A Logit Regression Model for the Prediction of Liver Ailments

Project Report 3: MBA652A (Statistical Modeling For Business Analytics)

Submitted by:

Nitesh Sharma(20114013)

Submitted to:

Prof. Devlina Chatterjee
Department of Industrial & Management Engineering
Indian Institute of Technology Kanpur

TABLE OF CONTENTS

1. INTRODUCTION

Precise diagnosis of disease symptoms can help in proper and suitable treatment of the body ailments. Contaminated food or water, poor life style (smoking, alcohol or other forms of intoxications), presence of comorbidities are the few factors that could led to the liver ailments. Devising a model that could precisely predict the state of being a "patient" or "non-patient" could help the field of medicine in myriad ways and will also be very beneficial for the subjects undergoing diagnosis or treatments.

2. OBJECTIVE

In this project, we've attempted to create a model that could predict whether the subject is suffering from liver ailment or not (based on the outcomes of various relevant pathological parameters obtained from blood test).

3. METHODOLOGY

3.1. Data collection and its description

Dataset used in this study is a secondary dataset, originally it was collected from the north east of Andhra Pradesh; India. This data set contains information for 583 subjects (416 patients and 167 non patients; 441male and 142 female) [**1, 2**]. Authors made it publically available on machine learning repository of the University of California, Irvine [**3**].

Information corresponding to 4 subjects was incomplete and thus those subjects were not included in this study for further analysis. Hence, the actual numbers of observations in this study are 579.

Data Description

The output (Dependent variable) is as follow:

Y= result, which is liver patient or not liver patient

The inputs (Independent variables) are as follows:

X1 = age: refers to the Age of the subject
X2 = gender: refers to the Gender of the subject
X3 = tb: Total Bilirubin level in blood.
X4 = db: Direct Bilirubin level in blood.
X5 = alkphos: Alkaline Phosphotase level in blood.
X6 = sgpt: Alanine Aminotransferase level in blood.
X7 = sgot: Aspartate Aminotransferase level in blood.
X8 = tp: Total Protein level in blood.
X9 = ab: Albumin level in blood.
X10 = agratio: Albumin-Globulin Ratio level in blood.

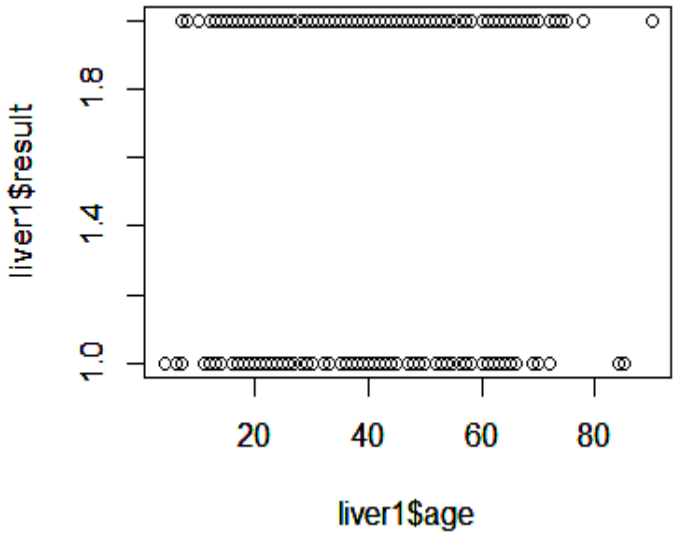
3.2. Descriptive Statistics of Data

Descriptive statistical data was obtained to observe the distribution of independent variables (in numerical terms). Descriptive statistical parameter indicates towards the presence of outliers in the dataset.

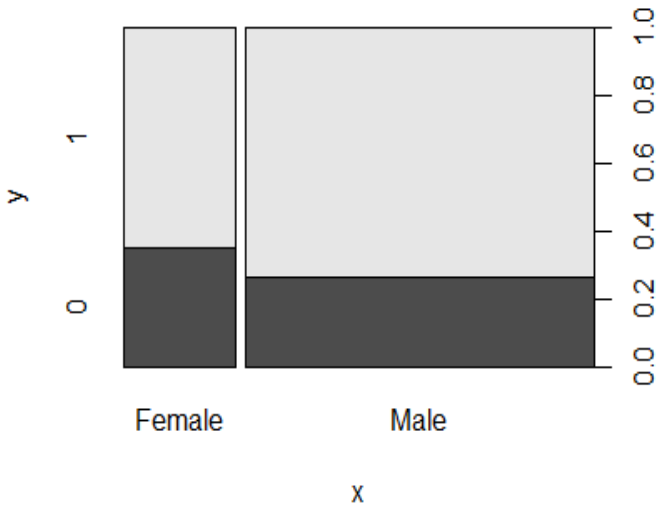
Variables	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.	Std. Deviation
age	4.0000	33.000	45.000	44.780	58.000	90.000	16.2217
tb	0.4000	00.800	01.000	3.3150	2.600	75.000	6.22771
db	0.1000	00.200	00.300	1.4940	01.300	19.700	2.81649
alkphos	63.000	175.50	208.00	291.40	298.00	2110.0	243.561
sgpt	10.000	23.000	35.000	81.130	61.000	2000.0	183.182
sgot	10.000	25.000	42.000	110.40	87.000	4929.0	289.850
tp	2.7000	5.8000	6.6000	6.4820	7.2000	9.6000	1.08464
ab	0.9000	2.6000	3.1000	3.1390	3.8000	5.5000	0.79443
agratio	0.3000	0.7000	0.9300	0.9471	1.1000	2.8000	0.31959

