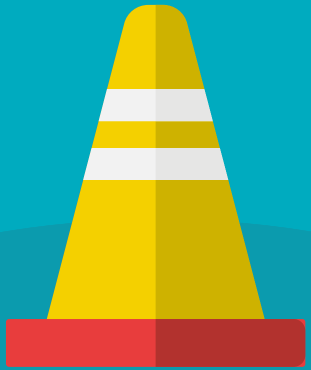
Novel Approach to deal with imbalance in Automobile Insurance Fraud Data

Abdus Saboor | Alizar Farhan | Khalifa Alshamsi

Table of Contents

01

Introduction



Automobile Insurance Fraud

Deceive the insurance company by providing false or misleading information, to receive financial benefits

- Inflated damage claims
- Staged accidents,
- Faking injuries, etc.

Impacts

- Increases the cost that insurance companies must cover
- Increases overall insurance premium for all customers

According to Forbes (2024) , insurance fraud costs the U.S. economy approximately $308 billion annually

Kilroy A. Insurance fraud statistics 2024. forbes.com Web site. https://www.forbes.com/advisor/insurance/fraud-statistics/. Updated 2024. Accessed Sep 23, 2024.

Challenges with Traditional Methods

- Time consuming and costly
- Inefficient use of resources
- Prone to human error
- Many genuine cases are scrutinized unnecessarily
- Many fraudulent claims go undetected

How are these issues being solved by researchers?

MACHINE LEARNING

Challenges in Research for applying ML to Insurance Data

Data Imbalance

- There are a greater number of non-fraud cases than the fraud cases
- This leads to the model overfitting to the non-fraud data

Availability of publicly available data

- Due to security concerns
- Leads to studies using synthetically generated data

RESEARCH QUESTION

> "How can we deal with the issue of data imbalance along with improving the performance of applied machine learning techniques?"

02

Literature Review

Use of Data Mining Techniques for Data Balancing and Fraud Detection in Automobile Insurance Claims [1]

Methods:

- Propose novel hybrid approach fuzzy C-means clustering + SMOTE to address imbalance issue
- Ensemble techniques of SVM, MLP and KNN with majority voting

Results:

- FCM: Under-sampled via removing outliers
- SMOTE: Generated synthetic samples to balance datasets
- Improved accuracy and recall across all classifiers with the balanced dataset

Limitation:

- High computational cost due to clustering (little samples are removed)
- Can introduce noise or synthetic data artifacts, leading to overfitting.

Detecting Fraudulent Insurance Claims Using Random Forests and Synthetic Minority Oversampling Technique [2]

Methods :

- Propose Random forest as classifier
- Tree-based ensemble learning algorithm
- Can handle high-dimensional tabular effectively
- Automatically identify important features
- Propose **SMOTE** to handle data imbalance

Results:

- Generated synthetic samples to balance dataset
- Outperformed SVM, DT and MLP

Limitation:

- Can introduce noise or synthetic data artifacts, leading to overfitting

Using a data mining approach to detect automobile insurance fraud [3]

Methods:

- Used RF as classifier
- Performance comparison between SMOTE and ROSE to address imbalance issue

Results:

- **SMOTE**: Generated synthetic samples to balance datasets
- **ROSE**: Generated more diverse synthetic samples.
- SMOTE outperformed ROSE for overall performance comparison

Limitation:

- SMOTE: Can introduce noise or synthetic data artifacts, leading to overfitting.
- ROSE: is computationally intensive and may add noise to the dataset.

Enhancing Auto Insurance Fraud Detection Using Convolutional Neural Networks [4]

Methods:

- Proposed 1-Dimension Convolution Neural Network for its excellent spatial feature extraction
- Used Tabular Generative Adversarial Networks (**CTGAN**) to address imbalance issue
- Performed comparison to SMOTE

Results:

- Outperformed SMOTE significantly for overall performance metrics (precision, recall and f1)
- Generated realistic synthetic data by mimicking data distribution
- Effectively handles categorical data

Limitation:

- More complex and computationally intensive than GAN.

Work	Balance technique	Advantages/Strengths	Disadvantages/Weakness
[1]	FCM + SMOTE	Under sampling via removing outliers + Generate synthetic samples to balance datasets	High computational cost due to clustering (little samples are removed) Can introduce noise or synthetic data artifacts, leading to overfitting.
[3],[1]	Randomly oversampling	Easy to construct	Can lead to overfitting by duplicating existing data
[3],[1]	Randomly undersampling	Easy to construct, reduces computational cost	Can lead to loss of valuable information from the majority class
[1],[2],[3],[4]	SMOTE	Generate synthetic samples to balance datasets, widely used.	Can introduce noise or synthetic data artifacts, leading to overfitting.
[3]	ROSE	Generates more diverse synthetic samples.	Computationally intensive and may add noise to the dataset.
[4]	ADASYN	Generate synthetic samples for harder-to-classify instances.	May skew data distribution, leading to overfitting in some cases.
[4]	GAN	Generates realistic synthetic data by mimicking data distribution	Requires extensive training and is computationally expensive.
[4]	CTGAN	Generates realistic synthetic data by mimicking data distribution, effectively handles categorical data,	More complex and computationally intensive than GAN.

