ETC1000 BUSINESS AND ECONOMIC STATISTICS

Group Project Semester 2 2021

By: Dhivyan, Ayush, Ivan, Eric, Lakshay

WHAT IS FRAUD?

Definition:

- Car insurance fraud is a serious deliberate dishonest act that causes an actual or potential financial loss to any person.

Purpose of Study:

- Insurance fraud has been identified as a major business challenge in the insurance industry.
- The costs of investigating potential fraud are very high, with the need to check documentation and costly physical inspections.
- So to cut down on these costs, we are seeking to develop a predictive model that allows them to classify a claim as more or less likely to involve fraud, based purely on the information provided in the initial claim documents.
- The company will then focus investigation efforts on the claims that are likely to be fraud based on the predicted model.

I. THE DATA

HOW WE UNDERTOOK THIS RANDOM SAMPLE

- We first created a separate column which contained random numbers from 0-1 for each claim of fraud.
- We created these random numbers using the rand() function which will generate a random number greater than or equal to 0 and less than 1. We then applied this formula to the whole column giving us a random number for each claim of fraud
- We then ranked the data based of the random number column in ascending order and take the first 8000 claims. This will ensure that our sample is random
- Selection of random numbers removes selection bias and is an appropriate method as it allows for more accurate representations in our analysis which otherwise may have had bias and would have lead to misleading outcomes when determining car insurance fraud.

Random Number
0.000208
0.000228
0.00025
0.000478
0.000543
0.000567
0.000677
0.000923
0.000944

2A.
A FRAUD DETECTION MODEL

EXPLORING THE NUMERICAL DATA

Regression Between Annual Income and Fraud Detected

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.222951135
R Square	0.049707209
Adjusted R Squa	0.049588392
Standard Error	0.390071971
Observations	8000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	63.65504532	63.65504532	418.3534373	1.12999E-90
Residual	7998	1216.94483	0.152156143		
Total	7999	1280.599875			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*.ower 95.0%*	*Jpper 95.0%*
Intercept	0.311364302	0.006971213	44.66429123	0	0.297698907	0.325029698	0.297699	0.32503
Annual Income (	-2.93327E-06	1.4341E-07	-20.45369007	1.12999E-90	-3.21439E-06	-2.65214E-06	-3.2E-06	-2.7E-06

Regression Between Claim and Fraud Detected

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.167598541
R Square	0.028089271
Adjusted R Squa	0.027967752
Standard Error	0.394483839
Observations	8000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	35.97111669	35.97111669	231.1508467	1.74657E-51
Residual	7998	1244.628758	0.155617499		
Total	7999	1280.599875			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*.ower 95.0%*	*Jpper 95.0%*
Intercept	0.278676542	0.0067931	41.02347238	0	0.265360296	0.291992788	0.26536	0.291993
Claim Amount	-9.85825E-06	6.48414E-07	-15.20364583	1.74657E-51	-1.11293E-05	-8.58719E-06	-1.1E-05	-8.6E-06

Regression Between Monthly Premium and Fraud Detected

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.092721749
R Square	0.008597323
Adjusted R Squa	0.008473366
Standard Error	0.398419945
Observations	8000

ANOVA

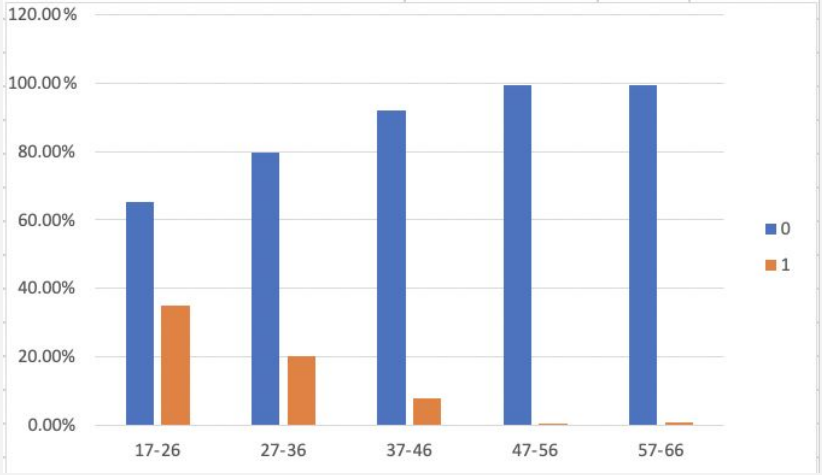
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	11.00973031	11.00973031	69.3576769	9.57866E-17
Residual	7998	1269.590145	0.158738453		
Total	7999	1280.599875			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*.ower 95.0%*	*Jpper 95.0%*
Intercept	0.102406366	0.012550655	8.159443987	3.87885E-16	0.077803811	0.127008921	0.077804	0.127009
Monthly Premi	0.000904309	0.000108585	8.328125653	9.57866E-17	0.000691454	0.001117164	0.000691	0.001117

