Machine Learning I (DATS 6202 - 11, Spring 2018)
George Washington University
Group - 16: Zhaoyang Chen, Zhen Zhang
Final Project Report
April 15, 2018

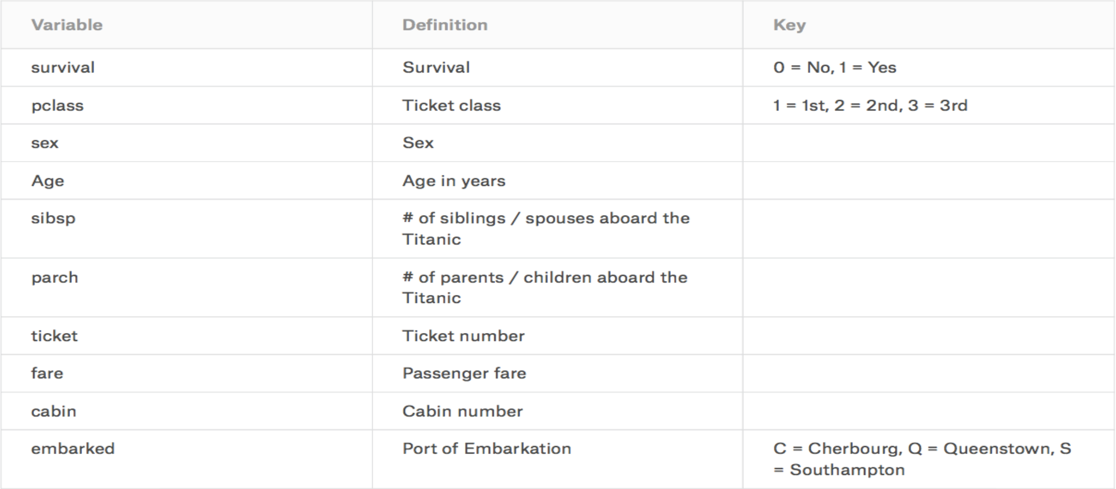
Predict Survival On the Titanic

1. Introduction

The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, causing 1502 death out of 2224 passengers and crew. In our project, we will apply machine learning techniques to build the optimal model, find the most valuable predicators, and predict the probability of survival on the Titanic.

2. Description of The Dataset

The dataset we used in from Kaggle. It contains the data of 891 passengers, and for each passenger, there are 12 attributes. Besides the attributes "Passenger ID" and "Name", the other 10 variables are used to build our model. The variable "survival" which is also the dependent variable, is given in binary form, "1" means survived.



Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

3. Data Preprocessing

The data preprocessing step is the most important step before building the model. Our data preprocessing is consisted of the following 5 steps:
1. **Drop the data**: we dropped the attributes "PassengerId", "Name" and "Ticket" from our dataset since they are not related to our prediction.
2. **Missing data imputation**: there are 2 missing values in the attribute "Embarked", we used the mode to impute the missing values since "Embarked" is a factor variable. In the "Age" variable, there are a total of 177 missing values. Since the distribution of "Age" in normal, we

use the median to impute the missing values. To make the imputation more accurate, we group the data by sex, and then calculate the median separately.

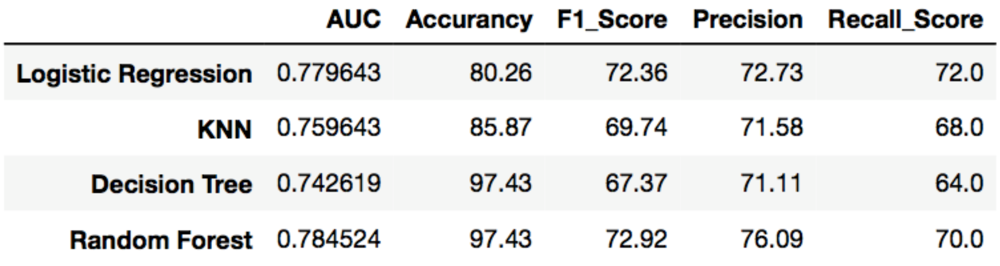
3. **Handling categorical data:** perform one-hot encoding on categorical variables “Sex” and “Embarked” to convert them into a form that is better to build our model.

