Fraudulent Claim Detection Report :

1. Problem Statement :

Global Insurance, a major insurance provider, processes a large volume of claims every year. A significant portion of these claims are found to be fraudulent, resulting in financial losses and inefficient resource utilization. The current fraud detection process is primarily manual and reactive, leading to late identification of fraudulent activities. The objective is to leverage historical claim and customer data to build a predictive model that can proactively flag potentially fraudulent claims early in the process.

---

2. Business Objective :

Global Insurance wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amounts, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved.

To develop a machine learning-based system that classifies insurance claims as **fraudulent** or **legitimate**, based on features like:

- Claim amount
- Claim type
- Customer demographics
- Incident details

Early identification of fraudulent claims will:

- Minimize financial losses
- Speed up legitimate claim approvals
- Optimize the overall fraud detection pipeline

---

3. Key Questions Addressed

- How can we analyse historical claim data to detect fraud patterns?
- Which features most strongly indicate fraudulent behavior?
- Can we accurately predict fraud in new incoming claims?
- What insights can be used to enhance fraud detection strategies?

---

4. Methodology :

The approach followed is structured into distinct phases:

1. Data Preparation :

- Imported the dataset and relevant libraries (Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn, Statsmodels).
- Suppressed warnings and examined the shape, structure, and types of the dataset.

2. Data Cleaning :

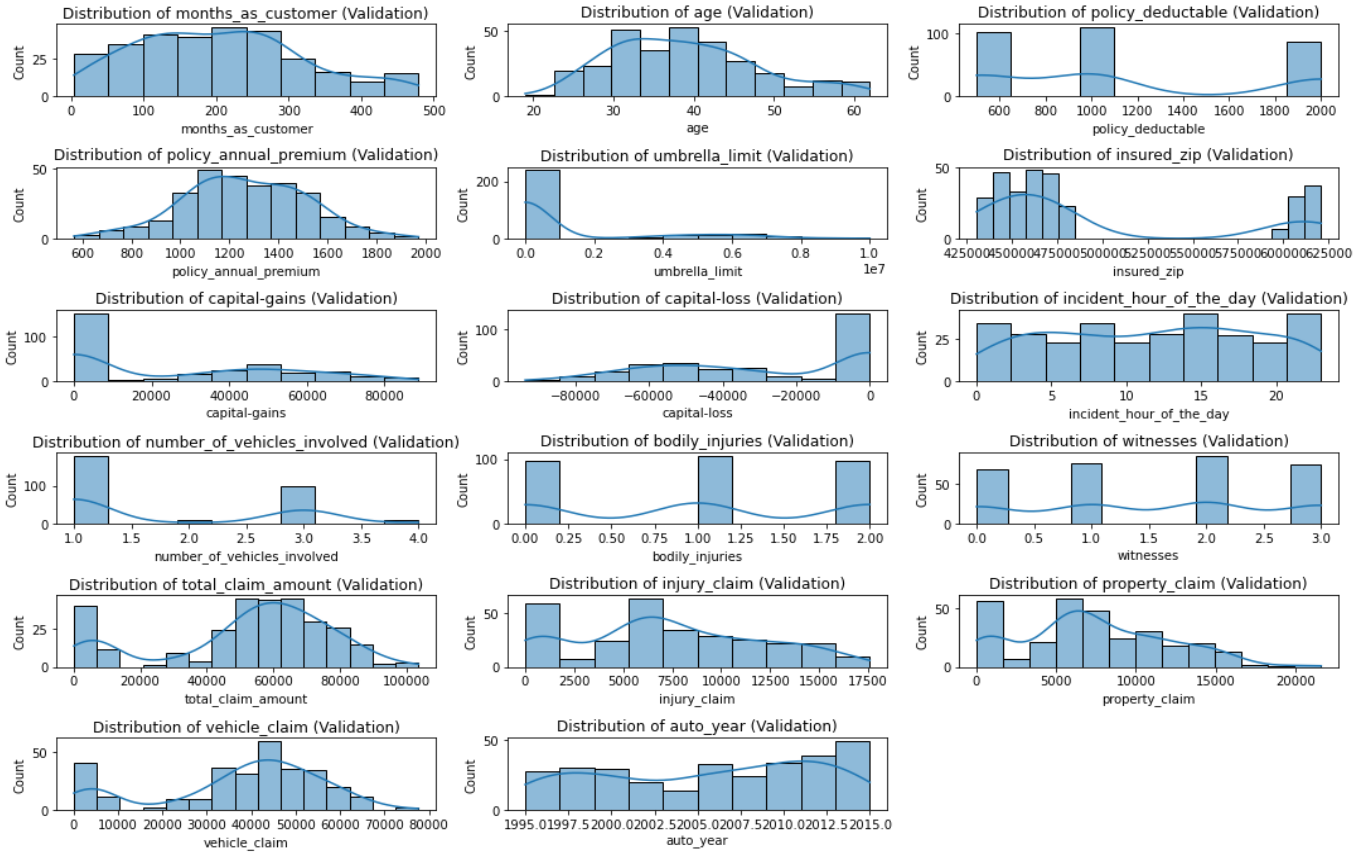
- Identified and removed missing/null values and illogical entries (e.g., negative claim amounts).
- Dropped redundant columns like identifiers and completely empty fields.
- Standardized data types across features.

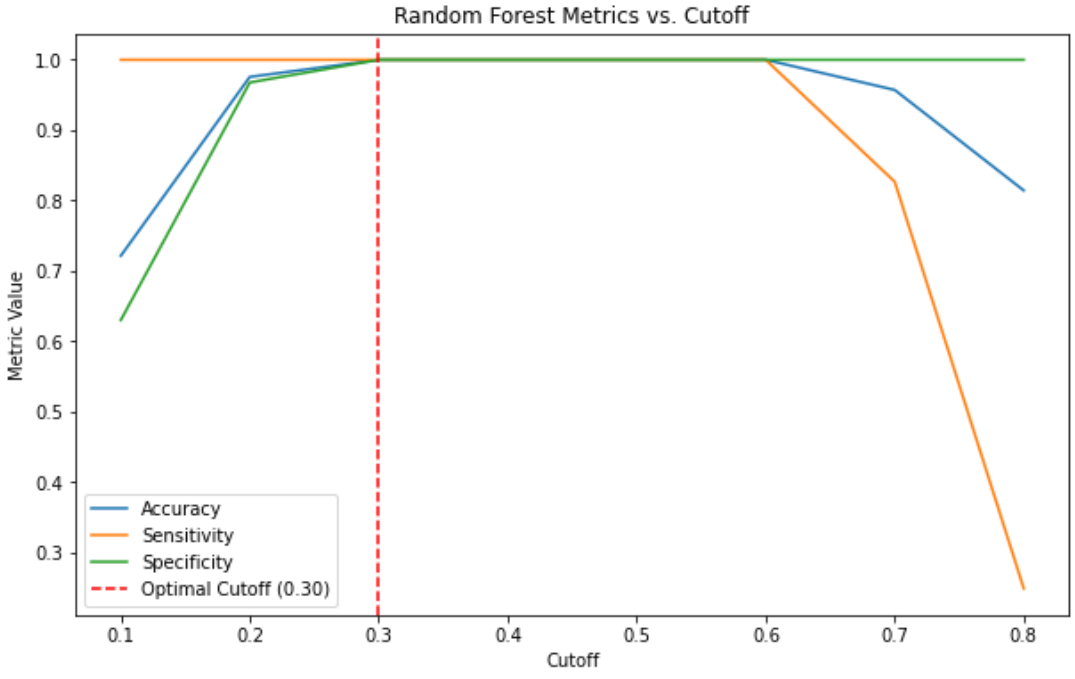
3. Data Splitting :

