**Final Project Report**

**Contents**

Topics--------------------------------------------------------------------------------------------------Page No.

1. Executive Summary 2
2. Exploratory Data Analysis 2
3. Data Partitioning 3
4. Modelling Insights 4
5. Summary 8
6. Appendix 1 9
7. Appendix 2 16

**Executive Summary**

This project is aimed at addressing the resource management issues of a health care system of 5 hospitals. By means of this project, we want to correctly understand whether a patient will return to the hospital or not within 30 days of discharge from the hospital and utilise this information to improve the quality of services rendered to the patients and lower the overall costs endured by the hospitals.

The primary goal of this analysis is to get the highest accuracy possible for predicting the return of a patient to the hospital within 30 days of discharge. We have considered accuracy as the most important parameter here as wrongly predicting the return of a patient would result in poor service to the patient or increasing the overhead expenses of the hospital . For instance, if we predict that a patient will return, however, in reality the patient does not then there would be excess resources like medicines etc. allocated resulting in loss. On the other hand, if we predict that a patient will not return, however, the patient returns within 30 days, then there will be a sense of customer dissatisfaction due to lack of resources and non preparedness.

For this purpose we got the dataset from the emergency department of the hospital. The dataset consists of 38221 entries and 27 variables. A number of models like Logistic, LDA, LASSO, Ridge, Trees, KNN and ensemble methods like Bagging, Gradient Boosting, Random Forest and XGBoost have been run in order to identify the model which gives the best accuracy. Based on the results we got we can clearly say that XGBoost model gave us highest accuracy of 78.70 percent on the unseen data set which is 3 percent more than having no model at all (random guessing). This means our model has a good predicting power and can be put into implementation. Also, based on our results we can clearly say that variables such as *ED\_RESULT(The disposition of the patient from the emergency department, indicating what happened to the patient after leaving the emergency room), Charges (paid by the patient),financial class* are among the most important variables that play a significant role in predicting whether the patient would return back to hospital or not, therefore we recommend the hospitals must focus on these factors and must take extra caution when filing data related to these fields as any mismatch or wrong entries may give misleading results and might affect the accuracy of our predictions. These models and their outcomes are discussed in greater depth in the report .

Managing resources plays a significant role in hospital operations. Effective use of resources allows hospitals to efficiently provide high-quality care to patients. Unlike many other industries, health care decisions deal with highly sensitive information, require timely information and action, and sometimes have life or death consequences. Proper analysis of the above data would enable the healthcare providers to ‘do more with less’ and help them efficiently allocate resources like medicines, equipments, qualified physicians, nurses and caregivers. This is a critical need for hospitals to remain viable because there is an ever-increasing shortage of such resources.

**Exploratory data analysis**

The data preprocessing is an essential step in getting better accuracy and also to run the models well. We initially found out that 5500 rows have no data at all. Hence, we dropped these rows. Furthermore, the data has a lot of missing entries. Therefore, the data preprocessing (imputation of missing values) has to be done with respect to each column and also we have to convert many variables into factors. The main reason for performing the imputation against deleting the columns is that we don’t want any loss of information.

The pre-processing of the columns is as follows-

1)INDEX - This column is not useful as it does not provide any additional predictive power. So we dropped it.

2)HOSPITAL, FINANCIAL\_CLASS, CONSULT\_ORDER, CONSULT\_CHARGE, DIAGNOSIS, DIAG\_DETAILS - We did not perform any data manipulation as they didn’t have any missing values.

3)GENDER, ED\_RESULT - We imputed the missing data points using the most common class which is Male, Discharge respectively.

4)RACE - We placed the missing into the unknown category. As the we can intuitively say the missing values most likely would be an Unknown.

5)ETHNICITY - We placed the missing into the unknown category. As the we can intuitively say the missing values most likely would be an Unknown.

8)WEEKDAY\_ARR, HOUR\_ARR - - We would not include this variable as it is an exact copy of the Week Day Departure variable and having a duplicated row does not add any predictive.

