Titanic Data Analysis

Name: Vankadari Sivanjanesh, **Reg No**: 11802994, **Ph.no**: 6303669968

Email id: sivanjaneesh@gmail.com

Computer Science and Engineering, Lovely Professional University, Phagwara, India

Abstract: The sinking of the Titanic caused the death of thousands of passengers and crew is one of the deadliest maritime disasters in history. One of the reasons that the shipwreck led to such loss of life was that there were not enough life boats for the passengers and crew. The interesting observation which comes out from the sinking is that some people were more likely to survive than others, like women, children were the one who got the priority to rescue. The objective is to first explore hidden or previously unknown information by applying classification models on available data set to complete the analysis of what sorts of people were likely to survive. After this the results of applying machine models are compared and analyzed on the basis of accuracy.

Keywords: Logistic regression, Random Forest, SVM, Confusion matrix, Accuracy.

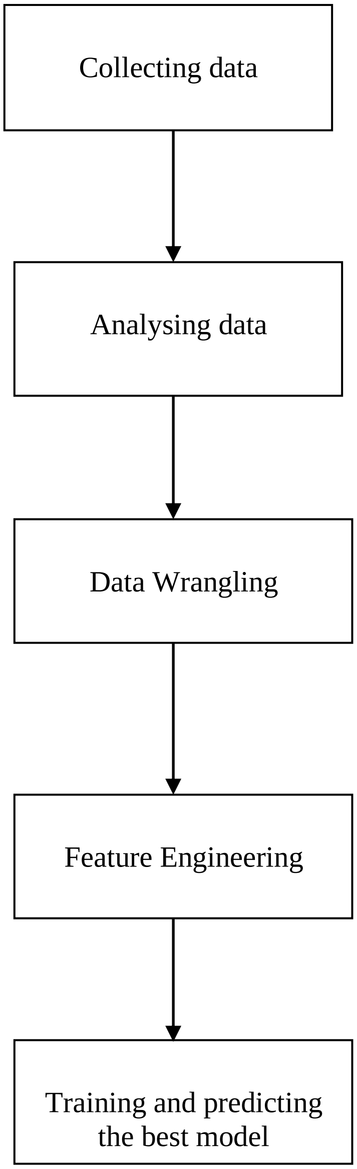
INTRODUCTION

The most infamous disaster which occurred over a century ago on April 15, 1912, that is well known as sinking of "The Titanic". The collision with the iceberg ripped of many parts of the titanic. Many classes of people of all ages and gender where present on that fateful night, but the bad luck was that there were only few life boats to rescue. The data included a large number of men whose place was given to many children on board. The men travelling in second class were dead on the vine.

Machine learning algorithms are applied to make a prediction which passengers survived at the time of sinking of the Titanic. Features like ticket, fare, age, sex, class will be used to make predictions. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data. Using the machine learning algorithms, survival is predicted on different combinations of features.

The objective is to perform exploratory data analysis to various information in the data set available and to effect of each field on survival of passengers by applying plots between every field of data set with survival field. The predictions are done for newer data sets by applying machine learning models. The data analysis will be done on applied algorithms and accuracy will be checked. Different algorithms are compared on the basis of accuracy and the best performing model is suggested.

Steps:



Step 1: Description of data set

Attribute	Desc
Survival	Survival Pasenger
Pclass	Ticket class
Age	Age of pass in years
Sex	Male or Female
SibSp	Siblings, Spouses
Parch	Parents of child
Ticket	Ticket number
Fare	Ticket cost
Cabin	Cabin no.
Embarked	From where pass. embarked

Step2: Analyzing Data

We are going to perform exploratory data analysis for our problem in the first stage. In exploratory data analysis dataset is explored to figure out the features which would influence the survival rate. The data is deeply analyzed by finding a relationship between each attribute and survival.

