STAT 745: Team Project - Titanic Dataset
STAT.745-F22.RIT.Team4: Cate Renner, Prasanna Rodrigues, Erik Schuster, Nick Smith

0.0 Project Overview

The Titanic dataset is derived from the information on the manifest of the ship's passengers. This includes whether or not the passenger was to survive. The group is tasked with developing a predictive model to determine the survival of a passenger, given the factors. The objective is to maximize the predictive accuracy of the model onto the test dataset. This generates a score via the Kaggle competition interface.

0.1 Data Understanding/Overview

This research paper examined the Titanic dataset, which contains 891 observations and 12 variables. The target variable "Survived" indicates whether the passenger survived the sinking of the Titanic or not. Around 38.4% of passengers survived the Titanic in the training data (shown in Figure 0.1 in appendix). Upon initial analysis, it was revealed that approximately 19.87% of the age feature was missing in the training data, and 20.57% of the age entries missing, and .24% of the fare entries missing for the test data. Most (687/891 = 77%) of the value of the Cabin predictor were missing, and the predictor had to be dropped from the dataset. It was found that the majority of the passengers did not survive and that the majority of the passengers were male. The plots revealed that the majority of the passengers who did not survive were in Pclass = 3 and that Pclass may be valuable in terms of feature importance. The kernel density plots showed that there is a "peak" in younger passengers who survived, indicating that age may be a valid feature for modeling. There are graphs in the appendix that visualize this data understanding.

1.0 Data Cleaning

There were several techniques tried throughout this project. Each technique that was tried and tested will be detailed. The first technique was extracting the title of a person's name from the Name variable. This was done by splitting the string and pulling out the title. Character-type variables were converted to factor variables. Dummy variables were created for nominal values in some cases. There were several imputations tried since many Age variables were missing. Another technique tested for data cleaning was unwanted columns like Passenger Id, Cabin, and Names were deleted from the data set. Predictors of type characters were converted to numbers. Sex was coded as 1 for males and 0 for females, Embarked was coded as (1, 2, 3) for ports (C, Q, S) respectively, and P-class was re-ordered as ("1" = 3, "2" = 2, "3" = 1) so that class 1 will have the highest weight.

The first imputation method tried was the step functions imputed the mode of nominal predictors, and imputed the median of continuous predictors. A correlation filter was tuned to de-correlate continuous predictors above a threshold of .85. For consistency, a series of steps then transformed the predictors via range reduction, centering, and scaling continuous values. The next imputation method tried was computing the median age for each gender and group and using that median to fill in the missing ages. After that, caret has a method that takes missing values into a model and passes them to the model technique for that package to deal with. Next, mean age and missing values of Embarked were imputed by the most occurred embarked value. Then, the second technique was Amelia imputation which was done

using amelia package. Finally, the third technique was the MissForest imputation using the MissForest package. Our final model ended up using mean imputation for age.

1.1 Modeling

Many different modeling techniques were tried with each imputation and different data cleanings mentioned in the section above. First, Model Set 1 consisting of Logistic Regression (regularized), XGBoost, Ranger Random Forest, and C5.0 Decision Trees were fitted via tune Bayes and Gaussian Process Regression using the beginning imputation techniques. Each model was k-fold cross validated with 10 folds. The performance was relatively poor in the test set. There was approximately a test accuracy of .75 overall for each of the models. Each of these models had an ROC above .83 on the training set, suggesting it was overfit. It should be noted the transforms above allowed the logistic regression model to meet ht assumptions. C5.0 was the next best performing model, with the same procedure, and cross validation methods. This scored .765 on the test set. Random forest with the ranger computational engine was the best performing model in comparison to the other models in the model set tested. It received a .778 test score, which while better, still implies it was overfit to the training data.. This model surprisingly showed ROC scores above .86 on the training set, further implying overfitting. Many of the models that were run and tested were found to be overfitted.

