

FARM BUREAU MUTUAL INSURANCE COMPANY INSURANCE FRAUD

by Rikka Vivekanand Reddy Operation Research Analyst

Executive Summary

NC-FBMIC had suspicion of increase in insurance fraud cases in the late 2000's. Manager Jill Smith had been involved in several cases in which claimants had won their claims under dubious circumstances. She had access to Fraud Investigation agency which gave her access to extra information and understanding of the process. She was going through the article in the Harvard Business review about the new crew of "Data Scientists" and remembered meeting Charles. She then approached him asked him to select the data from past few years and see if he could develop a process where they could predict when the fraud might happen. Charles assessed the data which was provided to him. He then used the document management system of NC-FBMIC for it's electronic claim records. He was looking for cases where he could make sure that there were bigger claims being paid and were they and lawyers being sued for all such claims. So he made sure that claims which he retrieved had recorded amounts. He tried going through the 20,000 cases using the keywords "Attorney" & "Lawyer" and see if he could find a lawyer or group of lawyers involved in such claims. When he saw any and he would go through the Lead lawyer and recorded that in his dataset. After days of going through the data he put up 20,153 records for analysis.

Using visualization we have come across lot of interesting facts. We observed that the number of fraud cases are actually on a rise. We have plotted different charts and found that newly insured claimants commit more fraud than existing customers, Claims submitted for past five years are fraud if there are no prior claims associated with it. Also, we observed that teens are more likely to commit fraud compared to other age groups whereas claimants under eighties and nineties have zero probability to commit fraud. Finally using quality and process chart we found that Fraud rate is increasing when claimants use at least one Attorney. We used Logistic Regression and Decision Tree to build model to predict the outcome (Fraud/Not Fraud). Both models concluded that Attorney, Gender, Credit score and total monthly income strongly predict outcome variable. We have compared results from both models and concluded that Logistic Regression not only predict fraud cases accurately but also minimized.

When it comes to picking lawyers we observed Individuals who have made prior claims up to one would pick Attorneys Lawyer One and others. If they have two prior claims made, they prefer Lawyer team. The most interesting of all, if the claimant had prior claims of more than 3, Smith, Gold and Jones are the Attorneys of choice. Looks like these guys are some experts when it comes to dealing with complex cases. We also found that people who do not have insurance have committed the fraud. Another factor which affects is the credit score. Depending on the credit score the specific lawyer was approached. Henceforth we can collect evidence against these lawyers who have been helping these people.

Author: Rikka Vivekanand Reddy Technology Practice Lead Vuesol Technologies Inc

Index

•	Executive Summary	2
•	Data Cleansing	4
•	Data Visualization	4
	1. Visualization 1: Pie chart (Outcome vs Newly Insured)	4
	2. Visualization 2: Bar Chart (Outcome vs prior claims)	5
	3. Visualization 3: Histogram (Age vs Outcome)	5
	4. Visualization 4: Pareto Plot (Outcome vs Attorney)	6
	5. Visualization 5: Box plot outliers (Claim Amount & Credit Score)	6
•	Logistic Regression:	7
•	Decision Tree Analysis:	11
•	Confusion Matrix for different probability cut offs for Logistic Regression	14
•	Confusion Matrix for different probability cut offs for Decision Tree	15
•	Comparing and choosing the Best Model	16
•	Recommendations	. 17

Data Cleansing

- 1. We have identified 2 outliers, one in claimAMT(19000000) which has been excluded before starting model.
- 2. Outlier on Credit score(999) is ignored because both are Not Fraud claims.
- 3. We have partitioned Data into Training and Validation data sets. We will use Training data to build the model and Validation data to evaluate the model. 60% of the data is training data and remaining 40% of the data is validation data.

Data Visualizations



Visualization 1: Pie chart (Outcome vs newly insured)

It is observed from the pie chart that Fraud percentage in newly insured (53.7%) is more compared to fraud percentage to the people who are not newly insured (46.3%). This shows that newly insured people are more susceptible to make a fraud.