1.No. of persons survived and number of not survived:

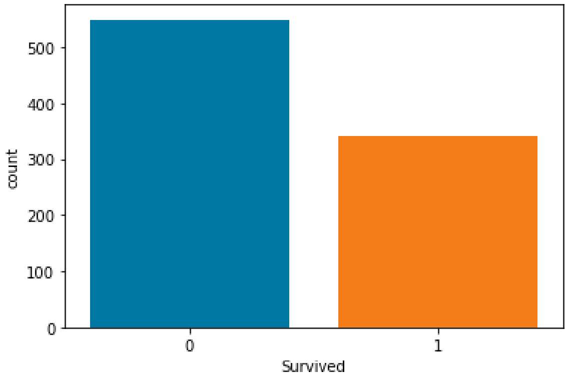


Fig 1: No. of persons survived

2.Sex vs Survival:

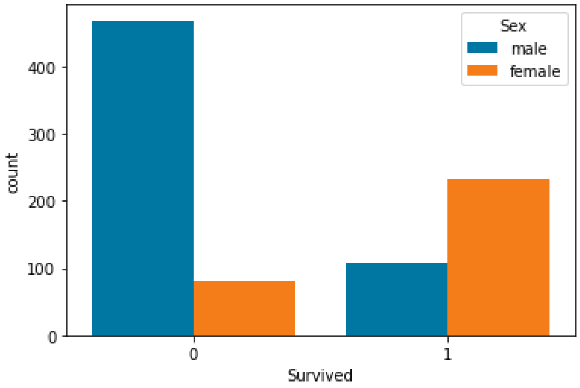


Fig:2 Sex vs Survival

From Fig.2 it is clear that females are more likely to survive than males. We calculated that survival rate of female and male are 74.20382% and 18.89081% respectively.

In similar way relationship between other attributes like fare, cabin, title, family, Pclass, Embarked and survival is found. We extracted the title from attribute ‘name’. We combined parch and sibsp. In this way we will be able to decide emphasis of each attribute on survival of passenger.

3.Survived vs Pclass:

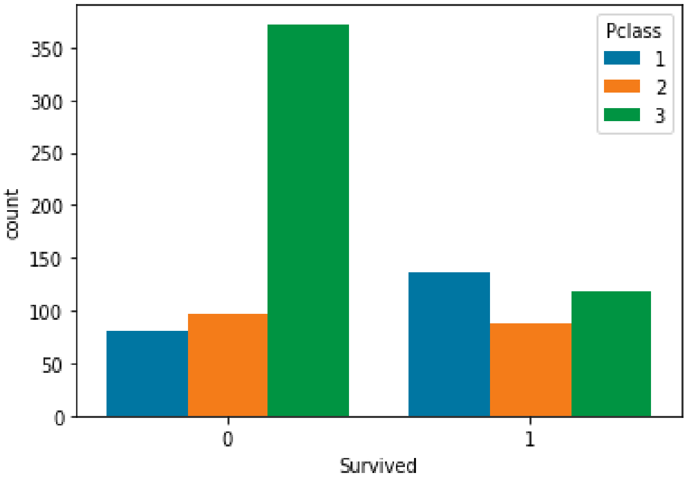


Fig 3: Survived vs Pclass

4.Age histogram:

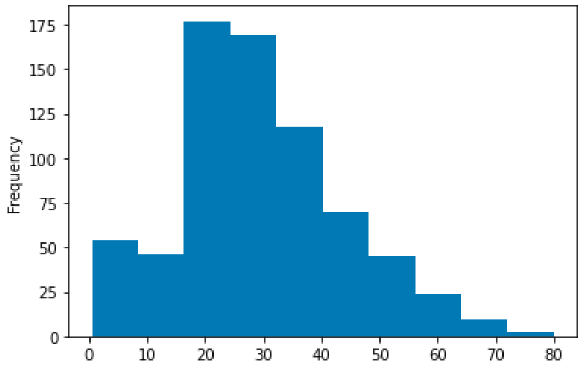


Fig 4: Age histogram

In Fig 4 we can see most of the persons are in the 20 to 30.