GRAPHICAL TECHNIQUES

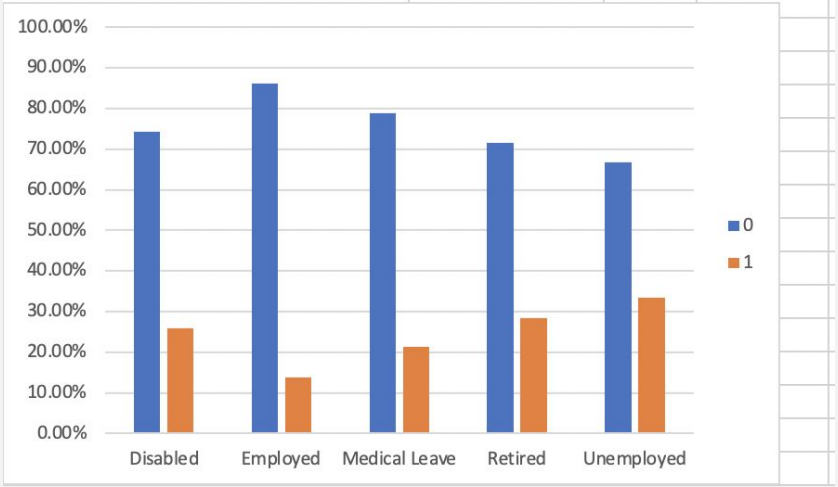
Age

Count of Fraud Detected? Yes=1, No=0	Column Labels		
Row Labels	0	1	Grand Total
17-26	65.05%	34.95%	100.00%
27-36	79.83%	20.17%	100.00%
37-46	92.10%	7.90%	100.00%
47-56	99.43%	0.57%	100.00%
57-66	99.35%	0.65%	100.00%
Grand Total	**79.99%**	**20.01%**	**100.00%**



Employment Status

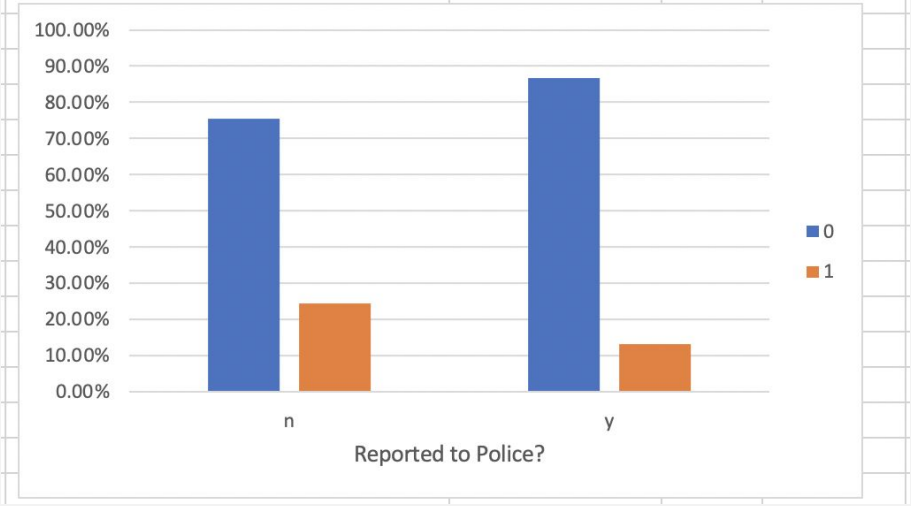
Count of Fraud Detected? Yes=1, No=0	Column Labels		
Row Labels	0	1	Grand Total
Disabled	74.23%	25.77%	100.00%
Employed	86.23%	13.77%	100.00%
Medical Leave	78.74%	21.26%	100.00%
Retired	71.60%	28.40%	100.00%
Unemployed	66.68%	33.32%	100.00%
Grand Total	**79.99%**	**20.01%**	**100.00%**



