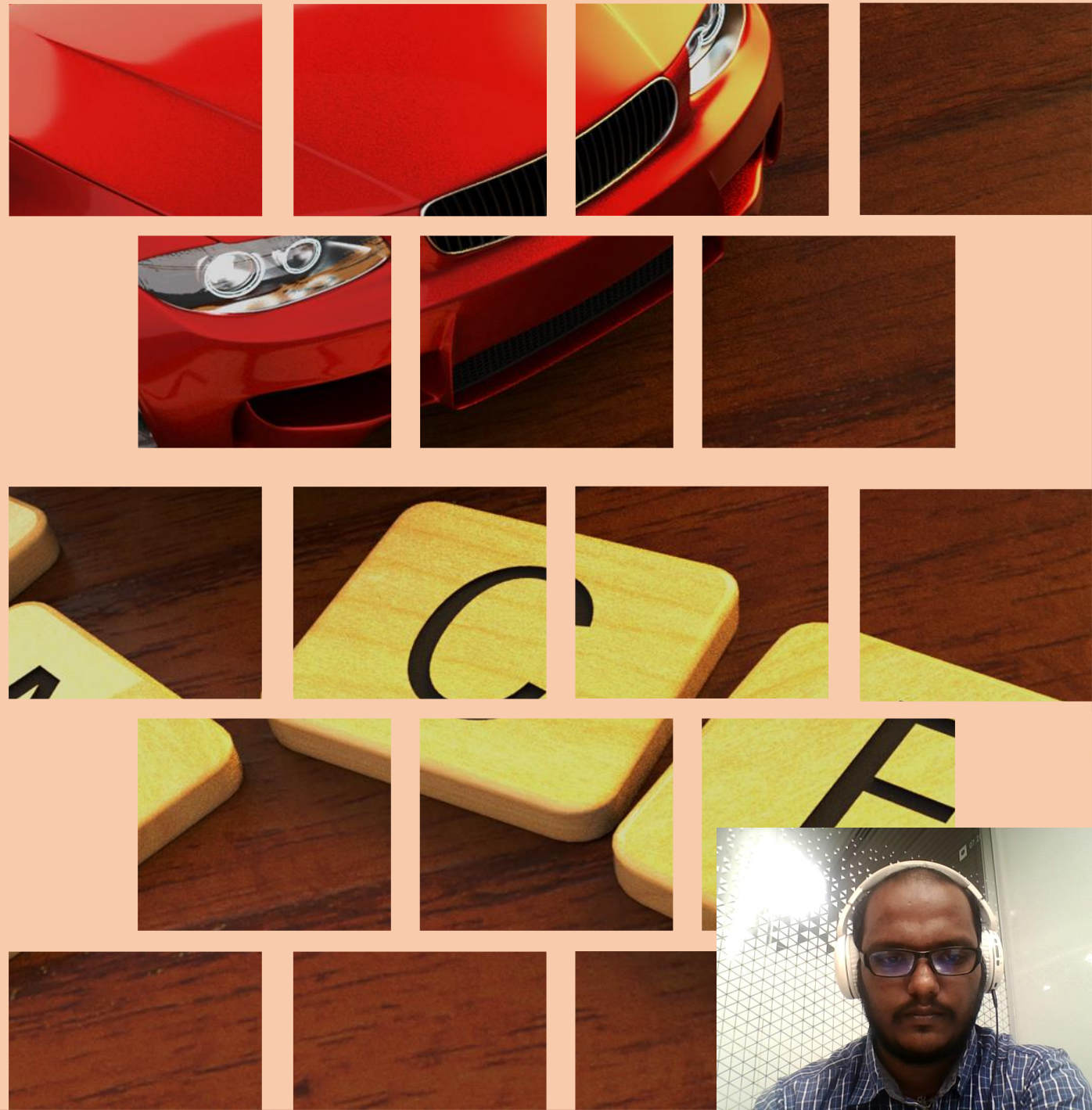
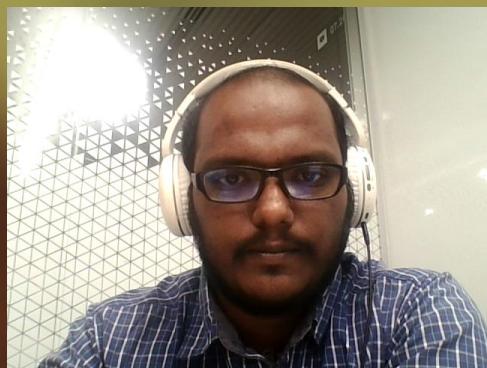
Enhancing Fraud Detection in the Automobile Insurance Sector: A Comparative Analysis Utilizing PyCaret



By
N. Ambarish

AGENDA



1

INTRODUCTION

In this chapter, we can learn about insurance fraud in the automobile sector and how machine learning helps in reducing fraud.