5. Box plot for Pclass vs Age:

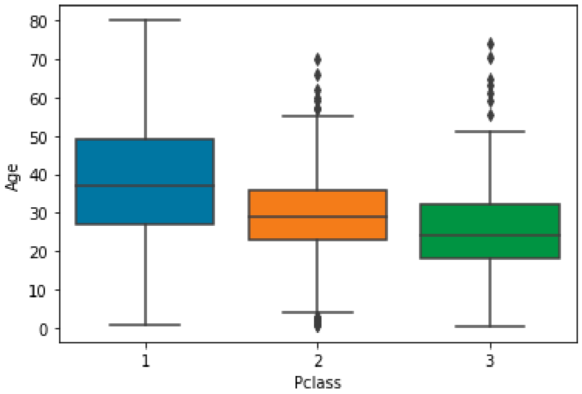


Fig 5: Box plot

6. Count plot for Siblings and Spouses:

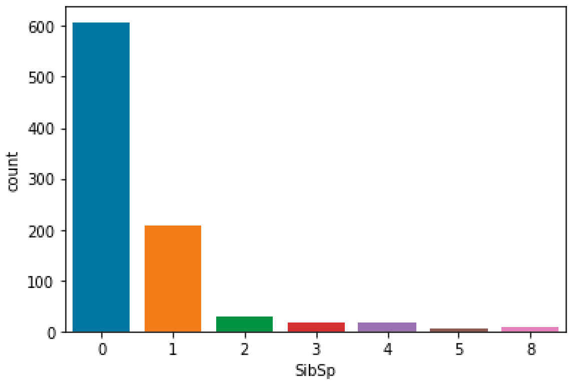


Fig 6: SibSip plot

Step 3: Data Wrangling:

Before applying any model we have to first cleaned the data. Data cleaning in sense removing unnecessary columns, checking for any null values. If any null values present in the data set either we have to delete those rows or replaced with some random values. In my case I deleted the rows which having null values as my data set is huge.

For sex column: Converted male to 1 and female to 0.

For Embark, Pclass column: Used get dummies method and attached to original data set.

Removed passenger id, name, ticket , cabin column as they are not necessary.

Step 4: Feature Engineering:

Feature engineering is the most important part. It deals with selecting the features that are selecting the features that are used in training and making predictions. In feature engineering the domain knowledge is used to find features in the dataset which are helpful in building machine learning model. It helps in understanding the dataset in terms modeling. A bad feature selection may lead to less accurate or poor predictive model. The accuracy and the predictive power depend on the choice of correct features. It filters out all the unused or redundant features.

Based on the exploratory analysis above, following features are used age, sex, sibsp, parch, fare, embarked, pclass. And then survival column is chosen as response column. These features are selected because their values have an impact on the rate of survival. These features will be the value of ‘x’ in the bar plots. If wrong features were selected then even the good algorithm may produce the bad predictions. Therefore, feature engineering acts like a backbone in building an accurate predictive model.

Step 5: Machine Learning Models:

Various machine learning models are implemented to validate and predict the survival.

1.Logistic Regression:

Logistic regression is the technique which works best when dependent variable is dichotomous (binary or categorical). The data description and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables is done with the help of logistic regression. It is used to solve binary classification problem, some of the real life examples are spam detection predicting if an email is spam or not, health-predicting, if a given mass of tissue is benign or malignant, market predicting if a given user will buy an insurance product or not.

2. Decision Tree:

Decision tree is supervised learning algorithm. This is generally used in problems based on classification. It is suitable for both categorical and continuous input and output variables. Each root node represents a single input variable and a split point on that variable. The dependent variable is present at leaf nodes. For example: Suppose there are two independent variables. i.e, input variables which are height in centimeter and weight in kilograms and the task is to find gender of a person based on given data.

There are two types of decision tree based on the type of target variable.

1. Categorical Variable decision Tree: the tree in which target variables have categorical values.
2. Continuous Variable decision tree: The tree in which the target variables has continuous values.

3. Support Vector Machine:

Support vector machine falls in supervised machine learning algorithm. This algorithm is used to solve both classification and regression problems. The classification is performed by constructing hyper planes in a multidimensional space that separates cases of different class labels. For categorical data variables a dummy variable is created with values as either 0 or 1. So, a categorical dependent variable consisting three levels, say (A, B, C) can be represented by a set of three dummy variables.

Confusion Matrix:

A confusion matrix is a method to verify how accurately the classification model works. It gives the actual number of predictions which were correct or incorrect when compared to the actual result of the data. The matrix is of the order N*N, here N is the number of values. Performance of such models is commonly evaluated using the data in the matrix.

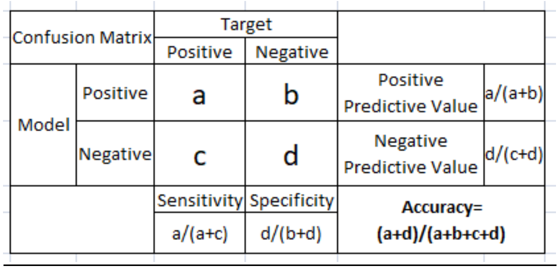
Sensitivity: It defines the percentage of actual positive which are correctly identified, and is complementary to the false negative rate. Sensitivity= true positive/(true negative + false positive). The ideal value for sensitivity is “1.0” and minimum value is “0.0”

Specificity: It measures the proportion of negatives which are correctly identified, and is complementary to the false positive rate. Specificity= true negatives/(true negatives + false positives). The ideal value for specificity is “1.0” and least value is “0.0”.

