Auto Insurance Fraud Detection Report

1 Introduction

Insurance fraud poses a major challenge for the financial sector, resulting in significant monetary losses. The objective of this project is to leverage machine learning models to detect fraudulent auto insurance claims using a publicly available dataset. The effectiveness of various algorithms is evaluated based on performance metrics, and the most suitable model is selected.

2 Dataset Overview

- **Source:** Figshare
- **Shape:** 15,420 rows and 61 columns
- **Target Variable:** `FraudFound` (Yes / No)

The dataset includes both categorical and numerical features related to customer profiles, claim details, and policy information.

3 Data Preprocessing

3.1 Handling Missing Values

- A single row contained placeholder values (0), which was identified and corrected.
- Categorical missing values such as `DayOfWeekClaimed` and `MonthClaimed` were imputed using the mode.
- Missing numerical values, particularly `Age`, were filled with the mean.

3.2 Feature Removal

`PolicyNumber` showed high correlation with `Year`, but did not contribute meaningfully to prediction. It was removed to prevent noise and potential leakage.

4 Encoding and Scaling

4.1 Encoding

- Binary categorical variables were label encoded.
- Multi-class categorical variables were one-hot encoded.

4.2 Scaling

`StandardScaler` was used to scale features for logistic regression. Tree-based models were used without scaling.

5 Modeling and Evaluation

The following models were trained and evaluated using standard classification metrics:

- Logistic Regression
- Decision Tree
- Random Forest
- LightGBM
- XGBoost

5.1 Performance Before Hyperparameter Tuning

Table 1: Model Performance Before Tuning

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
Logistic Regression	0.9358	0.3333	0.0051	0.0100	0.8268
Decision Tree	0.9099	0.3081	0.3299	0.3186	0.6397
Random Forest	0.9361	0.0000	0.0000	0.0000	0.8698
LightGBM	0.9471	0.9250	0.1878	0.3122	0.9545
XGBoost	0.9543	0.7800	0.3959	0.5253	0.9736

5.2 Performance After Hyperparameter Tuning

Table 2: Model Performance After Tuning

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
LightGBM	0.9540	0.7895	0.3807	0.5137	0.9794
XGBoost	0.9484	0.8167	0.2487	0.3813	0.9692
Random Forest	0.9361	0.0000	0.0000	0.0000	0.8630
Logistic Regression	0.9358	0.0000	0.0000	0.0000	0.8259

6 Visual Analysis

- **ROC Curve Comparison:** Visual comparison of true and false positive rates for each model.
- **Feature Importance:** Tree-based models such as LightGBM and XGBoost were used to extract and plot important features.

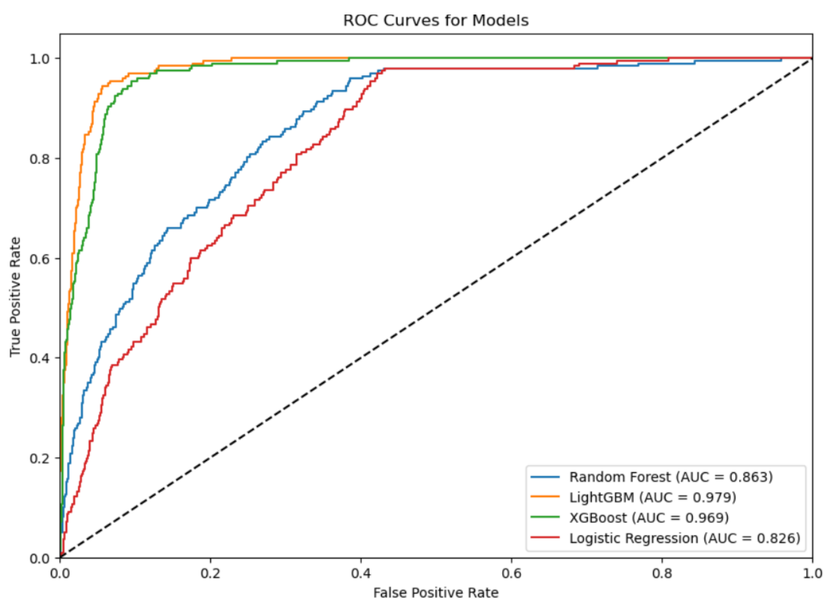

Figure 1: ROC Curve Comparison of All Models

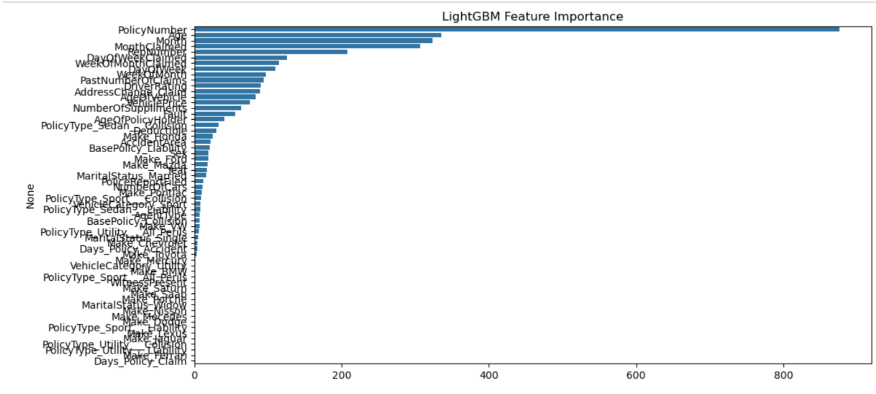

Figure 2: Feature Importance from LightGBM

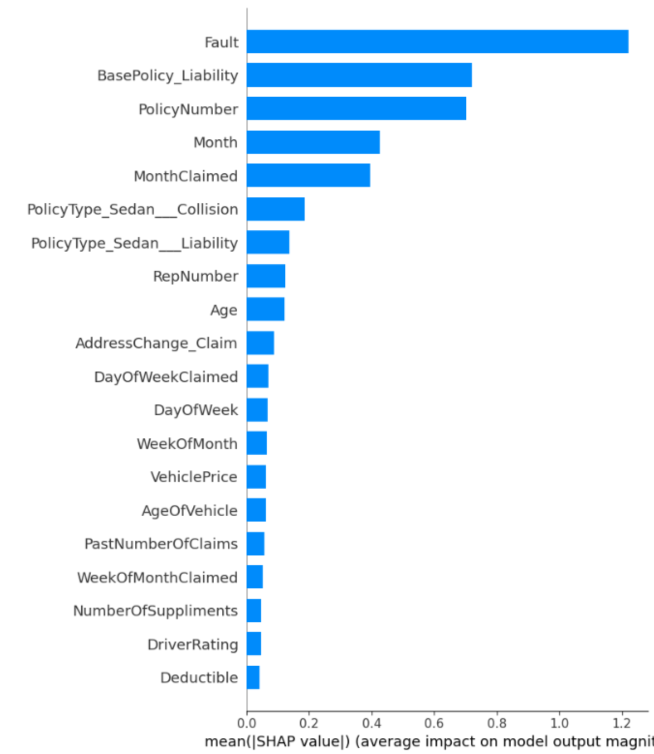

Figure 3: SHAP Summary Plot Showing Feature Importance and Impact on Model Predictions

7 Conclusion

This project developed a robust pipeline for detecting fraudulent insurance claims using multiple machine learning models. LightGBM was found to be the most effective, with the highest ROC AUC score (0.9794) and a balanced trade-off between precision and recall. Proper handling of missing values, encoding, and model tuning played a critical role in achieving high performance.