A comparative study of using various Machine Learning and Deep Learning based fraud detection models for Universal Health Coverage schemes and assessing the impact of COVID-19 in healthcare fraud

Rohan Yashraj Gupta
Doctoral Research Scholar, SSSIHL

rohanyashrajgupta@sssihl.edu.in

+91 - 9593256368

Agenda

- Health Scheme
- Methodology
- Key Concepts
- Results
- COVID-19 impact
- Observations

Universal Group Health Insurance Scheme

World’s largest

500 Million +
Covered Population in India

15 Million +
Claims paid till date

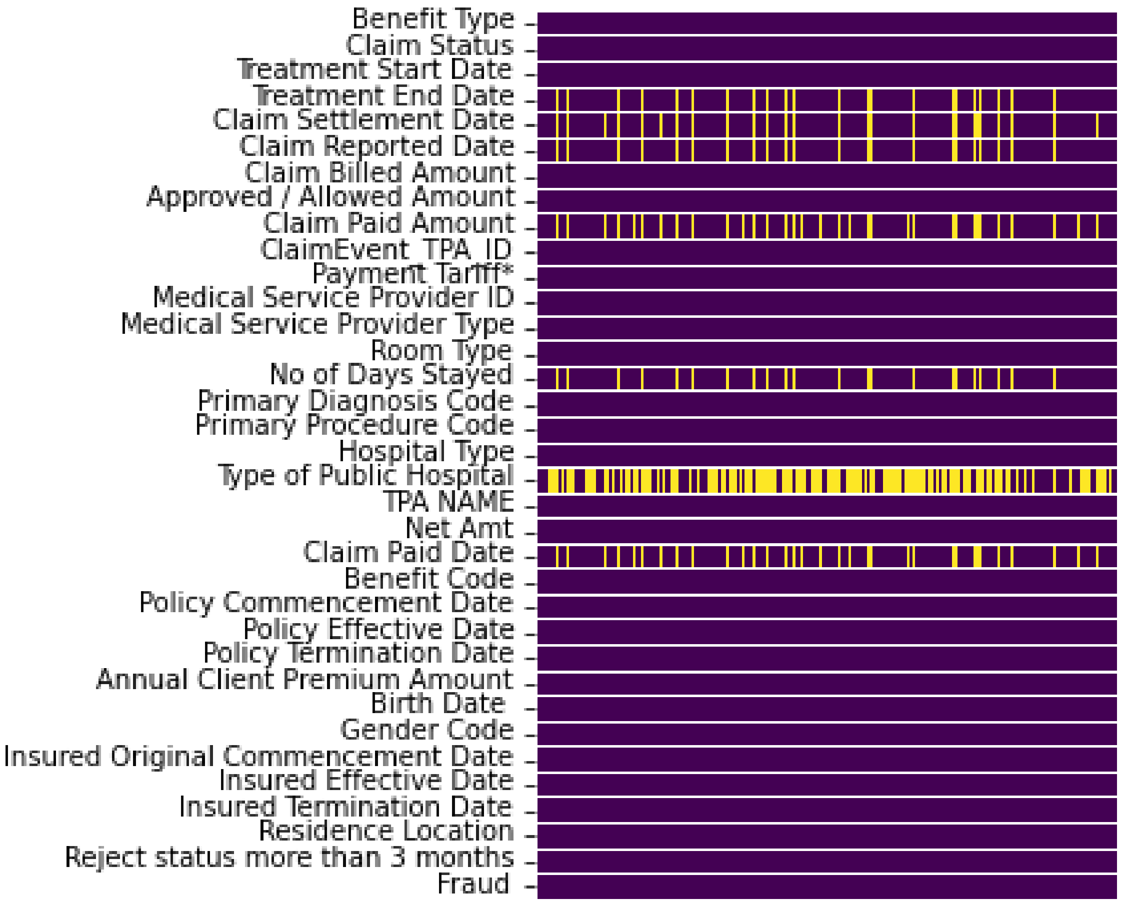
50%
Contribution to insurance penetration in India

Hospitals empaneled
24,334 +

Packages
2500+

- Aug-2019 to Aug-2020
- Policy and claims data used
- 380,000 record +
- 51 features – policy and claims data combined
- 12% fraudulent claims

Data challenges - missing values



Other missing values in the dataset were handled using statistical methods such as – mean value imputation, median value imputation, average value imputation and random selection.