DATA LEAKAGE

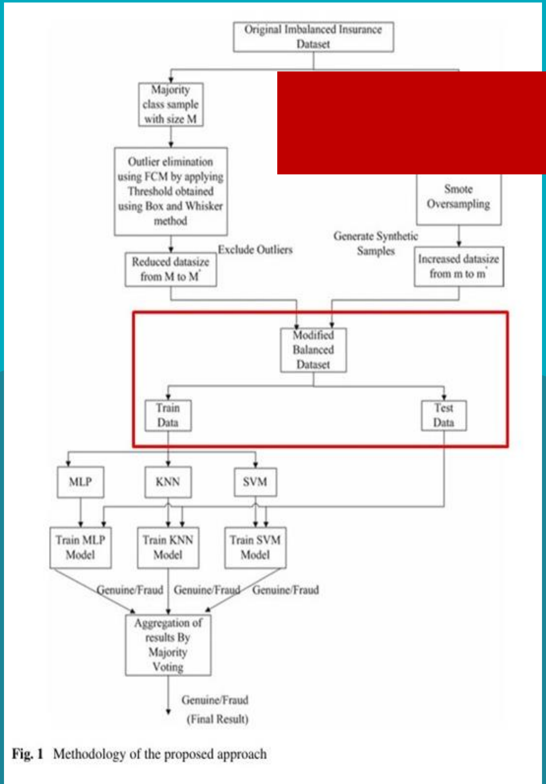
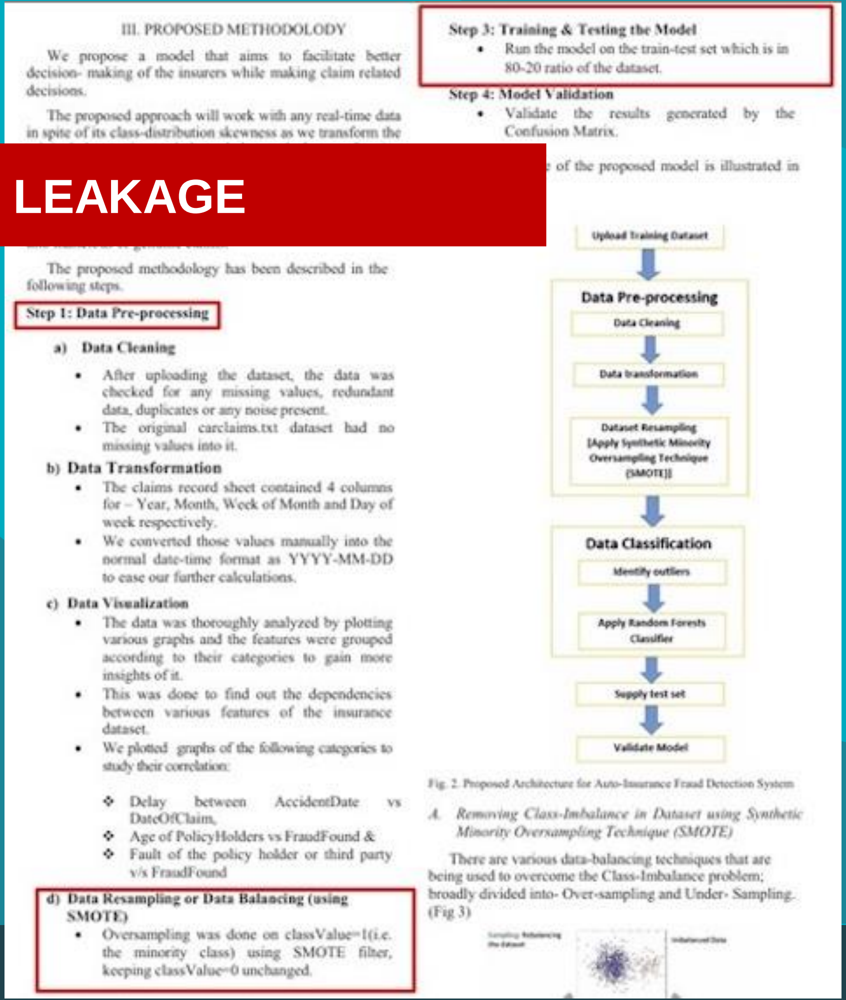


Fig. 1 Methodology of the proposed approach



III. PROPOSED METHODOLODY

We propose a model that aims to facilitate better decision- making of the insurers while making claim related decisions.

