Data Mining Project Analysis on Liver Disease Prediction using Machine Learning

ISM 6136 Data Mining

Dr. Kiran Garimella

University of South Florida

Final Group Project Team 13:

Maruthi Vinay Jampani

Venkata Satya Sai Lakshmi Lavanya Rekkala

Lavanya Chundu

Background of the problem:

Human blood acts as a carrier for parasites, harmful cells, carcinogenic substances, and tumours. The fact that blood from all over the body eventually circulates through liver makes it more prone to diseases. The hepatitis virus that enters the human blood through semen, contaminated needles, contaminated food, and water etc… is known to damage the liver tissue. Apart from the above-mentioned reasons, too much fat accumulated in the liver by excessive alcohol consumption and unhealthy food habits also causes liver inflammation. Liver diseases can also be caused by weak immune systems which again are a result of genetic disorders, diet issues, exposure to chemicals or drug overdoses. A major Source of reports suggest that liver diseases can be inherited from parents or other members in the family line too. Therefore, it is important to curb the number of cases in the current generation to prevent the risk of exposing our future generations to liver diseases.

The reports, articles published and cited by WHO, NCBI, CDC and several other major websites suggest an increasing trend in liver diseases. From the above paragraph, we can draw inferences that these increasing trends might be a result of increase in the alcohol consumption and unhealthy food habits. The improvement in technology has improved the accessibility to food and alcohol with the aid of online platforms, thereby, leading to an increase in food and alcohol consumption rates. Further, the reduction in the necessity to drive to office and, the increase in stress levels due to the lack of social life during the pandemic situation, the alcohol consumption rates and drug consumption rates even more increased. Therefore, during the pandemic, hospitals saw a rise in the alcohol related Liver diseases. In developing countries, most of the people consume cheap liquor, which is the main reason for more liver disease cases in developing countries. Also, people with liver diseases are subjected to high risk if in case they are tested positive for covid-19. WHO and several health experts already announced that the covid 19 situation is going to have a long-lasting impact on the world. Hence, it is high time that we address all our major health concerns and the reasons behind them. Even if the impact of covid 19 dies down, we might fail to bring down the alcohol consumption rates that cause liver problems as we are living in an era of door deliveries and online payments. As mentioned in the earlier paragraph, the fat cells that build up in the liver can also cause scarring of liver tissues. Hence, people who are obese are easily prone to liver diseases. However, we are living in a busy world where most of the people rarely find time for exercises, and this also creates a scope for liver diseases.

Motivation for solving the Problem:

People with liver diseases have less work productivity and a lot of them end up unemployed. Hence, the rise in liver diseases pushes up the attrition rate and this will have a bad effect on our economy. Further, the increase in such major health concerns will use up a large portion of government funds that are allocated for the national health insurance programs. The hike in liver diseases can turn out to be a major concern for private health insurance companies too. Most of the symptoms of liver disease remain hidden for a long time and, even if the symptoms do appear, they are initially very mild and this is because liver has a property to repair itself when the damage is small. So, People find it extremely difficult to recognise the

liver diseases in the initial stages and, by the time one realises and starts the treatment, some irreparable damage would have already been done. Therefore, most of the liver diseases end up damaging the liver to an extent that it fails to function properly, and, in some cases, it damages the entire liver which is a life-threatening situation. Some liver diseases might require a life-long treatment i.e., mostly liver diseases lead to chronic conditions. The damage caused by liver diseases and its negative impact on the economy especially on developing countries, motivated us to work on the prediction of liver diseases. Also, in the current situation, it deemed more appropriate to address the liver problems out of all the other health concerns because of the alarming increase in the number of liver disease cases. The effect of any health concern can be minimized by concentrating on prevention and cure. Disease recognition is the first step towards cure and this report describes the usage of data mining techniques in predicting the liver diseases using Liver function Tests results (series of blood tests). It discusses various algorithms used and their evaluation metrics to predict liver diseases in early stage.