4. **Partitioning training and testing data:** we chose 70% of our data as training set and 30% of the data as testing set.

5. **Bring feature onto the same scale:** bring the data onto the same scale for fast convergence of our model.

4. Models and Model Comparison

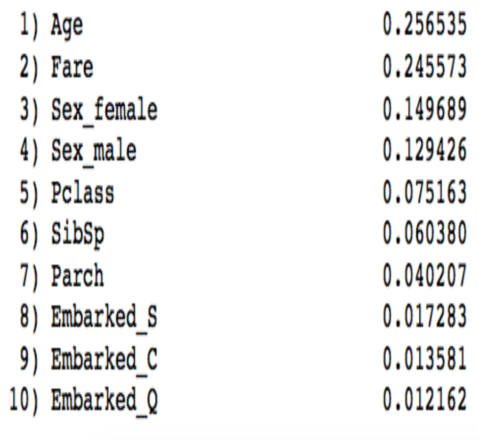
After completed the data preprocessing part, we started building models for our dataset. The first model we considered to build was the logistic regression model since there are several categorical variables in the dataset. Then we built the other two models: KNN and Decision Tree and compared with the first model. Finally, we built a better model using Random Forest. For each model, we have calculated the precision, recall score, f1 score, accuracy and AUC in order to compare the efficiency. By comparing the models, we have concluded that Random Forest has the best result.



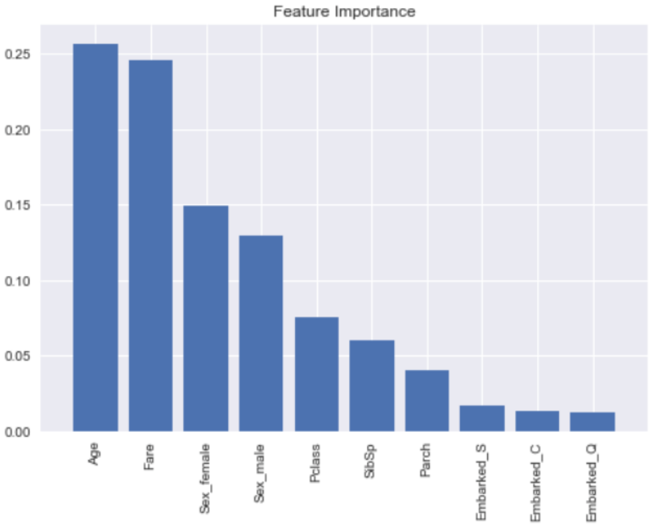
	AUC	Accurancy	F1_Score	Precision	Recall_Score
Logistic Regression	0.779643	80.26	72.36	72.73	72.0
KNN	0.759643	85.87	69.74	71.58	68.0
Decision Tree	0.742619	97.43	67.37	71.11	64.0
Random Forest	0.784524	97.43	72.92	76.09	70.0

5. Model Optimization:

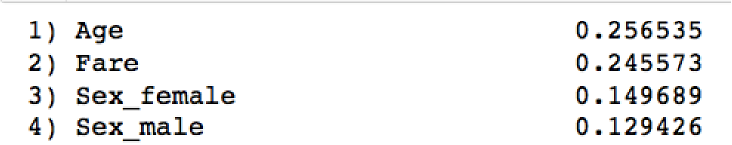
After comparing the models, we need to optimize our random forest model. By doing this, we first analyzed the importance of different features based on our random forest model.



```
 1) Age                         0.256535
 2) Fare                        0.245573
 3) Sex_female                  0.149689
 4) Sex_male                    0.129426
 5) Pclass                      0.075163
 6) SibSp                       0.060380
 7) Parch                       0.040207
 8) Embarked_S                  0.017283
 9) Embarked_C                  0.013581
10) Embarked_Q                  0.012162
```



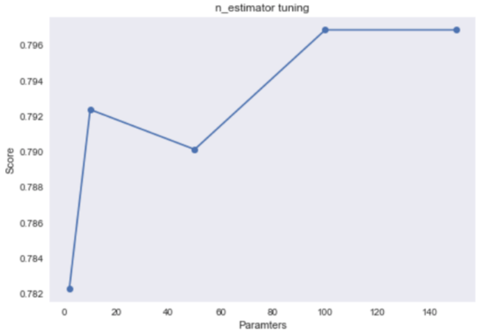
Next, we will try to select the most important features using SelectFromModel from sklearn.feature_selection. It turns out that only 4 variables will be needed to build the model which will give a relatively accurate prediction, and they are “Age”, “Fare”, “Sex_female” and “Sex_male”.



