**Titanic – Machine Learning from Disaster**

* **Predictor Table**

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Rationale** | **Effect** |
| Pclass | Travelers with higher socio-economic status might be given preference over lower socio-economic status people in case of emergencies | + |
| Sex | Ideally this may or may not have any effect on survival, but we will use this column to check any gender bias on survival. | + or - |
| Age | Young travelers might have greater strength and endurance which might help them when the ship wreaks. | - |
| SibSp | Having siblings or spouses will affect the survival probability of the travelers as he/she might have to make sure that their siblings or spouses are safe as well | - |
| Parch | Having parents or children aboard on titanic might reduce the chances of survival. | - |
| Fare | More expensive fares may indicate that in case of emergencies these passengers may be provided priority. | + |
| Embarked | Ideally there should be no effect on survival but since each embark port is from different country passenger might have different survival strategies which might help them. | + or - |
| Excluded Variables | | |
| PassengerId | This is unique value column therefore it won’t have any effect on their survival. | NA |
| Name | Since every passenger has unique name or even if it repeats it won’t have any effect on their survival. | NA |
| Ticket | Ticket number will not have any effect on the survival of the passenger. | NA |
| Cabin | Majority of data is blank in this column, therefore there it will not add any value to the model. | NA |

* **Feature Engineering**

1. SibSp – The count of siblings or spouses won’t help the model because the data is biased. We have maximum number of travelers with 0 count. Therefore, changing the column to binary values as follows,

0 – No siblings or Spouses

1 – Have siblings or Spouses

1. Similarly, for Parch, changing the column to binary values as follows,

0 – No parents or children

1 – Have Parents or children

* **Build ML Classification Models**

Below are the top 3 ML Models,

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 77% |
| XGBoost | 77% |
| SVM Classifier | 75% |