GRAPHICAL TECHNIQUES

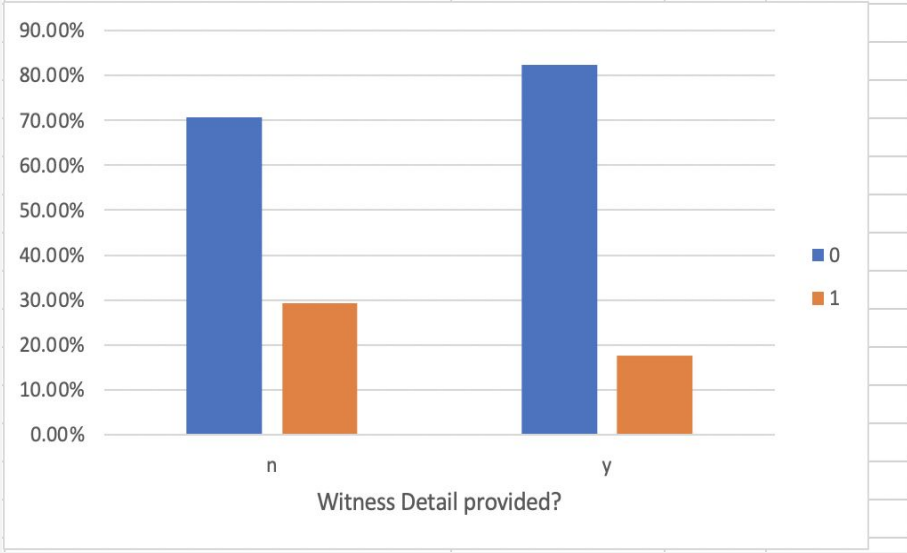
Reported to Police?

Count of Fraud Detected? Yes=1, No=0	Column Labels		
Row Labels	0	1	Grand Total
n	75.52%	24.48%	100.00%
y	86.76%	13.24%	100.00%
Grand Total	79.99%	20.01%	100.00%



Witness details provided?

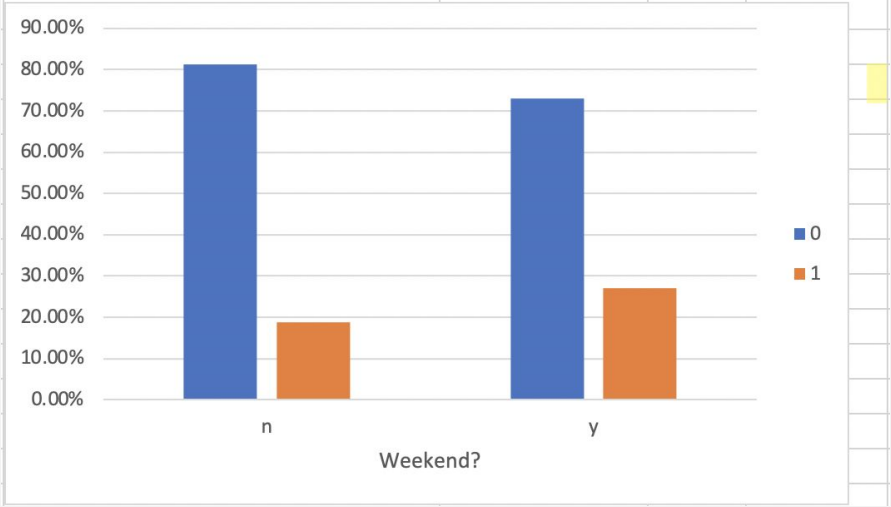
Count of Fraud Detected? Yes=1, No=0	Column Labels		
Row Labels	0	1	Grand Total
n	70.64%	29.36%	100.00%
y	82.30%	17.70%	100.00%
Grand Total	79.99%	20.01%	100.00%



GRAPHICAL TECHNIQUES

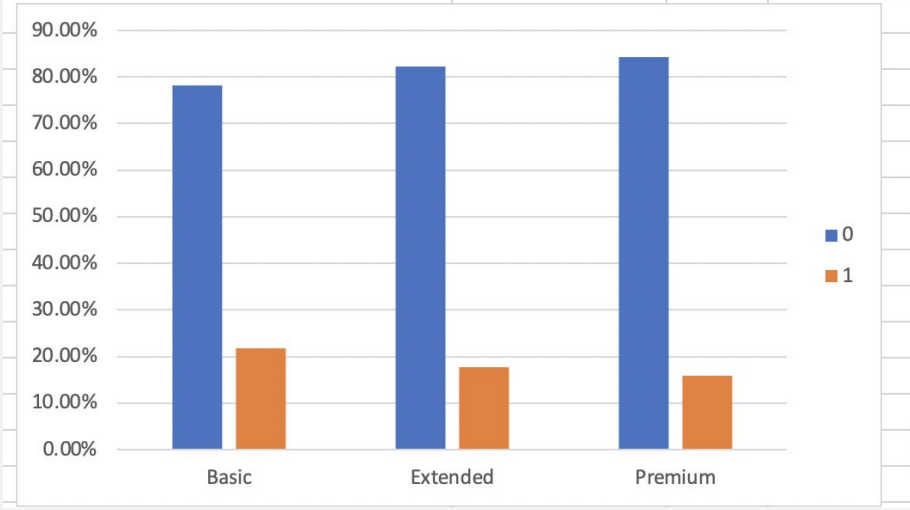
Reported on weekend?

