1. Problem Statement

RMS Titanic was a British luxury passenger liner that sank on 15th April 1912, during its maiden voyage. The sinking of Titanic was a result of a collision with an iceberg, resulting in the death of 1502 out of 2224 passengers. In addition, one of the contributing factors to the large number of death was due to insufficient number of lifeboats on Titanic.

In this term paper, a machine learning model will be constructed to predict if a passenger will survive based on a set of features given in the dataset from Kaggle Competition – Titanic: Machine Learning from Disaster

The features included in the dataset are given in the table below:

Table 1 showing all the features in Titanic training dataset

Features	Definition	Remarks
Passenger ID	Passenger identification	
Survival	Passenger's survival	0 = No 1 = Yes
Name	Passenger's name	
Sex	Passenger's gender	Female / Male
Age	Passenger's age	Given in years
PClass	Passenger's ticket class	1 = first class 2 = second class 3 = third class
SibSp	Number of siblings/spouses aboard Titanic	
Parch	Number of parents/children aboard Titanic	
Fare	Passenger's Fare	

Port of Embarkation	Port which passenger embarked on	C = Cherbourg
		Q = Queenstown
		S = Southampton

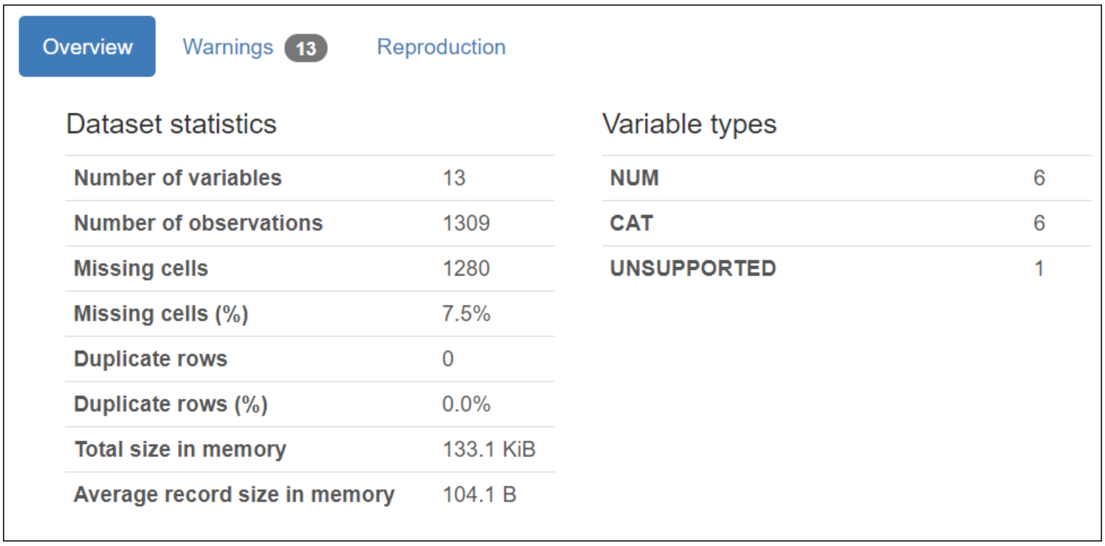
Cabin	Cabin number	
Ticket	Ticket number	

Exploratory data analysis on the features will be done in the subsequent section to analyse on potential challenges faced in the dataset.

2. Challenges of the Dataset

Prior to training a machine learning algorithm to predict if a passenger will survive, the given data has to be analysed and processed to understand the nature of data given. Exploratory data analysis was carried out using basic statistical tools and visualization aids from python to gather insights of training dataset.

In Figure 1 below, it shows the total number of features, types of data and the percentage of missing values in the dataset.



Dataset statistics

Number of variables	13
Number of observations	1309
Missing cells	1280
Missing cells (%)	7.5%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	133.1 KiB
Average record size in memory	104.1 B

Variable types

NUM	6
CAT	6
UNSUPPORTED	1

Figure 1 shows the general information extracted from dataset

In total, there was 7.5% of data in the dataset that has no value, contributing primarily by 'Age', 'Cabin', 'Embarked' and 'Fare' features (refer to Figure 2). The method in dealing with missing values will be discussed in section 3 of the report.

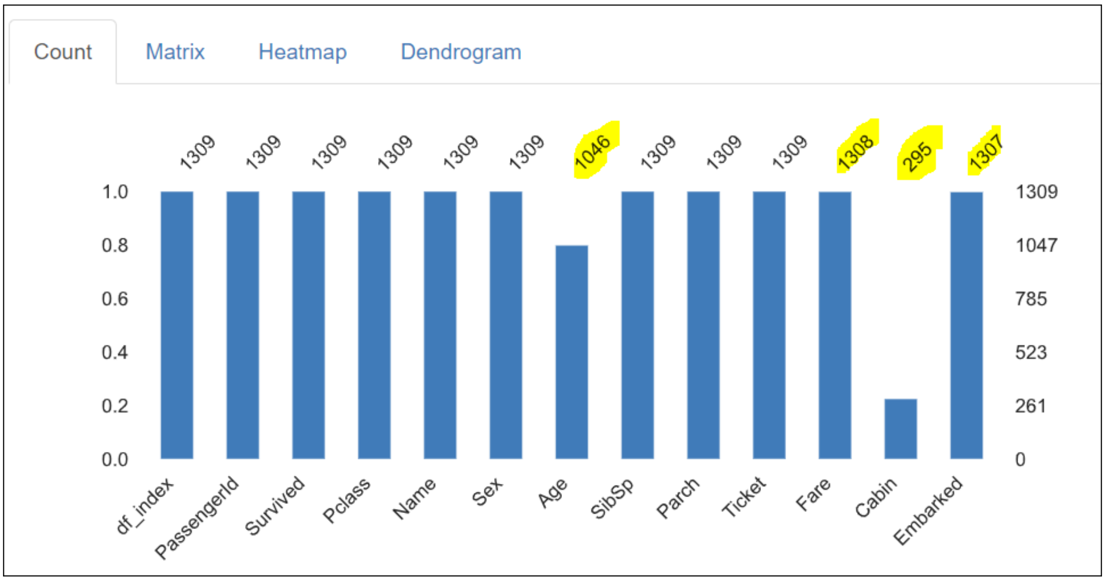


Figure 2 shows the statistics of missing values from all the features in dataset

2.1 Numerical Features

Pearson Correlation Coefficient (PCC) was used to evaluate the interaction between various numerical features and the results are shown in Figure 3. Based on PCC, 'SibSp' and 'Parch' have some degree of correlation, possibility of constructing new a new feature from them. While 'Fare' shows to have a positive correlation to survival probability, it does not mean other features are not useful. It will be looked into in Section 3 of the report.