2

LITERATURE REVIEW

In this chapter, we can learn about what other research has been carried out for fraud detection and how our research is different from theirs.

3

OBJECTIVES

This thesis aims to identify an optimal machine learning model that can efficiently and affordably address insurance fraud in the automobile sector.

4

METHODOLOGY

We can learn how our research has been carried out in this section.

5

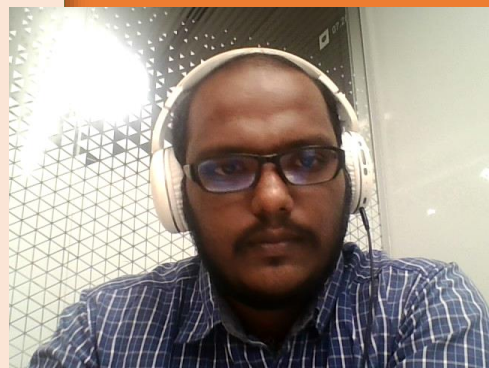
RESULTS & DISCUSSION

The result of our research is listed here.

6

CONCLUSION & FUTURE WORK

Conclusion from our research and future works to be looking for.



1

INTRODUCTION

In this chapter, we can learn about insurance fraud in the automobile sector and how machine learning helps in reducing fraud.

- **What is Insurance?**

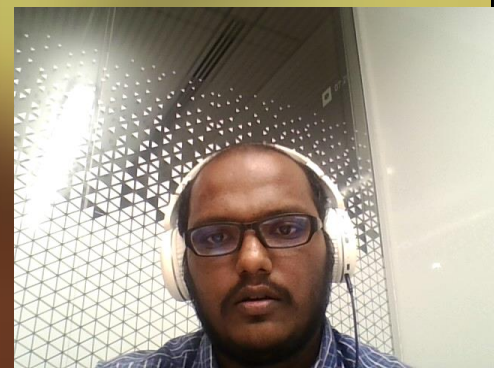
  A contract between an insurance company and an individual to be protected from sudden financial loss is known as an insurance policy.

- **What are the types of Insurance?**

  There are two main types of insurance: LIFE insurance and GENERAL insurance. LIFE insurance provides financial protection for your loved ones in the event of an untimely death. On the other hand, GENERAL insurance, also referred to as non-life insurance, protects us against unexpected financial losses in various areas such as auto, home, health, and travel.

- **What is insurance fraud and how it came to be known?**

  Insurance fraud is when someone lies to an insurance company to get money they shouldn't. This became a problem in the late 1980s after the US insurance industry lost $15 billion in 1987. Insurance fraud has caused $308.6 billion in losses in the US as of 2022.



1

INTRODUCTION

In this chapter, we can learn about insurance fraud in the automobile sector and how machine learning helps in reducing fraud.

- **How has insurance fraud affected the automobile insurance sector?**

  In 2021, $196.82 billion has been lost due to auto insurance fraud in US. In 2019, India faced a total of $6.25 billion in loss due to insurance fraud and 90% of it was due to automobile insurance.

- **When was machine learning introduced into the insurance sector?**

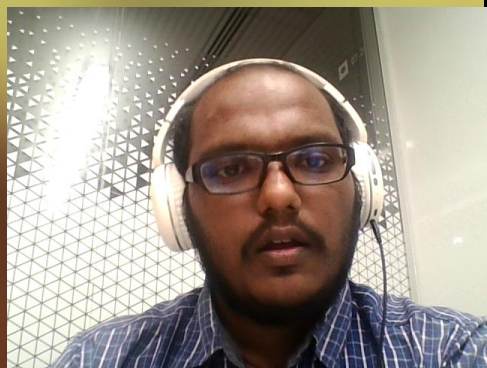
  To take measures against fraud claims Machine Learning was introduced in the early 2000s. The fraud claims have been reduced between 20% to 60%.

- **How was the research in fraud detection using ML so far?**

  Multiple types of research have been conducted over the decade to increase fraud prediction. Many ML models have been introduced for that purpose. Fraud detection is a classification problem which is a supervised learning method.

- **What is this study focused on?**

  Our study focuses on detecting the best model for fraud detection in automobile insurance using machine learning methods. We are implementing PyCaret for that purpose.



2

LITERATURE REVIEW

In this chapter, we can learn about what other research has been carried out for fraud detection and how our research is different from theirs.