Visualization 2: Bar Chart (Outcome vs Prior Claims)

It is observed from the bar chart that fraudulent claims are high if there are no prior claims as we see that there were 1375 fraudulent claims when the prior claims were 0.



Visualization 3: Histogram (Age vs Outcome)

It is observed from the Histogram that, Teens and people in their twenties and thirties and also people in sixties and seventies are more likely to commit fraud as the numbers in the graph say.



Visualization 4: Pareto Plot (Outcome vs Attorney)

It is observed from the graph that Fraud rate is increasing when people are likely to use one lawyer or a Lawyer Team.

Also, for all the claims which are not Fraud they had no attorney or has one lawyer.



Visualization 5: Box plot Outliers (Credit Score & Claim Amount)

For the outlier on the credit score, it is observed that the outlier has a credit score 999 which is unusual.

For the outlier on the ClaimAMT, it is observed that the outlier has a huge claim amount 19000000\$ which is very unusual and hence we excluded that record from our analysis.

Logistic Regression

Dependent Variable : Outcome (Fraud/NonFraud)

Independent Variables :

1. Age	2. ClaimAMT
3. ticketsLast3yrs	4. Prior Claims
5. Attorney	6. Gender
7. Credit Score	8.Newly insured
9. Total Monthly income	10.Year

Whole Model Test :

Δ	Whole N	lodel Tes	t				
	Model	-LogLikeli	hood		DF	ChiSquare	Prob>ChiSq
	Difference	993	.7622		16	1987.524	<.0001*
	Full	5826	.8290				
	Reduced	6820	.5912				
	RSquare (U)		0	.1457		
	AICc			11	687.7		
	BIC			1	1822		
	Observation	ns (or Sum V	Vgts)	2	0016		
	Measure		Train	ing	Defini	tion	
	Entropy RS	quare	0.1	457	1-Logl	ike(model)/L	.oglike(0)
	Generalized	RSquare	0.1	913	(1-(L(0)/L(model))/	^(2/n))/(1-L(0)^
	Mean -Log	р	0.2	911	∑-Log	(p[j])/n	
	RMSE		0.2	817	√Σ(y[j]-p[j])²/n	
	Mean Abs I	Dev	0.1	600	Σly[j]-	ρ[j]/n	
	Misclassific	ation Rate	0.0	934	Σ(ρ(j);	≠pMax)/n	
	N		2001	6	n		
Þ	Lack Of	Fit					

Parameter Estimates

⊿ Effect Likelihood Ratio Tests

			L-R	
Source	Nparm	DF	ChiSquare	Prob>ChiSq
Attorney	5	5	1343.74166	<.0001*
Credit_score	1	1		
Tot_mthly_incm	1	1	13.3382184	0.0003*
Gender	1	1	5.26849044	0.0217*
Age	1	1	3.68042914	0.0551
ClaimAMT	1	1	5.93869833	0.0148*
TicketsLast3Yrs	1	1	2.23184959	0.1352
PriorClaims	4	4	8.01062678	0.0912
Newly Insured	1	1	1.29033839	0.2560

H_0 : The coefficients of the regression equation = 0

H_a : Atleast one of the coefficients of the regression equation $\neq 0$

- It is observed from the Test that P value for the whole model is less than alpha and the model is significant. Hence, we are rejecting the null hypothesis i.e at least one of the coef-ficients of the regression equation is not equal to zero.
- From Effect Likelihood ratio tests we can see that variables Age, Tickets last 3 years, Newly Insured, and Year have P-Values greater than Alpha whereas ClaimAMT, Attorney, Gender, Credit score and Totally Monthly Income have P-values less than Alpha.
- We have deleted the insignificant variables with the highest p value and ran the logistic re-gression with all significant variables. We have identified ClaimAMT as insignificant varia-ble here.
- Our final LR model has Outcome as dependent variable and Attorney, gender, Credit score and total monthly income as independent variables.

