“Who has the chance of survival?”

Submitted by: Basabdatta Ray

Business Objective

An Introduction to ‘Titanic’:

On April 15, 1912 a tragic shipwreck incident happened due to the ship’s collision with heavy pack ice and large icebergs. The ship was known as ‘Titanic’. This unfortunate incident took approximately 1635 lives including boarded passengers and crew members.

Aim of the study:

Based on the available data about ‘Titanic’ passengers, and using the hidden patterns found in that historical data, we need to predict whether a passenger survived or not.

About The Data:

Independent features/variables:

PassengerId: Passenger unique ID, Survived: Survival, Pclass: Passenger Class, Name: Name of passenger, Sex: Gender of passenger, Age: Age of passenger, SibSp: Number of Siblings/Spouses Aboard, Parch: Number of Parents/Children Aboard, Ticket: Ticket Number, Fare: Passenger Fare, Cabin: Cabin, Embarked: Port of Embarkation

Dependent/Target variable:

Survived:

0 = No, Negative class

1 = Yes, Positive class

Shape of the dataset:

There are total 891 rows and 12 columns in the train dats set
There are total 418 rows and 11 columns in the train dats set

Project Workflow:

Phase-1: Business understanding

Identify the business problem clearly. Once the problem statement is defined then convert it to analytical problem.

Phase-2: Data Gathering

Once the business problem is framed ready for data ingestion from the organization sources (Data repository, file, DB, real-time streaming data, sensor data from IoT).

Phase-3: EDA

This phase is all about how well we can understand the data and hence, is the most time consuming and critical phase of the entire data science life-cycle.

Phase-4: Data pre-processing

Cleaning the raw data so that it can be used to train the ML model. We need data pre-processing to achieve good results from the model.

Phase-5: Model building

Researching the model that will be best for the type of data. Main goal is to train the best possible model using the pre-processed data. Then split the data into 3 three sections - Training data ,Validation data and Testing data. Train the model using training data set, tune the parameters using validation set and finally then test the performance of the model on the unseen test data set.

Phase-6: Model evaluation and final conclusion

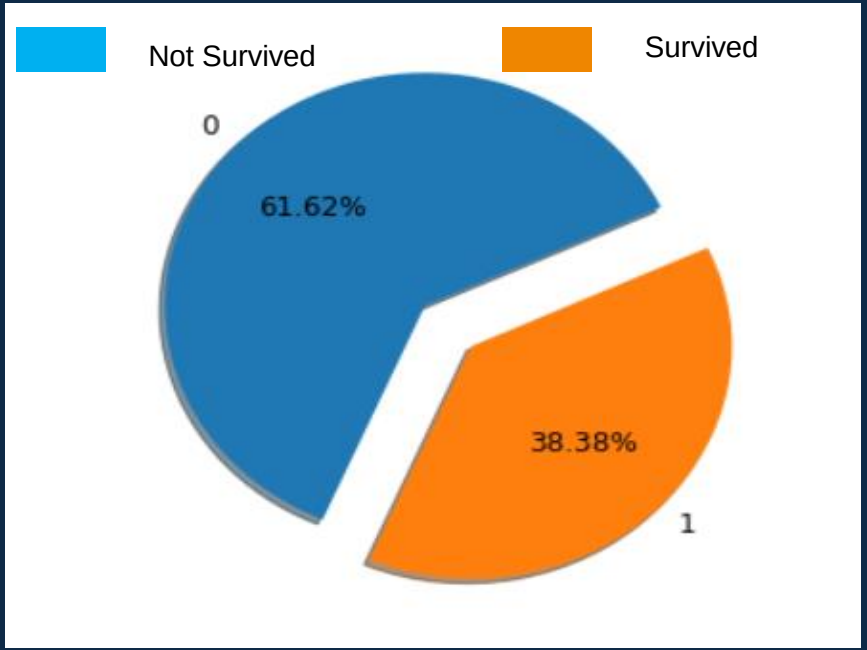
It is an integral part of the model development process. It helps to find the best suitable model that represents the data and how well the chosen model will work in the future.

