STA546 STATISTICAL DATA MINING-II

Final Project Proposal

Instructor: Rachael Hageman Blair

March 28, 2021

Titanic Passenger Survival Prediction using R

(3) analyze a complex data set

By

Manasa Challa – [manasasa@buffalo.edu](mailto:manasasa@buffalo.edu) (50356441)

Nishant Kapoor– [nkapoor2@buffalo.edu](mailto:nkapoor2@buffalo.edu) (50354536)

PROPOSAL

In this project we will use the Titanic dataset, which is used by many people all over the world. It provides information on the fate of passengers on the Titanic, summarized according to economic status (class), sex, age and survival. In this challenge, we are asked to predict whether a passenger on the titanic would have been survived or not. In this competition, we gain access to two similar datasets that include passenger information like name, age, gender, socio-economic class, etc. One dataset is titled `train.csv` and the other is titled `test.csv`. Train.csv will contains the details of a subset of the passengers on board (891 to be exact) and importantly, will reveal whether they survived or not, also known as the “ground truth”. The `test.csv` dataset contains similar information but does not disclose the “ground truth” for each passenger. It’s our job to predict these outcomes. Using the patterns we find in the train.csv data, we predict whether the other 418 passengers on board (found in test.csv) survived. Survival prediction is done using different methods like logistic regression, random forest, support vector machine, Naive Bayes are performed and then we bag the results.

**Objectives:-**

I. Overview of Data

The Titanic dataset has the following attributes:

* Survival: 0 = No, 1 = Yes
* pclass (Ticket class): 1 = 1st, 2 = 2nd, 3 = 3rd
* Sex: male and female
* Age: In years
* sibsp: # of siblings / spouses aboard the Titanic
* parch: # of parents / children aboard the Titanic
* ticket: Ticket number
* fare: Passenger fare
* cabin: Cabin number
* embarked (Port of Embarkation): C = Cherbourg, Q = Queenstown, S = Southampton

II. Data pre-processing

It is necessary to convert the **raw data**into a **clean data** set and dataset must be converted to **numeric data**. You have to encode all the **categorical lables** to column vectors with binary values. **Missing values** or NaNs in the dataset is an annoying problem. You have to either drop the missing rows or fill them up with a mean or interpolated values.

III. Data visualization and analysis

Exploring the data to find patterns and ideas which could help with the prediction .

IV. Predictive Modeling

Methods like logistic regression, random forest, support vector classifier, Naive Bayes are performed with default parameters. With a validation set baseline.

V. Conclusion/Result