9/13/2014

A Study on Insurance Fraud using Advanced Analytics

Capstone Project Report

Polaris Financial Technology Limited

Team

Name	Student ID
Arun Prakash	71310100
Deeparani	71310007
Lakshmi Narasimha Rao	71310039
Nagarchana Mathuraiveeran	71310063
KameshwarSambamurthy	

Acknowledgements

We would like to take this opportunity to express our profound gratitude and deep regards to our project mentor, Professor Thriyambakam Krishnan for his guidance, monitoring and constant encouragement throughout the course of this project work, Reema, for her efforts in facilitating this capstone opportunity and last but not the least, our project sponsor, Polaris Financial Technology Limited, for providing the opportunity to learn and work on challenging insurance fraud detection model building.

We are deeply indebted to all of them and welcome this opportunity to benefit further from their contribution. In particular, we wish to express our special thanks to batch mates who provided their valuable suggestions.

Our several well-wishers who helped us directly or indirectly; we virtually fall short of words to express our gratefulness to them. Therefore, we are leaving this acknowledgement incomplete…… in their reminiscence.

Arun, Deepa, Lakshmi, Nagarchana and Kamesh

Table of Contents

A Study on Insurance Fraud using Advanced Analytics

Overview

Insurance fraud is the second biggest white-collar crimes in the U.S. after tax evasion, according to the National Insurance Crime Bureau.1 As insurers deal with an uncertain economic climate and intense competition, they must also grapple with the increasing incidence and sophistication of fraud, not to mention the resulting losses. The traditional methods of identifying fraud are no longer sufficient.

Advanced analytics can help insurers identify and reduce fraud-related losses, as well as condense the claims cycle, resulting in improved customer satisfaction. Historical claims data, combined with industry data, can be a starting point for insurers to identify common types of fraud early in the claims process.

We have chosen this project offered by ***"Polaris Financial Technology Limited".***

About our sponsor

Founded in 1993, Polaris Financial Technology Limited (BSE: 532254 | NSE: POLARIS) is a global leader in Financial Technology (FT) for Banking, Insurance, and other Financial Services. The organization offers superior technology solutions through its two specialized divisions that enable clients' unprecedented operational efficiency – FT Services and FT Products.

Polaris' FT Services is guided by powerful platforms and high performance practices. Its techno-functional capabilities lead industry standard on several parameters. The organization's specialist capability in providing solutions through delivery is apparent across its full spectrum offerings that include Testing, Infrastructure Management, Business Efficiency, Business Transformation, Data & Analytics, Mobility & Channels, and Risk & Compliance. Today, Polaris' high performance FT solutions run in over 250 financial institutions around the world.

About the dataset

The health insurance claims dataset is provided by Insurance Information Bureau of India. Insurance Information Bureau of India was promoted in year 2009 by IRDA, with the participation of stakeholders of the insurance sector, with the objective of supporting the insurance industry with sector-level data to enable data-based and scientific decision making including pricing and framing of business strategies. The Bureau is also expected to provide key inputs to the Regulator and the Government to assist them in policymaking. The Bureau has in its brief period of existence generated insightful reports, both periodic and one-time, for the benefit of the industry. IIB handles the Central Index Server which acts as a nodal point between different Insurance Repositories and helps in de-duplication of demat accounts at the stage of creation of a new account. The Central Index Server also acts as an exchange for transmission/routing of information pertaining to transactions on each policy between an insurer and the insurance repository.

The health insurance claim dataset is downloaded from IIB website. Tariff Advisory Committee (TAC) created a National Data Repository of Health Insurance. All Insurers and Third-party Administrators (TPAs) shall furnish data in respect of health insurance to the Repository. Tariff Advisory Committee is the custodian of the Repository. The claim dataset which we are using is provided by IIB

Attributes

The claim dataset has got 100,000 records in total with 56 attributes. The claim record has details related to the policy, insurer, TPA, claim amount, medical procedure, disease diagnosis etc.

The following is the list of variables

#	**Attribute Name**
1	Boo_hospital_is_networked
2	Boo_Whether_Claim_Made_Under_Alternate
3	Date_Claim_Intimation
4	Date_of_Admission
5	Date_of_Birth
6	Date_of_Discharge
7	Date_of_Payment
8	Date_Policy_End
9	Date_Policy_Start

10	Num_Age_of_Insured
11	Num_Amount_of_Co_Payment_or_Excess_if_applicable
12	Num_Bonus_Sum_Insured
13	Num_Consultation_Charges
14	Num_Investigation_Charges
15	Num_Medicine_Charges
16	Num_Miscellaneous_Charges
17	Num_Other_Non_Hospital_Expenses
18	Num_Percentage_of_Co_Payment_or_Excess_if_applicable
19	Num_Post_Hospitalisation_Expenses_included_under_150035
20	Num_Pre_Hospitalisation_Expenses_included_under_150035
21	Num_Room_Nursing_Charges
22	Num_Sum_Insured
23	Num_Surgery_Charges
24	Num_Total_Amount_Claimed
25	Num_Total_Claim_Paid
26	Txt_Claim_Number_Masked
27	Txt_Diagnosis_Code_Level_I
28	Txt_Diagnosis_Code_Level_II
29	Txt_Diagnosis_Code_Level_III
30	Txt_Gender
31	Txt_Hospital_Code
32	Txt_Insurer_Code_Masked
33	Txt_Medical_History_Level_I
34	Txt_Medical_History_Level_II
35	Txt_Medical_History_Level_III
36	Txt_Member_Reference_Key_Masked
37	Txt_Name_of_the_Hospital_Masked
38	Txt_PAN_of_Hospital_Masked
39	Txt_Payment_Reference_Number_Masked
40	Txt_Pincode_of_Hospital_Masked

41	Txt_Policy_Number_Masked
42	Txt_Procedure_Code_Level_III
43	Txt_Procedure_Code_Level_I
44	Txt_Procedure_Code_Level_II
45	Txt_Procedure_Description_Level_I
46	Txt_Procedure_Description_Level_II
47	Txt_Procedure_Description_Level_III
48	Txt_Product_Type
49	Txt_Reason_for_Reduction_of_Claim
50	Txt_Reason_for_Rejection_of_Claim
51	Txt_Registration_Number_of_Hospital_Masked
52	Txt_Remarks_of_TPA
53	Txt_System_of_Medicine_Used
54	Txt_TPA_Code_Masked
55	Txt_Type_of_Claim_Payment
56	Txt_Type_of_Policy

The dataset also has lot of fields masked and they are the following,

- Txt_Claim_Number_Masked
- Txt_Insurer_Code_Masked
- Txt_Member_Reference_Key_Masked
- Txt_Name_of_the_Hospital_Masked
- Txt_PAN_of_Hospital_Masked
- Txt_Payment_Reference_Number_Masked
- Txt_Pincode_of_Hospital_Masked
- Txt_Policy_Number_Masked
- Txt_Registration_Number_of_Hospital_Masked
- Txt_TPA_Code_Masked

The abbreviated form of the data type of all the attributes are prefixed in the attribute name. For instance Boo_, Txt_, Num_, Date_ indicate that the attribute is of Boolean, textual, number and date format respectively.

Challenges Faced

In order to build advanced analytical models to perform fraud detection, the dataset should contain a fraud indicator. This fraud indicator will help in training the various models for fraud detection. Any classification model or discriminant analysis mandates the need of an indicator. The biggest challenge faced by us is the missing fraud indicator in the claim dataset. This was the biggest roadblock faced and we started brainstorming various ways of arriving at a response variable (fraud indicator) for modelling.

In addition to that domain knowledge was a challenge faced. Even though the team had a fair understanding on the Insurance domain, the team lacked in depth domain skills which is mandatory for resolving the road block faced. Hence we took up a different track to resolve the issue faced.

Solution to the challenge faced

We went through publicly available research papers, current industry trends, existing fraud management practices etc. to figure out a way out of this issue. Then we came across a research work by ***Dr. Ashish Dogra*** on ***Trigger based scoring System for health insurance claims*** (reference 1). This is a business rule based scoring method which is a result of extensive research of health insurance claim data. A collection of business rules along with a score for each rule are defined.

We ultimately fine-tuned the scoring model after acquiring the necessary domain knowledge and guidance from our mentor. The implementation of the scoring model and the selection of variables for the scoring model are detailed below.

Exploratory Data Analysis and Data Preprocessing

In this section of the document we will focus on the exploratory data analysis of the claims dataset. We shall also look at the preprocessing/ cleaning performed on the same.

General note

All the numeric variables are maintained in int or num format and text/ categorical variables in factor format.

All the dates are converted into POSiXct format for the convenience of calculation.

Imputation for missing values

The below mentioned numeric attributes contain NA values which are impute using zero.

Attribute	# of NA values
Num_Amount_of_Co_Payment_or_Excess_if_applicable	41618
Num_Consultation_Charges	17920
Num_Investigation_Charges	21071
Num_Medicine_Charges	16521
Num_Miscellaneous_Charges	25577
Num_Other_Non_Hospital_Expenses	45039
Num_Post_Hospitalisation_Expenses_included_under_150035	35978
Num_Pre_Hospitalisation_Expenses_included_under_150035	37017
Num_Room_Nursing_Charges	20305
Num_Surgery_Charges	30622
Num_Total_Amount_Claimed	6

The categorical variable Boo_hospital_is_networked contains 10 NA values and are replaced using zero

Num_Age_of_Insured has 309 NA values imputed using the corresponding values of Policy_Start_Date - Date_of_Birth.

Condition checks based on domain knowledge

Condition	# of records that does not satisfy
Date_of_Discharge>= Date_of_Admission	10898
Policy end date > Policy Start date	1741

Date_Policy_Start<Date_of_Admission	5074
Date_of_Discharge<Date_Policy_End	7331
Num_Age_of_Insured<=100	130

Miscellaneous checks

Removed 343 negative values from Num_Miscellaneous_Charges present in dataset.

Removed 131 negative values from Num_Age_of_Insured created due to imputation.

Variables Selected for Modeling and Scoring

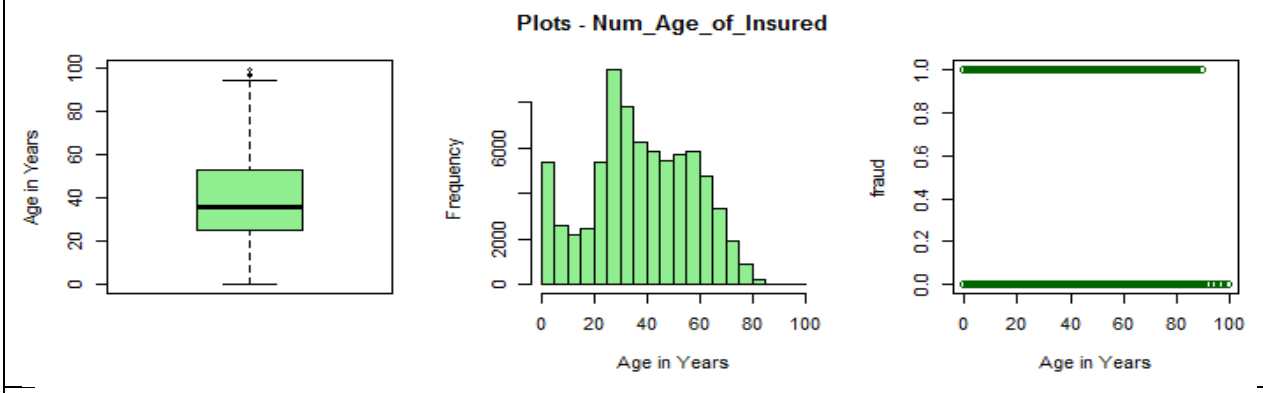
The following table shows the complete list of variables along with the indication of whether it is

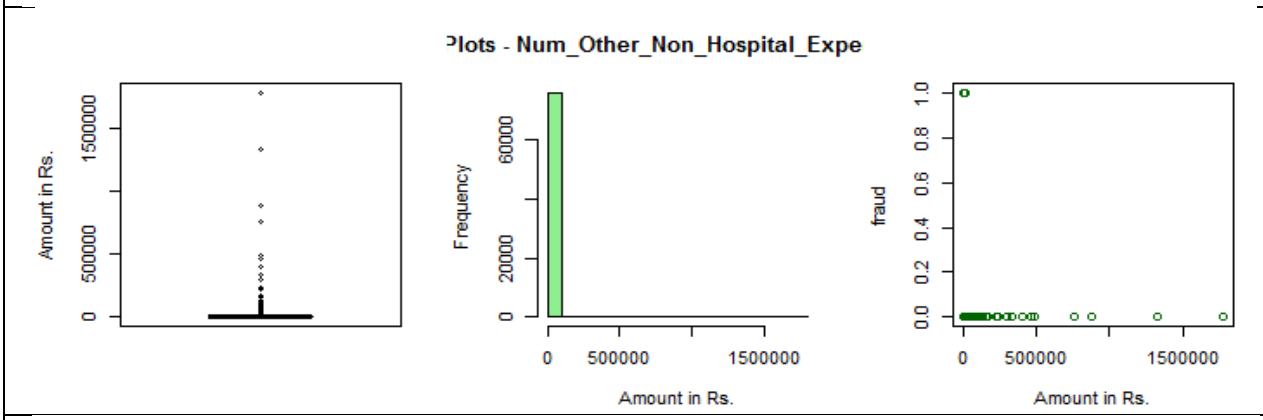
a) selected for modeling
b) selected for business rules of the scoring model
c) selected for both
d) unselected

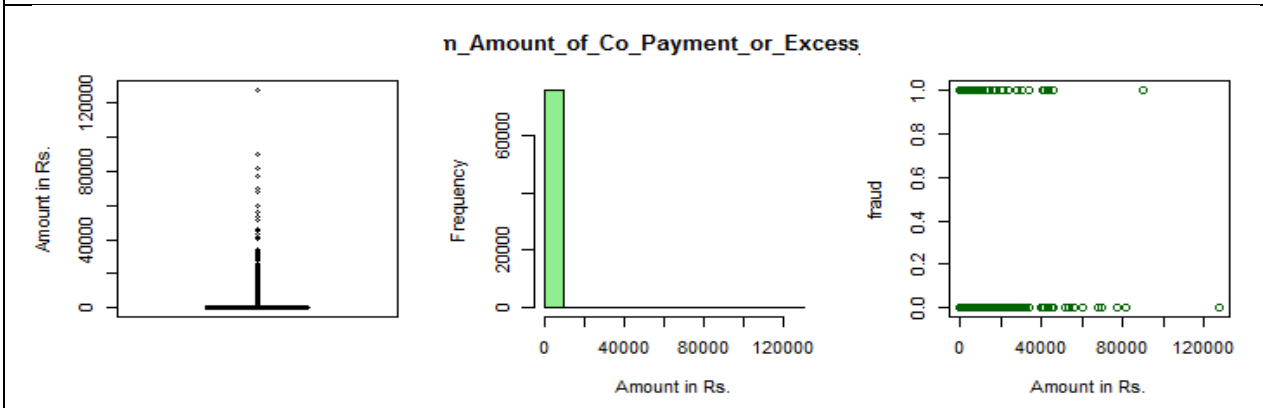
Usage	Attribute Name	n	missing	unique	Mean
Scoring Model - Business Rule Only	Date_Claim_Intimation	100000	0	1474	
	Date_of_Admission	100000	0	1628	
	Date_of_Discharge	100000	0	1635	
	Date_Policy_Start	100000	0	1605	
	Num_Bonus_Sum_Insured	58377	41623	374	5780
	Num_Total_Amount_Claimed	99994	6	41369	25050
	Txt_Claim_Number_Masked	100000	0	96931	
	Txt_Diagnosis_Code_Level_I	99999	1	7066	
	Txt_Diagnosis_Code_Level_II	99996	4	4409	
	Txt_Diagnosis_Code_Level_III	99996	4	1800	
	Txt_Hospital_Code	100000	0	13308	
	Txt_Medical_History_Level_I	100000	0	6927	
	Txt_Medical_History_Level_II	100000	0	475	
	Txt_Medical_History_Level_III	100000	0	326	
	Txt_Member_Reference_Key_Masked	100000	0	86295	
	Txt_Pincode_of_Hospital_Masked	100000	0	2815	
	Txt_Policy_Number_Masked	100000	0	46552	
	Txt_Procedure__Code_Level_III	99996	4	715	
	Txt_Procedure_Code_Level_I	100000	0	1886	
	Txt_Procedure_Code_Level_II	99996	4	979	
	Txt_Procedure_Description_Level_I	100000	0	2143	
	Txt_Procedure_Description_Level_II	100000	0	924	
	Txt_Procedure_Description_Level_III	99997	3	6603	
	Txt_Remarks_of_TPA	100000	0	4621	
	Txt_Type_of_Policy	1.00E+05	0	5	
Modeling Only	Num_Age_of_Insured	99691	309	100	37.12
	Num_Amount_of_Co_Payment_or_Excess_if_applicable	58382	41618	1920	192.3
	Num_Other_Non_Hospital_Expenses	54961	45039	1958	613.9
	Num_Post_Hospitalisation_Expenses_included_under_150035	64022	35978	4355	799.8
	Num_Pre_Hospitalisation_Expenses_included_under_150035	62983	37017	2889	275.6
	Txt_Gender	1.00E+05	0	3	
	Txt_Product_Type	1.00E+05	0	8	