Data challenges – some more

Feature engineering

Reject_status_more_than_3_months

Claims status month on month was tabulated and claims with reject status more than 3 month was labelled

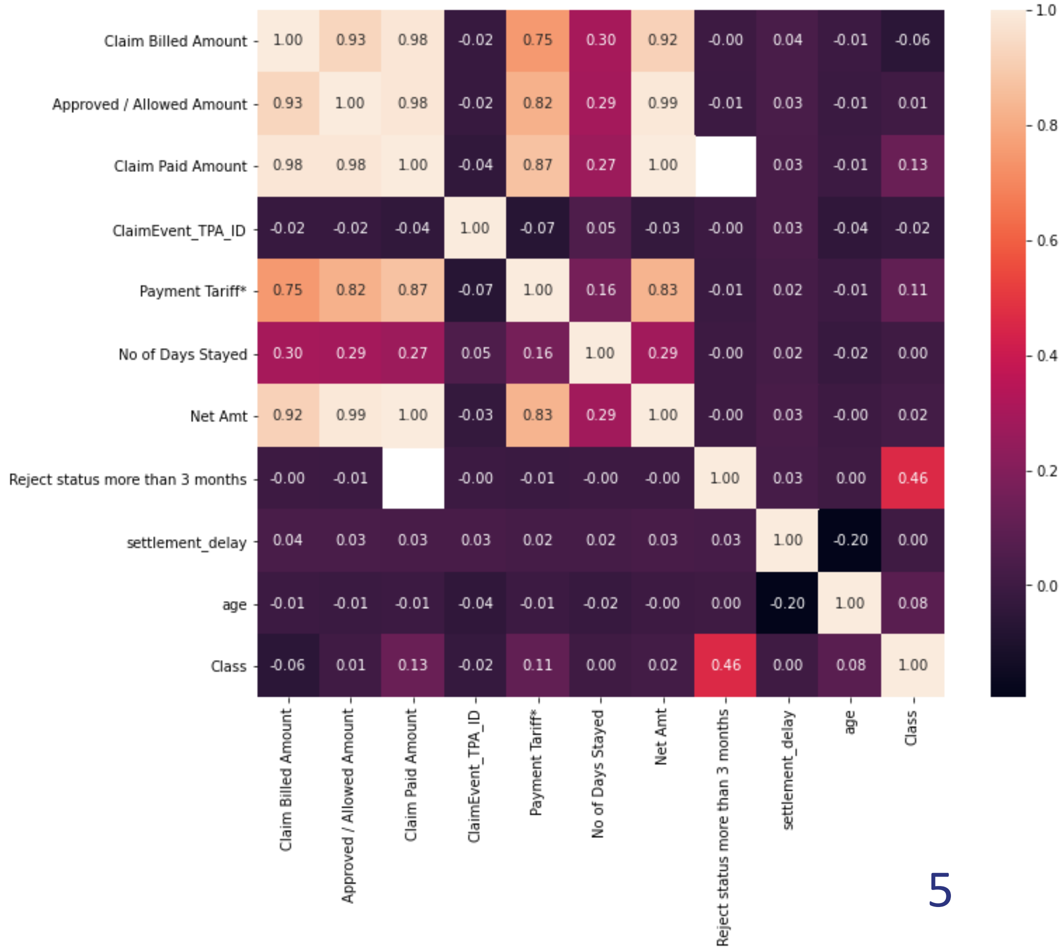
Reclaimed

The status of the claim changed from cancelled to paid

Diagnosis code

Derived diagnosis code from the list of procedures

Correlation



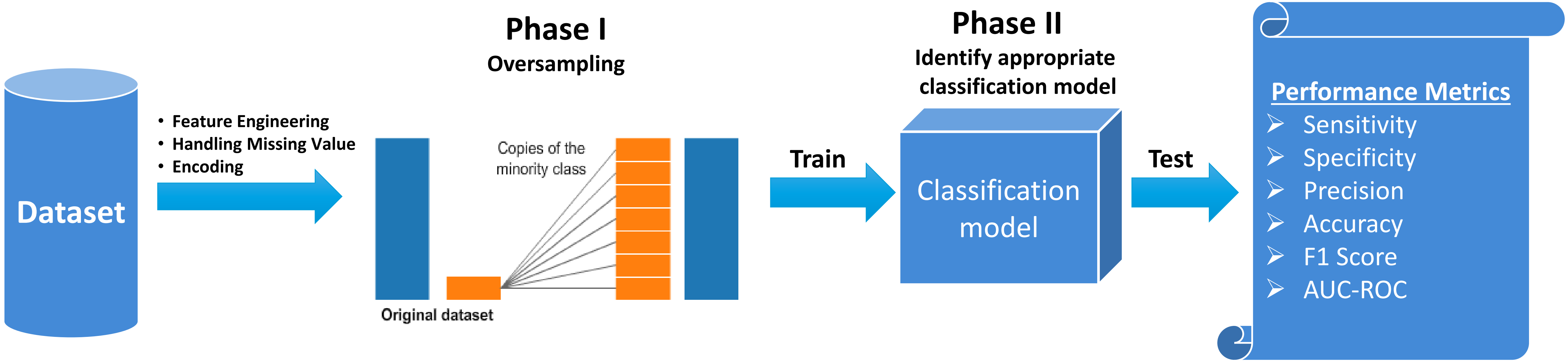
Methodology

Phase I

- Deals with data pre-processing
- Major challenge of data imbalance is also handled