Count of Fraud Detected? Yes=1, No=0	Column Labels		
Row Labels	0	1	Grand Total
n	81.25%	18.75%	100.00%
y	73.02%	26.98%	100.00%
Grand Total	**79.99%**	**20.01%**	**100.00%**



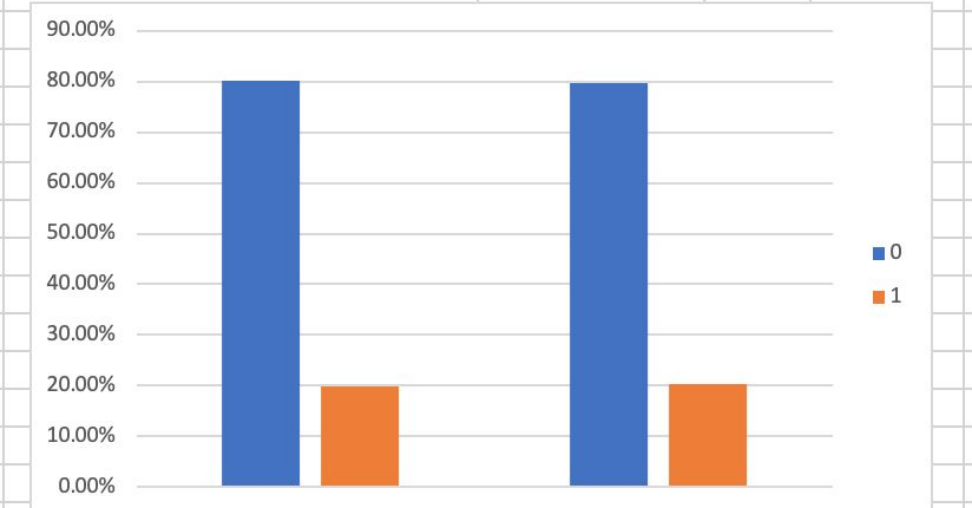
Coverage

Count of Fraud Detected? Yes=1, No=0	Column Labels		
Row Labels	0	1	Grand Total
Basic	78.20%	21.80%	100.00%
Extended	82.33%	17.67%	100.00%
Premium	84.22%	15.78%	100.00%
Grand Total	**79.99%**	**20.01%**	**100.00%**

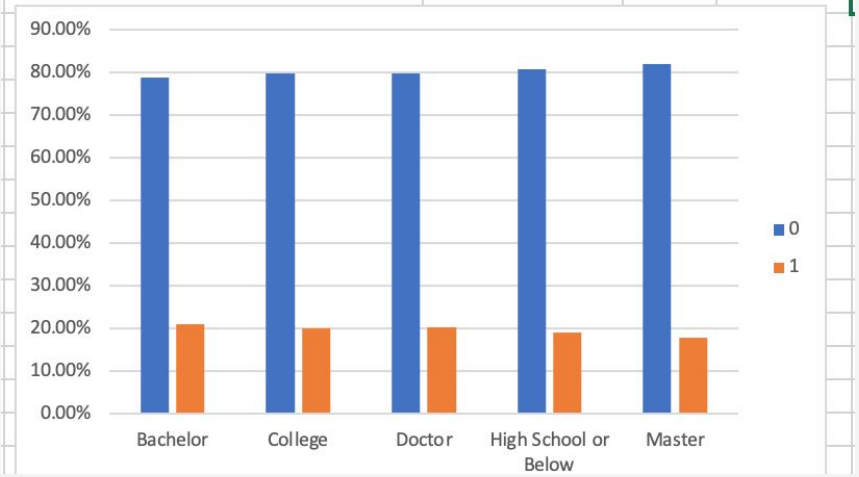


GRAPHICAL TECHNIQUES

Count of Fraud Detected? Yes=1, No=0	Column Labels		
Row Labels	0	1	Grand Total
F	80.22%	19.78%	100.00%
M	79.75%	20.25%	100.00%
Grand Total	**79.99%**	**20.01%**	**100.00%**