9)MONTH\_DEP, SAME\_DAY,"HOSPITAL", "GENDER","RACE", "ETHNICITY", "FINANCIAL\_CLASS", "WEEKDAY\_ARR", "HOUR\_ARR", "MONTH\_ARR", "WEEKDAY\_DEP", "HOUR\_DEP", "ED\_RESULT", "ACUITY\_ARR", "ADMIT\_RESULT", CONSULT\_ORDER", "CONSULT\_CHARGE", "CONSULT\_IN\_ED", "DIAGNOSIS", "RISK", “SEVERITY"- We converted these columns into a factor.

10)ACUITY\_ARR - Replaced the 1 Purple entry with most common class (3-urgent)

Replaced the missing values with (5-non urgent) as for around 2000 of these cases have charges as 0. Also, to support all of them were discharged on the same day. Also Diagnosis was ‘No’ for most of these patients.

11)DC\_RESULT - This one missing value deleted when we deleted the 141 missing rows for the target variable Return.We are classifying the data into 4 bins. The first three bins are the categories with highest number of data points and we classified the rest as the other bin. We did this as the we run into problems when running the model as when we split the data, few categories can only exist in the validation when that particular category is not in the training data. This

can be attributed to having very less data points in some categories.

12)ADMIT\_RESULT -When compared it with ED\_RESULT and observed they have a trend. Most of the missing values here indicated categories like ‘Discharge’, ‘left before completing the treatment’ etc which gives us a very strong reason that those patients were not admitted to the hospital. Hence, we created a new category named ‘ Not Admitted to Hospital’.

13)CONSULT\_IN\_ED - We could see that this categorical variable could only have 2 values that is 0 or 1. In the dataset we only have 1’s and no 0. So, by using the variable definition we can intuitively say that the missing values could only mean that the consulting did not originate in an emergency room.

14)RISK, SEVERITY - We found out that there is a relationship between Risk and Accuity arrival, Risk and ED\_Result. However, we are not sure if we could use this relation to impute the missing values in risk. So for now we created a new category ‘Others’ were we placed in all the missing values.

15)CHARGES -We imputed the missing values with median. We did this because, we were not able to find any meaningful relationship of this variable with any other variable and also where not able to intuitively say anything about this variable. Hence, we assumed the values are missing at random. Therefore, we wanted to impute the data either with median or mean. However, we used median as the data has outliers.

16)AGE- We replaced age of the patients in excess of 120 by the median value of the age in the dataset.

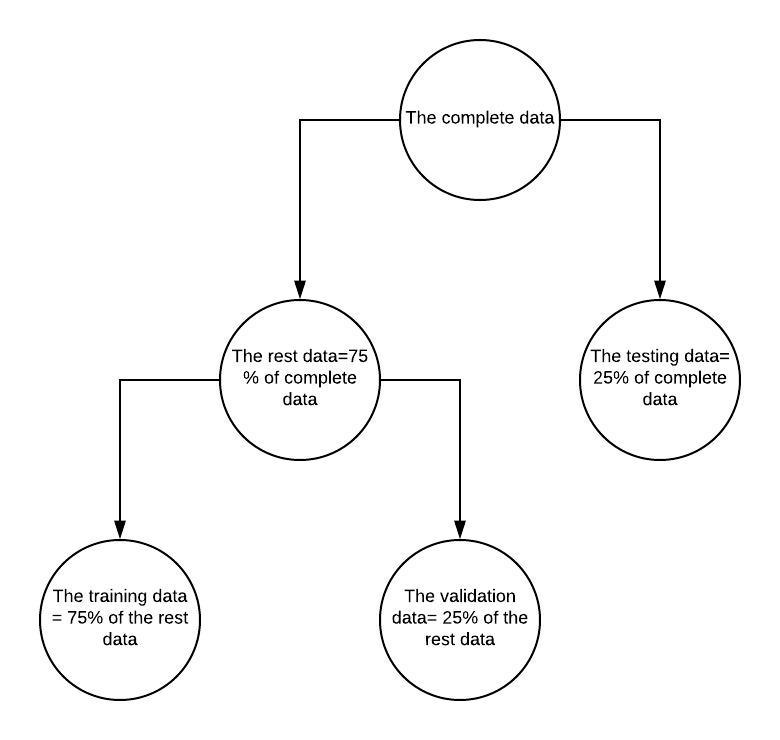
17)RETURN- We replaced the 141 missing values by the majority class.

