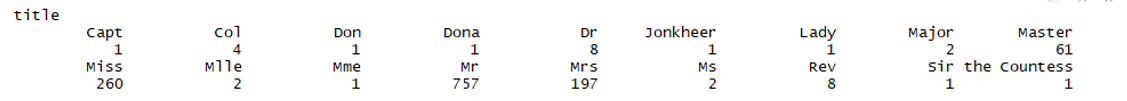
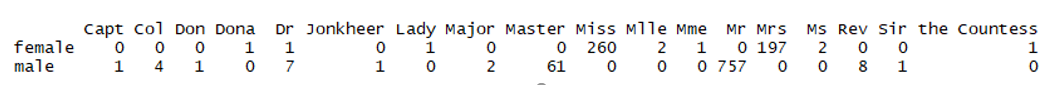
**Feature Engineering** (missed to include in my last document)

I missed to add feature engineering for titles in my last document (Milestone 3). Sex variable doesn’t tell us whether a person is married or single but the name title does. Title helps us to predict if they are in family or not. Let’s extract title from names.

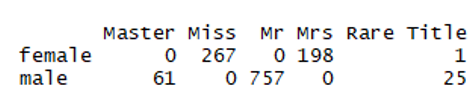
So here we now have a nice new discrete column of titles as shown below:



If sorted with respect to sex it looks as below:



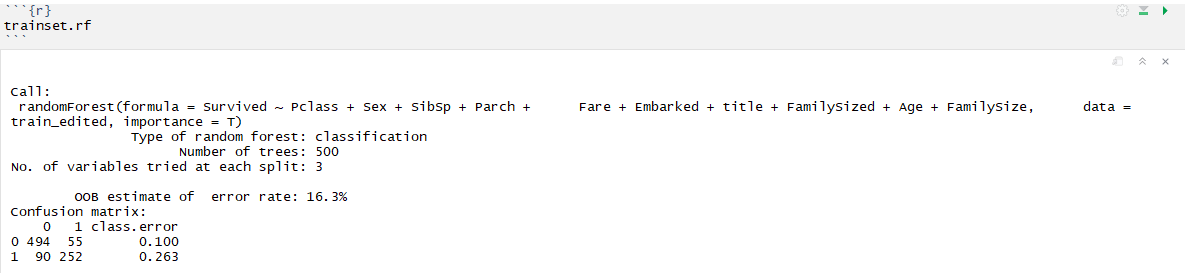
Here titles with very low cell counts can be combined to "Rare Title" level and I have also reassigned mlle, Ms, countess and Mme accordingly. The below title count shows by sex again



**RandomForest Model**

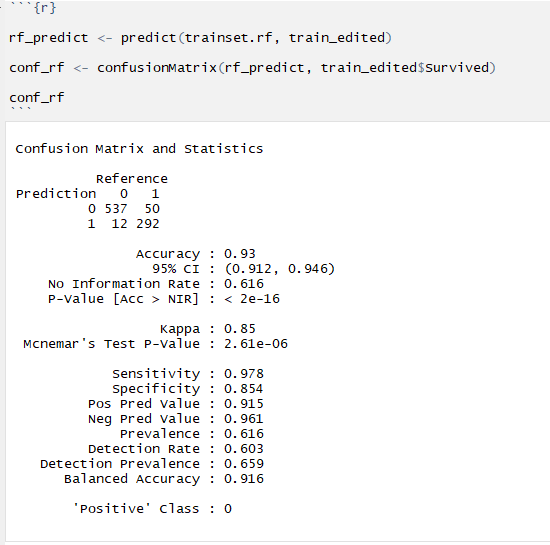
My last data modeling document had Ctree and a decision tree model for the train dataset. Now will try using Randomforest. Randomforest is an ensemble learning method that grows multiple decision trees and each decision tree will produce its own prediction results. Will use the same variable to build the random model





Out of bag (OOB) error is the overall classification error in the above model and as we can see, OOB error still is more than 15%. This means that we still can reduce the OOB error in the model.

The below Confusion matrix is created based on the train data setusing RandomForest - the overall confusion matrix accuracy is 93%



Now it time to get the error rate which is as shown below in Fig1. Below plot shows mean square error rate. The green line shows the error rate for survival whereas red lines shows for the dead. Plot is telling us that we are more successful predicting death than survival.

