SUNY PLATTSBURGH

Health Care Analytics

MSA 580 - Final Report

Liver Transplantation

Authors:
Danish Rai
Bridget Hanley
Siddant Mehta

Supervisor:
Dr. Zahra Gharibi

May 07, 2020

1 Introduction and Motivation

Our project will be focusing on two different research questions focusing around patient survival time after a liver transplant. There are nearly 17,000 individuals on the liver transplant wait list, many as a result of End Stage Liver Disease (ESLD). Currently, there is no cure for ESLD other than a liver transplant. So for many of these patients, a liver transplant is their last option at living a longer life. Our goal is giving patients a better understanding of their chances of survival after they receive the transplant.

2 Data

The data we are using for this project comes from the Scientific Registry of Transplant Recipients (SRTR)[1]. The data contains variables that go back as far as 1987, though some variables in the data don't start collecting information until later years. The data-set initially started of with 404 variables and we were able to narrow it down to 46 variables that are associated to our quantitative and qualitative questions. We used a regression model[2] to generate Liver Size variable for recipients and donors. The new variable uses body weight(BW), age and gender to calculate it. The regression formula of Liver Size for age 16-50, 452 + 16.34 x BW + 11.85 x age + (-166.5 x gender) and for age 51-70, 1390 + 15.94 x BW + (-12.86 x age). Moreover, the data also contains demographic information of both the recipient and the donor; whether the donor was living or dead, and other health indicators for both subjects. The data also contains key diagnostic information of the recipients such as MELD Score (Model for End-Stage Liver Disease) (12 and older) and PELD score (Pediatric End-stage Liver Disease) (younger than 12). The calculation of these scores uses DoB, Gender, Bilirubin, INR, Albumin, height and weight to generate values between 6-40. Essentially, MELD score assess the severity of ESLD in a patient, giving a higher score for the more severe a patients condition is. The data also contains what region and OPO center that the liver was transplanted in.

The data set can be categorized in four main parts:

1. Donor information
2. Recipient information
3. Region
4. OPO

Data Cleaning and Preparation

After the final variables were selected, the data was filtered to only have patients with a MELD Score between 6 and 40 as our data contained negative values. Since our focus of study is adult patients who received a transplant from a deceased donor, we removed patients with a PELD Score (younger than 12) and living donors from the data set. We also filtered the data set so that Patient Status is in binary; alive or dead. For variable PTime, which contains information about the survival time in days for patient after transplant, we filtered it for greater than 365 such that the data only contains patients who survived the first year after their transplant. This was done to avoid patients who might have had unsuccessful surgery and/or got re-transplant within a year. Further more, 365 was removed from the variable PTime to see the number of days a patient survived one year after the transplant. This was done to easily filter the data set for the analysis. In addition, we normalized our data for our predictions since many of our variables were on vastly different scales. After everything, we had a data set with 47 variable and 61,600 observations.

3 Research Questions

1. Quantitative

Can we predict survival time for patients one-year after receiving a transplant.

2. Qualitative

Classifying if a patient survived one-year after receiving a transplant.

4 Preliminary Analysis

Correlation Plots

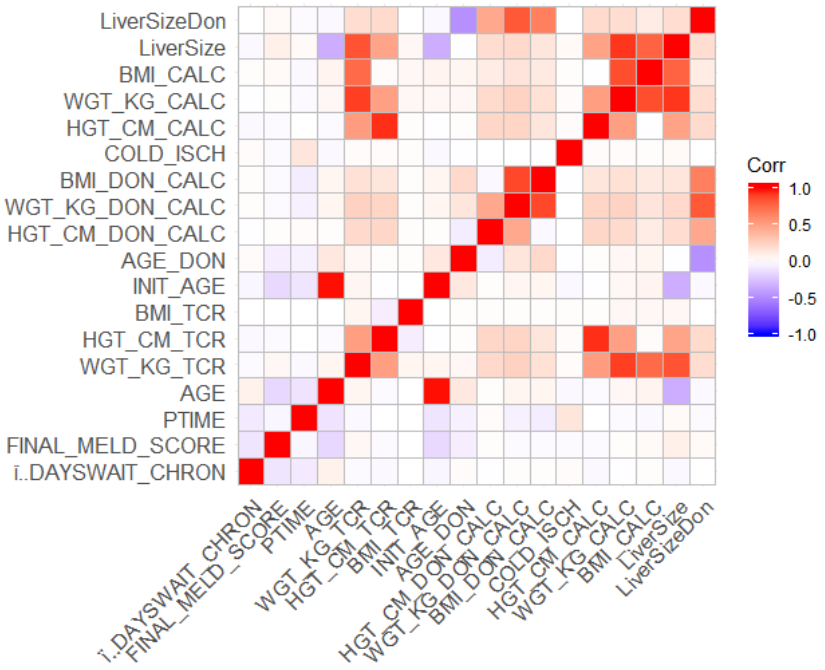
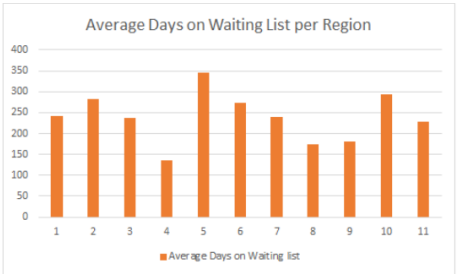


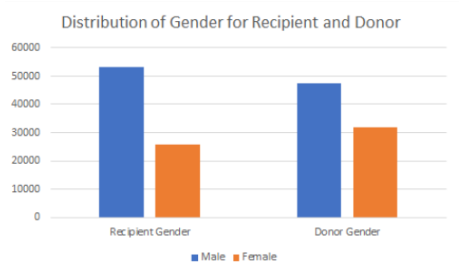
Figure 1: Correlation Heat map of normalized numerical variables

The correlation plot in Figure 1 shows most of the square boxes colored as while which shows that there is a very low correlation between the variables. There is a significant correlation between Liver Size and BMI for both recipient and donor. This could be due to the fact that both BMI and Liver Size take Body Weight as an input variable. We see a similar correlation happening with Liver Size and Age as Age is part of the liver size calculation. BMI and Weight have a high positive correlation as well, possibly due to the same reason.

As a part of our preliminary analysis, we also looked at the distribution of certain variables across the different transplant regions. We found that region 4 had the lowest waiting time, while regions 5,10, and 2 had the highest as seen in Figure 2 (a). We also found that there are more male recipients and male donors than there are female in Figure 2 (b). In addition, from our calculated liver size variables in Figure 3, we see that all regions have donors with larger liver sizes than recipients. The discrepancy in these variables indicates that it will be harder to match recipients and donors.