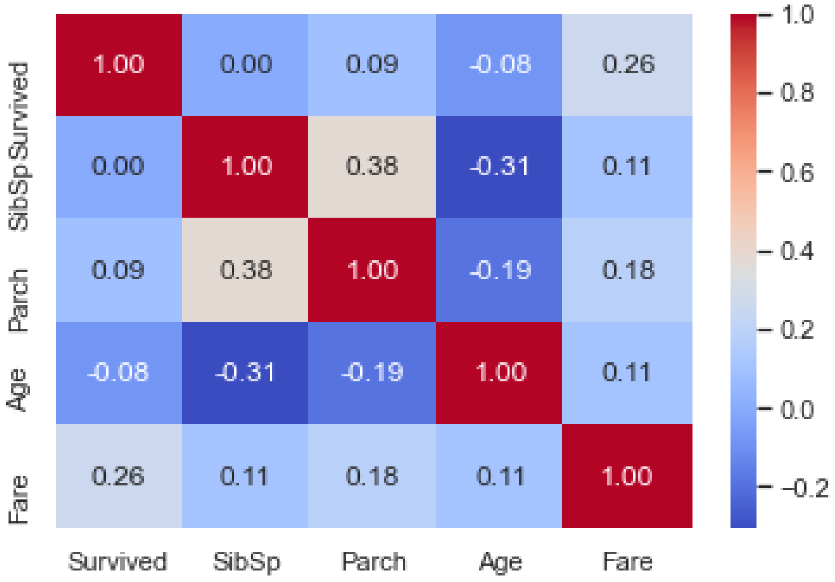


Figure 3 Pearson Correlation Coefficient heat map for various features in dataset

In Figure 4, it highlighted that Fare data is much skewed, and requires processing to reduce the skewness.

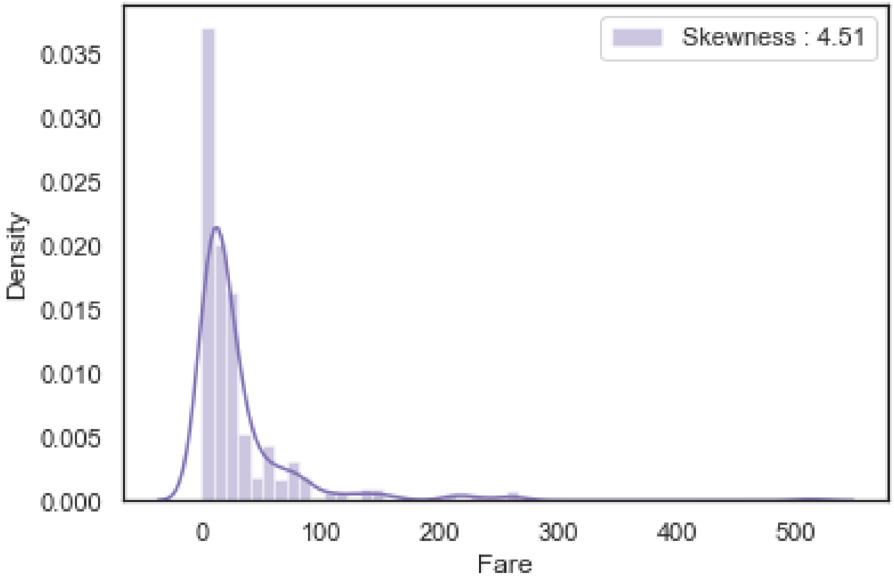


Figure 4 shows distribution of Fare feature

Passenger's age seem to follow Gaussian distribution (refer to Figure 5), and noticed that a larger proportion of passenger with age lesser than 5 surviving. Correspondingly, a larger proportion of older passenger above 60 did not survive and passenger between age 20 and 40 did not survive. As such, 'Age' is still an important feature as it can help to predict survival chance despite poor PCC with survival

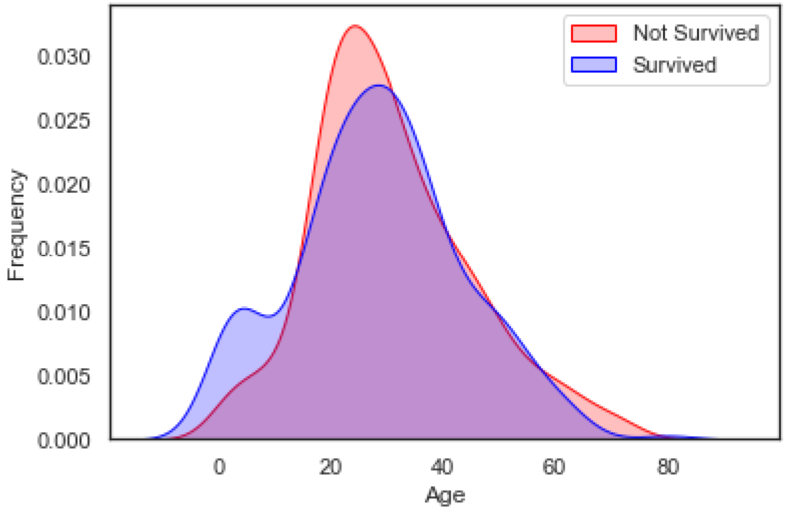


Figure 5 shows a distribution plot of passenger's age with respect to survival rate

2.2 Categorical Features

'Pclass', 'Sex' and 'Embarked' features suggested a better ticket class, being female and embarkation at Cherbourg will result in a higher survivability rate. Thus, all 3 features should be considered in the model evaluation.