Logistic Regression and Output interpretation for significant variables :

				_					
4	Whole Mo	del Tes	t						
	Model ·	LogLikeli	ihood		DF	ChiSq	uare	Prob>(ChiSq
	Difference	985.4204			8	197	0.841	<.0	001*
	Full	5836	.4186						
	Reduced	6821	.8390						
	DC (1)								
	RSquare (U)			11	.1445				
	RIC			1	1762				
	Observations	(or Sum \	Nats)	2	0027				
	Measure		Train	ina	Defin	ition			
	Entropy RSqu	are	0.1	445	1-Log	like(mo	del)/L	oglike(0)	
	Generalized R	Square	0.1	897	(1-(L()/L(mo	del))^	(2/n))/(1-L(0)^(2/n))
	Mean -Log p		0.2	914	Σ-Log	(ρ[j])/r	1		
	RMSE		0.2	819	√ ∑(y[i]-ρ[j])²/	/n		
	Mean Abs De	v	0.1	602	Σlv(i)	·ρ[j] /n			
	Misclassificat	ion Rate	0.0	935	<u>Σ</u> (ρ[j]	≠pMax)/n		
			2002	1	n				
Δ	Lack Of Fi	t							
	Source	DF	-LogL	ikeli	hood	ChiSqu	lare		
	Lack Of Fit	19277	5	701	.9549	1140	3.91		
	Saturated	19285		134	.4637	Prob>	ChiSq		
	Fitted	8	5	836	.4186	1.	0000		
\triangleright	Paramete	r Estima	ates						
⊿	Effect Like	lihood	Ratio	o Te	ests				
						L-R			
	Source	Npar	m	DF	ChiS	quare	Prob	>ChiSq	
	Attorney		5	5	1357	78102	<	.0001*	
	Credit_score		1	1	560.3	93866	<	.0001*	
	Tot_mthly_ind	:m	1	1	15.31	40185	<	.0001*	
	Gender		1	1	4.931	27279	(0.0264*	

 H_0 : The coefficients of the regression equation = 0

 H_a : Atleast one of the coefficients of the regression equation $\neq 0$

- The p value for the whole model equals to 0.0001, which is significantly smaller than alpha (0.05). Hence all these four independent variables are good predictors for the outcome and at least one of the coefficients have non-zero value.
- The R square value for this model is 0.1445, which means that the model explains 14.45% of the variation in the "outcome (Fraud or Not Fraud)" and Misclassification rate is only 0.0935.
- From Lack of Fit it is observed that the Prob > ChiSq equals to 1, which means we don't have to add more cross effects.
- Finally, P-Values from the effect likelihood ratio tests suggest Attorney, Gender, Credit score, total monthly income are smaller than alpha, thus they all significantly affect our outcome.
- Based on the above regression result, we believe that we find a good prediction model. The equation for the Logit is:

Unit and Range Odds Ratios Interpretation

4	Odds R	atios									
Fe	or Outco	me odds	of Fraud ve	ersus	NotFraud	ł					
Te	ests and (confidenc	ce intervals	on o	odds ratio	s are li	kelihood	ratio			
b	ased.										
4	Unit	Odds Ra	atios								
	Per unit	change i	n regresso	r 				-			
	Term		Odds Rat	tio l	lower 95	% Up	per 95%	Rec	iprocal		
	Credit_s	score	0.9952	18	0.9948	24	0.995612	1.0	048049		
	Tot_mu	niy_incm	0.9999	00	0.9999	55	0.999970	1.0	000441		
4	Rang	e Odds	Ratios								
	Per char	nge in reg	ressor ove	r ent	ire range						
	Term		Odds Rat	tio l	ower 95.	% Up	per 95%	Rec	iprocal		
	Credit_s	score	0.0350)63	0.0265	78	0.046223	28.	520225		
	lot_mt	hly_incm	0.0583	55	0.0132	63	0.245/46	17.	136609		
4	Odds	Ratios	for Atto	rne	y						
	Level1	/Le	evel2	Od	ds Ratio	Prob:	>Chisq l	ower	95%	Upper	95%
	Lawyer	Team Sm	iith	23	8.82025	<	.0001*	86.7	11952	849.5	6541
	Lawyer	Team Lav	wyerOne	13	8.14879	<	.0001*	96.1	54272	204.6	8594
	Lawyer	Team no	ne	12	8.50003	<	.0001*	89.69	97922	190.0	2779
	Lawyer	Team Go	10	18	.204910		.0001*	6.20	430/0	43.02	2798
	Lawyer	Sm	ies ith	12	/ 33235		0001*	3.48	10083	61.86	/00/ //738
	Gold	Sm	ith	13	.075355	<	.0001*	3.8	30986	54.02	6325
	Jones	Lav	wyerOne	7.	7816042	<	.0001*	2.99	53763	21.03	3746
	Gold	Lav	wyerOne	7.5	5636148	<	.0001*	3.474	43536	16.66	0587
	Jones	no	ne	7.3	2381116	<	.0001*	2.80	27206	19.46	8972
	Gold	no	ne	7.0	0353474	<	.0001*	3.25	27537	15.40	5815
	none	Sm	ith	1	.858523	0	.2126	0.72	58717	6.296	9876
⊿	Odds	Ratios	for Gen	der							
	Level1	/Level2	Odds R	latio	Prob>	Chisa	Lower	95%	Upper	95%	
	male	female	0.894	8727	0.	0264*	0.811	1894	0.9	87051	
	female	male	1.117	4774	0.	0264*	1.013	1188	1.232	27577	

We have come up with some useful conclusions based on the unit odds ratio,

• If Total monthly Income increases by one dollar, the customers are 0.9999 times more likely to be fraud than those customers with lower total monthly income. Total Monthly Income is a very important factor the management should check when they plan to decrease the fraud rate.

• If the credit score increases by 1, the customer are 0.99 times more likely to be fraud than those customers with lower credit scores. That means the higher the credit score, the less likely the customers will commit fraud.

• Comparing with male and female, we can see that female is 1.11 times more likely than male to commit fraud. Hence management to pay more attention on female customers than on male customers.

• Comparing among different attorney, we can see that if customers are using lawyer team, then they are more likely to commit fraud comparing with only one attorney or none. Customers who use lawyer team are 138.14 times

more likely to commit fraud than customers with no attorney.

• Furthermore, when we compare individual attorney, we can find that customers choosing Gold as their attorney is 18.26 more likely to commit fraud than customers choosing Smith. Pretty much the same when we compare the odds ratio for Gold and Lawyer one, and Gold with no lawyer. Thus, if the new customers are with attorney Gold, the management should be clear that they are more likely commit fraud.

Confusion Matrix for training set and Validation set:

•	 Contingency Table 									
	Most Likely Outcome									
	Count	Fraud	Fraud NotFrau							
	Total %		d							
	Col %									
	Row %									
e	Fraud	188	1096	1284						
E E		1.56	9.12	10.68						
utc		89.10	9.28							
Ō		14.64	85.36							
	NotFraud	23	10713	10736						
		0.19	89.13	89.32						
		10.90	90.72							
		0.21	99.79							
	Total	211	11809	12020						
		1.76	98.24							

•	Conting	ency T	able							
Most Likely Outcome										
	Count	Fraud	NotFrau	Total						
	Total %		d							
	Col %									
	Row %									
e	Fraud	133	730	863						
B		1.66	9.12	10.78						
utc		85.26	9.30							
Ō		15.41	84.59							
	NotFraud	23	7120	7143						
		0.29	88.93	89.22						
		14.74	90.70							
		0.32	99.68							
	Total	156	7850	8006						
		1.95	98.05							

With 0.5 probability cutoff, we analyze confusion matrix for both training set and validation set.

- For Training set, we can find its explanatory accuracy power. For example, the overall accuracy equals to 90.69%. The sensitivity for training set is 89.09 and the specificity is 90.71%.
- For Validation set, we can see that its predictive accuracy is 90.59%. The sensitivity for validation set is 85.25 and the specificity is 90.70%



Interpretation from Prediction Profiler:

- We can see from the prediction profiler, as Total Monthly Income and Credit Score increase, the probability of fraud will decrease.
- Female customers are more likely to commit fraud.
- The probability of fraud varies with attorney.