Usage	Attribute Name	n	missing	unique	Mean
Both Business Rule & Modeling	Boo_hospital_is_networked	99990	10	2	
	Num_Consultation_Charges	82080	17920	6616	3141
	Num_Investigation_Charges	78929	21071	7222	2266
	Num_Medicine_Charges	81524	18476	14711	4187
	Num_Miscellaneous_Charges	74423	25577	12586	5096
	Num_Room_Nursing_Charges	79695	20305	3643	2935
	Num_Sum_Insured	1.00E+05	0	1089	1693632
	Num_Surgery_Charges	69378	30622	5288	4656
	Num_Total_Claim_Paid	1.00E+05	0	37515	20624
	Txt_Type_of_Claim_Payment	1.00E+05	0	6	
Unused variables	Boo_Whether_Claim_Made_Under_Alternate	85380	14620	2	
	Date_of_Birth	100000	0	21975	
	Date_of_Payment	100000	0	899	
	Date_Policy_End	100000	0	1620	
	Num_Percentage_of_Co_Payment_or_Excess_if_applicable	39403	60597	53	1.061
	Txt_Insurer_Code_Masked	100000	0	16	
	Txt_Name_of_the_Hospital_Masked	100000	0	30078	
	Txt_PAN_of_Hospital_Masked	100000	0	3212	
	Txt_Payment_Reference_Number_Masked	100000	0	75866	
	Txt_Reason_for_Reduction_of_Claim	100000	0	664	
	Txt_Reason_for_Rejection_of_Claim	100000	0	972	
	Txt_Registration_Number_of_Hospital_Masked	100000	0	4225	
	Txt_System_of_Medicine_Used	90701	9299	3	
	Txt_TPA_Code_Masked	100000	0	23	

EDA plots

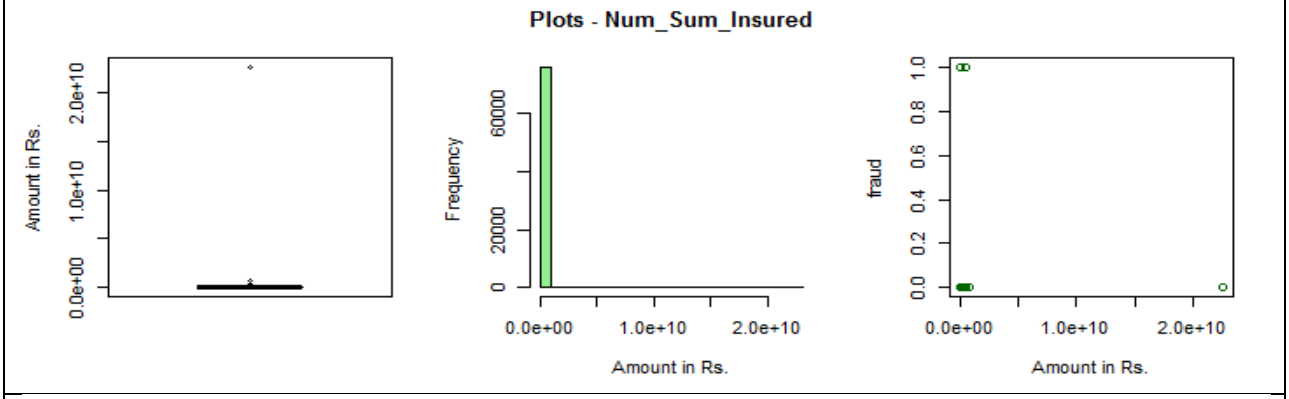


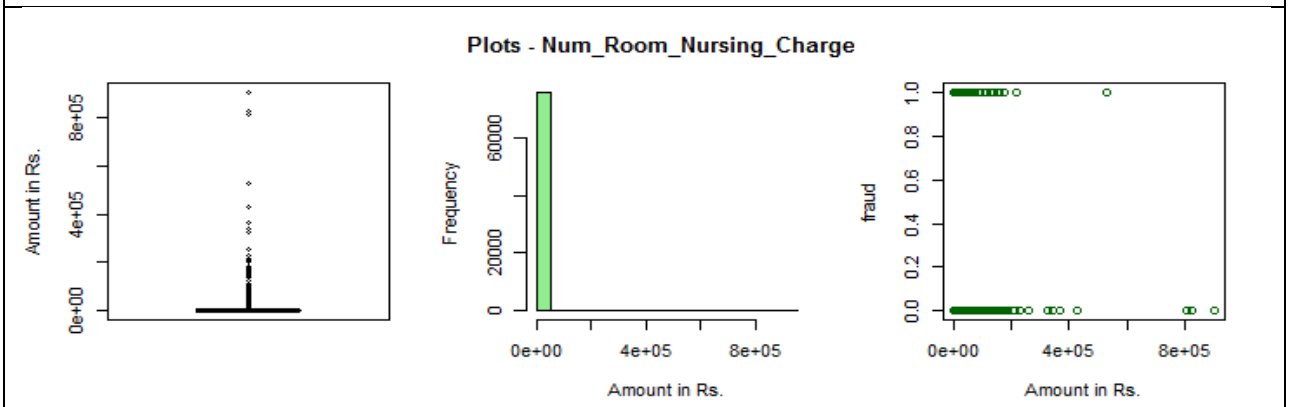


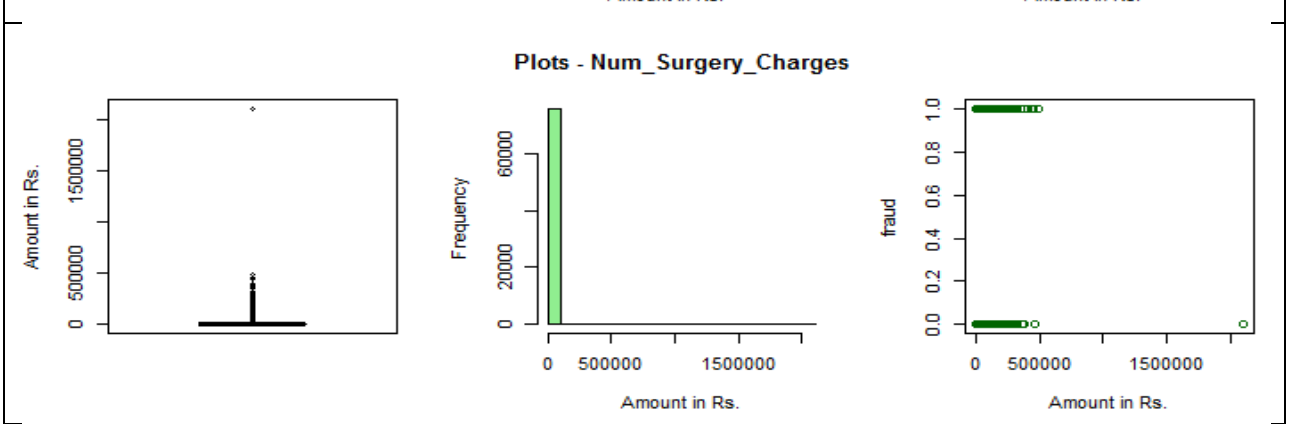


Comments:

Age in years is well distributed with very less outliers.Frequency varies rapid. People above 90 years of age do not perform fraud.Other Hospital expenses and Copayment both have lots of zeroes and left skewed.0 amount other hospital expense records are fraudulent.Lesser than 60000 Amount of Payment is fraudulent.

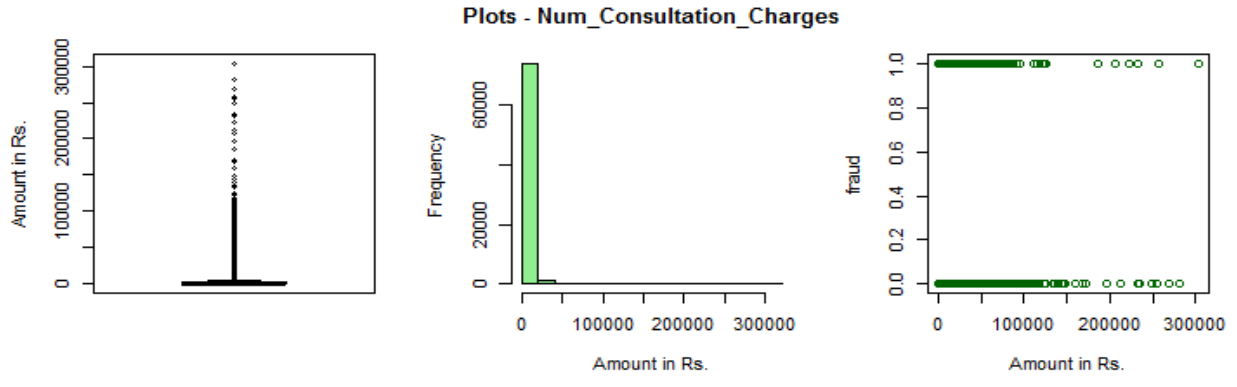


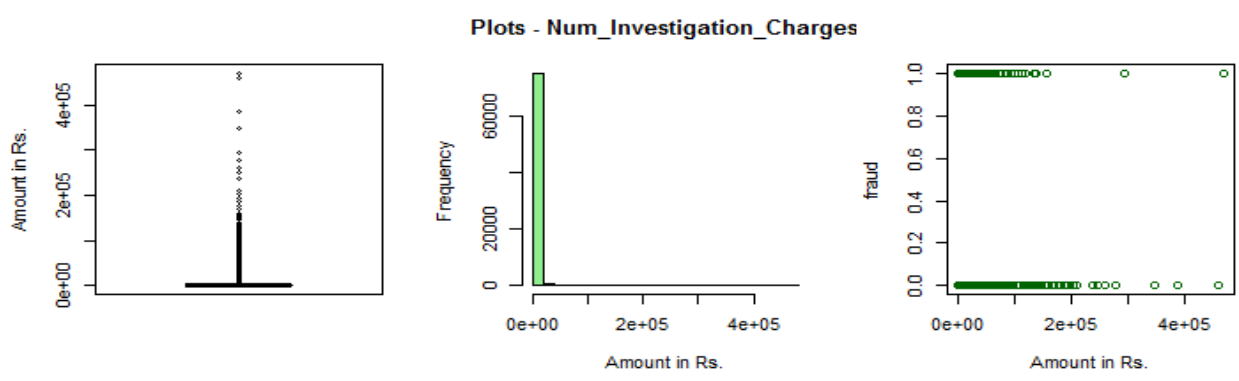


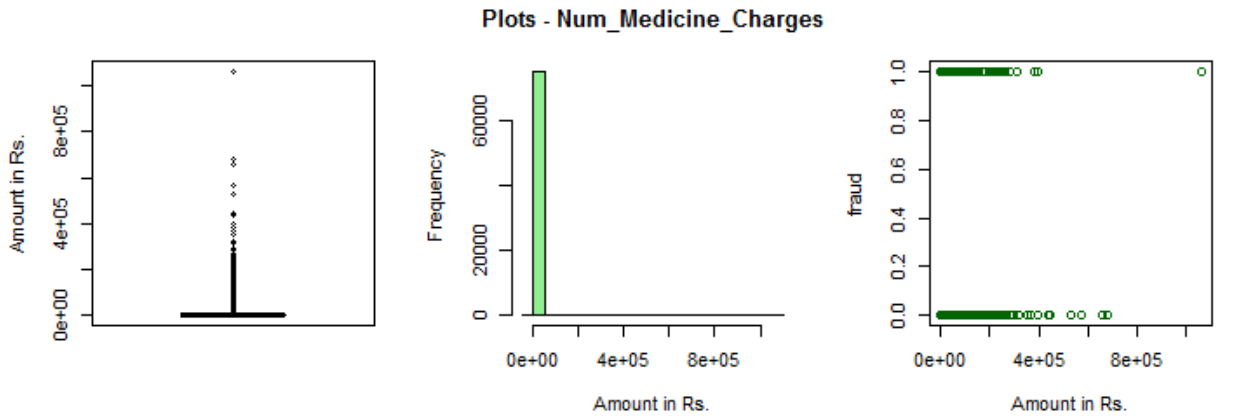


Comments:

Sum insured, surgery and room nursing charges are all left skewed and contain ots of zeroes. High extreme sum insured people are not fraudulent.

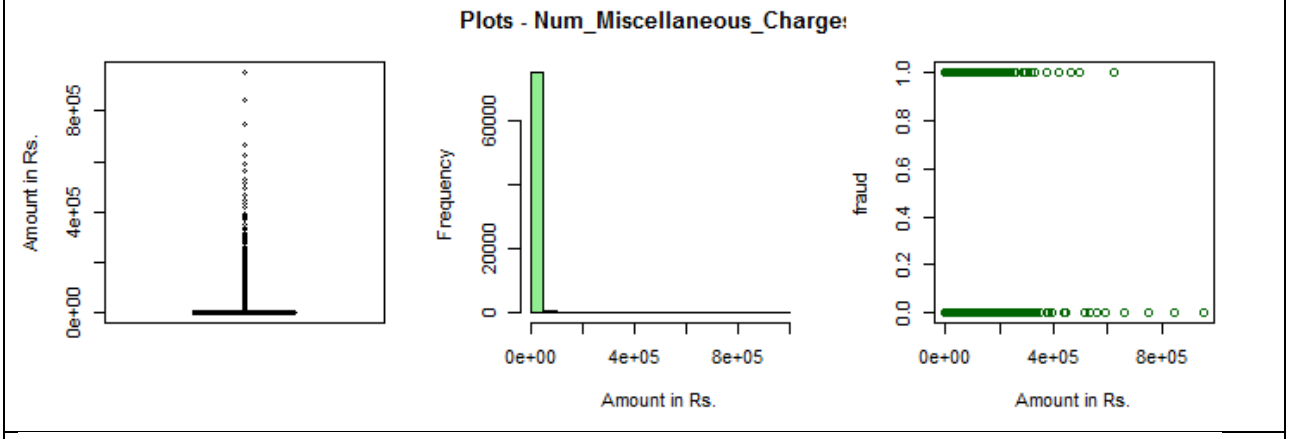


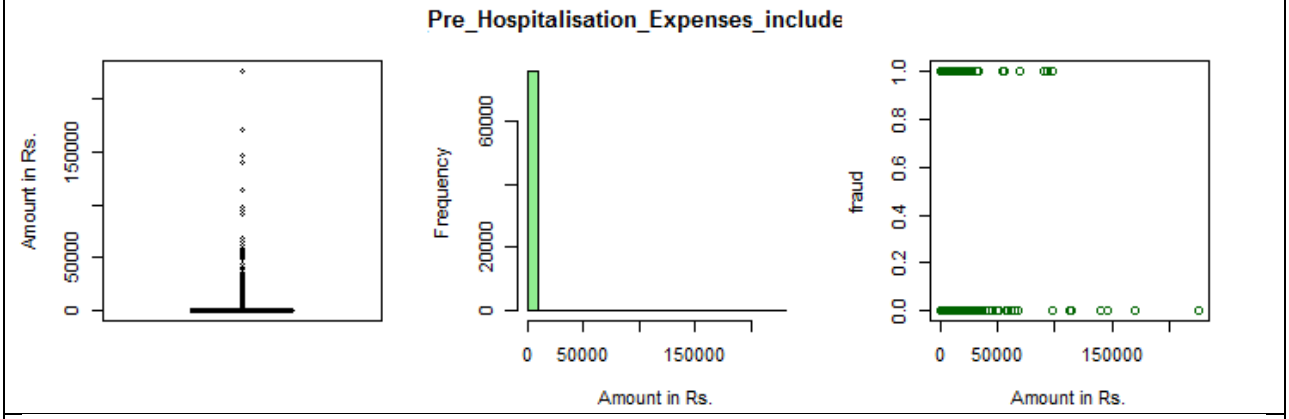


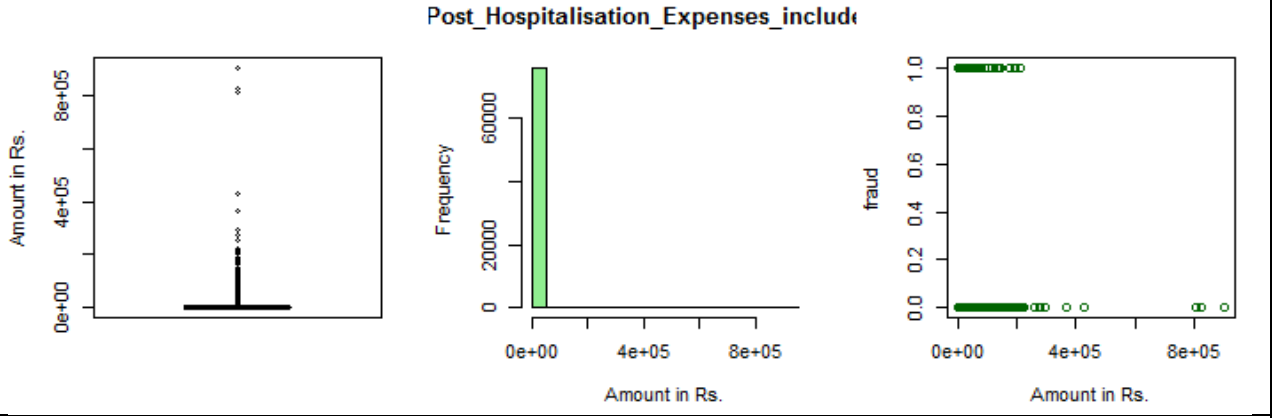


Comments:

Consultation,Investigation and Medicine charges are all left skewed. Both fraud and non fraudulent records are sparsely distributed after mid values of these charges.