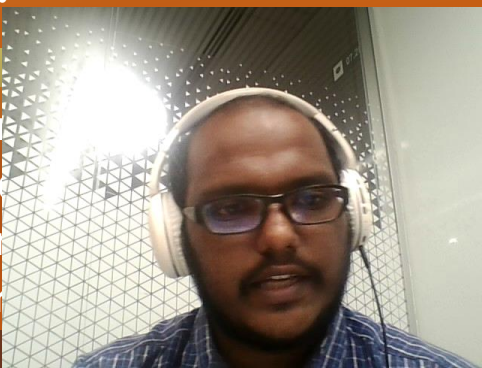
1. Derrig,1991 was one of the first-ever research papers published on fraud detection in the automobile insurance sector, based on 597 claims from accidents that occurred in 1985,1986 and 1989.

2. Brockett, 1998 was the earliest one to present a paper on automobile fraud detection using a machine learning-based approach. He then published another paper on fraud detection in 2002 using the same dataset that we are using in this study using the PRIDIT method.

3. Artis (2002) and Brockett (2002) both emphasized the significance of binary choice model for fraud detection, using similar data.

4. Derrig, an upper-echelon member of the insurance industry, authored several papers on automobile insurance fraud. In 2002, he wrote a comprehensive paper on insurance fraud, covering various types of fraudulent claims and how to classify them.

5. In 2002, Viaene presented a paper that discussed the issues a
insurance fraud. The paper also provided some insights into insuran
presented another paper that focused on methods for distinguishi
provided brief insights into the process of making an insurance clai



2

LITERATURE REVIEW

In this chapter, we can learn about what other research has been carried out for fraud detection and how our research is different from theirs.

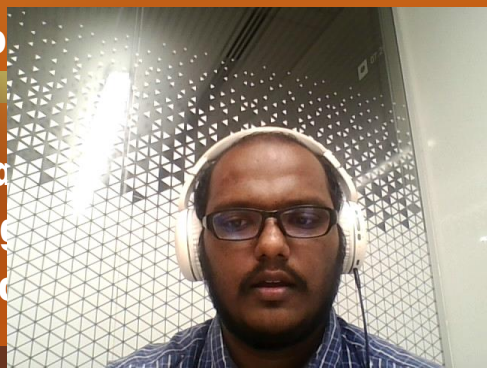
6 In 2018, Nikhil Rai presented a research paper on automobile insurance fraud detection. The study utilized machine learning methods and data balancing techniques, resulting in a 99% accuracy rate. In 2023, Debener presented a research paper on the same dataset, but this time, they combined both supervised and unsupervised learning methods for fraud detection.

7 Cai Jie presented feature selection methods in ML in 2018. Elssied used the ANOVA test from the Filter method for feature selection in 2013. In 2018, Sapna found the Wrapper method best suited for classification problems after trying all three feature selection methods on a dataset.

8 In 2018, Sharmila presented a paper on the impact of class imbalance on data. She used ADASYN and SMOTE methods on the dataset and compared the results, finding that ADASYN performed better. Later, in 2021, Rohan used a similar comparative approach on the same dataset. He tried various balancing techniques and found that the TGAN method was most effective in solving classification problems.

9 In 2021, Hanafy conducted a comparative study on various SMOTE techniques for classification problems. They introduced hybrid methods in balancing technique which proved to be better than normal methods. In 2023, David used the SMOTE method on the same dataset with different ML models, resulting in an accuracy of 95% fo

10 In 2020, Moez proposed a new method for ML known as PyCa proposed a low-code ML process using PyCaret and Streamlit making 2023, Pawan Whig utilized PyCaret for the diabetes classification pr



3

OBJECTIVES

This thesis aims to identify an optimal machine learning model that can efficiently and affordably address insurance fraud in the automobile sector.

Aims and objectives

- **To compare and analyze various machine learning methods and approaches for automobile insurance fraud detection.**

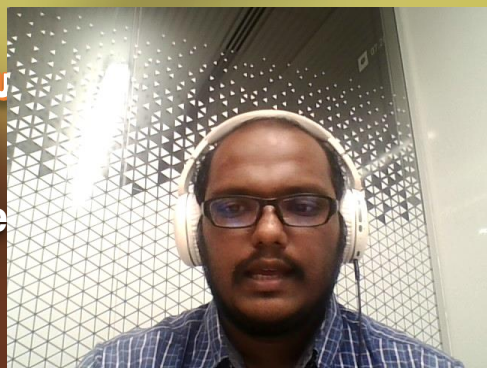
We can use PyCaret, which integrates multiple machine learning models, to achieve our objective. It provides a detailed comparison and analysis of the models and can be tuned for better results.