Positive Predictive Value**:** It gives the performance measure of the statistical test. It is a ratio true positive (event that makes true prediction and subject result is also true) and the sum of true positive and false positive (event that makes false prediction and subject result is also false).

Negative Predicted Value: It is the ratio of true negatives (the event which makes negative prediction and result is also false) and sum of true negative and false negative (event that makes false prediction and subject result is positive).

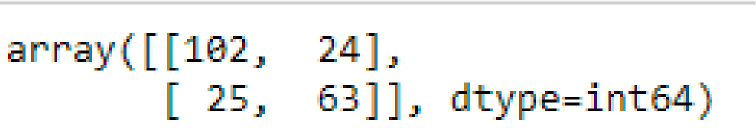
Accuracy: It gives the measure of percentage of correct prediction done by the model/algorithm. The best value is “1.0” and the worst value is “0.0”.



Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	Positive Predictive Value	a/(a+b)
	Negative	c	d	Negative Predictive Value	d/(c+d)
		Sensitivity	Specificity	**Accuracy=**	
		a/(a+c)	d/(b+d)	**(a+d)/(a+b+c+d)**	

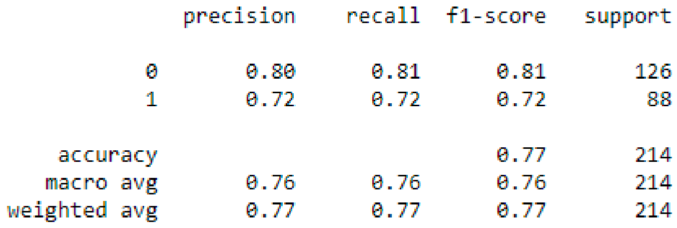
Fig 7: Generalization of confusion matrix

In R mathematical calculations are performed and accuracy using each model is found. Here are the accuracies we achieved for Logistic regression model.



```
array([[102,   24],
       [ 25,   63]], dtype=int64)
```

Fig 8: Confusion matrix for a L.R model

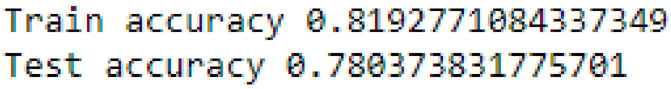


```
              precision    recall  f1-score   support

           0       0.80      0.81      0.81       126
           1       0.72      0.72      0.72        88

    accuracy                           0.77       214
   macro avg       0.76      0.76      0.76       214
weighted avg       0.77      0.77      0.77       214
```

Fig 9: classification report



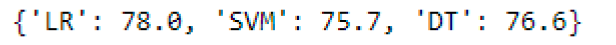
```
Train accuracy 0.8192771084337349
Test accuracy 0.780373831775701
```

Fig 10: Accuracy of a model

Performance measure of diff. Classifiers:

Performance measure of Logistic regression, Support Vector Machine (SVM), Decision tree classifier

Accuracy scores of models:



```
{'LR': 78.0, 'SVM': 75.7, 'DT': 76.6}
```

Bar Graph of models:

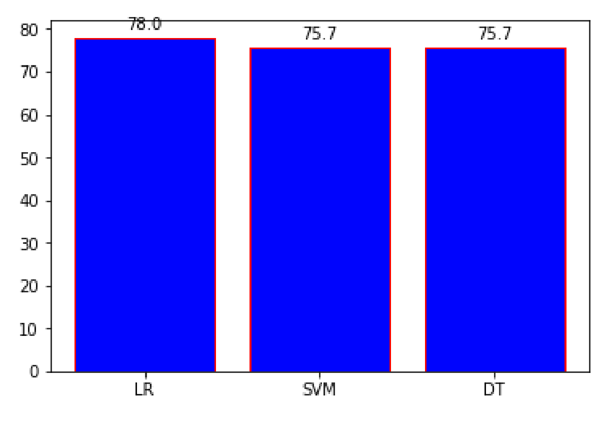


Fig 11: Performance comparison of a model

Prediction:

Here, we can choose any of the above 3 models to predict survival of test sample. Since, we have evaluated above 3 models we will predict by using model which has highest accuracy.

By seeing Fig11. We can say that Logistic Regression model is the best model as it is giving high accuracy compare to SVM, Decision tree.