---

title: "Texas DSHS Project"

pagetitle: "Texas DSHS Project"

author: "Husein Ali"

date: "September 27, 2019"

output:

pdf\_document: default

html\_document: default

word\_document: default

---

## Introduction

The overall objective of the project was to predict the financial risk of a patient as he or she is admitted. Both hospitals and insurance companies will benefit from these predictions. Hospitals will be able to forecast the patient’s total bill prior to any treatment and determine what portion of the bill insurance will pay for. By utilizing these models, insurance companies will be able to ensure a patient’s insurance plan is adequate and can plan their financials accordingly.

In order to generate this predictive model, the Texas Department of State Health Services Center for Health Statistics dataset was used. This dataset contained 743,626 records and has a total of 194 possible variables, including hospital type, location, length of stay, insurance coverage, age, ethnicity, day of the week, a variety of admittance codes, patient status, and a variety of diagnoses codes. The total charge variable was divided into two levels (0 and 1) to indicate whether the financial risk of the patient is high or low. After examining the dataset, the team determined that type of admission, source of admission, length of stay, discharged, age, risk mortality, and illness severity were the most impactful variables. Next, the team generated three different types of models: decision tree, KNN, and logistic regression.

Through the research and analysis, the team was able to successfully generate a model to determine the financial risk of a patient at admittance. As healthcare costs rise each year with the cost of insurance, the usage of predictive analytics can provide employers, hospitals, and insurance companies the ability to improve their future operations and financial budgeting.

## Previous Studies

After doing some research online, our predictive data mining problem turned out to be one others have attempted to tackle in the past. Predicting if a patient will turn out to cost a lot for the hospital is one of great value because it provides critical budgeting decision information and a detailed map for reducing those costs in the future.

The first study we came across was very similar to our problem of predicting the financial risk of patient. They came to the conclusion that a small amount of the population, 5% of users in Ontario, make up the majority of healthcare costs, 61%, and wanted to identify theses HCUs, High Cost Users, in order to intervene before substantially more costs were created(1). Building upon a previous study, length of stay, illness severity, and emergency department use were the key predictors which come very close to our predictors. They performed a logistic regression model with 97 variables ranging from demographics like age and sex to illness indicator variables. They were able to predict the highest cost 5% of patients with a 94 percent accuracy so their model worked very well.

The second study came to the same conclusion that the top 1% of health care users accounted for ⅕ of the medical costs. Using a texan medical dataset, they concluded that costs were directly related to texas medicaid users which makes a lot of sense. They used similar predictors at the first study however a few different models including ordinary least squares regression, regularized regression, gradient boosting machine, and recurrent neural networks. The best models were regularized regression and gradient boosting machine with the most interpretable results. They determined predictive models work well for predicting healthcare related costs and that health care costs are temporally consistent(2).

The final observation came from an article reviewing the potential of predictive analytics in healthcare. A notable fact in the article was attempting to use analytics to figure out what causes patients to be readmitted within a 30 day window(3). The patients with high chances for quick readmission cost the hospitals a lot of money and if a predictive model can find patients like these doctors can intervene to make sure the likelihood of readmission is much lower. As a result, this should lower healthcare costs significantly. This could be a major application of our predictive model after determining which patients qualify for being a high financial risk.

## Data Collection & Analysis

\* Dataset Analysis

The Hospital Discharge Dataset was obtained from the Texas Department of State Health Services website. The dataset included records from 558 Texas hospitals during the first quarter of 2012. This dataset contained 740,817 records from 194 variables. This variables ranged from hospital type, location, patient gender, day of the week, to patient status. A majority of these variables were hospital-specific codes which were not usable in building predictive models.

Unfortunately, this dataset was very inconsistent and required extreme cleaning. Initially, there were 97,752,607 empty cells. This was mainly due to the fact that 152 variables contained mostly blank cells. Due to this, these variables were excluded from the predictive models and removed from the dataset. Next, the records with blank cells were deleted from the dataset. The dataset was reduced by 600,866 records, resulting in a total of 139,951 records in the dataset. Next, all of the out of state resident patients were deleted, resulting in solely Texan patients. The sex variable was transformed into dummy variables. Initially, the age variable was comprised of 21 categories. However, this variable was reduced to 11 more inclusive categories for the predictive models.