EDA:

Univariate Analysis: ‘Survived’

Understand the proportion of the survived passengers:

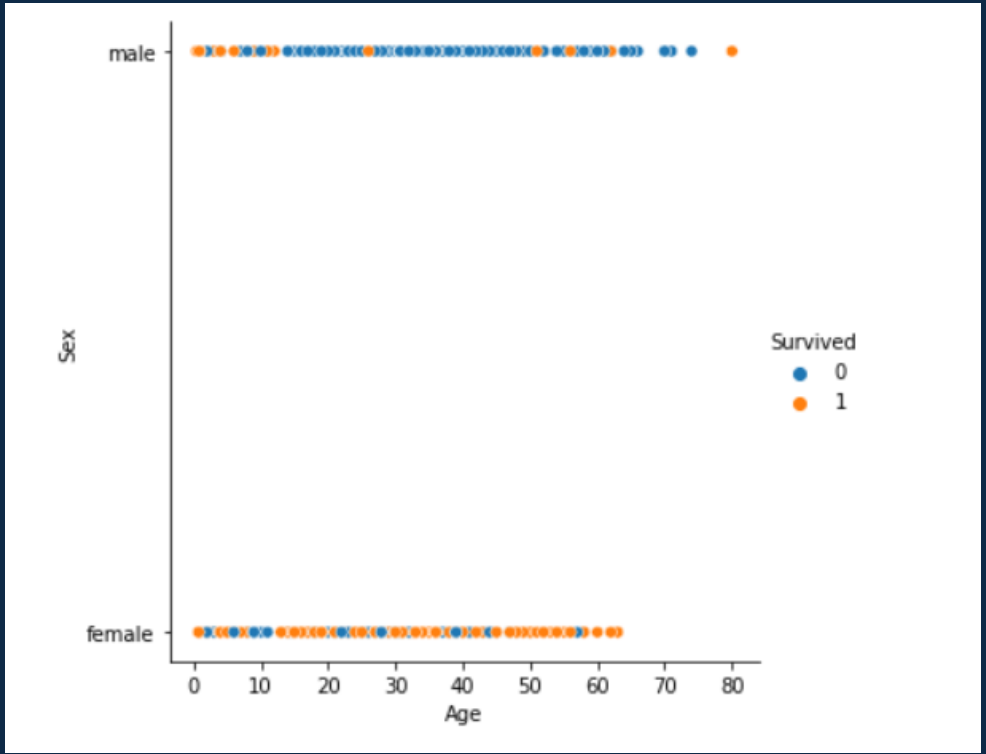
- Unfortunately, out of all the 891 passengers only 38% survived and 62% did not
- Also, we can say that the dataset is not disproportionate, hence in the further processing we need not apply any up-sampling or down-sampling technique to make the data proportionate



EDA:

Multivariate Analysis: Age and Gender of the passengers related to survival or not

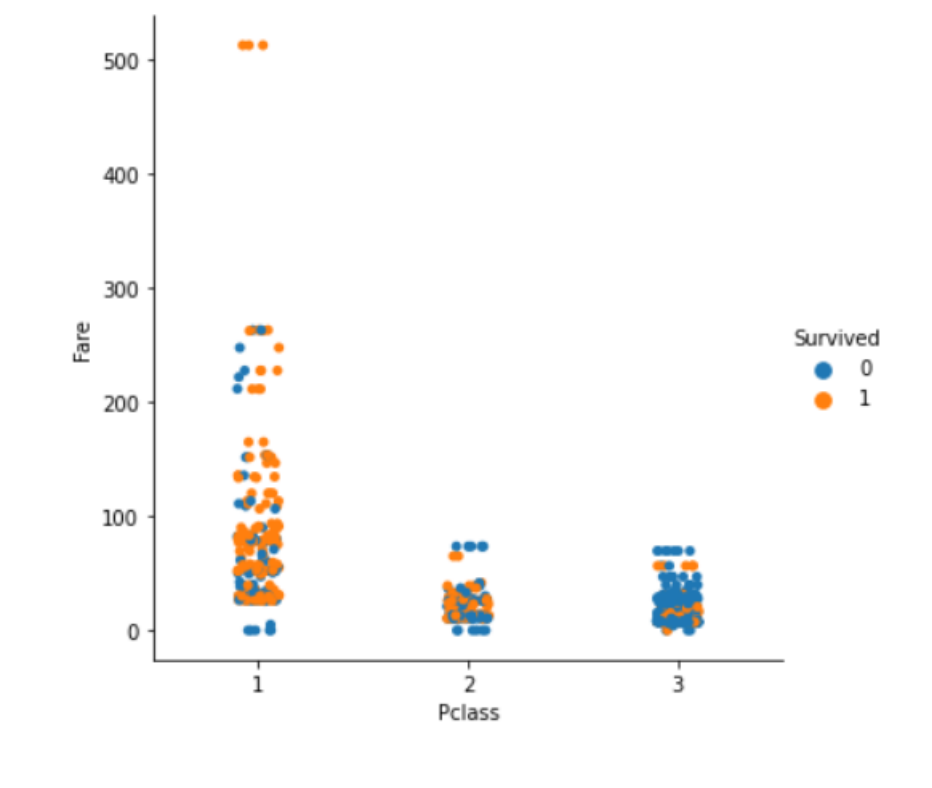
- From this graph, it is clear that female passengers have higher rate of survival than male passengers
- Also, aged female shows higher chance of survival than younger female, which is opposite trend male passengers



