Group Name: Datacians
Group Members: Oyindamola Obisesan, Micheal Olubode, Judah Odiachi

Topic: Hospital Readmission Analysis

- **Model Insights**

#the intercept indicate the log odds of the whole population of interest to be readmitted #with no predictor variables in the model. The model is a multivariate logistic model.

1. Race: From the p-value, this is less than 0.05, which is statistically important and useful in the model prediction. The result from the analysis of the logistic regression, show the importance of some of the races to the being readmitted or not. If the patient is Asian, the odd ratio is 0.7946 which says that there is a 20% decrease in the odd of being readmitted as an Asian. If the patient is Caucasian, the odd ratio is 1.073 which shows that the race of a Caucasian does not really matter as there is a 7% of being readmitted as a Caucasian. If the patient is Hispanic, there is a 12% decrease in the odd of the patient being readmitted. Insurance companies can increase the rate for races with higher readmittance rate. The hospital administration can find the root causes for readmission rate in the races with high readmission. Are they being cared for less? Or maybe they need more attention and tests run on them.
2. Age: From the p-value, this is less than 0.05, which is statistically important and useful in the model prediction. The result from the analysis of the logistic regression, show the importance of some of the age groups to the being readmitted or not. If the patient is between the age of 10-20, the odd ratio is 2.0184, which means there is a 100% increase in odd of the patient being readmitted and if they are 20 -30, there is an odd ratio of 1.89, which means there is a 89% increase in odds for patients within that age group. For patients between 30-40, the odd ratio is 1.86, which means there is an 86% increase in the odds for patients in that age group. The odds ratio increases for patients in the age group 40-50, there is 100% increase in the odd for this age group. This increase continues for 50-60, 60-70, 70-80. The highest increase in the odd of being readmitted is in 70-80 age group; 142% which means that they are very highly likely to be readmitted to the hospital. The age group [80-90] and [90-100] is 132% and 65% increase in odd. This can be important to insurance companies, the insurance for the patients in the 70-80 age group can be the highest as they are more likely to be readmitted. While those for the 90-100 and 20-30, 30-40 can be made to be lower. Hospitals staff can also pay more attention to this people in the category with high readmission rate to ensure they are properly heal.
3. Inpatients: From the p-value (1.72*10^-8), this is less than 0.05, which is statistically important and useful in the model prediction. The odd ratio is 1.49, the odd increases by 49% for every increase in the number of times the patient was admitted. The shows that patients that have been previously admitted have a higher probability of being readmitted in the coming year. Insurance companies can include the number of times admitted yearly in their rate cost analysis and give a higher rate to patients who have a higher number of previous admissions. Doctors should pay more attention to patients who have had

recurring admission and get to the root cause of the readmissions. This can help lower the rate of readmissions significantly.

4. Number of Emergency: From the p-value (<2*10^-16), this is less than 0.05, which is statistically important and useful in the model prediction. This has an odd ratio of 1.25% which shows that the odd of readmissions increases by 25% for each increase in the number of emergency visits. The shows that patients that have previous emergency visits have a higher probability of being readmitted in the coming year. Insurance companies can include the number of times emergency visits yearly in their rate cost analysis and give a higher rate to patients who have a higher number of emergency visits. Doctors should pay more attention to patients who have had emergency visits and get to the root cause of the emergency visit. This can help lower the rate of readmissions significantly.

5. Number of Procedures: From the p-value (4*10^-13), this is less than 0.05, which is statistically important and useful in the model prediction. This has an odd ratio of 0.95% which shows that the odd of readmissions reduces by 5% for each increase in the number of procedures. The hospitals can increase the number of procedures done on patients to ensure the patients is totally cured of the ailment and doesn't come back to the hospital. This reduces the rate of readmissions. The patients should do all the procedures recommended by the doctor to avoid a readmission to the hospital.