Solution methodology:

After finalising the topic, we went through the symptoms, treatments, current diagnosis, and severity of illness caused by liver diseases to make sure that it is worth working on a liver disease dataset. UCI and Kaggle are deemed to be the most popular websites that provide datasets collected in real world scenarios. We have searched a lot of datasets, but some of them did not have the right attributes and some of them had very few rows. After going through many datasets, we eventually decided to work on the Indian Liver Patient Dataset (ILPD) as it has all the attributes taken from liver function tests (blood tests) and it has a decent size. All the attributes in the dataset were clearly examined before proceeding with further analysis using the machine learning algorithms. All the attributes of the dataset have been thoroughly explained in the later sections of this report.

Before the application of any data mining algorithm, it is important to clean the data and it is also important to determine the response variable. In our case, the response variable is the selector column, and it is a binary column that represents the existence or absence of a liver disease. Data cleaning includes filling of missing values and removal of irrelevant columns. For determining irrelevant columns, we used three techniques.

1. Determining the rate using manual calculation
2. Correlation plot
3. Anova analysis using R

Correlation plots and linear regression are one of the most common methods used for determining the impact of numerical explanatory variables on the response variable and Anova analysis is one of the most common methods used for finding the impact of categorical explanatory variables on the response variable. The gender column in our dataset is a categorical explanatory variable and using Anova analysis, we figured out that the gender has no impact on the response variable. Also, we calculated the positivity rate of women and men by using the formulae (Number of women having a liver disease/ Total number of women), (Number of Men having a liver disease/ Total number of Men) respectively and, there was

not much difference in the percentage of Women and Men having liver diseases. After proper analysis, we could conclude that the gender has nothing to do with developing or not developing a liver disease. Therefore, we obtained the final dataset for analysis by removing the Gender column from the ILPD dataset taken from UCI repository.

After data cleaning, we worked on determining the algorithms. As our problem is a classification one and output variable has two categories only, we have chosen "two class models" to predict the output. In this report we will be displaying the results obtained from the two-class boosted decision tree, Two Class Decision Forest and two class Neural networks.

Reasons For choosing Two Class Decision Tree:

Two class decision trees are easier to interpret than most of the other machine learning algorithms. Given the fact that our data is cleaned, and the size of our training dataset is decent, we did not have to worry about overfitting which is one of the most seen problems while choosing two class boosted decision trees.

Reasons For choosing Two Class Neural Networks:

Both Neural networks and boosted decision trees deal with non-linear relationships. "The neural networks" take the weightage of different factors into consideration in determining the response variable while "the two class boosted decision trees" do level by level classification using one factor after the other i.e., Neural networks have the capacity to do parallel processing. Neural networks once trained make faster predictions and can be used to deal with complex relationships and large datasets.

Reasons For choosing Two Class Decision Forest:

Two class decision Forest is applicable in most of the cases where Two class Decision tree is applicable. As per our understanding, in decision tree model, the next decision tree corrects the errors that were committed by the previous decision tree and therefore, usually the last decision tree is usually better than the previous ones and the same is used in decision making process. However, in Random Forest decisions are made based on multiple decision trees. The Decision Forest is known to work better on large datasets, and it works better usually on the linear data.