The final action to prepare the data was creating the outcome variable, financial risk. This variable was created by creating two distinct categories from the “Total Charge Level” variable. A financial risk of 1 represented a patient with a total charge less than $25,000. A financial risk of 2 represented a patient with a total charge greater than or equal to $25,000. The distribution is shown in the table below. There are 78,889 patients with a low financial risk of 1 (56.4% of the records) and 61,062 patients with a high financial risk (43.6% of the records). This distribution is shown in the graph above. The categories were not split into equal 50% divisions so that simpler threshold numbers could be used for easier and quicker comprehension.

## Feature Selection

```{r Feature Selction}

library(ISLR)

library("readxl")

library(leaps)

hospital<-read\_xlsx("Final Data\_Report 3.xlsx")

set.seed(1)

train = sample(1:nrow(hospital), nrow(hospital)/2)

colnames(hospital, do.NULL = FALSE)

colnames(hospital) <- c("TypeA","SourceA","Discharged","LengthStay","Age","RiskMortality","IllnessSev","FRisk","Beds")

#forward stepwise

regfit.fwd=regsubsets(hospital$FRisk~., data=hospital, nvmax=19,method="forward")

summary(regfit.fwd)

#backward stepwise

regfit.back=regsubsets(hospital$FRisk~., data=hospital, nvmax=19,method="backward")

summary(regfit.back)

#best subset selection

regfit.best=regsubsets(hospital$FRisk~., data=hospital, nvmax=19)

summary(regfit.best)

```

## Variable Analysis

Next, each of the eight variables were analyzed for understanding. The first variable evaluated was type of admission. The relationship between type of admission and financial risk is demonstrated in the chart above. There were five categories for this variable: emergency (1), urgent (2), elective (3), newborn (4), and trauma center (5). The newborn category appears to have the lowest financial risk, with an average financial risk of 1.03. The type of admission with the highest financial risk is trauma center, which has an average financial risk of 1.89.

The next variable analyzed was source of admission. There were nine categories for this variable: non-healthcare facility (1), clinic referral (2), transfer from another home health agency (3), transfer from hospital (4), transfer from skilled nursing facility (5), transfer from another healthcare facility (6),transfer from surgery center (7), court/law enforcement (8), and transfer from home health agency (9). This bar graph above demonstrates that different sources of admission have different average financial risk levels. The source of admission with the lowest financial risk is transfer from a skilled nursing facility, with a risk of 1.05.

The third variable analyzed was discharged. A value of 1 means that a patient expired in the hospital, while a value of 0 means the patient was discharged from the hospital. The chart above shows that patients who were discharged have an overall higher financial risk.

The fourth variable analyzed was length of stay. Length of stay represents the number of days the patient was in the hospital. This number ranged from 1 to 595 days. The chart above displays the relationships between financial risk and length of stay. The line graph clearly shows that once a patient stays at the hospital for more than about 50 days, the patient will be at a financial risk of 2 and therefore will be paying more than $25,000. The drop in financial risk on day 372 occurred because there was only one record with that length of stay and this patient had a financial risk of 1.

The graph above displays the relationship between financial risk and age. The age variables is broken up into 11 categories: infant, 1-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, and 90+. The graph clearly shows that as a patient becomes older, his or her financial risk increases.

The sixth variable examined was risk mortality. This variable has five categories, 0 to 4, with 0 being low risk of mortality and 4 being a high mortality risk. The graph above clearly shows that as the risk of mortality increases, so does the financial risk.

The graph above displays the relationship between illness severity and financial risk. Similar to risk mortality, illness severity has five categories, 0 to 4. An illness severity of 0 denotes that the patient does not have a severe illness, while an illness severity of 4 means that the patient has a very severe illness. The graph shows that the financial risk steadily increases as a patient’s illness severity increases.

The final variable is number of beds. This variable is referring to the number of beds in the hospital where the patient was admitted. The number of beds per hospital ranges from 4 to 189. The graph above shows the relationship between the number of beds in a hospital and the financial risk of a patient. There is no clear trend shown in the graph.

## Variable Selection

A correlation matrix was created in order to learn how the variables were related. The correlation matrix is shown above. The most highly correlated variables are risk mortality and illness severity. These variables have a correlation of 1.0, showing that these variables have overlap in information. Patient age and risk mortality have a correlation of 0.60, which is fairly high. Additionally, type of admission and source of admission have a high correlation of 0.53.

Next, forward stepwise selection was conducted. The results of this method are shown above. The process concluded that risk mortality, beds, and discharged are the least significant predictor variables. The selection process also showed that illness severity, age, length of stay, and source of admission were all important variables.

Backward stepwise selection was also run. The results of this process are shown above. The results are extremely similar to forward stepwise selection, as age, length of stay, illness severity, and source of admission are significant variables.