(a) [ii] **Summary of Model Performance**

Logistic Regression

```
fit_hosp <- glm(data=hospdata, readmitted~race+gender+age
        +number_outpatient+number_diagnoses+num_medications
        +discharge_disposition+number_inpatient+number_emergency+
       num_lab_procedures+num_lab_procedures+num_procedures,
      family="binomial")
```

Logistic Regression with Lasso and cross validation

```
fit_hosp1 <- cv.glmnet(x_train, y, family="binomial", alpha = 1, lambda = lambdas,
            nfolds = 5)
```

Decision Trees

```
fit_hosp2<-rpart(data=hospdata, readmitted~race+gender+age+admission_source
         +number_outpatient+number_diagnoses+num_medications
         +discharge_disposition+number_inpatient+number_emergency
         +num_lab_procedures+
          num_lab_procedures+num_procedures,
       parms=list(split="information"),
       control=rpart.control(xval=20))
```

SVM

```
classifier = svm(formula = readmitted ~race+gender+age+admission_source
         +number_outpatient+number_diagnoses+num_medications
         +discharge_disposition+number_inpatient+number_emergency
```

```
            +num_lab_procedures+
              num_lab_procedures+num_procedures,
            data = hospdata,
            type = 'C-classification',
            kernel = 'radial')
```

Random Forest

```
ran_class = randomForest(readmitted ~ race+gender+age+admission_source
                +number_outpatient+number_diagnoses+num_medications
                +discharge_disposition+number_inpatient+number_emergency
                +num_lab_procedures+
                  num_lab_procedures+num_procedures, data = hospdata,
                ntree=1200, mtry = 3, importance=TRUE)
```

LDA

```
lda_class = lda(readmitted ~race+gender+age+admission_source
            +number_outpatient+number_diagnoses+num_medications
            +discharge_disposition+number_inpatient+number_emergency
            +num_lab_procedures+
              num_lab_procedures+num_procedures, data = hospdata)
```

Model	Method	Package	Hyperparameter	Selection	Accuracy	Kappa
Logistic Regression	glm	glmnet	NA	NA	0.6191	0.2215
LDA	lda	LDA	NA	NA	0.6149	0.2128
Random Forest	randomforest	randomForest	ntrees	1200	0.6261	0.2437
			mtry	3		
Decision Tree	rpart	rpart	cp	0.01	0.6143	0.213
			cross-val	50 fold		
SVM	C-classification		c	5	0.6351	0.2588
			kernel	radial		