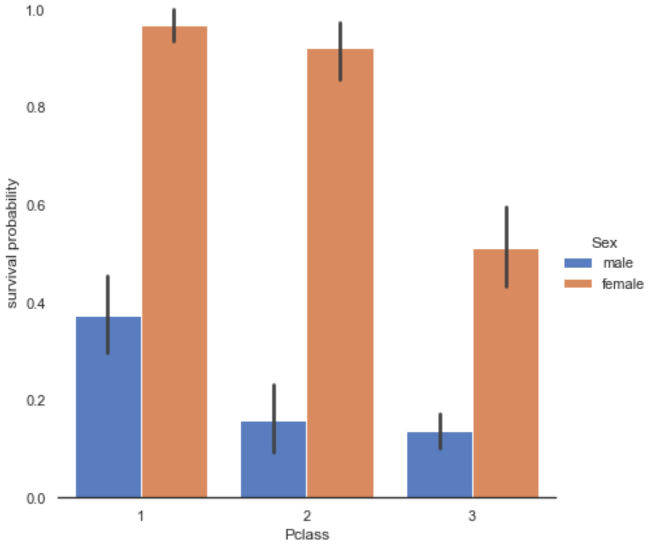


Figure 6 shows a plot of survival probability against passenger ticket class and gender

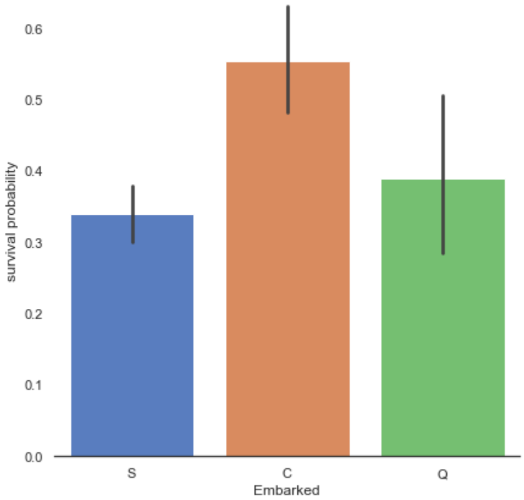


Figure 7 shows a plot of survival probability against port of embarkation

‘PassengerID’ is highly distinct/unique value as it is identification number of passenger, it serves no purpose in increasing the model accuracy, and will be removed from evaluation.

As previously highlighted, ‘Cabin’ has a large number of missing value (77%) and removing it will be an easy decision. However, this feature is likely to indicate the probable location of passenger in Titanic, therefore, processing and retaining it may help in model performance.

‘Name’ feature is highly unique as it represent individual passenger’s name and will not affect the model accuracy in prediction. However, it contains title of passenger, which may be useful to gauge the socioeconomic status and age of passenger, hence, may be useful for model prediction

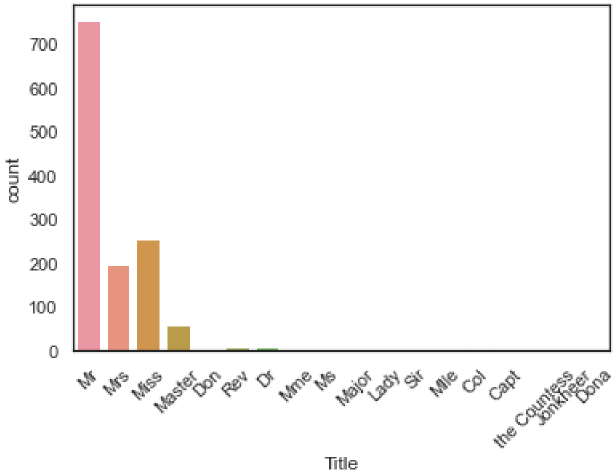


Figure 8 shows a plot of title count vs various passenger title

In essence, the dataset requires preprocessing before using it to train a machine learning model and some of the challenges are listed below:

1. Missing values in dataset namely 4 features, 'Age', 'Cabin', 'Embarked 'and 'Fare'
2. New features construction to improve model performance (family size, passenger's title, cabin type)
3. Removing redundant features (passenger ID and name)
4. Feature transformation for string data into categorical data.

3. Proposed Solution

It was highlighted in Section 2 of the report, data preprocessing has to be carried out before using the data for training. In this section, the motivation behind features engineering and machine learning method will be discussed in greater details.

3.1 Missing Values

'Age', 'Cabin', 'Embarked' and 'Fare' features in dataset has some form of missing values, with 'Age' and 'Cabin' having 20% and 77% of missing values respectively.

'Embarked' and 'Fare' features are defined as passenger's port of embarkation and the ticket price paid by passenger. As highlighted in Section 2, Figure 2 of the report, the missing values are few, 2 missing values of 'Embarked' and 1 missing value of 'Fare' out 1309 data instances. Therefore, a simple statistical method of using the mode and median will be used to complete the missing data for 'Embarked and 'Fare' feature respectively.

Previously, 'Fare' feature was noted to be skewed, thus a logarithmic function will be applied to it. Skewness was reduced from 4.51 to 0.57 after logarithmic transformation.

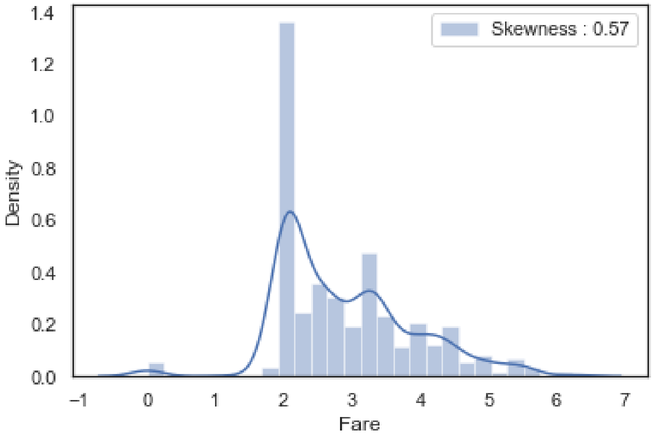


Figure 9 shows fare distribution post logarithmic transformation

‘Age’ feature contains approximately 20% missing values and noting that it has a certain degree of negative correlation with ‘Pclass’, ‘Parch’ and ‘SibSp’ (refer to Figure 10).

Based on Figure 11, 12 and 13, it seems to suggest, passenger’s age decreases as ticket class increases, number of parents/children increases, and number of siblings/spouse decreases

Therefore, the strategy is to fill Age with the median age of similar rows according to Pclass, Parch and SibSp.