We carried out the major exploratory data analysis in Phase 1, therefore we have attached the complete data analysis in Appendix A (attached)

**Data Partitioning**

We partitioned the training data into three parts - training data, validation data and testing data. We selected 25% of the

data as the testing data. We have taken 25% of the rest data as validation data and the remaining of it as the training data.

****

**MODELING AND MODEL EVALUATION**

Since accuracy is our primary goal, we ran as many classification models as possible to achieve the highest accuracy. We tuned the parameters for each model using the validation data and finally compared the testing data accuracy to decide the best model to use in practice.

Out of all the models we ran we got the most convincing results on Bagging with trees, Boosting , XGBoost, Random Forest and Logistic regression since these models gave us the highest accuracy.These models are discussed below in detail..

**LOGISTIC REGRESSION :**

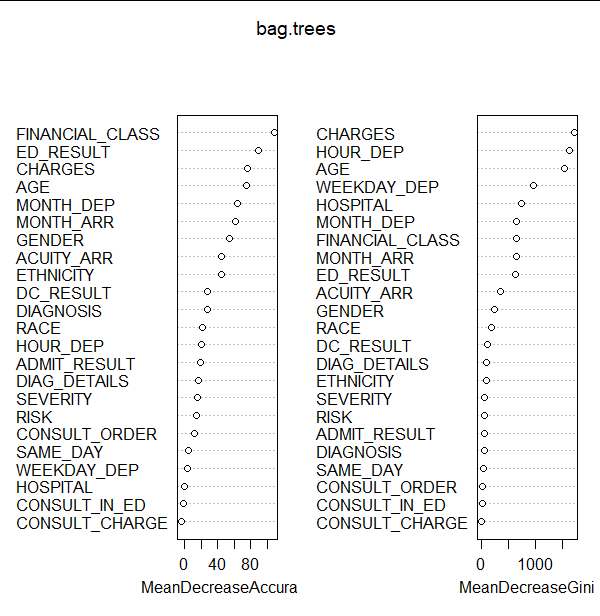
We tried logistic regression because it’s a classification model and computationally less demanding given our huge data set. So, we first ran a logistic model using all the variables (log1) to predict RETURN. Then we tried logistic model (log2) removing the variables from LASSO (Diagnosis & Diag\_Details) . We also tried backward step-wise (log3) and forward step-wise model (log4). Both these methods gave exactly the same final output/model. As many as 8 variables were eliminated - Hospital, Month\_Arr, Consult\_Charge, Consult\_in\_ed, Diagnosis, Diag\_Details, Risk and Severity.

Although AIC decreased from log1 to log2 and log3/log4, the highest validation accuracy of 76.86% occured for log1 model at a cutoff of 0.47. All the models were tried with different cut-offs from 0.1 to 0.9 and then zoomed in on 0.4 to 0.6 to achieve the highest validation accuracy for the respective model. So, we re-trained model log1 using the rest data and got a testing accuracy of 77.10 which is higher than our testing baseline accuracy. The variables that were statistically the most significant were Hospital, Gender, Financial Class, Weekday\_Dep, Hour\_Dep, ED\_Result and Charges. We expect most of these variables to be important in our subsequent models as well.

**BAGGING :**

We tried bagging with trees because trees are data driven. Although our single tree gave just a baseline model, bagging uses bootstrap sampling of the data and we expect the aggregated result to provide a better accuracy. We used the RandomForest library in R to run our bagging models by specifying mtry = 23 (it’s just like running a random forest using all the variables). We trained our bagging models on the training data. We tried different combinations of ntree from 100 to 2000 in multiples of 100 and also tried different cutoff values on the validation data. We obtained the highest validation accuracy (77.3%) with 1000 trees and a cutoff of 0.49. We re-trained this bagging on our rest data and obtained a testing accuracy as 77.4%.

