**Titanic Dataset Analysis**

Sankalp Gupta

**Contents**

**1 Introduction**

**1.1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .2**

**1.2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2**

**2 Data Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**

**2.1 Data pre-processing. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4**

**2.1.1 Missing Value Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .4**

**2.1.2 Feature Selection. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .4**

**2.1.3 Feature scaling. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .5**

**2.2 EDA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6**

**2.2.1 Univariate Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .6**

**2.2.2 Bivariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .7**

**2.3 Data modelling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8**

**2.3.1. Model Selection . . . .. . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . . .. . . .. .8**

**2.3.2 Logistic Regression . . . . . . . . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . 8**

**2.3.3 Descision Tree Classifier . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9**

**2.3.4 Kneighbours Classifier. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .9**

**2.3.5 Gradient Boosting Classifier. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10**

**3 Conclusion. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10**

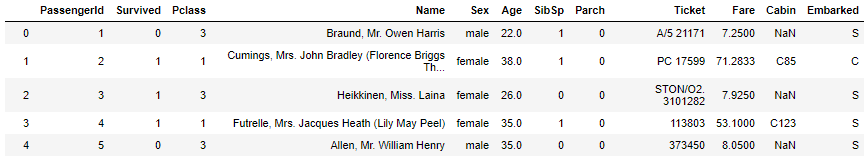
1. **Accuracy. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .10**
2. **Cross Validation Score. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11**
3. **Introduction:**
   1. **Problem Statement:**

* On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. Translated 32% survival rate.
* One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew.
* Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.
  1. **Data:**

We are given 12 features and 891 points(Rows) in the data.

Given below is a sample of the data set, we are using to cluster the behaviour of survival of the passenger.

**Table 1.1: Titanic Sample Data (Columns: 1-12)**



Below are the given variables, will help us to cluster the different behaviours of data.

**Table 1.2: Predictor variables**

|  |  |
| --- | --- |
| **Sr.no.** | **Variables** |
| 1 | PassengerId |
| 2 | Survived |
| 3 | Pclass |
| 4 | Name |
| 5 | Sex |
| 6 | Age |
| 7 | SibSp |
| 8 | Parch |
| 9 | Ticket |
| 10 | Fare |
| 11 | Cabin |
| 12 | Embarked |

**2. Data Analysis:**

**2.1 Data Pre-Processing**

**2.1.1 Missing Value Analysis**

Missing values in data is a common real world problem, we face while analysing and fitting the model. In this case, I have imputed missing values using the medianof the active feature.

The table below shows us missing value counts for each feature in our data.

|  |  |
| --- | --- |
| **Variables** | **Total #. Of missing data** |
| PassengerId | 0 |
| Survived | 0 |
| Pclass | 0 |
| Name | 0 |
| Sex | 0 |
| Age | 0 |
| SibSp | 0 |
| Parch | 0 |
| Ticket | 0 |
| Fare | 0 |
| Cabin | 687 |
| Embarked | 2 |

**2.1.2 Feature Selection**

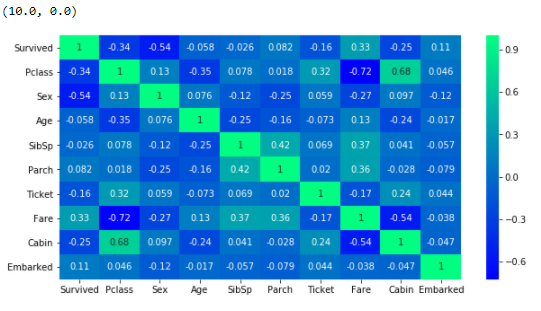
When the number of features are very large. We can’t visualize or create correlation heat map to observe which features are important and which are not. In our case we have known that have only 12 features out of which 10 features(Relevant ones).

**Observations:**

1. Fare is Positive correlated
2. Cabin and Pclass are highly correlated
3. We can drop cabin column as it is less negatively correlated to target.

Below fig 2.1 illustrates that relationship between all numeric variables using Correlation heat map

Figure 2.1 correlation heat map of numeric variables



**2.1.3 Feature Scaling**

Feature scaling is one of the most important pre-processing techniques. It majorly involves two techniques named as Normalization and standardization.

