**Introduction**

In this assignment we are going to use Titanic dataset and check what were factors that decided who had the best chances of surviving. This study will help us to understand different situations and how we can predict the past even and draw a possibility of what could be the reason behind a particular situation or event. We all know how the ultimate sinking of titanic was and how it is still a question on how people survived the tragic incident. While we try to see various factors to understand what the reasons of one surviving could be.

We are going to build different models and compare what are factors that will help us to best decide the best fit model for out dataset.

Steps:

1. Check data for cleanliness and any outliers

2. Choose the appropriate model to predict whether or not someone survived

3. Build a Decision Tree Model and then either a Random Forest or Gradient Boosted Model.

4. Build a Logistic Regression Model

5. Compare the three models.

6. Communicate to historians which features decided who lived or died, which model best represents this and what feedback can we use from this to understand survivorship of other disasters.

**Problem Statement**

Looking at the dataset we have multiple variables that we have to look into to see what could be the one reason that decided on who survived and who died. This prediction will help us to understand the historic events.

**Dataset Details**

|  |  |
| --- | --- |
| **Dataset Key points** | **Values** |
| **Source** | Provided |
| **Tentative Entries** | 888 |
| **Number of Attributes** | 8 |
| **Dependent Variable** | Survived |
| **Independent Variable** | Pclass, Name, Sex, Age, Siblings/Spouses abroad, Parents/Children Abroad, Fare |

**Data Analysis**

After loading the dataset, we check dataset for outliers and missing values. We noticed that there were no outliers and missing values in the dataset. Therefore, we move forward with the analysis and we check for the class of the variables. We notice that sex is not numeric value to perform the correlation. So, we change the labels into levels for performing correlation between the variables. We replace male and female with 0, 1 level. Now we perform correlation to check the variables. We notice that variables Age, Pclass, Parents/ Children and fare have high correlation with variables. We drop the Name column as it is not of any us for the prediction.

We built graphs to check the variable relation with the targeted variable. We check relationship between the four variables selected with Survived variable to check which could be one of the major reasons that had higher chances of surviving.

We start with the models now. First, we perform logistic Regression using these 4 variables and see that the accuracy of the logistic model is 65.16% only. We have to compare this model with couple of other models. We build decision tree with max depth =5, we get different results depending on the levels. By 5 max depth we get the accuracy of 70.7% for decision tree model.

We build another model Gradient boosting for comparison with decision tree and logistic regression model. We see that gradient boosting model has the highest accuracy of 71.9% as compared to other two models. So, we can conclude that we should Gradient boosting model for this dataset and it is the best fit model for titanic dataset.