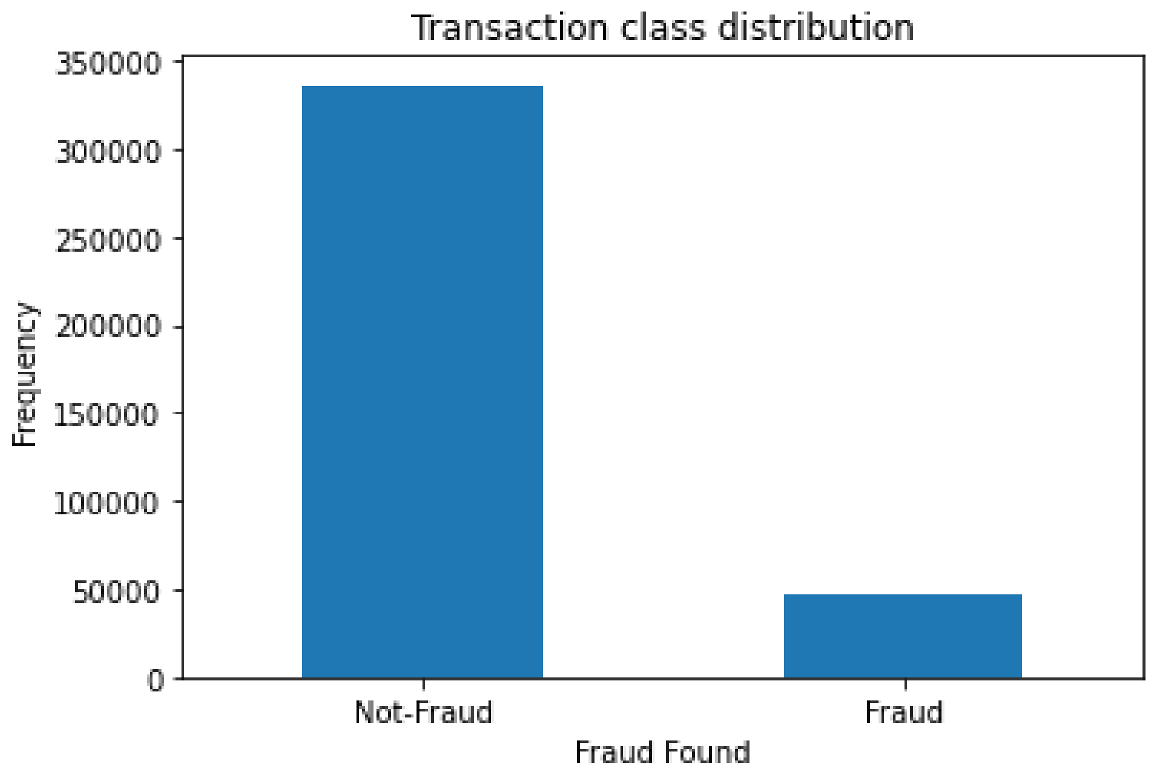
Phase II

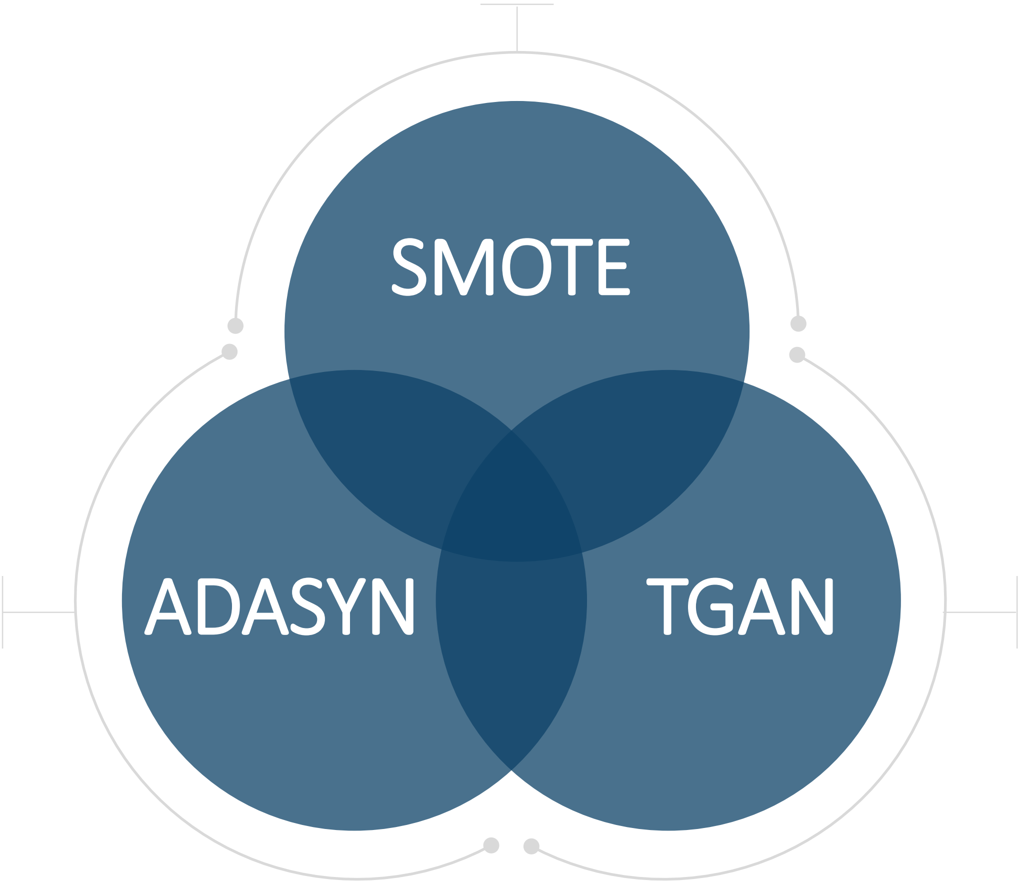
- Build a classification model to classify claims as fraudulent or not



