Fraud Detection Medicare Claims Dataset

HOANG NHA NGUYEN

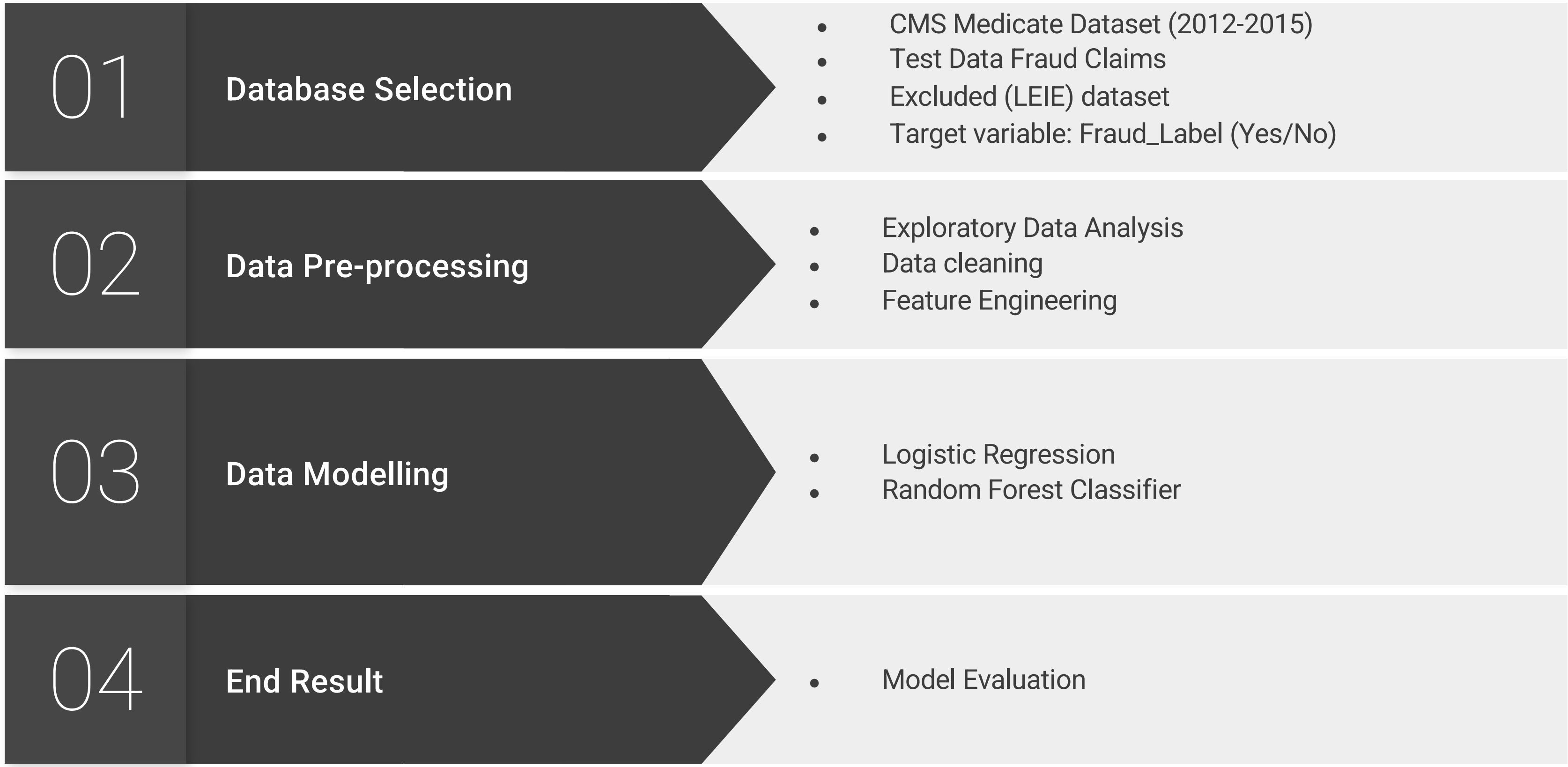
WHY ?

US Healthcare spending has increased.

High-cost health related services leave patients with limited medical care .US has established and funded programs such as Medicare

Issues facing healthcare such as fraudulent or potentially fraudulent physicians or providers.

Workflow



01 **Database Selection**

- CMS Medicate Dataset (2012-2015)
- Test Data Fraud Claims
- Excluded (LEIE) dataset
- Target variable: Fraud_Label (Yes/No)

02 **Data Pre-processing**

- Exploratory Data Analysis
- Data cleaning
- Feature Engineering

03 **Data Modelling**

- Logistic Regression
- Random Forest Classifier

04 **End Result**

- Model Evaluation

Goal

Develop a classifier model with certain features that can help detect fraud.

Target variable:
Fraud_Label (Yes/No)

Dataset Selection

01 List of Excluded Individuals and Entities (LEIE) database 2017

- List of individuals and entities excluded from Medicare due to previous healthcare fraud.
- Filter unique NPI (remove all null values)