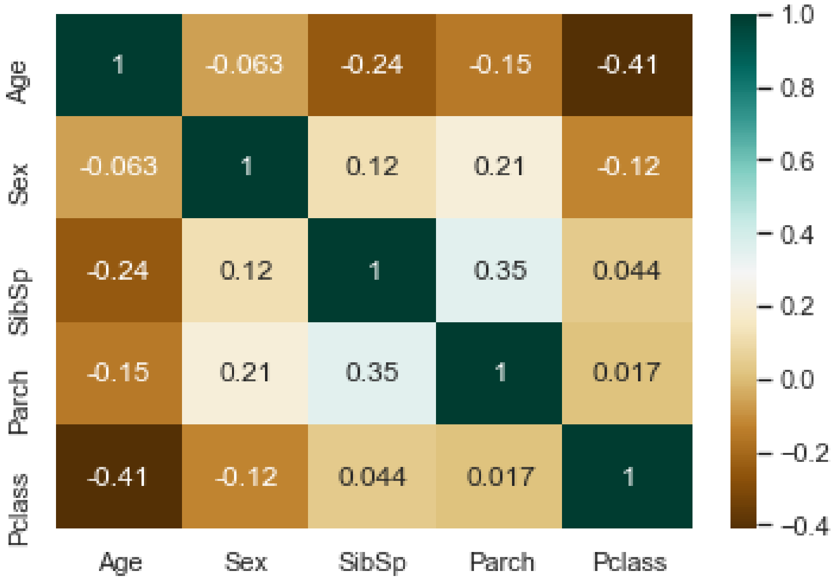


Figure 10 Pearson Correlation Coefficient heat map to gauge correlation of various features with respect to Age

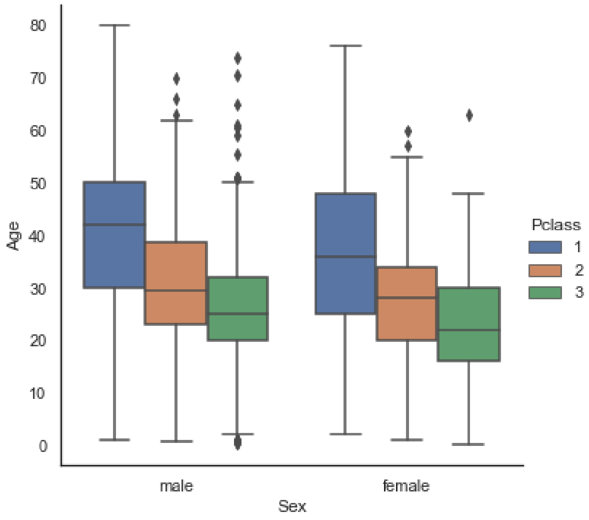


Figure 11 shows box plot of age, sex and passenger ticket class

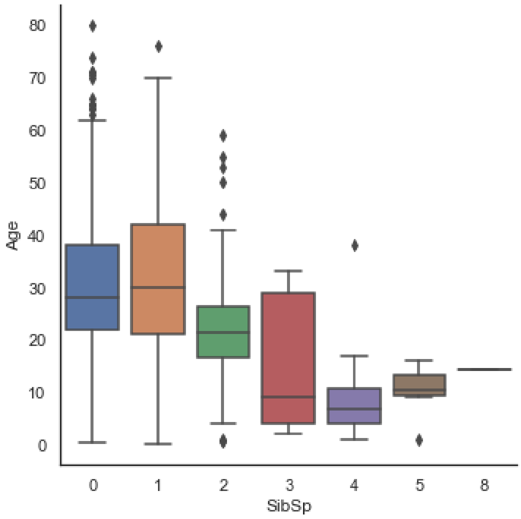


Figure 12 shows box plot of age and number of siblings/spouse

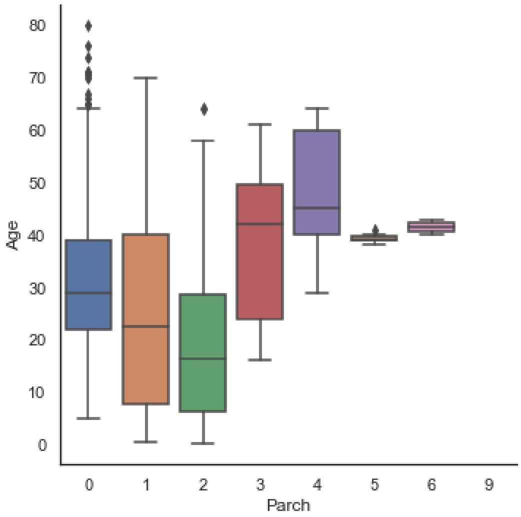


Figure 13 shows box plot of age and number of parents/children

3.2 Feature Construction

3.2.1 Family Size

‘SibSp’ and ‘Parch’ are defined as number of siblings/spouses aboard Titanic for each passenger and number of parents/children aboard Titanic for each passenger respectively. The two features suggest passengers with lesser siblings/spouse and passengers with lesser parents/children are equipped with a higher chance of surviving. As such, a new feature can be constructed to determine the size of family and hence, improving the model performance. The intuition behind the creation of this new feature is because larger families are more difficult to evacuate and the intuition is supported by Figure 14

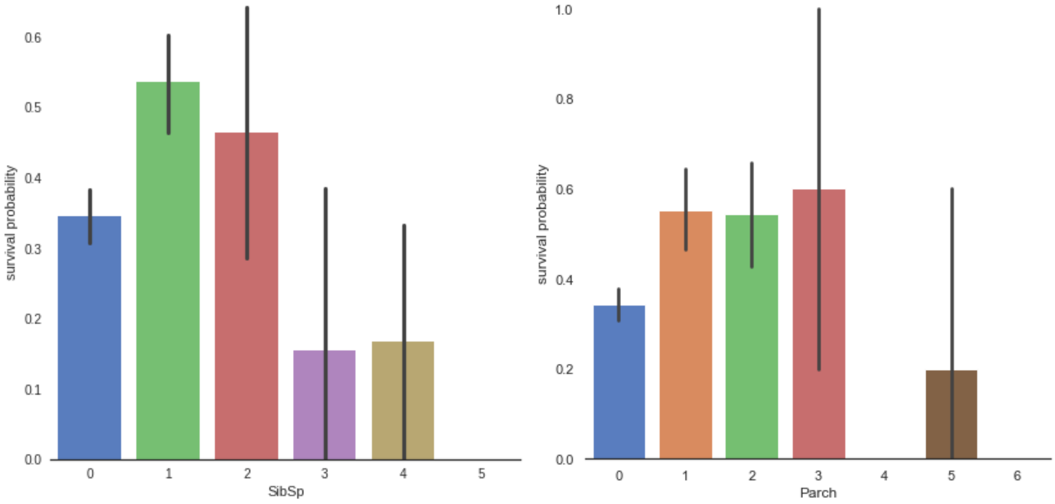


Figure 14 shows a plot of survival probability against # of siblings/spouse and # of parents/children features