The proposed approach will work with any real-time data in spite of its class-distribution skewness as we transform the

...

The proposed methodology has been described in the following steps.

Step 1: Data Pre-processing

a) Data Cleaning

- After uploading the dataset, the data was checked for any missing values, redundant data, duplicates or any noise present.
- The original carclaims.txt dataset had no missing values into it.

b) Data Transformation

- The claims record sheet contained 4 columns for – Year, Month, Week of Month and Day of week respectively.
- We converted those values manually into the normal date-time format as YYYY-MM-DD to ease our further calculations.

c) Data Visualization

- The data was thoroughly analyzed by plotting various graphs and the features were grouped according to their categories to gain more insights of it.
- This was done to find out the dependencies between various features of the insurance dataset.
- We plotted graphs of the following categories to study their correlation:
  - Delay between AccidentDate vs DateOfClaim,
  - Age of PolicyHolders vs FraudFound &
  - Fault of the policy holder or third party v/s FraudFound

d) Data Resampling or Data Balancing (using SMOTE)

- Oversampling was done on classValue=1(i.e. the minority class) using SMOTE filter, keeping classValue=0 unchanged.

Step 3: Training & Testing the Model

- Run the model on the train-test set which is in 80-20 ratio of the dataset.

Step 4: Model Validation

- Validate the results generated by the Confusion Matrix.

... of the proposed model is illustrated in

Upload Training Dataset
Data Pre-processing
Data Cleaning
Data transformation
Dataset Resampling [Apply Synthetic Minority Oversampling Technique (SMOTE)]
Data Classification
Identify outliers
Apply Random Forests Classifier
Supply test set
Validate Model

Fig. 2. Proposed Architecture for Auto-Insurance Fraud Detection System

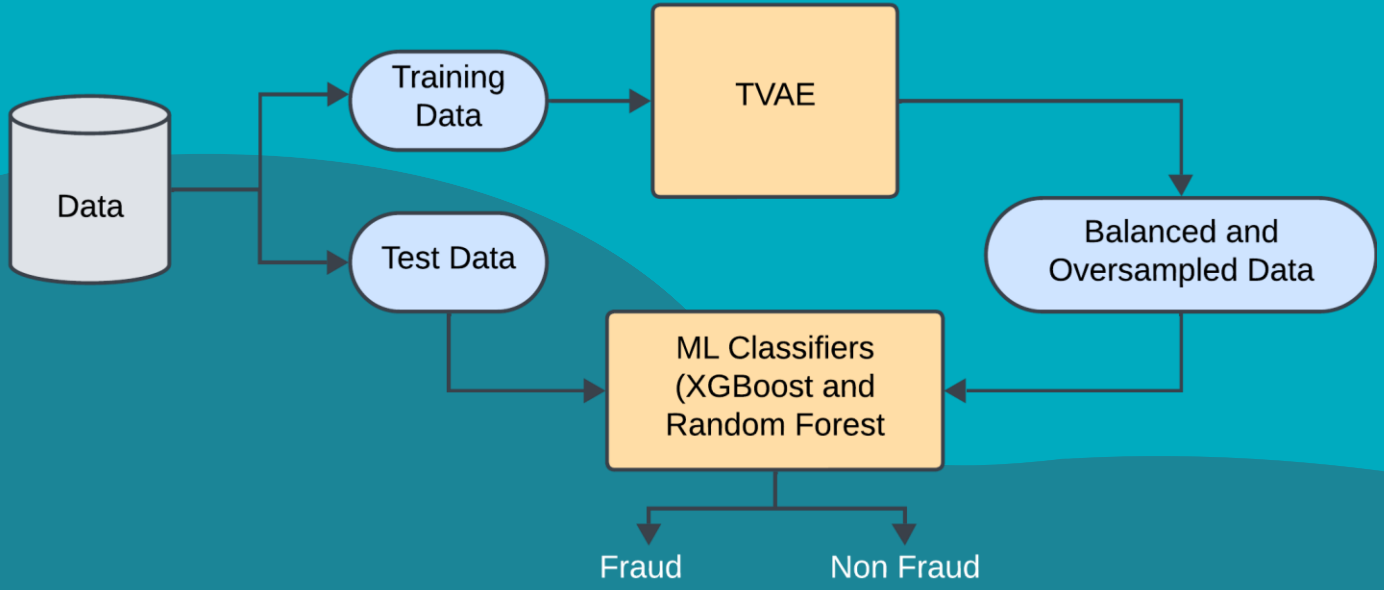
A. Removing Class-Imbalance in Dataset using Synthetic Minority Oversampling Technique (SMOTE)