- **To determine the best Machine Learning (ML) model for fraud detection in automobile insurance by emphasizing the advantages and limitations.**

PyCaret compares different machine learning models to select the best one. It transforms, selects features, and resamples the dataset for optimal performance. Evaluating the model is important to identify its limitations. One can further study the model to identify its pros and cons.

- **To evaluate the model's performance by conducting assessments and evalu**

Evaluated with pre-determined algorithms, PyCaret's best model was te
metrics. Further analysis can improve its performance.



4

METHODOLOGY

We can learn how our research has been carried out in this section.

Data Selection

- Data Gathering
- Data Cleaning

Data Pre-Processing

- EDA
- Feature Comparison

Feature Selection

- Feature Selection for Data Input.

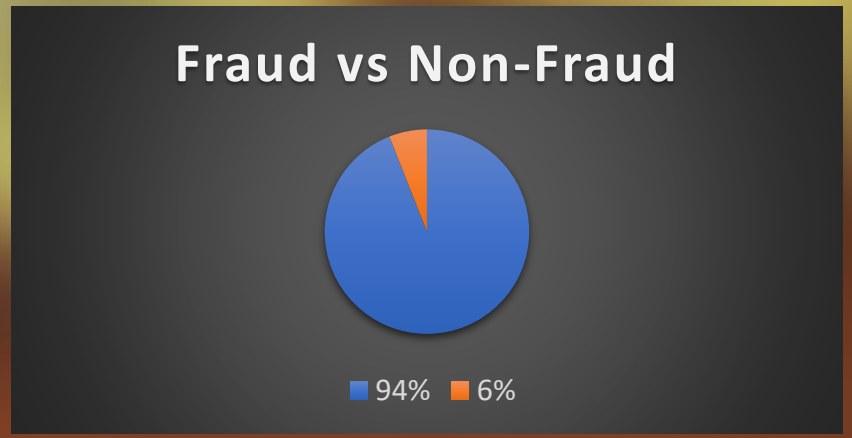
Data Implementation

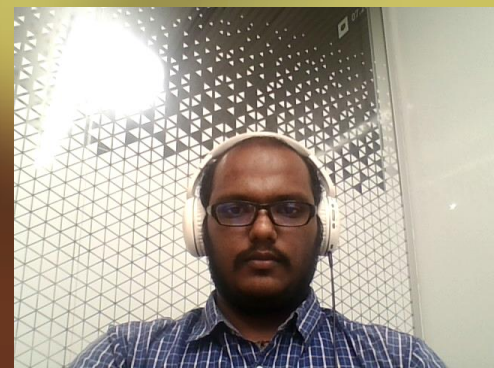
- Data transformation using Label encoding.
- Implementing in PyCaret.
- Testing the data again after data balancing techniques.

Model Output

- Checking parameters for model evaluation

We are using automobile insurance fraud detection data in our study from Kaggle. The data set has 15,420 samples with major data imbalance.





4

METHODOLOGY

We can learn how our research has been carried out in this section.

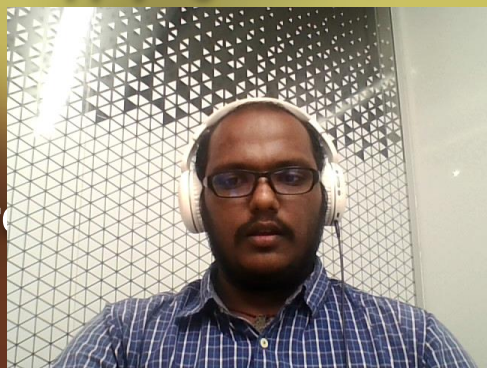
- Our data doesn't have any missing or Null values.

- Exploratory Data Analysis has been carried out in the dataset to provide insights for the feature selection process.
- Using Pearson's heatmap method values more than 0.7 is dropped.

- After subjecting the processed data to an ANOVA test, the K best model is selected using f-score.
- 17 features have been chosen based on their p-values, and alpha has been set to 0.05.

- Implementing the original encoded data as the first method then applying SMOTE technique.
- Implementing the feature-selected data as the second method then applying SMOTE technique.

- The PyCaret output is evaluated, then further tuned and boosted t
performance.



4

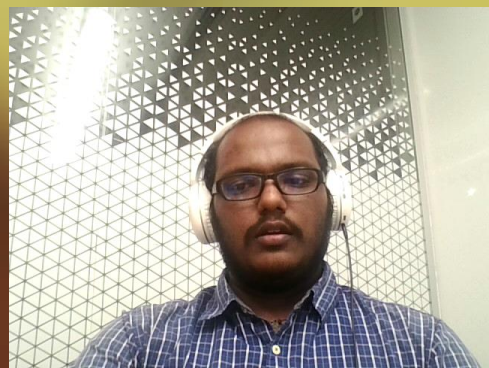
METHODOLOGY

We can learn how our research has been carried out in this section.

