Application of Machine Learning Techniques in Insurance Fraud Detection

Khutso Mphelo

BScHons in Applied Mathematics
University of Johannesburg

14 November 2024

UNIVERSITY
OF
JOHANNESBURG

Outline

- Objective: Enhance fraud detection in insurance claims using machine learning.
- Dataset: Vehicle claims fraud data (Bansal, 2021).

Figure: Image generated by ChatGPT

Literature Review

- Traditional methods: Audits and expert reviews.
- Machine learning in fraud detection: Logistic Regression, Random Forest, XGBoost.

Model Class	Methodology	Description
Clustering	K-means	Groups data points by similarity
Classification	Logistic Regression, Random Forest	Categorizes data based on labeled data
Outlier Detection	Isolation Forest	Identifies anomalous data points

Table: Overview of ML Models in Fraud Detection

Methodology

- **Data preprocessing:** Cleaning, encoding, standardization.
- **Class imbalance handling:** SMOTE and undersampling.
- **Models used:** Random Forest, Logistic Regression, XGBoost, and KNN.

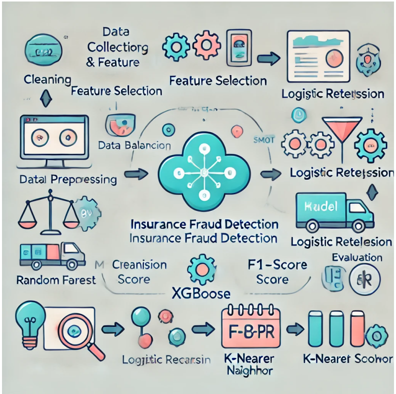

Figure: Image generated by ChatGPT

Data Analysis Insights

- Before modeling, an exploratory data analysis (EDA) was conducted to understand key patterns and trends in the data.
- Insights from EDA guided the feature selection and informed the class balancing approach.

Class Imbalance Visualization

- The dataset shows a significant class imbalance, with a smaller proportion of fraudulent claims.
- To address this, techniques like SMOTE were applied to balance the classes.

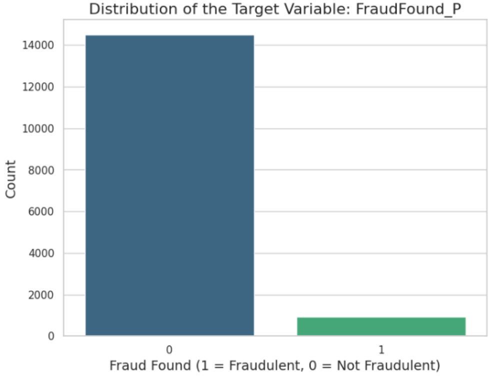

Figure: Class Imbalance

Fraud Rate by Month Claimed

- A decrease in fraud rates is observed prior to November, possibly due to effective fraud prevention measures or reduced claims as the year ends.
- December shows an increase in fraud, potentially linked to holiday season factors.

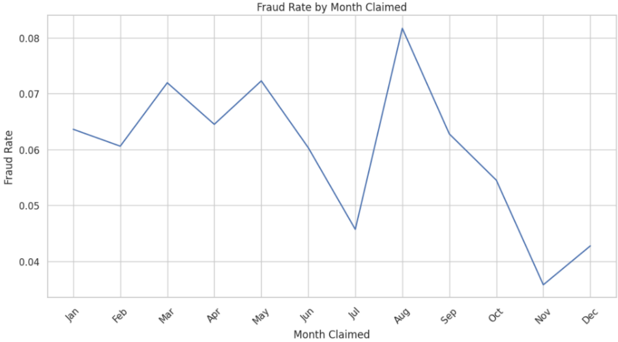

Figure: Fraud rate by month claimed

Correlation of Age of Policy Holder with Fraud Cases

- The most common fraudulent policy holders are middle aged, specifically in the 31-35 and 36-40 age group
- Likely due to more financial pressures, potentially leading to a greater number of bills and financial responsibilities.
- Understanding this relationship helps identify high-risk age groups in insurance claims.

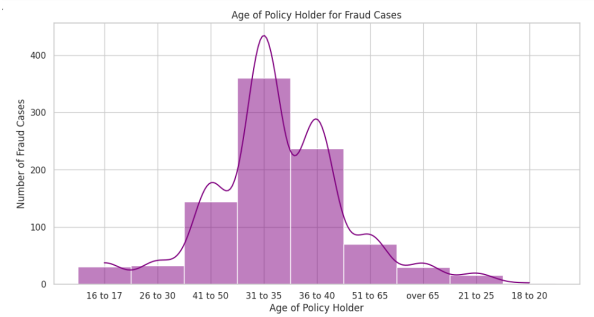

Figure: Correlation between Age of Policy Holder and Fraud Cases

Correlation of Age of Policy Holder with Fraud Cases

- Married individuals: Highest fraud cases (639), likely due to financial pressures.
- Single individuals: Moderate fraud cases (278).