(a) Average number of Days on the Wait list per Region



(b) Number of Males versus Females for both Recipient and Donor

Figure 2

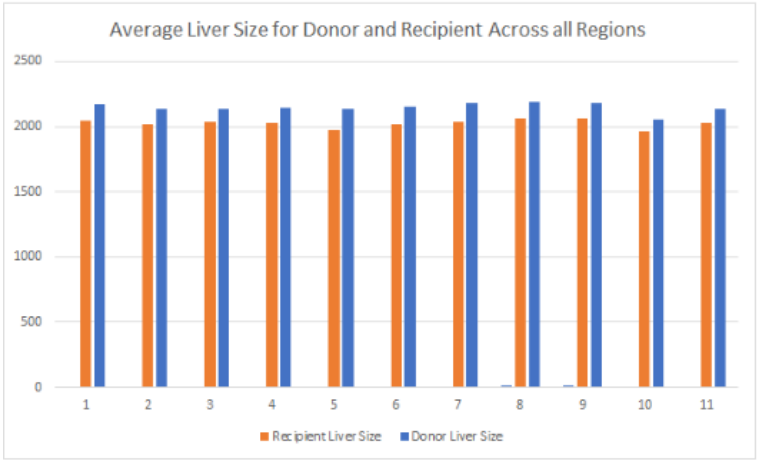


Figure 3: Average Liver Size per Region

We also looked at the distribution of our independent variables for our models. Figure 4 shows the patient status over different regions. We see that region 3 has the highest number of patients who have died after a transplant. Region 3 consists of Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, and Puerto Rico. When doing more variable exploration, we also found that region 3 had the highest rate of diabetes across the recipients. Diabetes is associated with a lot of other medical issues, and that combined with ESLD can cause more complications so we do not find it surprising that more patients in region 3 are not surviving after their transplants. Figure 5(b) shows the distribution of age across region 3, it shows that distribution is skewed more towards the older ages. We also looked at the survival time for less than 1 year (Figure 6 (a)), 1 to 3 years (Figure 6 (b)) , and 1 to 5 years (Figure 7) after a transplant. The distributions for under 1 year and 1-3 years shows that Region 9 has a low average survival time.

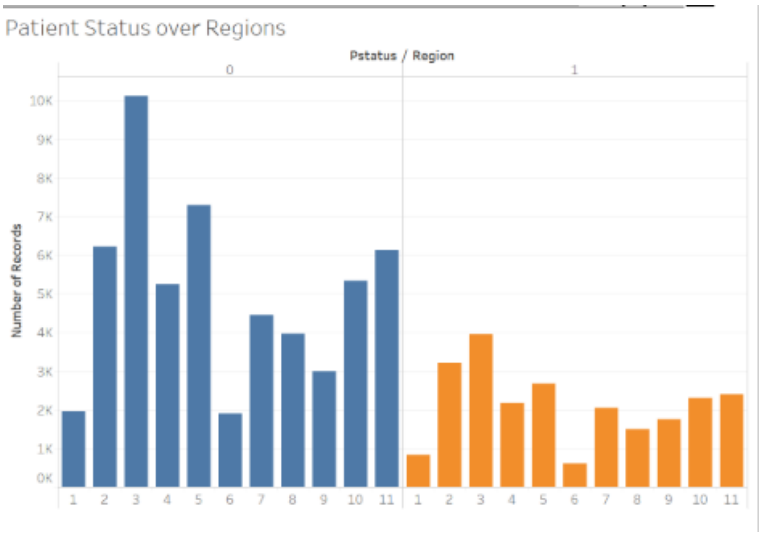
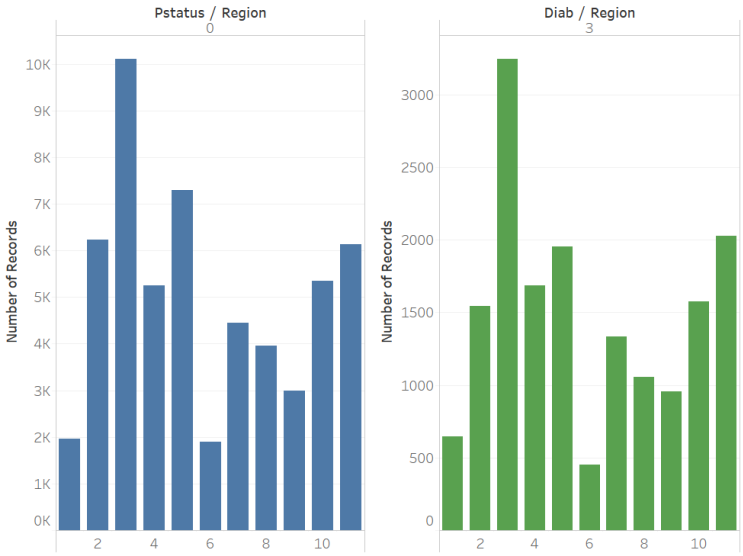
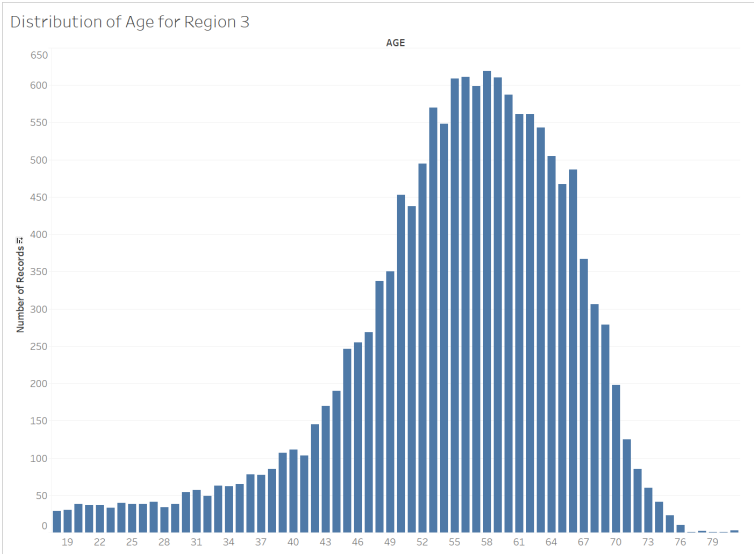