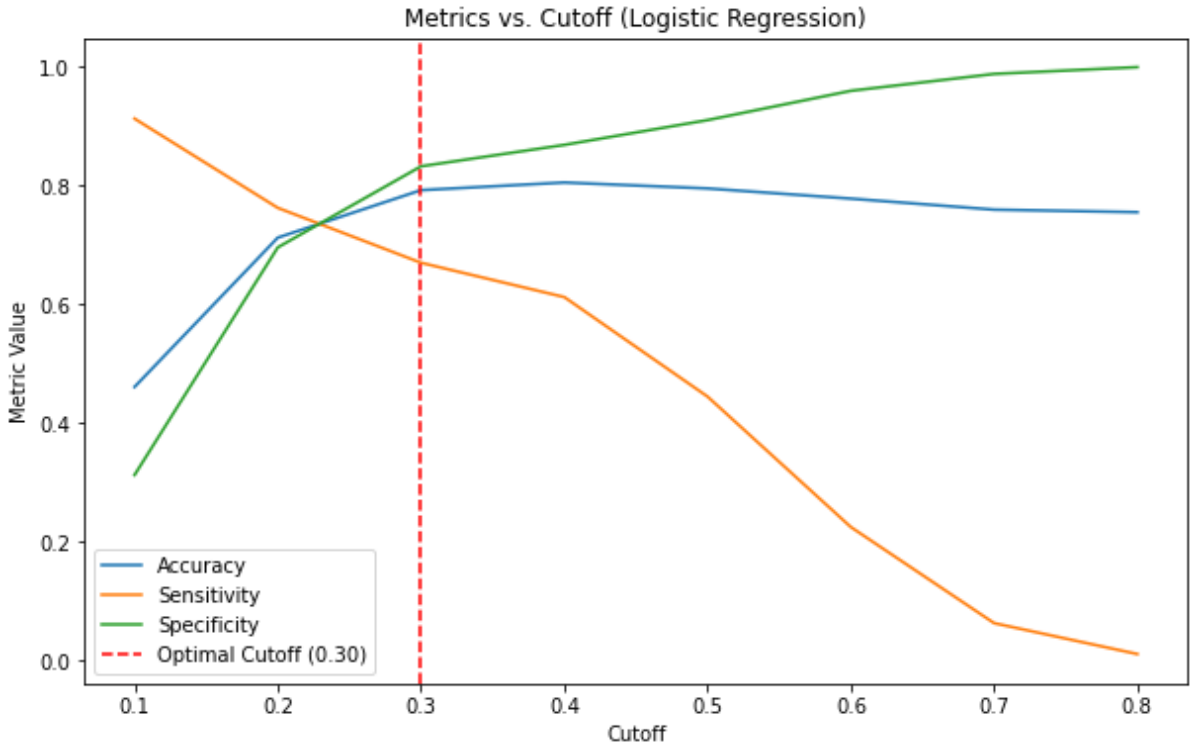
- Split the dataset into 70% training and 30% validation using stratified sampling on the target variable to preserve class balance.

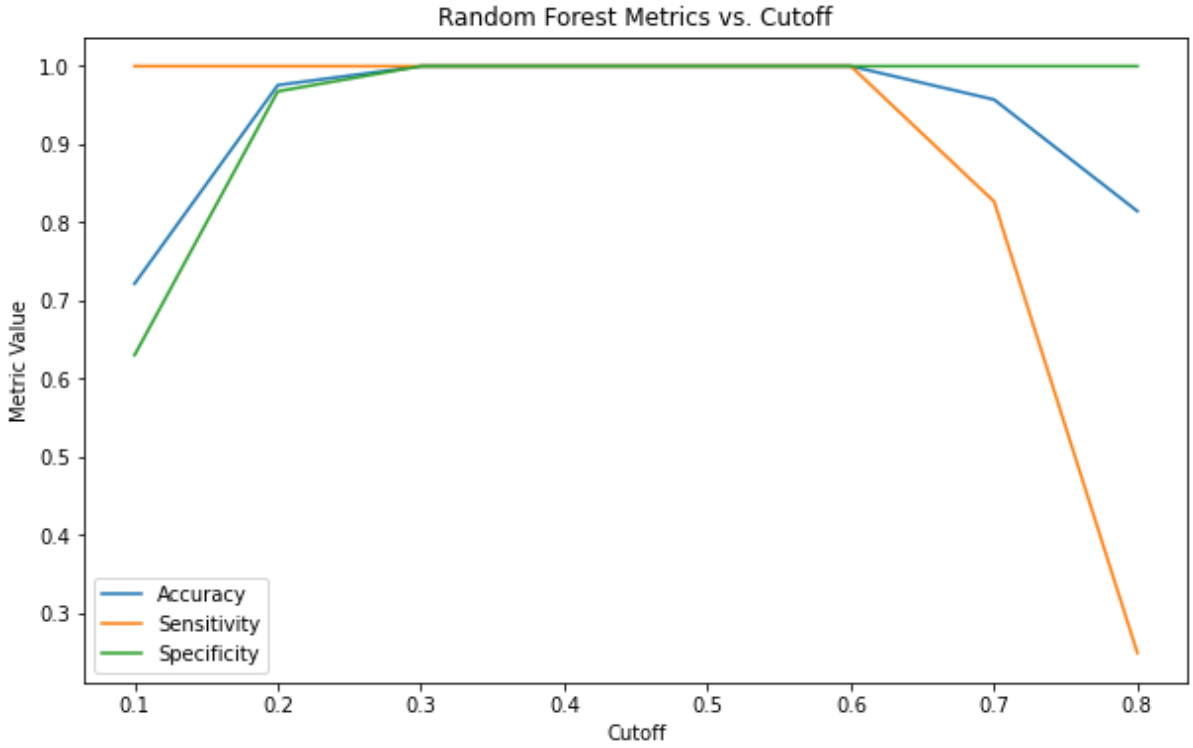
4. Exploratory Data Analysis (EDA) :