The new feature, family size, could be constructed by summing the feature ‘SibSp’ and ‘Parch’ to determine if each passenger’s family size. In addition, the family size is categorised into 4 categories, single (family size = 1), small family (family size = 2), medium family (2 < family size <= 4) and large family (family size < 4)

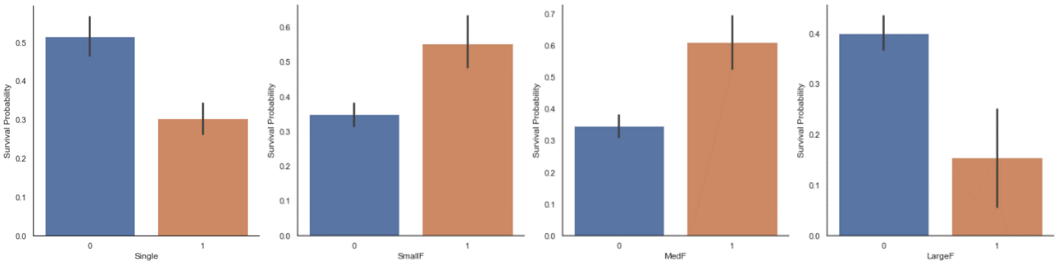


Figure 15 shows factorplot for various family sizes and their survival rate

3.2.2 Title

As previously mentioned in Section 2, ‘Name’ feature will be dropped. However, the title of passenger will be retained and will be used for model training. The motivation behind creating new title feature is because, it will provide insight to passenger’s socioeconomic status, hence possibly influence his/her survival rate.

As shown in Figure 8, there are various titles available and it will be further reclassify into 4 main categories, given below:

1. Master
2. Miss/Mrs
3. Mr
4. Rare – Titles that are not commonly used such as “Sir”, “Lady”, “Capt” and etc.

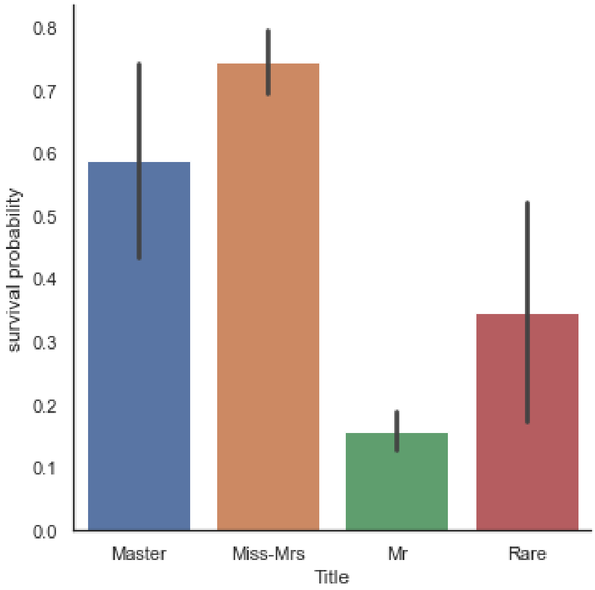


Figure 16 shows bar chart of various title against survival probability

In Figure 16, it is clear that younger passengers and female passengers stand a higher chance of surviving.

3.2.2 Cabin Type

‘Cabin’ feature is defined as cabin number of passenger abroad Titanic, while it may has a large number of missing data (77%), it provides insight on the probable location of passenger in Titanic, thus possibly improving model accuracy.

As such, we will be replacing the cabin number with the types of cabin. The first letter in the cabin feature indicates the type of cabin and its location, for all missing values in cabin feature, it will be replace with ‘X’

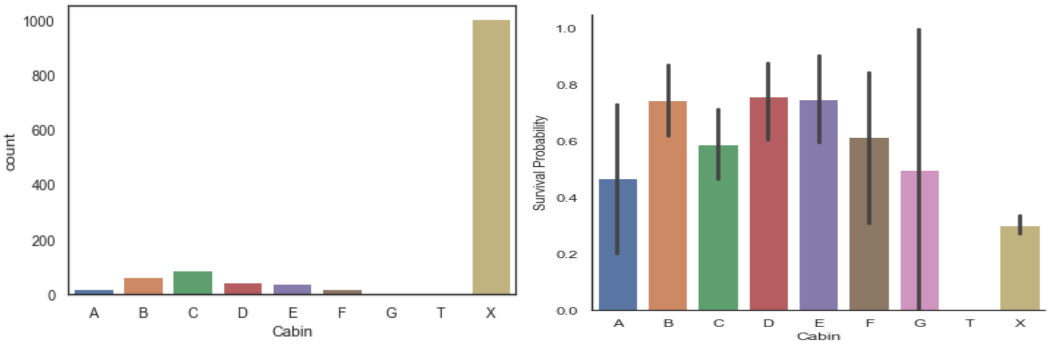


Figure 17 shows bar chart and factorplot of cabin type against frequency and survival probability

3.3 Feature Transformation

Some of the features ('Sex', 'Embarked', 'Title', 'Family Size', 'Cabin Type') in dataset are given as string text, it will be converted into categorical data type with discrete value

3.4 Machine Learning Algorithm

The nature of the task is a classification problem with binary output (0 = dead, 1 = survived). Therefore, classification algorithms will be employed. In this paper, support vector machine (SVM), decision tree, K-nearest neighbour (KNN) would be used for first pass evaluation.

Cross validation using K-fold cross validation method will be used to compare between the 3 classification algorithms, followed by grid search to optimize all the hyperparameters.

After determining the best classification algorithm with tuned hyperparameters, ensemble learning technique will be used to construct meta classifier to improve the accuracy of the model. The motivation behind constructing ensemble model is each classifier in ensemble model represents a single hypothesis or expert being trained differently and hence will be able to provide diverse response and improve model performance.