The goal is to find a golden combination of a technique in Phase I and a specific model in Phase II for assured best performance of a Fraud Detection Model

Key concepts: Phase I



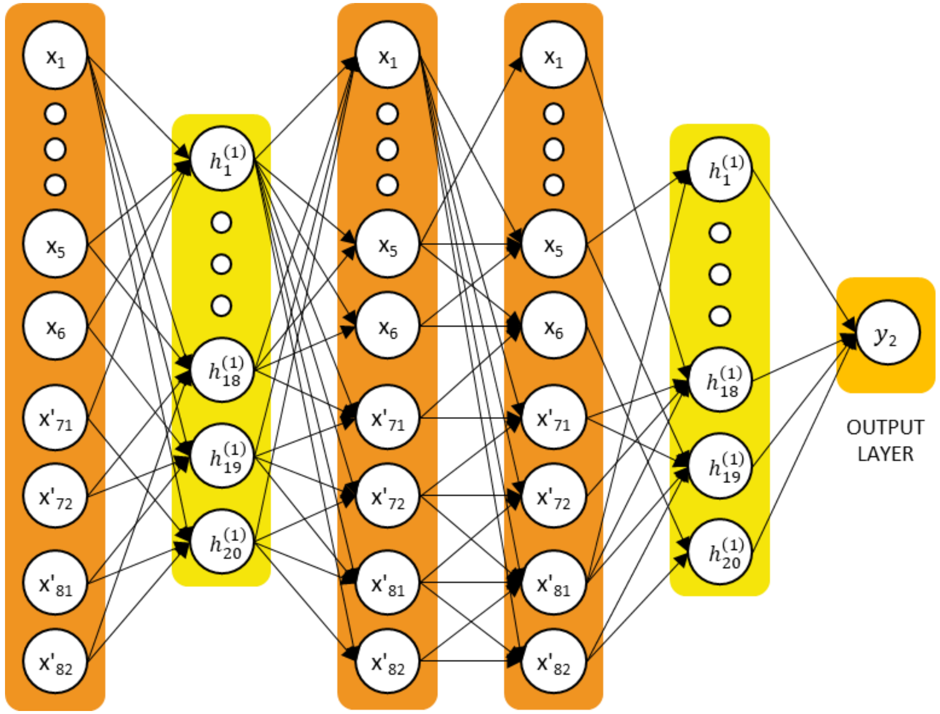


1. N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
2. Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.
3. L. Xu and K. Veeramachaneni, "Synthesizing Tabular Data using Generative Adversarial Networks," *arXiv*, Nov. 2018.

Key concepts: Phase II

- Decision tree
- Random forest
- XGBoost
- LightGBM
- GBM

Layer (type)	Output Shape	No. of parameters
dense_1 (Dense)	(None,49)	4,851
dense_2 (Dense)	(None,80)	4,000
dropout_1 (Dropout)	(None,80)	0
dense_3 (Dense)	(None,80)	6,480
dense_4 (Dense)	(None,49)	3,969
dense_5 (Dense)	(None,1)	50