(a) [iv] PERFORMANCE EVALUATION TECHNIQUES

Selected Model – Random Forest

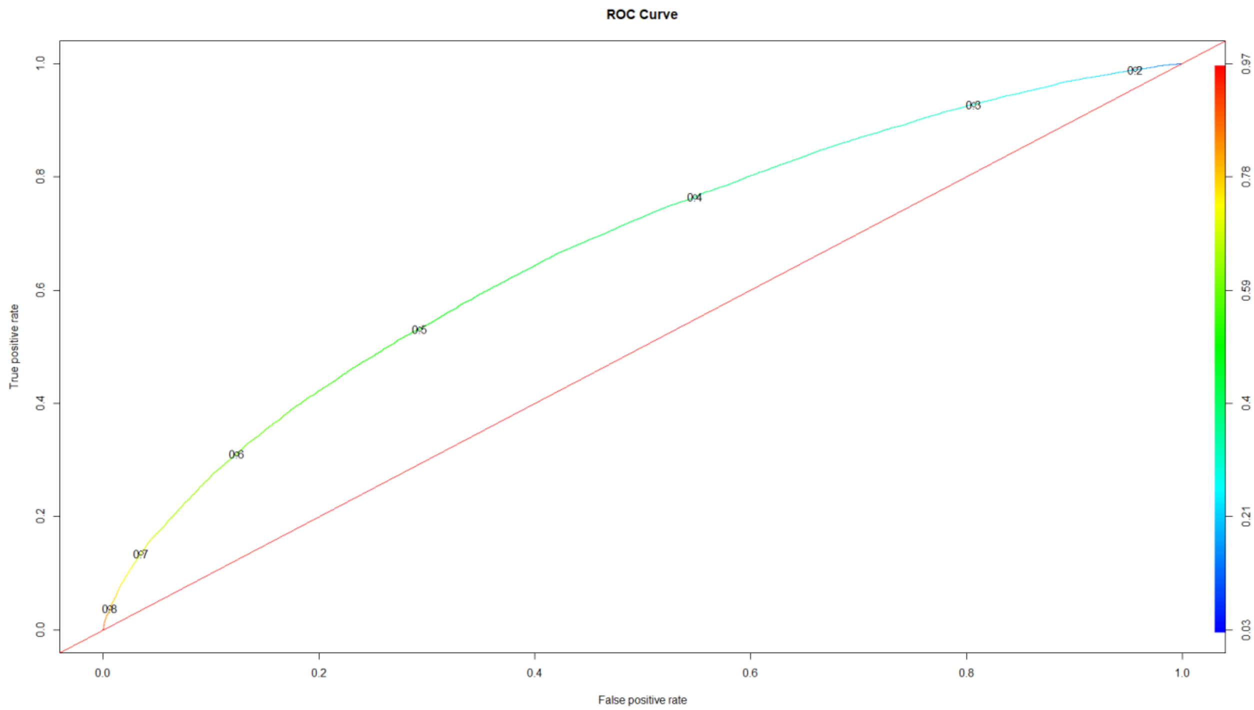
1. CONFUSION MATRIX, ACCURACY, KAPPA

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21651  8981
         1 12753 14470

               Accuracy : 0.6243
                 95% CI : (0.6204, 0.6283)
    No Information Rate : 0.5947
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2402

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

            Sensitivity : 0.6170
            Specificity : 0.6293
         Pos Pred Value : 0.5315
         Neg Pred Value : 0.7068
             Prevalence : 0.4053
         Detection Rate : 0.2501
   Detection Prevalence : 0.4705
      Balanced Accuracy : 0.6232