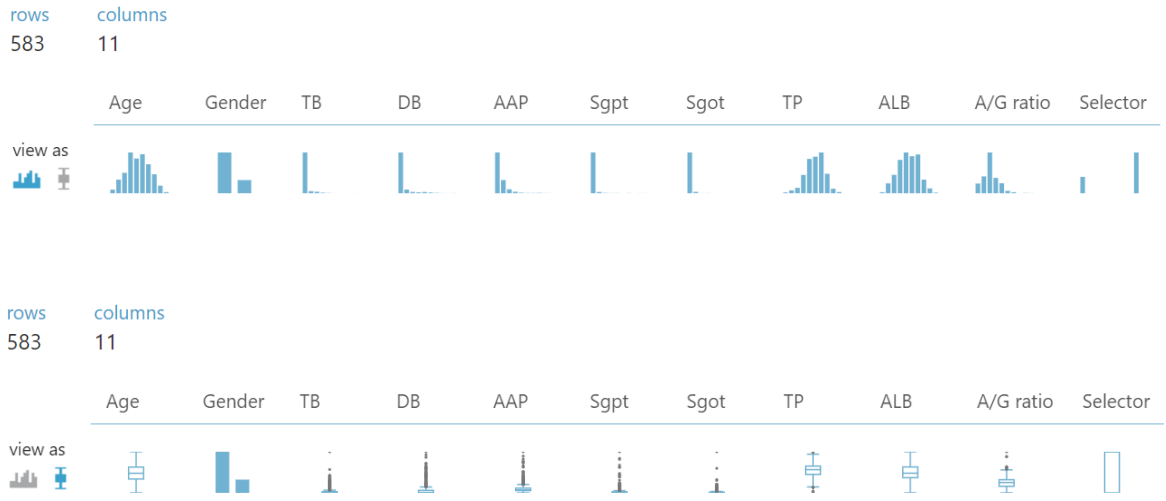
The nature of the dataset usually decides the metrics that can be used in the evaluation of performance. Initially our dataset was unbalanced and for unbalanced datasets, accuracy is not a measure of performance and hence we determined its performance using rest of the metrics. We initially ran our algorithms on an imbalanced dataset and later, we re-ran our algorithms on a balanced dataset that was obtained by applying SMOTE. We compared the metrics associated with the balanced dataset and unbalanced dataset. The comparison of different metrics and the conclusions drawn from these metrics are explained in the Metrics and Evaluation part of the report.

Description of Dataset:

This data set contains 416 liver patient records and 167 non liver patient records (total records is 583) and each variable is the measure of respective values in the Patient's blood sample. The data set was collected from northeast of Andhra Pradesh, India. Selector is a class label used to divide into groups (liver patient or not). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age "90". Clearly, our dataset is an imbalanced dataset and hence we have decided to not to consider Accuracy as a good evaluation metric for our model initially.

Reference Link:

https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)



Independent or Explanatory Variables:

Demographic variables:

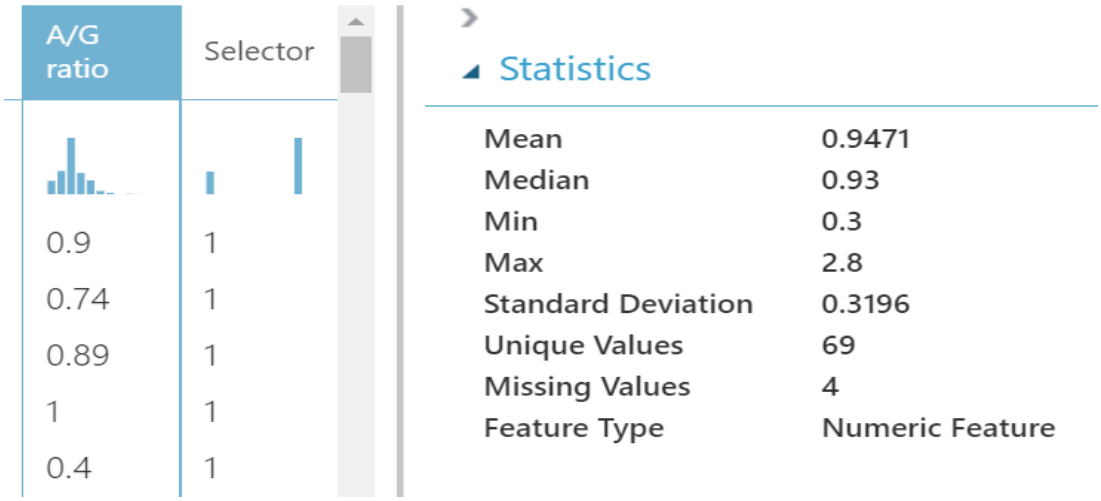
Age: Age of the Patient, Numerical variable which doesn't have any missing values.

Gender: Gender of the Patient, Categorical variable with two classes, Male and Female and no missing values.