Model Set 2, uses caret imputation and imputation based on gender and class as detailed in the Data Cleaning section. A XGBoost, C5.0 and a Random Forest were run for both imputation techniques. They each were hyperparameter tuned using caret, a grid search was done on the XGBoost and C5.0 on multiple hyperparameters. The Random Forest was tuned to only mtry. All models were cross validated on a k-fold with 5 repeats and 5 folds. The best model out of this set was a XGBoost imputed by caret with a train AUC of 0.874 and a test accuracy of 0.758. This model was also found to be overfitting on the training data. This can be seen more when looking at the variable importance for this model. The model relies heavily on only a couple variables, being a male was the most important variable by a lot. This shows that the model could be overfitted and is shown true when running the test accuracy. Since the model is overfitted, it cannot be trusted to have consistent results for any test set.

Model Set 3, the imputations used were Amelia, missForest and mean/mode that were discussed above. The Random Forest and C5.0 models were tested. The C5.0 Decision Tree model had better accuracy than the RF model. C5.0 with Amelia imputation gave the highest accuracy of 88.10% but since the Titanic data is not multivariate normal the Amelia imputation is not valid as it assumes MVN for the data. The C5.0 with MissForest imputation and Mean Mode imputation had the same accuracy of 87.21%. The C5.0 model with MissForest imputation, 10-fold CV, and parameter tuning gives an accuracy of 91.13%. C5.0 (ROC AUC = 0.9447081), It had a test score of 0.78468 and the C5.0 model with Mean Mode imputation, 10-fold CV, and parameter tuning gives an accuracy of 89.79%. C5.0 (ROC AUC = 0.9290922), It had a test score of 0.7679. Among the two potential models, the C5.0 model with MissForest Imputation gave a better precision on the test data. **But still both this model seems to be overfitted as they have a huge difference in the train and test accuracy measures.** Also tested a simple model that is a logistic model on the data to see if it performed any better. The accuracy of the GLM model was 82.15% at the threshold of 0.562 and the resulting test score was 0.77751.

Model Set 4 looked at mean imputation for age and mode imputation for Embarked. This model set began with a Logistic Regression model for its simplicity, but found that increasing the number of variables generally made this model perform worse. This concurs with the results of Random Forest models, which seemed promising, when breaking the training set into train-test samples, with a ROC AUC of 0.811 and an average out-of-sample accuracy score around 0.85, but gave a test score of 0.75837 – lower than the benchmark set by "gender_submission.csv" (prediction based solely on gender). Both Logistic Regression and Random Forest appeared to be overfitting quite a bit, regardless of data-prep.

Keeping simplicity in mind, the number of input parameters were reduced to a handful of the most important variables (Pclass, Sex, Parch, and Age), and run through a LDA function. The resulting model had a ROC AUC of **0.7855** on the training set, with a training accuracy around 0.8. On the test set, we saw a score of 0.78468, indicating that this model was well fit. The function giving us this score was of the form: Survived~1+(Pclass+Sex+AgeNA)*Parch+Age (See Figure 0.8 in the appendix for details)

Given the success of LDA, QDA was tested due to the underlying similarities, but did so poorly in testing, it was not worth submitting to Kaggle to assess the model's prediction. Likewise, boosting the successful LDA model was tried in an attempt to achieve incremental improvements. This method looked promising in training with an accuracy around 0.84, but the test score was 0.77511.

1.2 Evaluation

The best performing model was the LDA model, using the data cleaning and feature engineering noted in Model Set 4. This model performed at an accuracy of 0.78468. This was in the top 14% of the submissions in the current competition date range, and is better than a submission just using gender. I our most successful model , we see the interaction of Pclass, Sex, AgeNA with Parch and their impact on survival. The confusion matrix and model output can be seen in Figure 1.0 in the appendix.

To increase the accuracy of the models used, and prevent overfitting - further feature engineering may improve the scores. This can include more sophisticated imputation methods for missing data, model ensembles tuned for accuracy, and filtering the dataset for redundant features that may be contributing to the noise of the model.

