STA 5224: Final Project - Titanic Dataset

Huong Tran

3/27/2022

I. Project Proposal:

Objective:

This project is a competition on Kaggle with target of predicting which passengers survived the Titanic shipwreck by machine learning. Part 1 of the project will focus on cleaning data and obtaining some insights about data, together with variable selection process to create a meaningful dataset that can be used. Model application will be in the second part. Besides the statistical approach in logistic models, other machines learning model will be used. A typical model for supervised learning is decision tree, which is a series of sequential decisions represented as a tree to get a specific result. However, decision tree are prone to overfitting, especially when the tree is deep and random forest is a solution for this kind of problem. In general, a random forests model will creates several random decision trees and aggregate their result. Also, SVM works well with classification and regression, therefore it is good to represent SVM. Finally, the comparisions between these models will be derived. In the last section, statistical inference and interpretation from logistic models will be discussed.

About the dataset:

The data is obtained from the Titanic competition from Kaggle. While the train.csv will be used for model training, the test.csv and gender_submission.csv will be used to evaluate model performance.

The dependence variable is "Survived", which has value 0 or 1, indicates that the person survived after the disaster or not. The others are exploratory variables, with their meaning can be find at the website.

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```
## [1] 891
```

There are 891 observations and 10 features in this dataset, since PassengerId is unique, we can not extract any information from this variable and "Survived" is the dependent variable.