Recommendations:

- Using Logistic Regression model we found out that independent variables "attorney", "gender", "credit score" and "total monthly income" can predict outcome variable.
- Female customers are more likely to commit fraud than male customers.
- Customers with a lawyer team are more likely to commit fraud than customers with just one lawyer or no lawyers.
- Customers who have a lower income and credit scores are more likely to commit fraud.
- To sum up, the management can classify the existing customers according to these four identifica-tions and target at those people who have a larger rate of fraud. For future business consideration, they can also refuse the potential customers with higher possibility of committing fraud to de-crease the fraud rate.

Decision Tree

Dependent variable: Outcome (Fraud/Non Fraud)

Independent variables : Age, Claim amount, Ticket last 3 years, Prior claims, Attorney, Gender, Credit score, Newly issued, Total monthly income, Year.

Decision Tree :



- The R square value for training data is 0.216 that means 21.6% of outcome in training data is explained by the independent variables after 9 splits.
- Similarly 20.6% of outcome in validation data is explained by the independent variables after 9 splits.

Split History:



- From the above fig we can see that the validation decreases after the 9th split.
- The split history is a graph that shows the number of splits against the corresponding Rsquare value. This could be used as a reference to find the point when further splitting is to be stopped. The split history generated from our DT.

Column Con	tributior	ıs		
Term	Number of Splits	G^2		Portion
Attorney	4	854.290167		0.4826
Credit_score	1	811.376081		0.4583
Tot_mthly_incm	1	74.0881367		0.0419
Gender	1	18.101939		0.0102
Year	1	7.3750523		0.0042
Newly Insured	1	5.08339911		0.0029
Age	0	0		0.0000
ClaimAMT	0	0		0.0000
TicketsLast3Yrs	0	0		0.0000
PriorClaims	0	0		0.0000

Attorney, credit score and total monthly income are the three important variables in identifying the outcome (Fraud/not Fraud).

Leaf Report:

⊿ Leaf Report		
Response Prob		
Leaf Label	Fraud	NotFraud
Credit_score>=345 or Missing^&Tot_mthly_incm<10295&Attorney(LawyerTeam)	0.9958	0.0042
Credit_score<345 not Missing	0.7181	0.2819
Credit_score>=345 or Missing^&Tot_mthly_incm<10295&Attorney(Jones)	0.5329	0.4671
Credit_score>=345 or Missing^&Attorney(Gold)	0.4177	0.5823
Credit_score>=345 or Missing&Attorney(Jones, LawyerTeam)&Tot_mthly_incm>=10295	0.2577	0.7423
Credit_score>=345 or Missing^^&Attorney(none)&Gender(female,)	0.0919	0.9081
Credit_score>=345 or Missing^^&Attorney(none)&Gender(male)&Year<2010&Newly Insured(Y)	0.0859	0.9141
Credit_score>=345 or Missing^^&Attorney(none)&Gender(male)&Year<2010&Newly Insured(N)	0.0629	0.9371
Credit_score>=345 or Missing^^&Attorney(Smith, LawyerOne)	0.0617	0.9383
Credit_score>=345 or Missing^^&Attorney(none)&Gender(male)&Year>=2010	0.0468	0.9532

From the Leaf Report we can observe that:

- Claimants with credit score greater than or equal to 345 or missing and whose total monthly income is less than 10,295 and who has lawyer team as attorney have 99.58% probability in committing fraud.
- Claimants with credit score greater than or equal to 345 or missing value and whose total monthly income is less than \$10,295 and who has Jones as attorney has 53.29% probability in committing fraud.
- Female Claimants whose credit score greater than 345 or missing and attorney as Lawyer One have very low probability in committing fraud.
- Male Claimants whose credit score greater than 345 or missing and attorney as Smith, Lawyer One, none or jones and who are newly insured have very low probability in committing fraud.

