Final Report

In this report, we are going to show the full pipeline of data mining. We chose titanic dataset for classification problem

1. Problem Definition

- Given information about the passenger of titanic, can we predict whether or not they survived after crash?

2. Features

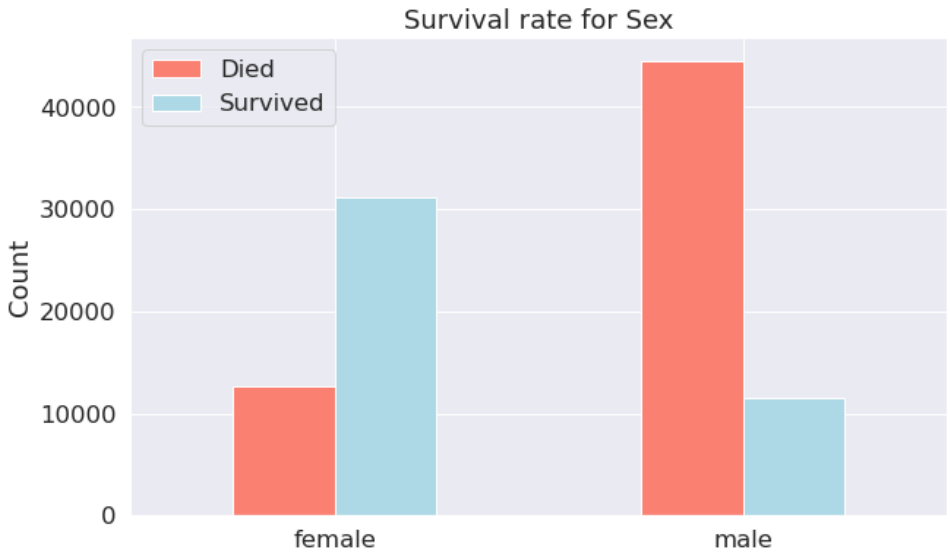
- **Survived**: Outcome of survival (0 = No; 1 = Yes)
- **Pclass**: Ticket class
- **Name**: Name of passenger
- **Sex**: Sex of the passenger
- **Age**: Age of the passenger
- **SibSp**: Number of siblings / spouses aboard the titanic
- **Parch**: Number of parents/ children aboard the titanic
- **Ticket**: Ticket number of the passenger
- **Fare**: Passenger fare
- **Cabin** Cabin number of the passenger
- **Embarked**: Port of embarkation of the passenger (C = Cherbourg; Q = Queenstown; S = Southampton)

3. Hypotheses

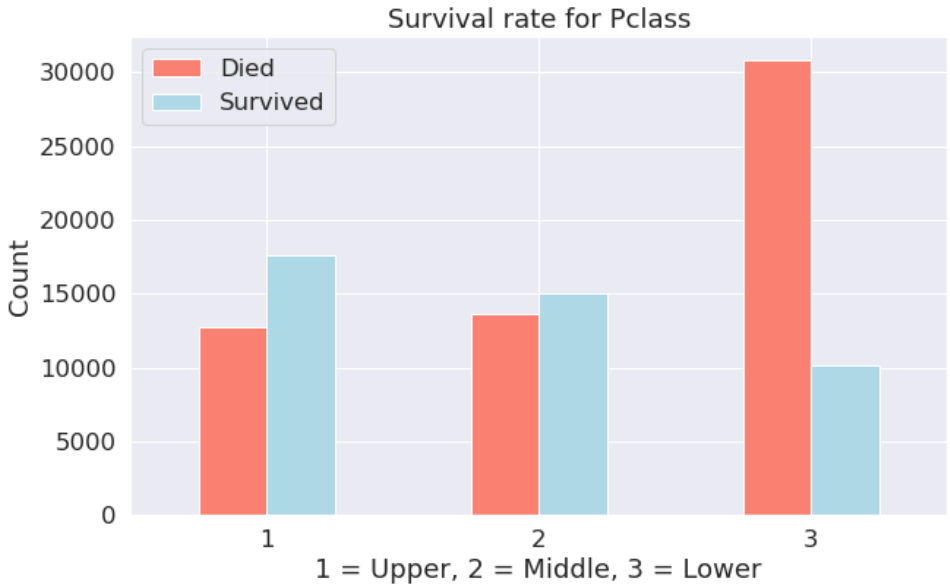
- **Women** were more likely to have survived in crash.
- **The passengers with high socio-economic status** were more likely to have survived in crash.
- There can be high correlation between **pclass** and **fare** as high-class passengers probably paid more for the ticket.
- **Senior people** were probably to have survived in crash
- **Men who traveled alone** were most likely to have died in crash
- **Men with low socio-economic status** were most likely to have died in crash
- **Women with high socio-economic status** were most likely to have survived in crash