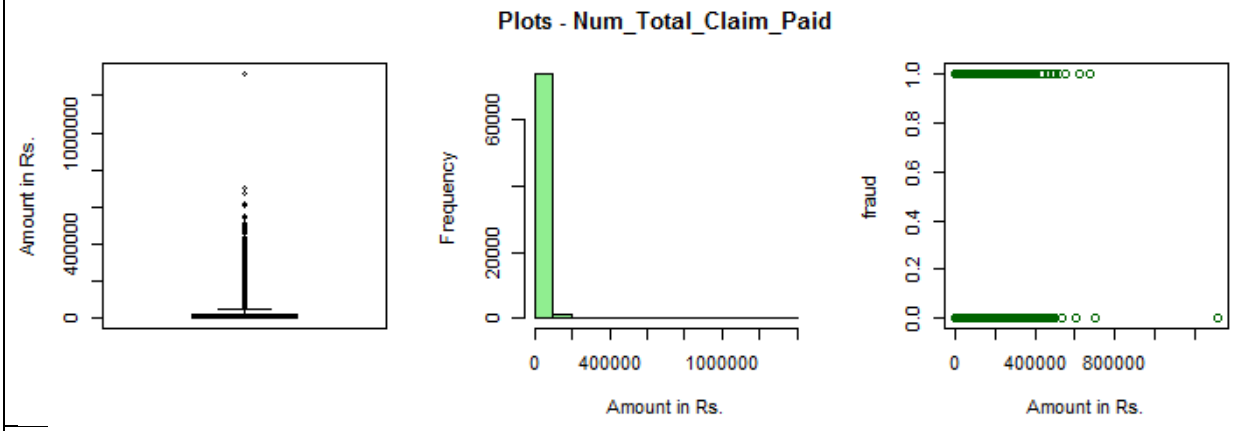


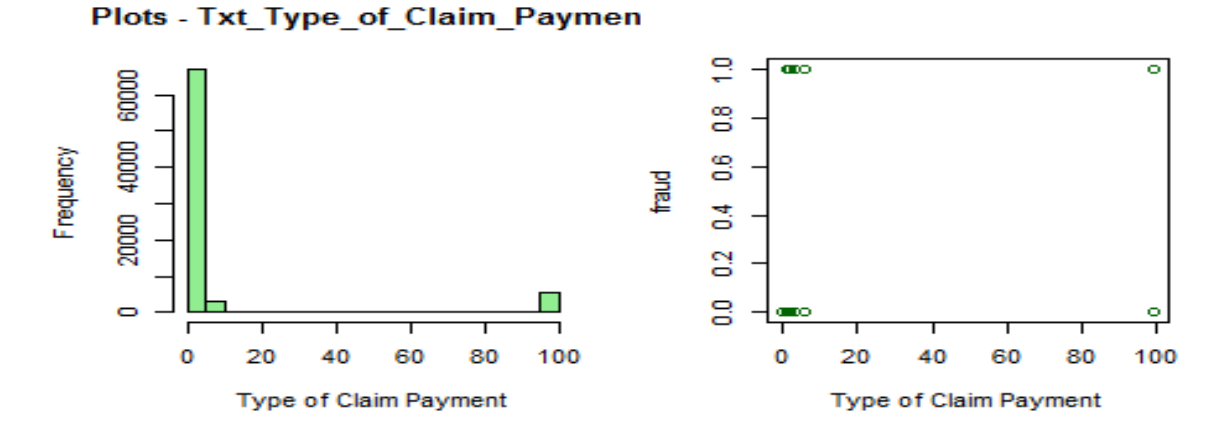


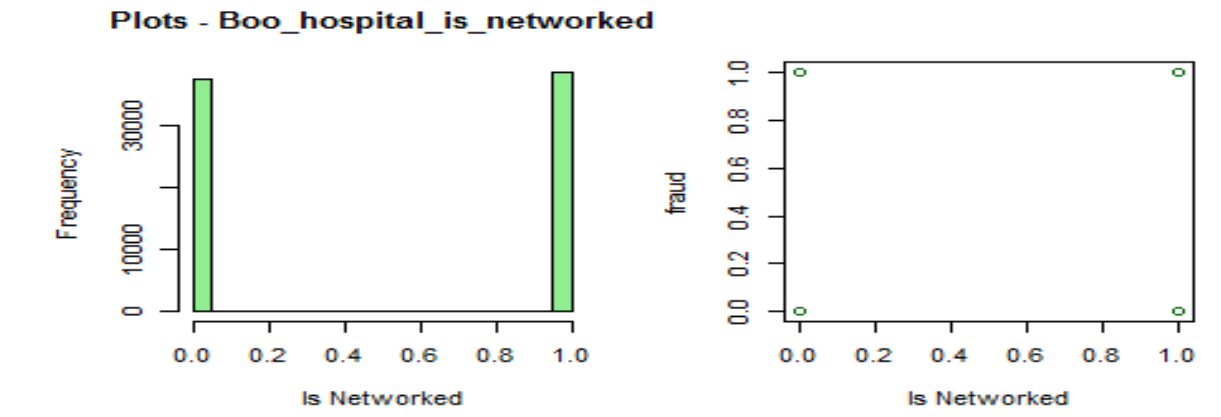


Comments:

Miscelleanous, Pre and Post hospitalization charges are all left skewed. Very high values of these charges are not fraud claims.Most of the below midvalues of these charges are fraudulent.

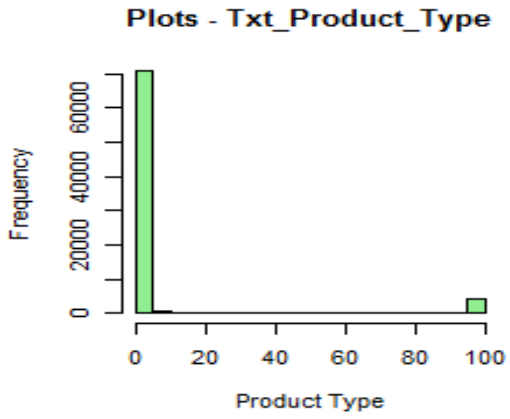


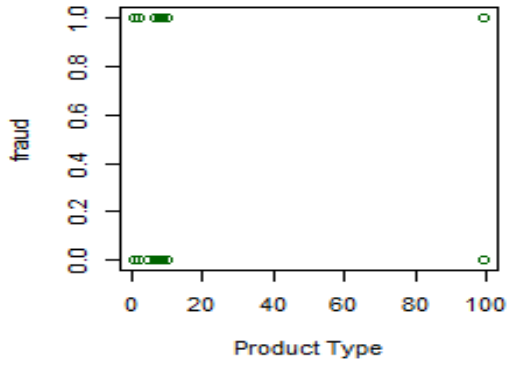


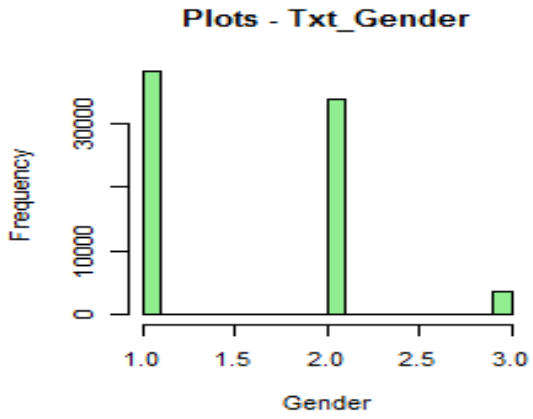


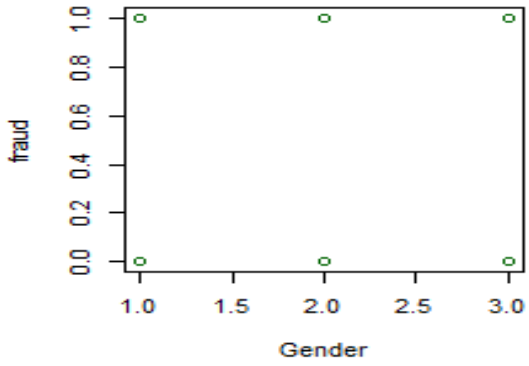
Comments:

All claim types and hospital networked or not has both fraudulent and non fraudulent claims.Total claim paid has both fraudulent and non fraudulent claims at almost for all its levels.





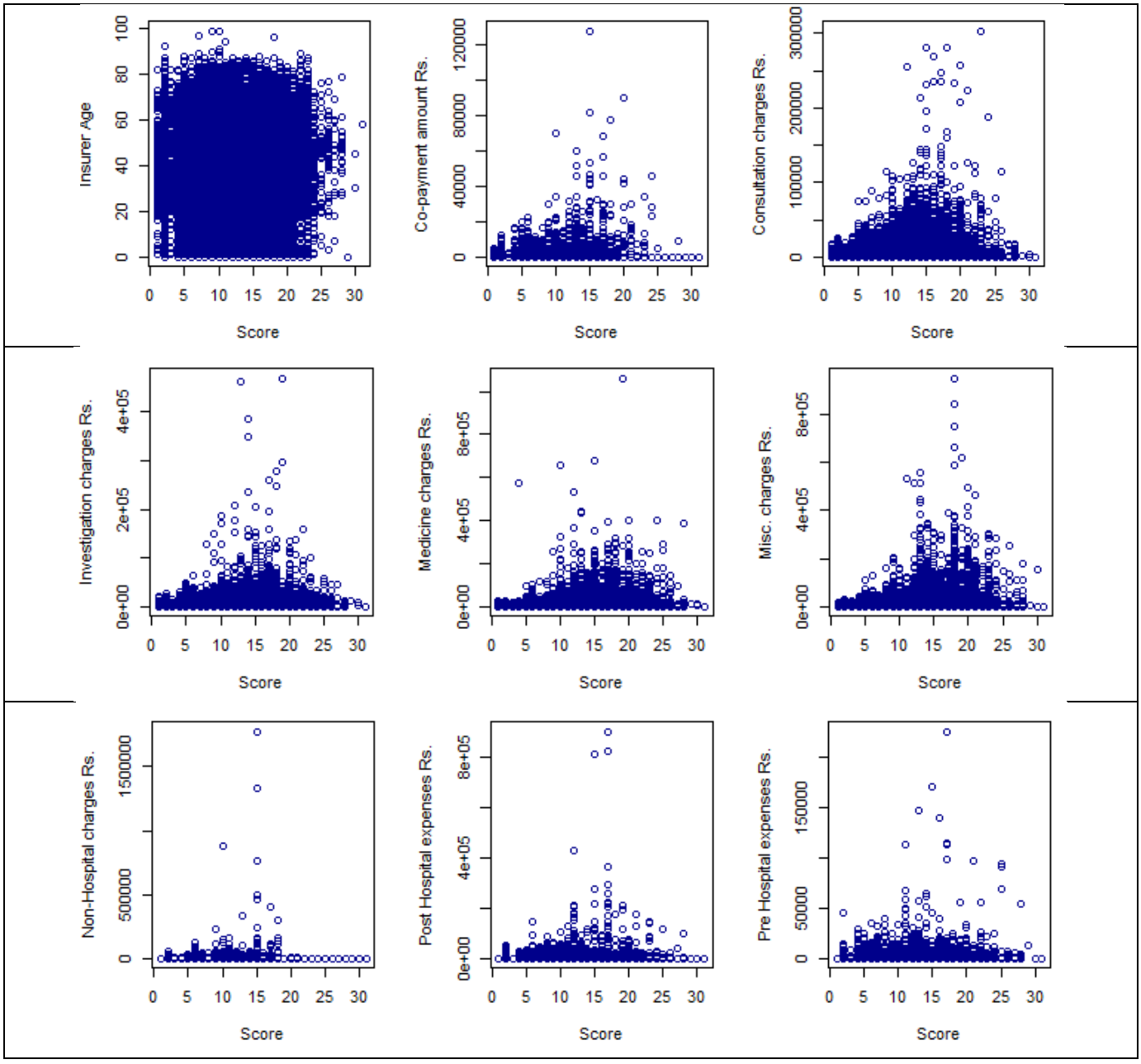


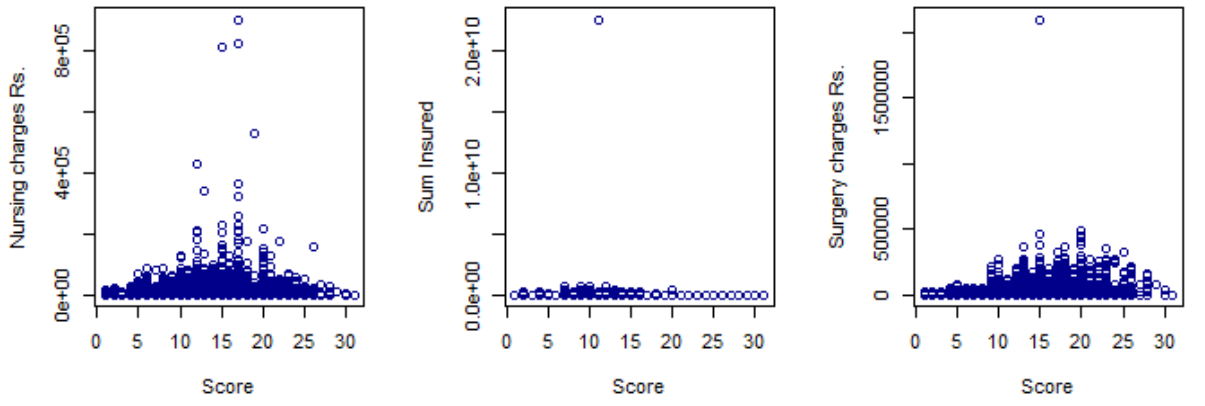


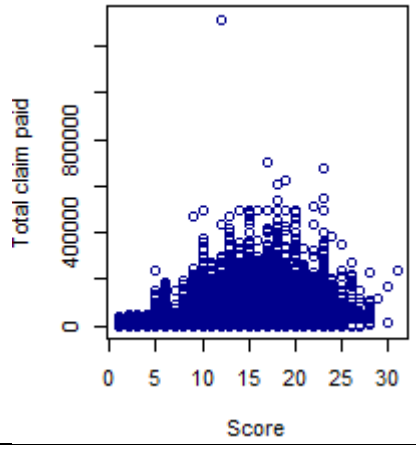
Comments:

All product types and Gender types have both fraudulent and nono fraudulent records in it.

Scatter Plot - Score Vs. Key quantitative variables







Comments:

Except Age of insured, non-hospital charges and sum insured all the other attributes have an almost similar curvy pattern of increasing from 0 to Midvale of the score and then decreasing till reaching the maximum score value. No linear relationship observed.

Scoring Model for Fraud Indicator Creation

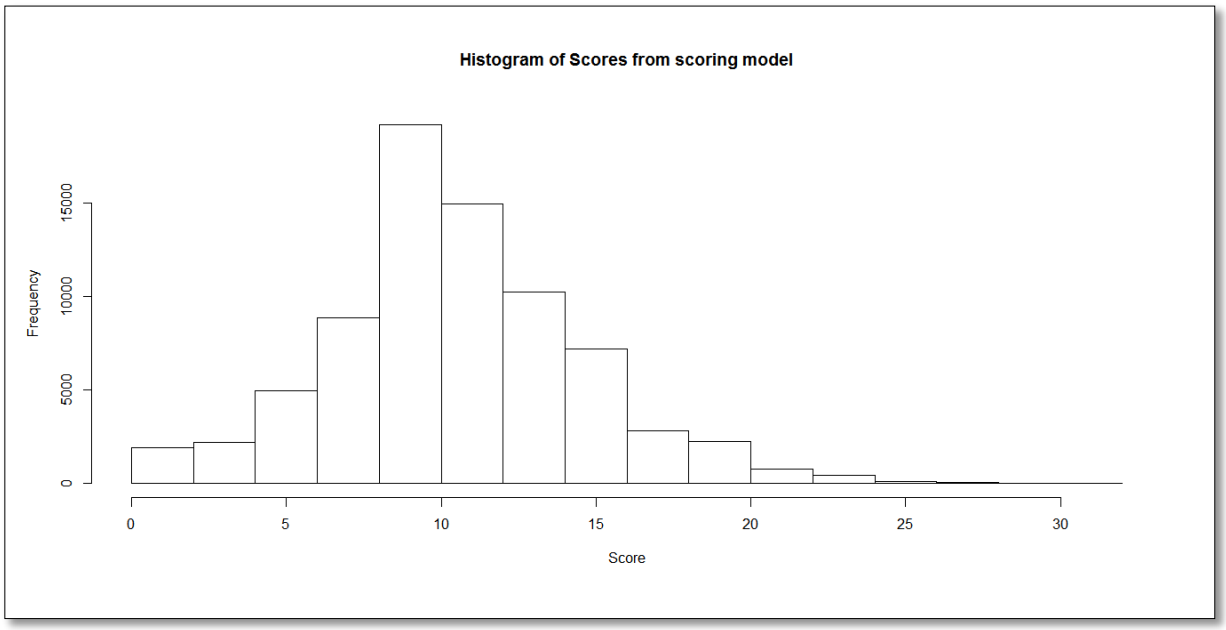
Business Rules and its weightage

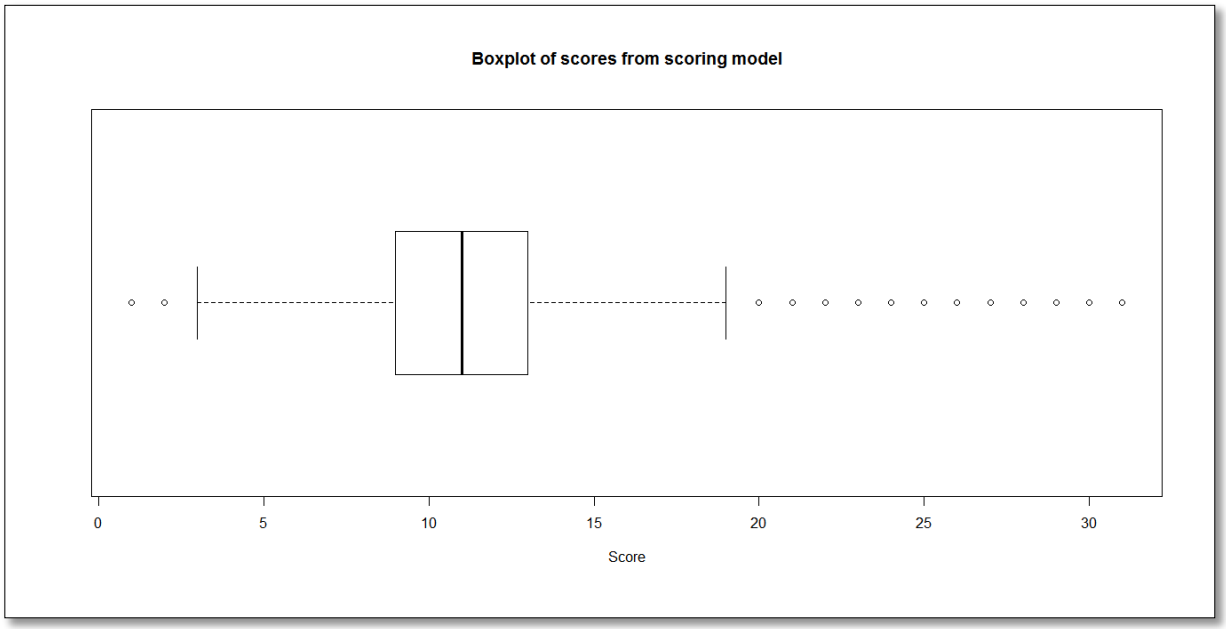
The following is the list of business rules used on the claims dataset to be used in the scoring model.