Variables related to blood tests which are collected using liver function tests (LFTs):

TB: Total Bilirubin, Numerical variable with no missing values. High bilirubin levels can cause liver damage.
DB: Direct Bilirubin, Numerical variable with no missing values. Bilirubin attached by the liver to glucuronic acid, a glucose-derived acid, is called direct, or conjugated, bilirubin. Bilirubin not attached to glucuronic acid is called indirect, or unconjugated, bilirubin. All the bilirubin in your blood together is called total bilirubin.
AAP: Alkphos Alkaline Phosphatase, Numerical variable with no missing values. High alkaline

phosphatase levels may mean there is damage to your liver.
Sgpt: Alamine Aminotransferase, Numerical variable with no missing values. Very high level of SGPT in the blood can be an indication of damage or problems related to the liver.
Sgot: Aspartate Aminotransferase, Numerical variable with no missing values. SGPT and SGOT are certain enzymes that are produced by the liver and its cells. Elevated SGPT and SGOT levels are an indication of liver cell injury or damage.
TP: Total Proteins, Numerical variable with no missing values. A total protein test measures the combined sum of all the different proteins in the blood. It is used to know if you have unexpected weight loss, fatigue, or the symptoms of a kidney or liver disease.
ALB: Albumin, Numerical variable with no missing values. The most common protein in the blood is albumin, which prevents fluid from leaking out of the blood and carries substances through the body.
A/G Ratio: Albumin and Globulin Ratio, Numerical variable with four missing values. High A/G ratio can be a sign of disease in your liver, kidney, or intestines.



Here, we have a total of 4 missing values in A/G ratio filed. 2 missing values of A/G ratio filed for liver disease patient have been replaced with the mean of available A/G ratio values of Liver disease patient records (mean of 414 records of liver disease patients) and the other 2 missing values for Non-Liver Disease patient, have been replaced with the mean or average of the existing A/G ratio values of non-liver disease patient (mean of 165 records of non-liver disease patients).

Age	Gender	TB	DB	AAP	Sgpt	Sgot	TP	ALB	A/G ratio	Selector
45	Female	0.9	0.3	189	23	33	6.6	3.9	missing	1
51	Male	0.8	0.2	230	24	46	6.5	3.1	missing	1
35	Female	0.6	0.2	180	12	15	5.2	2.7	missing	0
27	Male	1.3	0.6	106	25	54	8.5	4.8	missing	0

Age	Gender	TB	DB	AAP	Sgpt	Sgot	TP	ALB	A/G ratio	Selector
45	Female	0.9	0.3	189	23	33	6.6	3.9	0.914	1
51	Male	0.8	0.2	230	24	46	6.5	3.1	0.914	1
35	Female	0.6	0.2	180	12	15	5.2	2.7	1.03	0
27	Male	1.3	0.6	106	25	54	8.5	4.8	1.03	0

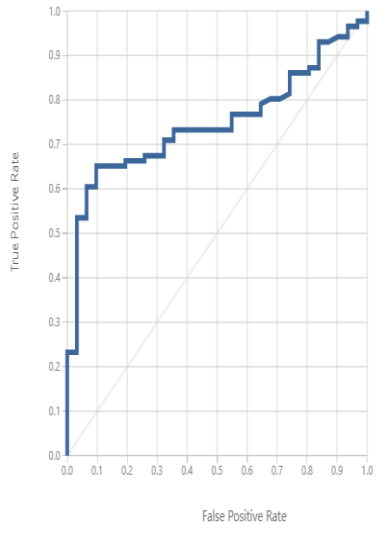
Response/Dependent Variable:

Selector: This field is used to distinguish between liver disease patient and non-Liver disease Patient (1 is used for Liver disease Patient and 0 is used for non-Liver disease Patient).

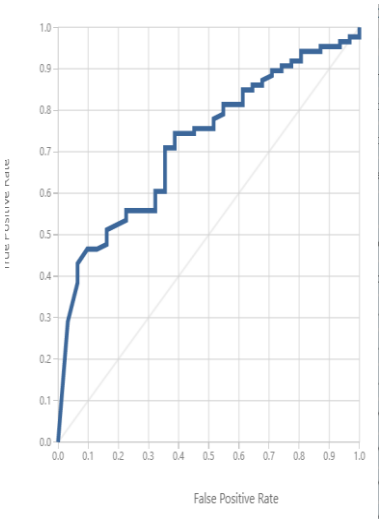
Comparison of Machine Learning Algorithms and Evaluation Metrics:

We split the data into training and testing datasets using ideal 80-20 rule. As our output (selector column) is binary, we have chosen different two class algorithms to understand the performance of models. We used three different two class algorithms here, but we got better results with the Two Class Boosted Decision Tree.