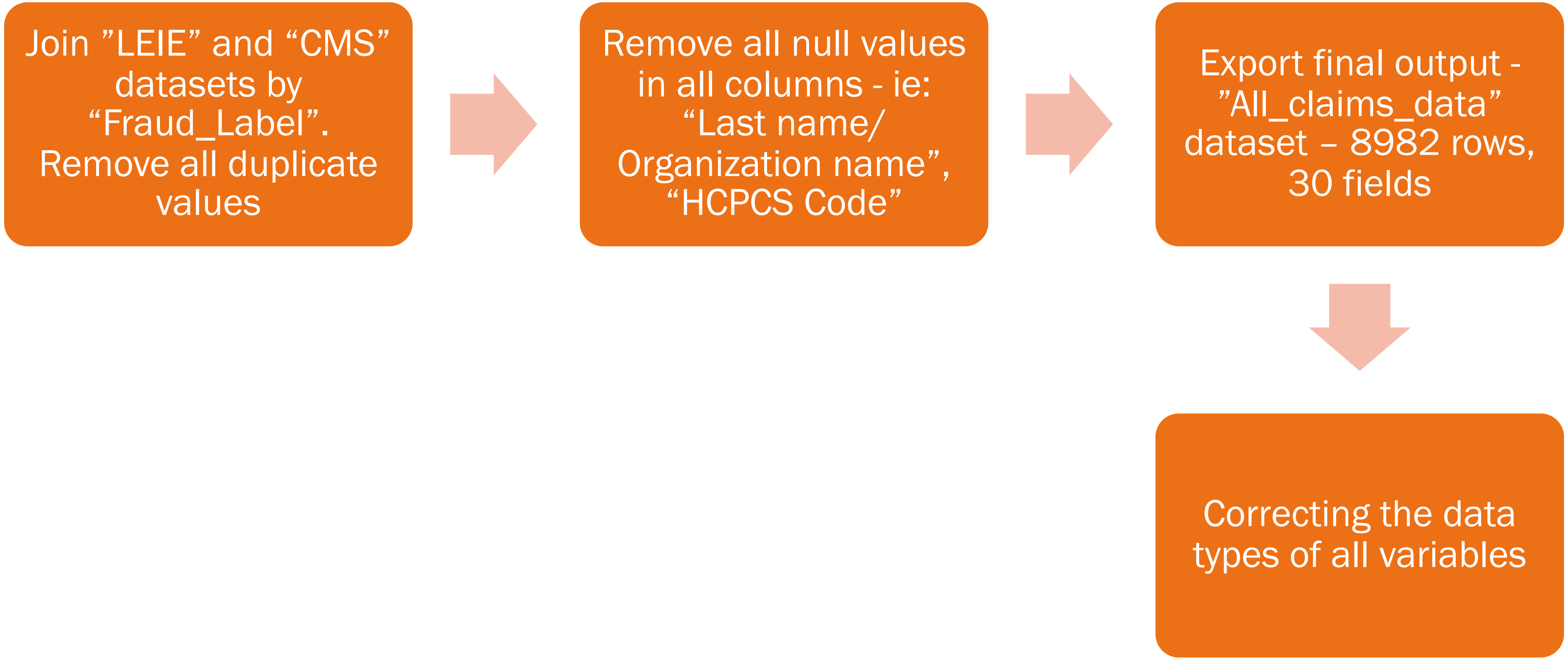
02 **CMS Dataset (2012-2015)**

- All information related to physicians and their payments, charges by NPI.
- Join these input data (2012-2015)

03 **Test data fraud claims**

- A test dataset with a smaller number of providers

Data Pre-processing



Join ”LEIE” and “CMS” datasets by “Fraud_Label”. Remove all duplicate values

Remove all null values in all columns - ie: “Last name/ Organization name”, “HCPCS Code”

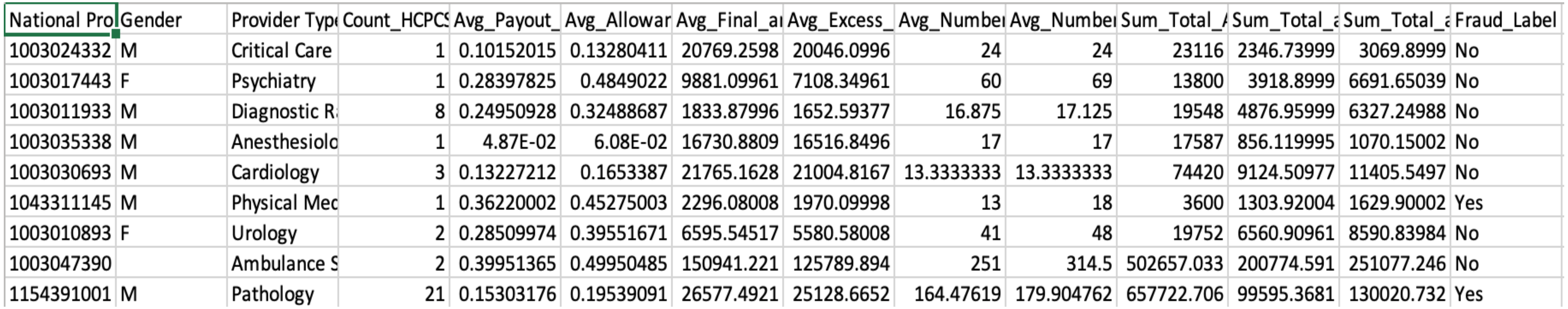
Export final output - ”All_claims_data” dataset – 8982 rows, 30 fields

Correcting the data types of all variables

Feature Engineering

Perform necessary transformations on the custom formulas/variables

- Total_amount_claimed -> Sum
- Total_amount_paid -> Sum
- Total_amount_allowed ->Sum
- Payout-ratio -> Average
- Allowance ratio -> Average
- Final_amount_received -> Average
- Excess_amount_claimed -> Average
- Number of medicare -> Average
- HCPCS -> Count
- Group by the following columns – “NPI”, “Gender”, “Provider Type”