Business Rules	Rule #	Weightage
Reimbursement claims from Network Hospitals	1	3
Claim within first year of coverage, Single person, Single Insured, Minimum Insurance	2	3
Multiple claims from single family.	3	4
Claims related to Group medi-claim policy from same hospital	4	2
Repeated Hospitalization in same hospital within specific policy period./ or end of the policy period	5	2
High Value Claims I	6	4
Poor medical history (complaints not mentioned, only diagnosis mentioned on claim document	7	2
Fraud Prone Area	8	4
Claim intimation not given.	9	1
Claim submission on weekend(especially in case of Pre Auth)	10	3
High value claims/ bills. (Doctor Charges 50% of total bill)	11	3
Frequency of claims increased during last two months of the Policy	12	5
Skin Diseases	13	1
Dental Claims	14	1
All Lens prescription (Ophthalmology)	15	1
High Value Claims II	16	4
Bill Breakup not filled in the Form	17	2
High value claim for Infectious origin	18	4
Diagnosis not filled	19	1
Past history not filled in the form	20	1
PA claim intimated one day prior to discharge of patient	21	5

First claim intimation received after 48hours of admission	22	5
Claim intimation immediately within 30 days of date_policy_start	23	5

Distribution of score





Threshold for score

Summary of score

Min	1st Quartile	Median	Mean	3rd Quartile	Max
1	9	11	10.94	13	31

It is deemed in the world of insurance industry that the number of fraudulent claims are very less. Thus evaluating the scores above 3rd Quartile - 13 and Below Maximum value - 31 of Score.

Score - percentile above the given score

Score	# Of data points > score (%)
16	8.38
17	6.73
18	4.71
19	2.64
20	1.75

Comments on Threshold - Scoring

Using our knowledge on health insurance domain, our mentor's view and a little bit of browsing, we have fixed the threshold at score 18 in order to obtain the fraud records around 5% of the total records. Fraud indicator is created with fraud = 1 for records with score greater than 18 and fraud=0 for records with score lesser than or equal to 18.Total number of fraudulent records is 3569 out of 75838 records *Thus 4.71% of the records are fraud in the cleaned dataset.* This fraud attribute is the dependent variable in the model built to detect fraud in health insurance claims.

Scoring Model Validation using Multiple Linear Regression

Model 1

Model Properties

Property	Values
Attributes	score ~ Boo_hospital_is_networked + Num_Consultation_Charges + Num_Investigation_Charges + Num_Medicine_Charges + Num_Miscellaneous_Charges + Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges + Num_Total_Amount_Claimed + Num_Total_Claim_Paid + Txt_Type_of_Claim_Payment + Txt_Type_of_Policy

Model output

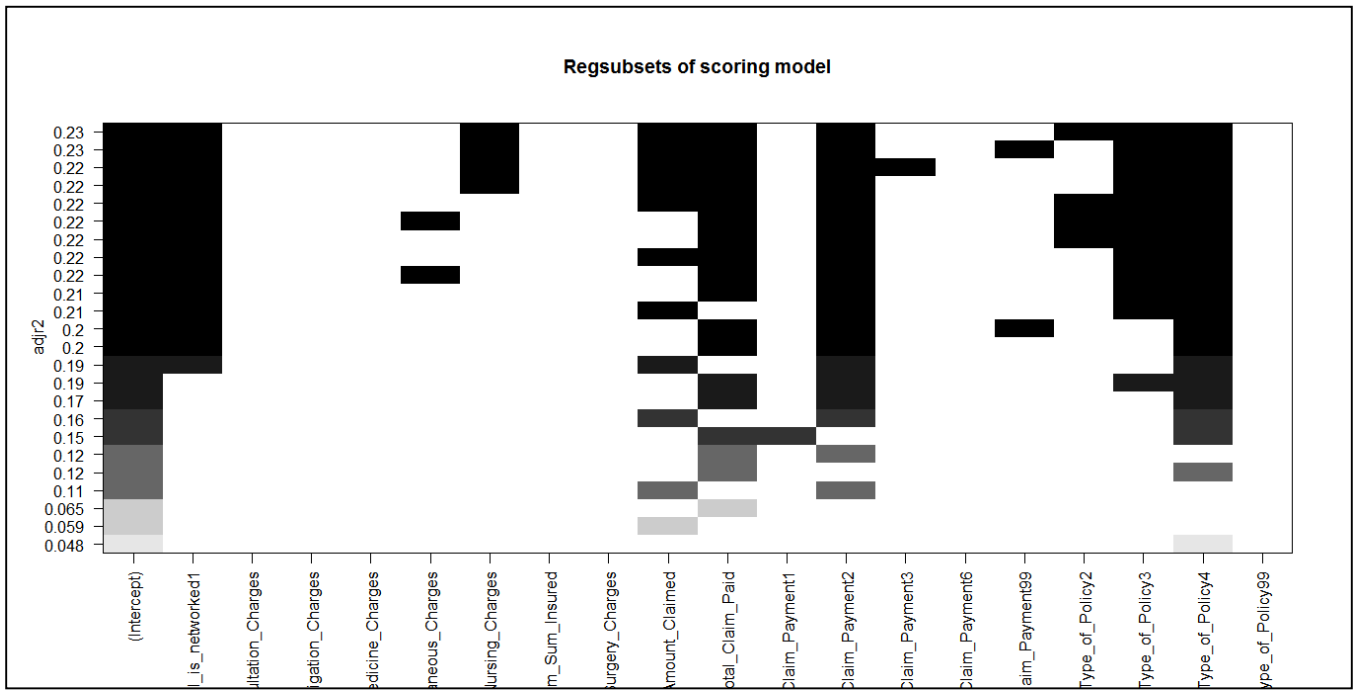
```
Call:
lm(formula = score ~ Boo_hospital_is_networked + Num_Consultation_Charges +
    Num_Investigation_Charges + Num_Medicine_Charges + Num_Miscellaneous_Charges +
    Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges +
    Num_Total_Amount_Claimed + Num_Total_Claim_Paid + Txt_Type_of_Claim_Payment +
    Txt_Type_of_Policy, data = claim)

Residuals:
    Min      1Q  Median      3Q     Max
-35.800  -2.377  -0.028   2.102  18.327

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 6.665e+00  3.553e+00   1.876   0.0607 .
Boo_hospital_is_networked1  1.946e+00  3.723e-02  52.278  < 2e-16 ***
Num_Consultation_Charges   -1.275e-05  1.903e-06  -6.699 2.11e-11 ***
Num_Investigation_Charges  -1.134e-05  2.178e-06  -5.206 1.93e-07 ***
Num_Medicine_Charges       -6.935e-06  1.189e-06  -5.835 5.40e-09 ***
Num_Miscellaneous_Charges   7.802e-06  8.989e-07   8.679  < 2e-16 ***
Num_Room_Nursing_Charges   -3.150e-05  1.757e-06 -17.923  < 2e-16 ***
Num_Sum_Insured             3.017e-11  1.102e-10   0.274   0.7842
Num_Surgery_Charges        -1.121e-06  9.074e-07  -1.236   0.2165
Num_Total_Amount_Claimed    1.381e-05  6.237e-07  22.148  < 2e-16 ***
Num_Total_Claim_Paid        2.275e-05  7.517e-07  30.262  < 2e-16 ***
Txt_Type_of_Claim_Payment1  2.237e+00  3.553e+00   0.630   0.5290
Txt_Type_of_Claim_Payment2  5.381e+00  3.553e+00   1.514   0.1299
Txt_Type_of_Claim_Payment3  3.525e+00  3.555e+00   0.991   0.3215
Txt_Type_of_Claim_Payment6  1.621e+00  3.553e+00   0.456   0.6483
Txt_Type_of_Claim_Payment99 3.171e+00  3.553e+00   0.892   0.3721
Txt_Type_of_Policy2        -7.654e-01  4.601e-02 -16.634  < 2e-16 ***
Txt_Type_of_Policy3        -2.204e+00  5.270e-02 -41.820  < 2e-16 ***
Txt_Type_of_Policy4        -2.566e+00  3.678e-02 -69.758  < 2e-16 ***
Txt_Type_of_Policy99       -4.956e-01  4.969e-02  -9.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.553 on 75818 degrees of freedom
Multiple R-squared:  0.2329, Adjusted R-squared:  0.2327
F-statistic:  1212 on 19 and 75818 DF,  p-value: < 2.2e-16
```

Model Interpretation

Boo_hospital_is_networked1, Num_Consultation_Charges, Num_Investigation_Charges, Num_Medicine_Charges, Num_Miscellaneous_Charges, Num_Room_Nursing_Charges, Num_Total_Amount_Claimed, Num_Total_Claim_Paid, all of Txt_Type_of_Policy categories are highly significant variables.

23.29% of the variation of score attribute is explained by the subset of covariates used to build the scoring model. Adjusted R squared is 0.2327.

Regsubsets plot



RegSubsets Interpretation

We have chosen the fourth model from above with Adjusted R squared=0.22 as it has minimal variables and a comparatively high Adjusted R squared. The selected variables are Boo_hospital_is_networked, Num_Total_Amount_Claimed, Num_Total_Claim_Paid, Txt_Type_of_Claim_Payment and Txt_Type_of_Policy from Regsubset.

Model 2

Model Properties

Property	Values
Attributes	Score~Boo_hospital_is_networked+Num_Total_Amount_Claimed+Num_Total_Claim_Paid+Txt_Type_of_Claim_Payment+Txt_Type_of_Policy

Model output

```
Call:
lm(formula = score ~ Boo_hospital_is_networked + Num_Total_Amount_Claimed +
    Num_Total_Claim_Paid + Txt_Type_of_Claim_Payment + Txt_Type_of_Policy,
    data = claim)

Residuals:
    Min      1Q  Median      3Q     Max
-39.607  -2.382  -0.008   2.105  18.338

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.581e+00  3.570e+00   1.844   0.0652 .
Boo_hospital_is_networked1 1.944e+00  3.737e-02  52.010   <2e-16 ***
Num_Total_Amount_Claimed   8.363e-06  4.917e-07  17.010   <2e-16 ***
Num_Total_Claim_Paid       2.456e-05  7.345e-07  33.437   <2e-16 ***
Txt_Type_of_Claim_Payment1 2.311e+00  3.570e+00   0.648   0.5173
Txt_Type_of_Claim_Payment2 5.491e+00  3.570e+00   1.538   0.1240
Txt_Type_of_Claim_Payment3 3.655e+00  3.572e+00   1.023   0.3062
Txt_Type_of_Claim_Payment6 1.725e+00  3.570e+00   0.483   0.6290
Txt_Type_of_Claim_Payment99 3.176e+00  3.570e+00   0.890   0.3736
Txt_Type_of_Policy2       -8.097e-01  4.586e-02 -17.654   <2e-16 ***
Txt_Type_of_Policy3       -2.262e+00  5.264e-02 -42.967   <2e-16 ***
Txt_Type_of_Policy4       -2.623e+00  3.613e-02 -72.598   <2e-16 ***
Txt_Type_of_Policy99      -4.276e-01  4.978e-02  -8.590   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.569 on 75825 degrees of freedom
Multiple R-squared:  0.2256, Adjusted R-squared:  0.2255
F-statistic:  1841 on 12 and 75825 DF,  p-value: < 2.2e-16
```

Model Interpretation

Though the Adjusted R squared has decreased to 0.2255 in Model 2 compared to that of 0.2327 in Model 1, the number of covariates is reduced to 5 in Model 2 than that of 12 in Model 1. We have achieved a parsimonious model. All the variable used in the model except Txt_Type_of_Claim_Payment is highly significant. But if we remove Txt_Type_of_Claim_Payment from model the Adjusted R squared gets reduced to 0.14 from 0.22 thus it's not advisable to remove Txt_Type_of_Claim_Payment from the linear scoring model as it contributes as whole rather than in categorical form. Thus we can conclude that all the covariates mentioned in the model significantly affect the score.

Advanced Analytical Modelling

After having performing the tasks of: exploratory data analysis, data cleaning, creation of fraud indicator using business rule based scoring model method, validation of scoring (model) using a multiple linear regression model, we have reached the stage of making fraud detection model. We brainstormed the various possible classification models that can be used in fraud detection. Post which we discussed the same with our mentor as well and finalized the list of models to be tried. They are

a) Logistic regression
b) Neural networks
c) Random forests – Bagging

Each of the above mentioned models are tried on the cleaned claim dataset and different variants of the same models are documented below.

Sampling

After cleaning the dataset the raw dataset which contained 100,000 records is reduced down to 75,838. In order to train the models the cleaned data set is broken down to train and test datasets in the ratio of 70:30 respectively.

Dataset	Number of records	Number of fraudulent records
Raw	100,000	NA (fraud indicator was missing)
Cleaned Data set	75,838	3569 (4.7% are classified as fraud)
Train	53,086	2524 (4.7%)
Test	22,752	1045 (4.6%)

Logistic Regression

As part of model building, we have applied logistic regression with fraud variable as dependent variables and other variables as independent variables. The independent variables which are included in the model along with model parameters are mentioned below.

- Two iterations of logistic regression have been applied. In first iteration, all the 17 variables have been included as independent variables.
- In second iteration, all insignificant variables based on 5% cutoff have been removed.