There are various data-balancing techniques that are being used to overcome the Class-Imbalance problem; broadly divided into- Over-sampling and Under- Sampling. (Fig 3)

03

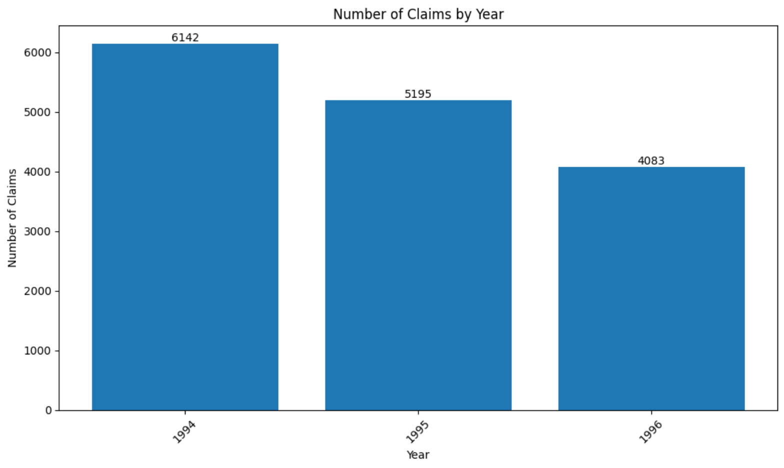
Methodology

Proposed Approach



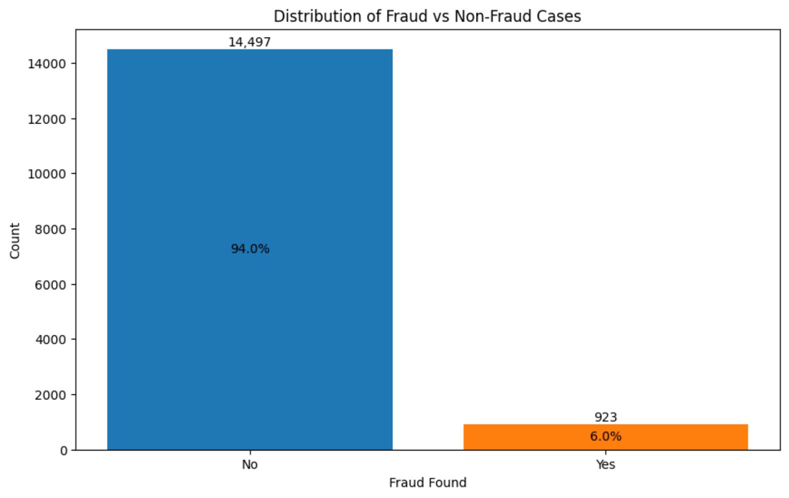
Dataset

- Automobile insurance fraud details dataset named “carclaims.txt” that is used by many research papers
- Originally made available as part of “Angoss Knowledge Seeker” product
- The original copy is lo longer available for download
- Publicly available copy is accessible on GitHub (https://github.com/Rashmi-77/Vehicle-Insurance-Fraud-Detection)
- It contains claim records from 1994 to 1996



Dataset

- High imbalance of fraud and non-fraud cases
- The size of dataset is very small
- Out of 15,420 claims, only 923 are fraudulent



Preprocessing

- The primary key column ‘PolicyNumber’ is dropped as it add no value to the prediction
- The feature ‘DayOfWeekClaimed’ contained one missing record, so it was dropped for convenience
- One-Hot Encoded the nominal features like – Make, Year, Month, Day of Week, Accident Area, etc.
- Label encoded the ordinal features like - Past number of claims, Number of cars, Age of Vehicle (These features may sound like numerical, but in the dataset, they are binned to create categorical features)
- No normalization was performed as only Ensemble Decision trees were used for the classification task and the implementation of TVAE used does its own data preprocessing.

Tabular VAE (TVAE)

L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling tabular data using conditional GAN,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 7335–7345.

- For synthetic data generation by employing Variation Autoencoder for tabular data
- Continuous columns are normalized, and discrete columns are one-hot encoded
- Handle mixed data types (continuous and discrete), non-Gaussian distributions, and multimodal distributions
- Uses Gaussian distributions for continuous columns and PMF for discrete columns.

TVAE – Synthetic Data Vault

Tunable parameters:

- Latent dimension size
- Encoder layers and width
- Decoder layers and width
- Epochs
- Etc.