Appendix:

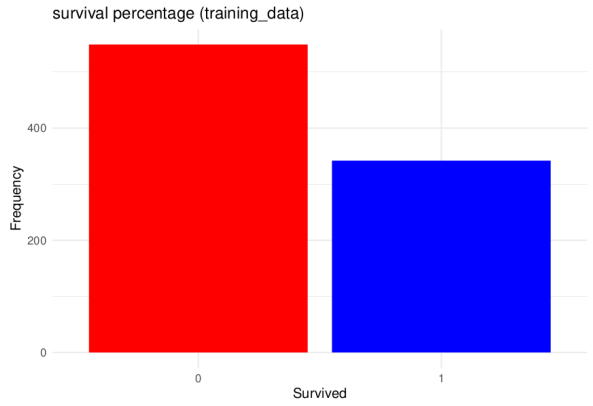

Figure 0.1: Proportion of Survival Percentage

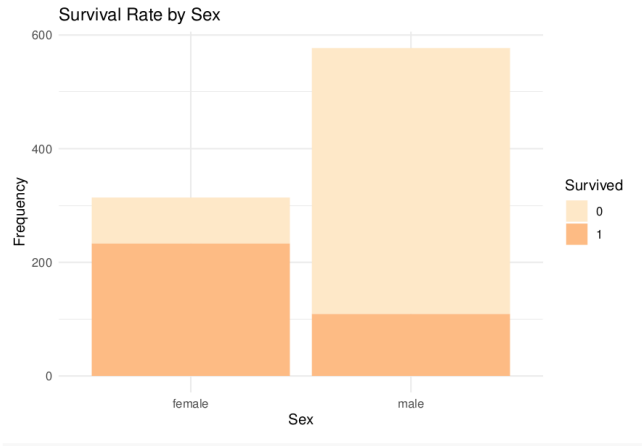

Figure 0.2: Male to Female Survival Rates

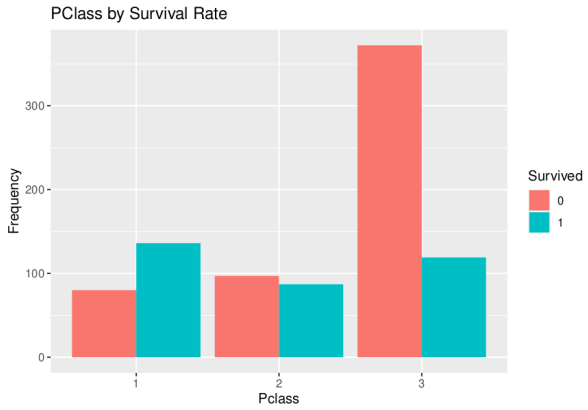

Figure 0.3: Passenger Class and Proportional Survival rate

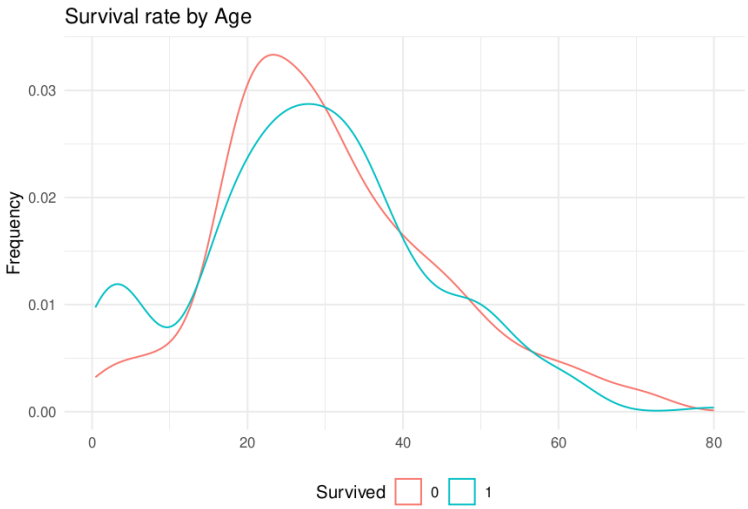

Figure 0.4: Survival Kernel Density with Age