The final variable selection process conducted was stepwise selection. The results of this process are shown above. Again, the results are very similar to both forward and backward stepwise selection as the same variables are shown to be significant in all three models.

## KNN Classification

```{r KNN Code, cache=TRUE}

rm(list = ls())

library("readxl")

library(class)

ds<--read\_excel("Final Data\_Report 3.xlsx")

head(ds)

ds= ds\*-1

ds$`Financial Risk` <- ifelse(ds$`Financial Risk`== 2, 1, 0)

ds$`Financial Risk` = as.factor(ds$`Financial Risk`)

set.seed(123)

trainingRowIndex <- sample(1:nrow(ds), 0.5\*nrow(ds))

train.x=ds[trainingRowIndex,c(1:7,9)]

train.y=ds[trainingRowIndex,8]

test.x=ds[-trainingRowIndex,c(1:7,9)]

test.y=ds[-trainingRowIndex,8]

test.x = test.x[-1,]

test.y = test.y[-1]

knn.model<-knn(train.x,test.x,train.y,k=10, prob=T)

table(knn.model,test.y)

mean(knn.model==test.y)

library(pROC)

roc(test.y, attributes(knn.model)$prob)

plot(roc(test.y, attributes(knn.model)$prob),

print.thres = T,

print.auc=T)

acc<-rep(0,10)

for (knum in 1:10) {

knn.fit<-knn(train.x,test.x,train.y,k=knum)

acc[knum]<-mean(knn.fit==test.y)

}

print(paste("Highest accuracy of",as.character(acc[which.max(acc)]),"at k =",as.character(which.max(acc) )))

plot(1:10,acc,type="p",xlab="k",ylab="Test Acc", pretty=0)

knn.model<-knn(train.x,test.x,train.y,k=7, prob=T)

table(knn.model,test.y)

mean(knn.model==test.y)

library(pROC)

roc(test.y, attributes(knn.model)$prob)

plot(roc(test.y, attributes(knn.model)$prob),

print.thres = T,

print.auc=T)

```

The next model that we built was the KNN Classification model. We included all the variables that we selected after variable selection method that we also used in the logistic regression model. Since the KNN algorithm requires an even test-train split, we used 50% of the data as training data set and the remaining 50% was used for testing data set. Since there was an odd number of rows, one extra observation was eliminated from the test data set.

In the dataset, the Financial Risk is 1 or 2. However, the KNN algorithm requires the response to be either 0 or 1, so we made a transformation. All of the financial risk that was originally level 1 was changed to 0, and those which were level 2 were changed to 1.

Initially, we took the value of K to be 10. The confusion matrix that we obtained is as follows:

Financial Risk Prediction

0 1

0 34,799 5,995

1 4,683 24,498

The accuracy of this model turned out to be 84.74% with a misclassification rate of 15.26%. This was the most accurate model since the test error reduced drastically as compared to the other two models. The AUC curve was plotted for this model which is shown below:

The area under this curve was lesser as compared to the logistic model because the proportion of the observations were more for false positives in case of the KNN model. This made the AUC curve flatter with a lesser area of sensitivity versus specificity.

Next we calculated all the possible KNN models for the value of K between 1 and 50.

The highest test accuracy of the models was found to be 84.85% at K=7. This accuracy was more than that of the KNN model at K=10, and has a misclassification rate of 15.15%. The confusion matrix for this model is as follows:

Financial Risk Prediction

0 1

0 34,776 5,898

1 4,706 24,595

We then plotted the AUC curve and found that the AUC also improved to 55.8% as compared to the KNN classification model with k=7. This is shown as follows:

The false positives were higher in the KNN model as compared to the logistic and decision tree models, which is why the AUC curve here shows a slightly flattened feature. Also, since the confidence intervals in KNN are only between 0 and 1, this curve is unattractive and is flattened but shows a correct ROC plot [4].