Data Exploration

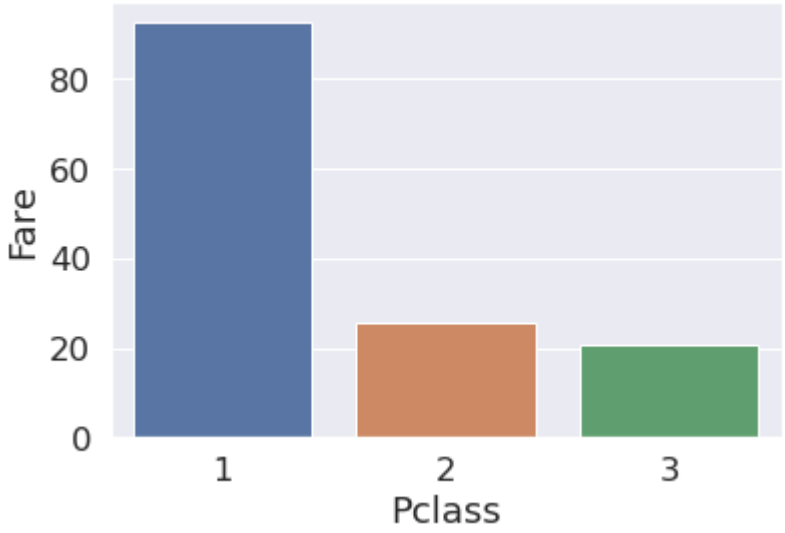
- We used bar plot to show the relation between **Sex** and **Survival rate** and as can be seen from the graph, our first hypothesis is correct



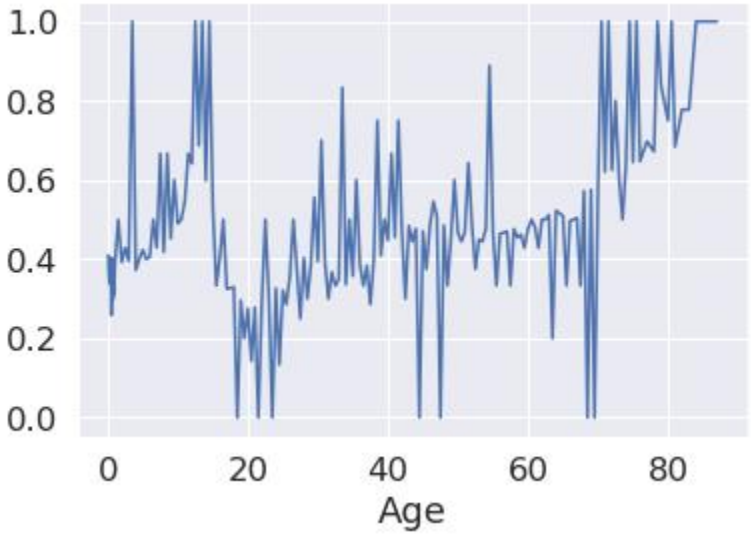
- We used bar plot to show the relation between **Socio economic status** and **Survival rate** and as can be seen from the graph, our second hypothesis is correct



- We used bar plot to show the relation between **Fare** and **Socio-economic status** and as can be seen from the graph, our third hypothesis is correct. As we move down in status the ticket price decreases.