Count of Fraud Detected? Yes=1, No=0	Column Labels		
Row Labels	0	1	Grand Total
Bachelor	78.86%	21.14%	100.00%
College	79.82%	20.18%	100.00%
Doctor	79.74%	20.26%	100.00%
High School or Below	80.80%	19.20%	100.00%
Master	82.06%	17.94%	100.00%
Grand Total	**79.99%**	**20.01%**	**100.00%**



SUMMARY OF FRAUD DATA

AGE GROUPS 17-26 AND 27-26

34.95% and 20.17% committed fraud in these categories

UNEMPLOYED,RETIRED AND DISABLED

33.32%, 28.40% and 25.77% committed fraud in these categories

NOT REPORTED TO POLICE

24.48% committed fraud in this category

NO WITNESS DETAILS

29.36% committed fraud in this category

REPORTED ON WEEKEND

26.98% committed fraud in this category

BASIC COVERAGE

21.8% committed fraud in this category

INITIAL FINAL MULTIPLE REGRESSION MODEL

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.4690895
R Square	0.220045
Adjusted R Square	0.2187753
Standard Error	0.3536526
Observations	8000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	13	281.78956	21.67612	173.31168	0
Residual	7986	998.81032	0.1250702		
Total	7999	1280.5999			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.0210671	0.02372	0.8881575	0.3744828	-0.0254303	0.0675645	-0.0254303	0.0675645
Dummy Age(17-26)	0.2810008	0.0103574	27.130435	3.94E-155	0.2606976	0.301304	0.2606976	0.301304
Dummy Age(27-36)	0.1372943	0.010191	13.472133	6.373E-41	0.1173173	0.1572713	0.1173173	0.1572713
Dummy Unemployed	0.0652067	0.0202971	3.2126179	0.0013205	0.0254192	0.1049942	0.0254192	0.1049942
Dummy police report (n)	0.1090492	0.0080832	13.490859	4.972E-41	0.093204	0.1248943	0.093204	0.1248943
Dummy Basic Coverage	0.0434417	0.0086781	5.0059056	5.679E-07	0.0264304	0.060453	0.0264304	0.060453
Monthly Premium	0.0005565	0.000118	4.714053	2.47E-06	0.0003251	0.0007879	0.0003251	0.0007879
Claim Amount	-9.782E-06	6.239E-07	-15.678751	1.37E-54	-1.101E-05	-8.559E-06	-1.101E-05	-8.559E-06
Annual Income ($)	-2.297E-06	2.257E-07	-10.179386	3.442E-24	-2.74E-06	-1.855E-06	-2.74E-06	-1.855E-06
Weekend Dummy (y)	0.0863457	0.0109791	7.864548	4.189E-15	0.0648238	0.1078676	0.0648238	0.1078676
Dummy Witness Details Provided(n)	0.1143893	0.0099206	11.530437	1.615E-30	0.0949423	0.1338364	0.0949423	0.1338364
Dummy Retired	0.0483807	0.0290441	1.6657694	0.0957986	-0.0085532	0.1053147	-0.0085532	0.1053147
Dummy Disabled	0.0249491	0.0260606	0.9573486	0.3384203	-0.0261365	0.0760348	-0.0261365	0.0760348
Dummy Employed	-0.0023779	0.0205087	-0.1159458	0.9076984	-0.0425802	0.0378245	-0.0425802	0.0378245

- A p-value less than 0.05 is considered statistically significant
- Therefore we rejected the highlighted variables which had a P-value greater than 0.05, indicating they do not have a significant effect on fraud occurring