Figure 4: Patient Status (alive or dead) Over Every Region

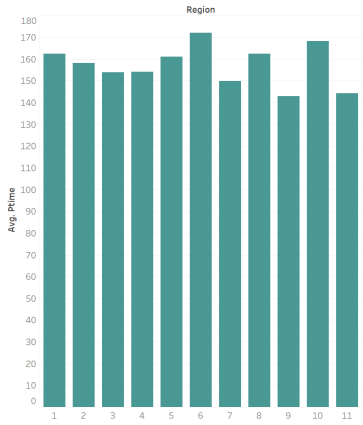


(a) Deceased and Type 2 Diabetes Patients over all Regions

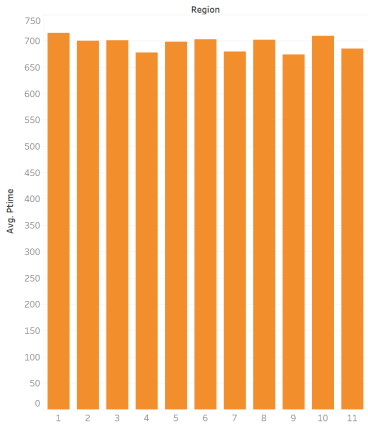


(b) Age Distribution for Region 3

Figure 5: Distribution of Deceased Patient, Type 2 Diabetes over 11 regions and Age over Region 3

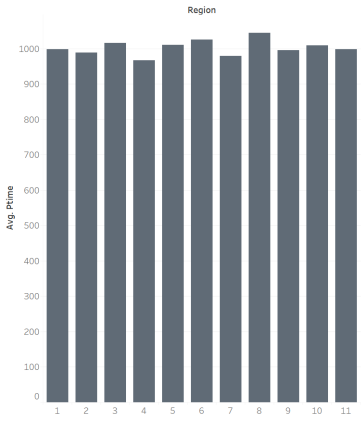


(a) Average Patient Survival Time Less than 1 year Over Every Region



(b) Average Patient Survival Time 1 to 3 years Over Every Region

Figure 6



(a) Average Patient Survival Time 1 to 5 years Over Every Region

Figure 7

5 Quantitative Models

Linear Regression

A linear model was generated to predict PTime for two different intervals. One for survival time between one to three years and another one for PTime between one to five years. We fit PTime over MELD Score, Region, Height (donor recipient), Alcohol Heavy (donor), Diabetes (donor), Liver size (donor and recipient), MALIG (Malignancy of recipient), Blood Type Match and transplant year for both intervals. We observed that the variable MALIG which describes if the recipient had a presence of malignant tumor or cancer had a huge impact on the predictability of both linear models.

Model Summary for 1-3 Years

The adjusted R square for our first linear model for one to three years was 0.56, F-Value of 402 and Residual Standard Error(RSE) of 0.6628. We later removed the diabetes, height and Blood Type variables to see its impact on our overall model as those weren't significant individually. The linear model was fit again, this time on fewer variables which resulted in a Adjusted R Square of 0.5603, an F-value of 541 and an RSE of 0.6628 shown in the Model Summary. An increase in the F-value shows an overall improvement in the prediction ability in the model. The p-value for the model is very low which shows the F-statistic is significant.

```
Residual standard error: 0.6628 on 13543 degrees of freedom
  (432 observations deleted due to missingness)
Multiple R-squared:  0.5613,    Adjusted R-squared:  0.5602
F-statistic: 541.4 on 32 and 13543 DF,  p-value: < 2.2e-16
```

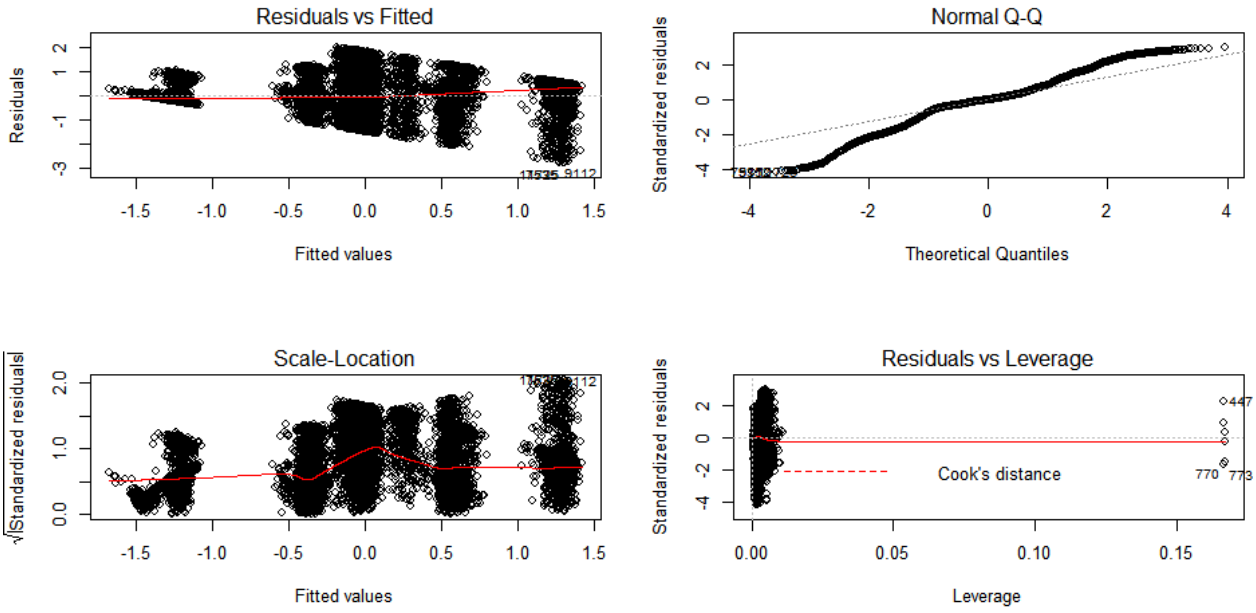


Figure 8: Model Summary

In the Model Summary Figure 8, the Residual vs Fitted plot shows a some-what a straight line and no distinctive pattern which means that there is no non-linear relationship between

the predictor variable and an outcome variable. The plot also shows that the residuals are normally distributed in The Normal Q-Q plot. Moreover, the Scale-location plot shows that the residuals randomly spread equally, which is a good sign. Finally, we see that there are no variables falling under Cook's distance in the Residual vs Leverage plot and the line is fairly straight which shows there are no leverage points, however, there is a presence of outliers. We removed those outliers but it had no impact on the model.

```
                      ME      RMSE       MAE      MPE     MAPE
Test set -9.647898e-16 0.6620231 0.4676241 37.36667 106.739
```
Train Data

```
                      ME      RMSE       MAE      MPE     MAPE
Test set -0.001916154 0.6756506 0.4768257 17.13365 105.935
```
Test Data

Figure 9: Prediction Accuracy

The predictive accuracy of this model is shown for both Training and Test Data. As expected, training data, had a higher accuracy as it has more values. The ME for test data is below 0 as well as the Mean Absolute Error is below 0.5 which is good sign of a high accuracy since these numbers are normalized.