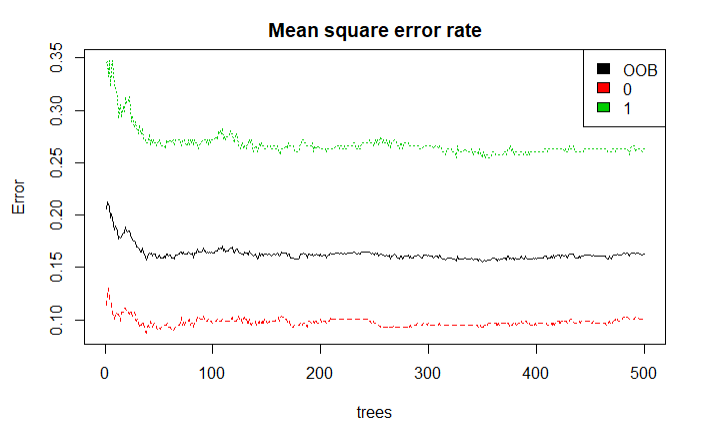
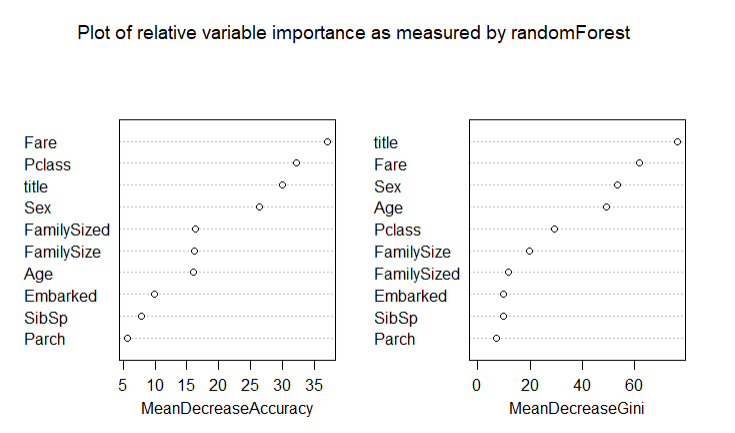
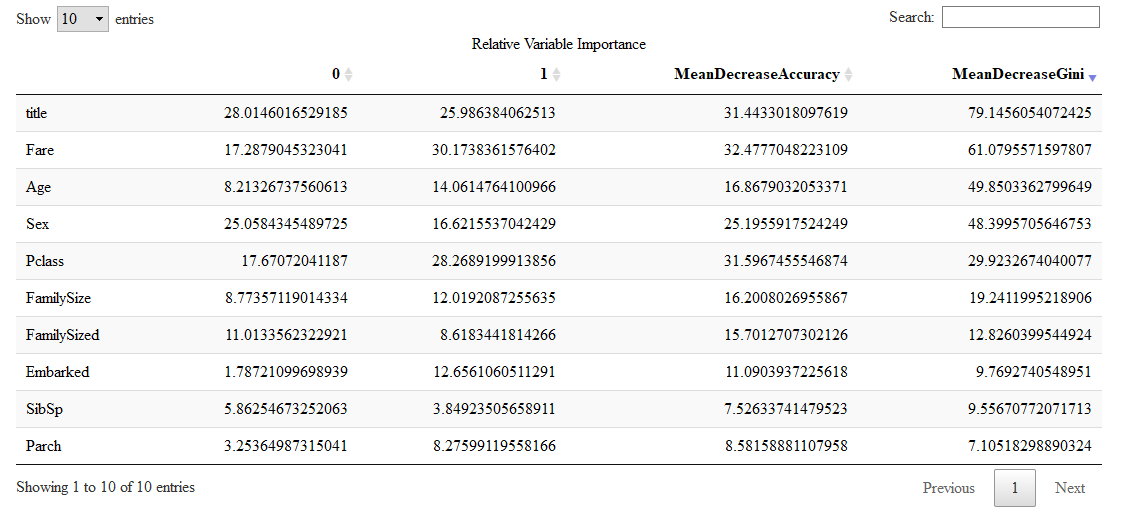


Fig1: Error rate

Now we can move onto finding the important variable. The below plot validates what ctree predicted that Title is the most important variable. Sometime it is necessary to rank a variable based on its importance. Here I feel it is Title. Now we are ready for the final prediction using the training\_rf we have created on the train dataset and apply on the test dataset. But before that I would go through the extractor function for variable importance measures as produced by randomforest



We can try eliminating the least important variable (parch and SibSp) but the variable we have used here to build the model is what I think are the one which can predict the people most likely to survive in such an accident.