Model 1

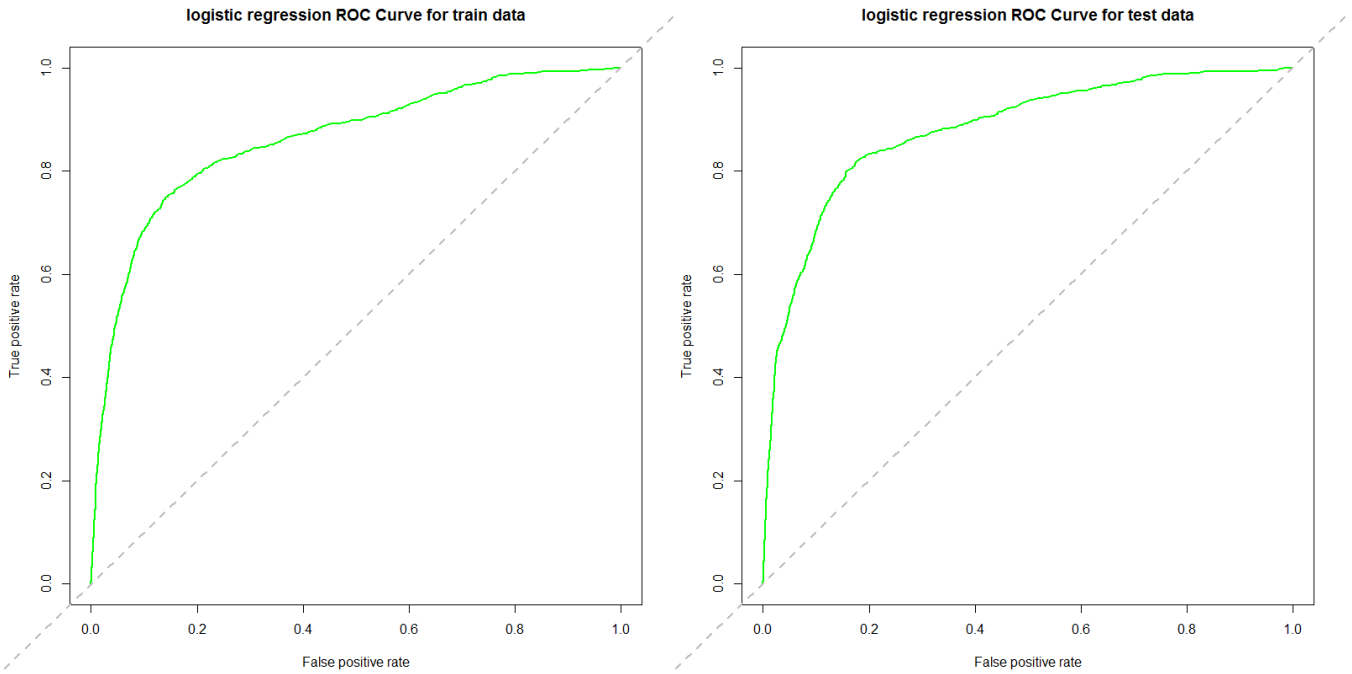
Model Properties

Property	Values
Attributes	Claim$fraud~Boo_hospital_is_networked + Num_Age_of_Insured + Num_Amount_of_Co_Payment_or_Excess_if_applicable + Num_Consultation_Charges + Num_Investigation_Charges + Num_Medicine_Charges + Num_Miscellaneous_Charges + Num_Other_Non_Hospital_Expenses + Num_Post_Hospitalisation_Expenses_included_under_150035 + Num_Pre_Hospitalisation_Expenses_included_under_150035 + Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges + Num_Total_Claim_Paid + Txt_Gender + Txt_Product_Type + Txt_Type_of_Claim_Payment
No. of variables	17
No. of Records	53086
No. of Fraudulent cases	2524

Model Results

	Logistic Regression - Model 1 - Train	Logistic Regression – Model 1 – Test
Specificity	0.9	0.9
Sensitivity	0.58	0.58
Confusion Matrix	Actual / Predicted 0: 0 → 29401, 1 → 250 1: 0 → 21161, 1 → 2274	Actual / Predicted 0: 0 → 12708, 1 → 100 1: 0 → 8999, 1 → 945

ROC Curve



	Train	Test
Area of ROC Curve	0.875	0.879

Model Interpretation

- Decent output in terms of specificity and sensitivity which will help in identifying fraudulent cases as well as minimizing false non-fraudulent cases.
- The ROC curve area is better for test rather than the train dataset.
- In case of fraudulent claim classification, the model prediction accuracy is very good going by confusion matrix and area under roc curve.
- However, there are some insignificant variables in the model which we will remove and run another iteration.

Model 2

Model Properties

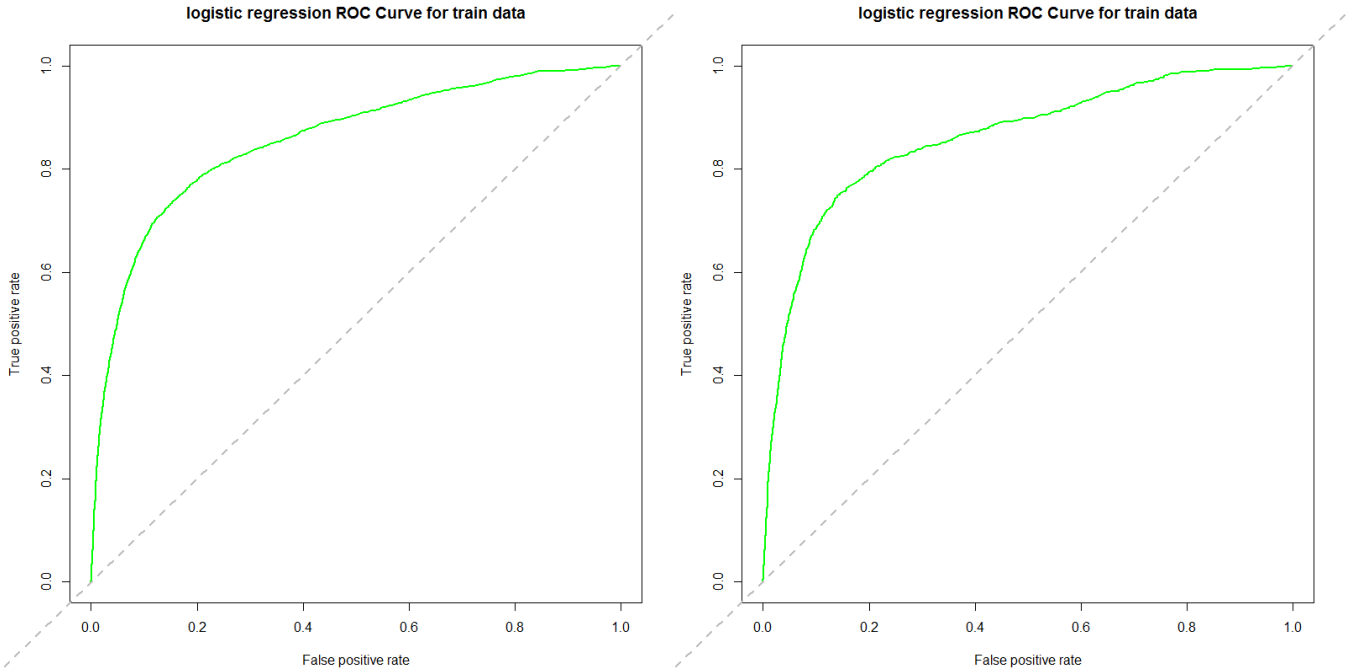
In second iteration, 3 insignificant variables have been removed from the model.

Property	Values
Attributes	Claim$fraud~Boo_hospital_is_networked + Num_Age_of_Insured + Num_Consultation_Charges + Num_Investigation_Charges + Num_Medicine_Charges + Num_Other_Non_Hospital_Expenses + Num_Post_Hospitalisation_Expenses_included_under_150035 + Num_Pre_Hospitalisation_Expenses_included_under_150035 + Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges + Num_Total_Claim_Paid + Txt_Gender + Txt_Product_Type
No. of variables	14
No. of Records	53086
No. of Fraudulent cases	2524

Model Results

	Logistic Regression - Model 2 - Train	Logistic Regression – Model 2– Test
Specificity	0.9	0.8985
Sensitivity	0.49	0.4965
Confusion Matrix	Actual / Predicted 0: 24986, 236 1: 25576, 2288	Actual / Predicted 0: 10778, 106 1: 10929, 939

ROC Curve



	Train	Test
Area of ROC Curve	0.85	0.86

Model Interpretation

- Decent output in terms of specificity and sensitivity which will help in identifying fraudulent cases as well as minimizing false non-fraudulent cases.
- The ROC curve area is better for test rather than the train dataset.
- In case of fraudulent claim classification, the model prediction accuracy is very good going by confusion matrix and area under roc curve.

Neural Network

For model building, we have incorporated neural network with fraud variable as dependent variable. The covariates which are included in the model along with model parameters are mentioned below.

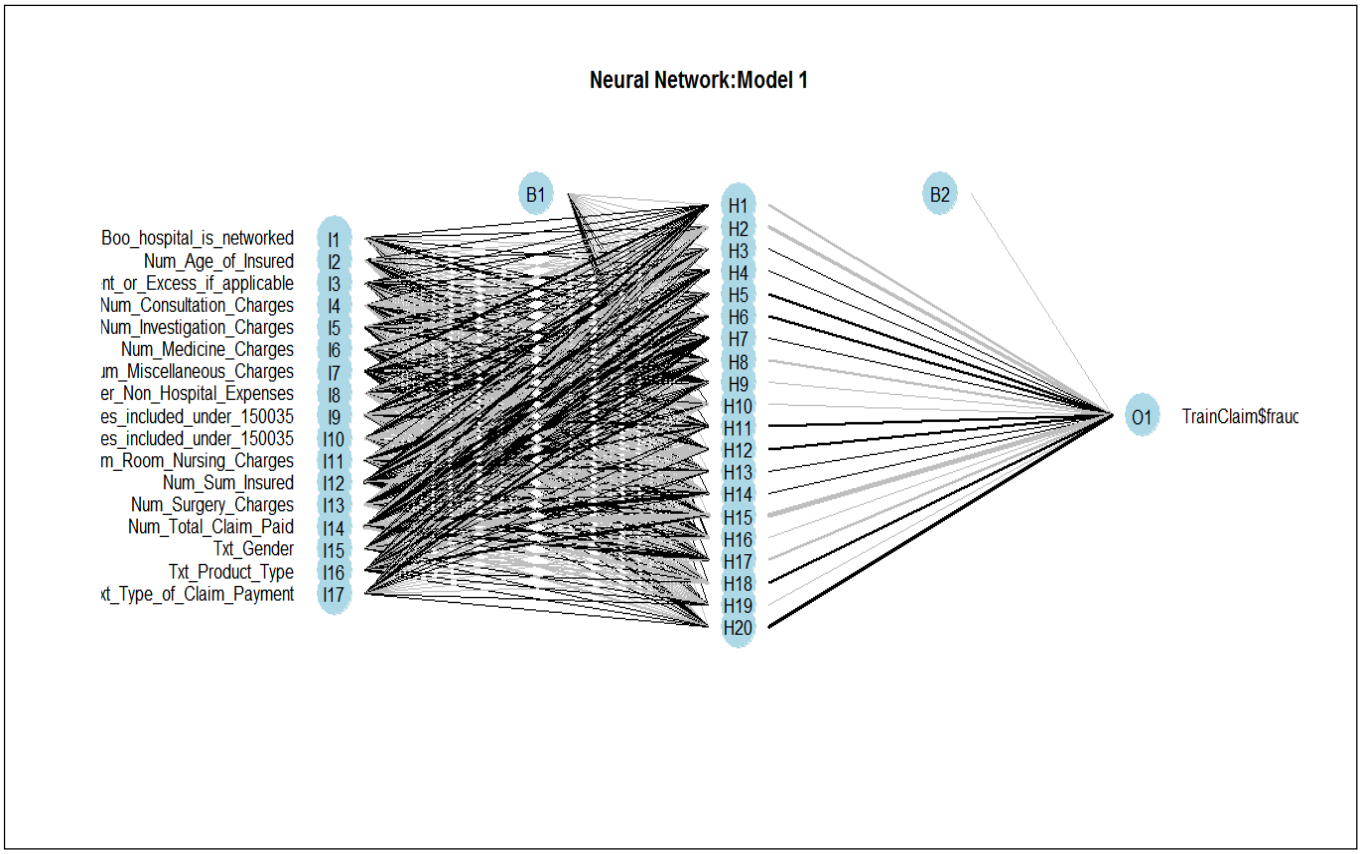
- Two models of neural network have been built. Both models contain 17 attributes as independent variables.
- Major change in neural network Model 2 compared to that of Model 1 is weighing cases.

Model 1

Model Properties

Property	Values
Attributes	TrainClaim$fraud ~ Boo_hospital_is_networked + Num_Age_of_Insured + Num_Amount_of_Co_Payment_or_Excess_if_applicable + Num_Consultation_Charges + Num_Investigation_Charges + Num_Medicine_Charges + Num_Miscellaneous_Charges + Num_Other_Non_Hospital_Expenses + Num_Post_Hospitalisation_Expenses_included_under_150035 + Num_Pre_Hospitalisation_Expenses_included_under_150035 + Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges + Num_Total_Claim_Paid + Txt_Gender + Txt_Product_Type + Txt_Type_of_Claim_Payment
Size	20
Max iterations	10000
Decay	0.001

Model Results

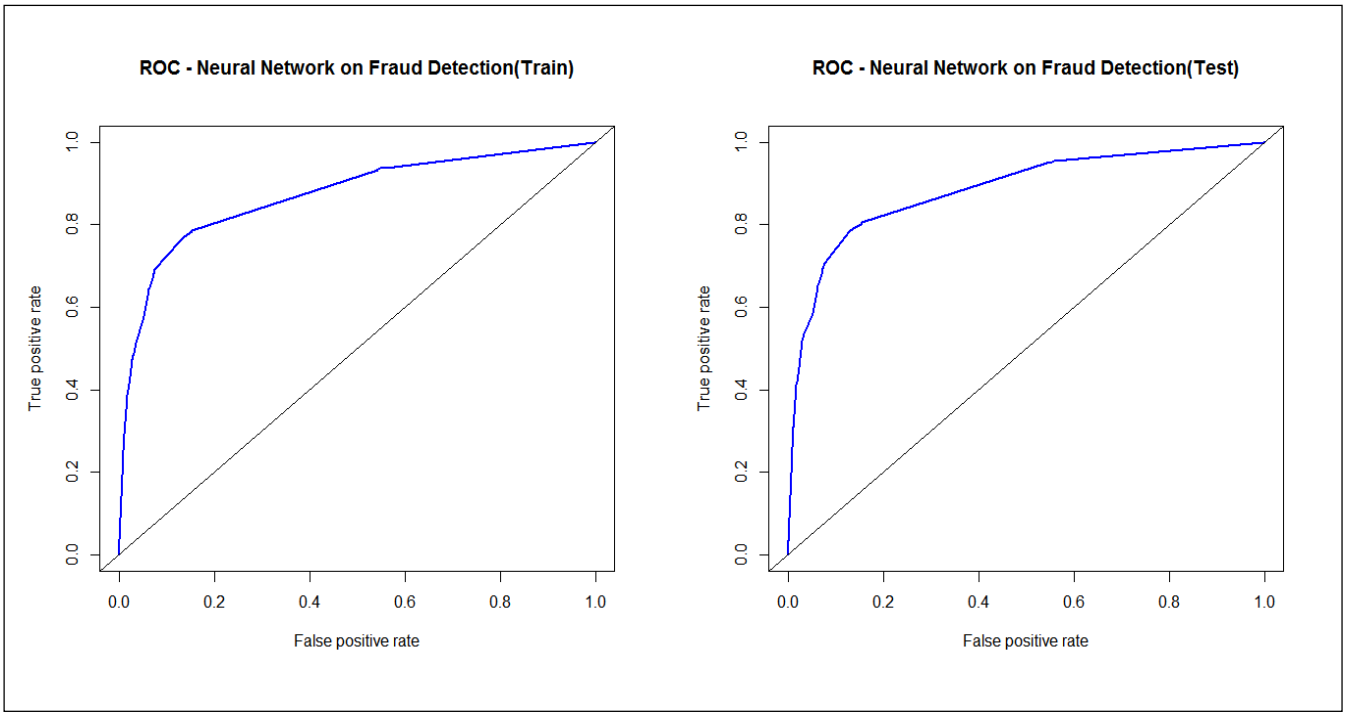


	Neural Network - Model 1 - Train	Neural Network – Model 1 – Test
Specificity	0.25	0.26
Sensitivity	0.99	0.99
Confusion Matrix	**Actual**	**Actual**

Predicted \ Actual	0	1
0	50083	1885
1	479	639

Predicted \ Actual	0	1
0	21503	776
1	204	269

ROC Curve



	Train	Test
Area of ROC Curve	0.871	0.885

Model Interpretation

- Reasonable output in sensitivity and identifying the non-fraudulent cases.
- The ROC curve area is better for test rather than the train dataset.
- In classification of fraudulent cases, the model prediction accuracy and the specificity is very nominal.
- Area of ROC curve is not bad but it the ROC curve is not smooth

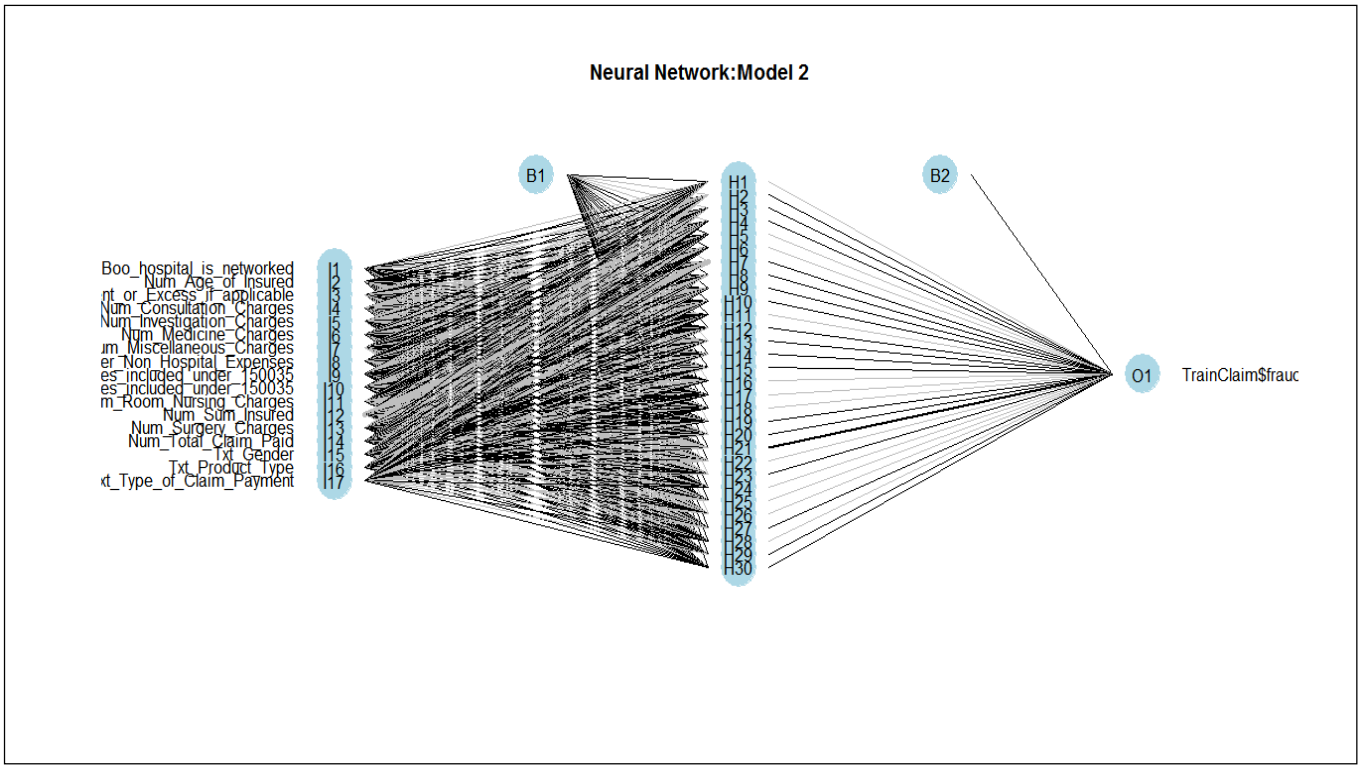
Model 2

In this model we have introduced weighing of fraudulent cases in model building. Increase in number of nodes will result in improved model performance is the conceived notion. Thus we have increased the number of nodes in Model 2 compared to that of Model 1. In order to give space for the same we have increased the maximum iterations limit too.