- We used line plot to show the relation between **Age** and **Survival rate** and as can be seen from the graph, our fourth hypothesis is correct because survival rate for the ages after around 75 is almost close to 1 which implies that most of them are survived

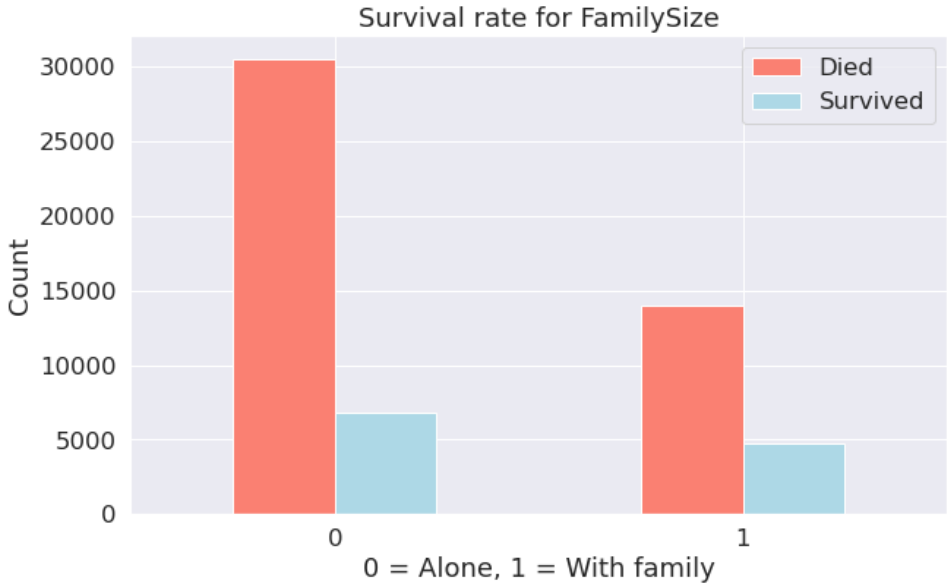


For the other hypotheses, we will show the graphs (or tables) after data processing and feature engineering.

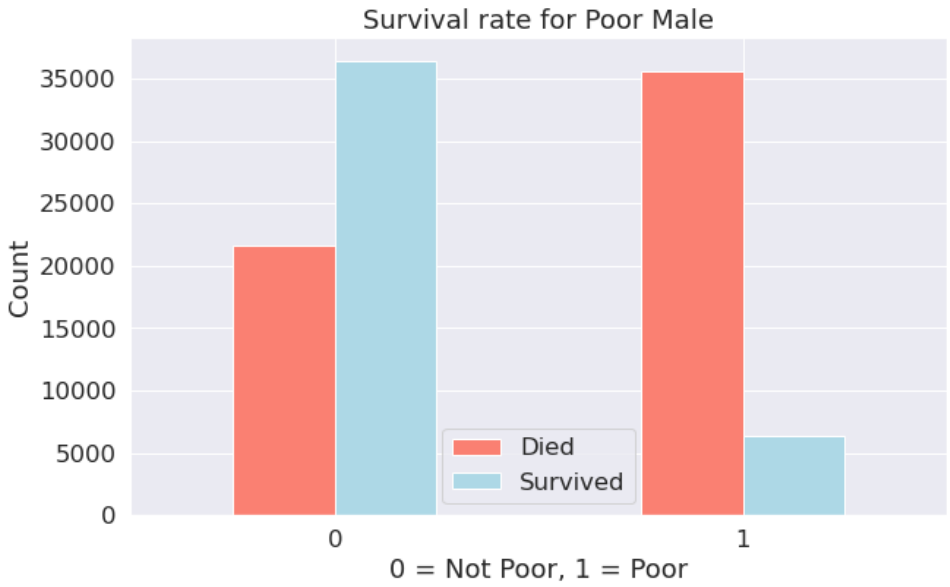
Data Preprocessing and Feature Engineering

