Assignment 2

Carlo Dalla Quercia Jingyi Li Kewei Wang

Kaggle team name : Carrot with a coin

1) Feature engineering

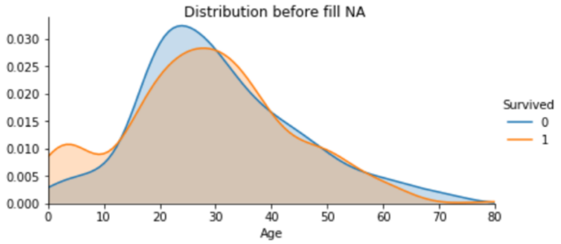
The general strategy of our feature engineering and creation is to study each of the variables of the initial data set one by one, visualize the relationship between each variable and the survival rate to help figure out if the variable has potential impact on the prediction target which is survived or not in this case. Then we contained all the relevant features in the initial model to test if the model can reach a high accuracy, and if not, dimensionality reduction can be applied to cut features with lowest contribution and optimize the combination of feature subset.

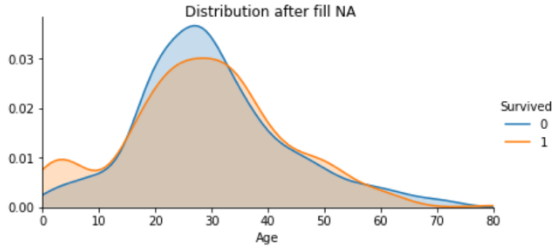
The original data set contains 11 features including PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin and Embarked. The passenger ID varies from each of the passenger, so it's meaningless to include it in key features. The feature engineering process is indicated as below.

a) Age

The initial data set has a huge loss of information on Age and Cabin with 20% and 77% missing rate respectively. If we simply fill the missing value with mean, median or mode, the result can lead to a huge bias.

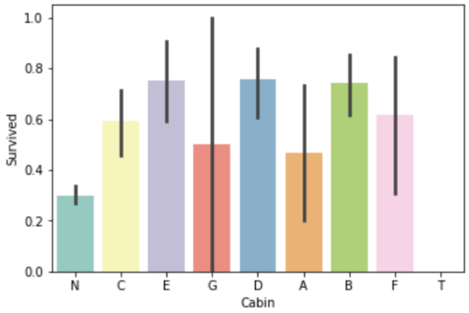
To retain the distribution of age, we built a random forest regressor to estimate the missing value on current numerical values containing Pclass, SibSp, Parch and Fare. As is shown in the following two graphs, the method ensures the age follows similar distribution before and after filling the missing value. Besides, from the distribution graph, it's clearly shown that children are more likely to survive than others, so age can be a key feature to the model.





b) Cabin

Most of the Cabin information is missing, so instead of using the original feature, we transformed the feature by extracting the first alphabetical value of the cabin such as A, B and C and replacing the empty ones with "N". The survival rate of class E, D and B are higher than others, so the feature is likely to have impact on the model.



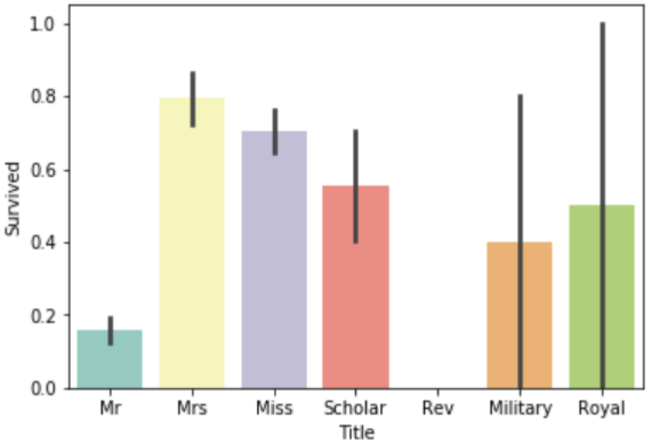
c) Numerical features

The rest of the numerical features are Pclass, SibSp, Parch and Fare. By analysing the distribution of these features, we found that passengers who bought first class tickets with 1 or 2 siblings or spouses or with 1 to 3 parents or children are more likely to survive than others. People who bought the ticket in lowest prices suffered from a high death rate.