National Pro	Gender	Provider Typ	Count_HCPC	Avg_Payout_	Avg_Allowar	Avg_Final_a	Avg_Excess_	Avg_Numbe	Avg_Numbe	Sum_Total_	Sum_Total_a	Sum_Total_a	Fraud_Label
1003024332	M	Critical Care	1	0.10152015	0.13280411	20769.2598	20046.0996	24	24	23116	2346.73999	3069.8999	No
1003017443	F	Psychiatry	1	0.28397825	0.4849022	9881.09961	7108.34961	60	69	13800	3918.8999	6691.65039	No
1003011933	M	Diagnostic R	8	0.24950928	0.32488687	1833.87996	1652.59377	16.875	17.125	19548	4876.95999	6327.24988	No
1003035338	M	Anesthesiolo	1	4.87E-02	6.08E-02	16730.8809	16516.8496	17	17	17587	856.119995	1070.15002	No
1003030693	M	Cardiology	3	0.13227212	0.1653387	21765.1628	21004.8167	13.3333333	13.3333333	74420	9124.50977	11405.5497	No
1043311145	M	Physical Mec	1	0.36220002	0.45275003	2296.08008	1970.09998	13	18	3600	1303.92004	1629.90002	Yes
1003010893	F	Urology	2	0.28509974	0.39551671	6595.54517	5580.58008	41	48	19752	6560.90961	8590.83984	No
1003047390		Ambulance S	2	0.39951365	0.49950485	150941.221	125789.894	251	314.5	502657.033	200774.591	251077.246	No
1154391001	M	Pathology	21	0.15303176	0.19539091	26577.4921	25128.6652	164.47619	179.904762	657722.706	99595.3681	130020.732	Yes

Final Dataset: “Physician_level_aggregate”

292 ROWS, 14 FIELDS -> “FRAUD_LABEL” Y/N

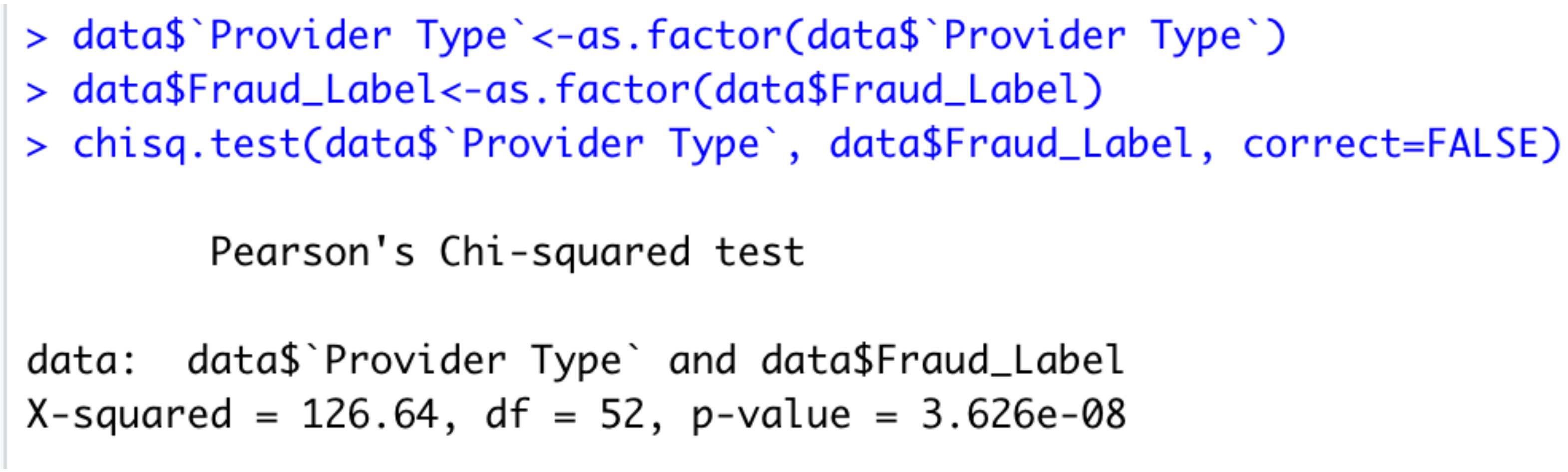
T-Test (Statistical Significance Test 5%)

Variables	P-Values
Avg_Payout_ratio	7.7717e-06
Avg_Allowance_ratio	6.8443e-07
Avg_Final_amount	0.9049
Avg_Excess_Amount	0.76564
Sum_Total_Amount_Claimed	0.046566
Sum_Total_Amount_Paid	0.0023565
Sum_Total Amount Allowed	0.0018021
Avg_Number_Medicare.Beneficiaries	0.028824
Avg_Number_Medicare.Beneficiaries.Day.Services	0.069605
Count_HCPCS	0.00063374

- Statistically significant variables:

“Avg_Payout_Ratio”
“Avg_Allowance_ratio”
“Sum_Total_Amount_Claimed”
“Sum_Total_Amount_Paid”
“Sum_Total_Amount Allowed”
“Avg_Number_Medicare.Beneficiaries”
“Count_HCPCS”

Chi-squared Test



```
> data$`Provider Type`<-as.factor(data$`Provider Type`)
> data$Fraud_Label<-as.factor(data$Fraud_Label)
> chisq.test(data$`Provider Type`, data$Fraud_Label, correct=FALSE)

	Pearson's Chi-squared test

data:  data$`Provider Type` and data$Fraud_Label
X-squared = 126.64, df = 52, p-value = 3.626e-08
```

- “Provider Type” is statistically significant

Features Selection

- Significant variables:

“Avg_Payout_Ratio”

“Avg_Allowance_ratio”

“Sum_Total_Amount_Claimed”

“Sum_Total_Amount_Paid”

“Sum_Total_Amount Allowed”

“Avg_Number_Medicare.Beneficiaries”