3.3. Data Exploration: Scatter Plots

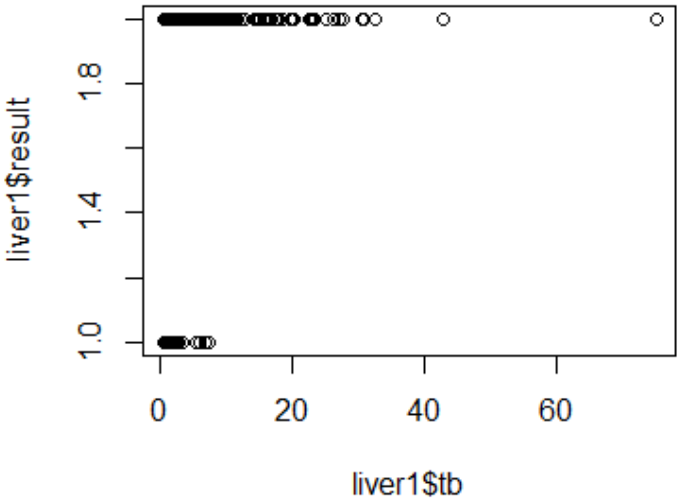
Scatterplots were generated to visually observe the distribution of the independent variables. Visuals from scatterplot confirm the presence of outliers; however, upon visualizing their distribution pattern it was decided not to drop them from dataset. Hence dataset will be used for the further analysis without any specific treatments for outliers.