       'Positive' Class : 1
```

Only Three (3) evaluation metrics will be considered from the result of the confusion matrix

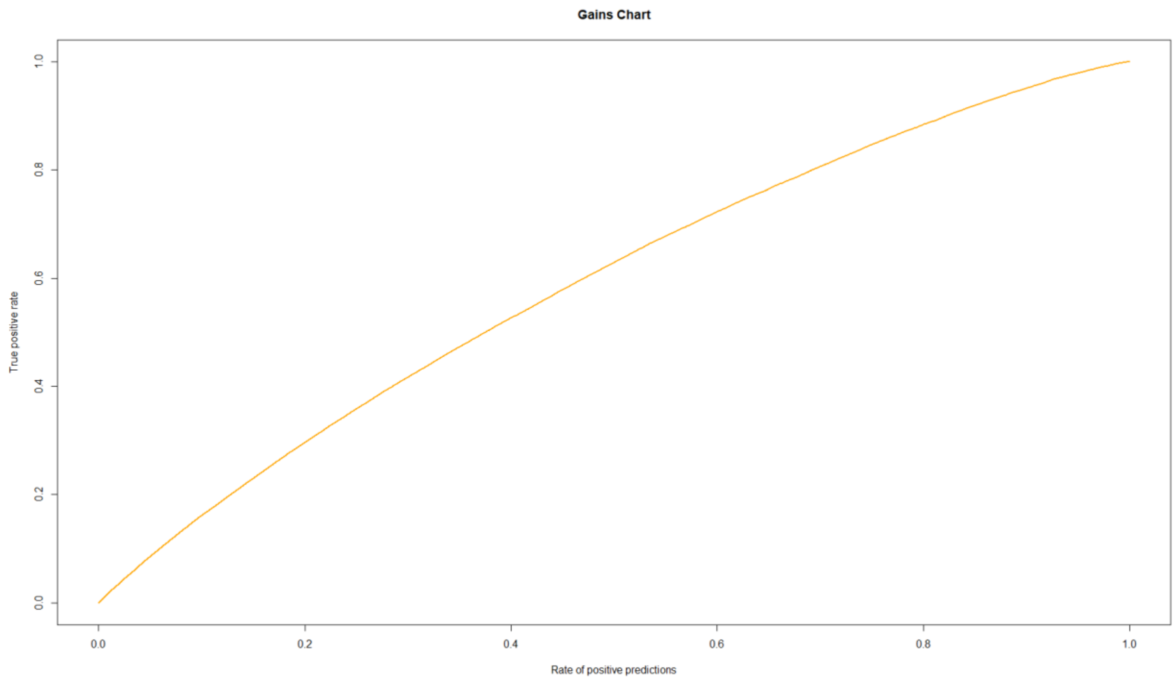
(a) The model has a fair sensitivity value of 61.7%. Sensitivity is the True positive rate, i.e the proportion of the actual readmitted patients that the model correctly identified. This value is not very impressive as about 38% of actual readmitted patients would be declared safe from the risk of readmission and the cost could be as high as the death of a patient.
(b) The accuracy of the model is 62.2%. This is the proportion of true positive and true negative in the entire dataset.
(c) The Kappa Statistic is 24% which is a fair performance as the value of 1 is a perfect agreement between the model and the actual label and 0 indicates that the model is just by the same as a random prediction