Results

Machine learning models			AUC-ROC	Recall	Specificity	Precision	Accuracy	F1 Score
Decision Tree	Baseline	M1	**0.9566**	0.9248	0.9885	0.9174	0.9808	0.9211
	SMOTE	M2	0.9534	0.9208	0.9860	0.9006	0.9781	0.9106
	ADASYN	M3	0.9508	0.9155	0.9862	0.9016	0.9776	0.9085
	TGANs	M4	0.9548	0.9214	0.9883	0.9155	0.9801	0.9185
Random Forest	Baseline	M5	0.9462	0.8947	0.9977	0.9818	0.9852	0.9362
	SMOTE	M6	0.9493	0.9027	0.9959	0.9682	0.9846	0.9343
	ADASYN	M7	0.9500	0.9057	0.9942	0.9556	0.9834	0.9300
	TGANs	M8	0.9460	0.8942	0.9977	0.9820	0.9852	0.9361
XGBoost	Baseline	M9	0.9307	0.8615	0.9999	0.9989	0.9831	0.9252
	SMOTE	M10	0.9458	0.8970	0.9945	0.9572	0.9826	0.9262
	ADASYN	M11	0.9270	**0.9835**	0.8705	0.5119	0.8842	0.6733
	TGANs	M12	0.9111	0.8223	**1.0000**	**1.0000**	0.9784	0.9025
LightGBM	Baseline	M13	0.9486	0.8977	0.9994	0.9952	**0.9871**	**0.9440**
	SMOTE	M14	0.9499	0.9014	0.9988	0.9905	0.9869	0.9438
	ADASYN	M15	0.9523	0.9105	0.9940	0.9547	0.9839	0.9320
	TGANs	M16	0.9482	0.8970	0.9994	0.9950	0.9870	0.9435
GBM	Baseline	M17	0.9425	0.8852	0.9997	0.9975	0.9858	0.9380
	SMOTE	M18	0.9451	0.8958	0.9945	0.9576	0.9825	0.9257
	ADASYN	M19	0.9288	0.9779	0.8796	0.5288	0.8916	0.6864
	TGANs	M20	0.9282	0.8566	0.9992	0.9992	0.9224	0.9224

Deep learning models			AUC-ROC	Recall	Specificity	Precision	Accuracy	F1 Score
Neural Networks	Baseline	M21	0.9406	0.8826	0.9986	0.9885	**0.9845**	0.9325
	Weighted	M22	**0.9557**	0.9418	0.9644	0.7852	0.9617	0.8564
	Undersampled	M23	0.9525	0.9374	0.9676	0.9663	0.9526	**0.9516**
	SMOTE	M24	0.9496	0.9533	0.9459	0.7087	0.9468	0.8130
	ADASYN	M25	0.9389	**0.9822**	0.8955	0.5650	0.9061	0.7173
	TGANs	M26	0.9392	0.8795	**0.9989**	**0.9908**	0.9844	0.9318

COVID-19 Impact on Healthcare Fraud

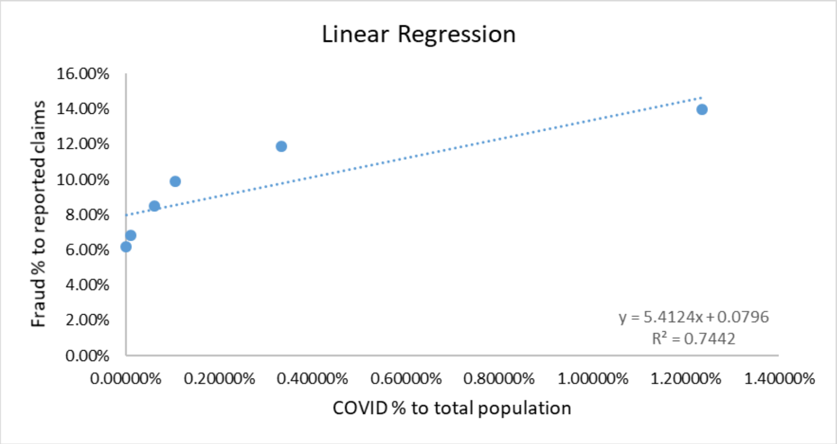
Month	Fraud rate	COVID-19 rate
Aug-19	0.58%	0.00000%
Sep-19	1.08%	0.00000%
Oct-19	1.19%	0.00000%
Nov-19	2.65%	0.00000%
Dec-19	5.14%	0.00000%
Jan-20	6.86%	0.00000%
Feb-20	4.21%	0.00003%
Mar-20	6.16%	0.00137%
Apr-20	6.84%	0.01053%
May-20	8.51%	0.06253%
Jun-20	9.89%	0.10617%
Jul-20	11.89%	0.33460%
Aug-20	13.96%	1.23567%