Combination of Variables

FINAL REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.468532885
R Square	0.219523065
Adjusted R Squa	0.218546125
Standard Error	0.353704474
Observations	8000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	10	281.1212093	28.11212093	224.7048804	0
Residual	7989	999.4786657	0.125106855		
Total	7999	1280.599875			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Jpper 95.0%*
Intercept	0.032146257	0.018227003	1.763661136	0.077827266	-0.003583426	0.067875939	-0.00358	0.067876
Dummy Age(17-	0.281474722	0.010356653	27.17815565	1.1954E-155	0.26117298	0.301776464	0.261173	0.301776
Dummy Age(27-	0.137455359	0.010191496	13.4872603	5.21317E-41	0.117477367	0.157433351	0.117477	0.157433
Dummy Unempl	0.053766367	0.013189032	4.076596899	4.61427E-05	0.027912422	0.079620312	0.027912	0.07962
Dummy police r	0.109178293	0.008083996	13.505485	4.09349E-41	0.09333155	0.125025036	0.093332	0.125025
Dummy Basic Cc	0.043693957	0.008676862	5.035686769	4.86555E-07	0.026685044	0.06070287	0.026685	0.060703
Monthly Premiu	0.000557749	0.000118055	4.724465507	2.34662E-06	0.000326329	0.000789168	0.000326	0.000789
Claim Amount	-9.80873E-06	6.23767E-07	-15.72497231	6.75165E-55	-1.10315E-05	-8.58598E-06	-1.1E-05	-8.6E-06
Annual Income	-2.49444E-06	1.88042E-07	-13.26535292	9.65931E-40	-2.86305E-06	-2.12583E-06	-2.9E-06	-2.1E-06
Weekend Dumm	0.086527391	0.01097917	7.881049989	3.67479E-15	0.065005351	0.10804943	0.065005	0.108049
Dummy Witness	0.114216598	0.009921239	11.51233222	1.98556E-30	0.094768381	0.133664816	0.094768	0.133665

- For the majority of our categorical variables they have positive coefficients which suggest that if they are true, the probability of fraud will increase (e.g. if unemployed/ have basic coverage it is likely to lead to fraud)

- Whereas Annual Income and Claim Amount have negative coefficients , this shows that as Annual Income and Claim Amount increase the probability of fraud will decrease

REGRESSION MODEL IN ACTION

I. MODEL IS USED TO PREDICT FRAUD

Probability of Fraud
0.28197779
0.139778058
0.113439615
-0.178380991
0.351977595

- We used the regression model generated to determine these values
- These values gave us the probability of fraud.

2. METHOD TO CLASSIFY FRAUD

Predicted Values where fraud occurred
0.549780552
0.543330088
0.366663974

Mean of Probability of fraud occurring
0.375716011

- We created a new column which gave us the predicted values of fraud of when fraud actually occurred
- We then took average of this value to give us the basis of the model
- We used the average as our Key Threshold Variable, predicting that fraud occurred if the probability of fraud was greater than this value

3. OUR OUTCOME VS ACTUAL OUTCOME

Predicted if Fraud (using mean)	Fraud Detected? Yes=1, No=0	Did our model predict correctly?	% of claims predicted correctly
0	0	1	82.33%
0	0	1	
0	0	1	
0	0	1	
0	0	1	
0	0	1	

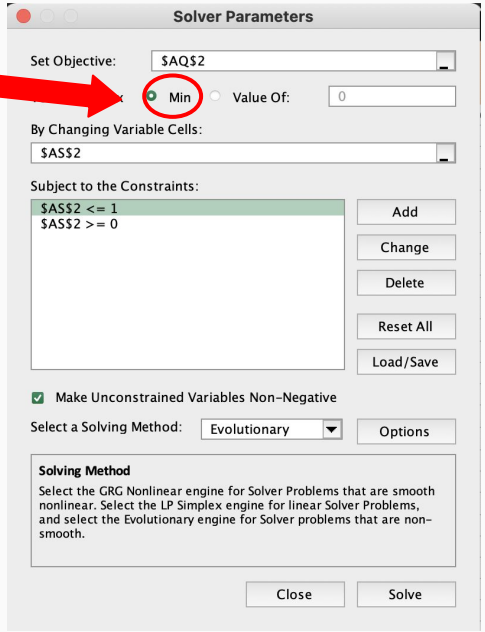
- The actual data set had a total of 1601 cases of fraud
- Our predicted model predicted 1443 cases of fraud.
- Although there were instances of incorrectly identifying fraud,
- The model was successful in predicting if fraud did occur or not 82.33% of the time

2B.
AN INVESTIGATION DECISION RULE

How did we approach creating a decision model?

Key Threshold Variable
0.356571291

Type 1 Errors (1 if ERROR)	Type 2 Errors (1 if ERROR)	Total Type 1 Errors	Total Type 2 Errors	Total Cost of Errors ($)
0	0	788	669	$3,189,000.00
0	0			
0	0			
0	0			



- This was our initial setup, where we utilised the solver function to alter the Key threshold variable in order to give the minimum total cost of errors ($).
- After using the solver function we came to the conclusion that thew shown key threshold value (0.3565...) was the most optimal, since it gave us the lowest cost of errors ($3,189,000).