4. **ROC CURVE**

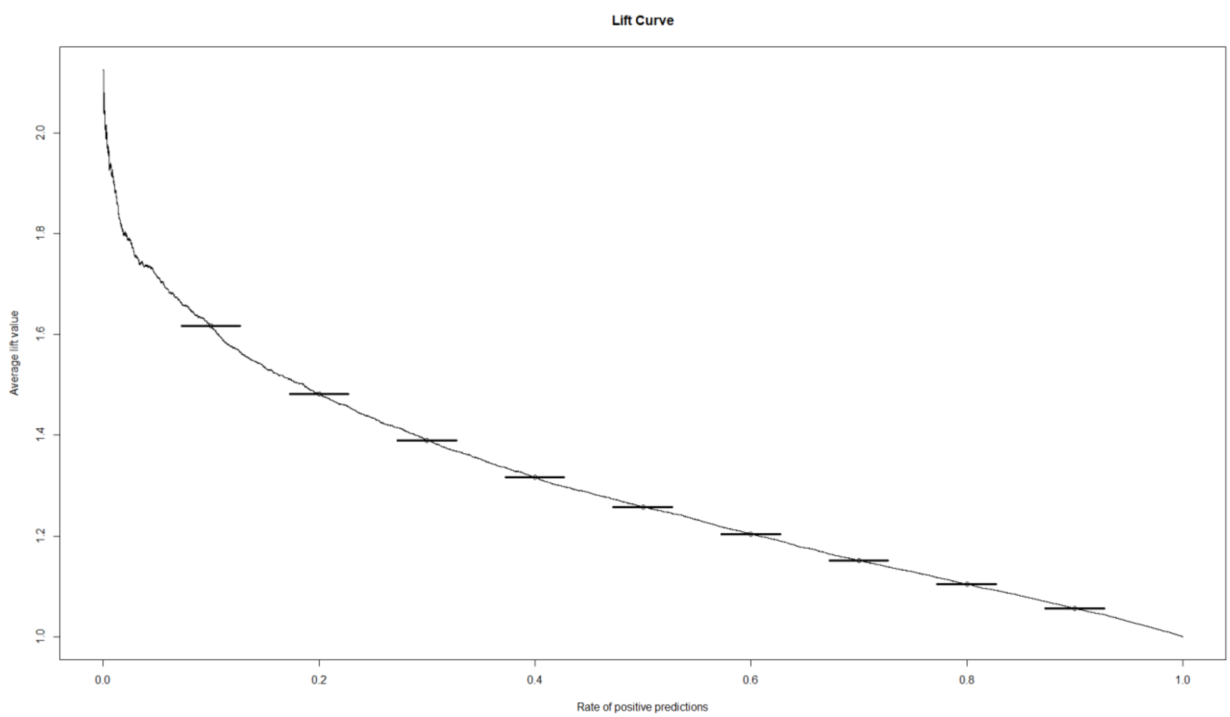


The ROC curve is a plot of sensitivity against the False positive rate. The diagonal line is the line of random chance and the farther away from it the curve, the better the model. The curve is not very far from the random chance line nevertheless, it is a fair prediction. We also see that cut-off value of 0.5 is the closest to the edge of the plot and that will give an optimum separation between the two classes/labels.

5. **Gains Chart**

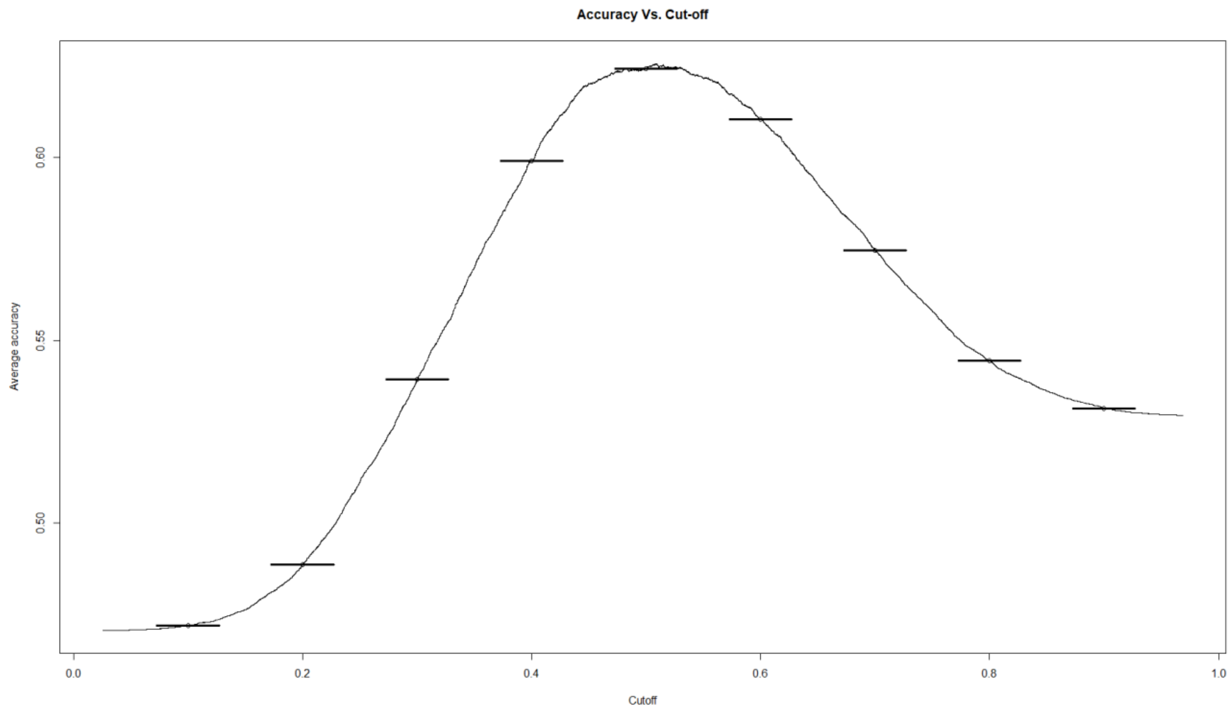


6. **Lift Curve**

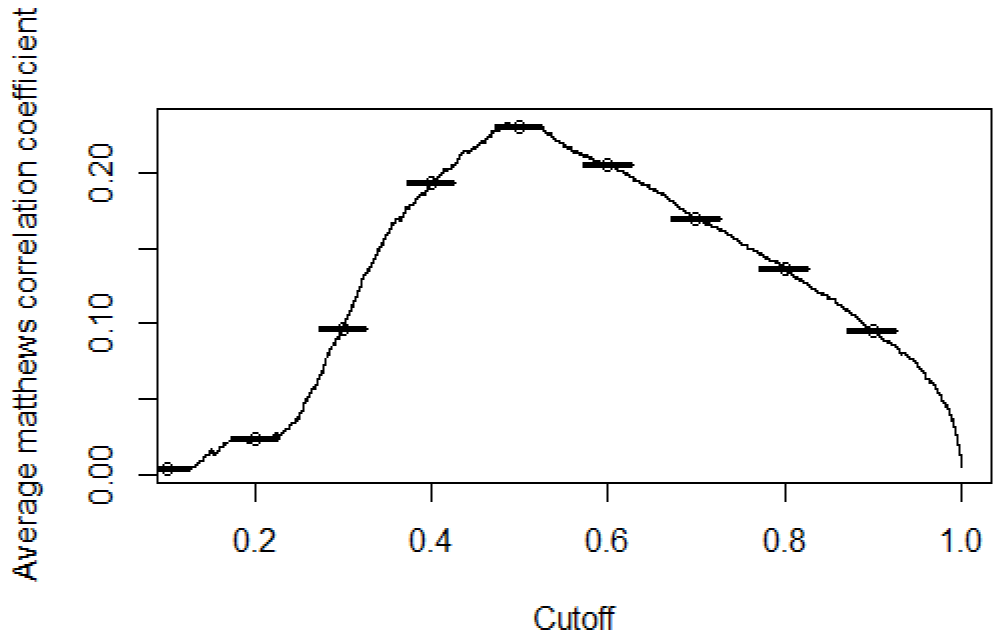


From the Lift chart, we see that the top 4% (approx.) probability values (when arranged in descending order) account for more than two (2) times of the average case prediction of the positive values. In another words, the positive case value is 1, and