PROJECT REPORT

Predicting Titanic survival using various machine learning algorithms and comparing the scores

Made By-

Madhura Shegaonkar – 1813120

TY-EXTC (2020-21)

m.shegaonkar@somaiya.edu

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

INTRODUCTION:

The goal of the project was to predict the survival of passengers based off a set of data. I used Kaggle competition "Titanic: Machine Learning from Disaster" (see https://www.kaggle.com/c/titanic/data) to retrieve necessary data and evaluate accuracy of the predictions. The historical data has been split into two groups, a 'training set' and a 'test set'. The training-set has 891 examples and 11 features + the target variable (survived). For the training set, we are provided with the outcome (whether or not a passenger survived). I used this set to build the models to generate predictions for the test set. For each passenger in the test set, I had to predict whether or not they survived the sinking. The score was the percentage of correctly predictions. Various machine learning algorithms were used to predict the survival chances of passengers on the Titanic.

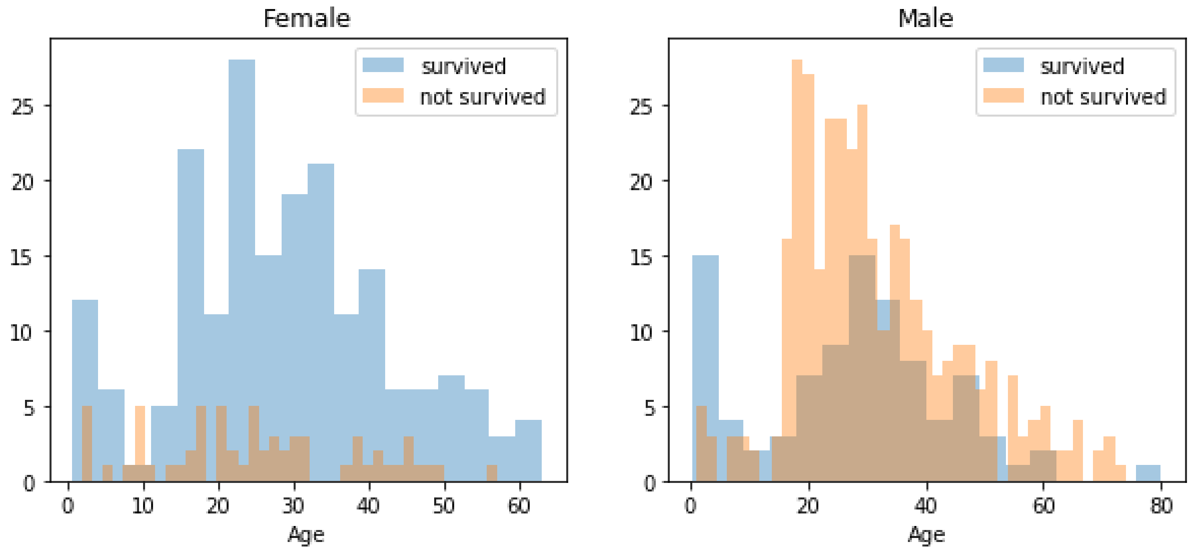
Below are the features with a short description-

Feature	**Description**
survival	Survival
PassengerId	Unique Id of a passenger.
pclass	Ticket class
sex	Gender
Age	Age in years
sibsp	no. of siblings / spouses aboard the Titanic
parch	no. of parents / children aboard the Titanic
ticket	Ticket number
fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation

DATA VISUALIZATION AND ANALYSIS:

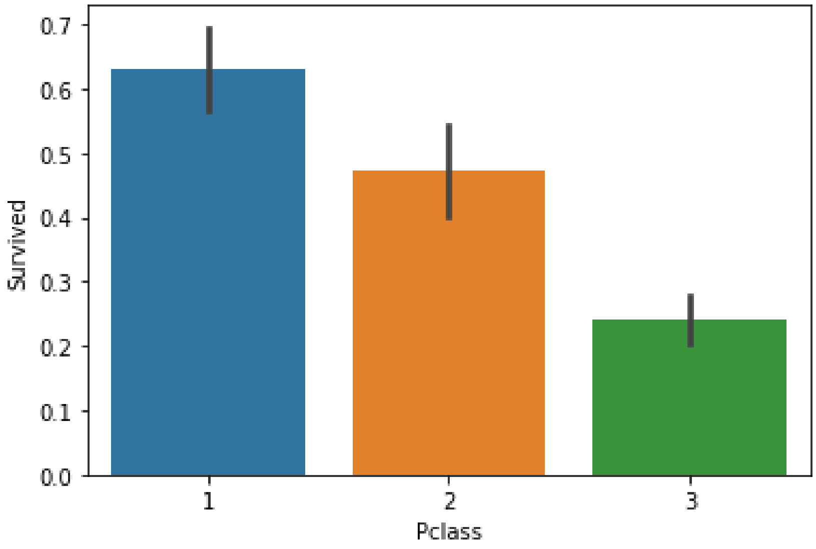
For checking the correlation of all these features with 'survival' of the passenger, the below given graphs were plotted-