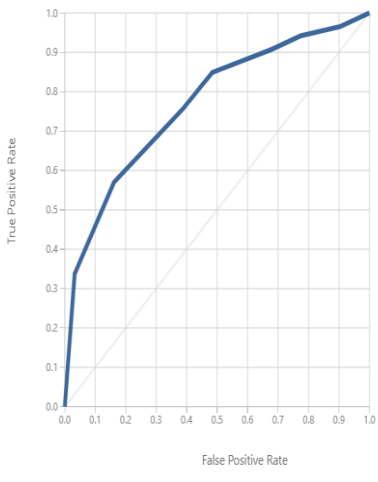
Two class Neural Network



Two Class Boosted Decision Tree



Two class Decision Forest



Below are the metrics for the three algorithms:

Algorithm	TP	FN	FP	TN	Accuracy	Precision	Recall	f1 score	AUC
Two-class decision forest	67	19	18	13	0.684	0.788	0.779	0.784	0.692
Two-class neural network	85	1	31	0	0.726	0.733	0.988	0.842	0.762
two class boosted decision tree	70	16	21	10	0.684	0.769	0.814	0.719	0.707

Our dataset was initially unbalanced (72% -positive (1), 28% - Negative (0)) and hence, instead of accuracy, we chose to look at other metrics such as AUC, FN, FP, TP, TN etc. to determine the best model.

Though the Two Class Neural Network has highest AUC, the FN and TN are negligible when compared to the TP and FP. From this, it is evident that the Neural Network is biased towards the positives in the given data. Therefore, despite high AUC, we stopped considering the Neural Network to be the best model. As per our results, out of the remaining two algorithms, two class Boosted decision tree has high AUC. So, we can say that the Two Class Boosted Decision Tree performs better than the remaining two models.

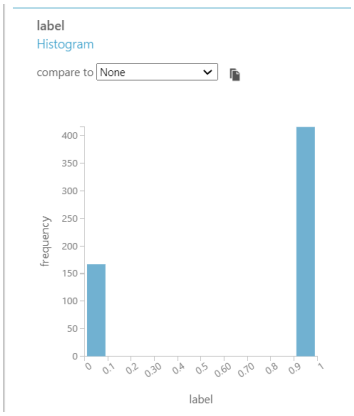
Over sampling technique:

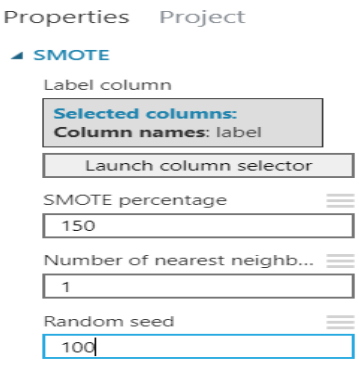
As our data is unbalanced, all the algorithms are biased towards positive label as our dataset has 70% positive cases. We wanted to remove this bias by making the dataset balanced. So, we have removed random positive cases (randomly selected a sample) from the current dataset in such a way that the data becomes balanced i.e., we have removed some rows which are positive.

Although, the above process removes the Bias to some extent, we are at risk of losing statistical power of our model. Therefore, as suggested by our professor, we have used the oversampling technique to create the neighbours of the negative cases and as a result we ended up having a balanced dataset without losing a lot a valid data.

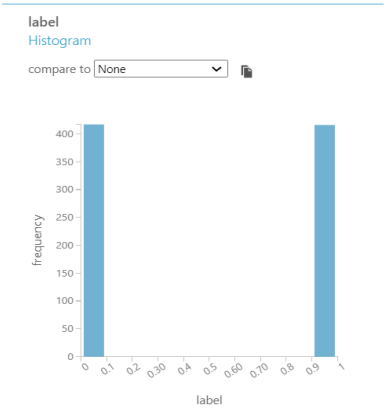
We have used SMOTE in azure ML studio for performing over sampling. We have used 150% as SMOTE percentage to make the dataset balanced.

