Assignment 3 Write-Up

The Titanic dataset provided consists of a total of 1309 passengers on board along with the characteristics of each passenger. These characteristics include ID, Name, Sex, Age, Survival, Ticket Number, Ticket Class, Passenger Fare, Number of Siblings and Spouses Aboard the Titanic, Number of Parents and Children Aboard the Titanic, Cabin Number, Port of Embarkation.

Data Preparation:

1. The field of predictive interest in my dataset is “Whether a passenger would survive”.
2. I looked at the initial summary statistics and checked for missing values. Missing values are found in ‘Age’ and ‘Cabin’, After removal, the dataset contains a total of 1043 passengers.
3. I decided to drop ‘Cabin Number’, ‘Ticket Number’, ‘Name’, ‘ID’ from the dataset because these characteristics are not meaningful nor needed, and details will be provided below.

Exploratory Data Analysis:

1. I used various types of graphs to look at how each characteristic is distributed as well as converting some of these into group indicators.
2. I noticed that over 60% of the passengers are embarked from Southampton, over 50% of the passengers are between age of 20 to 39.
3. I created a correlation matrix as well as calculating the correlation between each characteristic and ‘Survival’.
4. From correlation matrix, I found that ‘Fare’ and ‘Class’ are highly correlated, so I dropped ‘Fare’ from the dataset. It makes sense because the fare is usually priced accordingly to each class.
5. I also found that female has a significantly higher survival rate than male, which indicates that female passengers might have priority to use lifeboat during evacuation.
6. Both ‘Number of Parents and Children Aboard the Titanic’, ‘Number of Siblings and Spouses Aboard the Titanic has very low correlation to survival rate’, so I dropped it.
7. At the end, I only decided to use ‘Gender’, ‘Age’, ‘Class’, ‘Embarked’ as features in my model.

Model Implementation and Results:

1. The two model (Decision Tree and Probit) were implemented using a split of 32% testing to 68% training data sample. The 32:68 ratio is based on the original split ratio provided.
2. The classification tree estimated a 60.2% death rate in the Titanic data with 85% accuracy rate. ‘Gender’ is the strongest predictor of outcome.
3. The probit regression model further confirmed that ‘Gender’, ‘Age’, ‘Class’ and ‘Embarked’ are four significant predictors to the outcome, and ‘Number of Siblings and Spouses Aboard the Titanic’ and ‘Number of Parents and Children Aboard the Titanic’ are not important. Lastly, ‘Gender’ is also the strongest predictor in probit regression model. The accuracy rate is 33% which is really bad.

Comparison of Classification Tree and Probit Regression:

1. Classification tree is easier to explain for audiences with no technical background, but time consuming to implement.
2. Probit regression is very simple to implement but difficult to understand for audiences with no technical background.
3. Each method has its trade-offs, but I think in this case, the classification tree performs better.