Model Properties

Property	Values
Attributes	TrainClaim$fraud ~ Boo_hospital_is_networked + Num_Age_of_Insured + Num_Amount_of_Co_Payment_or_Excess_if_applicable + Num_Consultation_Charges + Num_Investigation_Charges + Num_Medicine_Charges + Num_Miscellaneous_Charges + Num_Other_Non_Hospital_Expenses + Num_Post_Hospitalisation_Expenses_included_under_150035 + Num_Pre_Hospitalisation_Expenses_included_under_150035 + Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges + Num_Total_Claim_Paid + Txt_Gender + Txt_Product_Type + Txt_Type_of_Claim_Payment
Size	30
Max iterations	20000
Weights	3 for fraudulent cases
Decay	0.001

Model Results

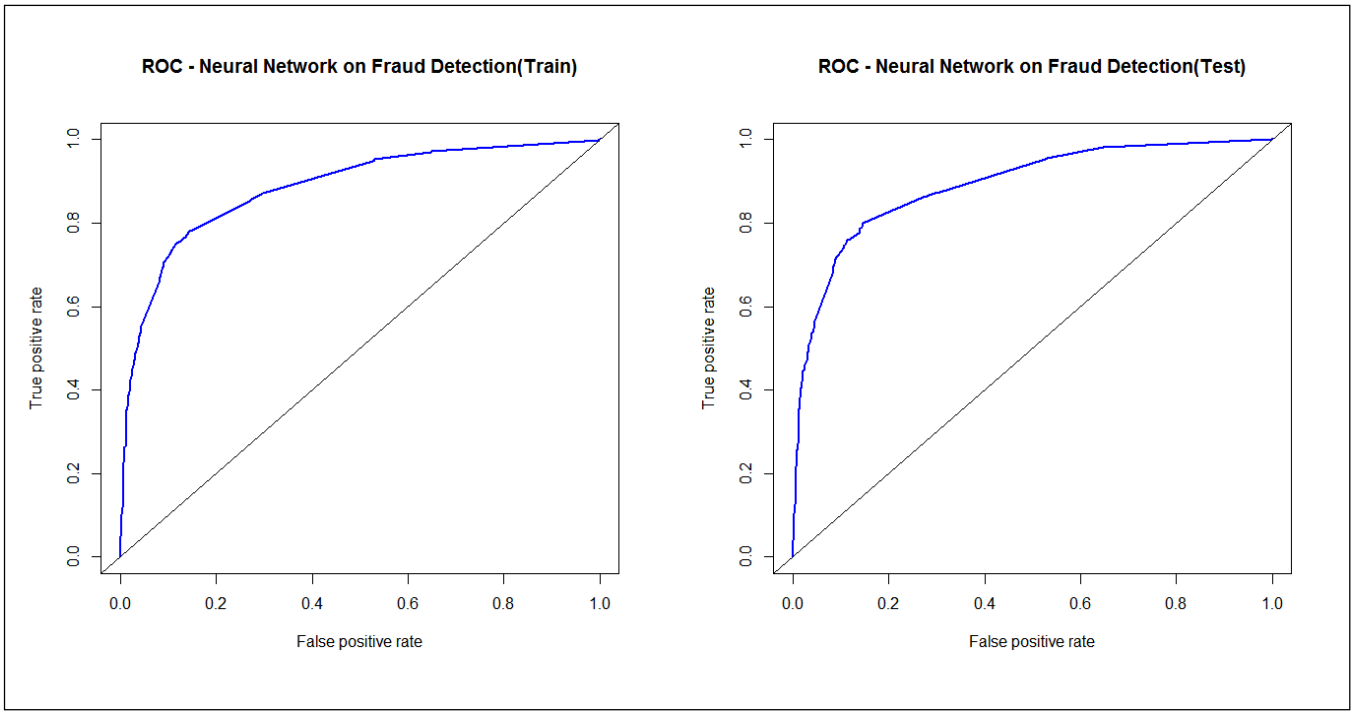


	Neural Network - Model 2 - Train	Neural Network – Model 2 – Test
Specificity	0.38	0.40
Sensitivity	0.98	0.98
Confusion Matrix	**Actual**	**Actual**

		Actual 0	Actual 1
	Predicted 0	49692	1555
	Predicted 1	870	969

		Actual 0	Actual 1
	Predicted 0	21334	625
	Predicted 1	373	420

ROC Curve



	Train	Test
Area of ROC Curve	0.884	0.890

Model Interpretation

- Specificity of fraudulent claims has increased in Model 2 compared to that of Model 1.
- Sensitivity is decreased around 0.01 in Model 2 compared to that of Model 1.But it's manageable as decrease in sensitivity in Model 2 is very less.
- Area of ROC curve has increased along with smoother ROC curve which is a good indication that Model 2 is definitely better.
- For Neural network, Model 2 has performed better than Model 1.

Random Forest

As part of model building, we have applied Random Forest technique with fraud variable as dependent variables and other variables as independent variables. The independent variables which are included in the model along with model parameters are mentioned below.

Three iterations of Random Forest model with different tree sizes (100,500,750) have been carried out. When the tree size grows to 1000, system will not able to allocate memory to process such a huge no. of trees.

Model 1

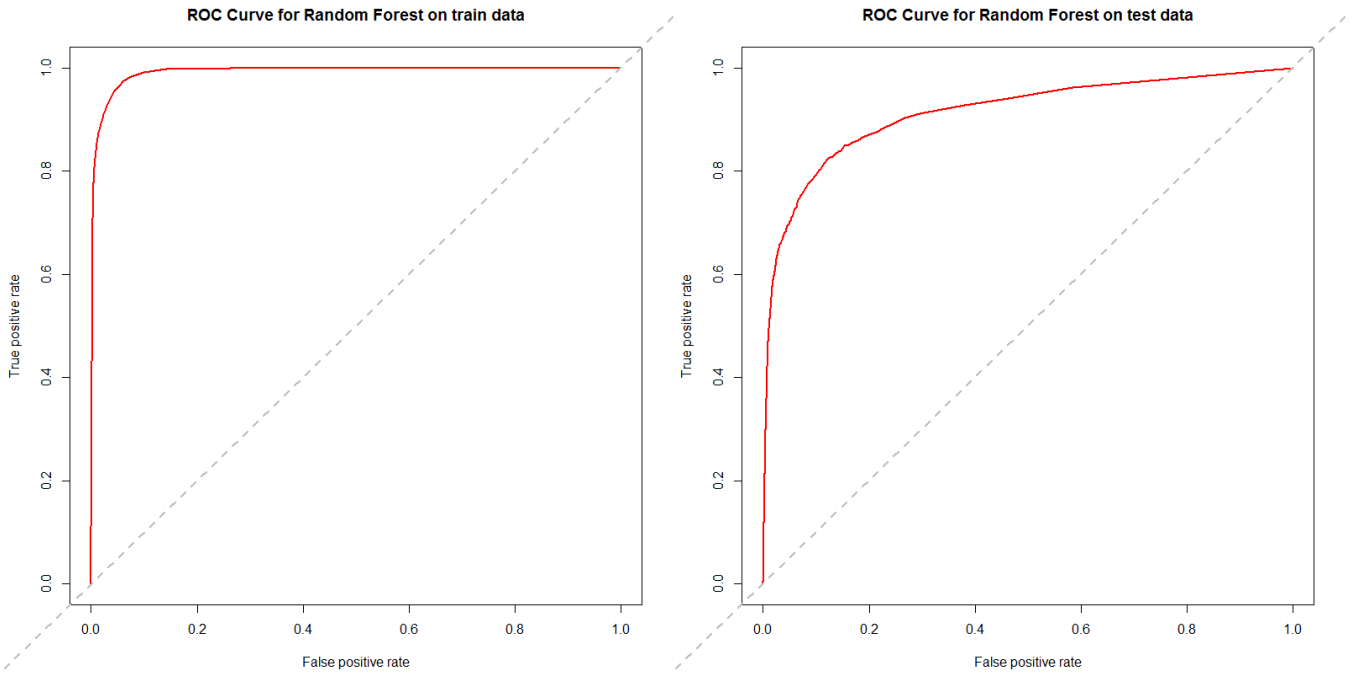
Model Properties

Property	Values
Attributes	Claim$fraud~Boo_hospital_is_networked + Num_Age_of_Insured + Num_Amount_of_Co_Payment_or_Excess_if_applicable + Num_Consultation_Charges + Num_Investigation_Charges + Num_Medicine_Charges + Num_Miscellaneous_Charges + Num_Other_Non_Hospital_Expenses + Num_Post_Hospitalisation_Expenses_included_under_150035 + Num_Pre_Hospitalisation_Expenses_included_under_150035 + Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges + Num_Total_Claim_Paid + Txt_Gender + Txt_Product_Type + Txt_Type_of_Claim_Payment
No. of variables	17
No. of Records	53086
No. of Trees	100
Mtry	3

Model Results

	Random Forest- Model 1 - Train	Logistic Regression – Model 1 – Test
Specificity	0.9965	0.9912
Sensitivity	0.7282	0.3885
Confusion Matrix	Actual / Predicted 0: 0 → 50124, 1 → 686 1: 0 → 178, 1 → 1838	Actual / Predicted 0: 0 → 21512, 1 → 639 1: 0 → 195, 1 → 406

ROC Curve



	Train	Test
Area of ROC Curve	0.9137	0.8814

Model Interpretation

- Random forest model has very good accuracy compared to logistic regression models. Though there is decrease in accuracy on test data set, still the model outperforms logistic regression model.
- The ROC curve area is better for train data set rather than the train dataset. However, there no significant reduction in AUC for test data set compared to train data set. This indicates good accuracy of classification.

- We will try to increase the no. of trees to see if we can obtain any improvement in the model.

Model 2

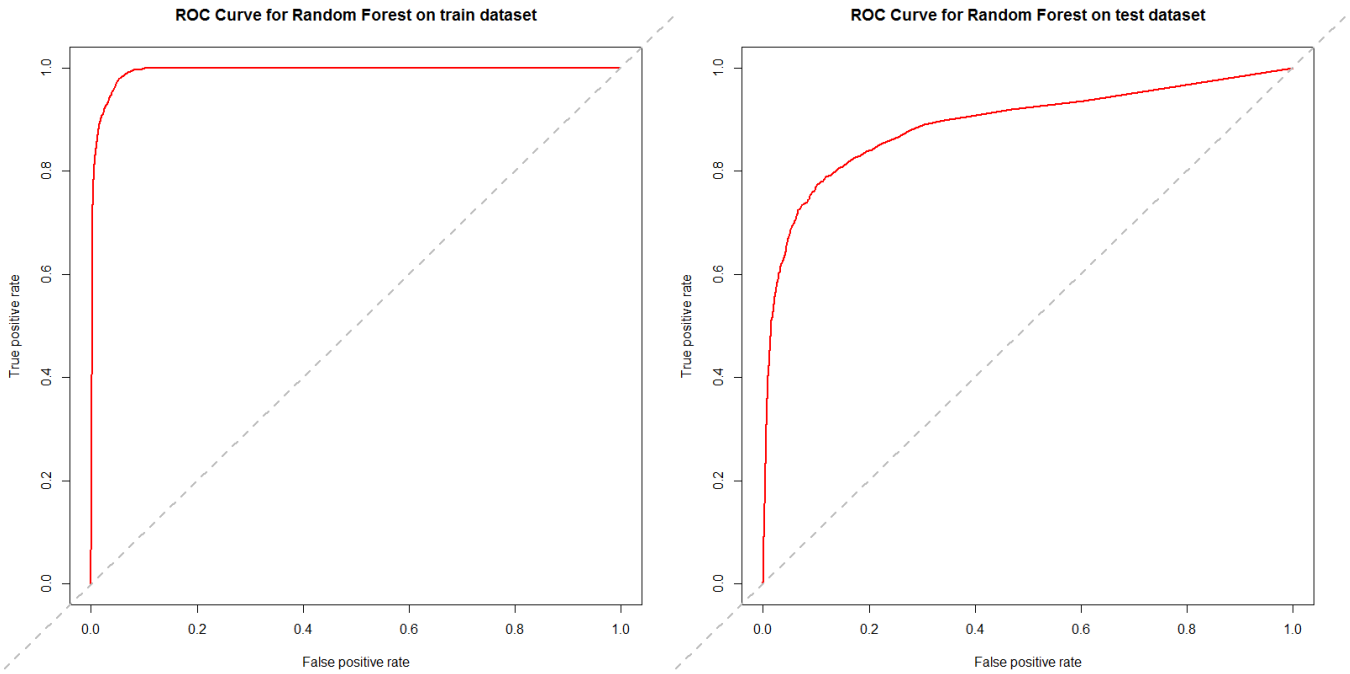
Model Properties

Property	Values
Attributes	Claim$fraud~Boo_hospital_is_networked + Num_Age_of_Insured + Num_Amount_of_Co_Payment_or_Excess_if_applicable + Num_Consultation_Charges + Num_Investigation_Charges + Num_Medicine_Charges + Num_Miscellaneous_Charges + Num_Other_Non_Hospital_Expenses + Num_Post_Hospitalisation_Expenses_included_under_150035 + Num_Pre_Hospitalisation_Expenses_included_under_150035 + Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges + Num_Total_Claim_Paid + Txt_Gender + Txt_Product_Type + Txt_Type_of_Claim_Payment
No. of variables	17
No. of Records	53086
No. of Trees	500
Mtry	3

Model Results

	Random Forest- Model 2 - Train	Random Forest – Model 2– Test
Specificity	0.7365	0.3742
Sensitivity	0.9967	0.9915
Confusion Matrix	Actual / Predicted 0 / 0: 50395 0 / 1: 665 1 / 0: 167 1 / 1: 1859	Actual / Predicted 0 / 0: 21522 0 / 1: 654 1 / 0: 185 1 / 1: 391

ROC Curve



	Train	Test
Area of ROC Curve	0.9933	0.89

Model Interpretation

- The ROC curve area is better for train data set rather than the train dataset. However, there no significant reduction in AUC for test data set compared to train data set. This indicates good accuracy of classification.
- We will try to increase the no. of trees to see if we can obtain any improvement in the model.