"label/classifier" statistics Before SMOTE





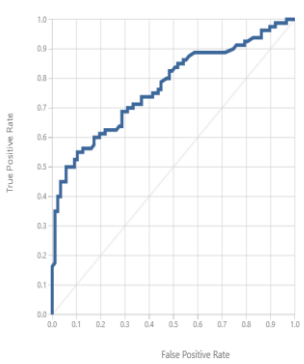
"label/classifier" statistics after SMOTE

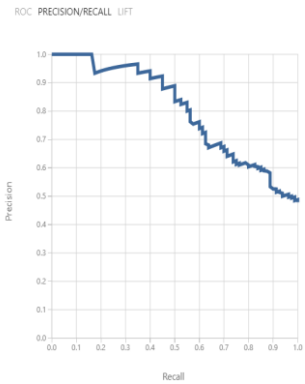


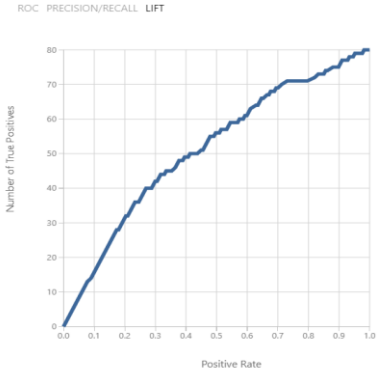
Below are the metrics for the three algorithms after applying SMOTE:

With Default parameters for the algorithms:

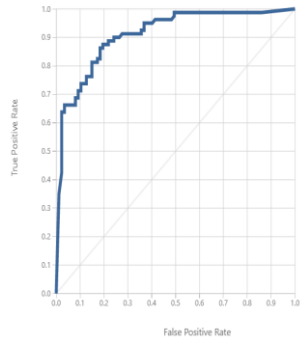
Two class Neural Network :

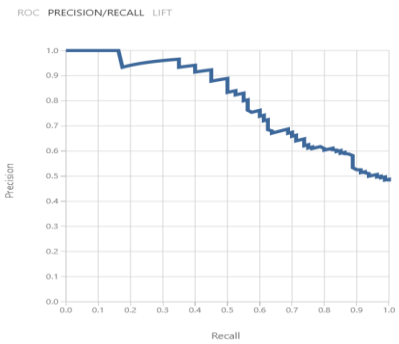


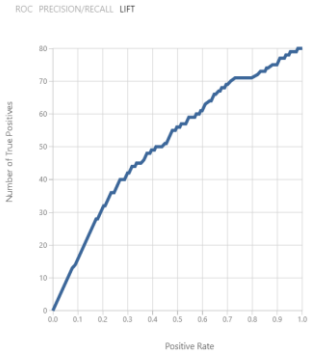




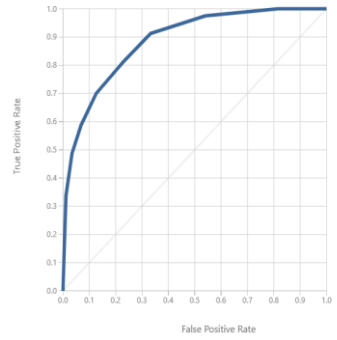
Two Class Boosted Decision Tree:

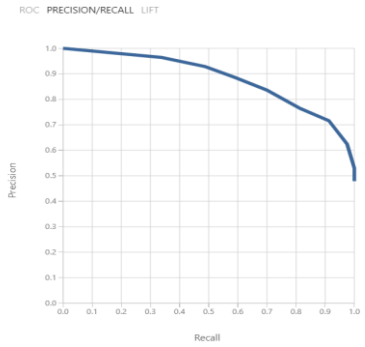


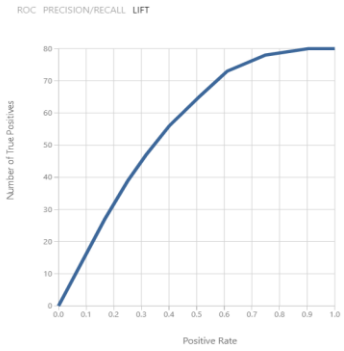




Two Class Decision Forest:



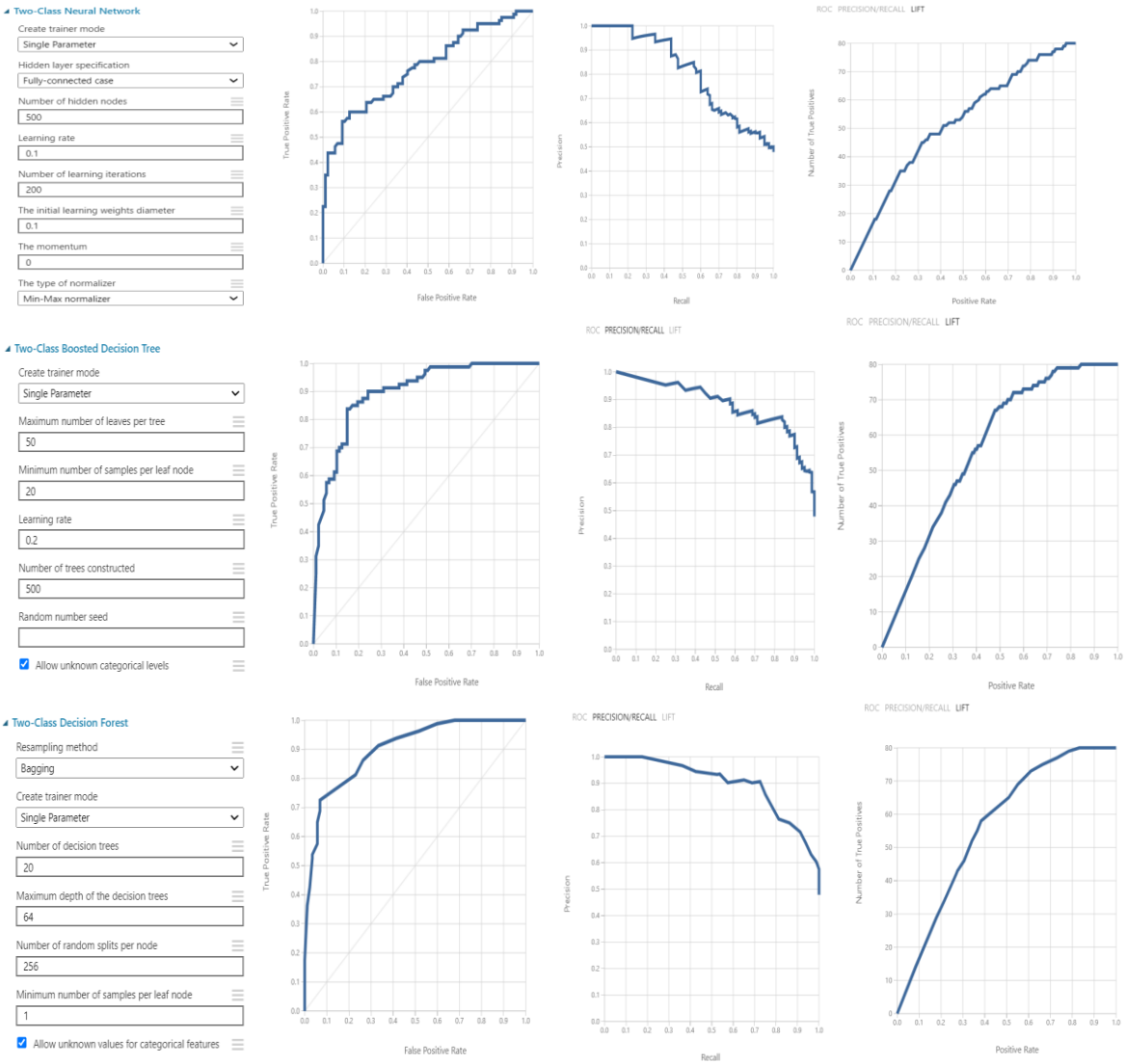




Algorithm	TP	FN	FP	TN	Accuracy	Precision	Recall	f1 score	AUC
Two-class decision forest	56	24	11	26	0.79	0.836	0.7	0.762	0.884
Two-class neural network	48	32	16	71	0.713	0.75	0.6	0.667	0.771
two class boosted decision tree	69	11	16	71	0.838	0.812	0.863	0.836	0.912