1. Age and gender-



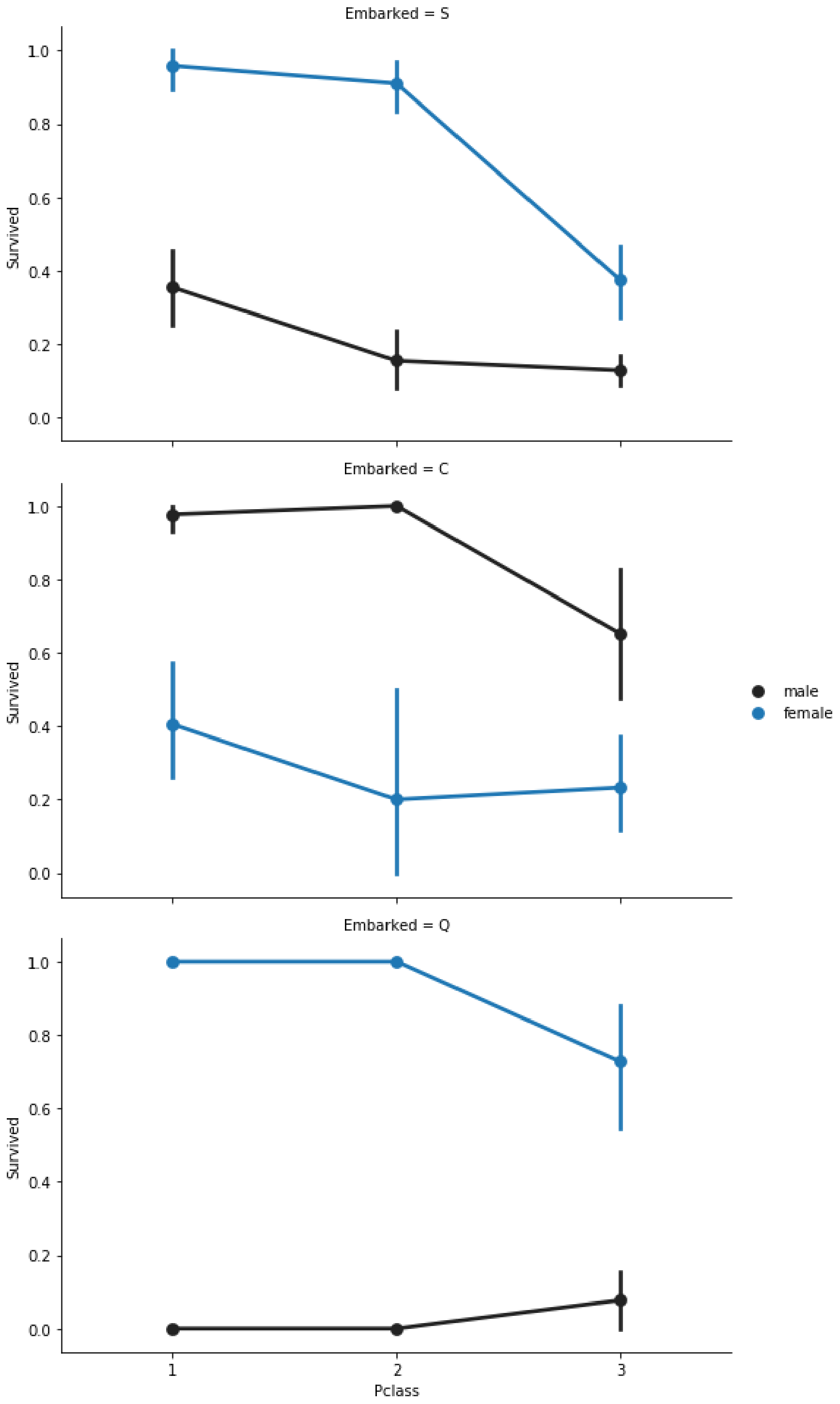
- It is clearly visible that for women the survival chances are higher between 15 and 42.
- Whereas, men have a high probability of survival when they are between 18 and 30 years old.
- For men the probability of survival is very low between the age of 5 and 18, but that isn't true for women. Also, the infants also have a little bit higher probability of survival.