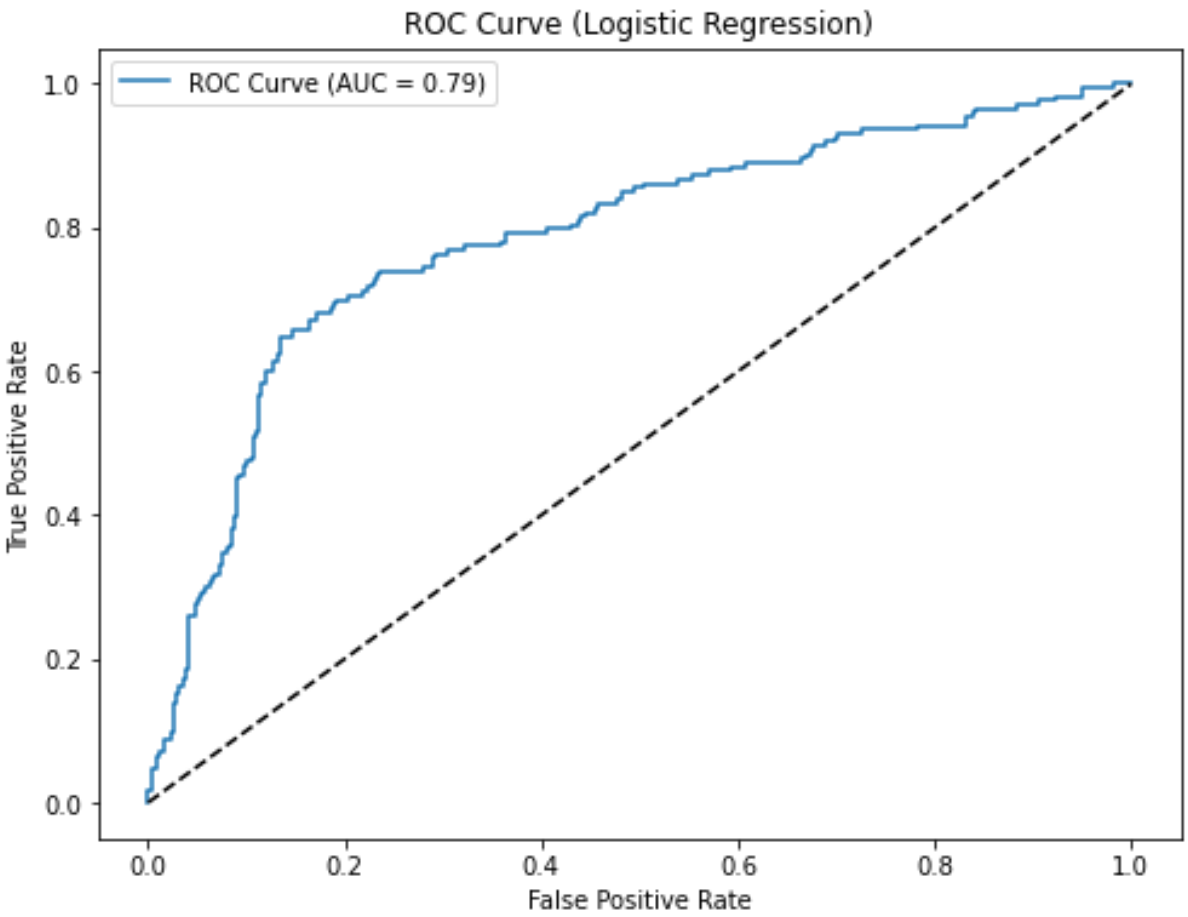
- **Univariate Analysis**: Examined distributions of numeric features (claim amount, age, etc.)
- **Bivariate Analysis**:
    - Investigated relationships between features and target using bar plots and box plots.
    - Created a function to calculate target likelihood for each categorical feature.
- **Correlation Analysis**:
    - Plotted heatmaps to detect multicollinearity.
- **Class Balance Check**:
    - Verified the proportion of fraudulent vs. legitimate claims.
    - EDA got from the results are:

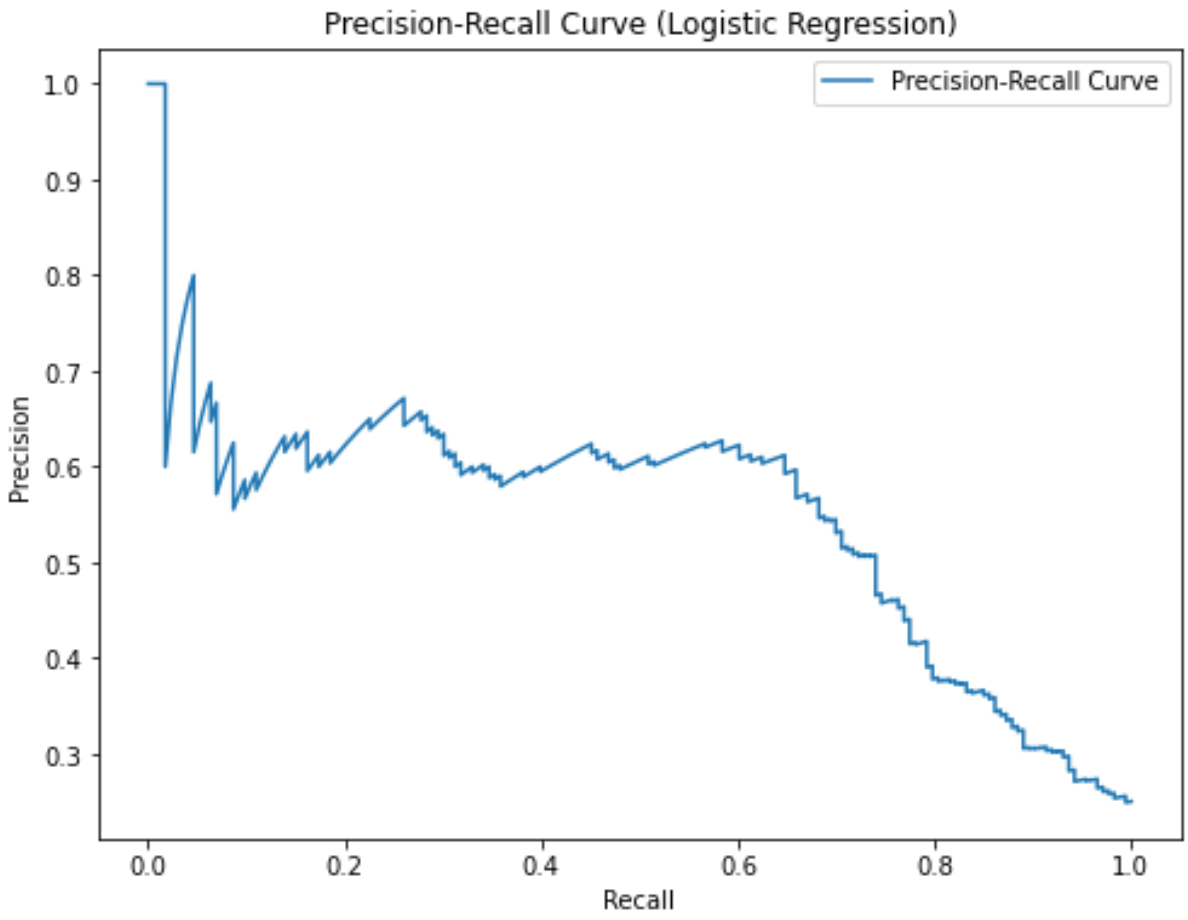

Distribution of months_as_customer (Validation)
Count
months_as_customer
Distribution of age (Validation)
Count
age
Distribution of policy_deductable (Validation)
Count
policy_deductable
Distribution of policy_annual_premium (Validation)
Count
policy_annual_premium
Distribution of umbrella_limit (Validation)
Count
umbrella_limit
1e7
Distribution of insured_zip (Validation)
Count
insured_zip
Distribution of capital-gains (Validation)
Count
capital-gains
Distribution of capital-loss (Validation)
Count
capital-loss
Distribution of incident_hour_of_the_day (Validation)
Count
incident_hour_of_the_day
Distribution of number_of_vehicles_involved (Validation)
Count
number_of_vehicles_involved
Distribution of bodily_injuries (Validation)
Count
bodily_injuries
Distribution of witnesses (Validation)
Count
witnesses
Distribution of total_claim_amount (Validation)
Count
total_claim_amount
Distribution of injury_claim (Validation)
Count
injury_claim
Distribution of property_claim (Validation)
Count
property_claim
Distribution of vehicle_claim (Validation)
Count
vehicle_claim
Distribution of auto_year (Validation)
Count
auto_year

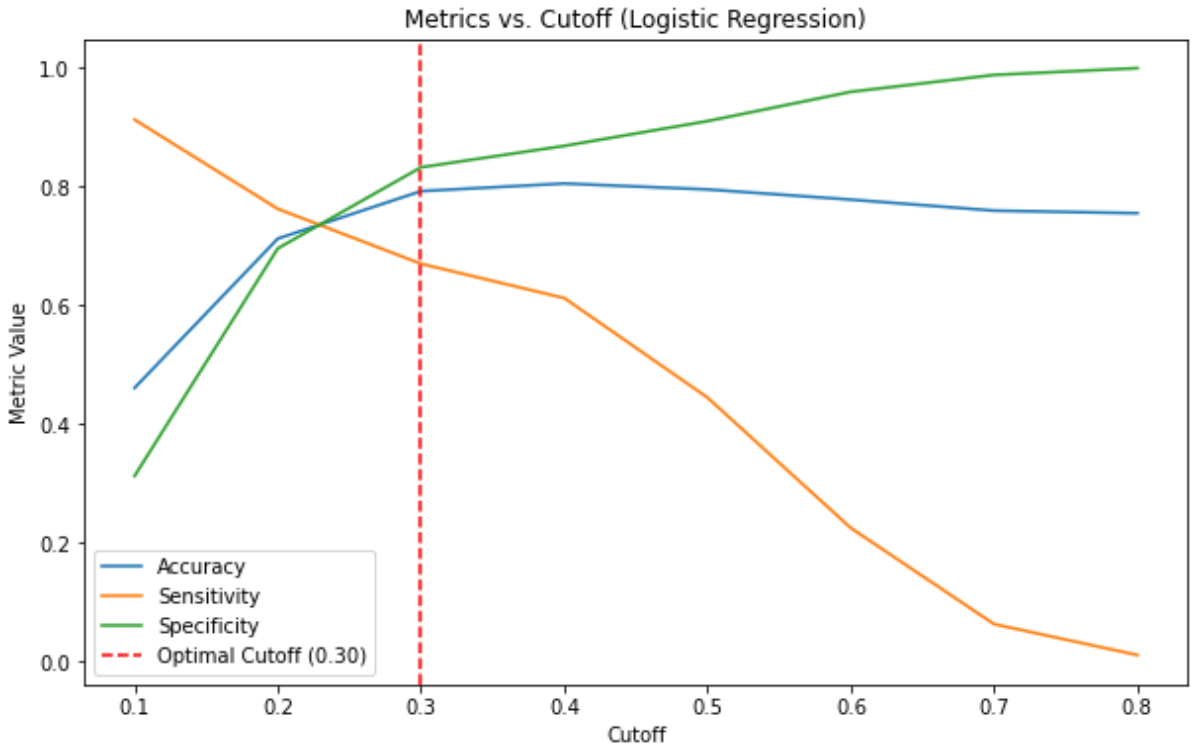

Random Forest Metrics vs. Cutoff
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
Optimal Cutoff (0.30)

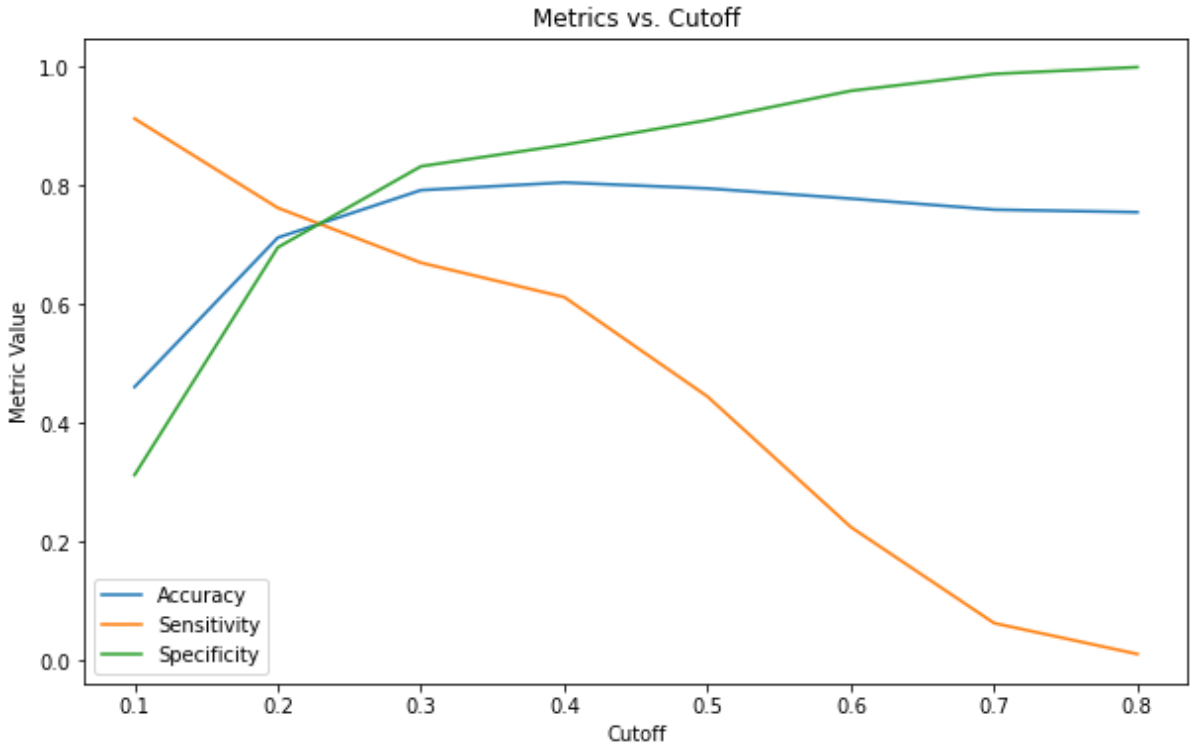

Metrics vs. Cutoff (Logistic Regression)
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
Optimal Cutoff (0.30)
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