d) Title

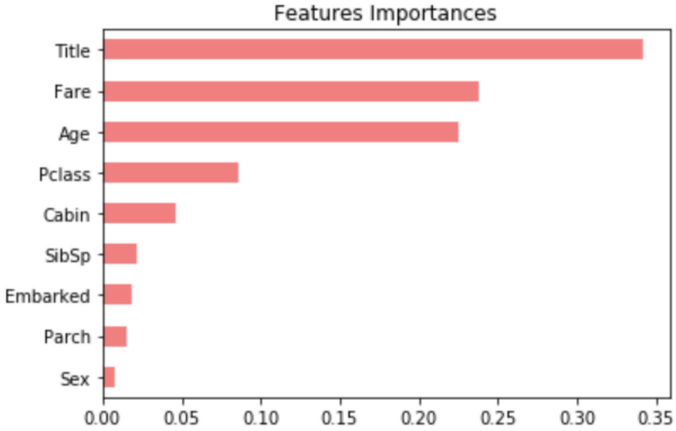
We created a new feature called Title from the original feature Name. We extracted the title in each name such as Mr., Mrs, Capt. and so on. We found some overlap and common points in the title, so we grouped them into 7 categories in the following way. The survival rate varies from each group, so it may have some impact on the model.

Category	Title
Mr	Sir, Don, Mr
Mrs	Mme, Mrs
Miss	Ms, Miss, Lady, Mlle, Dona
Scholar	Dr, Master
Rev	Rev
Military	Capt, Col, Major
Royal	Jonkheer, the Countess



Summary

In total, nine features are used in our first model with categorical features including Sex, Cabin, Embarked, Title and numerical features including Pclass, Age, SibSp, Parch and Fare. Our model has reached a quite high accuracy of 98.7% and the number of features is not that large, so dimensionality reduction is not necessary. The following graph shows the feature importance and the Title, Fare and Age show significant influence on the survival result of a passenger.



2) Model tuning and comparison

When it comes to the model selection, we choose models for classifiers definitely, and we test different models in order to choose the best one.

a) Random Forest

Random forests or random decision forests are an ensemble learning method for classification, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Here, we use the RandomForestClassifier function and apply the other function RandomizedSearchCV to decide the optimal parameter.

The final accuracy rate for the train set is 0.9057, and we also perform the cross-validation, the result CV score is: Mean is 0.8384485 and Std is 0.0424357.

b) Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Here we use the classification method. We use the DecisionTreeClassifier function to build it.

The final accuracy rate for the train set is 0.9876, and we also perform the cross-validation, the result CV score is: Mean is 0.794628 and Std is 0.04530771. The decisiontree's image is attached in the appendix.

c) KNN

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. Here we use it for classification. We use the KNeighborsClassifier function to build it.

The final accuracy rate for the train set is 0.8215, and we also perform the cross-validation, the result CV score is: Mean is 0.7194564 and Std is 0.01417451.

d) Logistic Regression

The simplest model, and we use the function LogisticRegression to get the model.

The final accuracy rate for the train set is 0.8148, and we also perform the cross-validation, the result CV score is: Mean is 0.8125545 and Std is 0.02716524.

3) Final Submission Results

We exported the results with different methods and submitted them to the Kaggle. The model with best performance is the Random-forest model, with the accuracy of 0.79425, while the decision tree model which has the best accuracy in training set only has 0.6953 accuracy in the test set, and other models' accuracies are all below 0.65. Apparently, the decision tree model has overfitting problems after testing in the test set, therefore, we regard the Random-forest model as the best predicting model in this assignment.

Appendix: Decision Tree