From the variable importance plot we found that the most important variables in bagging with trees models are ED\_Result, Financial\_Class, Month\_Arr and Age.This can be clearly seen from the model output also.



**RANDOM FOREST:**

Because there are huge number of variables in our dataset and it may be that we have very strong predictors along with moderately strong predictors. Then, bagging may lead to trees that all look the same.So,we use random forests to decorrelate the trees.

The basic procedure is to train the model on training data and choose parameters based on validation error.the parameters are number of trees, subset of variables for each split in the tree.then retrain the model on the rest data and check the accuracy on testing data.The best parameters are variables= 4 and number of trees =1000.Here, we achieved validation accuracy of 77.45 on the validation data and 77.60 on the testing data using the same parameters, this accuracy was higher than the one we got with Bagging.

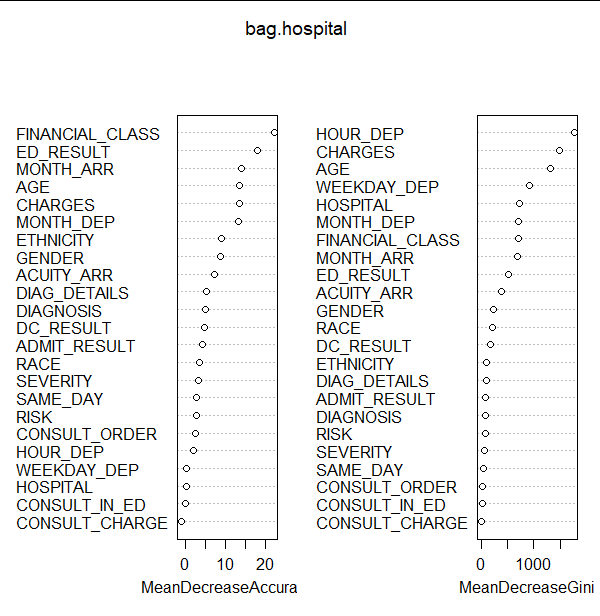
Few major trends observed :

# The variable importance plot remained almost the same ( the top 6 most important variables remained in top 6 almost in the same order) for various values of mtry.

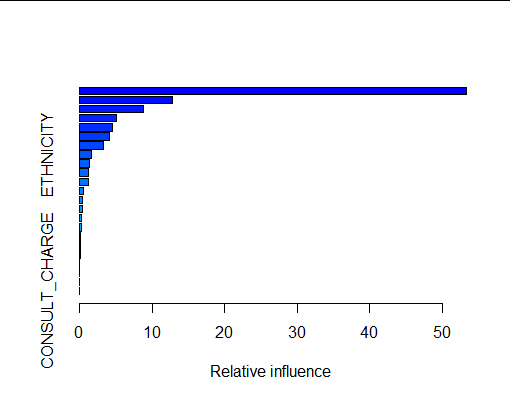
# The variable importance plot was very similar to that of Bagging.

# For a given mtry the variable importance almost remained the same for ntrees tried.

#The top 6 importance variables are ED\_Result, Financial\_Class, Month\_Arr ,Age, Charges, Month\_Dep.



**BOOSTING :**We tried Boosting model since it gives more weightage to the misclassified points and since the aim of our project is to maximize the overall model accuracy. We trained our model on the training data , used validation data to get the best value of cut off which we got as 0.55 and number of trees as 3000. Here , we used different combinations of the cut off value and the n.trees by using a for loop and chose the combination that gave us the best value of validation accuracy which was 77.76% and then used these parameters to perform predictions on the test data. and achieved accuracy of 77.70% .



It can be clearly seen that in case of Boosting, variables such as CHARGES,HOUR\_DEP,FINANCIAL\_CLASS play major role in classifying the misclassified points.