COVID-19 rate		Fraud rate	
Mean	0.002918111	Mean	0.09543601
Standard Error	0.001952094	Standard Error	0.01224767
Median	0.0008435	Median	0.09201897
Standard Deviation	0.004781635	Standard Deviation	0.030000542
Sample Variance	2.2864E-05	Sample Variance	0.000900033
Range	0.012343	Range	0.078036001
Minimum	1.36667E-05	Minimum	0.061601365
Maximum	0.012356667	Maximum	0.139637366
Count	6	Count	6

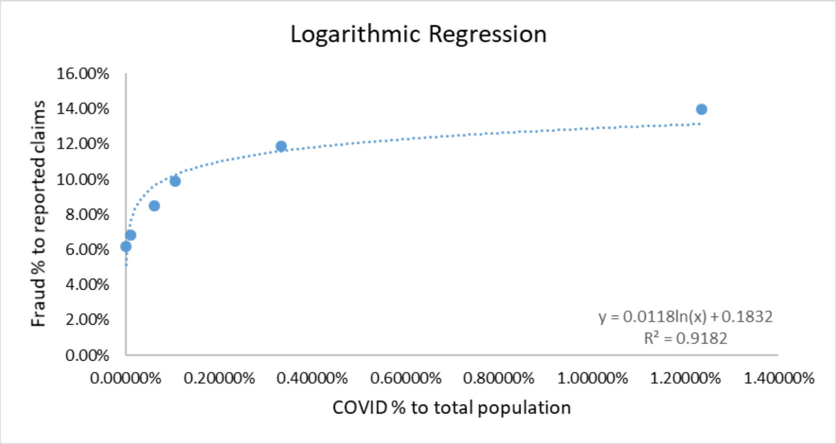
Observations

Found a correlation of **86.26 %** between increase in COVID-19 cases in India and healthcare fraud based on this data

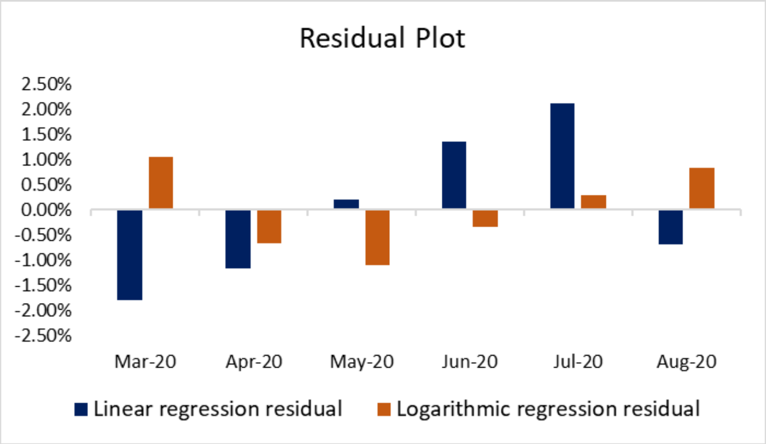
Linear regression
y = 5.4124x + 0.0796



Logarithmic regression
y = 0.0118ln(x) + 0.1832



Residual plot of the linear and logarithmic model



where,

y → Predicted fraud cases %
x → Infected COVID-19 cases %

	Linear Regression	Logarithmic Regression
Multiple R	0.5538	0.8431
R Square	0.7442	0.9182

Acknowledgements

- Bhagawan Sri Sathya Sai Baba, Founder chancellor, SSSIHL
- Research Supervisors:
  - Satya Sai Mudigonda, Honorary professor (Actuarial Science), Department of Mathematics and Computer Science, SSSIHL
  - Dr. Pallav Kumar Baruah, Department of Mathematics and Computer Science, SSSIHL
  - Phani Krishna Kandala, Visiting faculty, SSSIHL