2. Pclass-



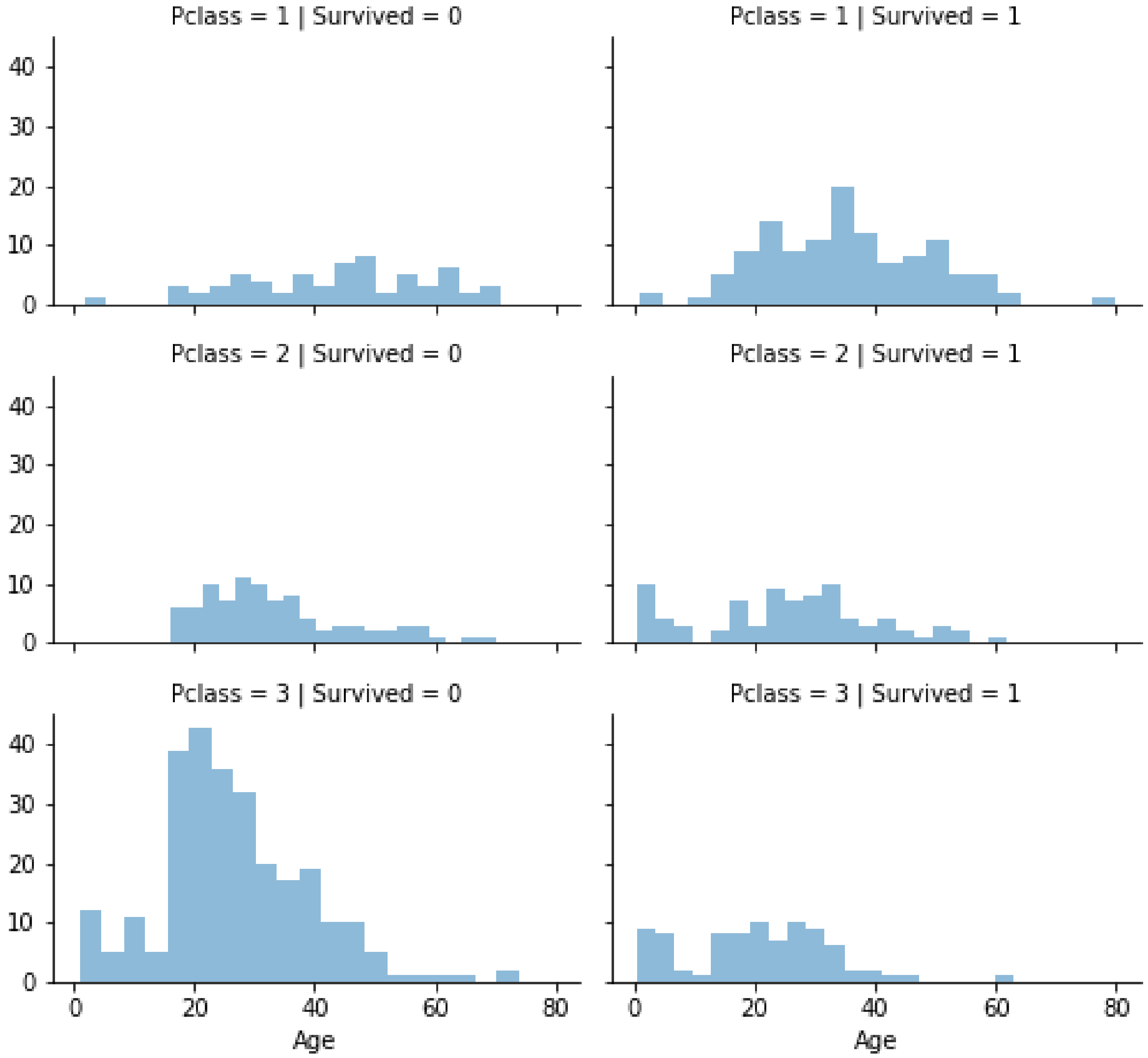
- For passengers in Pclass 1 have a higher probability of surviving than pclass 2 and 3
- For passengers in Pclass 3, the probability is the lowest.