**XGBOOST :** Going by the results we achieved on Boosting which superseded our previous best model(Bagging) with highest accuracy, we wanted to improve our results further and reduce the total computation time which was too high in case of boosting, we therefore ran XGBOOST, here we trained our data on training data and achieved maximum accuracy of 78.40% on our validation data at cut off value of 0.56 and using 200 nrounds(number of trees), eta(learning parameter) as 0.1 and max\_depth(depth of the tree) as 6. On applying this model to our testing data , this model gave us highest accuracy of 78.70% , which was 1 percent more than what we got with Boosting.

Though we ran other models too like LDA,LASSO,RIDGE,DECISION TREES but they were excluded from further consideration due to poor value of accuracy. Some of those models are discussed in Appendix B(attached).

**Summary :**

Based on our results we can clearly say that the Hospitals must put XGBOOST model into practice since this model gave us the highest value of accuracy on the unseen data which is approximately 2.82 percent higher than the baseline accuracy and is expected to yield best results in predicting our variable *RETURN.* This model`s results can be further used by the hospitals in managing their resources well, keeping a track of the stock of medicines and other medical equipments and plan their funds accordingly. Also, based on our results the most important predictors proved to be *ED\_RESULT,CHARGES (paid by the patient),FINANCIAL\_CLASS* Therefore , the Hospital staff must ensure these variables are given top most priority while entering the patient details in the database so as to accurately predict the future patients RETURN variable so as to better optimize their resources and improve patients service.