Confusion matrices Interpretation for Training and Validation Data Set:

Fit Details			
Measure	Training	Validation	Definition
Entropy RSquare	0.2155	0.2063	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2762	0.2655	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2662	0.2712	∑ -Log(p[j])/n
RMSE	0.2659	0.2686	√ ∑(y[j]-p[j])²/n
Mean Abs Dev	0.1415	0.1428	Σ [y[i]-p[i]]/n
Misclassification Rate	0.0788	0.0805	∑(ρ[j]≠ρMax)/n
Ν	12091	8061	n
Confusion Matr	ix		
Trainin	g		Validation
Actual Pr	edicted	Actu	ual Predicted
Outcome Fraud	I NotFrau	d Outco	ome Fraud NotFraud

Fraud

NotFraud

281

62

587

7131

Confusion Matrix

Fraud

NotFraud

• The explanatory power of Training data is (441+10697)/(441+848+105+10697) = 92.11%

848

10697

• The explanatory power of Validation data is (281+7131)/(281+587+62+7131) = 91.94%.

Recommendations from Decision Tree:

- The decision tree model developed by us has an accuracy rate of 91.94% and hence could be used to make a decision whether to issue a policy or not to a customer.
- These are the recommendations from our decision tree:

441

105

- » Fraud investigation team of NC-FBMIC should concentrate more on the customers with credit score greater than or equal to 345 or missing and whose total monthly income is less than 10,295 and who has lawyer team as attorney.
- » NC-FBMIC can trust the female Claimants whose credit score greater than 345 or missing and attor-ney as Lawyer One and also the newly insured male Claimants whose credit score greater than 345 or missing and attorney as Smith, Lawyer One, none or jones.

Confusion Matrix for different(0.25 & 0.75) prob cutoffs for LR

1 💌	Contingency Table										
	LR0.25 cutoff										
	Count	Fraud	Not	Total							
	Total %		Fraud								
	Col %										
	Row %										
e	Fraud	465	819	1284							
E E		3.87	6.81	10.68							
utc		58.56	7.30								
Ō		36.21	63.79								
	NotFraud	329	10407	10736							
		2.74	86.58	89.32							
		41.44	92.70								
		3.06	96.94								
	Total	794	11226	12020							
		6.61	93.39								

Contingency Table									
	Most Likely Outcome								
[Count	Fraud	NotFrau	Total					
	Total %		d						
	Col %								
	Row %								
e	Fraud	188	1096	1284					
ω		1.56	9.12	10.68					
utc		89.10	9.28						
0		14.64	85.36						
	NotFraud	23	10713	10736					
		0.19	89.13	89.32					
		10.90	90.72						
		0.21	99.79						
	Total	211	11809	12020					
		1.76	98.24						

Contingency Table								
	LR 0.75 Cutoff							
	Count	Fraud	Not	Total				
	Total %		Fraud					
	Col %							
	Row %							
e	Fraud	183	1101	1284				
шo		1.52	9.16	10.68				
utc		91.96	9.31					
0		14.25	85.75					
	NotFraud	16	10720	10736				
		0.13	89.18	89.32				
		8.04	90.69					
		0.15	99.85					
	Total	199	11821	12020				
		1.66	98.34					

Contingency tables:

- We set up three confusion matrixes for training group and validation group for cutoffs equal to 0.25, 0.5 and 0.75 and compare the result for three validation sets.
- As you can see from the three validation graphs above, the misclassification for

0.25 probability = 0.0865 0.5 probability = 0.08913

0.75 Probability = 0.08918

- Also, as the cutoff decreases, the false negative is increased from 819 to 1096 and then to 1101.
- We can see that the miss classification rate for 0.25 probability is less and False negative is less than other probabilities. Hence we can say that 0.25 probability in Decision tree is the best model.