Model 3

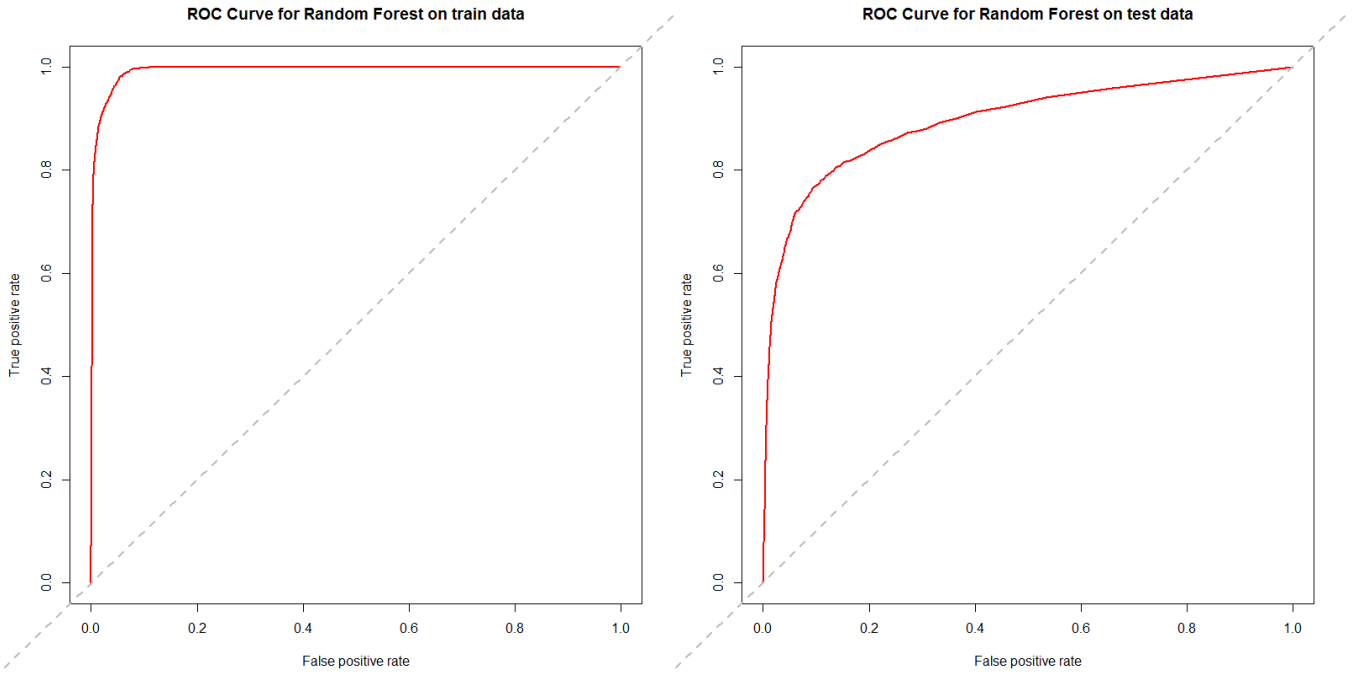
Model Properties

Property	Values
Attributes	Claim$fraud~Boo_hospital_is_networked + Num_Age_of_Insured + Num_Amount_of_Co_Payment_or_Excess_if_applicable + Num_Consultation_Charges + Num_Investigation_Charges + Num_Medicine_Charges + Num_Miscellaneous_Charges + Num_Other_Non_Hospital_Expenses + Num_Post_Hospitalisation_Expenses_included_under_150035 + Num_Pre_Hospitalisation_Expenses_included_under_150035 + Num_Room_Nursing_Charges + Num_Sum_Insured + Num_Surgery_Charges + Num_Total_Claim_Paid + Txt_Gender + Txt_Product_Type + Txt_Type_of_Claim_Payment
No. of variables	17
No. of Records	53086
No. of Trees	750
Mtry	3

Model Results

	Random Forest- Model 3 - Train	Random Forest – Model 3– Test
Specificity	0.7381	0.37
Sensitivity	0.9967	0.9915
Confusion Matrix	Actual / Predicted 0: 0 → 50395, 1 → 661 1: 0 → 167, 1 → 1863	Actual / Predicted 0: 0 → 21522, 1 → 658 1: 0 → 185, 1 → 387

ROC Curve



	Train	Test
Area of ROC Curve	0.9934	0.8954

Model Interpretation

- Random forest model has very good accuracy compared to logistic regression models. Though there is decrease in accuracy on test data set, still the model outperforms logistic regression model.
- The ROC curve area is better for train data set rather than the train dataset. However, there no significant reduction in AUC for test data set compared to train data set. This indicates good accuracy of classification.
- Based on variable importance output and plot, we can conclude that the following variables are key to detection of fraudulent claims. Out of these variables, having sum insured and claim paid amount being important variables is not surprising. However, inspection of other variables reveals that the following factors would drive fraudulent claims or help the claims processing team to suspect a potential fraud.
    a) Whether hospital is network hospital or not
    b) Consultation charges amount
    c) Post hospitalization expenses
    d) Non-Hospital expenses
    e) Nursing charges
    f) Surgery charges
    g) Total claim paid

h) Sum insured

Model Comparison

The best performing variant of each type of model is picked and compared against each other in the below table.

Model Name	Logistic		Neural Net		Random Forest	
	Train	Test	Train	Test	Train	Test
ROC Area	0.85	0.86	0.88	0.89	0.99	0.90
Specificity	90%	90%	38%	40%	74%	37%
Sensitivity	49%	50%	98%	98%	100%	99%
Accuracy	51%	52%	95%	96%	98%	96%
Misclassification rate	49%	49%	5%	4%	2%	4%

Individual model performance of supervised learning methods is often assessed using a confusion matrix. The objective, typically, is to increase the number of correct predictions (sensitivity) while maintaining incorrect predictions or the false alarm rate (specificity) at an acceptable level. The two goals, getting as much of the target field correctly predicted versus keeping the false alarm rate low, tend to be inversely proportional. A simple example can illustrate this point: to catch all the fraud in a data set, one need only call health care claims fraudulent, while to avoid any false alarms one need only call all claims non-fraudulent. Reality resides between these two extremes. The business question typically defines what false alarm rate is tolerable versus what amount of fraud (or other target) needs to be caught.

As ROC curve represents relationship between True positive rate and false positive rate and the area under ROC curve represents the tradeoff between these two measures. Hence, we have chosen area under ROC as the criteria for selecting a model from multiple models.

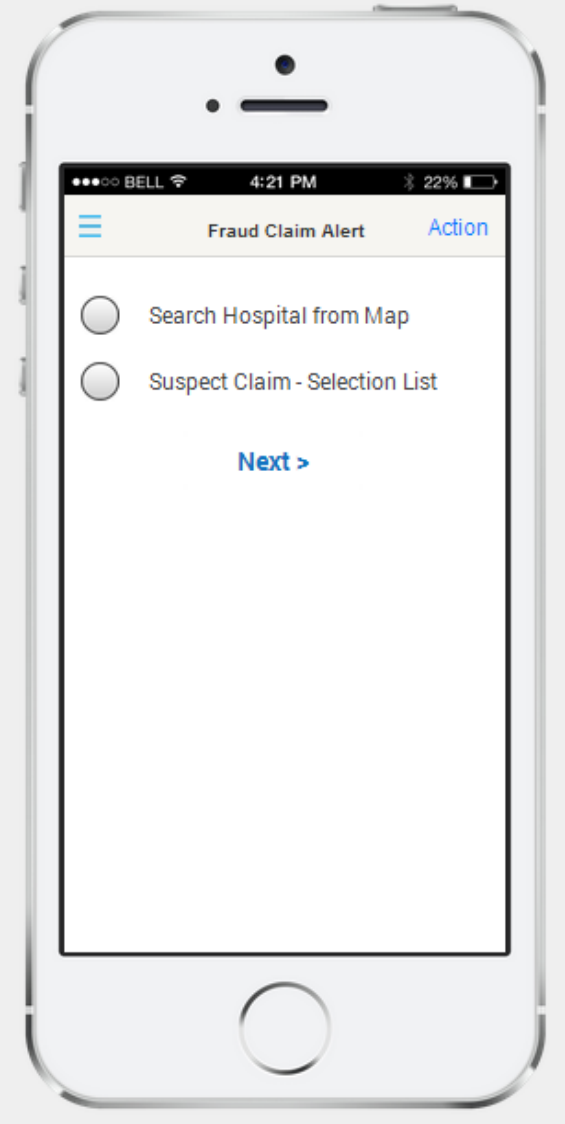
Random forest is slightly better than neural network with an area > 0.9 and hence either of them can be used in scoring any new claim data.

Mobile App Wireframe - Fraudulent Claim Alert

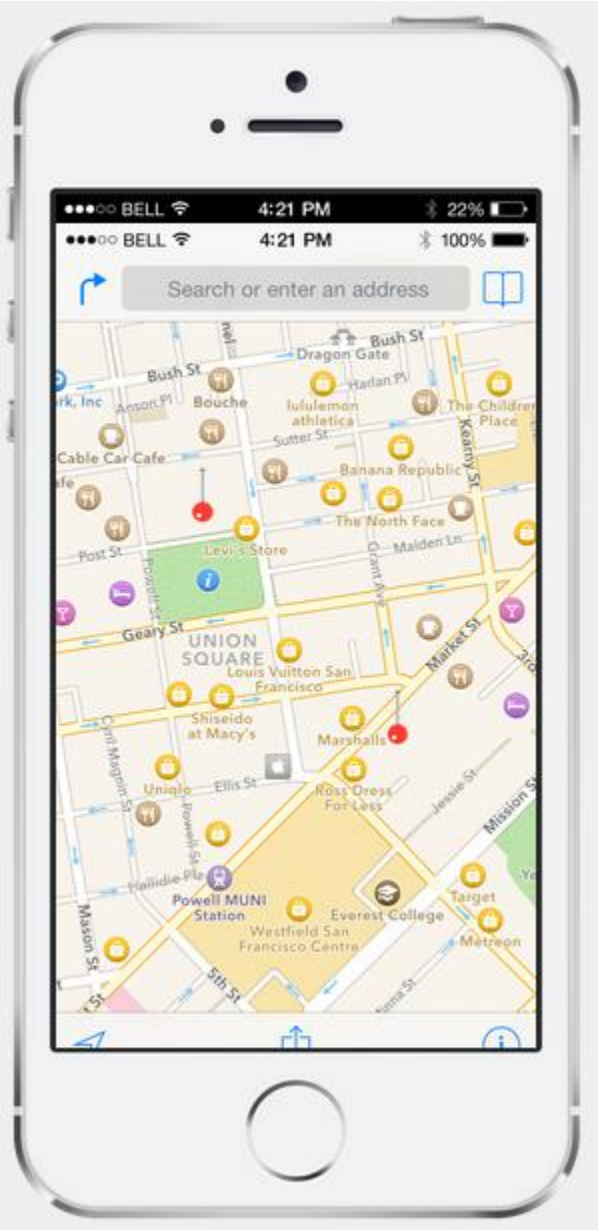
We have formulated an idea of a mobile application which can make use of the fraud detection models built. The app will serve any insurance officer in viewing the details of those claims which are identified as fraudulent by the models. Whenever a claim is submitted to the insurance company the data is fed to the models and the claims which are possibly fraudulent are sent to the app using a “push” mechanism. The insurance officer can make a decision on the necessary action. The wireframe of the “Fraudulent Claim Alert” app is shown below.

1. Insurance Claim Approver will use the Fraud claim alert for evaluating/approving the insurance claims.

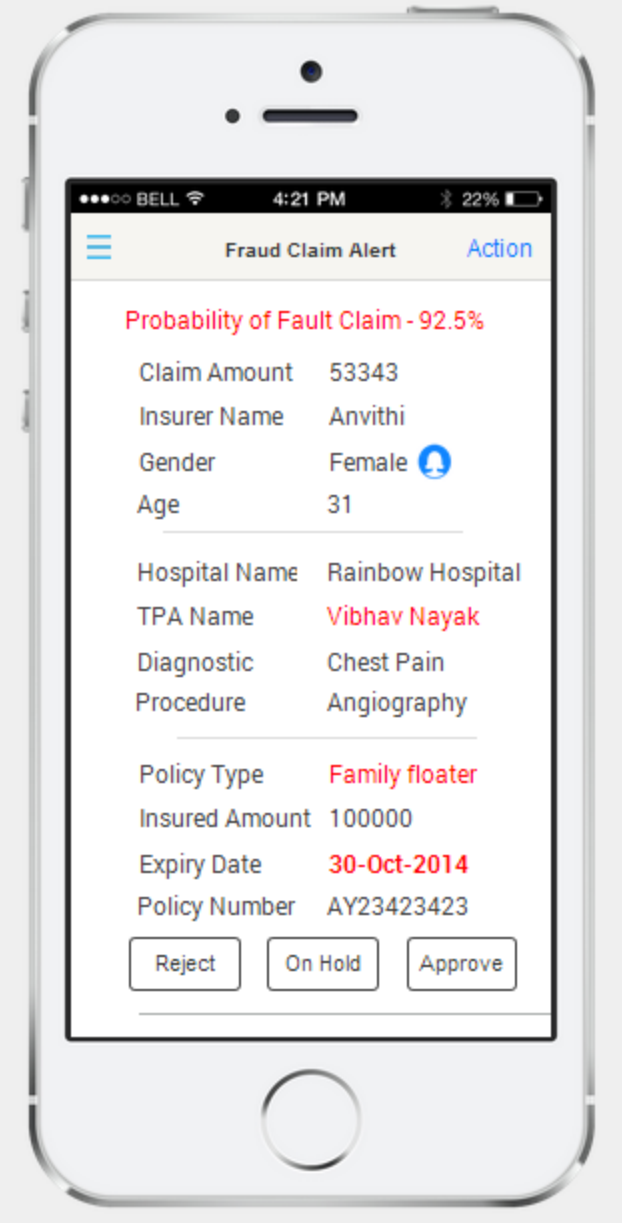
2. There are 2 options which the approver can choose from
    a. Choosing to approve the claim based on hospital search from Map
    b. Choosing the claim request directly



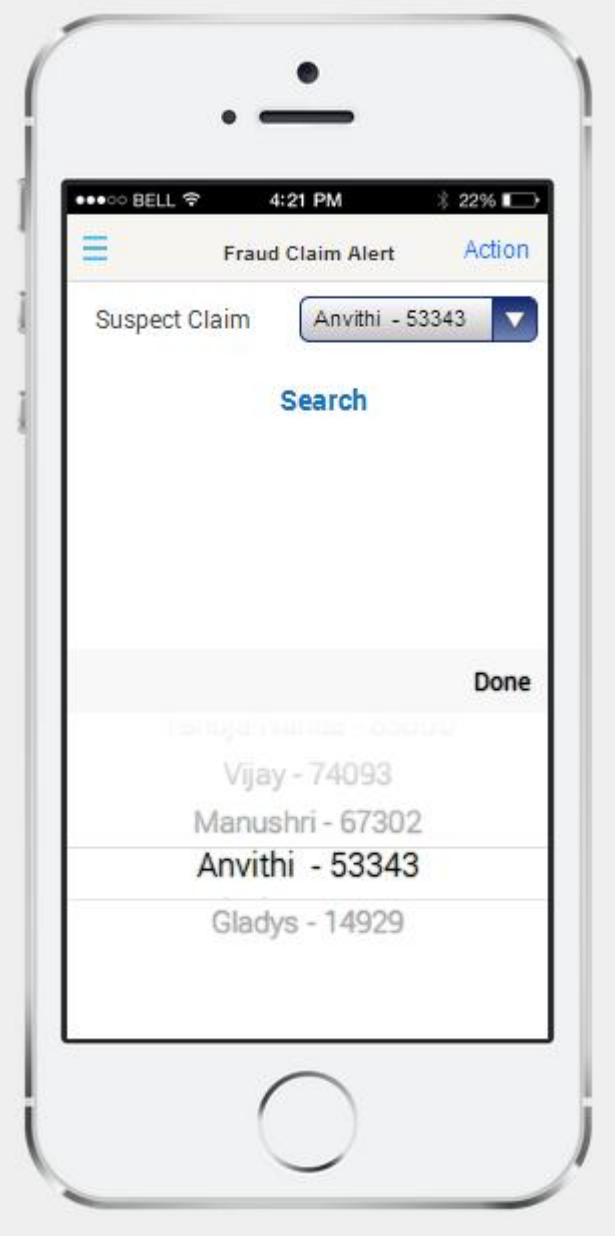
3. Let's say the user has chosen map to search the hospital, the below map will be shown to choose the exact hospital



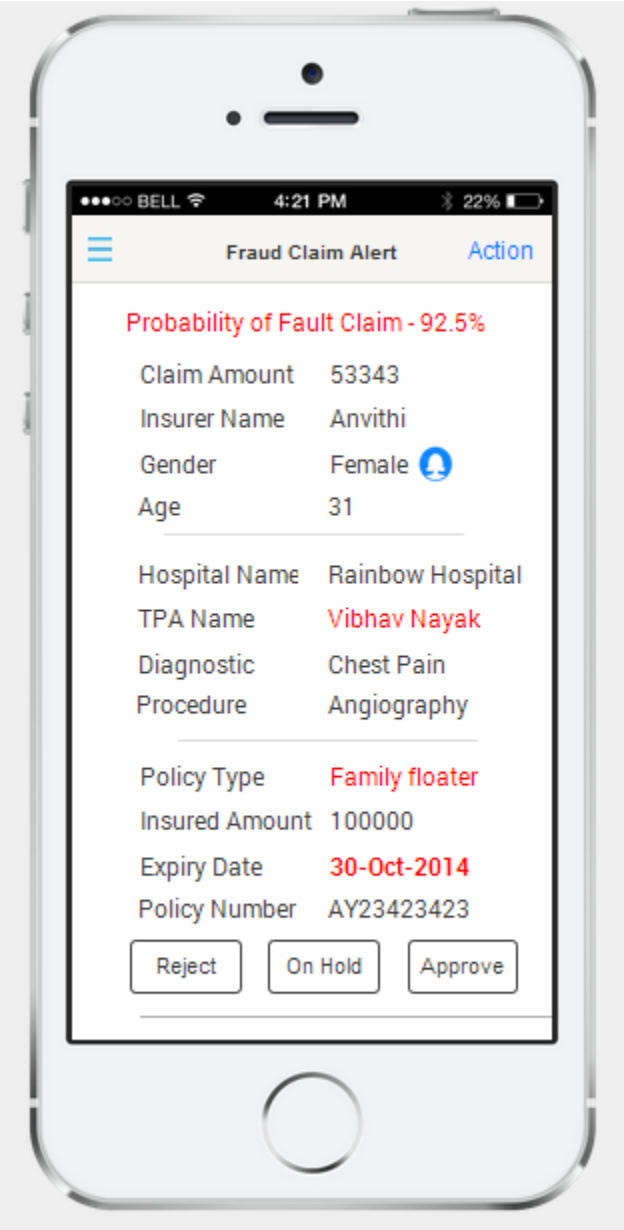
4. On clicking the pinned hospital location, the claims raised from hospital will be shown as below. Approver will get the below information
    a. Probability of the fault claim ranging 0 – 100%
    b. Attributes which are contributing to identify the claim as fraudulent in red color
    c. Approver can either approve, reject or put on hold of the claim approval request



5. In case if the approver wants to go thru' the suspected claims based on the order by claim amount, the below snapshot will be chosen. In this screen, patient name along with the claim amount will be shown to the user to choose.

