Project Report

WIDS DATATHON 2020

Final Report

Team - 10

Team Members:

M.Devi sri Chandana(20wh1a6624)

N.Sahithi Sri(20wh1a6631)

V.Namitha(20wh1a6635)

C.Sucharitha(20wh1a6654)

N.Manvitha(20wh1a6655)

K.Satvika(20wh1a6660)

PROBLEM STATEMENT :

The challenge is to create a model that uses data from the first 24 hours of intensive care to predict patients survival.

Problem on Kaggle:

<https://www.kaggle.com/c/widsdatathon2020>

TYPE OF THE DATA SET :

The type of data set we are considering is Training data set . Training data set in Machine earning is used to create a module . This data set helps to reach the ultimate goal of predicting the results.

SIZE OF THE DATA SET :

The given data set contains of 186 rows and 186 columns.

PREPROCESSING METHOD OF THE DATA :

The Datathon has already provided us a data set of more than 130000 hospital intensive care unit visits from patients spanning a one year time frame. The data provided is from different parts of the world like Argentina, Sri Lanka, New Zealand, Brazil and more than 200 hospitals in US.

Our data set contains 186 rows and 186 columns containing patient details and his medical history . In this 186 columns we are dropped the columns which contains more than 80 percent of null values. After dropping we are left with 150 columns. In this 150 columns all the null values are replaced with the mean values of the respective columns.

Target and Prediction :

Our target column is the ‘hospital\_death’ column, we have to predict the value of patient’s survival rate.

Models :

1. Linear Regression Model:

Accuracy(80 - 20) :0.25121268902762054

This equals to 25%.

1. RandomForest Model:

Accuracy(80 - 20) :0.92417892801280283

This almost equals to 92%

1. XGBoost Model:

Accuracy(80 - 20) :0.93521719127311239

This almost equals to 93%

Conclusion :

In this project, using machine learning algorithms we have predicted the survival of a patient. We have developed two to three models. Linear Regression model is the basic model this always gives minimum accuracy . For this problem the accuracy is near to 25% in Linear Regression model. And XGBoost always gives us the maximum accuracy value. For this problem statement the accuracy in XGboost model is almost equal to 93%.And we have set to patient’s encounter id in the deployment part. In the input region we should be giving the patient’s encounter id as provided and output received is the survival rate of the patient.