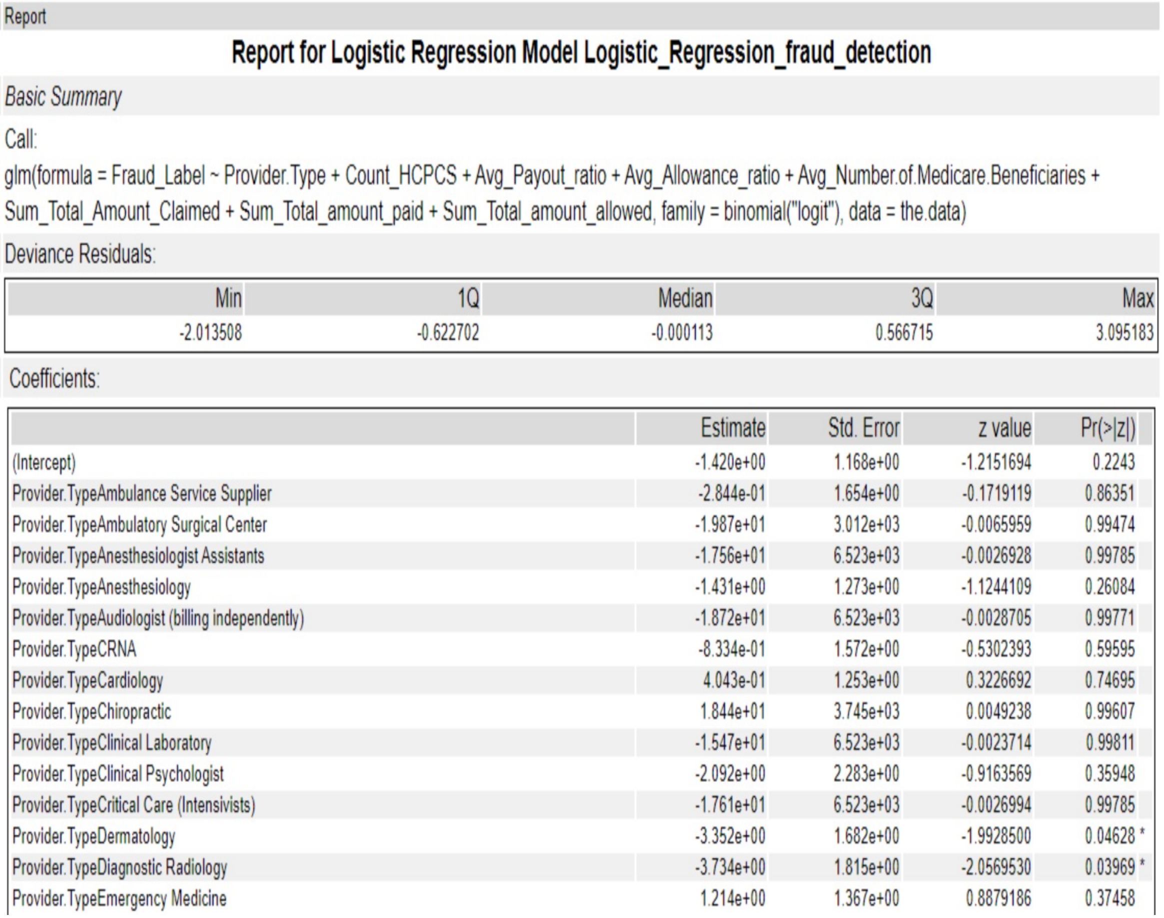
“Count_HCPCS”

“Provider Type”

- Also, there is no need to consider the following variables for the model:

“NPI” and “Gender”

Model 1
Logistic Regression



Report

Report for Logistic Regression Model Logistic_Regression_fraud_detection

Basic Summary

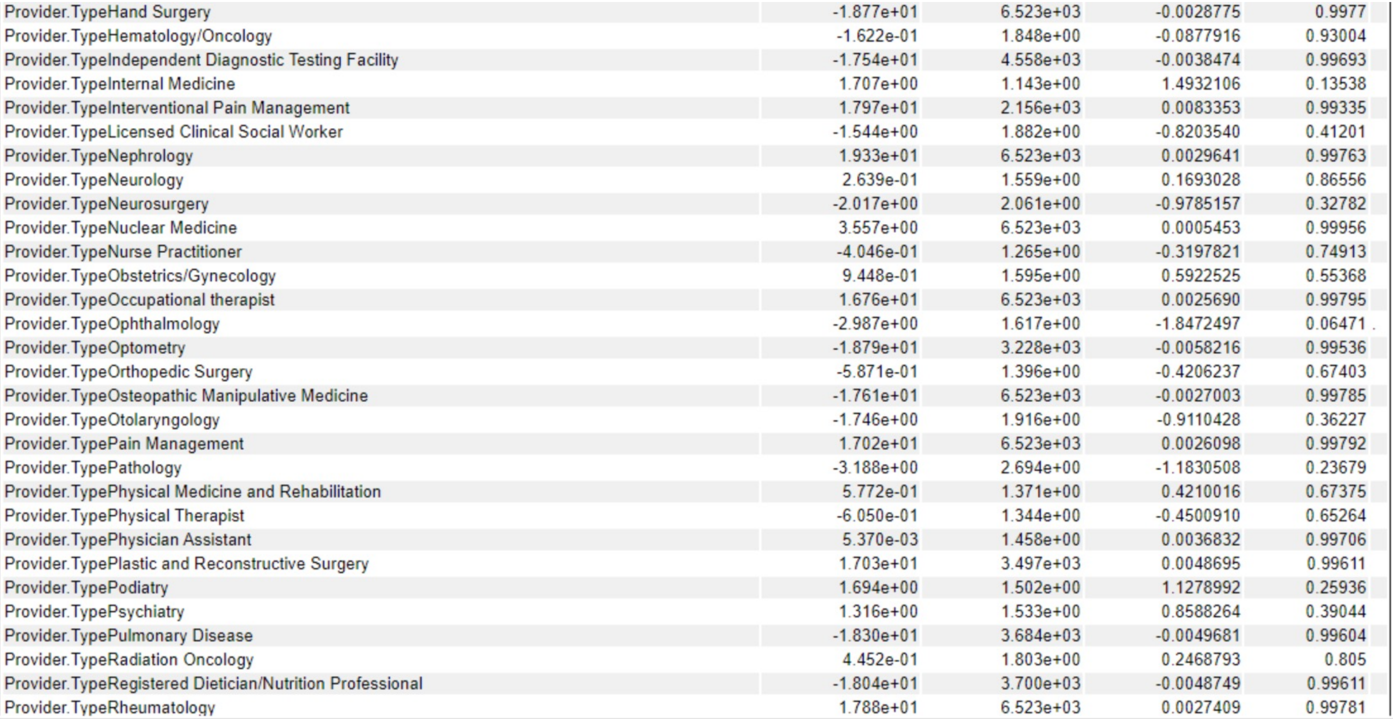
Call:

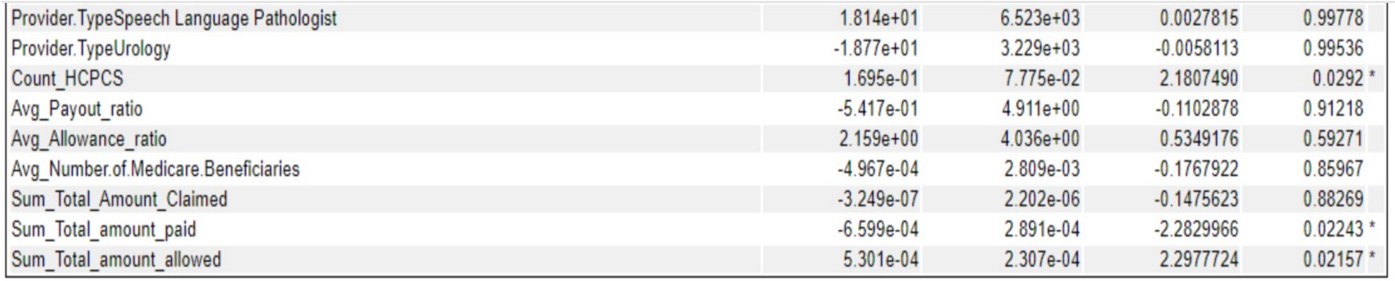
glm(formula = Fraud_Label ~ Provider.Type + Count_HCPCS + Avg_Payout_ratio + Avg_Allowance_ratio + Avg_Number.of.Medicare.Beneficiaries + Sum_Total_Amount_Claimed + Sum_Total_amount_paid + Sum_Total_amount_allowed, family = binomial("logit"), data = the.data)

Deviance Residuals:

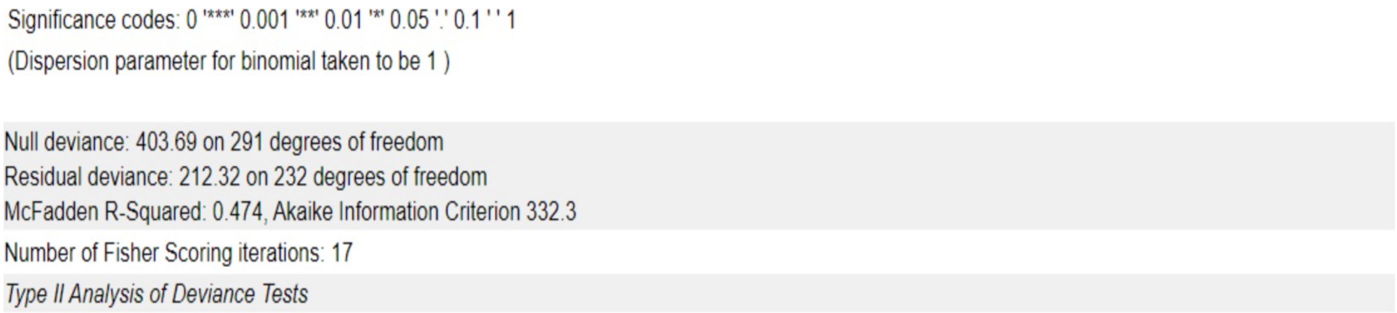
Min	1Q	Median	3Q	Max
-2.013508	-0.622702	-0.000113	0.566715	3.095183

Coefficients:





	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.420e+00	1.168e+00	-1.2151694	0.2243
Provider.TypeAmbulance Service Supplier	-2.844e-01	1.654e+00	-0.1719119	0.86351
Provider.TypeAmbulatory Surgical Center	-1.987e+01	3.012e+03	-0.0065959	0.99474
Provider.TypeAnesthesiologist Assistants	-1.756e+01	6.523e+03	-0.0026928	0.99785
Provider.TypeAnesthesiology	-1.431e+00	1.273e+00	-1.1244109	0.26084
Provider.TypeAudiologist (billing independently)	-1.872e+01	6.523e+03	-0.0028705	0.99771
Provider.TypeCRNA	-8.334e-01	1.572e+00	-0.5302393	0.59595
Provider.TypeCardiology	4.043e-01	1.253e+00	0.3226692	0.74695
Provider.TypeChiropractic	1.844e+01	3.745e+03	0.0049238	0.99607
Provider.TypeClinical Laboratory	-1.547e+01	6.523e+03	-0.0023714	0.99811
Provider.TypeClinical Psychologist	-2.092e+00	2.283e+00	-0.9163569	0.35948
Provider.TypeCritical Care (Intensivists)	-1.761e+01	6.523e+03	-0.0026994	0.99785
Provider.TypeDermatology	-3.352e+00	1.682e+00	-1.9928500	0.04628 *
Provider.TypeDiagnostic Radiology	-3.734e+00	1.815e+00	-2.0569530	0.03969 *
Provider.TypeEmergency Medicine	1.214e+00	1.367e+00	0.8879186	0.37458
Provider.TypeHand Surgery	-1.877e+01	6.523e+03	-0.0028775	0.9977
Provider.TypeHematology/Oncology	-1.622e-01	1.848e+00	-0.0877916	0.93004
Provider.TypeIndependent Diagnostic Testing Facility	-1.754e+01	4.558e+03	-0.0038474	0.99693
Provider.TypeInternal Medicine	1.707e+00	1.143e+00	1.4932106	0.13538
Provider.TypeInterventional Pain Management	1.797e+01	2.156e+03	0.0083353	0.99335
Provider.TypeLicensed Clinical Social Worker	-1.544e+00	1.882e+00	-0.8203540	0.41201
Provider.TypeNephrology	1.933e+01	6.523e+03	0.0029641	0.99763
Provider.TypeNeurology	2.639e-01	1.559e+00	0.1693028	0.86556
Provider.TypeNeurosurgery	-2.017e+00	2.061e+00	-0.9785157	0.32782
Provider.TypeNuclear Medicine	3.557e+00	6.523e+03	0.0005453	0.99956
Provider.TypeNurse Practitioner	-4.046e-01	1.265e+00	-0.3197821	0.74913
Provider.TypeObstetrics/Gynecology	9.448e-01	1.595e+00	0.5922525	0.55368
Provider.TypeOccupational therapist	1.676e+01	6.523e+03	0.0025690	0.99795
Provider.TypeOphthalmology	-2.987e+00	1.617e+00	-1.8472497	0.06471 .
Provider.TypeOptometry	-1.879e+01	3.228e+03	-0.0058216	0.99536
Provider.TypeOrthopedic Surgery	-5.871e-01	1.396e+00	-0.4206237	0.67403
Provider.TypeOsteopathic Manipulative Medicine	-1.761e+01	6.523e+03	-0.0027003	0.99785
Provider.TypeOtolaryngology	-1.746e+00	1.916e+00	-0.9110428	0.36227
Provider.TypePain Management	1.702e+01	6.523e+03	0.0026098	0.99792
Provider.TypePathology	-3.188e+00	2.694e+00	-1.1830508	0.23679
Provider.TypePhysical Medicine and Rehabilitation	5.772e-01	1.371e+00	0.4210016	0.67375
Provider.TypePhysical Therapist	-6.050e-01	1.344e+00	-0.4500910	0.65264
Provider.TypePhysician Assistant	5.370e-03	1.458e+00	0.0036832	0.99706
Provider.TypePlastic and Reconstructive Surgery	1.703e+01	3.497e+03	0.0048695	0.99611
Provider.TypePodiatry	1.694e+00	1.502e+00	1.1278992	0.25936
Provider.TypePsychiatry	1.316e+00	1.533e+00	0.8588264	0.39044
Provider.TypePulmonary Disease	-1.830e+01	3.684e+03	-0.0049681	0.99604
Provider.TypeRadiation Oncology	4.452e-01	1.803e+00	0.2468793	0.805
Provider.TypeRegistered Dietician/Nutrition Professional	-1.804e+01	3.700e+03	-0.0048749	0.99611
Provider.TypeRheumatology	1.788e+01	6.523e+03	0.0027409	0.99781
Provider.TypeSpeech Language Pathologist	1.814e+01	6.523e+03	0.0027815	0.99778
Provider.TypeUrology	-1.877e+01	3.229e+03	-0.0058113	0.99536
Count_HCPCS	1.695e-01	7.775e-02	2.1807490	0.0292 *
Avg_Payout_ratio	-5.417e-01	4.911e+00	-0.1102878	0.91218
Avg_Allowance_ratio	2.159e+00	4.036e+00	0.5349176	0.59271
Avg_Number.of.Medicare.Beneficiaries	-4.967e-04	2.809e-03	-0.1767922	0.85967
Sum_Total_Amount_Claimed	-3.249e-07	2.202e-06	-0.1475623	0.88269
Sum_Total_amount_paid	-6.599e-04	2.891e-04	-2.2829966	0.02243 *
Sum_Total_amount_allowed	5.301e-04	2.307e-04	2.2977724	0.02157 *



Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 403.69 on 291 degrees of freedom

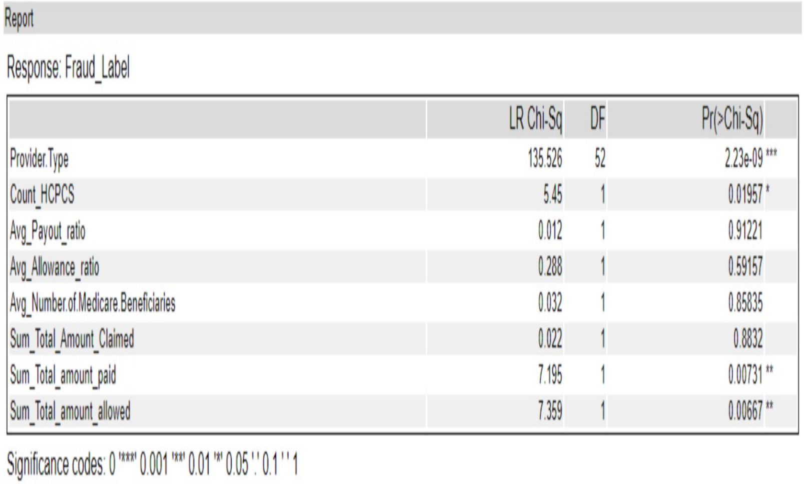
Residual deviance: 212.32 on 232 degrees of freedom