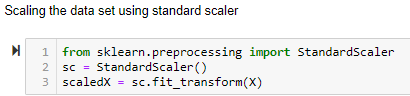
**Link to Random forest classifier based feature importance is given below,**

[**https://towardsdatascience.com/running-random-forests-inspect-the-feature-importances-with-this-code-2b00dd72b92e**](https://towardsdatascience.com/running-random-forests-inspect-the-feature-importances-with-this-code-2b00dd72b92e)

It’s important to rescale features else it may lead to wrong predictions, especially in the case of regression problems.

Rescaling data between 0 and 1 is known as feature scaling. In our case we will normalize all the features in giving data.

**Code snippet:**



**2.2 EDA(Exploratory Data Analysis)**

Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms *looking at data* refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process we will first try and look at all the distributions of the Numeric variables. Most analysis like regression, require the data to be normally distributed.

**2.2.1 Univariate Analysis**

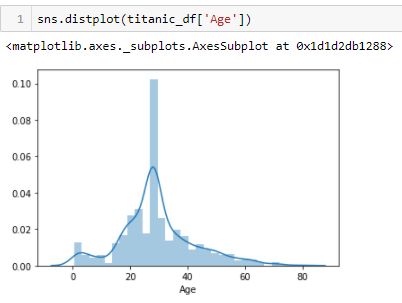
Univariate analysis of the numerical features are done which might help to predict the survival i.e. the target feature.

1. The distribution plot of fare feature.

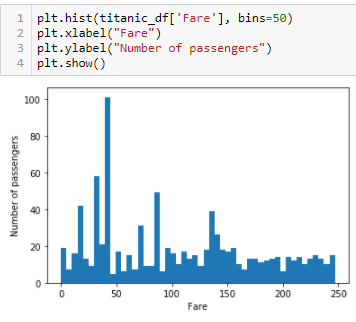
2. The histogram plot for fare feature.

3. The distribution plot for fare feature.

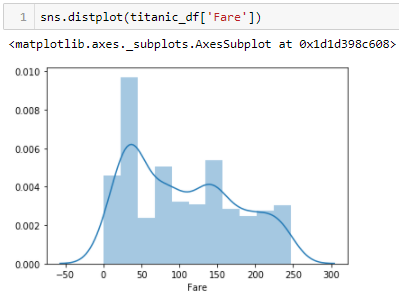
**Fig 2.2 Distribution plot for age feature**



**Fig 2.3 Histogram plot for fare feature**



**Fig 2.4 Dsitribution plot of fare feature**

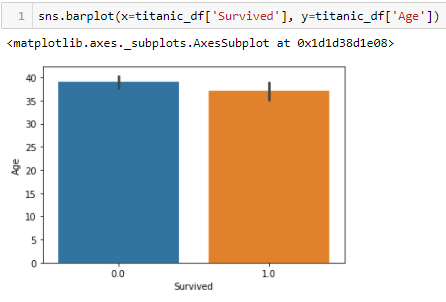


**2.2.2 Bivariate Analysis**

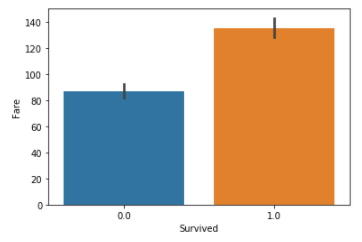
This is the analysis between the target variable and the features of the data set to analyse the contribution of the feature on the target variable.

Since the target column was categorical I preferred count plot to check if age is telling any information.