## Logistic Regression

```{r Logistic}

rm(list = ls())

library("readxl")

ds<--read\_excel("Final Data\_Report 3.xlsx")

head(ds)

ds= ds\*-1

ds$`Financial Risk` <- ifelse(ds$`Financial Risk`== 2, 1, 0)

attach(ds)

log.results=glm(`Financial Risk`~.,data=ds,family=binomial)

summary(log.results)

probs<-predict(log.results,type="response")

pred<-rep(0,nrow(ds)) # create a vector of predictions with default value of low risk

pred[probs>0.5]=1

table(`Financial Risk`,pred)

mean(`Financial Risk`==pred)

detach(ds)

set.seed(123)

trainingRowIndex <- sample(1:nrow(ds), 0.8\*nrow(ds))

train<-ds[trainingRowIndex,]

test<-ds[-trainingRowIndex,]

train.glm = glm(`Financial Risk`~.,data=ds,family=binomial)

summary(train.glm)

probs<-predict(train.glm,test,type="response")

m.pred<-rep(0,nrow(test))

m.pred[probs>0.5]<-1

table(test$`Financial Risk`,m.pred)

mean(m.pred==test$`Financial Risk`)

# With reduced # of observations

train.glm = glm(`Financial Risk`~.-RISK\_MORTALITY,data=ds,family=binomial)

summary(train.glm)

probs<-predict(train.glm,test,type="response")

m.pred<-rep(0,nrow(test))

m.pred[probs>0.5]<-1

table(test$`Financial Risk`,m.pred)

mean(m.pred==test$`Financial Risk`)

thresholds<-seq(from=0.1, to=0.9, by=0.05)

acc<-rep(0,19)

for (i in 1:19) {

newpred<-rep(0,nrow(test))

newpred[probs>thresholds[i]]<-1

table(test$`Financial Risk`,newpred)

acc[i]<-mean(newpred==test$`Financial Risk`)

acc[i]

}

acc

max(acc)

which.max(acc)

```

Another model that we decided to move forward with was one of the simplest classification models, the logistic regression model. We chose a multiple logistic regression model with all the predictors initially. The train test split for this model was considered to be 80-20 to achieve highest possible accuracy with less chances of the data being overfit. On the train data set, we first built a model with eight variables including the Type of Admission of the Patient, the Patient’s Source of Admission, whether or not the patient was discharged, the patient’s length of stay, the patient’s Age, patient’s risk of mortality, severity of illness of the patient, and number of beds in the hospital. After building the model on the training data set with these variables, we obtained the following results:

This model was built considering the probability between the classes to be 0.5. So, the values of the probabilities greater than 0.5 were considered to belong to class 1 whereas the other observations of response variables were made to belong to class 0. What can be interpreted from the output of this model is that these variables are vastly significant as far as their p-values are concerned, except for the variable Risk Mortality. Conceptually, this makes sense because a patient’s death risk should ideally not affect the amount that he/she is going to pay as medical expenses.

Upon testing the accuracy of the model on the new data, we obtained the following confusion matrix:

Financial Risk Prediction

0 1

0 12,568 3,189

1 5,026 7,298

The accuracy of this model was found out to be 70.66% with a misclassification rate of 29.34%. The accuracy of this logistic model is less than the decision tree model built previously, and thus to check if the Logistic model is any better for this data set, we decided to eliminate the unrelated variable of death mortality. The same training and testing data set was used again for this improved logistic model and we obtained the confusion matrix:

Financial Risk Prediction

0 1

0 12,568 3,189

1 5,026 7,298

The accuracy of this model improved slightly which came out to be 71.14% along with a misclassification rate of 28.86%. Although this accuracy was better than the logistic model with initial predictors, it was not comparable to the decision tree model accuracy. We decided to plot an AUC curve to plot the sensitivity versus specificity graph as follows:

The AUC curve that was obtained showed the area under curve to be 0.79, which is good but it still needs to be compared to other models to look for its significance. This AUC was obtained on the point (0.65, 0.76). Next, we will compare both these better models with another classification model to find which model gives the best predictions for this data set.

## Decision Tree

```{r Decision Tree}

library(tree)

library(ISLR)

library("readxl")

data<-read\_xlsx("Final Data\_Report 3.xlsx")

set.seed(1)

train = sample(1:nrow(data), nrow(data)/2)

colnames(data, do.NULL = FALSE)

colnames(data) <- c("TypeA","SourceA","Discharged","LengthStay","Age","RiskMortality","IllnessSev","FRisk","Beds")

data$FRisk<-as.factor(data$FRisk)

tree.data=tree(data$FRisk~.,data,subset=train)

summary(tree.data)

tree.data

plot(tree.data)

text(tree.data)

cv.data=cv.tree(tree.data)

plot(cv.data$size,cv.data$dev,type='b')

prune.data=prune.tree(tree.data,best=6)

plot(prune.data)

text(prune.data,pretty=0)

#the yhat plot is not functioning not sure if its necessaryl, xy lengths differ

yhat=as.numeric(predict(tree.data,newdata=data[-train,]))

data.test=data[-train,"FRisk"]

auc

roc

```

We believed the decision tree would be a good model due to its high interpretability and visualization capability. However, some downsides are its poor predictive ability, its high variance, and it is prone to overfitting. Because visualizing the variables was important to us, we chose decision tree over random forest despite the issues with variance and predictive ability. Using categorical data, a classification decision tree worked well at splitting nodes between low and high financial risk, highlighting the most important predictor variables and how they interrelate. After cleaning the data in the Texas Health Care Information Collection (THCIC) we loaded the set into RStudio and created the first decision tree using financial risk as our output variable. This is the initial tree we created prior to performing cost-complexity analysis.