Confusion Matrix for different(0.25 & 0.75) prob cutoffs for DT

	Contingency Table							
	DT 0.25 cutoff							
	Count	Fraud	Not	Total				
	Total %		Fraud					
Outcome	Col %							
	Row %							
	Fraud	454	835	1289				
		3.75	6.91	10.66				
		77.61	7.26					
		35.22	64.78					
	NotFraud	131	10671	10802				
		1.08	88.26	89.34				
		22.39	92.74					
		1.21	98.79					
	Total	585	11506	12091				
		4.84	95.16					

∠ ⊂ Contingency Table

	Most Likely Outcome 2						
	Count	Fraud	NotFrau	Total			
	Total %		d				
	Col %						
	Row %						
e	Fraud	441	848	1289			
omo		3.65	7.01	10.66			
utc		80.77	7.35				
0		34.21	65.79				
	NotFraud	105	10697	10802			
		0.87	88.47	89.34			
		19.23	92.65				
		0.97	99.03				
	Total	546	11545	12091			
		4.52	95.48				

Contingency Table

	DT 0.75 Cutoff						
	Count	Fraud	Not	Total			
	Total %		Fraud				
	Col %						
	Row %						
e	Fraud	175	1114	1289			
om		1.45	9.21	10.66			
utc		100.00	9.35				
0		13.58	86.42				
	NotFraud	0	10802	10802			
		0.00	89.34	89.34			
		0.00	90.65				
		0.00	100.00				
	Total	175	11916	12091			
		1.45	98.55				

Contingency tables:

- We set up three confusion matrixes for training group and validation group for cutoffs equal to 0.25, 0.5 and 0.75 and compare the result for three validation sets.
- As you can see from the three validation graphs above, the misclassification for 0.25 probability = 0.08826
 0.5 probability = 0.08947
 0.75 Probability = 0.08934
- Also, as the cutoff decreases, the false negative is increased from 835 to 848 and then to 1114.
- We can see that the miss classification rate for 0.25 probability is less and False negative is less than other probabilities. Hence we can say that 0.25 probability in Decision tree is the best model.

Model Comparision and Selecting BestModel:

Model	ТР	TN	FP	FN	Missclassification Rate	Sensitivity	Specificity
Logistic Reg							
Prob = 0.25	465	10407	329	819	0.0865	0.362	0.969
Prob = 0.5	188	10713	23	1096	0.08913	0.146	0.9978
Prob = 0.75	183	10720	16	1101	0.08918	0.142	0.9985
Decision Tree							
Prob = 0.25	454	10671	131	835	0.08826	0.352	0.9878
Prob = 0.5	441	10697	105	848	0.08947	0.342	0.9902
Prob = 0.75	15	10802	10	1114	0.08934	0.011	0.9990

LR and DT model comparison

• From the above table we conclude that LR model is best model with probability is 0.25 for predicting Fraudulent claims because of high True Positive (465) and low False Negative(819) and low misclassification rate which is 8.65%.

Recommendations:

In order to order to curb the fraudulent cases we suggest the following suggestions:

- Claimants with Credit score greater than or equal to 345 and who use either Lawyer team as Attorney and whose Total Monthly Income is greater than 10,295 have high probability in committing fraud.
- Claimants with Credit score less than 345 have 99% probability in committing fraud. Hence we recommend NC-FBMIC to investigate and challenge the validity of claim and ask claim-ants for more documentation before issuing the auto insurance.
- Male claimants with credit score greater than or equal to 345 and attorney as none have less probability in committing fraud. Hence no further documentation is required for these claimants.
- We need to monitor more closely the income group of \$1000-\$10,000 and make sure all the paperwork is properly scrutinized.
- The bank need to be more careful with the customers having low credit score and need to be careful while filing for their claims.
- As we know that Gold, Smith and Jones are helping the customers with low credit score, it is compulsory that next time for the court hearing, the company should go in with better lawyers and strong evidence.



Headquarters

4625 Alexander Drive Suite #245,Alpharetta, GA-30022, USA +1-678-862-0550

Sales office

2033 Gateway Place, Suite 605, San Jose, CA 95110 +1-678-862-0550

visit **www.vuesol.com**

Development Centre

Nagar, HITEC City, Hyderabad, Telangana 500081 040-29884477

E-mail