- We will make a new column called **FamilySize** that will be sum of **SibSp and Parch**.
- As **FamilySize** column is imbalanced so to make it simple, it will consist of binary values. (0 => people traveled alone, 1 => people traveled with family).
- There is unnecessary information about passengers. We will remove **Passenger ID, Ticket, Name, Cabin** from data
- **Age**, **Fare**, **Embarked** columns have NaN values. NaN values of **Age** and **Fare** will be filled with median of each respectively and NaN values of **Embarked** will be filled with the most frequent value
- We will make 3 columns (**Poor, Middle, Rich**) based on the values of **Pclass**
- We will make a new column called **Rich Female** who are the females with **Pclass** of 1 or 2
- We will make a new column called **Poor Male** who are the males with **Pclass** of 3
- We will make two columns (**High Fare**, **Low Fare**) based on the values of **Fare**
- We will make a new column called **Female C** who are the females that embarked from **Cherbourg**
- We will make a new column called **Male S** who are the males that embarked from **Southampton**
- We will make a new column called **Adult Male** who are the males with age between 20 and 40
- We will make a new column called **Senior People** who are over 75 years old
- Finally, we will remove **Age**, **Fare**, **FamilySize**, **Pclass, Embarked, Parch** and **SibSp** columns

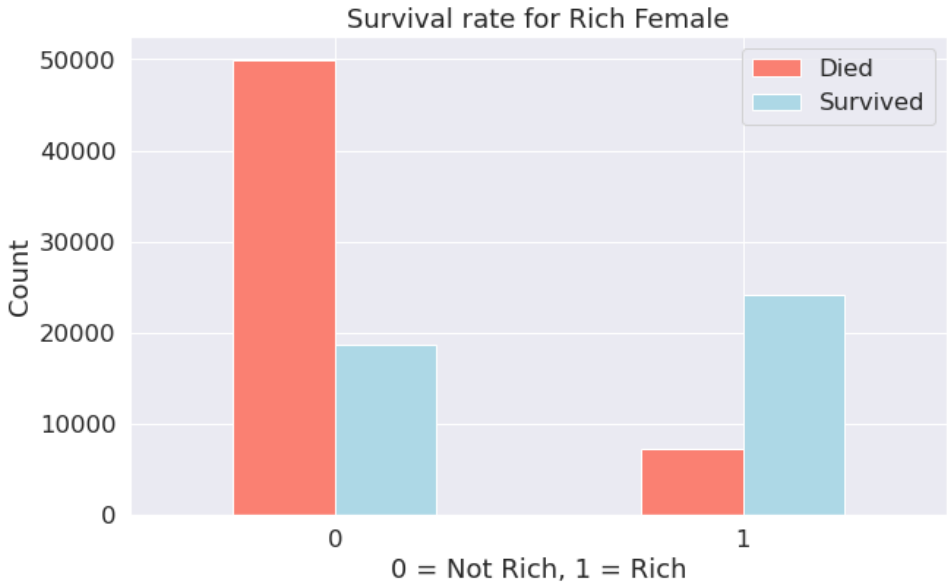
- We used bar plot to show the relation between **Alone men** and **Survival rate** and as can be seen from the graph, our fifth hypothesis is correct.



- We used bar plot to show the relation between **Poor Male** and **Survival rate** and as can be seen from the graph, our sixth hypothesis is correct.



- We used bar plot to show the relation between **Rich Female** and **Survival rate** and as can be seen from the graph, our seventh hypothesis is correct.