EDA:

Multivariate Analysis: Class of the passengers related to survival or not

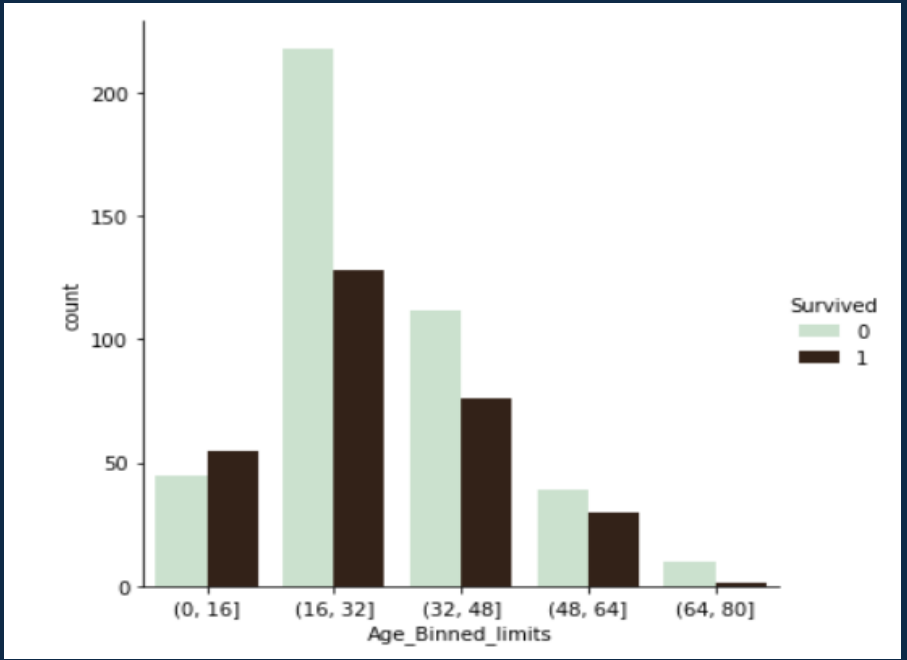
- It seems that 1st class or Class1 passengers paid Fare from very low to very high. Whereas, the passengers from other 2 classes have paid very low fare (all below 100).
- All those passengers who have paid the highest Fare (almost 512) did survive
- The second observation here is that the survival is remarkably high for the 1st class passengers. And the passengers from 3rd class mostly did not survive



EDA:

Bivariate analysis: How age alone impacted the chance of survival

- It seems that most of the on-board passengers were aged between 16 to 32 followed by age range 32 to 48. There were very few passengers of age above 64 and less than 16
- Only for the age group 0 to 16, the number of survived passengers is higher than the not-survived passengers



Few of the key takeaways from EDA part:

- Being a 'male' passenger, with passenger class as 3 reduced the chance for one to survive
- Passengers who paid the highest Fare, they all survived
- If the passenger travelled alone then there is less chance that he/she survived
- Older ‘female' passengers with passengers had more chance of survival than young 'female' passengers

Model Buidling:

Model building:

- For this specific type of data and given problem statement, built the model with Logistic Regression, Decision Tree, Random Forest and then finally based on different evaluation metrics compared these models.

	Model	roc_auc_validation	roc_auc_train
0	Logistic Regression	0.841173	0.773015
1	Decision Tree	0.776812	0.998861
2	Random Forest	0.808630	0.998861

Model Performance Evaluation:

Model performance comparison:

- Among all the models, Logistic regression model is giving the highest Recall accuracy on the validation set. From the accuracy on validation set of Decision tree, it is clear that it has overfitted the data. The random Forest model too has the same issue here.

Model Selection: Logistic Regression is selected based on Recall and Precision both

- Positive class (class 1) recall 86%: Out of all the actual positives 86% of the data is correctly predicted as positive
- Positive class (class 1) precision 76%: Out of all the predicted positives 76% of the data is actually positive
- Similarly, for Negative class (class 0), both recall and precision looks good and hence, F1-score as well
- So, with simple imputation technique, encoding techniques the base model, Logistic regression is performing well. Hence, we can consider it as a better model for this given dataset and the given problem statement

	precision	recall	f1-score	support
0	0.90	0.83	0.86	110
1	0.76	0.86	0.80	69
accuracy			0.84	179
macro avg	0.83	0.84	0.83	179
weighted avg	0.85	0.84	0.84	179

Thank you