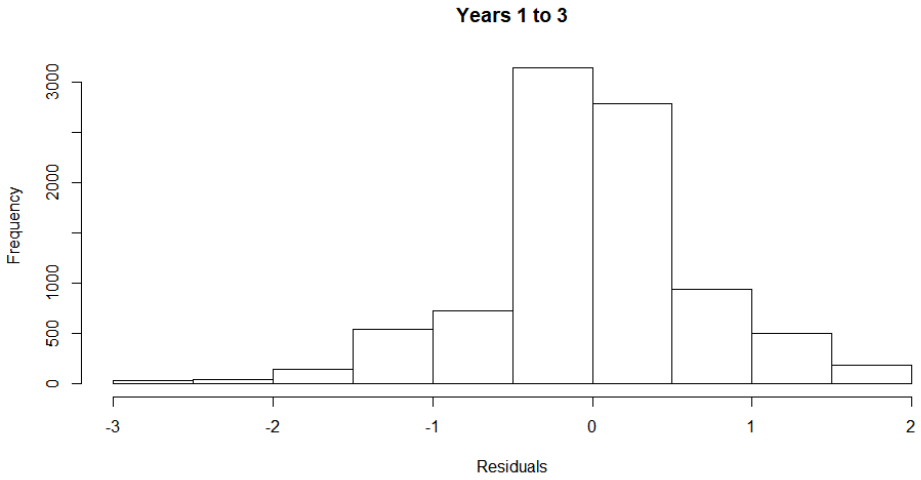


Figure 10: Validation Distribution

The Validation Distribution histogram shows that majority of the residuals i.e. the difference between the predicted values and the actual values are around zero which reflect a great accuracy of the model.

Model Summary for 1-5 Years

The adjusted R square for the second linear model for one to three years was 0.5782 and F-Value of 644.1, both of which are relatively higher than the first model. The RSE for this model was 0.6487. This shows an overall higher significance and predictability of the model. We later removed the diabetes, height and Blood Type variables to see its impact on our overall model as those weren't significant individually. The linear model was fit again, this time on fewer variables which resulted in a Adjusted R Square of 0.5778, an F-value of 818.1 and RSE of 0.649. A high F-value and low RSE refers to an overall significance of the model even though the Adjusted R square is not very high.

```
Residual standard error: 0.649 on 19668 degrees of freedom
  (665 observations deleted due to missingness)
Multiple R-squared:  0.5785,    Adjusted R-squared:  0.5778
F-statistic: 818.1 on 33 and 19668 DF,  p-value: < 2.2e-16
```

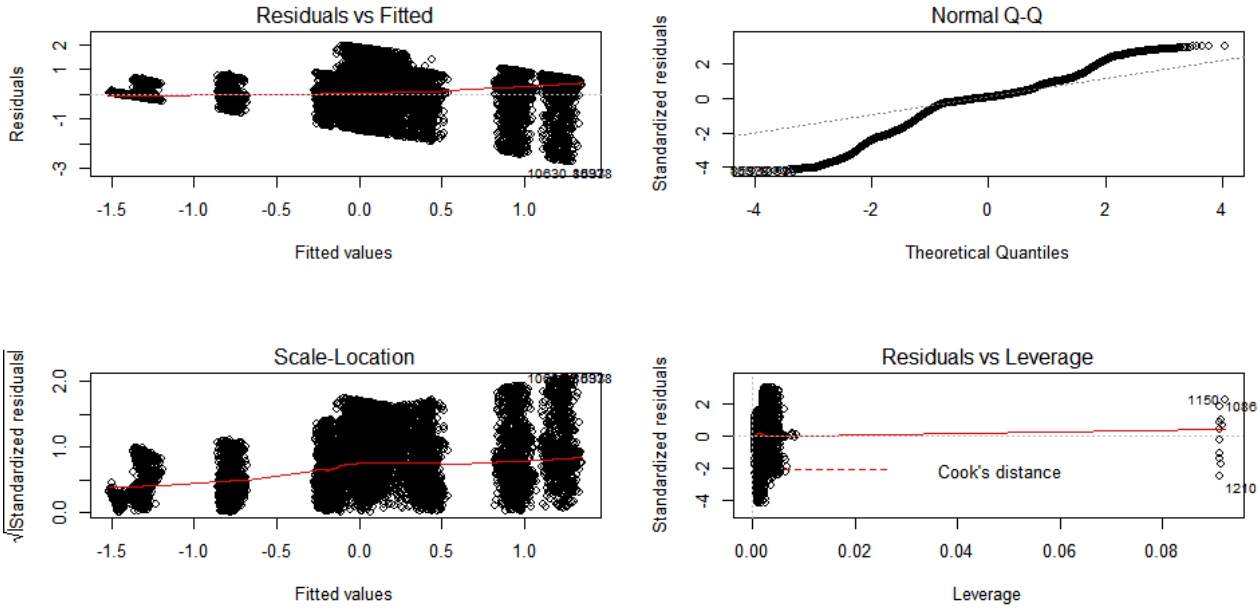


Figure 11: Model Summary

In the Model Summary figure, the Residual vs Fitted plot shows a some-what a straight line and no distinctive pattern which means that there is no non-linear relationship between the predictor variable and an outcome variable. The plot also shows that the residuals are normally distributed in The Normal Q-Q plot. In comparison to the 1-3 year model, the Normal Q-Q plot has worsened. Moreover, the Scale-location plot shows that the residuals randomly spread equally, similar to the first model. Finally, we see that there are no variables falling under Cook's distance in the Residual vs Leverage plot in this model as well and the line is fairly straight which shows there are no leverage points. However, there is a presence of outliers in this model as well. We removed those outliers, but it had no impact on the model as well seen in Figure 12.

The model has low mean error of lower than 0 for Train and Test data. Again, a prediction on Train data will have a low ME as it is predicted on a larger data set. Moreover, it is significant MAE of 0.43.