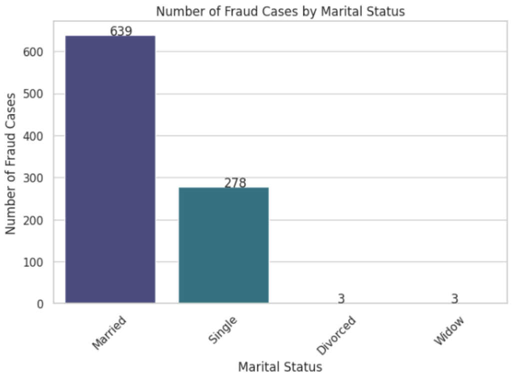

Figure: Number of Fraud Cases by Marital Status

Percentage of Fraud Cases by Accident Area

- Urban areas account for approximately 85.6% of fraud cases.
- Rural areas account only 14.4% of fraud cases.

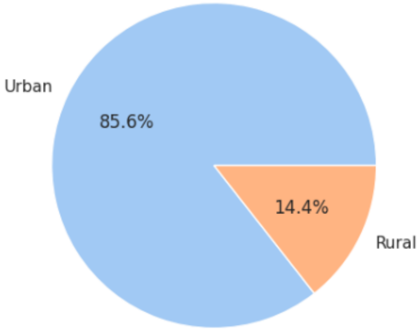

Figure: Fraud Cases by Accident Area

Summary of Data Insights

- **Marital status & Age:** Higher fraud rates among married and middle-aged individuals.
- **Seasonal peaks:** Fraud cases peak in August and December, potentially due to financial or seasonal factors.
- **Geographic patterns:** Urban areas sees more fraud cases (85.6%), while rural areas have fewer.
- **Class imbalance:** Fraudulent claims are rare, requiring techniques like SMOTE for model balance.

XGBoost Model Performance - Precision and Recall

- Best performing model is XGboost with an accuracy of 76.5%.

Accuracy: 0.7659

Classification Report:

	precision	recall	f1-score	support
0	0.99	0.76	0.86	2885
1	0.20	0.89	0.33	199
accuracy			0.77	3084
macro avg	0.60	0.83	0.59	3084
weighted avg	0.94	0.77	0.82	3084

Figure: Precision and Recall for XGboost model

Confusion Matrix for XGBoost Model

- XGBoost model performance with SMOTE application.

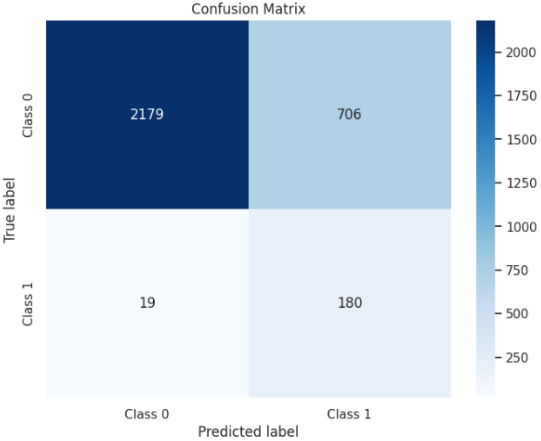

Figure: Confusion Matrix for XGBoost

Conclusion

- **Enhanced Detection:** Machine learning models enhance fraud detection accuracy.
- **Class Imbalance Challenge:** SMOTE helps address imbalance, boosting model effectiveness.
- **Key Risk Insights:** Higher fraud rates are seen among married, urban, and middle-aged groups, highlighting areas for targeted prevention.
- **Future Work:** Further explore ensemble models and advanced sampling techniques to refine detection capabilities.
- **Business Impact:** Effective fraud detection can significantly reduce losses and improve client trust in the insurance process.

References

- Viaene, S., Dedene, G. (2004). "Insurance Fraud: Issues and Challenges." The Geneva Papers on Risk and Insurance, 29(2), 313–333.
- Derrig, R. A. (2002). "Insurance Fraud." The Journal of Risk and Insurance, 69(3), 271–287.
- Bansal, S. (2021). "Vehicle Claim Fraud Detection Dataset." Kaggle. Available at: `https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection/data`

Thank you!

Questions?