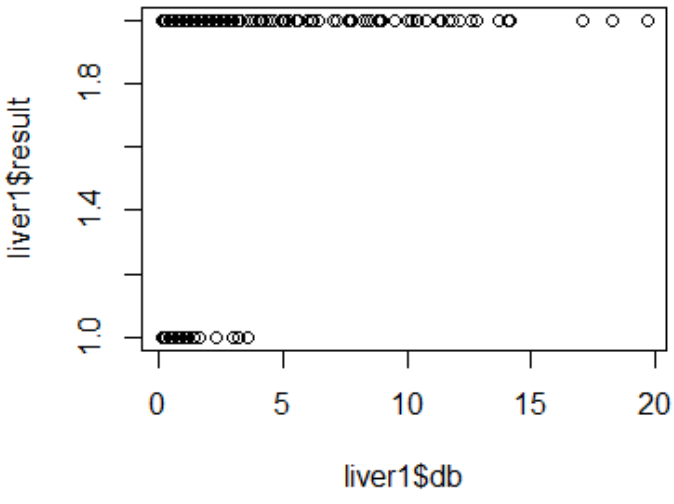
age vs **result**



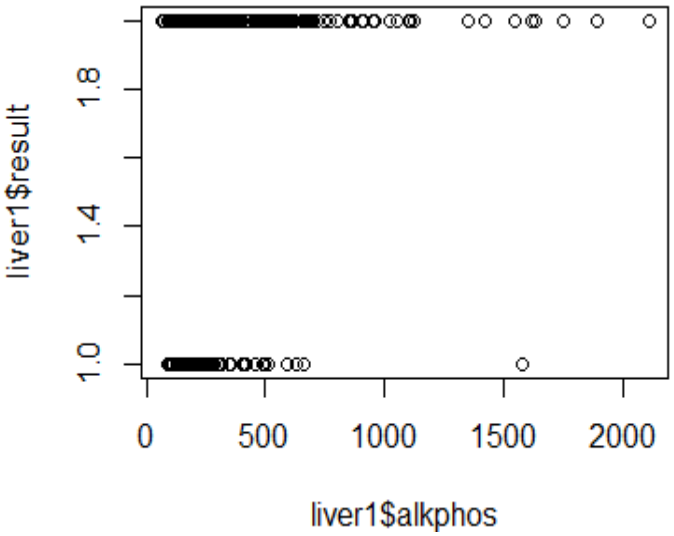
gender vs **result**



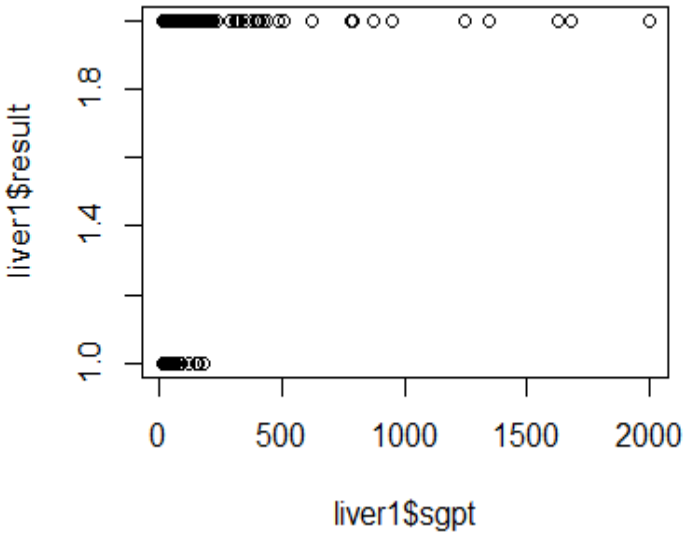
tb(Total Bilirubin) vs **result**



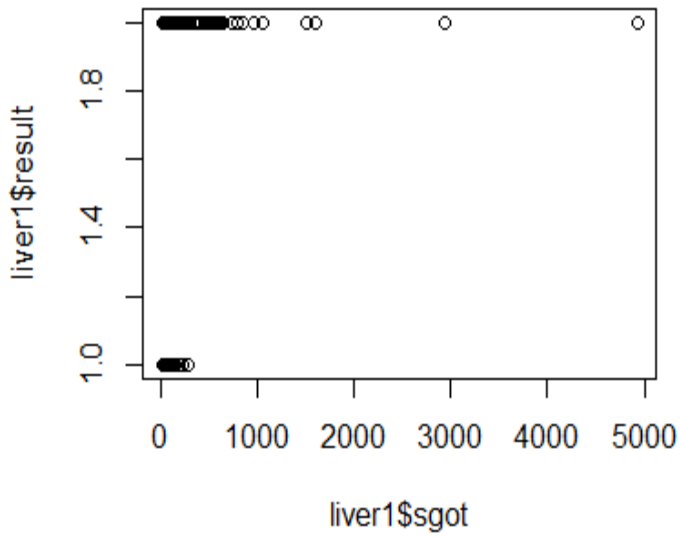
db(Direct Bilirubin) vs **result**



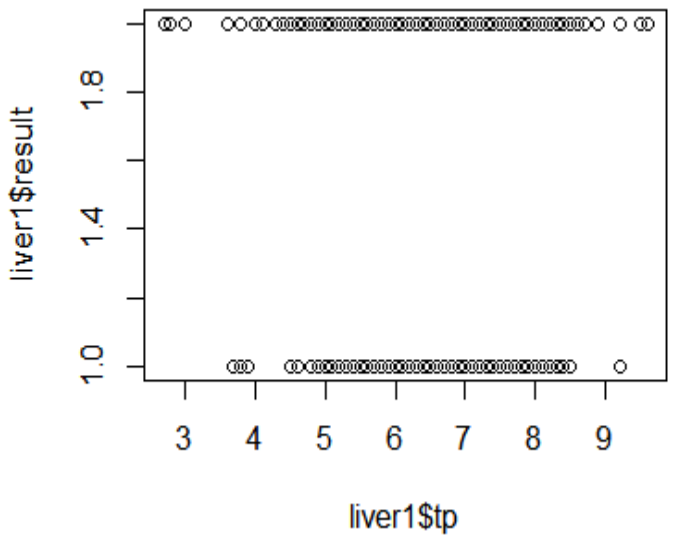
alkphos(Alkaline Phosphotase) vs **result**



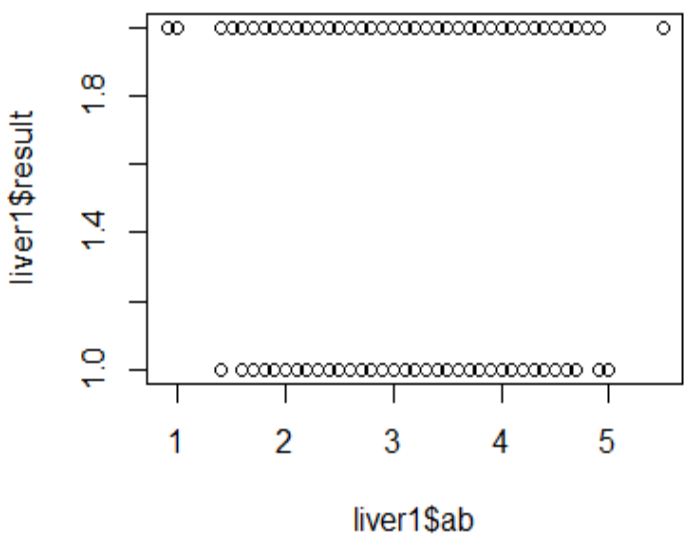
sgpt(Alamine Aminotransferase) vs **result**



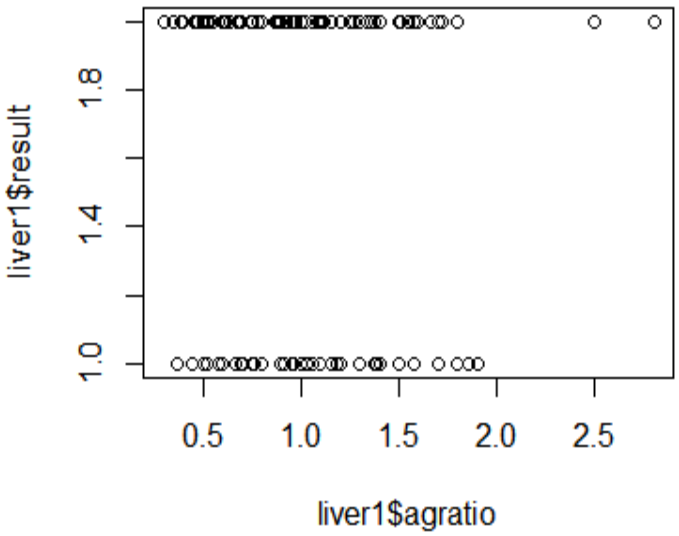
sgot(Aspartate Aminotransferase) vs **result**



tp(Total Proteins) vs **result**



ab(Albumin) vs **result**



agratio(Albumin and Globulin Ratio) vs **result**

3.4. Data Partitioning

Given data set was divided into the ratio of 3:1 (3 parts for training and 1 part for testing). Decision for such division was simply based on a thumb rule.