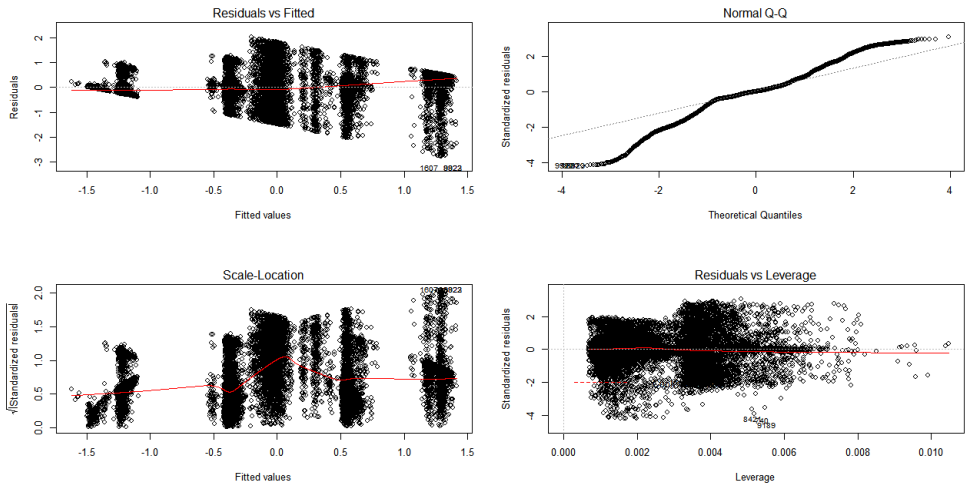


Figure 12: Model Summary Without Outliers

	ME	RMSE	MAE	MPE	MAPE
Test set	-3.318671e-14	0.6484449	0.4353351	171.3283	344.3526

Train Data

	ME	RMSE	MAE	MPE	MAPE
Test set	-0.002014685	0.6415864	0.4300143	79.52421	185.5455

Test Data

Figure 13: Prediction Accuracy

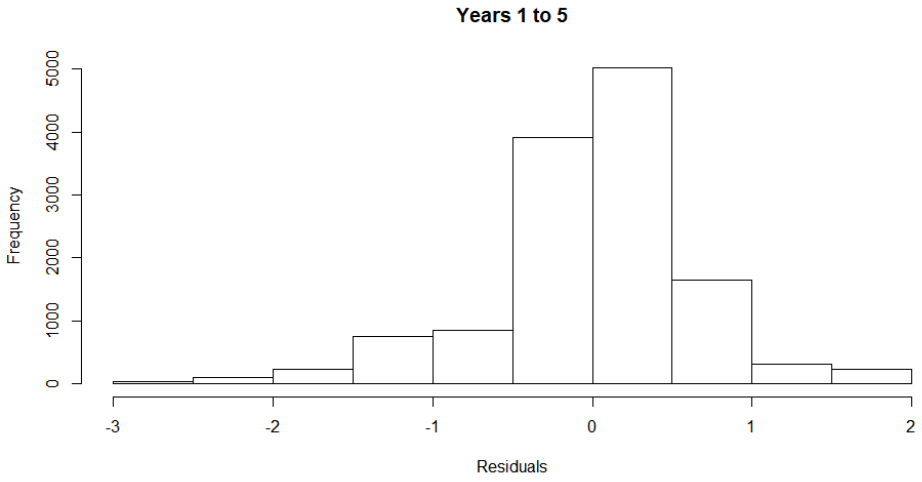
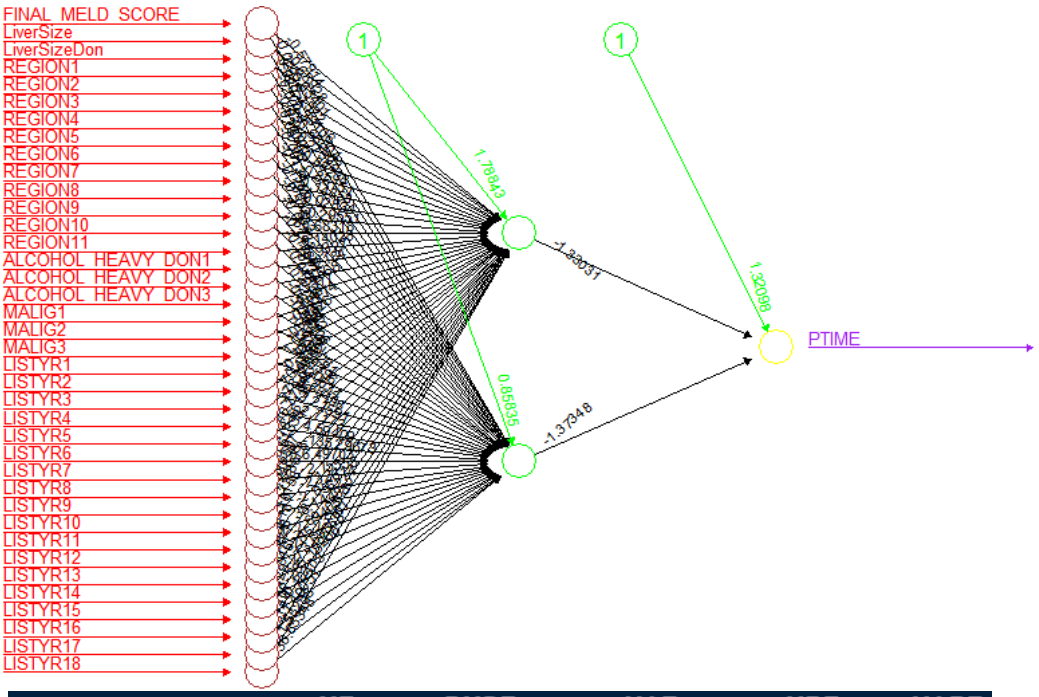


Figure 14: Validation Distribution

The Validation Distribution for this model shows a high frequency for residuals which are 0. This is again, a sign of good predictability and model accuracy.

Neural Network

Model Summary for 1-3 Years



	ME	RMSE	MAE	MPE	MAPE
Test set	-0.9809707	1.401026	1.174287	112.5939	356.117

Figure 15: Neural Network with 2 hidden layers

The Neural Network (NN) for the linear model was generated with 2 hidden layers to predict Patient survival time. The same set of variables were use as in the regression model to compare the performance. The NN has a very low mean error of -0.98 and a high RMSE (Root Mean Squared Error). RMSE is a measure of how spread out these residuals are. MAE is also high given the variables are normalized. This shows that model under predicts.