**APPENDIX A**

**Data Insights**

**Structure of data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **Variables** | **Structure** | **Observations( Relations with other variables or RETURN variable )** | **Data cleaning/Transformation** |
| 1 | INDEX | This column is not useful as it does not provide any additional predictive power | It is different for all the observations for all other variables in data. | Not required as the we are not going to use this in our model. |
| 2 | HOSPITAL | This column has 5 different categories that is hospital A,B,C,D,E.  This column does not have any missing values.  The values are evenly split across the hospitals. | All the levels in HOSPITAL are equally distributed with respect to RETURN. | No data cleaning required. |
| 3 | GENDER | This variable has two outcomes which is Male or Female.  The column has two missing values. | In relation to the return column the number of females who do not return to hospitals are greater than males. Moreover, the number of males who return to the hospital are greater than the number of females. This is a point to note as the number of males and females in the dataset are almost the same. | We imputed the missing data points using the most common class which is Male. |
| 4 | AGE | This variable has no missing values. | AGE of the patients tends to be almost normally distributed.The average age of patients who are returning back to hospital in the next 30 days from their discharge tends to be higher than that of patients who are not returning back.Majority of patients are aged between 20 and 60. | We analyzed the data and found that there were few instances where the age of the person was entered 150, but using the domain knowledge we know that in the history of mankind no one has survived beyond 120 years, therefore considering this we replaced all such values by the median age(as there are outliers in our data set |
| 5 | RACE | In this variable we have 177 missing values.  In total we have 11 different types of input including the missing values. | We have observed no specific or special pattern when compared RACE with RETURN.We also tried various combinations of RACE with ETHNICITY ,ACCUITY ARR etc.We weren’t able to find any specific patterns in RACE. | We placed the missing into the unknown category. As the we can intuitively say the missing values most likely would be an Unknown.( unable to find any pattern with respect to other variables to impute variables with a known category) |
| 6 | ETHNICITY | This variables has 5 categories including the missing values.  We have 373 missing values in this variable. | Similar to RACE , We have plotted ETHNICITY with various other variables.But,haven’t been able to find any specific pattern. | We placed the missing into the unknown category. As the we can intuitively say the missing values most likely would be an Unknown..( unable to find any pattern with respect to other variables to impute variables with a known category) |
| 7 | FINANCIAL\_CLASS | This variable has 13 types of input.  It has no missing values. | Intuitively, there appears to be a relation between FINANCIAL\_CLASS and CHARGES.We will explore this in our Phase 2 | No data cleaning required. |
| 8 | WEEKDAY\_ARR | This variable has 7 different categories.  It has no missing values. | It appears to be duplicated data of WEEKDAY\_DEP | We would not include this variable as it is an exact copy of the Week Day Departure variable and having a duplicated row does not add any predictive power. |
| 9 | HOUR\_ARR | This variable has 24 different categories.  It has no missing values. | It appears to be duplicated data of HOUR\_DEP | We would not include this variable as it is an exact copy of the Hour Departure variable and having a duplicated row does not add any predictive power. |
| 10 | MONTH\_ARR | This variable has 11 different categories.  It has no missing values. | It appears to be highly correlated with MONTH\_DEP | We converted this into a factor. |
| 11 | WEEKDAY\_DEP | This variable has 7 different categories.  It has no missing values. | It appears to be duplicated data of WEEKDAY\_ARR. | We converted this into a factor. |
| 12 | HOUR\_DEP | This variable has 24 different categories.  It has no missing values. | It appears to be duplicated data of HOUR\_ARR. | We converted this into a factor. |
| 13 | MONTH\_DEP | This variable has 12 different categories.  It has no missing values. | It appears to be highly correlated with MONTH\_ARR | We converted this into a factor. |
| 14 | SAME\_DAY | This variable has two categories which are 0 and 1.  It has no missing values | Even though majority of the patients left on the same day, we weren’t able to pinpoint it to any particular trend when compared with various variables as ‘SAME\_DAY=1’ is a majority class and is not exclusive to any category of level of other categorical variables. | We converted this into a factor. |
| 15 | ED\_RESULT | This variable has 17 different categories including the missing values.  It has 73 missing values. | We checked for relation between ED\_RESULT and DC\_RESULT and found that the categorical variables of both the variables have lots in common.(They tend to provide same information but in a much detailed way).It will further be explored in Phase 2. | We replaced the missing values with the majority class ‘Discharge’ |
| 16 | ACUITY\_ARR | This variable has 7 different categories.  It has 3263 missing values.  It also has 1 data point with value as 5 Purple | Tried to find patterns between ACUITY\_ARR and RISK and SEVERITY . Also, analysed it with RETURN and found no direct relation. | Replaced the 1 Purple entry with most common class (3-urgent)  Replaced the missing values with (5-non urgent) as for around 2000 of these cases have charges as 0. Also, to support all of them were discharged on the same day. Also Diagnosis was ‘No’ for most of these patients. |
| 17 | DC\_RESULT | This variable has 37 different categories including the missing values..  It has just one missing value . | We checked for relation between ED\_RESULT and DC\_RESULT and found that the categorical variables of both the variables have lots in common.(They tend to provide same information but in a much detailed way).It will further be explored in Phase 2. | This one missing value deleted when we deleted the 141 missing rows for the target variable Return.We are classifying the data into 4 bins. The first three bins are the categories with highest number of data points and we classified the rest as the other bin. We did this as the we run into problems when running the model as when we split the data, few categories can only exist in the validation when that particular category is not in the training data. This can be attributed to having very less data points in some categories. |
| 18 | ADMIT\_RESULT | This variable has 4 different categories.  It has around 30000 missing values. | Since,majority of this values are missing we felt that it’s not appropriate to conclude anything based on patterns with other variables.But in order to impute missing values we have checked for trends with ED\_RESULT | When compared it with ED\_RESULT and observed they have a trend. Most of the missing values here indicated categories like ‘Discharge’, ‘left before completing the treatment’ etc which gives us a very strong reason that those patients were not admitted to the hospital. Hence, we created a new category named ‘ Not Admitted to Hospital’. |
| 19 | CONSULT\_ORDER | This variable has 2 different categories.  It has no missing values.  The frequency of this variable is-  x freq  0 35668  1 2553 | We can combine CONSULT\_ORDER and CONSULT\_IN\_ED variables to create a new variable with 3 levels ( consult in ED ,consult in non ED , No consult ORDER ) --Will be explored and analysed in phase 2 | We did not make any changes in this variable. |
| 20 | CONSULT\_CHARGE | This variable has 2 different categories.  It has no missing values.  The frequency of this variable is-  x freq  0 37997  1 224 | We saw that approximately only 10 percent of people who had a CONSULT\_ORDER have been charged. | We did not make any changes to this column. |
| 21 | CONSULT\_IN\_ED | This variable has 2 different categories.  It has 37321 missing values.  The frequency of this variable is-  x freq  1 900  NA 37321 | We can combine CONSULT\_ORDER and CONSULT\_IN\_ED variables to create a new variable with 3 levels ( consult in ED ,consult in non ED , No consult ORDER ) --Will be explored and analysed in phase 2 | We could see that this categorical variable could only have 2 values that is 0 or 1. In the dataset we only have 1’s and no 0. So, by using the variable definition we can intuitively say that the missing values could only mean that the consulting did not originate in an emergency room. |
| 22 | DIAGNOSIS | This variable has 2 different categories.  It has no missing values.  x freq  No 9822  Yes 28399 | Checked for pattern with DIAG\_DETAILS and SAME\_DAY. | We did not make any changes to this variable. |
| 23 | DIAG\_DETAILS | This variable has 14 different categories.  It has no missing values. | We felt that this variable can be at very random and also checked patterns with DIAGNOSIS.Where there is DIAGNOSIS is 1, there are 0 details and at few places details are higher and also vice versa.So,we didn’t feel there is a relation between DIAGNOSIS and DIAG\_DETAILS | We did not make any changes to this column. |
| 24 | RISK | This variable has 5 different categories including the missing values.  It has 33045 missing values. | Since,the majority of values are missing,we felt that it’s not appropriate to come to conclusions based on minimal data we have in this column. | We found out that there is a relationship between Risk and Accuity arrival, Risk and ED\_Result. However, we are not sure if we could use this relation to impute the missing values in risk. So for now we created a new category ‘Others’ were we placed in all the missing values. |
| 25 | SEVERITY | This variable has 5 different categories.  It has 33045 missing values. | Since,the majority of values are missing,we felt that it’s not appropriate to come to conclusions based on minimal data we have in this column. | We found out that there is a relationship between Severity and Acuity arrival, Severity and ED\_Result. However, we are not sure if we could use this relation to impute the missing values in risk. So for now we created a new category ‘Others’ where we placed in all the missing values. |
| 26 | CHARGES | This variable has 500 different categories.  It has 100 missing values. | We feel that there can be an interaction between CHARGES and FINANCIAL\_CLASS. It will be explored and analysed in Phase 2. | We imputed the missing values with median. We did this because, we were not able to find any meaningful relationship of this variable with any other variable and also where not able to intuitively say anything about this variable. Hence, we assumed the values are missing at random. Therefore, we wanted to impute the data either with median or mean. However, we used median as the data has outliers. |

