# Personal Project No 2 Summary

# “Titanic : Machine Learning from Disaster Kaggle”

**Problem Definition**: Knowing from a training set of samples listing passengers who survived or did not survive the Titanic disaster, can our model determine based on a given test dataset not containing the survival information, if these passengers in the test dataset survived or not.

**Acquire Data**: As the data is given on Kaggle Website , I just downloaded the data . Different types of features (Categorical, Numerical) were present in the data .

| Types of Features | | | | |
|---|---|---|---|---|
| **Categorical** | **Nominal** | Survived | Sex | Embarked |
| | **Ordinal** | Pclass | | |
| **Numerical** | **Continuous** | Age | Fair | |
| | **Discrete** | SibSp | Parch | |
| **Mixed** | Ticket | Cabin | Names | |

## Exploratory Data Analysis:

- **These were some early insights after initial analysis.**

Total passengers were 891 and greater than 75% did not traveled with parents or children. About 30% had siblings and/or spouse abroad. Less than 1 % were paying as high as $512. And less than 1 percent were within age range 65-80.Cabin , Age , Embarked features contained null values. Ticket feature had high ratio (22%) of duplicate values (unique=681)

- **After doing analysis by visualization , I observed following things .**

I observed most passengers in Pclass = 1 survived and Pclass = 3 did not survived. significant correlation (>0.5) among Pclass=1 and Survived . females had very high survival rate at 74% with Exception in Embarked=C where males had higher survival rate. Infants and old passengers had high survival rate and Large number of people 15-25 year olds did not survive. Higher fare paying passengers had better survival and Port of embarkation correlates with survival rates

## Wrangling:

After collecting several assumptions and decisions regarding our datasets and solution requirements, I did some cleaning and wrangling .

- As ticket feature was providing no information , so I dropped it.

- Null values were filled with their median and mode values respective features.

## Feature engineering :

Feature engineering plays a vital role in improving our results . I made several features but found that some of them were useful .

| Feature | Description |
|---|---|
| has_cabin | True: if cabin value is avaliable |
| Fam_size | Family size feature was created by adding Parch and SibSp |
| Title | Title was extracted from names |
| Is _Alone | 0 : When Fam_size =1 |
| Age*Class | Multiply Age and Pclass |
| Fare_Per_Person | Fare/Fam_size+1 |
| Sex_target_enc | Mean encoding of Sex |
| E_freq | Frequency encoding of Embarked |

Non-Numerical Values were converted to numerical values as many libraries could only deal with numerical values. In the end feature looked like this.

| | Pclass | Sex | Age | Fare | Embarked | has_cabin | Fam_size | Title | IsAlone | Age*Class | E_freq | Fare_Per_Person | Sex_target_enc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 3 | 0 | 1 | 0 | 0 | False | 1 | 1 | 0 | 3 | 0.725028 | 0 | 0.383838 |
| **1** | 1 | 1 | 2 | 3 | 1 | True | 1 | 3 | 0 | 2 | 0.188552 | 3 | 0.383838 |
| **2** | 3 | 1 | 1 | 1 | 0 | False | 0 | 2 | 1 | 3 | 0.725028 | 1 | 1.000000 |
| **3** | 1 | 1 | 2 | 3 | 0 | True | 1 | 3 | 0 | 2 | 0.725028 | 2 | 1.000000 |
| **4** | 3 | 0 | 2 | 1 | 0 | False | 0 | 1 | 1 | 6 | 0.725028 | 1 | 0.000000 |

## Model used and Results :

I used following algorithms to get results .

| | |
|---|---|
| Logistic Regression | Extra Tree |
| Random forest | Xgboost |
| Decision Tree | Adaboost |

Decison Tree was underfitting the data while Xgboost and Adaboost was overfiting. The best accuracy of 0.803 was achieved using Extra Tree. This model placed me among top 11 percent on the leader board. Verify