Fraud Claim Detection Report

By Garima Chhabra, Harikrishnan R, Divyansh Singh

This analysis was conducted to help **Global Insure** to enhance its ability to detect fraudulent insurance claims. By leveraging historical claim, we aimed to identify key indicators that differentiate fraudulent claims from genuine ones and also develop predictive model that can proactively access likelihood of fraud and finally reducing the cost.

Key Findings & Approach

1. Data Cleaning

- Many of the variables have a level "?" which needs to be handled, was first replaced by null and then missing value imputation was done, mode for categorical columns
- Columns with very high unique values were dropped, as they were either identifiers or they show no meaningful patterns hence less predictive power
- Rows showing illogical values were dropped e.g. Umbrella limit with negative values
- Data type for incident date was converted to datetime
- C_39, Column was completely empty, so dropped

2. Train validation split

- Data splitting was done, with train size 0.7 and stratify y, for same distribution of target variable in train and Test data set and to ensure robust model validation.

3. Exploratory Data Analysis (EDA)

EDA revealed that continuous numerical column were slightly skewed, we cap the outliers to 0.01(lower limit) and .99(upper limit). The missing values were imputed with median to preserve the original distribution. Many categorical variables contained irrelevant or redundant elements. These were refined to ensure meaningful insights during modelling.

3. Feature Engineering

- Data imbalance was handled using **Random Over Sampler**
- **Binary Mapping**
- New levels were created for some features, some levels were merged and also a new column(incident_state_grouped with levels such as High, Medium, low risk) was created to reduce dimensionality
- **Dummy Variables:** Categorical variables were transformed using one-hot encoding and Frequency encoding
- **Scaling:** Numeric values were standardized using **MinMaxScaler**

4. Model Building & Optimization

- **Feature Selection:** RFECV was applied to identify the most relevant features. **Logistic Regression Model:** using 13 selected features and stats model. Further **Variance Inflation**

Factor (VIF < 5) and **p-values (< 0.05)** were checked to ensure features were statistically significant.

- **Random Forest Model:** Built using features with importance score = 0.015. Cross validation was used to check if model is overfitting. Hyper parameter tuning was done. Model with best Recall and F1 Score was used to build random forest model

6. Model Evaluation

- **Logistic Regression - Precision-Recall Trade-off:** Adjusting the cut-off to **0.5** optimized **precision (87%)** and **recall (90%)**, balancing false positives and false negatives.
- **High AUC (0.92)** indicates the model is highly effective at distinguishing between potential and non-converting leads.
- **Confusion Matrix & ROC Curve:** The **optimal cut-off threshold (0.5)** was identified, leading to an accuracy, sensitivity, and specificity of approximately **90%** on the training data.
- **Test Set Prediction:** The model achieved 85**% accuracy, sensitivity, and specificity** on unseen test data.

6. Model Evaluation

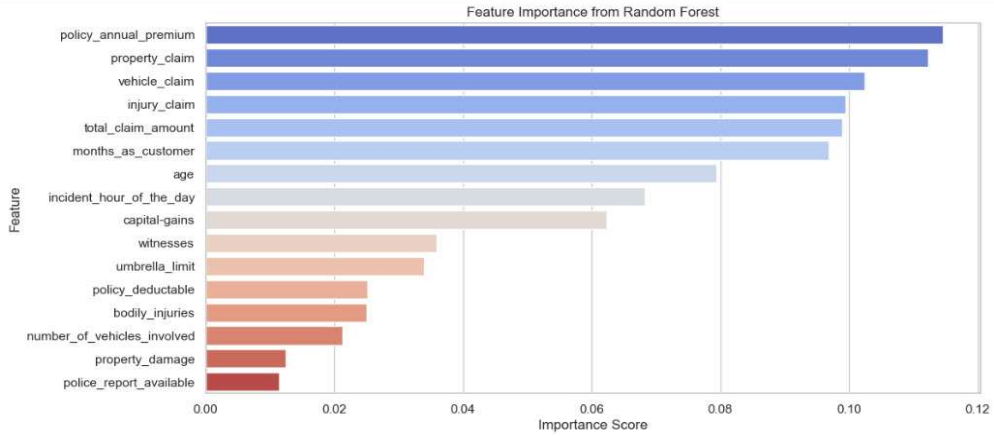
- **Random Forest - Precision-Recall Trade-off:** Adjusting the cut-off to **0.5** optimized **precision (87%)** and **recall (87%)**, balancing false positives and false negatives.
- **High AUC (0.93)** indicates the model is highly effective at distinguishing between potential and non-converting leads.
- **Confusion Matrix & ROC Curve:** The **optimal cut-off threshold (0.52)** was identified, leading to an accuracy, 77% sensitivity, and specificity of approximately **90%** on the training data.
- **Test Set Prediction:** The model achieved **83% accuracy, 77%sensitivity, and 85%specificity** on unseen test data.

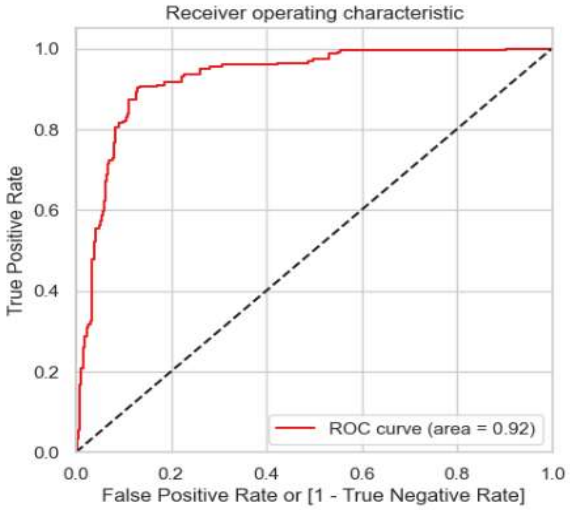
Top Recommendations for Global Insure

- Enhanced Scrutiny Based on Claim Severity - Both Minor Damage and Total Loss incidents are highly associated with fraud.
- Time-Based Risk Monitoring - Certain hours (e.g., 11 AM, 2 PM, 6 PM) had higher fraud rates.
- Behavioral Risk Profiling - Hobbies such as chess and cross-fit showed a significant correlation with fraud.
- Monetary Threshold-Based Audits - Features like vehicle_claim, property_claim, and total_claim_amount were top predictors
- Policy-Based Risk Scoring - Variables like umbrella_limit and policy_annual_premium contribute to risk prediction.

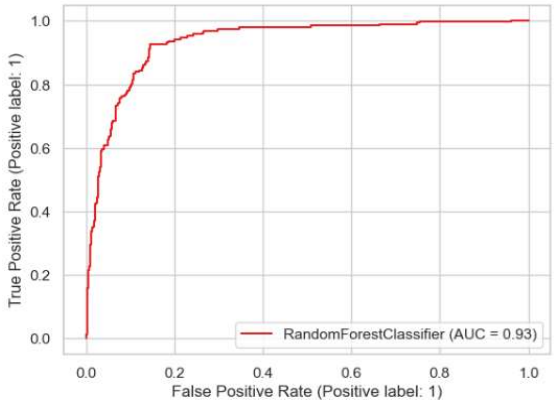
By implementing these **data-driven recommendations**, Global Insure can improve Fraudulent Insurance claims detection

Feature importance detection using Random forest classifier





ROC curve for Logistic Regression shows area under curve is 0.92



ROC curve for Random Forest shows area under curve is 0.93