3. Embarked, Pclass and Sex-



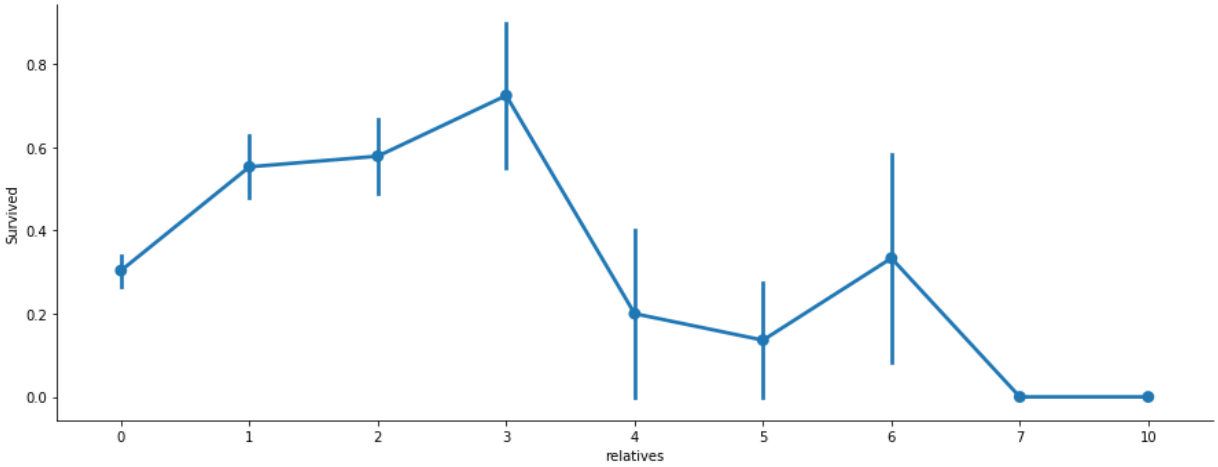
- It is clear that women on port S and on port Q have a higher chance of survival than men, but have a lower chance on port C.

4. Pclass and age-



- As visible from the graph, for passengers in Pclass 3 have a high probability of not surviving for almost all age groups.
- At the same time for Pclass 1, people from age 20 to 60 have higher survival chance.
- And in Pclass 2, people from age 0 to 10 have higher survival chance.

5. SibSp and Parch-



- Here it can be seen that a passenger has a high probability of survival with 1 to 3 relatives, but a lower one with less than 1 or more than 3 relatives.

DATA PRE-PROCESSING:

After initial data visualization, missing values were filled and numeric categories were created for cabin, name, sex, embarked, age groups and fare.

Two new features were also added in the dataset- age times class and fare per person.

MACHINE LEARNING MODELS AND RESULTS:

The final dataset was used to train various machine learning models.

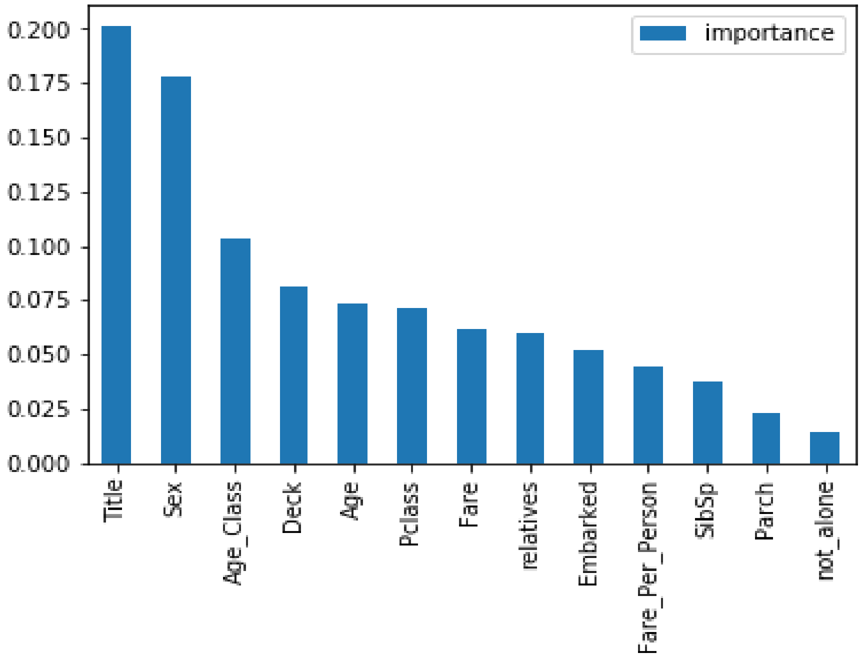
The following is the comparison table of all the algorithms used to make the predictions with their respective scores-

