction. Impute the missing value of age by its median and fill in the remaining missing values with -1.
b) Sample class imbalance is sometimes an issue with logistic regression. Clearly, the training dataset has a major imbalance, as the number of characters alive is greater than the number of characters dead. Represent this imbalance with a help of a pi-chart. Correct this imbalance by upsampling the records of 'dead' characters in the training dataset. Represent the division of records in upsampled data also with a pi-chart.

Building of model

c) How are categorical variables dealt with, in logistic regression? Explain.
d) 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
e) 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 and visualization

f) Use the two models built to predict the value of 'isAlive' parameter in the test data. Build the confusion matrix for both the predictions.
g) Plot the ROC curves for the predictions made by the two models and state which is a better model for the given test data.

Question – 2 (10 points)

The dataset in folder A5_Q2 contains information of viewership of each episode for the first 6 seasons. We will predict the viewership of the next 7 episodes of season 7.

Pre-processing of data

a) The viewership of each episode seems to depend on the year, episode number, number of deaths in each episode and critic ratings given before the airing of the episode. The representation of season and episode number is not very concrete (cannot be used for regression purposes). Convert it to a representation such that it is usable for regression. Name this column, 'episode_num'.
Hint: Convert it to a representation, season_num.episode.num (for example 1.01, 3.05, etc.)

Building of model

b) Use the column created by you, along with year, deaths in each episode and critic ratings and build a linear model to predict the viewership of an episode. Name the type of regression model built.
c) Most users do not look into the critic ratings before watching an episode (fear spoilers). Hence, use the remaining components from the dataset and build another model to predict the viewership of an episode.

Prediction using models and visualization

d) Use the models built to predict the viewership of the seven episodes of Season 7 (2017). Assume the number of deaths in all episodes to be the mean of the number of deaths across series and critics ratings of all episodes to be the all-time lowest of the series. The viewership ratings of all episodes of season 7 is given *here*. Calculate the root mean square of the difference in the actual and predicted viewership. Which model was better at predicting the viewership?
e) Visualize the predictions of the two models and the actual viewership of each episode of Season 7.