**Preprocessing:**

**Remove Null Values**

* The training and test data had 263 **null** values in the age column, which I replaced with median of ages = 28
* The test data contains 1 **null** value in the fare column which I replaced with mean fare values = 33.29

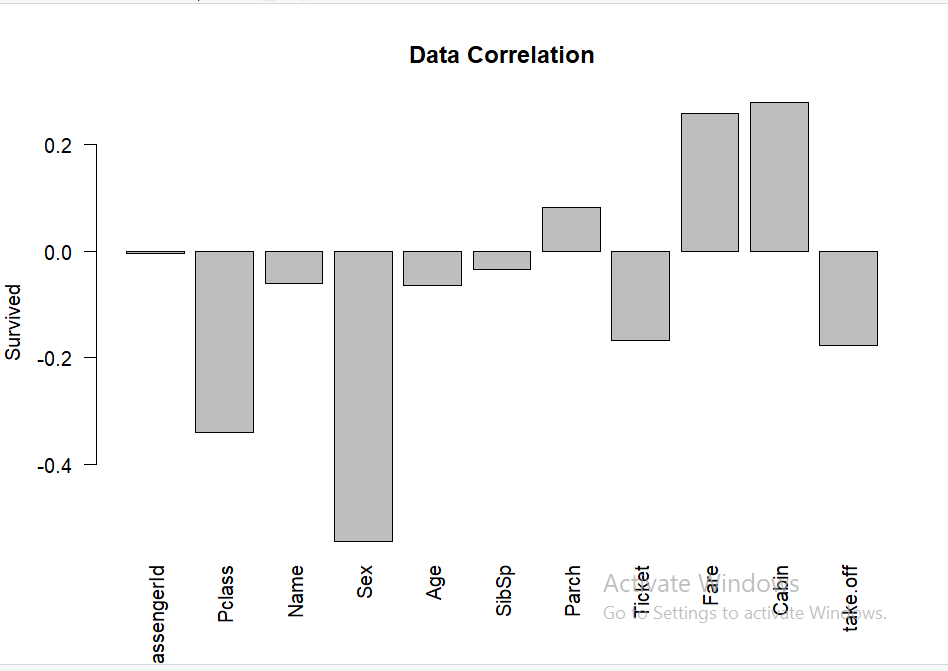
**Encoding**

* Convert categorical columns to numeric, namely: name, gender, take off, cabin, and ticket columns

**Data Scaling**

**Feature Selection**

* Extract the features that have high correlation with the Survived column which are Pclass, gender, ticket, fare, cabin and take off



* Apply Different Models and select the best accuracy which is Random Forest Algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Training Accuracy | Validation Accuracy | Testing Accuracy |
| Naive Bayes | 76.69 % | 77.65 % | 73.4 % |
| Logistic Regression | 79.21 % | 78.21 % | 75.8 % |
| Decision tree | 81.46 % | 78.77 % | 77.7% |
| SVM | 83.15 % | 79.33 % | 77.77% |
| Random Forest | 95.22 % | 85.47 % | 78.2% |