Based on our data analysis, we have found that male drivers between the ages of 31 and 35, who own sedans from Pontiac and Toyota, valued between $20,000 to $29,000 and live in urban areas, tend to opt for 'All Perils' and 'Collision' policies for their cars. However, we have also discovered that these drivers tend to file fraudulent claims when there is no police report or witness present for the incident.

- **Encoded our data using Label Encoding.**
- **Using the feature selection method, based on the p-values these features are selected.**

	Input_features	Score	P_value
2	DayOfWeek	4.699458	0.0302
3	Make	5.677963	0.0172
4	AccidentArea	17.321735	0.0000
7	Sex	13.845476	0.0002
9	Fault	270.838424	0.0000
10	PolicyType	50.356504	0.0000
11	VehiclePrice	58.614157	0.0000
13	Deductible	4.641420	0.0312
17	PastNumberOfClaims	8.420544	0.0037
18	AgeOfVehicle	7.627339	0.0058
19	AgeOfPolicyHolder	15.770398	0.0001
20	PoliceReportFiled	3.951655	0.0468
22	AgentType	8.144971	0.0043
24	AddressChange-Claim	21.874567	0.0000
26	Year	9.457663	0.0021
27	BasePolicy	390.046626	0.0000



5

RESULTS & DISCUSSION

The result of our research is listed here.

The first method of our study is conducted using encoded original data and SMOTE data.

PyCaret setup for encoded data.

	Description	Value
0	Session id	4627
1	Target	FraudFound
2	Target type	Binary
3	Original data shape	(28994, 33)
4	Transformed data shape	(28994, 33)
5	Transformed train set shape	(20295, 33)
6	Transformed test set shape	(8699, 33)
7	Numeric features	32
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10

PyCaret setup for encoded data.

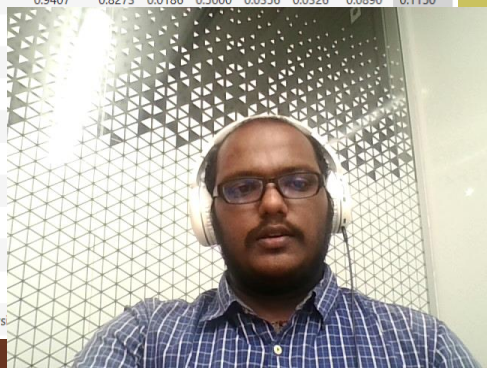
	Description	Value
0	Session id	5959
1	Target	FraudFound
2	Target type	Binary
3	Original data shape	(15420, 33)
4	Transformed data shape	(15420, 33)
5	Transformed train set shape	(10794, 33)
6	Transformed test set shape	(4626, 33)
7	Numeric features	32
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10
14	CPU Jobs	-1
15	Use GPU	False
16	Log Experiment	False
17	Experiment Name	clf-default-name
18	USI	c1fc

Encoded data output

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.9762	0.9973	0.9750	0.9774	0.9762	0.9524	0.9524	0.1190
catboost	CatBoost Classifier	0.9748	0.9976	0.9760	0.9738	0.9749	0.9496	0.9497	3.5800
lightgbm	Light Gradient Boosting Machine	0.9596	0.9937	0.9582	0.9610	0.9596	0.9193	0.9193	0.3170
et	Extra Trees Classifier	0.9546	0.9918	0.9625	0.9475	0.9549	0.9091	0.9093	0.2560
rf	Random Forest Classifier	0.9508	0.9898	0.9568	0.9454	0.9511	0.9016	0.9016	0.2930
gbc	Gradient Boosting Classifier	0.9196	0.9785	0.9403	0.9031	0.9212	0.8392	0.8400	0.5050
dt	Decision Tree Classifier	0.9064	0.9064	0.9278	0.8898	0.9083	0.8128	0.8136	0.0380
ada	Ada Boost Classifier	0.9000	0.9688	0.9307	0.8770	0.9030	0.8000	0.8017	0.1560
lda	Linear Discriminant Analysis	0.8673	0.9466	0.9280	0.8275	0.8748	0.7345	0.7401	0.0370
ridge	Ridge Classifier	0.8672	0.0000	0.9279	0.8275	0.8748	0.7344	0.7400	0.0210
lr	Logistic Regression	0.8498	0.9269	0.8672	0.8381	0.8524	0.6996	0.7001	0.4330
nb	Naive Bayes	0.8097	0.9097	0.9148	0.7560	0.8278	0.6194	0.6337	0.0250
qda	Quadratic Discriminant Analysis	0.8096	0.9421	0.9706	0.7343	0.8360	0.6192	0.6541	0.0280
knn	K Neighbors Classifier	0.7989	0.9089	0.9552	0.7278	0.8261	0.5978	0.6294	0.0900
svm	SVM - Linear Kernel	0.5329	0.0000	0.5883	0.6237	0.4561	0.0661	0.1282	0.1250
dummy	Dummy Classifier	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0220