Model name	**Accuracy (%)**
Random Forest	92.48
Decision Tree	92.48
KNN	87.77
Support Vector Machines	82.04
Logistic Regression	81.82
Perceptron	81.82
Stochastic Gradient Decent	78.11
Naive Bayes	78.0

FURTHER EVALUATION:

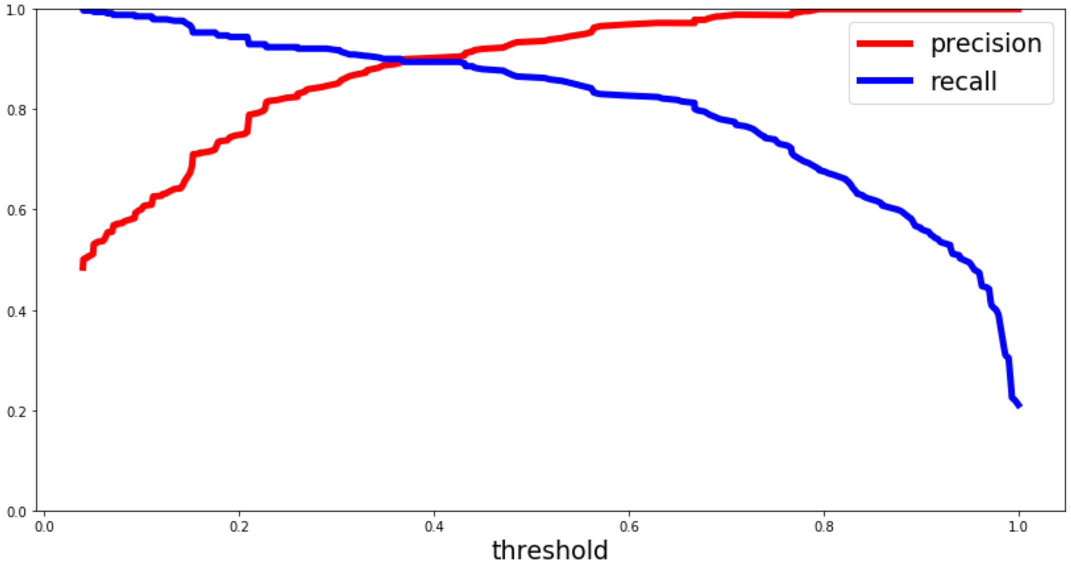
- As we can see, Random Forest has the highest accuracy, so the further evaluation was done on this model.
- Using k-fold validation on Random forest model for 5 folds, mean score came as 0.8114807607808675

- Checked the non-important features, removed ‘parch’ and ‘not alone’ as they were least important. (As shown in the below plot)



- Again, trained the model, the accuracy was 92.48%, thus overfitting was not happening.
- Calculated Precision, Recall and F1-score of the model-
  - Precision: 0.7919463087248322
  - Recall: 0.6900584795321637
  - F1-score: 0.7374999999999999

Below is the precision vs recall plot-



CONCLUSION AND FUTURE WORK:

It is evident that Random Forest Model is the most efficient algorithm for predicting Titanic survival chances. Other algorithms can also be used by increasing their accuracy through various methods such as finding the best features using SelectKBest (to get an optimal fit between bias and variance). Further improvement of Random Forest model can also be done using Hyperparameter Tuning and further evaluation can be done using ROC AUC curve.

REFERENCES:

1. https://towardsdatascience.com/predicting-the-survival-of-titanic-passengers-30870ccc7e8
2. https://www.kaggle.com/abhishekchhibber/predicting-titanic-survival-using-knn#5.-Find-the-best-features-using-SelectKBest-(to-get-an-optimal-fit-between-bias-and-variance)