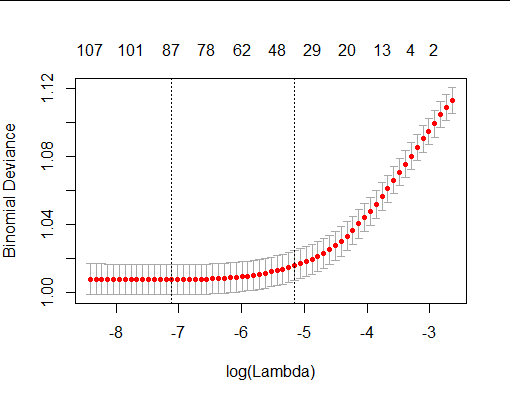
**APPENDIX B**

**LDA MODEL :**

Knowing the fact that LDA assumes similar variation in distribution of 1`s and 0`s which was not the case with our dataset since our RETURN column has majority of 0`s and data set is highly skewed but since LDA is a data driven model and we wanted to apply it on ensemble methods (based on the value of the accuracy achieved), we still ran LDA model on training data and picked the best cut off value (0.52) based on the highest achieved validation accuracy of 76.93% . Then using this cut off value we re-trained our data on the rest data , and achieved accuracy of 76.69% on the testing dataset which was much lower than the accuracy achieved on logistic regression. Therefore, we removed this model for further analysis.