McFadden R-Squared: 0.474, Akaike Information Criterion 332.3

Number of Fisher Scoring iterations: 17

Type II Analysis of Deviance Tests

Model Evaluation



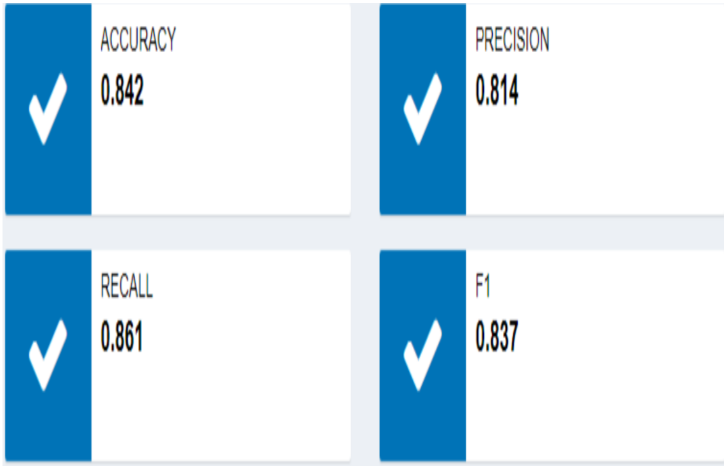
Report

Response: Fraud_Label

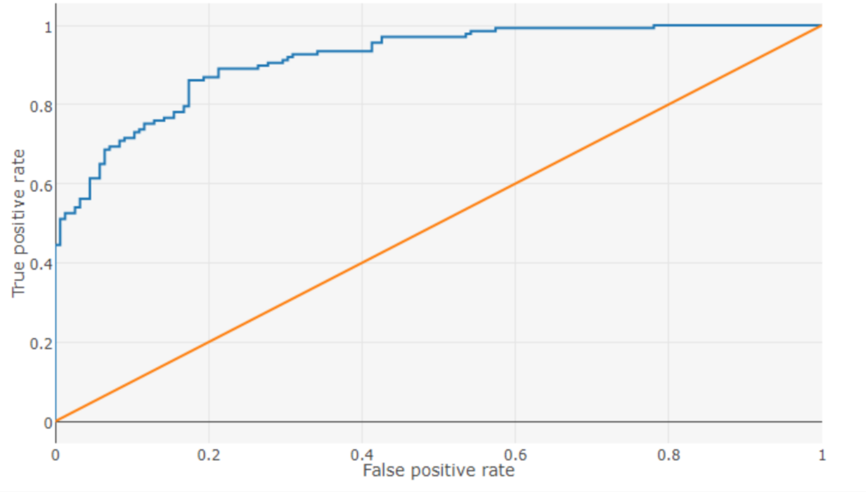
	LR Chi-Sq	DF	Pr(>Chi-Sq)
Provider.Type	135.526	52	2.23e-09 ***
Count_HCPCS	5.45	1	0.01957 *
Avg_Payout_ratio	0.012	1	0.91221
Avg_Allowance_ratio	0.288	1	0.59157
Avg_Number.of.Medicare.Beneficiaries	0.032	1	0.85835
Sum_Total_Amount_Claimed	0.022	1	0.8832
Sum_Total_amount_paid	7.195	1	0.00731 **
Sum_Total_amount_allowed	7.359	1	0.00667 **

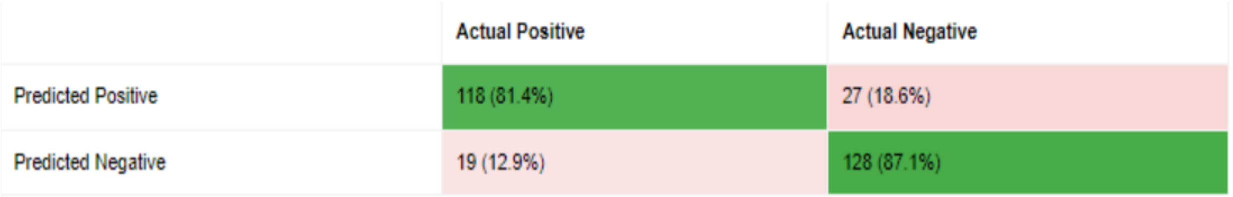
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Logistic Regression