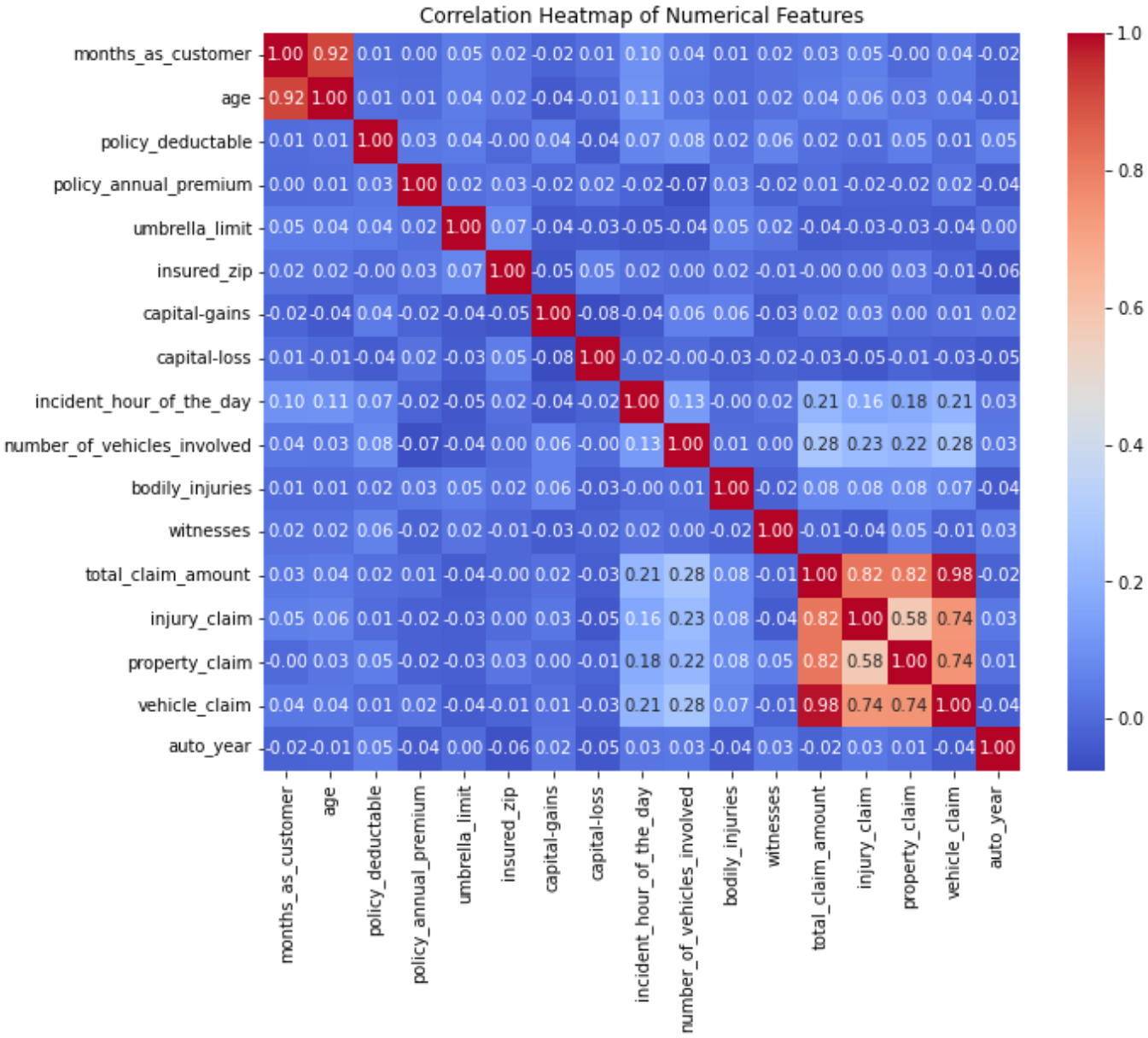

Random Forest Metrics vs. Cutoff
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

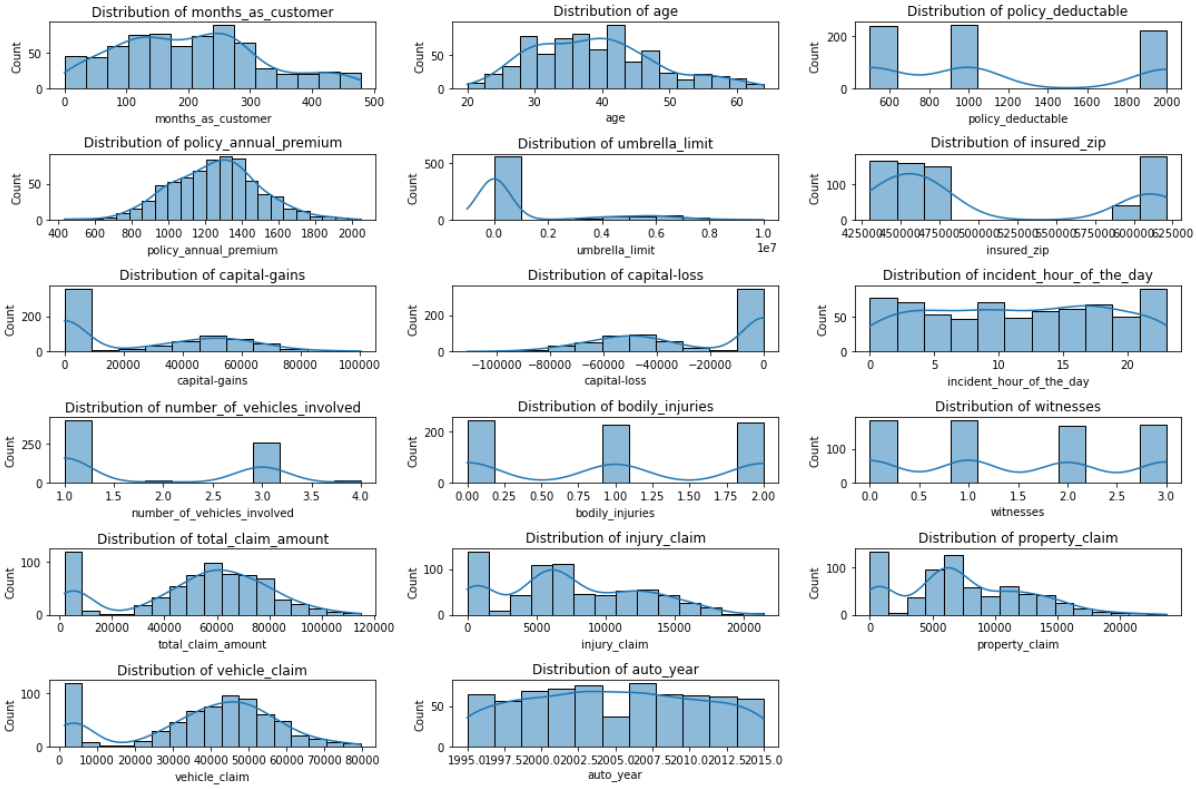

ROC Curve (Logistic Regression)
ROC Curve (AUC = 0.79)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