3.5. Logistic regression analysis:

As the dependent variable in our dataset is a binary (1: Patient, 0: Non-Patient) variable, "Logit" or "Probit" regression model are non-linear regression model designed specifically for binary dependent variables [**4**]. "Logit" and "Probit" regression models are known to produce similar results and there is no defined criterion to select one over other. We've decided to use "Logit" model for our study as many researchers have used it for similar analysis.

MODEL 1

result ~ age + gender + tb + db + alkphos + sgpt + sgot + tp + ab + agratio

VIF values (threshold = 10): Model 1

VIF Table									
Age	gender	Tb	db	alkphos	sgpt	sgot	tp	ab	agratio
1.069	1.052	2.079	2.215	1.138	1.949	1.918	18.056	33.783	10.408

The VIF value for '**ab**' is very high which implies that it is highly correlated with other independent variables. Thus, '**ab**' will not be considered for the further analysis.

MODEL 2

result ~ age + gender + tb + db + alkphos + sgpt + sgot + tp + agratio

VIF Values (threshold = 10): Model 2

VIF Table								
Age	gender	Tb	db	alkphos	sgpt	Sgot	tp	agratio
1.074	1.048	2.146	2.267	1.134	1.871	1.883	1.116	1.123

All the VIF values are < 10 (threshold limit), this indicates that independent variables are uncorrelated to each other.

Result Summary: Model 2

Coefficients	Beta	Std. error	z value	P value	Confidence Interval	
					2.5 %	**95 %**
(Constant)	-0.66397	0.96178	-0.690	0.49	-2.54904	1.22109
age	0.01791	0.00728	2.458	0.0140 *	0.00363	0.03219
gender	0.10473	0.26741	0.392	0.6953	-0.41939	0.62885
tb	0.01392	0.12395	0.112	0.9106	-0.22903	0.25686
db	0.62719	0.33620	1.866	0.0621 .	-0.03175	1.28612
alkphos	0.00163	0.00097	1.677	0.0935 .	-0.00027	0.00354
Sgpt	0.00988	0.00554	1.780	0.0751 .	-0.00099	0.02075
Sgot	0.00115	0.00341	0.336	0.7368	-0.00555	0.00784
tp	-0.00451	0.11387	-0.040	0.9684	-0.22770	0.21866
agratio	-0.69985	0.42298	-1.655	0.0980 .	-1.52889	0.12919
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Model Fitness: Model 2

Null Deviance	Residual Deviance	PseudoR square (McFadden)	AIC
519.07	424.21	0.1827583	444.21

Remarks:

As "**tp**" and "**tb**" are statistically most insignificant, they will not be considered for the further analysis.

MODEL 3

result ~ age + gender + db + alkphos + sgpt + sgot + agratio

VIF Values (threshold = 10): Model 3

VIF Table						
age	gender	db	alkphos	sgpt	sgot	agratio
1.052	1.037	1.119	1.122	1.859	1.863	1.069

All the VIF values are < 10 (threshold limit), this indicates that independent variables are uncorrelated to each other.

Results Summary: Model 3

Coefficients	Beta	Std. error	z value	P value	Confidence Interval	
					2.5%	**97.5%**
(Constant)	-0.68572	0.66666	-1.029	0.30368	-1.99236	0.62092
age	0.01795	0.00721	2.489	0.01281 *	0.00381	0.03208
gender	0.10588	0.26599	0.398	0.69059	-0.41545	0.62721
db	0.65535	0.23584	2.779	0.00546 **	0.19309	1.11759
alkphos	0.00163	0.00096	1.683	0.09230 .	-0.00026	0.00352
sgpt	0.00986	0.00553	1.781	0.07486 .	-0.00098	0.02071
sgot	0.00116	0.00340	0.340	0.73409	-0.00551	0.00782
agratio	-0.70423	0.41262	-1.707	0.08787 .	-1.51295	0.10448
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Model Fitness

Null Deviance	Residual Deviance	PseudoR square (McFadden)	AIC
519.07	424.23	0.1827220	440.23

Remarks:

PseudoR square remains almost the same as the previous model, however, improvements were observed in the AIC value which is indicative of the improvements in model fitness.

As "**gender**" and "**sgot**" are statistically insignificant, they'll not be considered for further analysis.

MODEL4

result ~ age + db + alkphos + sgpt + agratio

VIF Values (threshold = 10): Model 4

VIF Table				
Age	db	alkphos	sgpt	agratio
1.048	1.064	1.119	1.111	1.063

All the VIF values are < 10 (threshold limit), this indicates that independent variables are uncorrelated to each other.

Result Summary: Model 4

Coefficients	Beta	Std. error	z value	P value	Confidence Interval	
					2.5%	**97.5%**
(constant)	-0.64081	0.65735	-0.975	0.32964	-1.92919	0.64757
age	0.01816	0.00718	2.526	0.01153 *	0.004070	0.03225
db	0.68334	0.23226	2.942	0.00326 **	0.228118	1.13857
alkphos	0.01122	0.00096	1.687	0.09169 **.**	-0.00026	0.00353
sgpt	0.01122	0.00433	2.592	0.00953 **	0.00273	0.01971
agratio	-0.69138	0.41083	-1.683	0.09240 **.**	-1.49659	0.11383
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Model Fitness

Null Deviance	Residual Deviance	PseudoR square (McFadden)	AIC
519.07	424.51	0.1821708	436.51

Remarks:
AIC value decreased further, indicating improved model fitness; PseudoR square remains almost constant.