Model Summary for 1-5 Years

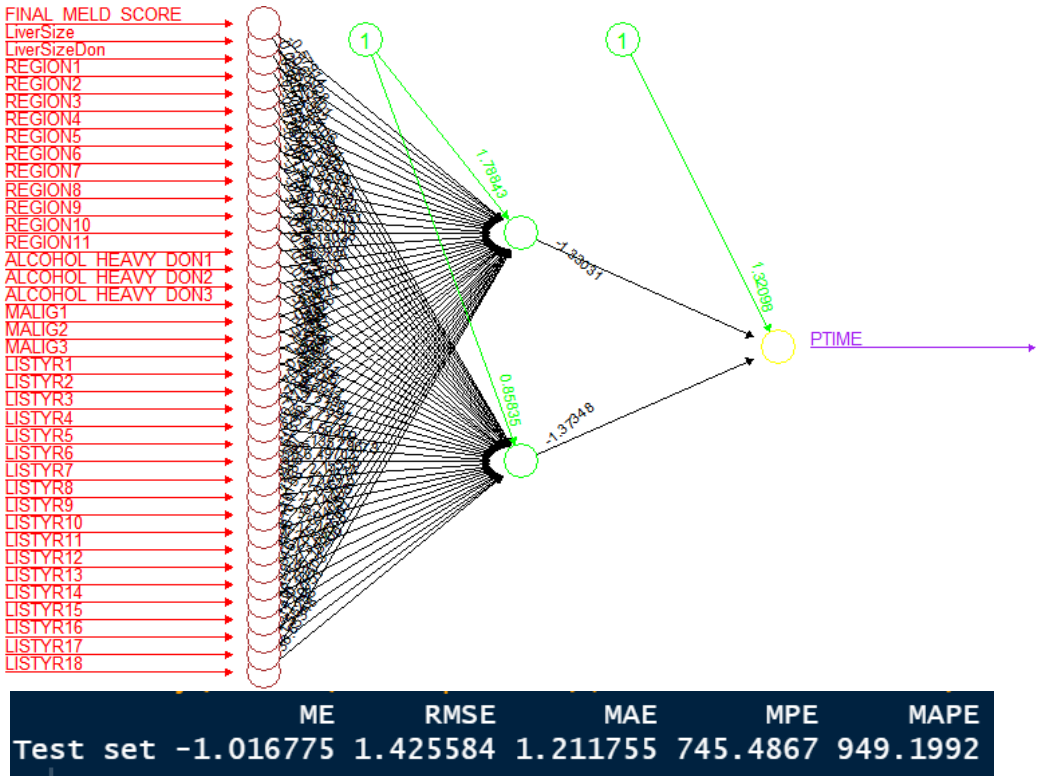


Figure 16: Neural Network with 2 hidden layers

The Neural Network for one to five years interval was also generated with 2 hidden layers. The predictive accuracy is quite similar to that of one to three years NN model. Mean error is -1.016 and RMSE is close to 1.5 which means the model under predicts.

Results and Conclusions from Linear Models

Overall, the linear models are able to predict with a decent accuracy. However, in a comparison of the two, linear regression stood our as the best model to predict patient survival time. Both models in linear regression have a significantly lower error and are able to predict more accurately. We also observed that the variable blood type had no impact on either of the models showing that it doesn't matter if the recipient's blood type matches with their donor.

6 Qualitative Models

Logistic

For our logistic models, we wanted to predict the patients status 3 years and 5 years after their transplant. Since we are predicting the patient status, we wanted to include variables that pertained to both the patient and the donor. Specifically, we wanted variables that indicated the patient and donors size, as well as the liver quality. Before we had our liver size and gender match variable, we included variables that were eventually used in calculating these variables. These include things like weight, age, and gender. The final variables we decided to include are MELD score, diabetes in the donor, diabetes in the recipient, the recipients medical condition at the time of transfer, region, donor height, recipient height,

recipient liver size, donor liver size, and gender match (if the recipient and donor have the same gender). For our model looking at the one to three years survival time, the logistic model had a classification rate of 71.52%. Our model for looking at the survival of one to five years after transplant is 71.63%.

LDA and QDA

Our LDA and QDA models both used the same variables that we used in our logistic model. Our LDA model predicted just about the same as the logistic model, with a 71.56% for one to three years survival. Our QDA model predicted worse than all models so far, with a 70.80% classification rate for one to three years survival. For one to five years after transplant, our LDA classification rate was 71.59% and our QDA classification rate was 71.56%.

Neural Network

For our neural networks, we ran the data through two different structures. First, we ran a network with 1 hidden layer and 2 hidden nodes, which gave an accuracy of 71.40% for one to three years. We then ran a model with 1 hidden layer and 3 nodes, which gave an accuracy of 71.11%. For five years survival, the network with 1 hidden layer and 2 nodes gave an accuracy of 71.21%. The network with 1 hidden layer and 3 nodes gave an accuracy of 71.46%.