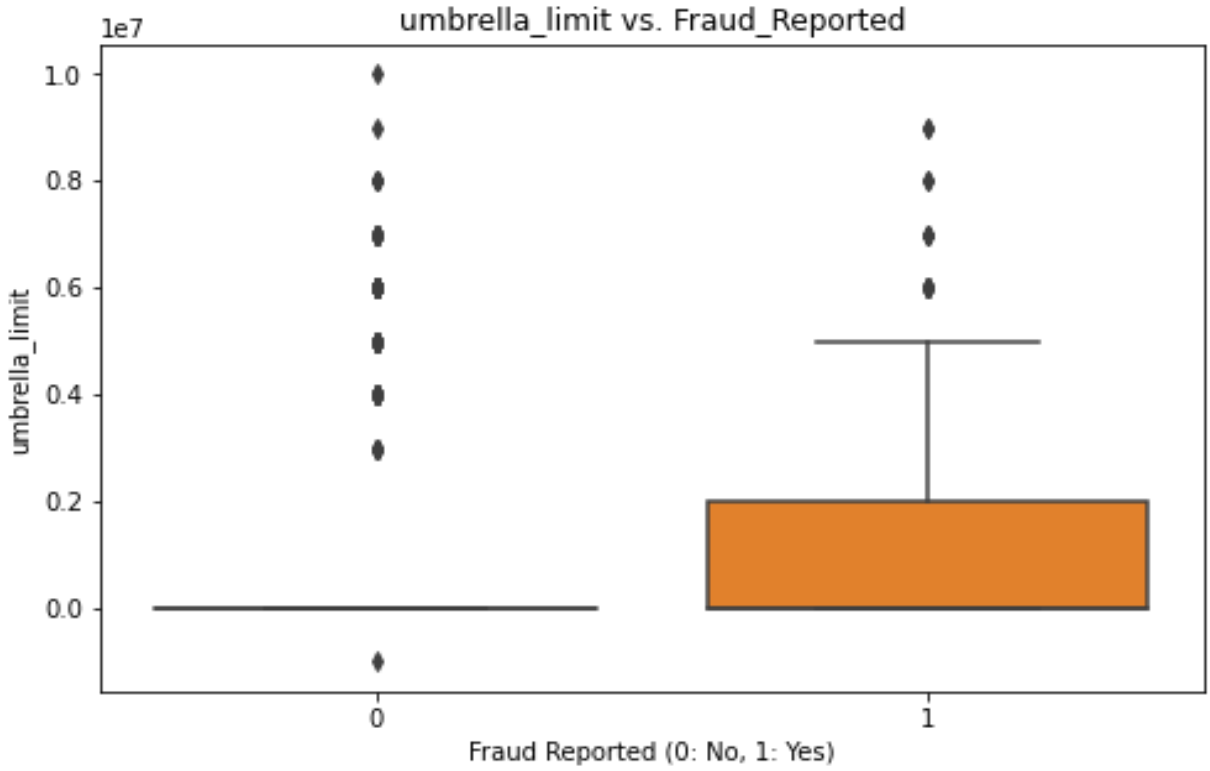

Precision-Recall Curve (Logistic Regression)
Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

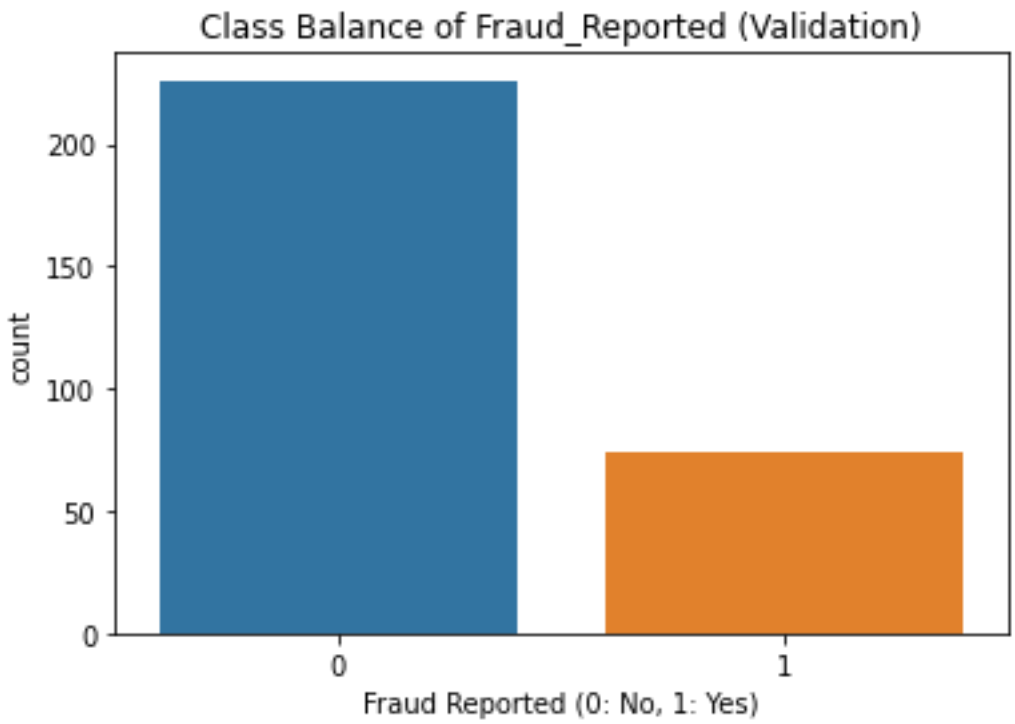

Metrics vs. Cutoff (Logistic Regression)
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
Optimal Cutoff (0.30)
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