All variables, except "**agratio"** and "**alkphos**" were statistically significant. Upon due consideration, "**agratio**" was dropped from the further analysis.

MODEL 5

result ~ age + db + alkphos + sgpt

VIF Values (threshold = 10): Model 5

VIF Table			
Age	Db	alkphos	sgpt
1.03292	1.05363	1.08146	1.10271

All the VIF values are < 10 (threshold limit), this indicates that independent variables are uncorrelated to each other.

Result Summary: Model 5

Coefficients	Beta	Std. error	Z value	P value	Confidence Interval	
					2.5 %	97.5 %
Intercept	-1.49122	0.43125	-3.458	0.00054 ***	-2.33646	-0.64599
age	0.01988	0.00709	2.801	0.00508 **	0.00597	0.03378
db	0.73318	0.23444	3.127	0.00176 **	0.27367	1.19267
alkphos	0.00195	0.00098	1.965	0.04937 *	5.28e-06	0.00388
sgpt	0.01086	0.00425	2.556	0.01060*	2.53e-03	0.01919
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Model Fitness

Null Deviance	Residual Deviance	PseudoR square (McFadden)	AIC
519.07	427.33	0.1767400	437.33

Goodness of Fit Test

Goodness of Fit Value: **5.618425e-19 (fits well)**

Remarks:

Though AIC value increased slightly and there is a slight decrease in the Pseudo square value, however, all the independent variables are statistically significant in this model and from "Goodness of Fit Test" it was observed that model fits well. Thus, the outcome of the "MODEL 5" is accepted as a final result and the regression model based on this result is given as:

Pr(Y = 1|age, db, alkphos, sgpt**)**

= F(-1.49122 + 0.01988 **age** + 0.73317 **db** + 0.00194 **alkphos** + 0.01086 **sgpt**)

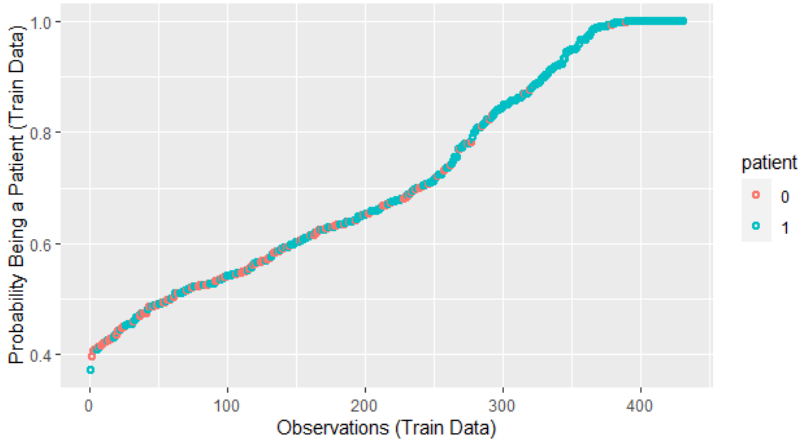
(0.43125) (0.00709) (0.23444) (0.00098) (0.00425)

$$= \frac{1}{1+e^{-(-1.49122 + 0.01988\,\mathbf{age} + 0.73317\,\mathbf{db} + 0.00194\,\mathbf{alkphos} + 0.01086\,\mathbf{sgpt})}}$$

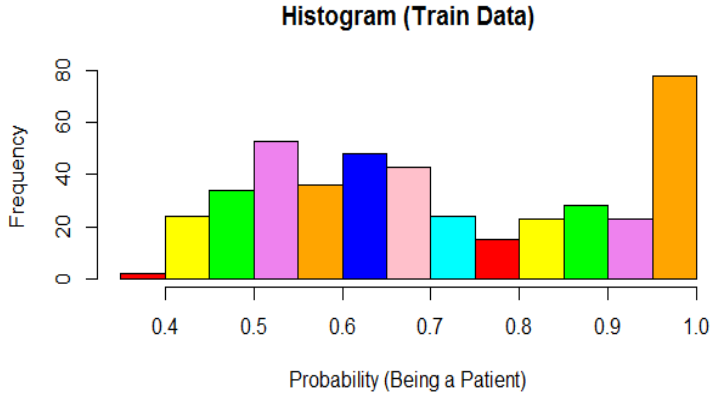
4. APPLICATION & RESULTS

4.1. Application of the Model on Train Data

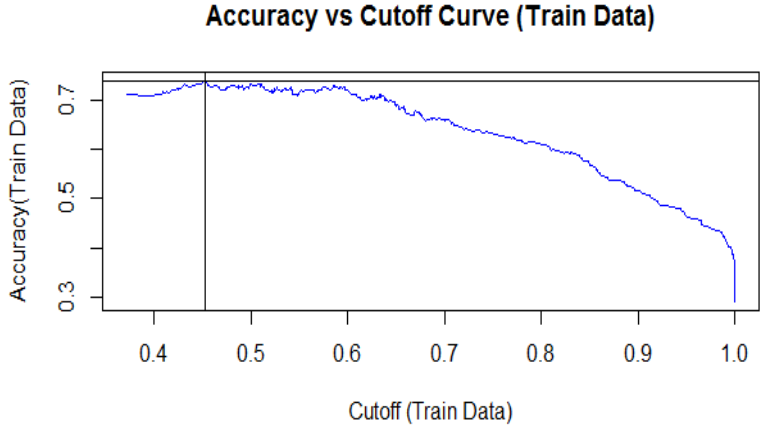
Predicted probabilities (based on Model 5) for "subjects" from "train data" is shown here below:



Probability (being a patient) and its frequency distribution for "subjects" from "train data" are plotted below:



Minimum probability (being a patient) lies somewhere around 0.4 and goes up to 1. In next step, attempts were made to find the most suitable cutoff value for the classification of "subjects" as patient or non-patient.