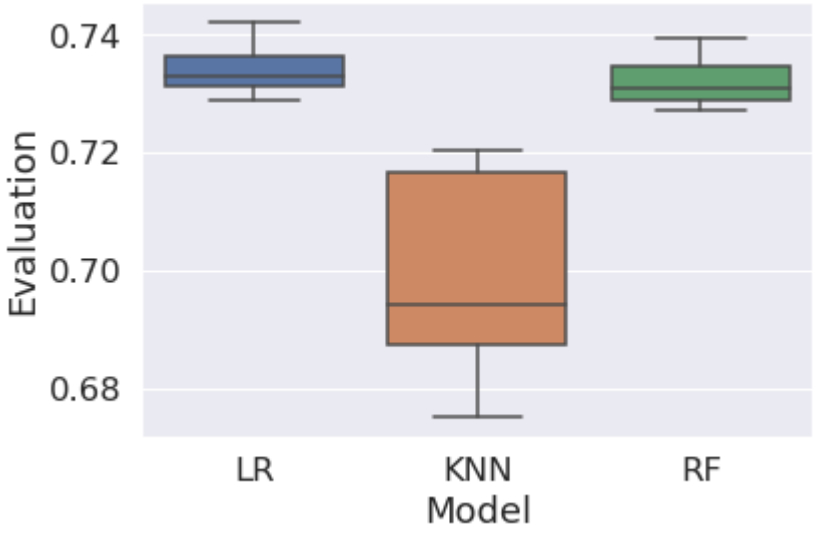
Modelling

We tried 3 different machine learning models for this dataset

- Logistic Regression

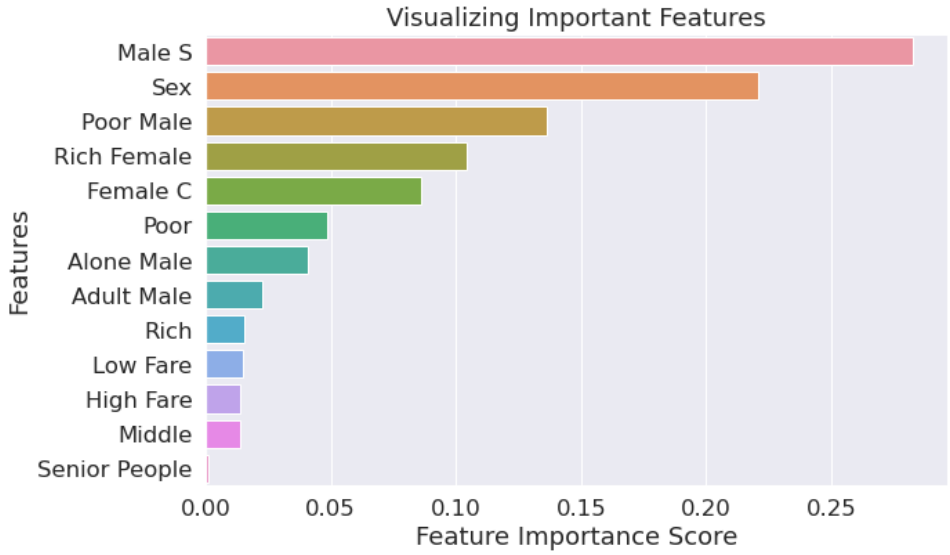
- K-Nearest Neighbours Classifier
- Random Forest Classifier

Firstly, we split the dataset to test and train sets. The test set contains 5% of the whole dataset. Next, we ran 3 models with default parameters in 10-fold cross validation so that the models do not overfit. As a scoring value, we chose F1 score since it is balancing the precision and recall.