4. Experiment

As highlighted in Section 3.4, comparison between 3 classification algorithms will be done over Kfold cross validation method. In Figure 18, it shows the 3 algorithms accuracy against training data set.

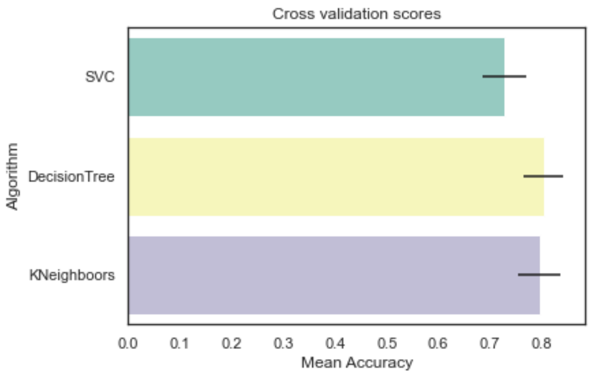


Figure 18 A plot of algorithm against its model accuracy

From cross validation, decision tree algorithm looks to be the best classification method to be used in this task, As such, further testing of decision tree algorithm with different ensemble learning method such as adaboost with decision tree, random forest, and extra trees will be employed. SVM (kernel) will also be used for further tuning, as it is likely that hyperparameters used for SVM are not optimised resulting in poor cross validation scores

Table 2 shows various algorithms with its cross validation scores and error

Algorithm	Cross Validation Score	Cross Validation Error
SVM	0.728	0.041
Decision Tree	0.803	0.038
KNN	0.796	0.041
AdaBoost	0.800	0.053
Random Forest	0.800	0.051
Extra Trees	0.803	0.045

The cross validation score between the new 3 ensemble learning methods are fairly similar to decision tree algorithm. Hyperparameters tuning will be done for the 3 ensemble learning method to further improve model performance.

Post hyperparameters tuning, all 3 ensemble learning methods outperformed decision tree algorithm. It is interesting to note that SVM model validation score improved drastically and it would be used in the model evaluation.

Table 3 shows various algorithms with its scores post hyperparameter tuning

Algorithm	Cross Validation Score
Decision Tree	0.814
SVM (kernel)	0.833
AdaBoost	0.824
Random Forest	0.834
Extra Trees	0.830

According to the growing cross-validation curves (Figure 19a to 19e), Adaboost seems to be overfitted, it could be due to the increased in weight for every wrongly classified data instances, hence resulting in higher chances of selecting such instances for training.

SVC, random forest and extra trees classifiers seem to better generalize the prediction since the training and cross-validation curves are close together.

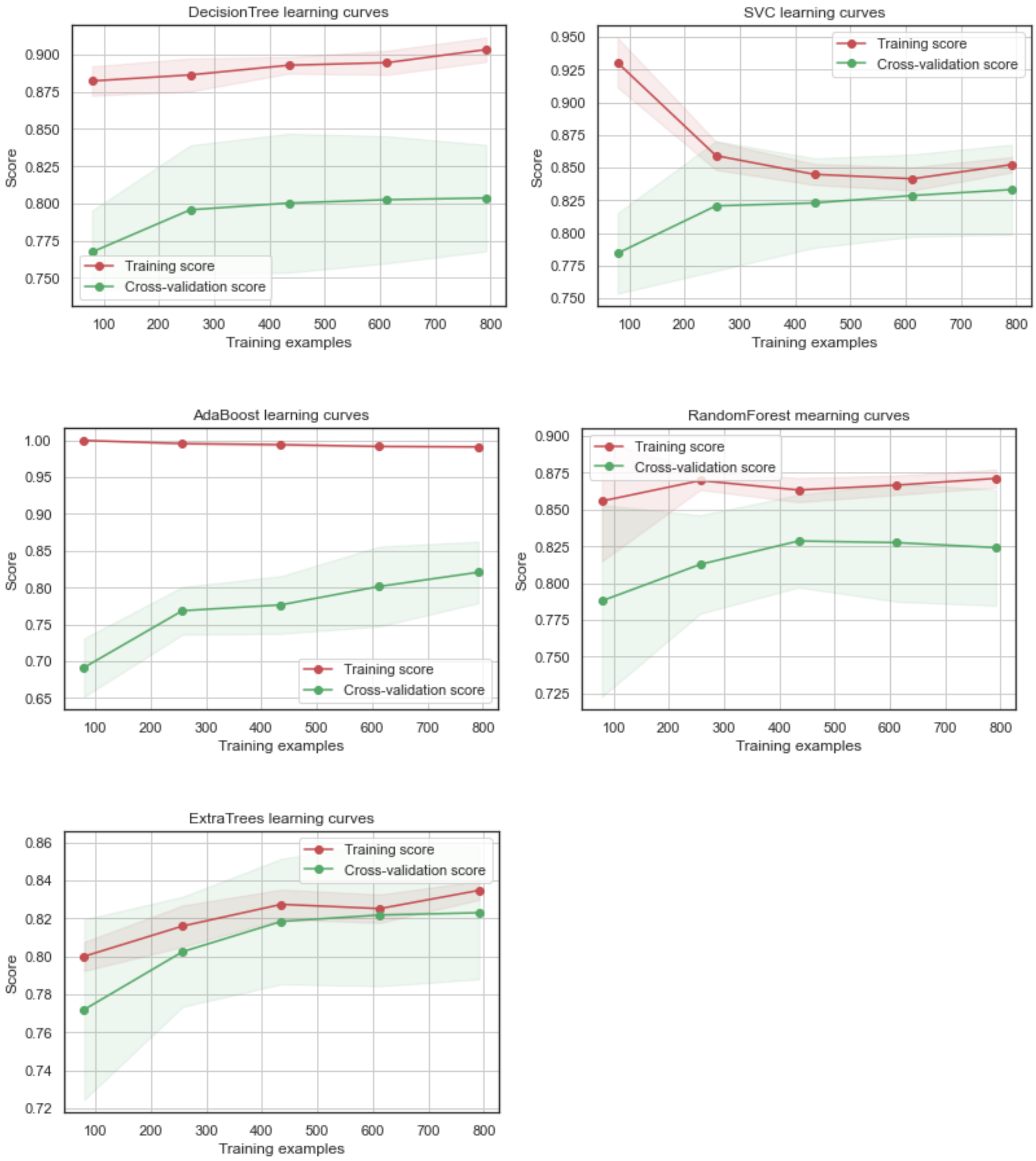


Figure 19 shows growing cross validation curves for various algorithms

To validate the hypothesis of various handcrafted features, a bar plot was created to show relative importance of each features in the various tree classifiers. From Figure 20, it is clear that family size and passenger title played an important role in the survival rate of each passengers.