From "Accuracy vs Cutoff curve", it was observed that the highest **accuracy** of **0.7378** occurs at the **cut off value** of **0.4519** in train data.

Train Data Confusion matrix
(*based on the cut off value from "Accuracy vs Cutoff curve" of "train data"***)**

		Actual	
		*Not Liver Patient **(0)***	*Liver Patient **(1)***
Predicted	*Not Liver Patient **(0)***	19	7
	*Liver Patient **(1)***	106	299

Model evaluation

Parameter	**Value**	**Formulae**
Accuracy	0.7378	(TP+TN))/total
Misclassification rate	0.2621	(FP+FN)/total
Recall (sensitivity)	0.9771	TP/(TP+FN)
Precision	0.7382	TP/(TP+FP)

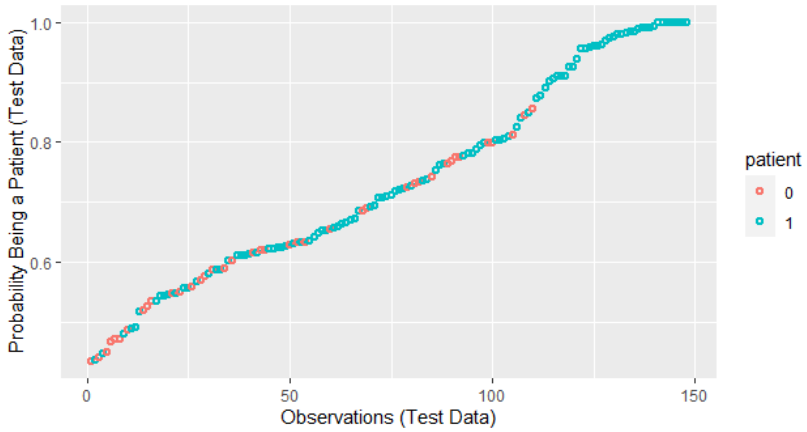
***TP**-True Positive, **TN**-True Negative, **FP**-False Positive, **FN**-False negative*
***Total**-TP+TN+FP+FN (sum of all observation)*

*"**Recall**"* is the most important evaluation parameter from our objective point of view and the model performance is fairly good (value = **0.9771**) in this regard.

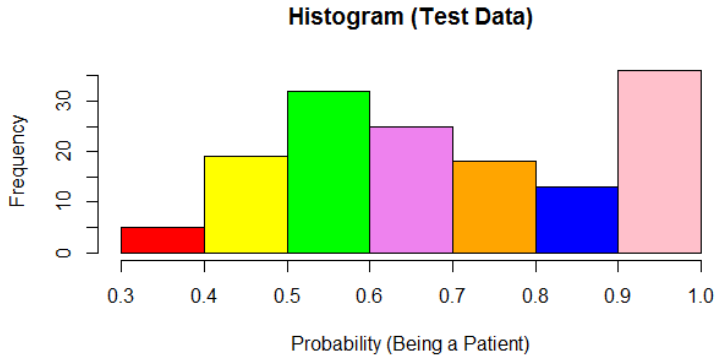
"*Precision*" is the 2nd most important evaluation parameter and has relatively (compared to "Recall") lesser value (value = **0.7382**); this could be because of the non-inclusion of variables that affects liver health.

4.2. Application of Model on Test Data

Predicted probabilities (based on Model 5) for the "subjects" from "test data" is shown here below:



Probability (being a patient) and its frequency distribution for "subjects" from "train data" are plotted below:



Test data Confusion matrix
(*based on* the cut off value from "*Accuracy vs Cutoff curve*" of "*train data*")

		Actual	
		Not Liver Patient ***(0)***	*Liver Patient* ***(1)***
Predicted	*Not Liver Patient* ***(0)***	5	4
	Liver Patient ***(1)***	35	104

Model evaluation

Parameter	**Value**	**Formulae**
Accuracy	0.7364	(TP+TN))/total
Misclassification rate	0.2635	(FP+FN)/total
Recall (sensitivity)	0.9541	TP/(TP+FN)
Precision	0.7482	TP/(TP+FP)