the output of the model is probability values; when these probability values are arranged in descending order, the top 4% account for more than twice the average value of the total positive case prediction.

7. **Accuracy Vs. Cut-off**

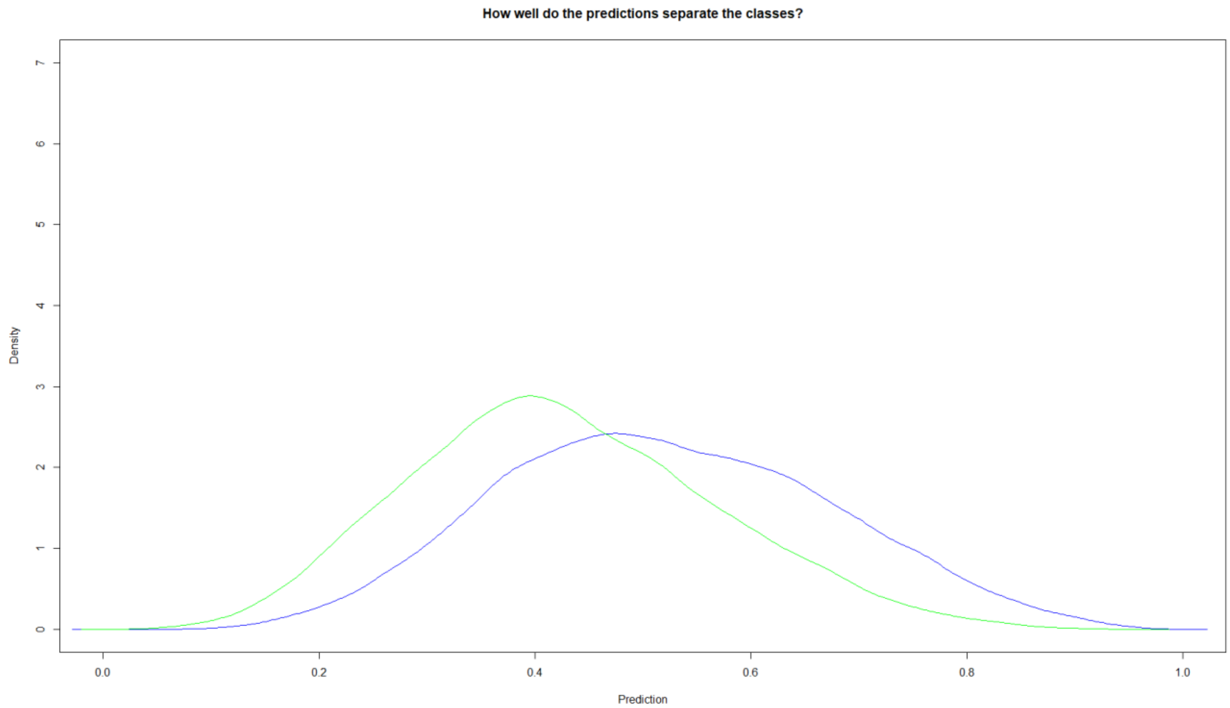


From the Accuracy vs. cut-off chart, the highest accuracy is at the cut-off value of 0.5. So we use a cut-off value of 0.5 for separating the classes.



From Matthews correlations, the highest correlation is at the cut-off value of 0.5. So we use a cut-off value of 0.5 for separating the classes.

9. **Distribution of Predicted Classes**



The above is the distribution of the predicted classes. The more separated they are, the better the accuracy of the model. The above shows a wide overlap between the two classes which indicates a lot of false negatives and false positives.