In general, ‘Fare’, ‘Title_2’, ‘Age’, ‘Fsize’ and ‘Sex’ played an important role to determine the survival rate of each passengers. Note that Title_2 indicates Mrs/Mlle/Mme/Miss/Ms.

It can be inferred that, socio-economic class of passenger, gender of passenger, family size as well as age played an important role in determining the survivability of each passenger.

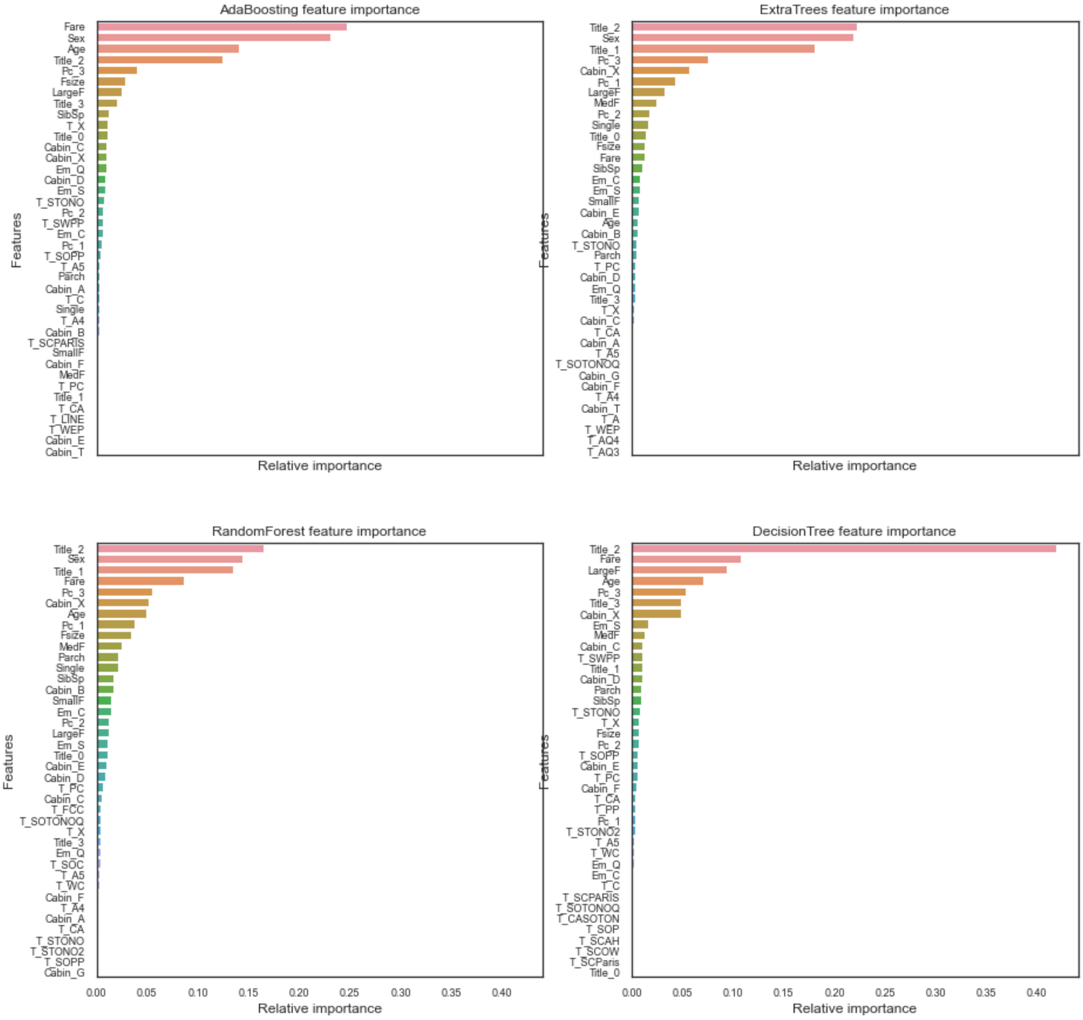
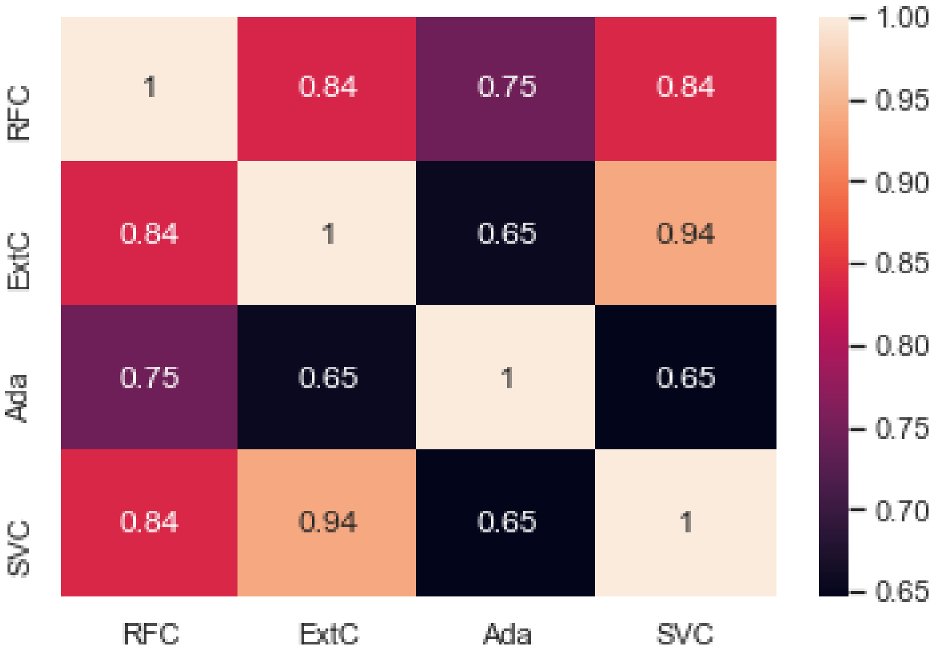


Figure 20 shows relative important of each features for various algorithm

In addition, it is important to check similarity in terms of each classifier predicted result. With reference to Figure 21, there are some differences between the predicted results from each classifier, this meant that, the classifiers are slightly different and able to provide diverse response. As such, meta classifier can be constructed and a simple majority voting system will be used for the model.