SMOTE data output

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.9520	0.9521	0.3034	0.7443	0.4299	0.4095	0.4555	0.1000
catboost	CatBoost Classifier	0.9457	0.9544	0.1455	0.7286	0.2420	0.2268	0.3095	1.0770
lightgbm	Light Gradient Boosting Machine	0.9450	0.9250	0.1376	0.7019	0.2289	0.2138	0.2937	0.2900
gbc	Gradient Boosting Classifier	0.9415	0.8754	0.0263	0.9333	0.0510	0.0478	0.1475	0.2510
rf	Random Forest Classifier	0.9407	0.8419	0.0139	0.5500	0.0270	0.0250	0.0810	0.1330
et	Extra Trees Classifier	0.9407	0.8273	0.0186	0.5000	0.0356	0.0326	0.0890	0.1150
ridge	Ridge Classifier								
dummy	Dummy Classifier								
lr	Logistic Regression								
lda	Linear Discriminant Analysis								
knn	K Neighbors Classifier								
ada	Ada Boost Classifier								
dt	Decision Tree Classifier								
nb	Naive Bayes								
svm	SVM - Linear Kernel								
qda	Quadratic Discriminant Analysis								



5

RESULTS & DISCUSSION

The result of our research is listed here.

The second method of our study is conducted using feature-selected data.

PyCaret setup for Feature-selected data.

1	Target	FraudFound
2	Target type	Binary
3	Original data shape	(15420, 17)
4	Transformed data shape	(15420, 17)
5	Transformed train set shape	(10794, 17)
6	Transformed test set shape	(4626, 17)
7	Numeric features	16
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10

PyCaret setup for Feature-selected data.

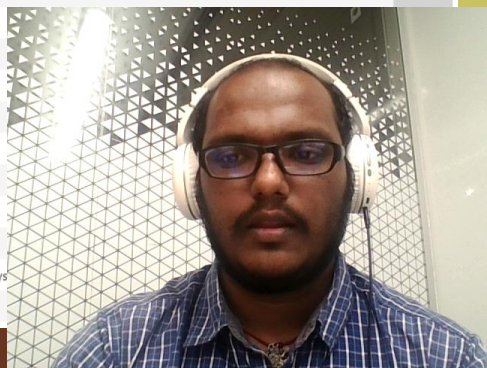
	Description	Value
0	Session id	2919
1	Target	FraudFound
2	Target type	Binary
3	Original data shape	(28994, 17)
4	Transformed data shape	(28994, 17)
5	Transformed train set shape	(20295, 17)
6	Transformed test set shape	(8699, 17)
7	Numeric features	16
8	Preprocess	True
9	Imputation type	simple
10	Numeric imputation	mean
11	Categorical imputation	mode
12	Fold Generator	StratifiedKFold
13	Fold Number	10

Feature-selected data output

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.9412	0.8245	0.0217	0.7667	0.0418	0.0389	0.1190	0.2980
catboost	CatBoost Classifier	0.9407	0.8194	0.0464	0.5371	0.0848	0.0765	0.1439	2.1520
ridge	Ridge Classifier	0.9402	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0310
lightgbm	Light Gradient Boosting Machine	0.9402	0.8193	0.0417	0.5442	0.0766	0.0683	0.1344	0.4720
dummy	Dummy Classifier	0.9402	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0260
lr	Logistic Regression	0.9401	0.7961	0.0000	0.0000	0.0000	-0.0002	-0.0008	1.6620
lda	Linear Discriminant Analysis								
knn	K Neighbors Classifier								
xgboost	Extreme Gradient Boosting								
rf	Random Forest Classifier								
ada	Ada Boost Classifier								
et	Extra Trees Classifier								
dt	Decision Tree Classifier								
svm	SVM - Linear Kernel								
qda	Quadratic Discriminant Analys								
nb	Naive Bayes								