Allows synthetic data generation with conditions
Provides tools for calculation and visualizing the quality of data

Ensemble Classification Models

XGBoost

Speed, efficient and supports regularization

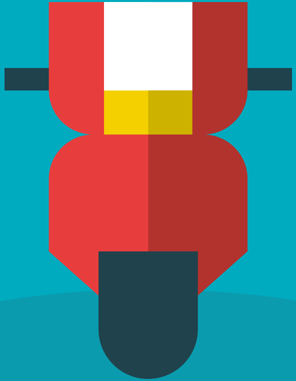
Random Forest

Simple & handles high dimensional data

Many papers that performed oversampling before the split also applied one or both of the techniques. And these models are proven to work well for the selected dataset

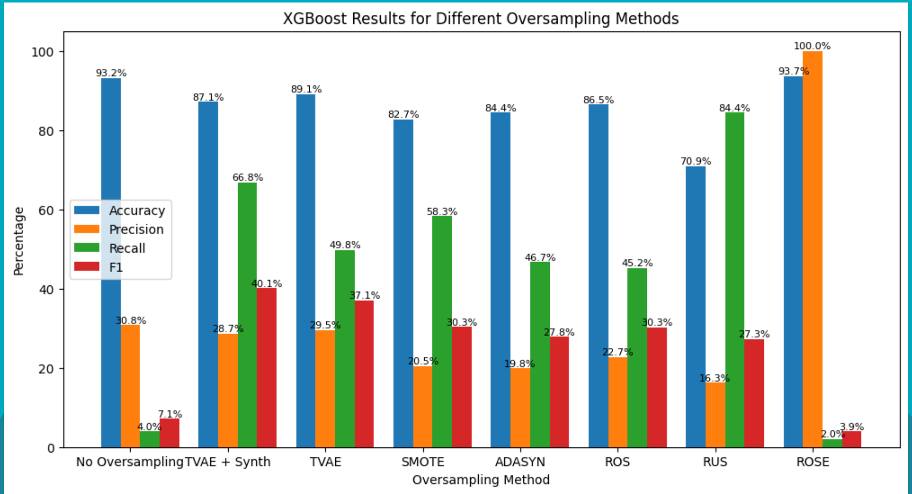
04

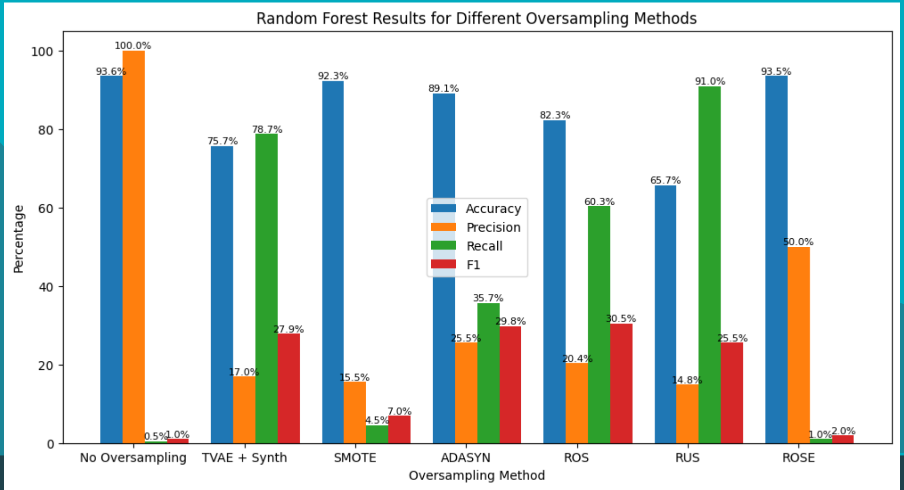
Result & Analysis



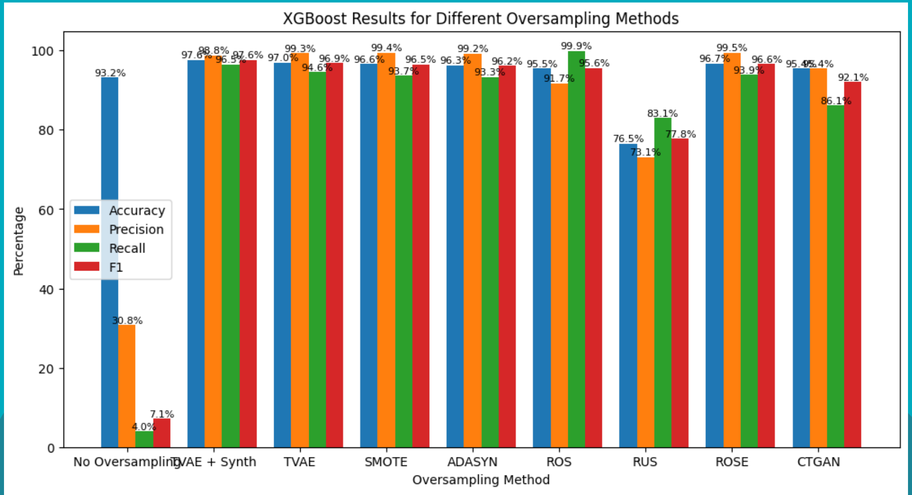
Work	Dataset	Methods	Balance Techniques	Order(Split/Balance)	Result(Accuracy, Precision, Recall & F1-score)
[6]	carclaims.txt	ELM	Not Mentioned	Balance the data then split	--,--,74.9%,--
[2]	carclaims.txt	Random Forest	SMOTE	Balance the data then split	94.3%,98.6%,45.1%,61.9%
[3]	carclaims.txt	Random Forest	SMOTE ROSE	Split Data then balance	64.3%,---,93.07%,23.8% 61.34%,14.1%,95.24%,22.79%
[1]	carclaims.txt	Ensemble of SVM, MLP & KNN	FCM (Under Sampling) SMOTE	Balance the data then split	81.2%,----,94.2%,----
[5]	carclaims.txt	4-layer 1D-CNN	SMOTE CTGAN	Split Data then balance	81.3%,13.6%,39.8%,20.3% 79.3%,16.7%,61.5%,26.2%
Proposed	carclaims.txt	Random Forest & XGboost	TVAE	Split Data then balance and oversample	87.13%,28.66%,66.83%,40.12%

