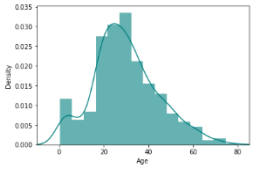
Introduction:

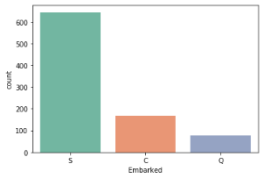
The purpose of the analysis is to do an initial exploration of the data set gathered for the passengers on-board on Titanic, apply various machine-learning techniques to develop predictive models and validate the most successful model to predict the survival of a passenger. The models developed for prediction will be used to predict the survival of all the passengers from the test data.

Exploratory Data Analysis:

There were two data sets provided. The first data set is train, that is used to create the model and the second data set is to test the model and predict the survival of passengers. The train data has 445 rows of data and 12 variables like the passenger, the ticket class, the ticket fare, gender, age, if they were travelling alone or with family, ticket no, the cabin assigned, where they boarded the ship and if they survived.



The data is a combination of categorical variables like sex, cabin, embark and more, continuous variables like age and binary variables like survived. The response variable survived is a binary variable with 0 or 1 value. The categorical variables that help us predict the survived are Pclass, Age, Sex, Embark, Sibsp, Parch. About 22% of Cabin data is missing and will not be helpful to predict the response variable. The Embarked variable has three possible values that S, Q and C. Most frequent value is S. One of the four variables chosen for the model, Age has some missing values.

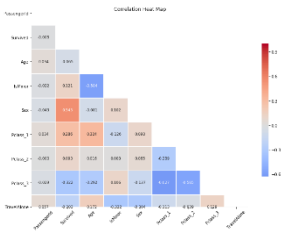


Data Preparation and Overview of Programming:

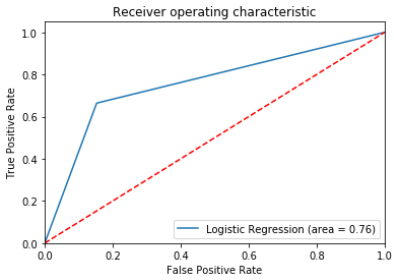
The first step to get the data ready to create the model is to fill the missing values. For the variable Age, I chose median age of the existing data to fill the missing values. For the

Embarked variable, I chose the most frequent value to fill the missing values. From the Age field, I created binary variable IsMinor that classifies the passengers as Minor/Not Minor based on Age. Created a new binary variable called TravelAlone combining SibSp and Parch. Created three different binary variables by breaking Pclass variable into Pclass_1, Pclass_2, Pclass_3. Converted the male/female values in Sex column into 0 and 1. The same data and transformations were applied to the test data set. The train_test_split function was used to split the train data.Using the columns Age, IsMinor, Sex, TravelAlone, Pclass_1, Pclass_2, Pclass_3 the Logistic Regression and Naïve_Bayes Regression models was applied on the train split data. The ROC_AUC_Score method was used to calculate the area under the curve.

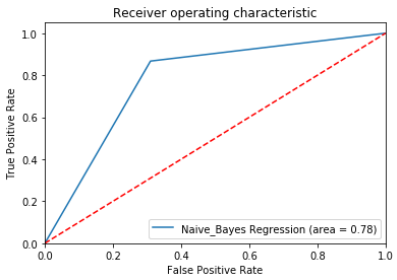
	PassengerId	Survived	Age	IsMinor	Sex	Pclass_1	Pclass_2	Pclass_3	TravelAlone
0	1	0	22.0	0	0	0	0	1	0
1	2	1	38.0	0	1	1	0	0	0
2	3	1	26.0	0	1	0	0	1	1
3	4	1	35.0	0	1	1	0	0	0
4	5	0	35.0	0	0	0	0	1	1
...	...	...	...	...	...	...	...	...	...
886	887	0	27.0	0	0	0	1	0	1
887	888	1	19.0	0	1	1	0	0	1
888	889	0	28.0	0	1	0	0	1	0
889	890	1	26.0	0	0	1	0	0	1
890	891	0	32.0	0	0	0	0	1	1



The accuracy score, log_loss score and auc score were calculated using roc_curve function.




```
accuracy is 0.782
log_loss is 0.454
auc is 0.756
```




```
accuracy is 0.754
log_loss is 0.473
auc is 0.779
```

Insights & Conclusion:

My recommendation to the historian would be to use Logistic regression model using the above binary variables as the accuracy for Logistic Regression is higher than Naives_Bayes Regression, although the roc_auc_score is higher for Naïve_Bayes regression model.

Appendix:

The ipynb notebook and an html version of the notebook along with the output and Kaggle submission scores are included in the submission.