Immediately, the most important predictive variables come to light and highlight how they are indicating for financial risk. Firstly, age appears to the most important predictor variable separating older patients (<5.5) as higher risk individuals. This is logical because as people age they are more susceptible to disease, especially expensive chronic conditions like cancer. Especially since high financial risk is classified as greater than $25,000 in charges, this includes patients that have large hospital bills to pay but also those that are not insured and are racking up large bills over time. Next, illness severity is the next important predictor. This makes sense because more severe illnesses typically involve expensive medications or surgeries to correct. If you are older, length of stay is critical to determining if your a risk (<15.5 days). This is also very logical because every day a patient takes up a bed in the hospital, he or she is racking up fees for chronic care. The next split is a little confusing because it suggests just being in the hospital for (<2.5 days) puts you at financial risk and appears to override the first rule. However, if you are in the hospital for less than 2.5 days, the type of admission appears to be the determining factor with being admitted through the emergency room as one that would indicate high risk. This makes a lot of sense since someone being admitted through the emergency room typically needs immediate attention and is most likely suffering from a severe illness that would require expensive treatment not to mention the cost of the ambulance on the way there.

Next we created a cost-complexity chart, which indicated that the lowest possible deviance would be a decision tree with 9 nodes. This chart can be seen below.

While this may be true, a simpler and more interpretable decision tree could be made at 8 nodes in size for a very small sacrifice in deviance. This also helps to prevent a model that suffers from overfitting to the data. Below is the decision tree we obtained from 8 nodes. Interestingly, this pruning process was actually a foliation process increasing the amount of nodes from 6 to 8. It appears the first tree produced was not fully grown.

This second model added a few extra predictors but these additional nodes actually provide little useful information. The tree attempted to further break up the categorical output with extra predictors but its not useful because those splits were already either 100 percent financial risk or not financial risk and therefore communicate zero extra information. The new tree produced most definitely suffered from overfitting, which is what one would expect increasing the nodes, which is confusing because the cost complexity analysis suggested to use 8 nodes.

Financial Risk Prediction

0 1

0 27,235 6,254

1 12,083 24,403

Financial Risk Prediction

0 1

0 27,432 6,458

1 11,886 24,199

The top correlation matrix is related to the 6 node tree and the bottom is related to the 8 node tree. The accuracy for the first model was .7379 and the second model was .7378. The second model also increased in type 1 error, classifying a patient as high risk when they are not, and decreased type 2 error, not classifying a high risk patient when they are. Failing to not classify a high risk patient when they are is a worse type of error in this instance so the model very slightly improved through pruning even though the accuracy overall remained basically the exact same.

## Model Comparison

The three different models created were logistic regression, KNN and decision tree. The logistic regression model gave an accuracy of about 70% with a good AUC of 0.79. In the decision tree model, pruning was conducted using the cost complexity analysis method to choose a simpler model with a low deviance. Even though the model with 9 leaves had a slightly lower deviance, the model with 8 leaves was chosen to avoid overfitting and to create a simpler model. This decision tree model gave an accuracy of 73%. The accuracies of the three models are displayed in the chart below. The KNN model showed the best accuracy rate which is 85% at K=7. The KNN model was also built for K= 10 but it had a higher misclassification error rate; therefore, K=7 was the best choice. Out of all the models built, KNN had the lowest error rate and the highest accuracy. Hence, the KNN model at K=7 is recommended for this prediction.

## Conclusion

The team was successful in using the Texas Department of State Health Services - Hospital Discharge Dataset to build several models to predict the financial risk of a patient. After the dataset was cleaned, different models were used to predict the outcome response using the predictors that were filtered by the variable selection method. The three models created were logistic regression, KNN, and decision tree. The models were compared and the best model predictors were determined. These were age, length of stay and severity of illness. Based on the accuracies obtained from confusion matrices, the best model was chosen and recommended. Out of all the models built, KNN had the lowest error rate and the highest accuracy. Hence, the KNN model at K=7 is recommended for this prediction. This model could be improved by using additional predictors like patient’s history and by expanding the data set to include data from hospitals nationwide. These improvements would result in a more extensive and accurate model.