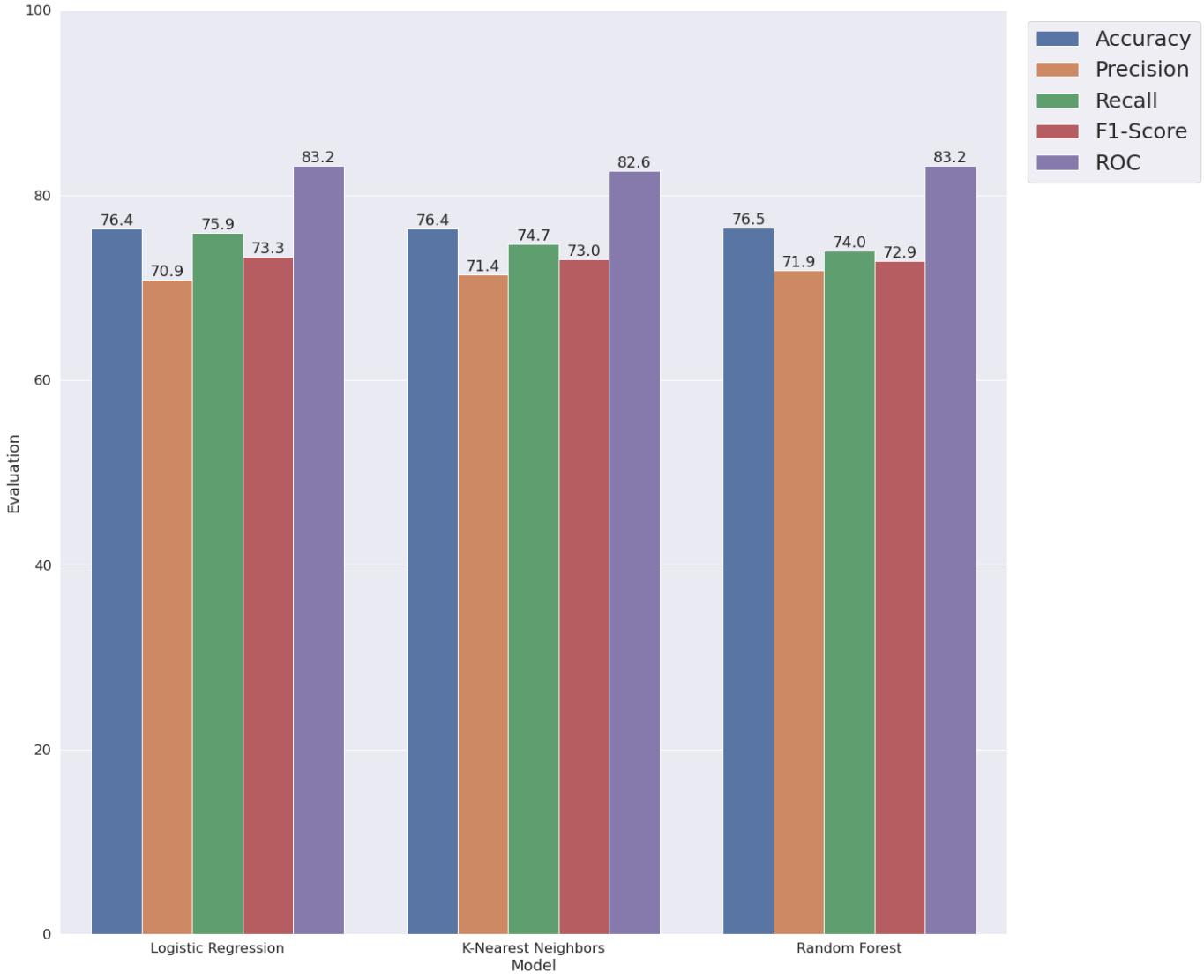
As can be seen from the graph, Logistic Regression (LR) and Random Forest (RF) did better than K-Nearest Neighbor in terms of F1 score on training dataset.

As an additional process, we trained Random Forest Classifier and visualized the importance of each feature as follows:



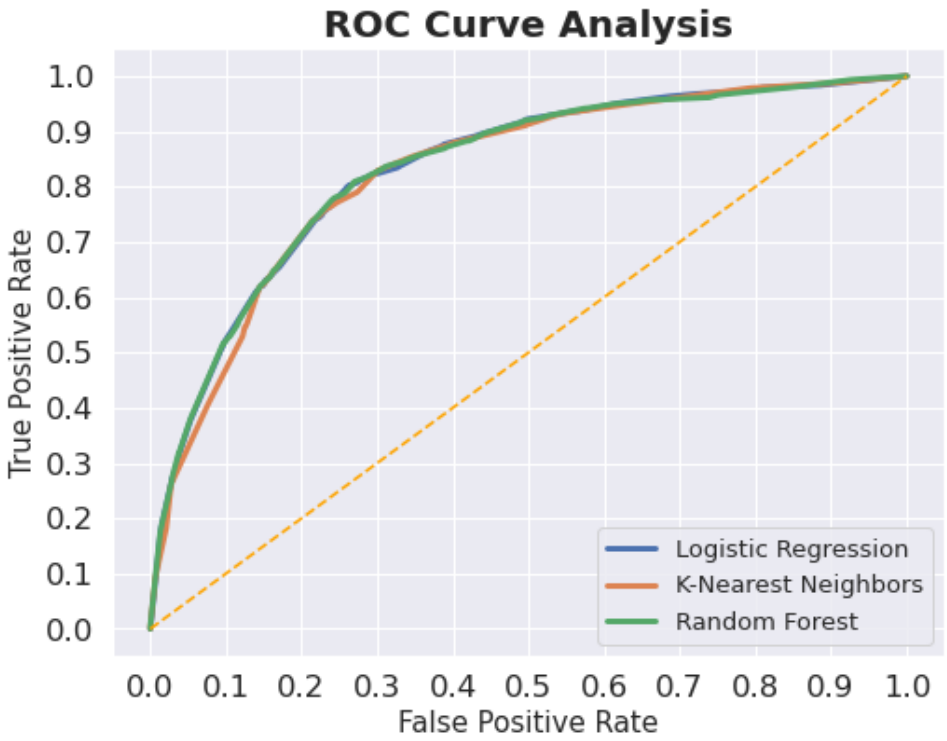
As can be seen from the graph, our generated new columns are affecting to the target value. This implies that our feature engineering works fine.

For the next step, we tried to find the best parameters for each classifier by using **RandomizedSearchCV** with 10-fold, F1 scoring and 30 iterations for each model. We did not go for **GridSearchCV** as the training of 3 models can take ages. After finding the best parameters we trained the models with these parameters and compared with various evaluation metrics.