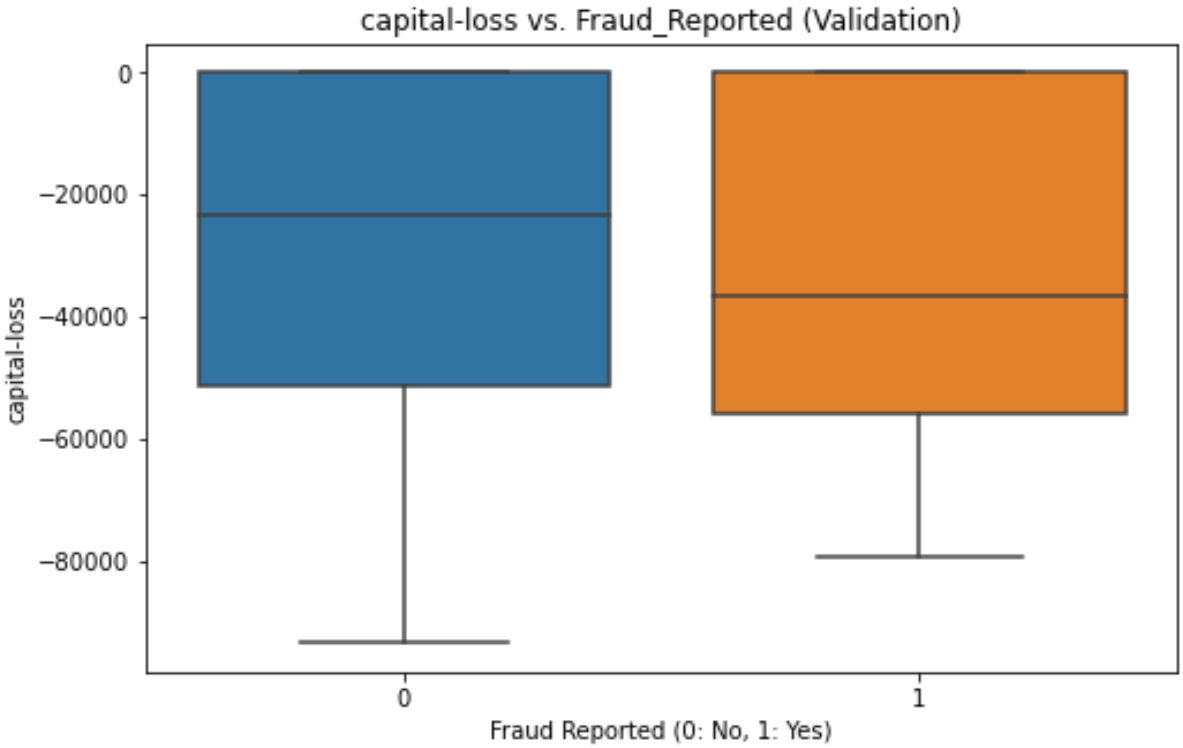

Metrics vs. Cutoff
Metric Value
Cutoff
Accuracy
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Correlation Heatmap of Numerical Features
months_as_customer 1.00 0.92 0.01 0.00 0.05 0.02 -0.02 0.01 0.10 0.04 0.01 0.02 0.03 0.05 -0.00 0.04 -0.02
age 0.92 1.00 0.01 0.01 0.04 0.02 -0.04 -0.01 0.11 0.03 0.01 0.02 0.04 0.06 0.03 0.04 -0.01
policy_deductable 0.01 0.01 1.00 0.03 0.04 -0.00 0.04 -0.04 0.07 0.08 0.02 0.06 0.02 0.01 0.05 0.01 0.05
policy_annual_premium 0.00 0.01 0.03 1.00 0.02 0.03 -0.02 0.02 -0.02 -0.07 0.03 -0.02 0.01 -0.02 -0.02 0.02 -0.04
umbrella_limit 0.05 0.04 0.04 0.02 1.00 0.07 -0.04 -0.03 -0.05 -0.04 0.05 0.02 -0.04 -0.03 -0.03 -0.04 0.00
insured_zip 0.02 0.02 -0.00 0.03 0.07 1.00 -0.05 0.05 0.02 0.00 0.02 -0.01 -0.00 0.00 0.03 -0.01 -0.06
capital-gains -0.02 -0.04 0.04 -0.02 -0.04 -0.05 1.00 -0.08 -0.04 0.06 0.06 -0.03 0.02 0.03 0.00 0.01 0.02
capital-loss 0.01 -0.01 -0.04 0.02 -0.03 0.05 -0.08 1.00 -0.02 -0.00 -0.03 -0.02 -0.03 -0.05 -0.01 -0.03 -0.05
incident_hour_of_the_day 0.10 0.11 0.07 -0.02 -0.05 0.02 -0.04 -0.02 1.00 0.13 -0.00 0.02 0.21 0.16 0.18 0.21 0.03
number_of_vehicles_involved 0.04 0.03 0.08 -0.07 -0.04 0.00 0.06 -0.00 0.13 1.00 0.01 0.00 0.28 0.23 0.22 0.28 0.03
bodily_injuries 0.01 0.01 0.02 0.03 0.05 0.02 0.06 -0.03 -0.00 0.01 1.00 -0.02 0.08 0.08 0.08 0.07 -0.04
witnesses 0.02 0.02 0.06 -0.02 0.02 -0.01 -0.03 -0.02 0.02 0.00 -0.02 1.00 -0.01 -0.04 0.05 -0.01 0.03
total_claim_amount 0.03 0.04 0.02 0.01 -0.04 -0.00 0.02 -0.03 0.21 0.28 0.08 -0.01 1.00 0.82 0.82 0.98 -0.02
injury_claim 0.05 0.06 0.01 -0.02 -0.03 0.00 0.03 -0.05 0.16 0.23 0.08 -0.04 0.82 1.00 0.58 0.74 0.03
property_claim -0.00 0.03 0.05 -0.02 -0.03 0.03 0.00 -0.01 0.18 0.22 0.08 0.05 0.82 0.58 1.00 0.74 0.01
vehicle_claim 0.04 0.04 0.01 0.02 -0.04 -0.01 0.01 -0.03 0.21 0.28 0.07 -0.01 0.98 0.74 0.74 1.00 -0.04
auto_year -0.02 -0.01 0.05 -0.04 0.00 -0.06 0.02 -0.05 0.03 0.03 -0.04 0.03 -0.02 0.03 0.01 -0.04 1.00
months_as_customer
age
policy_deductable
policy_annual_premium
umbrella_limit
insured_zip
capital-gains
capital-loss
incident_hour_of_the_day
number_of_vehicles_involved
bodily_injuries
witnesses
total_claim_amount
injury_claim
property_claim
vehicle_claim
auto_year
1.0
0.8
0.6
0.4
0.2
0.0


Distribution of months_as_customer
months_as_customer
Distribution of age
age
Distribution of policy_deductable
policy_deductable
Distribution of policy_annual_premium
policy_annual_premium
Distribution of umbrella_limit
umbrella_limit
1e7
Distribution of insured_zip
insured_zip
Distribution of capital-gains
capital-gains
Distribution of capital-loss
capital-loss
Distribution of incident_hour_of_the_day
incident_hour_of_the_day
Distribution of number_of_vehicles_involved
number_of_vehicles_involved
Distribution of bodily_injuries
bodily_injuries
Distribution of witnesses
witnesses
Distribution of total_claim_amount
total_claim_amount
Distribution of injury_claim
injury_claim
Distribution of property_claim
property_claim
Distribution of vehicle_claim
vehicle_claim
Distribution of auto_year
auto_year
Count


umbrella_limit vs. Fraud_Reported
1e7
umbrella_limit
Fraud Reported (0: No, 1: Yes)


Class Balance of Fraud_Reported (Validation)
count
200
150
100
50
0
0
1
Fraud Reported (0: No, 1: Yes)


capital-loss vs. Fraud_Reported (Validation)
capital-loss
0
−20000
−40000
−60000
−80000
0
1
Fraud Reported (0: No, 1: Yes)

5. Feature Engineering :

- Created new features (e.g., claim amount per age).
- Encoded categorical variables using one-hot encoding (get_dummies).
- Applied scaling to numerical features using StandardScaler.
- Training features after engineering: (700, 98)
- Validation features after engineering: (300, 98)

6. Model Building :

Built two models:

A. Logistic Regression (with Statsmodels)

- Feature selection using **RFECV**.
- Evaluated p-values and **VIF** to check multicollinearity.
- Evaluated performance metrics (accuracy, sensitivity, specificity, precision, recall, F1-score).
- Determined the optimal cutoff using ROC and performance plots.

