Your Name Jeremy Scaria

Subject Name Titanic Dataset Analysis

INTRODUCTION

This Project aims to obtain a dataset from kaggle regarding the survival of titanic tragedy and perform EDA on it to gather valuable insights from the dataset and design a ML model to predict the survival rate for a test dataset.

APPROACH

I started by importing all the important packages to be used in the analysis such as pandas,matplotlib,numpy and other sklearn packages.I did not require pipelines as I just used pandas fillna to fill missing values so I didn't see the point in pipeline for just 1 estimator.

I got an overview of what I was dealing with from the info() and assumed logically that sex,age,fare and passenger class would be the main predicting features.

Then I started on data preprocessing and saw that cabin was missing a lot of values, I took a plot between survived and cabin to see if cabins close to the wreckage had less survival rate but sure enough the plot was more or less random. With most values missing I decided to remove it from affecting the model.

Then I saw Age,which was a valuable feature had close to 20% missing this was more hard to neglect as children had more survival rate and taking median would just reduce the survival rate as median was an age of less survival. So I decided to group missing ages with their titles in their names and take mean age of respective titles as this could be more accurate.

I then filled the missing values of embarked with mode and fare with mean.

Then I created 3 new features:

Title: the title from the name

Deck:the first letter from cabin

Totfam: total family members from SibSp and Parch

I then dropped the rest and moved to EDA where I had various plots which gave the following insights:

-First class had more survival and it decreased with the rest of the classes.This showed a negative correlation in the correlation matrix.

-Females ofcourse had more survival rate than males

-So did Age with highest for children and it decreased from that

-Fare price was more meant more survival rate implying people of influence survived more.

The correlation matric gave such and such insights but it failed for some features as the survived was a binary feature and age and fare was a numeric feature making a linear relation hard this lead to low correlations than expected.

Then I encoded the categorical features and splitted the test and train sets and also the validation sets and started working on logistic regression which gave me good precision and recall but was less compared to RandomForestClassifier when I compared accuracy.Then I tuned hyperparams for RFC and got appropriate parameters which improved the accuracy,precision and recall.

So i predicted it on the test set and created a submission csv file.

IMPROVEMENTS

Could use pipelines to make the code more more neat and use imputators rather than fillna.

Could find a way to accommodate cabin into prediction without overfitting but getting an order corresponding to survival rates.

Could find a more efficient way to account for missing ages without overcomplicating.

Could find a more unique hyperparameters using trial and error but could take time.

Could find some relation with total family members and how grouped people had more survival rates.It was used in predicting but did not plot any graphs with survival rates.

CONCLUSION

We got insights on how survival in the tragedy was associated with different features of the people on board and analysed on how each feature affected survival rate visually.We created a model with solid accuracy and precision and predicted it on the testset we got a solid accuracy of 82% after hyperparameter tuning and precision and accuracy close to 0.95.