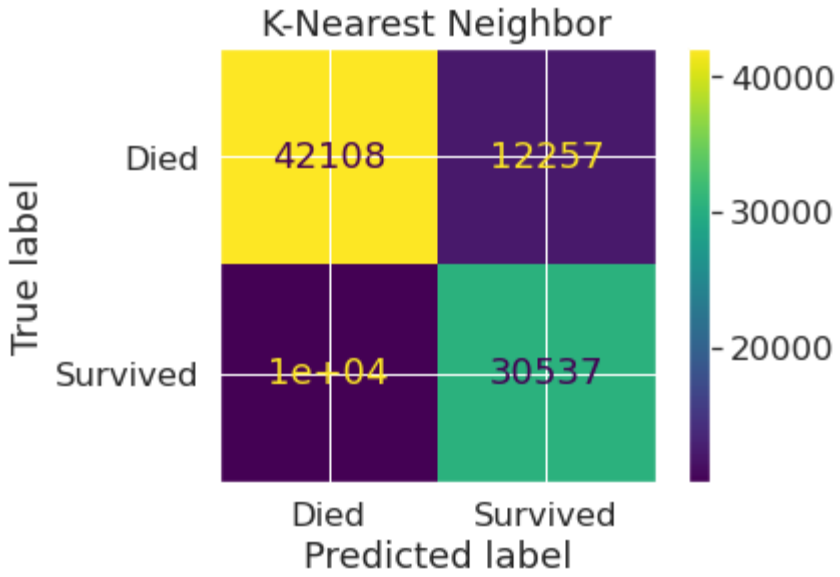


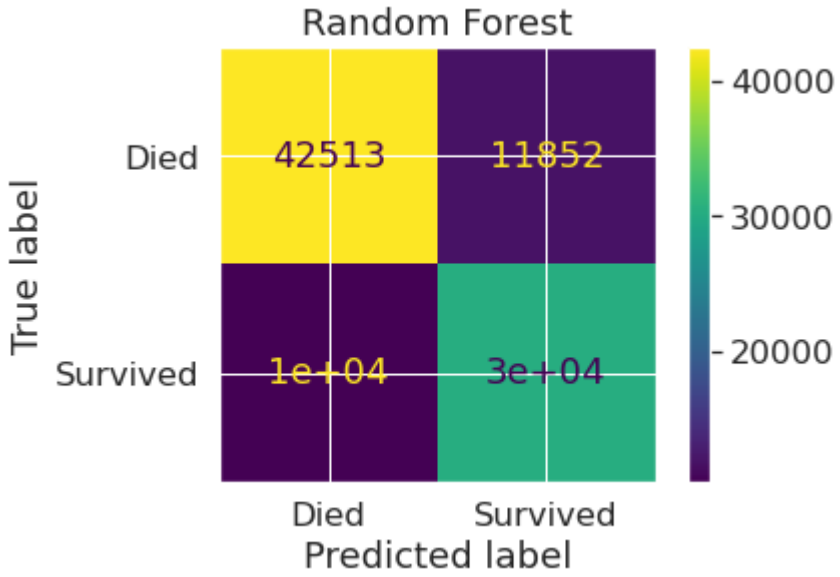
As can be seen from the graph, the performance of the models is quite like each other. In terms of accuracy and precision, Random Forest did the best (76.5%, 71.9%) whereas in terms of F1-score and recall Logistic Regression did better (73.3%, 75.9%). K-Nearest Neighbor, on the other hand, has no score better than all other.

We also plotted the ROC curve although the results of the models are very similar



We also plotted confusion matrix for each model on train set.

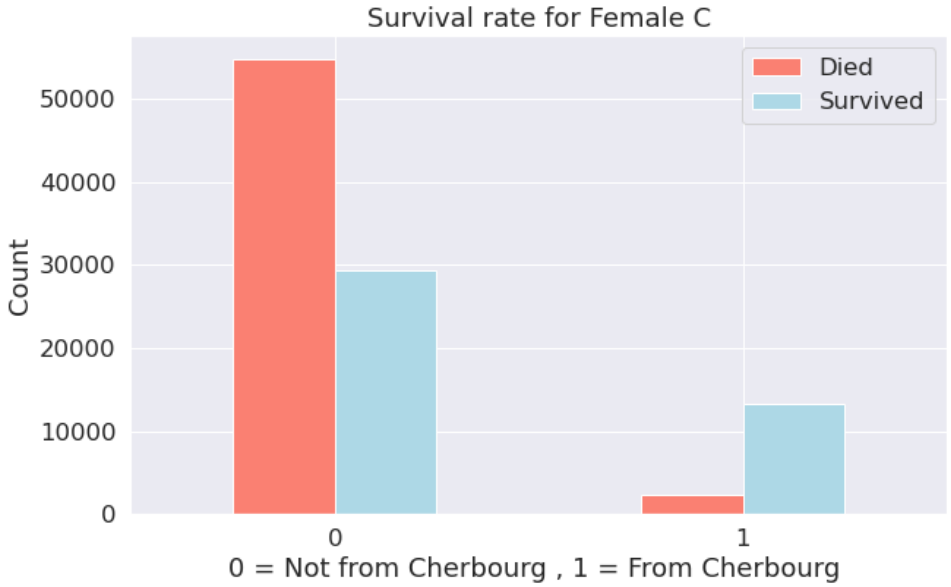




Insights

What we learned from this dataset is that during the crash, women, elder people, rich people were rescued first then the others. Probably lowest class alone males were the last priority. One interesting information we found is that most of the men that were embarked from Southampton died whereas most of the women that were embarked from Cherbourg survived. Following graphs can prove this clearly.


Survival rate for Male S
40000
35000
30000
25000
20000
15000
10000
5000
0
Count
Died
Survived
0
1
0 = Not from Southampton, 1 = From Southampton


Survival rate for Female C
50000
40000
30000
20000
10000
0
Count
Died
Survived
0
1
0 = Not from Cherbourg , 1 = From Cherbourg