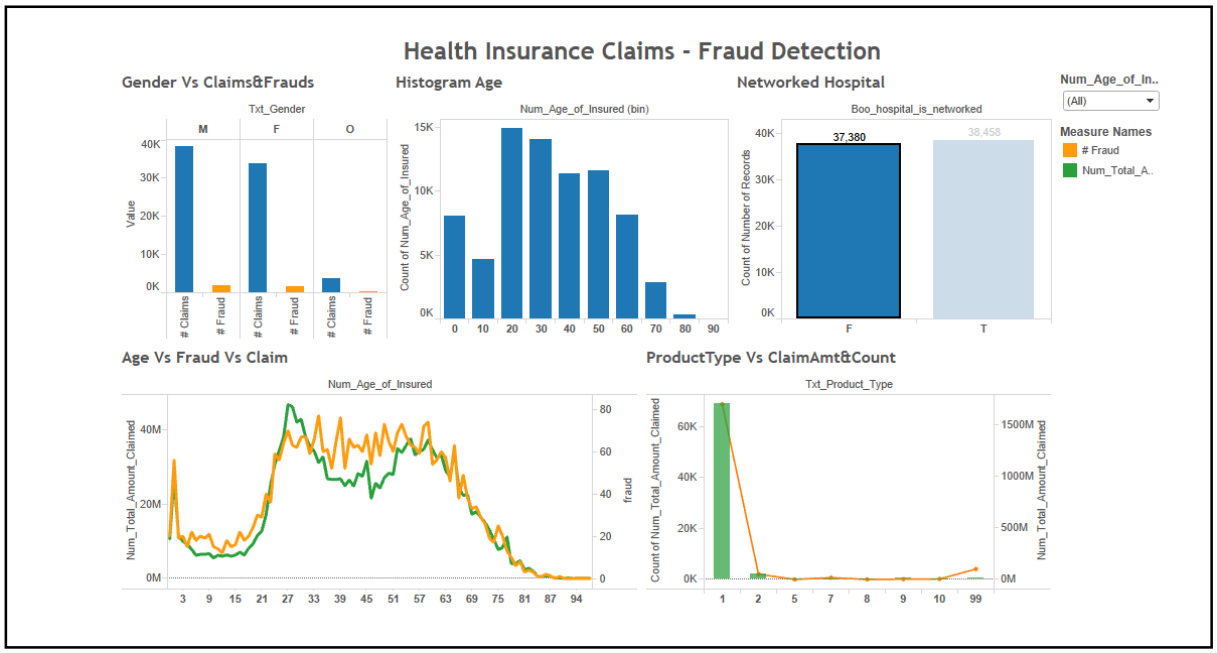


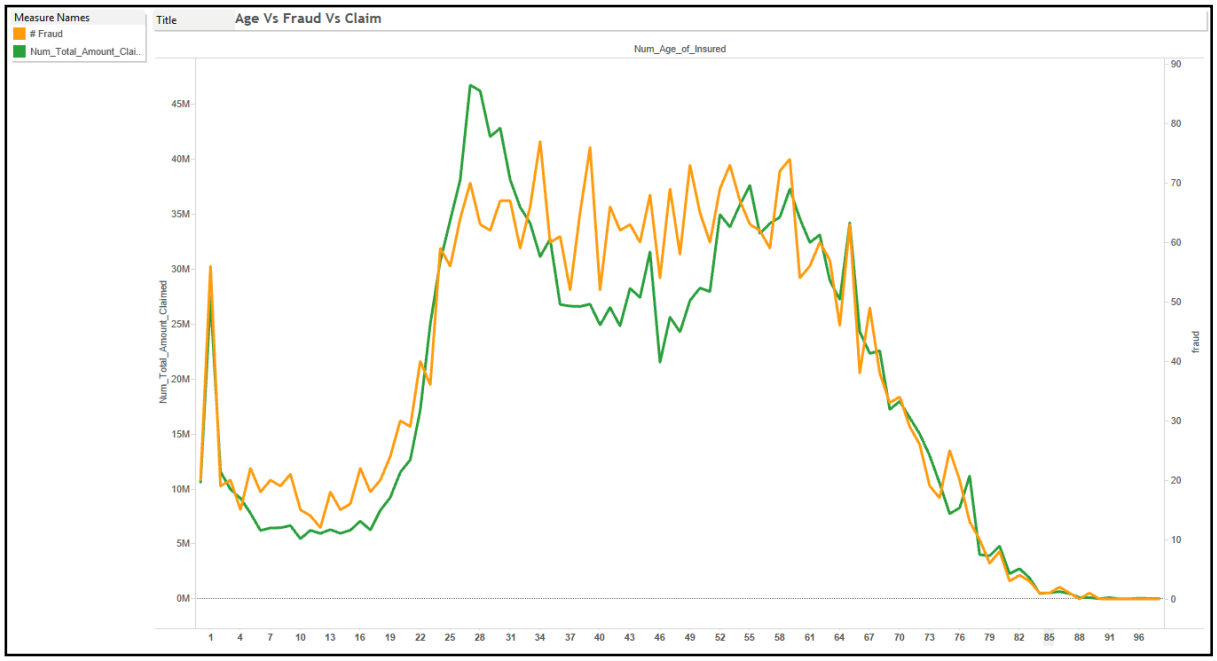
6. On clicking a particular patient name, the claims raised from the patient will be shown as below. Approver will get the below information
    a. Probability of the fault claim ranging 0 – 100%
    b. Attributes which are contributing to identify the claim as fraudulent in red color
    c. Approver can either approve, reject or put on hold of the claim approval request

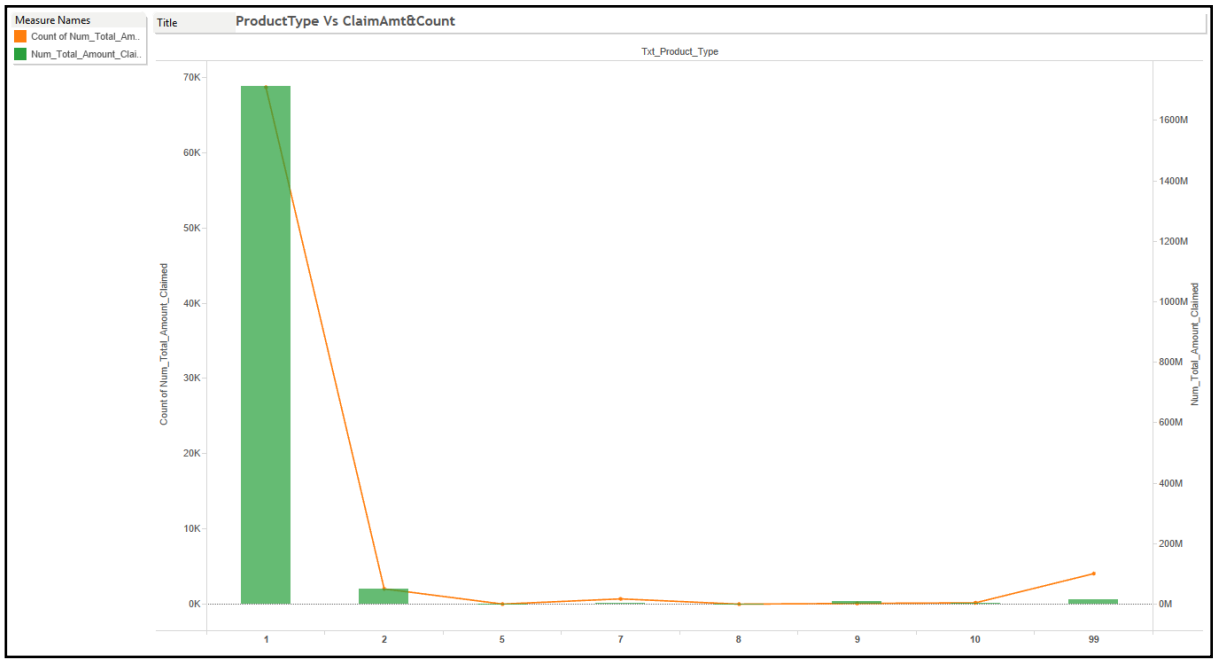


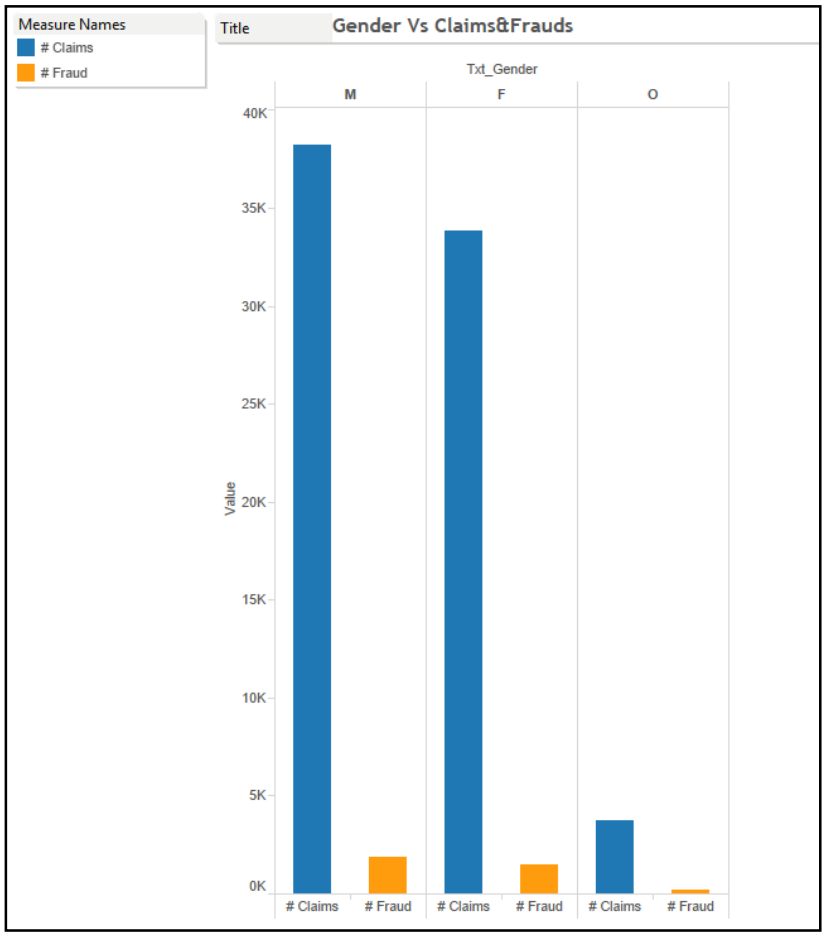
Dashboarding

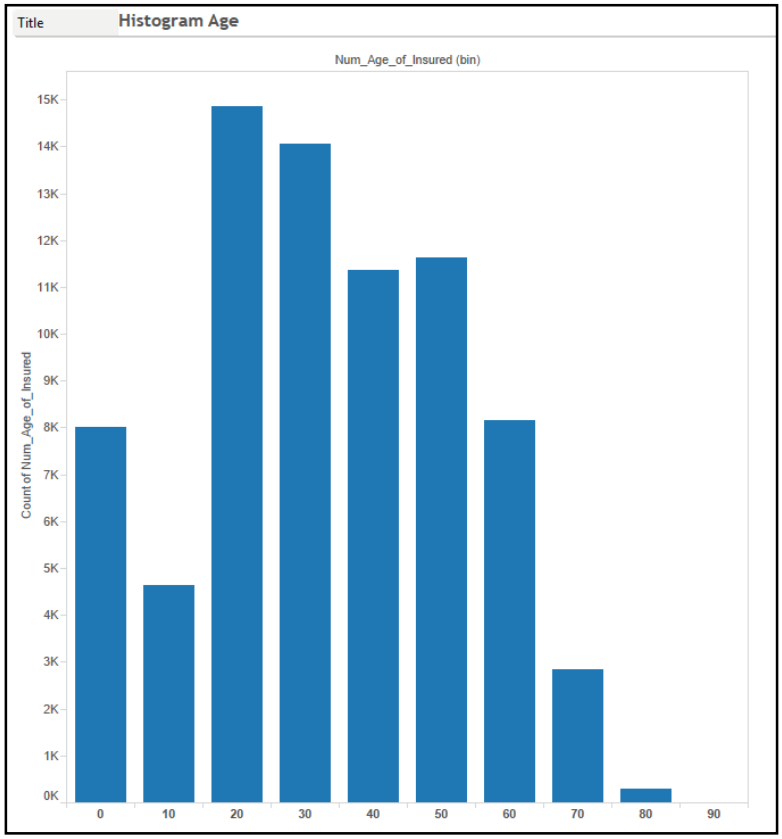
We have created a dashboard showcasing different visualizations, based on - age of the insured, sum insured, count of frauds, product types, gender, hospitals etc. The snapshots of the tableau dashboard and individual visualizations are furnished below.

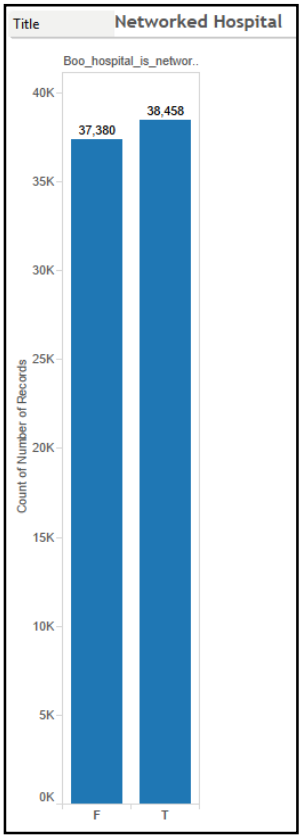





Measure Names
Count of Num_Total_Am..
Num_Total_Amount_Clai..
Title
ProductType Vs ClaimAmt&Count
Txt_Product_Type
Count of Num_Total_Amount_Claimed
Num_Total_Amount_Claimed
70K
60K
50K
40K
30K
20K
10K
0K
1600M
1400M
1200M
1000M
800M
600M
400M
200M
0M
1
2
5
7
8
9
10
99


Measure Names
Claims
Fraud
Title
Gender Vs Claims&Frauds
Txt_Gender
M
F
O
Value
40K
35K
30K
25K
20K
15K
10K
5K
0K
Claims
Fraud
Claims
Fraud
Claims
Fraud


Title
Histogram Age
Num_Age_of_Insured (bin)
Count of Num_Age_of_Insured
15K
14K
13K
12K
11K
10K
9K
8K
7K
6K
5K
4K
3K
2K
1K
0K
0
10
20
30
40
50
60
70
80
90


Title
Networked Hospital
Boo_hospital_is_networ..
Count of Number of Records
40K
35K
30K
25K
20K
15K
10K
5K
0K
37,380
38,458
F
T

Conclusion

We started off with a major challenge of developing a framework for detection of fraudulent health insurance claims without having fraud indicator in the data.

We have leveraged business rule based scoring and advanced analytical methods like logistic regression, neural networks and random forest to come up with a fraud detection framework. By adopting this framework, insurance companies will be able to significantly reduce the no. of suspected claims to be investigated by them. It will also help them in identifying the key drivers of fraudulent behavior which can help them in fine tuning their products and sales practices.

The major strength of the framework is that it is completely flexible and can be adopted to each insurer's business rules and practices. This can be achieved through addition or deletion of triggers, modification of weightage to the trigger, scoring range for each trigger etc.

Recommendations

We would also like to share our thoughts on this issue which will help the insurers in combating the ever increasing and innovating fraud in health care sector.

Data Quality Assurance

- We observed that even though standard templates are available to submit claims data to regulator, there were many data quality related issues E.g. Diagnosis details are not in standardized format making it difficult to use this information in either business rules or modeling. We strongly recommend that insurance company should review their data collection policies, establish data quality checks in place and review and enhance the data collection process on a periodic basis.

Social Media

- Apart from traditional data collected by insurance companies, they should leverage huge amount of unstructured data available from public and internal sources.

Policy member's social profile (Facebook, LinkedIn, Twitter etc.) would be captured along with other details during enrollment.

During claim processing, after running the scoring model, if a specific claim's probability of fraudulent claim is high /greater than a specified threshold, claim processor can retrieve social footprint of the policy member and use it as suggested below.

a) Using location intelligence (subject to local regulations) provided by social networks, check if the location of the member during hospitalization is different from hospital's location

b) Monitoring if policy member is active on social media during hospitalization period for specific disease types. Also check the device type on which the member is active (e.g. active on devices on other than mobile/tablet during hospitalization is suspicious)

c) Using text mining to analyze the posts / tweets on social media during hospitalization period and post hospitalization period to suspect fraudulent behavior

d) Using Social network analysis to analyze if the policy member is connected to suspicious persons/firms.

Reference

1. Trigger based scoring System, Dr Ashish Dogra, www.insuranceinstituteofindia.com/downloads/Forms/III/Important%20Notice/Fraud%20Control%20Workshop/Trigger%20based%20scoring%20System%20-%20Dr%20Ashish%20Dogra.pdf
2. http://en.wikipedia.org/wiki/Polaris_Financial_Technology_Limited
3. http://www.polarisft.com/about-us/about-us.asp
4. http://rocr.bioinf.mpi-sb.mpg.de/
5. scg.sdsu.edu/rf_r/