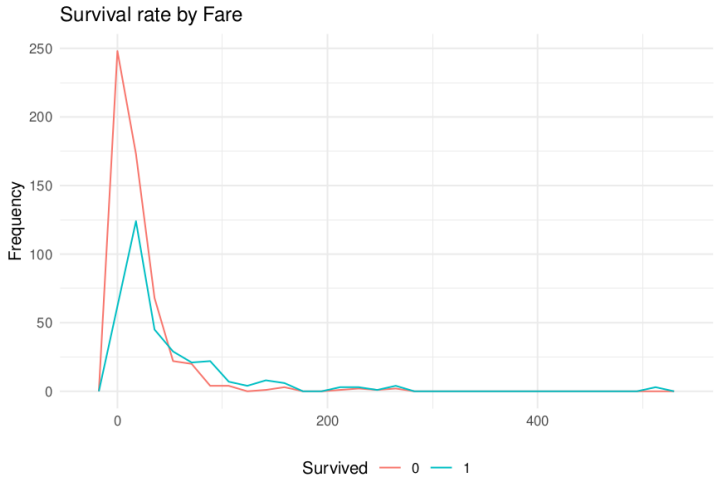

Figure 0.5: Survival Rate by Fare

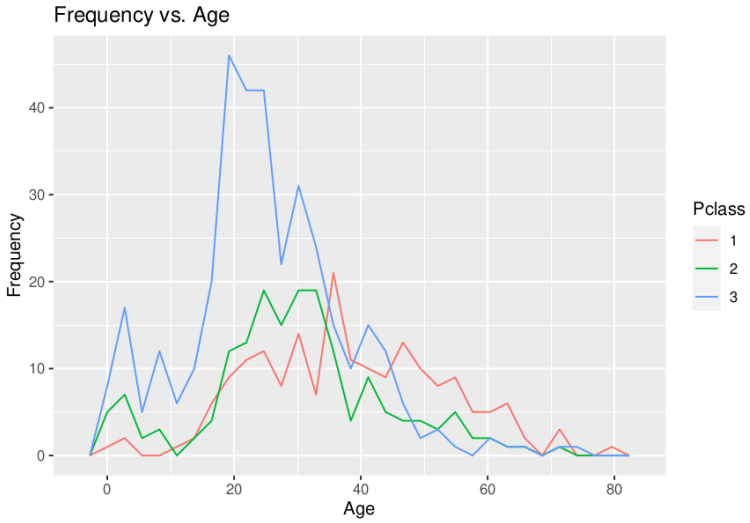

Figure 0.6: Passenger Class Age Frequency

```
predict(ranger_rand_forest_fit ,
        new_data = training,
        type = "prob") %>%
  bind_cols(select(training, Survived)) %>%
  roc_curve(Survived, .pred_Y) %>%
  autoplot() +
  labs(title = "ROC AUC with training data")
```