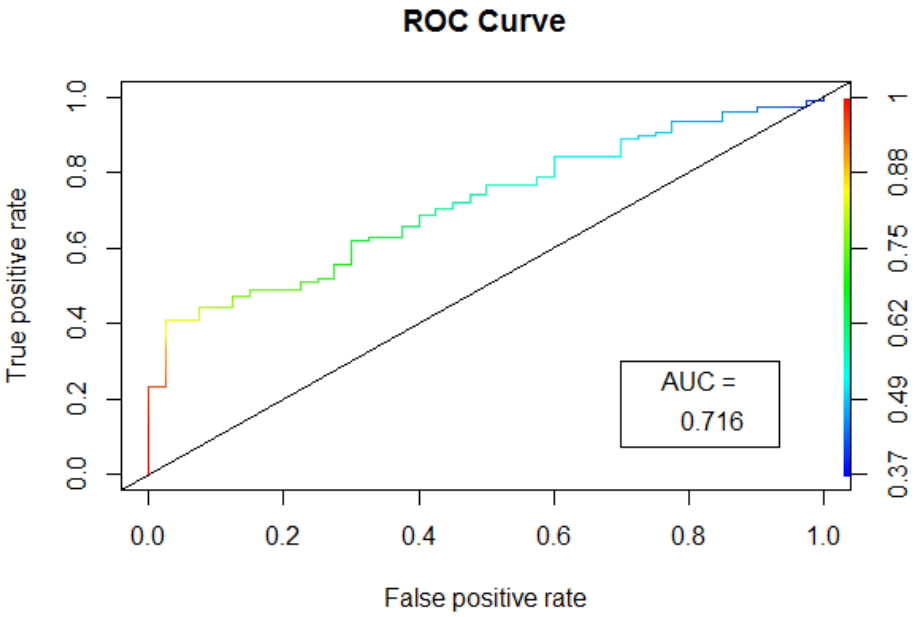
***TP**-True Positive, **TN**-True Negative, **FP**-False Positive, **FN**-False negative*
***Total**-TP+TN+FP+FN (sum of all observation)*

Model evaluation parameters for "**test data**" are **very similar to** the parameters of "**train data**"; which is indicative of a **statistical fit**.

5. ROC (Receiver Operating Characteristic) CURVE

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings. Higher the Area under the curve (AUC) better is the model at predicting the **1's** as **1's** and **0's** as **0's**.

The diagonal line (AUC = 0.5) represents the no-skill model which means model has no skill to predict the right values whereas the more we move to the north-west of the curve, the skill of our model improves which means it can predict more accurately.



In our case, the area under the curve for test data is **0.716**(> 0.5), which is a fairly decent value

6. CONCLUDING REMARKS

In our opinion, the model developed in this study performed decently, however, this model is based purely on the pathological parameters obtained through blood test. We believe that the performance of model can improve if other explanatory variables (like food preferences, consumption of liquor, physical activity level, heredity, quality of drinking water available etc.) are included.

Our model is based on the data collected from a particular region of Andhra Pradesh; it might not do well if the test data is from some other region. Therefore, in order to develop a robust model, train data should also have a geographical diversity.

7. REFERENCES:

1. Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2011). A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, *3*(2), 101-114.
2. Ramana, B. V., Babu, M. S. P., & Venkateswarlu, N. B. (2012). A critical comparative study of liver patients from USA and INDIA: an exploratory analysis. *International Journal of Computer Science Issues (IJCSI)*, *9*(3), 506.
3. Machine learning repository; University of California, Irvine https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)
4. Stock, J. H., & Watson, M. W. (2015). *Introduction to econometrics*.

APPENDIX I: R CODE

```
>library(readxl)
>liver<- read_excel("liver.xlsx")
>view(liver)
```

"Descriptive statistics"

```
>summary(liver)
>sd(liver$age)
>sd(liver$tb)
>sd(liver$db)
>sd(liver$alkphos)
>sd(liver$sgpt)
>sd(liver$sgot)
>sd(liver$tp)
>sd(liver$ab)
```

"Data Prepration& Exploration"

```
>str(liver1)
>liver1$gender <- as.factor(liver1$gender)
>liver1$result <- as.factor(liver1$result)
>str(liver1)
```

"Data Partition"

```
>set.seed(123456)
>ind<- sample(2, nrow(liver1), replace = T, prob = c(0.75, 0.25))
>train<- liver1[ind==1, ]
>test<- liver1[ind==2, ]
```

"Logistic Regression Model"

```
>library(car)
>library(BaylorEdPsych)

>model1 <- glm(result~age+gender+tb+db+alkphos+sgpt+sgot+tp+ab+agratio, data = train, family
= binomial(link = "logit"))
>vif(model1)

>model2 <- glm(result~age+gender+tb+db+alkphos+sgpt+sgot+tp+agratio, data = train, family = bi
nomial(link = "logit"))
>vif(model2)
>summary(model2)
>confint.default(model2)
>PseudoR2(model2)

>model3 <- glm(result~age+gender+db+alkphos+sgpt+sgot+agratio, data = train, family = binomia
l(link = "logit"))
>vif(model3)
>summary(model3)
>confint.default(model3)
>PseudoR2(model3)

>model4 <- glm(result~age+db+alkphos+sgpt+agratio, data = train, family = binomial(link = "logit"
))
>vif(model4)
>summary(model4)
>confint.default(model4)
>PseudoR2(model4)

>model5 <- glm(result~age+db+alkphos+sgpt, data = train, family = binomial(link = "logit"))
>vif(model5)
>summary(model5)
>confint.default(model5)
>PseudoR2(model5)
```

"Goodness of Fit Test: Model5"

```
>with(model5, pchisq(null.deviance - deviance, df.null - df.residual, lower.tail = F))
```