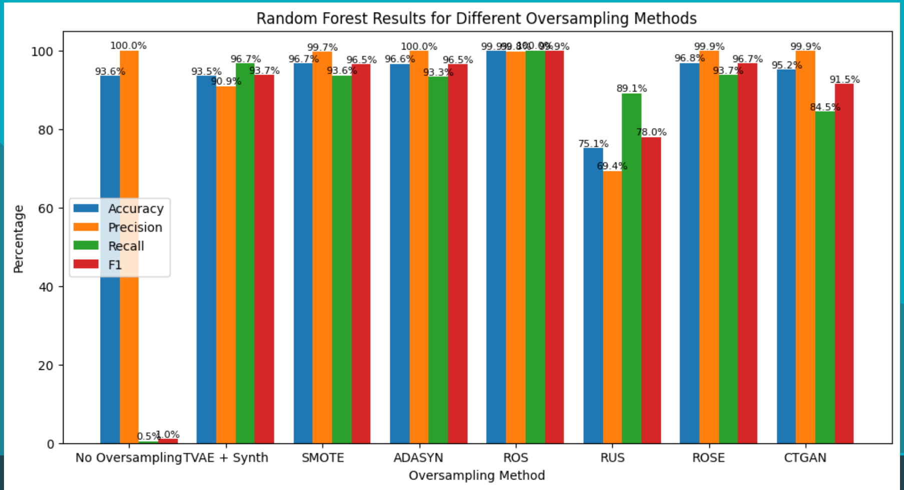
Oversampling after train-test split



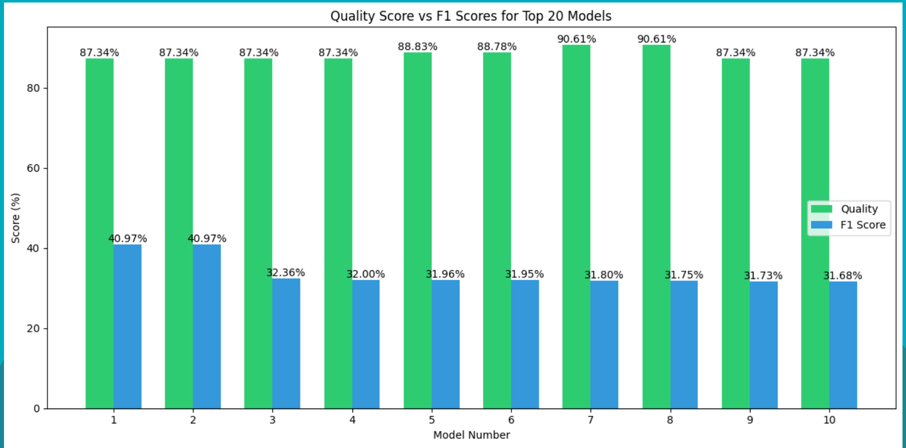


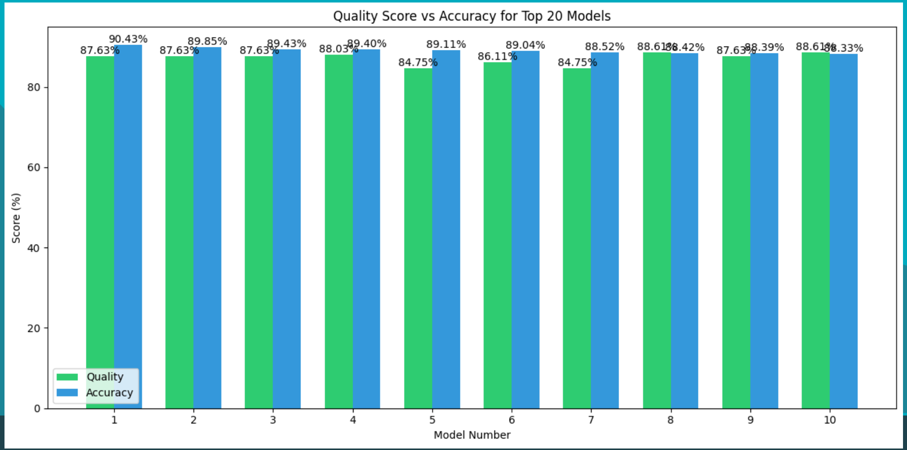
Oversampling before train-test split





Quality of Synthetic Data



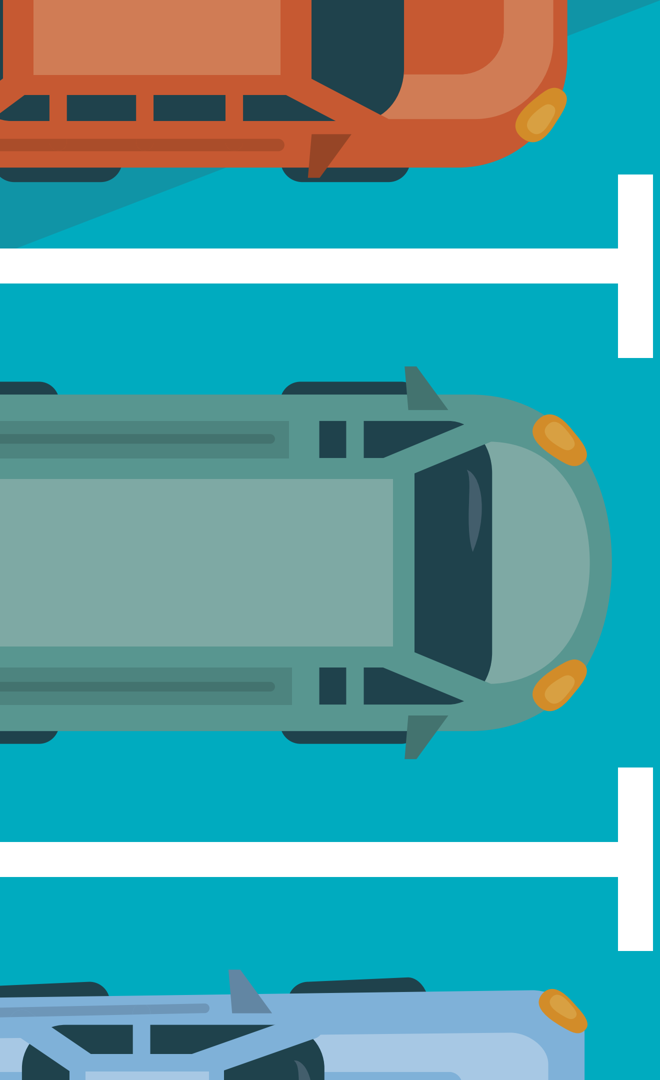


05

Challenges & Limitations

Data Availability

- Limited publicly available data
- Old data
- Small datasets



Model and resources

- Large number of hyperparameters to tune for CTGAN and TVAE
- Computational Resources and Power to get better computational result
- Time constraints

Limitations

Introduces Noise

It cannot be used in application where its preferred not to have noise in the data