SMOTE data output

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.9318	0.9755	0.9575	0.9107	0.9335	0.8635	0.8647	0.3720
rf	Random Forest Classifier	0.9292	0.9793	0.9648	0.9007	0.9316	0.8584	0.8606	0.2760
dt	Decision Tree Classifier	0.9083	0.9125	0.9452	0.8802	0.9115	0.8165	0.8188	0.0820
xgboost	Extreme Gradient Boosting	0.8968	0.9574	0.9529	0.8569	0.9023	0.7936	0.7988	0.1280
catboost	CatBoost Classifier	0.8872	0.9548	0.9495	0.8443	0.8938	0.7743	0.7806	4.3070
lightgbm	Light Gradient Boosting Machine	0.8744	0.9430	0.9551	0.8225	0.8838	0.7488	0.7589	0.3930
knn	K Neighbors Classifier	0.8666	0.9421	0.9803	0.7987	0.8802	0.7331	0.7529	0.1060
gbc	Gradient Boosting Classifier	0.8323	0.9099	0.9316	0.7774	0.8474	0.6646	0.6783	0.1920
ada	Ada Boost Classifier	0.8096	0.8875	0.8827	0.7703	0.8226	0.6192	0.6261	0.1010
ridge	Ridge Classifier	0.7795	0.0000	0.8177	0.7598	0.7876	0.5589	0.5607	0.0200
lda	Linear Discriminant Analysis	0.7795	0.8645	0.8178	0.7598	0.7876	0.5589	0.5607	0.0180
lr	Logistic Regression	0.7765	0.8644	0.7993	0.7646	0.7815	0.5530	0.5537	0.2770
qda	Quadratic Discriminant Analysis	0.7019	0.8416	0.9313	0.6384	0.7501	0.4038	0.4775	0.0190
nb	Naive Bayes	0.6980	0.8442	0.9620	0.6299	0.7612	0.3959	0.4659	0.0150
svm	SVM - Linear Kernel	0.6934	0.0000	0.7820	0.7111	0.7025	0.3867	0.4442	0.1200
dummy	Dummy Classifier	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0150



5

RESULTS & DISCUSSION

The result of our research is listed here.