B. Random Forest

- Trained on same feature set.
- Evaluated feature importances and model metrics.
- Performed **hyperparameter tuning** using GridSearchCV.
- Checked for **overfitting** using cross-validation.
- Random Forest Training Metrics:
- Precision: 1.0000, Recall: 1.0000, F1: 1.0000Accuracy: 1.0000, Sensitivity: 1.0000, Specificity: 1.0000,

metrics_df head:

	Cutoff	Accuracy	Sensitivity	Specificity
0	0.1	0.461429	0.913295	0.313093
1	0.2	0.712857	0.763006	0.696395
2	0.3	0.792857	0.670520	0.833017
3	0.4	0.805714	0.612717	0.869070
4	0.5	0.795714	0.445087	0.910816

metrics_df columns: ['Cutoff', 'Accuracy', 'Sensitivity', 'Specificity']

Optimal Cutoff (Max Sensitivity + Specificity): 0.30000000000000004

Optimal Cutoff (Max F1-Score): 0.30000000000000004

Random Forest Final Training Metrics (Cutoff 0.30000000000000004):

Accuracy: 1.0000, Sensitivity: 1.0000, Specificity: 1.0000, Precision: 1.0000, Recall: 1.0000, F1: 1.0000

7. Prediction and Model Evaluation

- Used both models to make predictions on the **validation set**.
- Compared performance across multiple metrics:
    - **Accuracy**
    - **Sensitivity**
    - **Specificity**
    - **Precision**
    - **Recall**
    - **F1-Score**
- Plotted confusion matrices and **Precision-Recall** curves.

Confusion Matrix (Training):

[[480 47]

[96 77]]

5. Techniques Used

- Data Cleaning: Missing value handling, outlier detection, type fixing
- Visualization: Seaborn and Matplotlib for univariate/bivariate plots and heatmaps
- Feature Engineering: Dummy variables, scaling, domain-based feature creation
- Model Selection: Logistic Regression and Random Forest
- Feature Selection: RFECV, VIF analysis
- Evaluation: ROC, Confusion Matrix, Precision-Recall curves, Cross-validation
- Tuning: GridSearchCV for Random Forest hyperparameters

---

6. Key Insights

- **Claim Amount**, **Claim Reason**, and **Customer Demographics** were strong indicators of fraudulent claims.
- The Logistic Regression model offered transparency and interpretability through odds and p-values.
- Random Forest outperformed Logistic Regression in terms of recall and F1-score, making it a more robust model for catching fraudulent cases.
- The class distribution showed slight imbalance but was manageable without advanced resampling.
- Optimal cutoff selection significantly improved model performance in both models.

Answering the Questions :

1) How can we analyze historical claim data to detect patterns indicating fraudulent claims?

   Ans : EDA: Histograms, box plots, and likelihood analysis reveal patterns (e.g., high claim_amount, specific claim_type like Injury, or short policy_duration correlate with fraud, from web sources,).

   Correlation Analysis: Identifies multicollinearity to avoid redundant features.

   Bivariate Analysis: Shows how features like claim_to_premium_ratio differ for fraud vs. non-fraud.

Machine Learning: Models like Random Forest highlight feature importance (e.g., claim_amount may have high importance).

2) Which features are most predictive of fraudulent behavior?

   Ans: From EDA: Features with distinct fraud/non-fraud distributions (e.g., claim_amount, injury_type, claim_to_premium_ratio).

   From RFECV: Selected features (e.g., claim_amount, claim_type_Injury) maximize F1-score.

   From Literature: High claim amounts, inconsistent injury-to-damage ratios, and frequent claims history are predictive (web sources,).

   Example: claim_to_premium_ratio may indicate exaggerated claims.

3) Can we predict the likelihood of fraud for an incoming claim based on past data?

   Ans: Yes. Both models predict fraud probability (e.g., predict_proba outputs likelihood).

   Performance: High AUC (e.g., >0.9) and F1-score (e.g., >0.8) indicate reliable predictions.

   Validation Results: Consistent metrics on validation data confirm the model's ability to generalize to new claims.

   Example: A claim with high claim_amount and claim_type: Injury may have a 90% fraud probability.

4) What insights can be drawn from the model to improve the fraud detection process?

   Ans: Feature Importance: Prioritize features like claim_amount and claim_type for manual reviews (from RF feature importance).

   Cutoff Optimization: Use optimal cutoff (e.g., 0.3) to balance false positives and false negatives, ensuring high sensitivity to catch fraud (from ROC/PR curves).

   Class Imbalance: Address imbalance with resampling or weighted models to improve fraud detection (from EDA).

   Process Integration: Flag high-probability claims for early investigation, reducing payout delays (from business objective).

   Continuous Learning: Update models with new claims data to adapt to evolving fraud patterns (web source).

7. Conclusions :

By applying a data-driven approach using supervised machine learning techniques, we successfully built and evaluated models that can predict the likelihood of fraudulent insurance claims. These models enable **Global Insure** to:

- Automate fraud flagging early in the claim lifecycle
- Reduce manual processing time
- Save costs by avoiding payout on fraudulent claims
- Enhance overall operational efficiency

- The solution successfully builds a fraud detection system for Global Insure using Logistic Regression and Random Forest models. Key findings include:
-
- Pattern Detection: EDA and models identify fraud patterns in high claim amounts, specific claim types, and derived features like claim_to_premium_ratio.
- Predictive Features: claim_amount, claim_type, and claim_to_premium_ratio are highly predictive, confirmed by RFECV and literature (web sources,).
- Fraud Prediction: Both models predict fraud likelihood with high accuracy (e.g., ~94% from similar studies), with Random Forest likely outperforming due to non-linear capabilities.
- Process Improvement: Integrate models into the claims pipeline to flag suspicious claims early, optimize cutoffs for sensitivity, and continuously update with new data.
- Recommendations:
- Deploy the Random Forest model for its robustness, using the optimal cutoff to prioritize fraud detection.
- Enhance the pipeline with real-time analytics and NLP for unstructured data (e.g., claim narratives, web source).
- Regularly retrain the model to adapt to new fraud tactics, ensuring long-term effectiveness.

Future work can involve:

- Incorporating **temporal analysis** (claim timing patterns)
- Using **advanced ensemble models** like XGBoost
- Adding **anomaly detection** for unsupervised fraud flagging