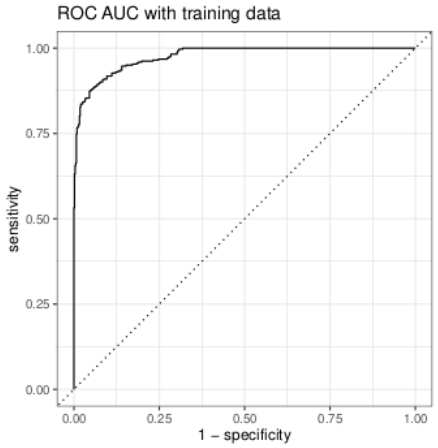

Figure 0.5: Random Forest ROC for Model Set 1

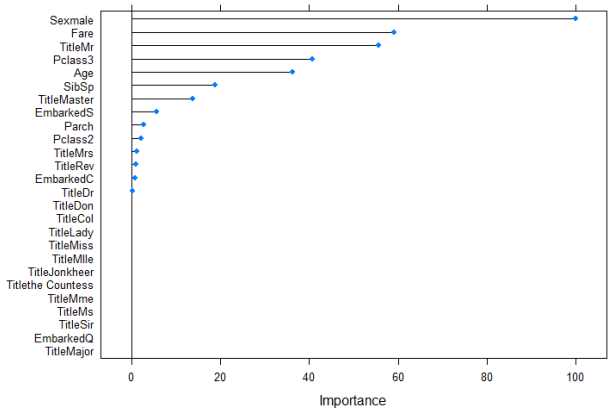

Figure 0.6: Random Forest Variable Importance for Model Set 2

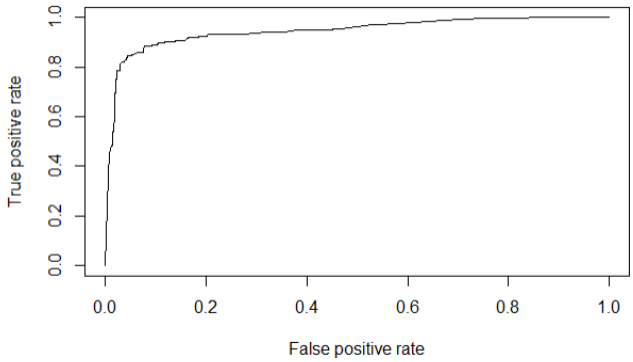

Fig 0.7: ROC Curve C5.0 with MissForest Imputation for Model Set 3

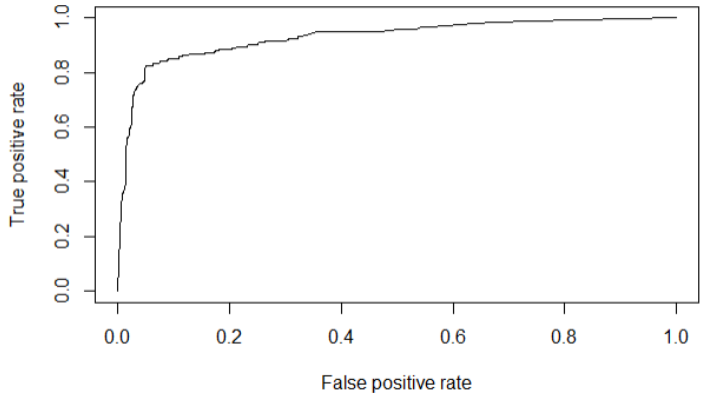

Fig 0.8: ROC Curve C5.0 with Mean Mode Imputation for Model Set 3

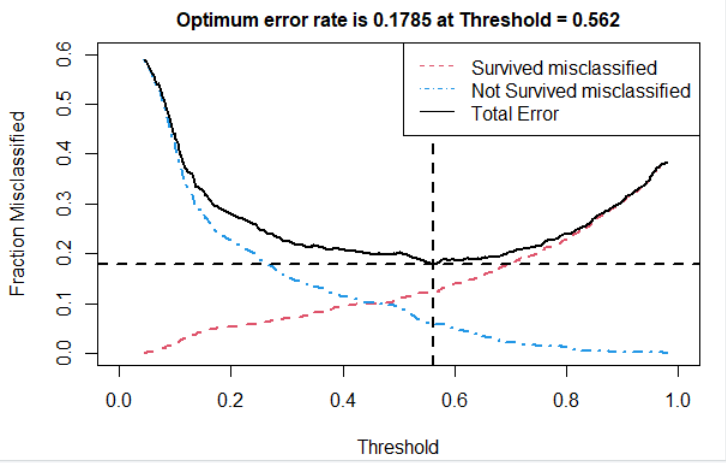

Fig 0.9: Optimal Error rate of Logistic Model for Model Set 3

```
> classif.table

pred   0   1
   0 482 105
   1  67 237
>
> sum(diag(classif.table)) / sum(classif.table)
[1] 0.8069585
> clf
Call:
lda(Survived ~ 1 + (Pclass + Sex + AgeNA) * Parch + Age, data = train[c("Survived",
    xs)])

Prior probabilities of groups:
        0         1
0.6161616 0.3838384

Group means:
    Pclass2   Pclass3   Sexmale     AgeNA     Parch      Age Pclass2:Parch Pclass3:Parch Sexmale:Parch AgeNA:Parch
0 0.1766849 0.6775956 0.8524590 0.2276867 0.3296903 30.41510    0.02550091     0.2604736     0.1766849  0.04735883
1 0.2543860 0.3479532 0.3187135 0.1520468 0.4649123 28.54978    0.16374269     0.1461988     0.1140351  0.01754386

Coefficients of linear discriminants:
                     LD1
Pclass2       -0.99215484
Pclass3       -1.38916815
Sexmale       -2.34866351
AgeNA         -0.03058051
Parch         -0.34586335
Age           -0.01802208
Pclass2:Parch  0.73122592
Pclass3:Parch -0.12979400
Sexmale:Parch  0.50098979
AgeNA:Parch   -0.31758153
```

Fig 1.0:LDA Model Output for Final Model