**Kaggle Titanic Project:**

**Author: Jianqiu Chen**

**Introduction:**

This project aims to provide basic statistical analyses for the Kaggle Titanic Data Competition, more specifically, these analyses attempt to provide a good statistical model for prediction survivals.

**Context:**

The dataset can be downloaded at <https://www.kaggle.com/shuofxz/titanic-machine-learning-from-disaster>

The datasets are also available in my git repository.

**Objective:**

The goal is simple, the person who comes up with the model that leads to the best prediction accuracy on the test data would win the competition. However, this is not the only goal of this project. This project also aims to provide some interpretable insights into the factors that are associated with survival through data visualization.

**Coding:**

A well commented R file is also available in the same repository as titanicProject.R which includes all coding implementation of the algorithms used in this project. The file can be run by any R IDE environment or R console.

**Data Inspection:**

Kaggle competition has split the entire titanic dataset into a train set and test set, both can be found in the same repository.

The train set has 891 observations and 12 variables with Survival status as response variable and all other variables as predictor variables. The test set has 418 and 11 variables with no response variable for obvious reasons.

**Data Description:**

A detailed data description is provided at <https://www.kaggle.com/c/titanic/data>

**Preliminary Data cleaning:**

Preliminary data inspection shows that there are several variables that are irrelevant to our analysis, namely Ticket Number, Passenger Id, Name and Embarked. Intuitively, one’s name or ticket number or ID assigned as a passenger or location at which embarked has no effect on the person’s survival status from the disaster. Hence, we immediately remove these variables to sharpen our analyses.

Further, the variable cabin as too many missing values such that no missing value retrieval method could be used. Thus, cabin is removed as well.

After removing the five variables mentioned above, one can see that the training data still has 177 missing values in the age variables. Intuitively, age could be an important variable when it comes to prediction because there could have been unspoken rules that favorize kids and elderlies when it comes to survival.

To save time and facilitate analysis, this project will simply remove all rows containing missing age information, although this approach leads to information loss and weakened predictive power. As a side note, one can repair age information by tracking family status, for instance, if 2 people have the same last name and 1 of them is in her 50s, one can infer the other could reasonably be in her mid 20s or early 30s. Furthermore, many statistical information retrieval techniques can be applied as well.

After removing the 177 rows, one’s left with 714 observations.

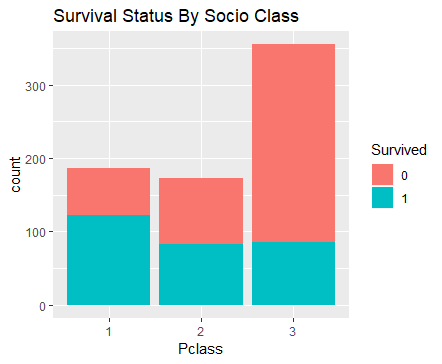
**Data Visualization:**

While there are many interesting questions one can ask about the dataset, this project attempts to look for insights regarding the effects of age, sex, social class have on survival status.

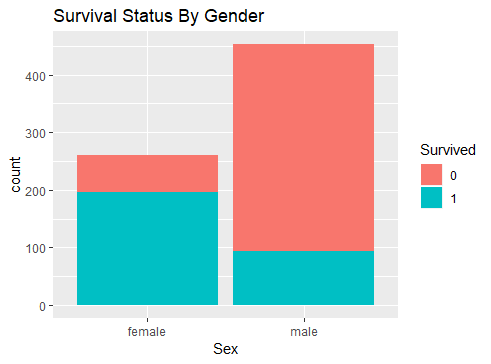
The followings are the graphs made in this project:



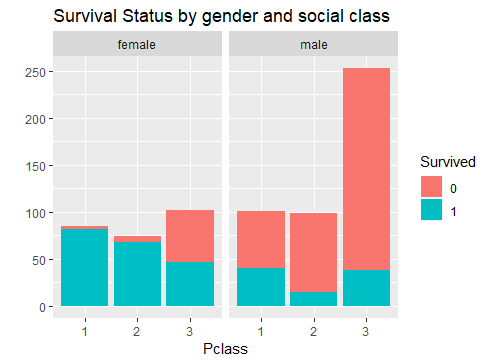
The age histogram shows that the passenger are generally young with the majority in working age.



The Survival Status by Socio Class Plot illustrates a correlation between social status and survival chances, namely, people with higher social status seem more likely to survive the disaster. Similarly, if we look at Survival status by gender:



The graph clearly indicates that females are more like to survive. The next natural question that comes to mind is, does the fact that females are more likely to survive change based on social class?



There many interesting features about this interaction plot, if one compares first and second-class females to third class females, turns out that most females in first or second-class survived while the third-class females are not as lucky, about half of them have not survived the disaster. Switching to males’ graph on the right, first, if the person was a male, he would have had not much chances to survive. The chances are slightly increased if the male is in first class.

A couple more plots are included in the project that would deliver more insights regarding the survival associations.

**Data Prediction:**

In this project, logistic regression, linear discriminant analysis, Adaboost, random forest and support vector machines will be used as our prediction models. For each model, the training set is randomly split into a training set and a test set to evaluate performance, then a 7-fold cross validation will be used to evaluate the stability of the prediction and thus help one come up with the most suitable model.

**Logistic Regression:**

The first natural thing to use is logistic regression for binary response variable. After implementing logistic regression with Survival Status as response variable and all remaining variables as predictors, if one use 0.5 as determination threshold, in other words, if one assigned survived to a prediction greater or equal to 0.5 and not survived to a prediction less than 0.5, one would reach a prediction accuracy of 80.5%. The seven-fold cross validation averages to 80.0%, the result is in the appendix.

**Linear Discriminant Analysis:**

Using the exact same approach, LDA gives 77.6% accuracy on a random split of training set and test set. The 7-fold Cross Validation confirms the average prediction rate to be 78.7%

Hence, LDA does not really provide any improvement over logistic regression.

**AdaBoost:**

Using Ada Boosting algorithm, one gets 81.3% prediction accuracy on a random split of training set and test set. The 7-fold Cross Validation leads to 81.9%. However, the variance for the 7-fold CV results is quite big, this makes perfect sense because each fold only contains 102 observations in the cross validation. Ada boost works more stably on larger datasets. Nonetheless, as the entire training set contains 714 observation, Ada boost should not be excluded from our repertoire.

**Random Forest:**

Using random forest with 2 parameters choice at each tree, and 100 trees in the algorithm, one obtains 85.5% prediction accuracy, however, cross validation gives an average of 81.8% accuracy. Once again, this decrease in performance is largely related to sample size, random forest outperforms when the sample size is larger. Notice that the choice of mtry and ntree is random, hence parameter tuning can be used for improving on random forest.

**Support Vector Machine:**

Using SVM with cost tuning which leads to cost =2.28 for best performance. The prediction on a random split between training and test set gives 80.8% accuracy, which is not as good as random forest. In theory, it makes sense because random forest outperforms SVM in a dataset with a mixture of numeric and categorical features, while in SVM, it is important to define the concept of distance between points.

**Conclusion:**

This project provides a preliminary model selection, it turns out that for the titanic dataset, random forest is the best model among the models used in the project. Parameter tuning should be used for random forest to yield even better prediction accuracy. Since this dataset is small, a neural network would very likely lead to overfitting and thus was not considered in the project.