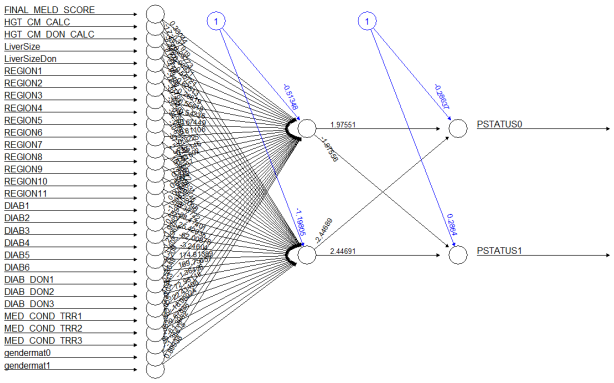


Figure 17: 3 year survival network with 1 layer and 2 nodes

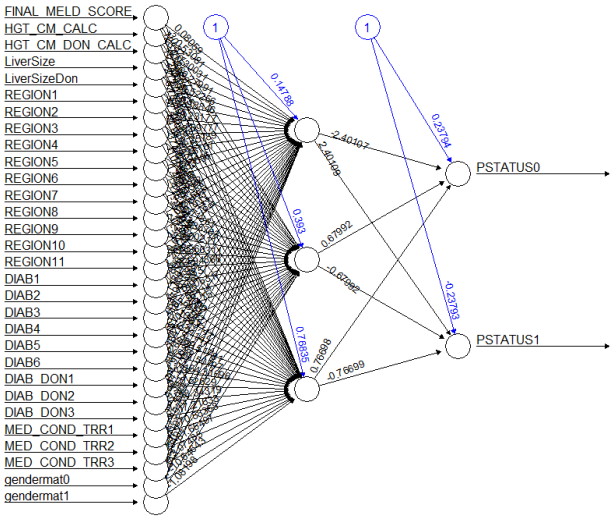


Figure 18: 3 year survival network with 1 layer and 3 nodes

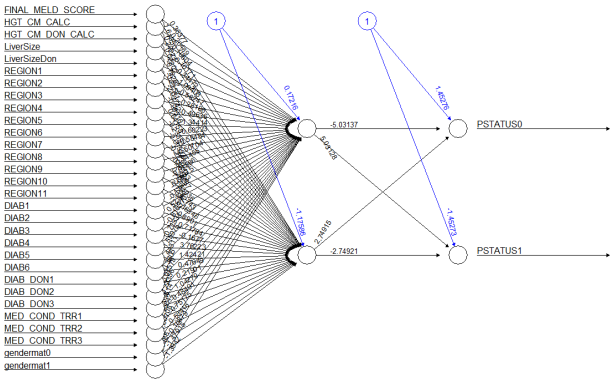


Figure 19: 5 year survival network with 1 layer and 3 nodes

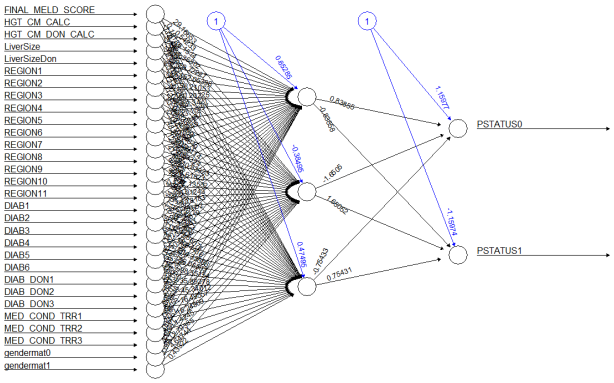


Figure 20: 5 year survival network with 1 layer and 3 nodes

Naive Bayes Model

For our Naive Bayes Model, we utilized the variable from our logistic model and performed analysis on patient survival from year one to three(lower bound) and year one to five(upper bound). This model aims to classify the probability of whether a patient would be alive or dead based on the variables used in our logistic model. Classification for our lower bound provided us with the accuracy of 71.22% in our training set. For our validation set, we received an accuracy of 71.48%. Taking a look at our upper bound classification, we received an accuracy of 71.96% for out training. Our validation set provided our with the accuracy of 72.65%. Please see the confusion matrices and statistics for our validation models(upper and lower bound).

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6443 2549
         1  114  231

               Accuracy : 0.7148
                 95% CI : (0.7055, 0.7239)
    No Information Rate : 0.7023
    P-Value [Acc > NIR] : 0.004073

                  Kappa : 0.0879

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.98261
            Specificity : 0.08309
         Pos Pred Value : 0.71653
         Neg Pred Value : 0.66957
             Prevalence : 0.70226
         Detection Rate : 0.69005
   Detection Prevalence : 0.96305
      Balanced Accuracy : 0.53285

       'Positive' Class : 0
```

Figure 21: Confusion Matrix for Validation Model(Lower)

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 9517 3506
         1  207  345

               Accuracy : 0.7265
                 95% CI : (0.7189, 0.734)
    No Information Rate : 0.7163
    P-Value [Acc > NIR] : 0.004318

                  Kappa : 0.0921

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.97871
            Specificity : 0.08959
         Pos Pred Value : 0.73078
         Neg Pred Value : 0.62500
             Prevalence : 0.71632
         Detection Rate : 0.70107
   Detection Prevalence : 0.95934
      Balanced Accuracy : 0.53415

       'Positive' Class : 0
```

Figure 22: Confusion Matrix for Validation Model(Upper)

Random Forest Model

For our Random Forest Model, we applied the same concept as our Naive Bayes Model in terms of utilizing variables from our logistic model and performing classification for lower and upper bound patient survival. For our lower bound classification, the model consisted of 500 trees and the number of variables for splitting at each tree node was set at 3(mtry). We received an estimated out-of-bag(OOB) error of 29.05%. There was a drastic difference in our class error when looking at patients that were classified as alive or dead. Class error for patients that were dead was 4.76% compared to patients that were alive at 85.22%. The chart below represents the class errors along with the estimated OOB error rate. We can see that our error predominately stabilizes at 200 trees and stays stable throughout. Our model was also able to determine variables of importance by providing us the mean decrease accuracy along with the mean decrease gini(please see illustration for Figure 15 below). We received a 70.97% accuracy for our training model and a 71.22% accuracy for our validation.