With parameters tweaked:

Two class Neural Network



Algorithm	TP	FN	FP	TN	Accuracy	Precision	Recall	f1 score	AUC
Two-class decision forest	60	20	10	77	0.82	0.857	0.75	0.8	0.9
Two-class neural network	48	32	18	69	0.701	0.727	0.6	0.658	0.776
two class boosted decision tree	67	13	13	74	0.844	0.838	0.838	0.838	0.897

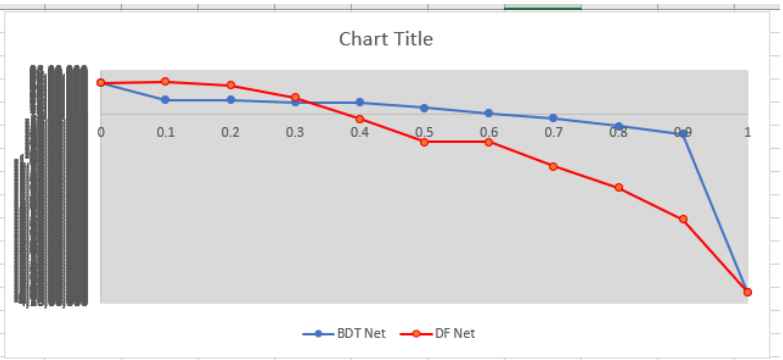
There is not much impact of tweaking on the accuracy and other metrics for the Neural network and the boosted decision tree but there is a small change in the metrics of the Decision Forest.

Cost benefit analysis:

We can move the threshold above or below and unlock business value by doing threshold analysis.

As Boosted Decision Tree and Decision Forest showed better results than the Neural Network, we performed the cost-benefit analysis on these two algorithms for different threshold values. Below is the summary of the cost-benefit analysis.

Threshold	BDT Net	DF Net
0	71,30,000	71,30,000
0.1	33,70,000	73,70,000
0.2	32,90,000	65,30,000
0.3	27,20,000	**38,10,000**
0.4	**27,50,000**	-8,65,000
0.5	15,95,000	-61,30,000
0.6	3,95,000	-61,30,000
0.7	-7,60,000	-1,14,55,000
0.8	-25,60,000	-1,62,10,000
0.9	-43,15,000	-2,33,80,000
1	-3,95,65,000	-3,95,65,000
At Maximum Net:	**BDT**	**DF**
Accuracy	0.83	0.79



we got good results with the Boosted decision tree, and at 0.4 threshold value, we can earn good profit without compromising on the accuracy and other metrics.

Conclusion:

Existing process of typical healthcare organization looks like this:

After preliminary diagnosis, based on symptoms, Doctors prescribe various blood tests for the patient and based on the amount or measures of various variables like sgpt, sgot and proteins in the blood, Doctors decide whether patient's liver is diseased or not.

In the second step/opinion, he/she needs to take multiple imaging tests like CT scan, Biopsy, and endoscopy to see if the liver is recoverable with medication or transplant is the only option. So, if there is a chance to add imaging tests data as new variables to our existing model, we can take our model to next level to predict if liver transplant is needed or not for a patient and this will be the future of the model and it becomes a revolution.

Initially, Our Model can assist the doctors to have 2nd opinion as even doctors can commit human errors sometimes and hospitals may incur huge penalties. As time goes, Doctors can input the real time data into the model, and this will help the model to learn from various examples and eventually the accuracy of the model will improve much more. Once the model become more accurate, we can deploy this in the corporate health chains, this helps the management to reduce the workforce in terms of working hours of doctors and it can be a huge milestone. If we can reduce the liver disease result wait time, it contributes to the good reputation of the health chains, and we believe that our model will be able to do that in the future once it learns with more examples from the real time data fed into the model by doctors.

References:

https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

www.nbcnews.com

https://pubmed.ncbi.nlm.nih.gov/30333543/

https://www.cdc.gov/nchs/fastats/liver-disease.htm

www.who.int