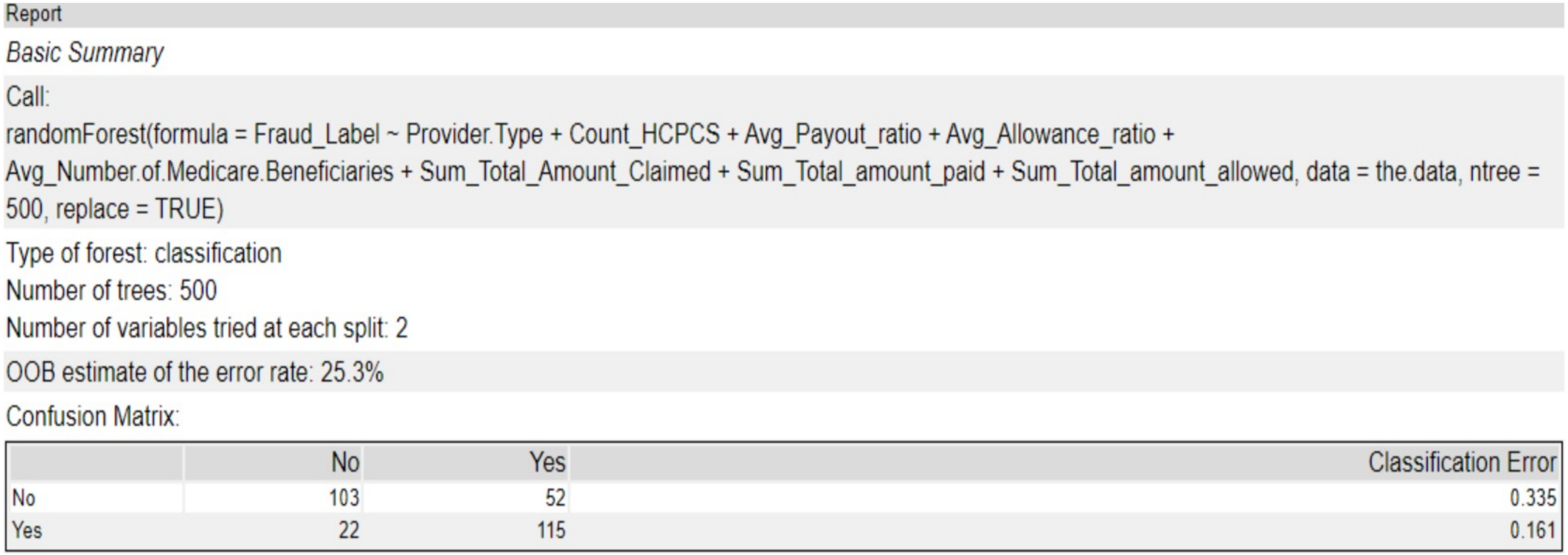
✓ ACCURACY 0.842	✓ PRECISION 0.814
✓ RECALL 0.861	✓ F1 0.837





	Actual Positive	Actual Negative
Predicted Positive	118 (81.4%)	27 (18.6%)
Predicted Negative	19 (12.9%)	128 (87.1%)

Model 2
Random Forest



Report

Basic Summary

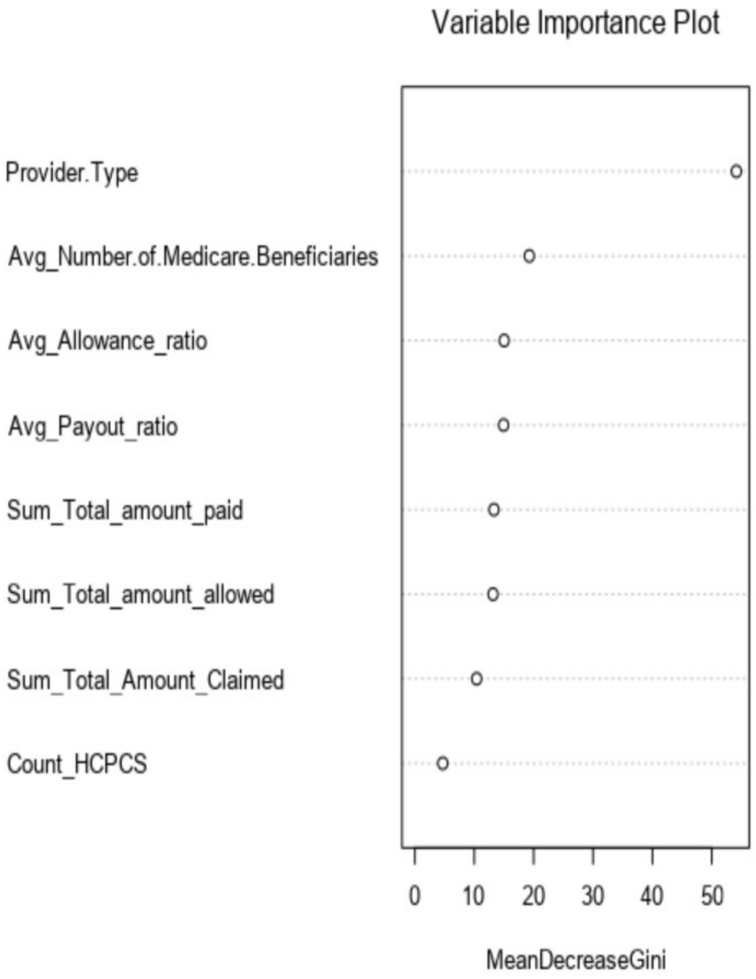
Call:
randomForest(formula = Fraud_Label ~ Provider.Type + Count_HCPCS + Avg_Payout_ratio + Avg_Allowance_ratio + Avg_Number.of.Medicare.Beneficiaries + Sum_Total_Amount_Claimed + Sum_Total_amount_paid + Sum_Total_amount_allowed, data = the.data, ntree = 500, replace = TRUE)

Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 2

OOB estimate of the error rate: 25.3%

Confusion Matrix:

	No	Yes	Classification Error
No	103	52	0.335
Yes	22	115	0.161


Variable Importance Plot
Provider.Type
Avg_Number.of.Medicare.Beneficiaries
Avg_Allowance_ratio
Avg_Payout_ratio
Sum_Total_amount_paid
Sum_Total_amount_allowed
Sum_Total_Amount_Claimed
Count_HCPCS
MeanDecreaseGini