"Preditions& Misclassification Error for Train Data"

```
>library(ROCR)
>library(ggplot2)

>model5.1 <- glm(result~age+db+alkphos+sgpt, data = train, family = binomial(link = "logit"))
>pred1.data<- data.frame(prob.patient=model5.1$fitted.values,patient=train$result)
>pred1.data<- pred1.data[order(pred1.data$prob.patient, decreasing=FALSE),]
>pred1.data$rank<-1:nrow(pred1.data)

>ggplot(data=pred1.data, aes(x=rank, y=prob.patient)) +
geom_point(aes(color=patient), alpha=2, shape=1, stroke=1.06) +
xlab("Observations (Train Data)") +
ylab("Probability Being a Patient (Train Data)")

>p1 <- predict(model5, train, type = "response")
>color1 = c("red" , "yellow", "green", "violet", "orange", "blue", "pink", "cyan")
>hist(p1, right = FALSE, col = color1, xlab = "Probability (Being a Patient)", main = "Histogram (Trai
n Data)")

>pred1 <- prediction(p1, train$result)
>eval1 <- performance(pred1,"acc")
>color2 <- c("blue")
>plot(eval1, col = color2, xlab = "Cutoff (Train Data)", ylab = "Accuracy(Train Data)", main = "Accur
acy vs Cutoff Curve (Train Data)")

>max1 <- which.max(slot(eval1, "y.values")[[1]])
>acc1 <- slot(eval1, "y.values")[[1]][max1]
>cut1 <- slot(eval1, "x.values")[[1]][max1]
>print(c(Acuraccy = acc1, Cutoff = cut1))
>abline(h=0.737819, v=0.451921)

>head(p1)
>head(train)
>pred.train<- ifelse(p1>0.4519, 1,0)
>tab1 <- table(predicted = pred.train, actual = train$result)
>tab1
```

"Model Performance Evaluation (Preditions& Misclassification Error for Test Data)"

```
>model5.2 <- glm(result~age+db+alkphos+sgpt, data = test, family = binomial(link = "logit"))
>pred2.data<- data.frame(prob.patient=model5.2$fitted.values,patient=test$result)
>pred2.data<- pred2.data[order(pred2.data$prob.patient, decreasing=FALSE),]
>pred2.data$rank<-1:nrow(pred2.data)

>ggplot(data=pred2.data, aes(x=rank, y=prob.patient)) +
geom_point(aes(color=patient), alpha=2, shape=1, stroke=1.06) +
xlab("Observations (Test Data)") +
ylab("Probability Being a Patient (Test Data)")

>p2 <- predict(model5, test, type = "response")
>hist(p2, right = FALSE, col = color1, xlab = "Probability (Being a Patient)", main = "Histogram (Test
Data)")
>head(p2)
>head(test)
>pred.test<- ifelse(p2>0.4519, 1,0)
>tab2 <- table(predicted = pred.test, actual = test$result)
>tab2
```

"Reciever Operating Characteristic (ROC) Curve"

```
>roc<- performance(pred2, "tpr", "fpr")
>plot(roc, colorize = T, main = "ROC Curve")
>abline(a = 0, b =1)
```

"Area Under Curve (AUC)"

```
>auc<- performance(pred2, "auc")
>auc<- unlist(slot(auc, "y.values"))
>auc<- round(auc, 3)
>auc
>legend(0.7, 0.3, auc, title = "AUC", cex = 1.0)
```

APPENDIX II: Sample Data

age	gender	tb	db	alkphos	sgpt	sgot	tp	ab	agratio	result
65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
58	Male	1	0.4	182	14	20	6.8	3.4	1	1
72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
26	Female	0.9	0.2	154	16	12	7	3.5	1	1
29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	0
55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
72	Male	2.7	1.3	260	31	56	7.4	3	0.6	1
64	Male	0.9	0.3	310	61	58	7	3.4	0.9	0
74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	0
38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1
33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92	0
40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1
51	Male	2.2	1	610	17	28	7.3	2.6	0.55	1
51	Male	2.9	1.3	482	22	34	7	2.4	0.5	1
62	Male	6.8	3	542	116	66	6.4	3.1	0.9	1
40	Male	1.9	1	231	16	55	4.3	1.6	0.6	1
63	Male	0.9	0.2	194	52	45	6	3.9	1.85	0
34	Male	4.1	2	289	875	731	5	2.7	1.1	1
34	Male	4.1	2	289	875	731	5	2.7	1.1	1
34	Male	6.2	3	240	1680	850	7.2	4	1.2	1
20	Male	1.1	0.5	128	20	30	3.9	1.9	0.95	0
84	Female	0.7	0.2	188	13	21	6	3.2	1.1	0
57	Male	4	1.9	190	45	111	5.2	1.5	0.4	1
52	Male	0.9	0.2	156	35	44	4.9	2.9	1.4	1
57	Male	1	0.3	187	19	23	5.2	2.9	1.2	0
38	Female	2.6	1.2	410	59	57	5.6	3	0.8	0
38	Female	2.6	1.2	410	59	57	5.6	3	0.8	0
30	Male	1.3	0.4	482	102	80	6.9	3.3	0.9	1
17	Female	0.7	0.2	145	18	36	7.2	3.9	1.18	0
46	Female	14.2	7.8	374	38	77	4.3	2	0.8	1
48	Male	1.4	0.6	263	38	66	5.8	2.2	0.61	1
47	Male	2.7	1.3	275	123	73	6.2	3.3	1.1	1
45	Male	2.4	1.1	168	33	50	5.1	2.6	1	1
62	Male	0.6	0.1	160	42	110	4.9	2.6	1.1	0
42	Male	6.8	3.2	630	25	47	6.1	2.3	0.6	0