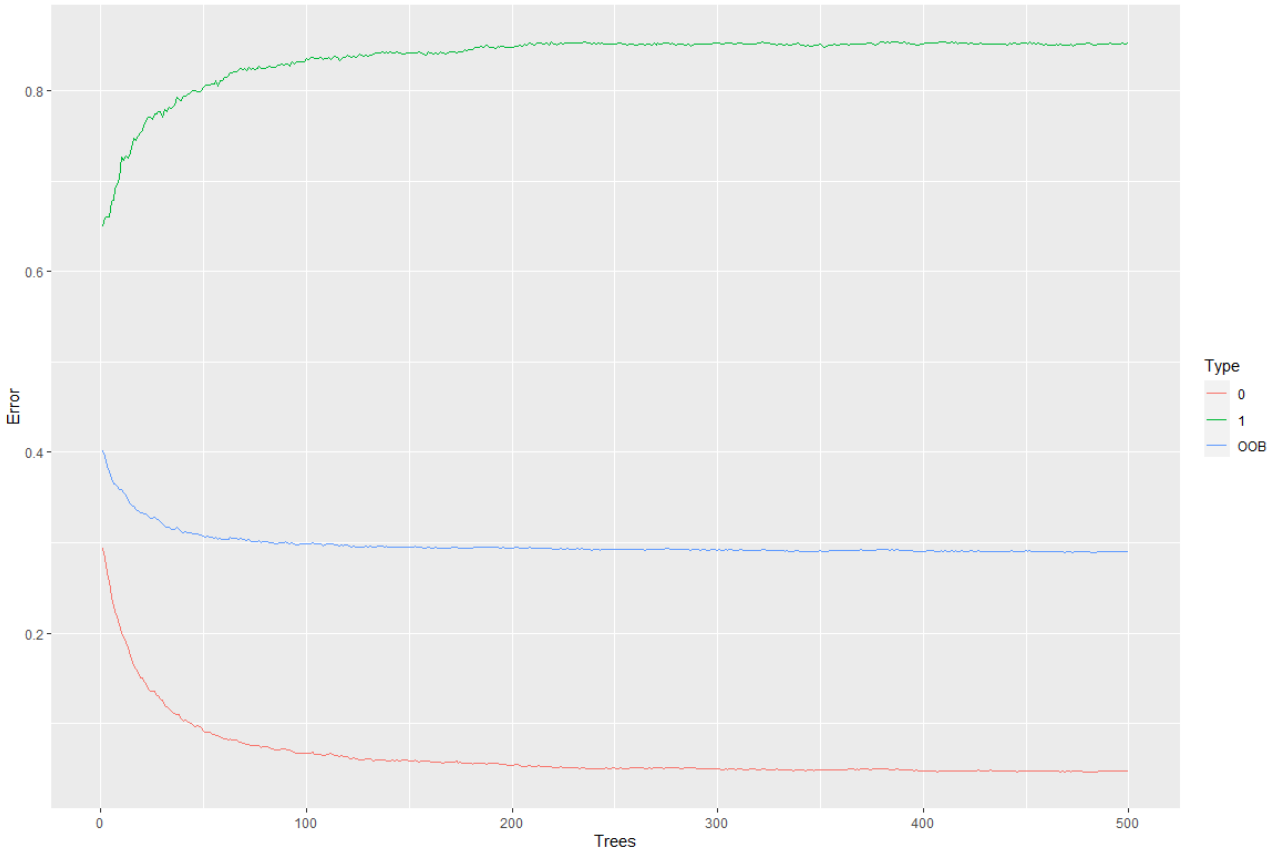


Figure 23: Class and Estimated Out-Of-Bag Error Rate

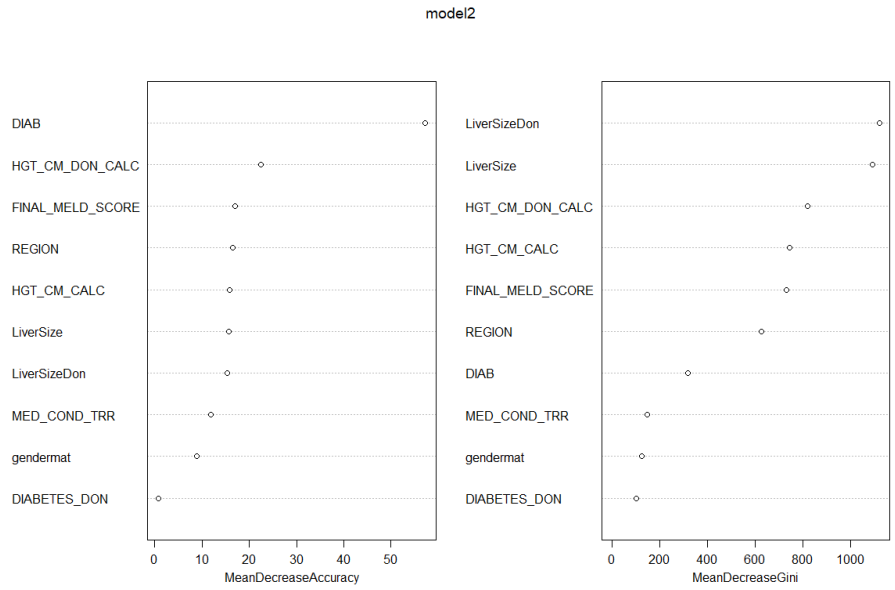


Figure 24: Mean Decrease Accuracy/ Gini

Moving forward, our upper bound classification consisted of 100 trees and the number of variables for splitting at each tree node was set at 7(mtry). The result from the graphs above were close to identical in relation to the upper bound classification. We received an estimated OOB error rate of 28.87%. Our class error for patients dead was 7.32% compared to patients that are alive which was 81.71%. The chart below represents the class errors along with the estimated OOB error rate. We received a 71.63% accuracy for our training model and a 72.54% accuracy for our validation.

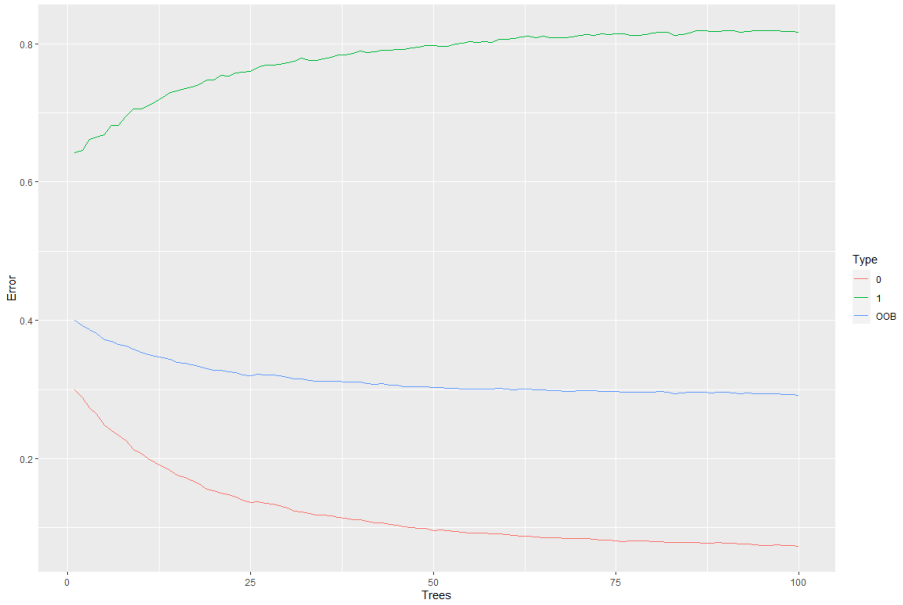


Figure 25: Class and Estimated Out-Of-Bag Error Rate

Results and Conclusions from Logistic Models

Overall, our classification models gave us a classification percentage that we feel confident in. Our best logistic model to predict patient status for three year survival time was our LDA model. This models had a 71.56% classification rate.The best model we found for predicting five year survival time was our Naive Bayes with a 72.65% classification rate. However, we would like to note that all of our models were around the same range with no particular model showing a significant improvement in prediction. We also ran our models twice in the beginning of our analysis to see what impact blood type would make on our outcome. When it comes to blood type, we had three variables in our data set pertaining to this information: the blood type of the recipient, the blood type of the donor, and a variable that indicated whether the blood type of the donor and recipient matched. In our research, we found that blood type does not have to be an exact match for liver donors, but rather the blood types have to be compatible. To see if blood type is significant, we decided to use blood type match since it takes into account both the donor and the recipients blood type. When running a logistic model with blood type match we found that the variable was not significant in the results of the model and did not make an impact on the overall results. In fact, in all cases where we included it, the models accuracy decreased by a small percentage. Due to this result, we made a decision to exclude blood type match moving forward in our analysis. This decision was done before we finalized our filtering for survival time, so you will only see models run with blood type match for three years survival time. We were also interested in looking at the regions significance in prediction. In our model, we found that all regions except region 9 were not significant in our predictions. This does not surprise us since region 9 is known for having longer wait times, and longer wait times can inevitably lead to a worsening condition while waiting for a liver.

References

[1] SRTR: Scientific Registry of Transplant Recipients,
https://www.srtr.org/requesting-srtr-data/data-requests/

[2] Alexander Choukèr,Andre Martignoni, Martin Dugas, Wolfgang Eisenmenger, Rolf Schauer, Ines Kaufmann, Gutav Schelling, Florian Löhe, Karl-Walter Jauch, Klaus Peter, and Manfred Thiel. Estimation of Liver Size for Liver Transplantation: The Impact of Age and Gender, 2002.
https://aasldpubs.onlinelibrary.wiley.com/doi/pdf/10.1002/lt.20113