**LASSO MODEL:**

Since our data points are skewed leading to high variance , we wanted to remove some of the variables from our dataset that had no predicting power and then use the final coefficients to run logistic regression to improve our logistic regression accuracy, so we ran LASSO model on the train data , performed cross validation and found best lambda value of .00081 and then used this lambda value to perform prediction on the testing data to achieve accuracy of 76.79% at 0.48 value of cutoff found using validation data (by trying different combinations of cut off on validation data and chose the one that gave highest value of accuracy . Since, this is a classification problem so we used logistic type model for this. The model only removed 2 of the 24 variables i.e. Diagnosis, Diag\_Details completely from the model, which was contrary to our expectation as we expected more variables to get eliminated, now in order to check our results we ran a logistic regression removing those 2 variables and checked the accuracy on the logistic regression, which did not improve our accuracy by any considerable amount, therefore we ruled out this model too from further analysis . Plot showing the corresponding values of lambda and variables is attached

# Kindly note there were 121 levels in all and after performing Lasso we got rid of only a few levels.

**RIDGE MODEL:**

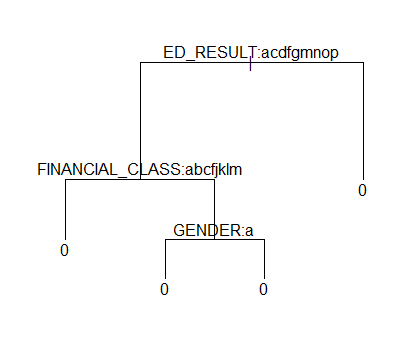
We ran RIDGE Model in order to check its accuracy on the testing data and see if the results are in accordance with LASSO and followed the same procedure as we did with LASSO. As expected the testing accuracy of RIDGE model almost gave the same accuracy as that of the LASSO model and did not show any appreciable improvement over Lasso since the value of accuracy achieved was 76.82%.. Since the model did not do us any better , we decided to rule it out too from our consideration as logistic model that we ran before still had better accuracy than any of these.

Attached the Ridge plot against the lambda values showing the number of variables used by our model.

**CLASSIFICATION TREE** :

Because we are dealing with a classification problem ,Classification Trees is a good option .Also our data has outliers and trees are robust to outliers. Moreover, since trees require large dataset and we have a large dataset.The basic procedure was to train the trees on training data and then select the parameters based on validation data .The parameters to be selected are cutoff and number of splits. Then choose the parameters that gives the least validation error and retrain the model on the rest data. Then we calculate the accuracy of the model on testing data.

Few Important observations from the model:

The first fully grown tree we obtained is as follows

# The first split occurs on ‘ED\_RESULT’ & then ‘FINANCIAL\_CLASS’ AND ‘GENDER’.But all the nodes are predicting 0’s at a cutoff of 0.5.So,this tree doesn’t seem very useful.After checking for parameters on validation data,we saw that pur at .5 cutoff and best=2 it’s giving validation accuracy equal to baseline accuracy.Once retrained on rest data ,the test accuracy we obtained was equal to it’s baseline. i.e ( 75.2381)

**KNN ( K Nearest Neighbours)** :

We tried a classification KNN as KNN is data driven and we have a huge dataset. Most of the variables in our dataset are categorical and the numerical variables like charges dominate the dataset as KNN works on euclidean distance. We didn’t normalise our variables.And as we have 23 variables , the calculation gets complicated.As knn requires numerical inputs,we used model matrix .

The basic procedure was to train the model on our training data and select the parameter : k ,based on validation error. Then retrain the model on rest dataset and check the testing accuracy.We tried k values between 0 and 50 and obtained best model at k=31 .Our testing accuracy was 75.91 which is slightly higher than the baseline accuracy of our testing data.