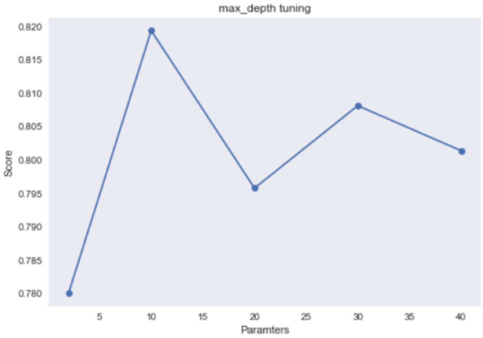
```
 1) Age                             0.256535
 2) Fare                            0.245573
 3) Sex_female                      0.149689
 4) Sex_male                        0.129426
```

Therefore, we will update our training set and testing set, which will only contain 4 features.

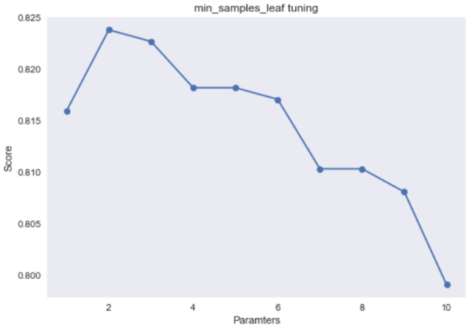
After we have updated our dataset, we will start working the parameters of our random forest model. For n_jobs, we found out that -1 have a better efficiency compared to n_jobs = 1.
Also, when n_estimators is set to 100, it will provide a higher score:



When max_depth is set to 10, it will result in the highest score

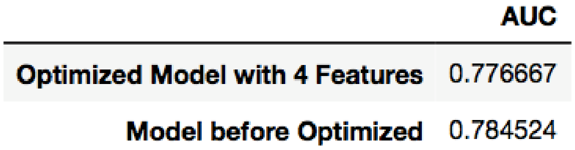


When min_sample_leaf is set to 2, it will result in the highest score



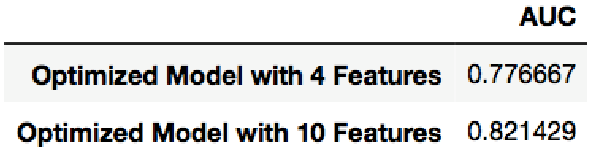
Finally, with the optimized model and updated dataset, we will be able to build our optimal random forest model. The new random forest model we built was given 4 parameters: n_estimators = 100, max_depth = 10, min_samples_leaf = 2 and n_jobs = 1. We will apply the updated random forest model on the updated training set and testing set, which only contain 4 features: “Age”, “Fare”, “Sex_Female” and “Sex_Male”.

After comparing the optimized model with 4 features with the previous random forest model, we found out that the model with 4 features has a lower AUC value which might not give better prediction.



	AUC
Optimized Model with 4 Features	0.776667
Model before Optimized	0.784524

The reason we assumed is because although 4 features can give a relatively good prediction, but it is not accurate as a 10 features model. Therefore, we will apply our optimized model on a 10 features dataset. It turns out that we have a AUC value of 0.82, which is much higher than all the models we had. Therefore, we concluded that when applied the optimized random forest model on a 10 features dataset, it will give the most accurate prediction.



	AUC
Optimized Model with 4 Features	0.776667
Optimized Model with 10 Features	0.821429

5. Limitation and Conclusion

Due to the fact that there are too many missing values in the variable “Age”, it is unlikely to make a perfect imputation, which might affect the prediction result. Overall, the last model we made has came up with a very high AUC value, which we believe it will give a relatively accurate prediction on the survival probability on the Titanic.