Hyperparameter Tuning

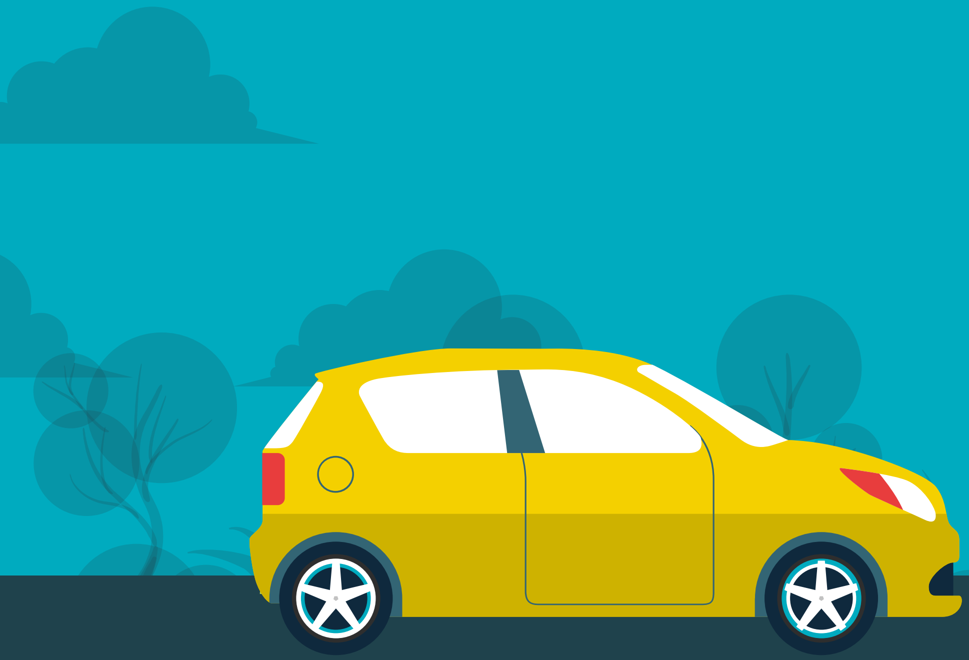
Models like TVAE and CTGAN need hyperparameter tuning to produce better results

Interpretability

Generative models like GAN,TVAE and advanced classifiers like XGBoost are known as Black box as the investor cannot interpret or trust the results generated

06

Conclusion & Future Work

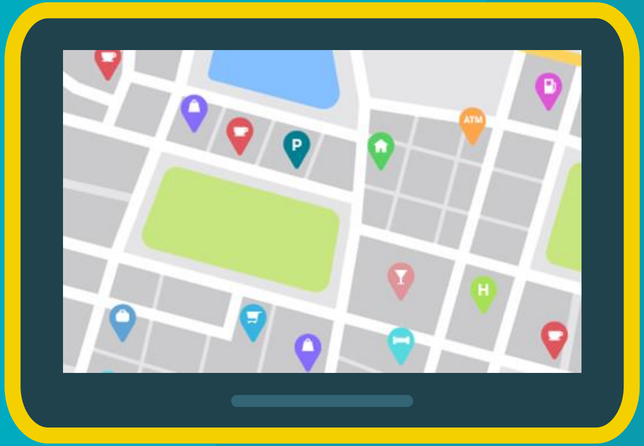


Conclusion

- Data balancing using TVAE achieves better F1 score when the balancing is done before and after the split
- It outperforms other techniques when done after the split
- It can be used to generate more synthetic data in application with limited data
- These technique can be applied to other industries with imbalance, specially for insurance fraud.

Future Direction

- Explore advanced hyperparameter optimization for XGBoost, Random Forest, TVAE and CTGAN
- Utilize modern clustering techniques for nuanced fraud datasets.
- Integrate ensemble learning methods to enhance fraud detection.
- Use this study on different fraudulent data sets and detect the pattern



THE END

THANKS!

Do you have any questions?

References

1. S. Padhi and S. Panigrahi. Use of data mining techniques for data balancing and fraud detection in automobile insurance claims. In Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS 2019), volume 1044, pages 259–268, Singapore, 2020. Springer. doi: 10.1007/978-981-15-1084-7 22. URL http://link.springer.com/10.1007/978-981-15-1084-7_22.
2. S. Harjai, S. K. Khatri, and G. Singh. Detecting fraudulent insurance claims using random forests and synthetic minority oversampling technique. In 2019 4th International Conference on Information Systems and Computer Networks (ISCON), pages 123–128, Mathura, India, 2019. IEEE. doi: 10.1109/ISCON47742.2019.9036162
3. M. Salmi and D. Atif. Using a data mining approach to detect automobile insurance fraud. In Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR), pages 55–66, Cham, Switzerland, 2022. Springer. doi: 10.1007/ 978-3-030-96302-6 5. URL https://doi.org/10.1007/978-3-030-96302-6_5.
4. R. Wongpanti and S. Vittayakorn. Enhancing auto insurance fraud detection using convolutional neural networks. In 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), pages 294–301, Phuket, Thailand, 2024. IEEE. doi: 10.1109/JCSSE61278.2024.10613702.
5. D. K. Patel and S. Subudhi. Application of extreme learning machine in detecting auto insurance fraud. In 2019 International Conference on Applied Machine Learning (ICAML), pages 78–81, Bhubaneswar, India, 2019. IEEE. doi: 10.1109/ICAML48257.2019.00023.