Titanic Survival Prediction

Pratikshya Parajuli

Ministry of Finance
Government of Nepal

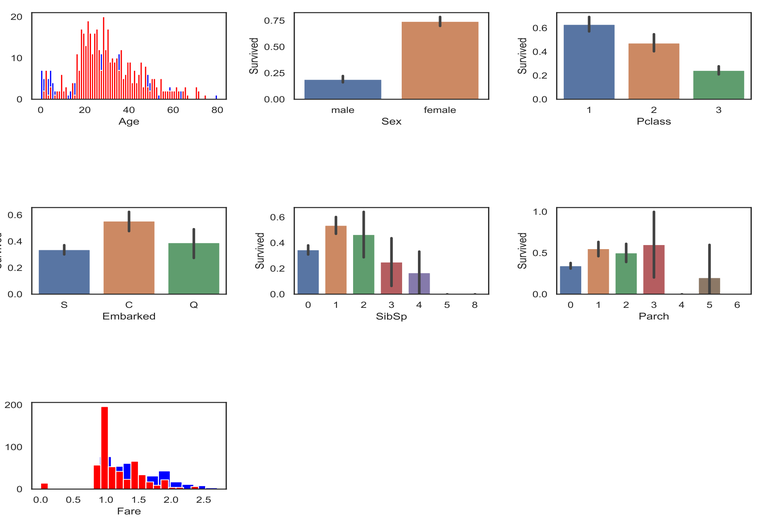
Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history. n this project, we aim to build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data (ie name, age, gender, socio-economic class, etc).
Problem Definition: Knowing from a training set of samples listing passengers who survived or did not survive the Titanic disaster, can our model determine based on a given test dataset not containing the survival information, if these passengers in the test dataset survived or not.

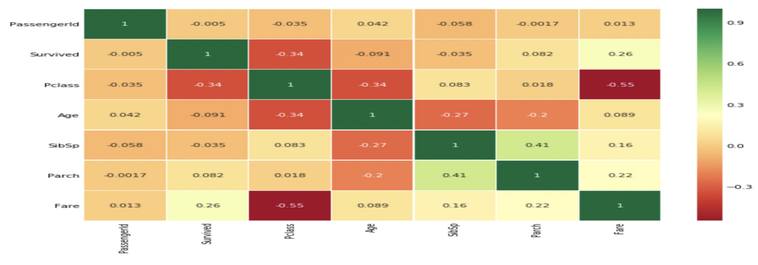
- Useful features are Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
- Target feature is Survived

Analysis of the Features

- Categorical Features in the dataset - Sex, Embarked
- Ordinal Features in the dataset - Pclass
- Continuous Features in the dataset - Age



Correlation Matrix



Interpreting the heatmap

Data Preprocessing

- Drop null values from Embarked (only 2)
- Include only relevant variables
- Define new features based on the existing ones that allow for a split into survived/not-survived with higher confidence than the existing features Eg.; Fare Cat, Age band

Removing Redundant Features

- Name: We don't need name feature as it cannot be converted into any categorical value.
- Ticket: It is any random string that cannot be categorised.
- Fare: We have the Fare cat feature, so unneeded.
- Cabin: A lot of NaN values and also many passengers have multiple cabins. So this is a useless feature.
- Fare Range: We have the fare cat feature.
- PassengerId: Cannot be categorised.

Predictive Modeling

Total sample size = 623; training sample size = 623, testing sample size = 268

	Acuracy
Radial Support Vector Machines(rbf-SVM)	0.835820895522388
Linear Support Vector Machine(linear-SVM)	0.8171641791044776
Logistic Regression	0.8134328358208955
Decision Tree	0.8059701492537313
K-Nearest Neighbours(KNN)	0.832089552238806
Gaussian Naive Bayes	0.8134328358208955
Random Forests	0.8208955223880597

At face value, some classifiers perform better than others. However, the differences between the methods are relatively small and more likely due to more or less over-fitting than anything else. There a bit more tuning might be appropriate.)

Conclusion

This is only the basic modeling of the data. The results can be further enhanced. To overcome the model variance, and get a generalized model,we can use Cross Validation.

Acknowledgement
- TULIP lab