Performance of our model?

Proportion of investigated claims	Succesfully identified fradulent claims (1 is succesful)	TOTAL SUCCESFUL IDENTIFIED CLAIMS	TOTAL UNDETECTED FRAUDLENT	Investigated but not fraudulent
21.50%	0	932	669	788
	0			
	0			
	0			

- Our model chose to investigate 21.50% of the 8000 claims (1,720 claims investigated)
- Out of these 1720 claims, 932 successfully investigated a claim with fraudulent activity
- However 669 fraudulent claims were not investigated (Type II Error)
- 788 claims were investigated based on our model, and these claims were found to have no fraud (Type I Error)

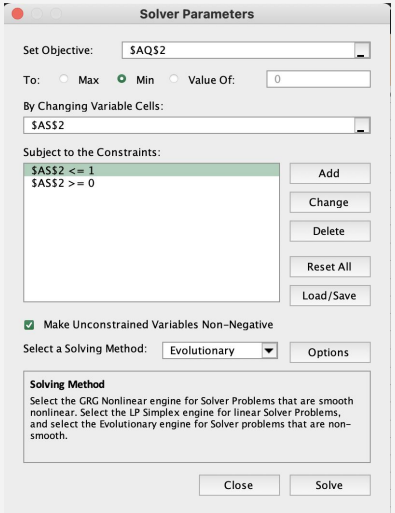
Sensitivity of our Decision Rule:

Decision Rule Sensitivity

	Cost of Type 1 Error	Cost of Type 2 Error	Total Type 1 Cost	Total Type 2 Cost	Total Cost	Key Threshold Variable	Total Type 1 Errors	Total Type 2 Errors
1	2000	1000	170000	1286000	1456000	0.510184487	85	1286
2	2000	6000	3090000	2316000	5406000	0.278751426	1545	386
3	2000	8000	3090000	3088000	6178000	0.278751426	1545	386
4	2000	10000	4488000	2160000	6648000	0.227435154	2244	216
5	5000	25000	11220000	5400000	16620000	0.227435154	2244	216
6	25000	5000	325000	7555000	7880000	0.6018971	13	1511

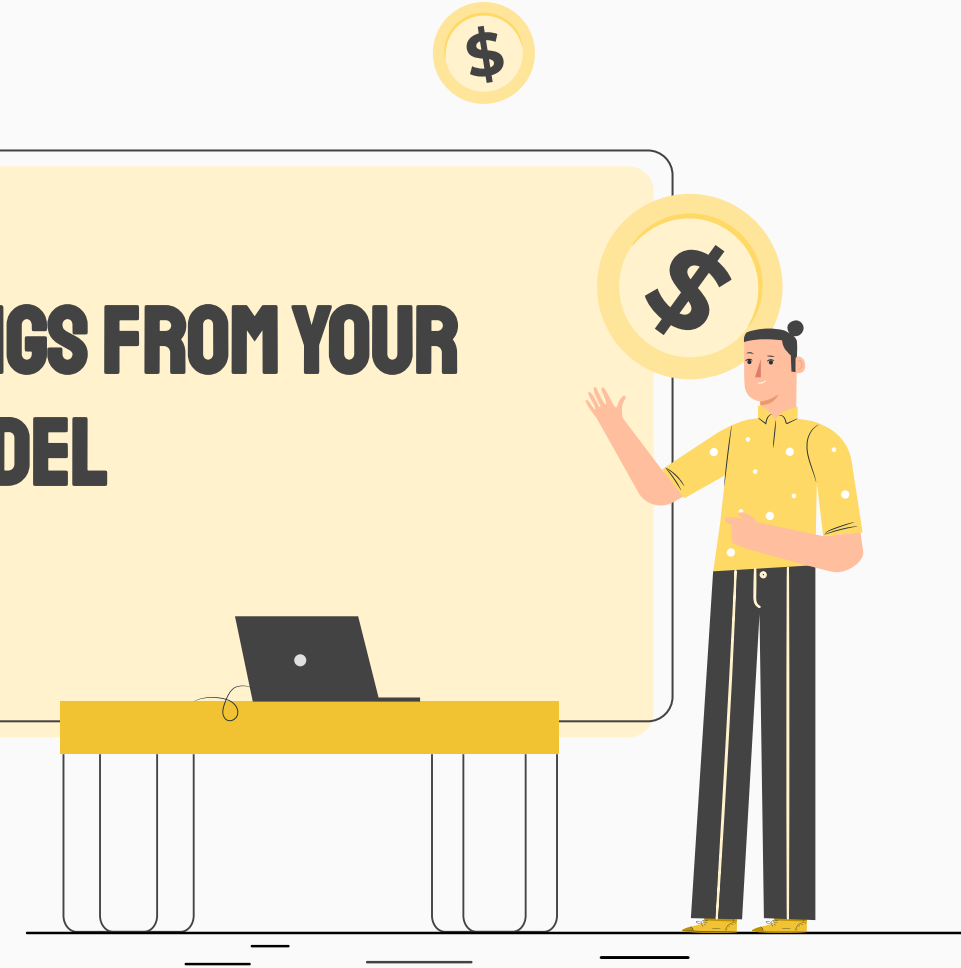
Original Key Threshold Value: **0.35657 when Type 1 cost = $1500 and Type 2 cost = $3000**