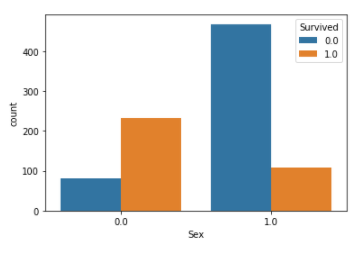
**Fig 2.5 Countplot for Age with target variable**



**Fig 2.6 Countplot for fare with target variable**



**Fig 2.7 Countplot for sex feature with target variable**



**2.3 Modeling**

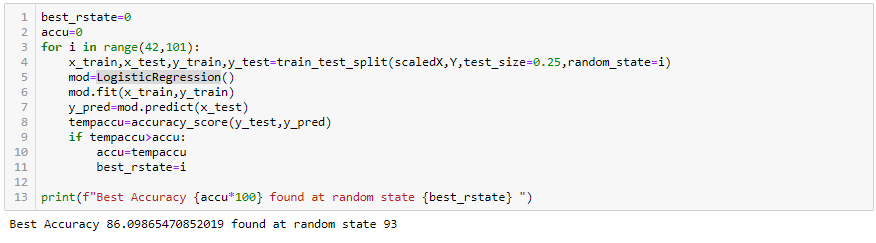
**2.3.1 Model Selection**

Since our target variable is categorical so we will go for classification methodology to go for model training.

We will start our model building from the most simplest to more complex. Therefore we use Logistic Regression at first.

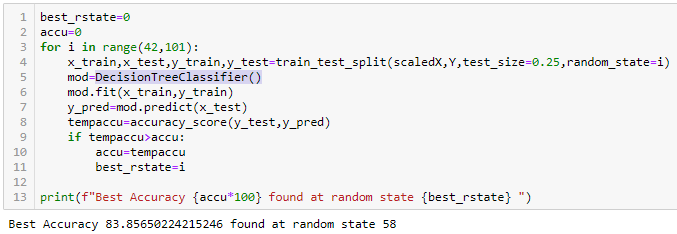
**2.3.2 Logistic Regression**

Logistic regression gave us 86% of accuracy with the train and test data. Lets see the other algorithms also.



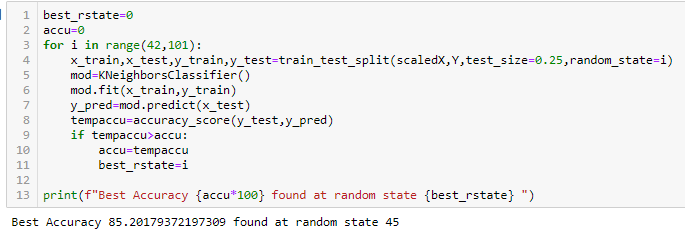
**2.3.3 Decision Tree Classifier**

Decision tree classifier gave us 84% accuracy based on the test and train data set which is less than linear regression score.



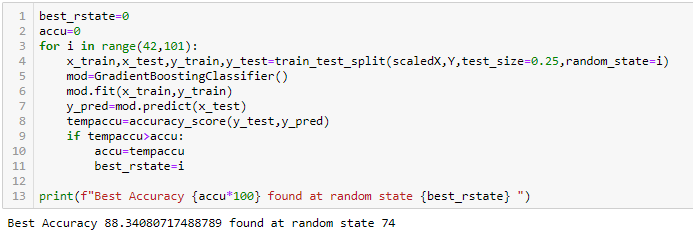
**2.3.4 KNeighbours Classifiers**

This is giving 85% accurcy which is still less than linear regression. Lets try other algorithms.



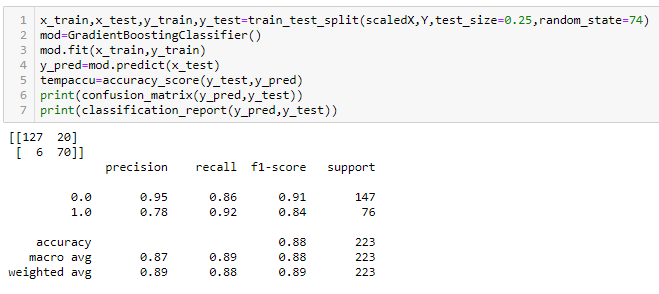
**2.3.5 Gradient boosting classifier**

This classification algorithm is giving us 88% accuracy. Will go with this to finalize the model.



**Conclusion**

**3.1 Model valuation:**

**a. Accuracy:** Gradient boosting model was having an accuracy of 88% at random state 74. So checking the accuracy with the confusion matrix.

**b. Cross Validation Score:** After validating the cross val score it says that we are neither overfitting nor underfitting.