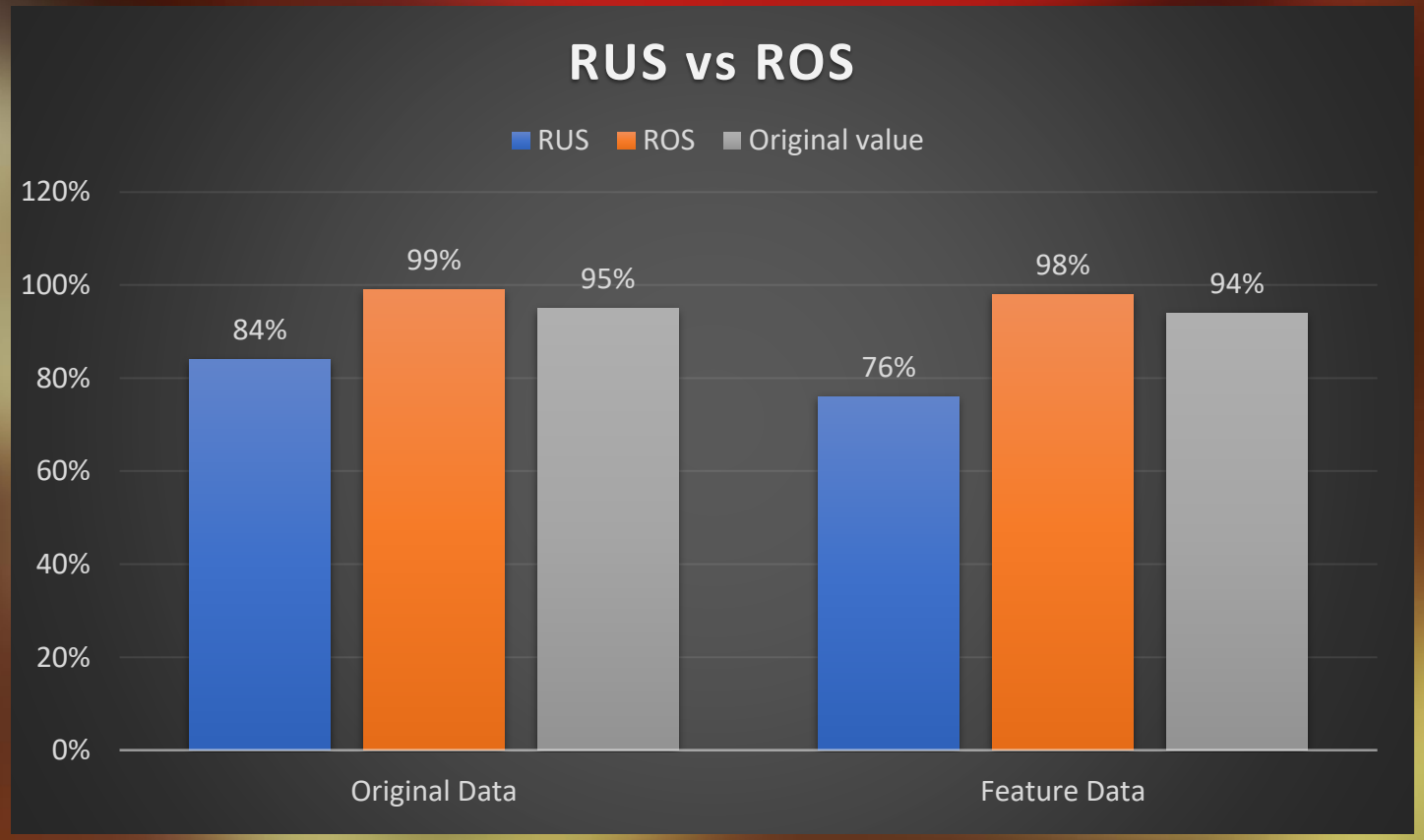
Result Discussion

5

RESULTS & DISCUSSION

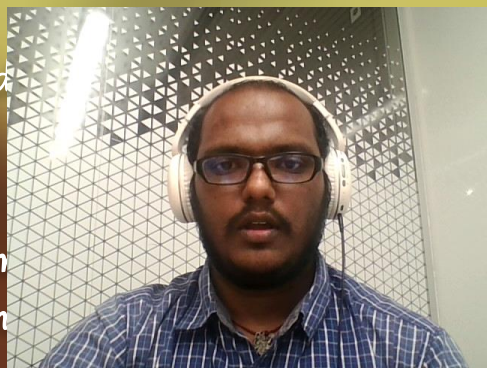
The result of our research is listed here.

For our understanding of the sampling technique, we used ROS and RUS methods on both the data



The Extra Trees Classifier's randomization, ensemble learning, and noise robustness make it ideal for ad datasets and implementing techniques like Random Over-Sampling.

The synergy between RUS and boosting enhances class imbalance handling, allowing boosting algorithm minority class patterns without being overwhelmed, resulting in improved predictive performance on im



6

CONCLUSION & FUTURE WORK

Conclusion from our research and future works to be looking for.

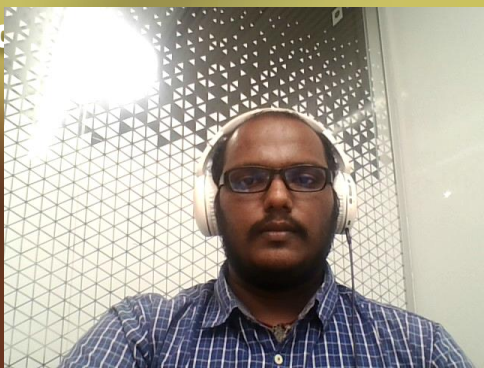
Conclusion

Our study concluded that the XGBoost Machine Learning model performs well after using the SMOTE technique.

Future Works & Recommendation

The following points are worth considering:-

- **In the future, newly acquired data can be used to work on the entire model using PyCaret, instead of programming separately.**
- **While machine learning is capable of reading given data, it may not be able to identify new fraud patterns. Deep learning methods such as Neural Networks can be employed for fraud detection to identify new patterns.**
- **The upcoming Generative AI method can also be researched for fraud detectio frauds are being attempted using this method.**



1

INTRODUCTION

In this chapter, we can learn about insurance fraud in the automobile sector and how machine learning helps in reducing fraud.

2

LITERATURE REVIEW

In this chapter, we can learn about what other research has been carried out for fraud detection and how our research is different from theirs.

3

OBJECTIVES

This thesis aims to identify an optimal machine learning model that can efficiently and affordably address insurance fraud in the automobile sector.

4

METHODOLOGY

We can learn how our research has been carried out in this section.

5

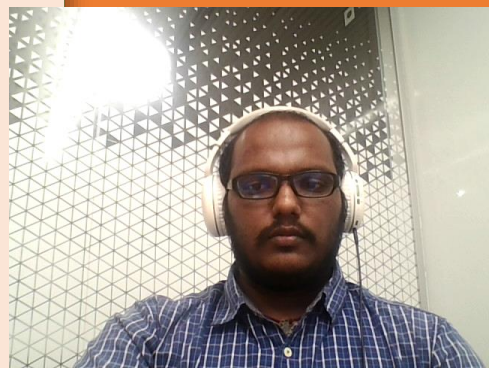
RESULTS & DISCUSSION

The result of our research is listed here.

6

CONCLUSION & FUTURE WORK

Conclusion from our research and future works to be looking for.




FRAUD