- Based on our sensitivity changes as shown above, changing the cost of type 2 did not lead to a significant change in out key threshold variable
  (This is evident in **test case 3 and 4** in the above figure)
- However when significantly changing type 1 cost to $25,000 and type 2 cost to $5000, it lead to a significant **increase** in the key threshold value
- This was not the case when type 1 cost was $5000, and type 2 cost was $25,000 (inverted), this lead to a significant **decrease** in the key threshold value
- When the ratio of type 1 cost to type 2 cost is >0 it yields a higher key threshold value compared to our original key threshold (e.g. test case 1 and test case 6)



**Key threshold variables were calculated using excel solver*

2C. COST SAVINGS FROM YOUR MODEL



Random Number	Random Sample to be investigated	Type 1 Error	Type 2 Error	Total Type 1 Errors	Total Type 2 Errors	Total Cost ($)	Total Cost Found in Part B	Total Cost Saved	Total Cost Saved (%)
0.00021623	1	0	0	1374	1255	$5,826,000.00	$3,189,000.00	$2,637,000.00	45.26%
0.0002509	1	1	0						
0.00035819	1	1	0						
0.00037149	1	1	0						
0.00095071	1	1	0						
0.00096704	1	1	0						
0.00113429	1	0	0						
0.00143161	1	1	0						
0.00143491	1	1	0						
0.00155038	1	0	0						
0.00182424	1	1	0						
0.00197605	1	1	0						
0.00205283	1	1	0						
0.00239506	1	1	0						
0.00252696	1	0	0						

- We obtained a random sample of 1733 pieces of data from our original chosen sample of 8000 customers
- This is because our previous model determined that there was 1733 claims to investigate further for potential fraud
- They were randomly selected using our previous method where we utilised the rand() function to give each data piece a random number.
- After assigning a random number to each claim, we sorted the whole sample in ascending order.
- This randomly sorted our sample, and then we selected the first 1733 pieces of data as claims to investigate.
- This will ensure that we have randomly selected 1733 claims to investigate, so that no bias exists.

Random Number	Random Sample to be investigated	Type 1 Error	Type 2 Error	Total Type 1 Errors	Total Type 2 Errors	Total Cost ($)	Total Cost Found in Part B	Total Cost Saved	Total Cost Saved (%)
0.00021623	1	0	0	1374	1255	$5,826,000.00	$3,189,000.00	$2,637,000.00	45.26%
0.0002509	1	1	0						
0.00035819	1	1	0						
0.00037149	1	1	0						
0.00095071	1	1	0						
0.00096704	1	1	0						
0.00113429	1	0	0						
0.00143161	1	1	0						
0.00143491	1	1	0						
0.00155038	1	0	0						
0.00182424	1	1	0						
0.00197605	1	1	0						
0.00205283	1	1	0						
0.00239506	1	1	0						
0.00252696	1	0	0						
0.00254975	1	1	0						
0.00266906	1	0	0						

- After selecting our sample data, we then calculated the type 1 errors, type 2 error, and the final total cost ($5,286,000), based on if we had selected to investigate claims randomly instead of using our model
- Once we had the total cost calculated, we subtracted the total cost found in part b using our model to investigate claims that were only predicted to have fraud
- The result ($2,637,000) is the total cost we saved if we had used our model instead of randomly selecting claims to investigate
- This can also be represented as a total cost saving of 45.26%.

2D. CONTINUOUS IMPROVEMENT

- Utilise new data (larger sample space) to improve regression model
- Include more variables in the regression model based on new data
- Update the key threshold variable based on new regression model