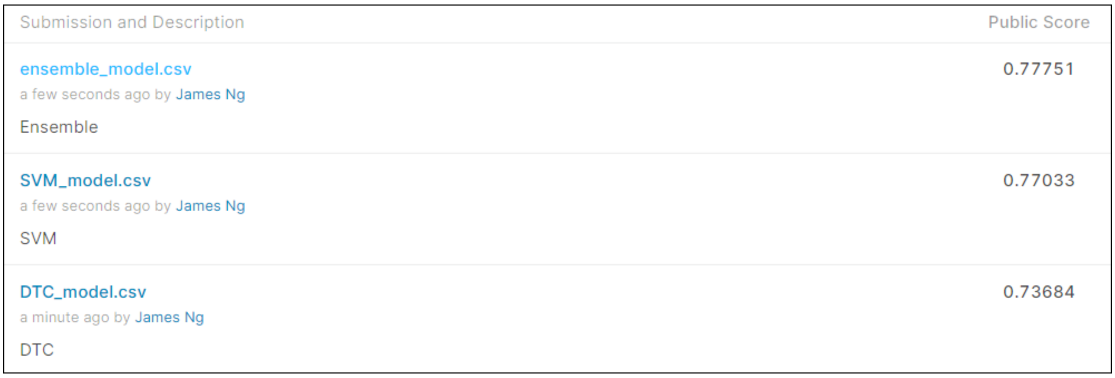
The final result from the ensemble model, based on 4 classifiers (AdaBoost, Extra Trees, Random Forest and SVM) have a mean accuracy of 92.04% on training dataset vs a simple classifier like SVM which has a score of 83.3% and decision tree classifier which has a score of 81.4%.



	RFC	ExtC	Ada	SVC
RFC	1	0.84	0.75	0.84
ExtC	0.84	1	0.65	0.94
Ada	0.75	0.65	1	0.65
SVC	0.84	0.94	0.65	1

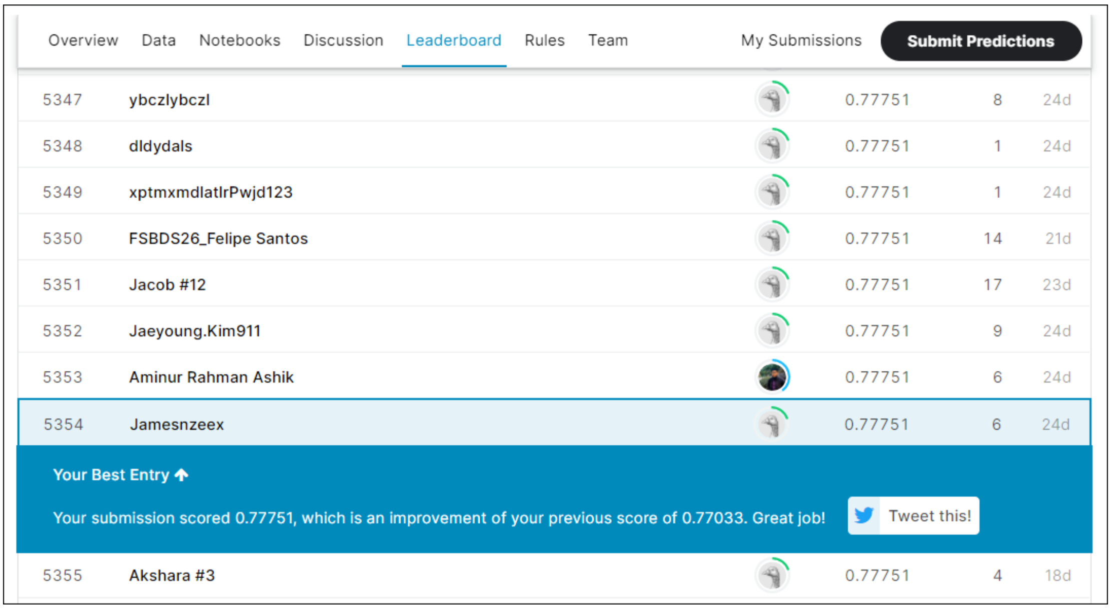
Figure 21 shows a heat map of the similarities of various classification predicted output

The final result based on submission to Kaggle is highlighted in the Figure 22 below. As expected, ensemble model is the best as compared to individual classifier such as SVM model and decision tree mode.



Submission and Description	Public Score
ensemble_model.csv a few seconds ago by James Ng Ensemble	0.77751
SVM_model.csv a few seconds ago by James Ng SVM	0.77033
DTC_model.csv a minute ago by James Ng DTC	0.73684

Figure 22 shows the score of various model in the Kaggle competition



5347	ybczlybczl	0.77751	8	24d
5348	dldydals	0.77751	1	24d
5349	xptmxmdlatlrPwjd123	0.77751	1	24d
5350	FSBDS26_Felipe Santos	0.77751	14	21d
5351	Jacob #12	0.77751	17	23d
5352	Jaeyoung.Kim911	0.77751	9	24d
5353	Aminur Rahman Ashik	0.77751	6	24d
5354	Jamesnzeex	0.77751	6	24d

Your Best Entry ↑

Your submission scored 0.77751, which is an improvement of your previous score of 0.77033. Great job!

5355	Akshara #3	0.77751	4	18d

Figure 23 shows the ranking in Kaggle Leaderboard

5. Conclusion

In this competition, exploratory data analysis was done with various statistical tools and visualisation aids which helps me further understand the importance of handcrafting good features and data cleaning before using the data for training. Poorly processed data will results in erroneous model and a good engineered feature will improve the model performance much better than a perfectly tuned model.

In addition, traditional machine learning techniques are being explored in this competition and an ensemble learning model was eventually constructed for the competition, achieving a score mean score of 92.04% in training dataset and 77.78% in the competition.