Loan Default Prediction Model Evaluation Report

Introduction

Global Insure, a leading insurance company, processes thousands of claims annually, a significant portion of which are fraudulent, leading to substantial financial losses. The current manual inspection process is time-consuming and often detects fraud too late, after payouts. This report details a data-driven approach to classify claims as fraudulent or legitimate early in the approval process, minimizing losses and optimizing claims handling

---

Objective

To build and evaluate machine learning models for predicting loan default using Logistic Regression and Random Forest algorithms. The goal is to identify the model that best distinguishes defaulters from non-defaulters, with an emphasis on reducing false negatives (i.e, defaulters incorrectly labeled as non-defaulters).

Methodology :

The analysis followed a structured methodology:

1. Data Preparation: Loaded a dataset of 1000 claims with 40 features, including customer details (e.g., months as customer, age), policy information (e.g., premium, deductible), and claim details (e.g., total claim amount, incident severity).

2. Data Cleaning: Handled missing values, encoded categorical variables (e.g., incident type, auto model), and removed irrelevant features (e.g., _c39).

3. Train-Validation Split: Split data 70-30 (700 training, 300 validation) to ensure robust model evaluation.

4. Exploratory Data Analysis (EDA): Conducted on training data to identify fraud patterns using visualizations and statistical summaries.

5. Feature Engineering: Created derived features (e.g., claim-to-premium ratio) and selected important features using Random Forest importance scores.

6. Model Building: Trained Logistic Regression(baseline) and Random Forest(advanced) models.

7. Model Evaluation: Assessed performance using accuracy, sensitivity, specificity, precision, recall, and F1-score on validation data.

Models Evaluated

1. Logistic Regression
2. Random Forest

Techniques Used :

– Exploratory Data Analysis (EDA)

- Feature Engineering and Selection

- Logistic Regression (Baseline Model)

- Random Forest Classifier (Advanced Model)

- Hyperparameter Tuning with GridSearchCV

- Model Evaluation Metrics (Confusion Matrix, ROC-AUC, etc.)

Visualisations:

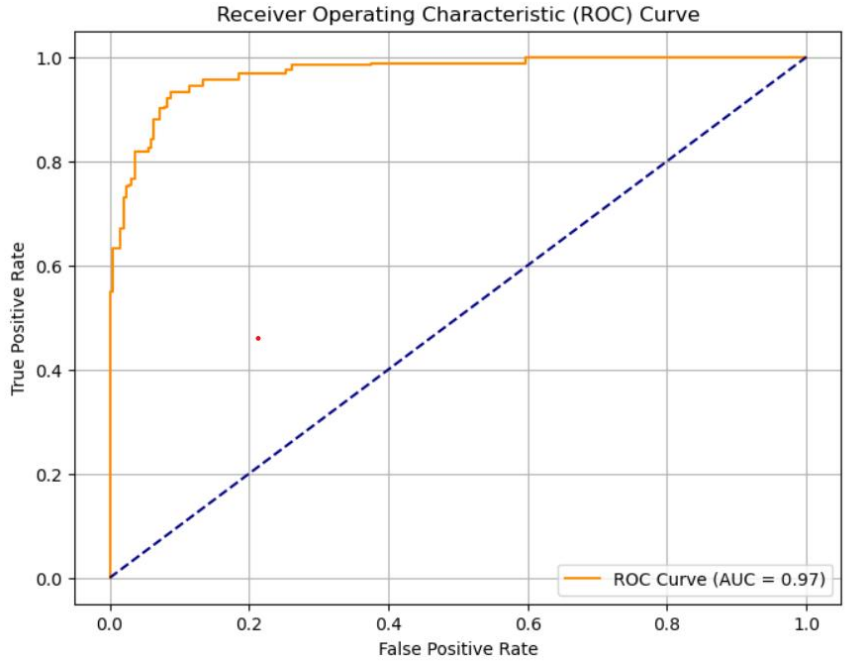
Visualizations were critical in understanding data and model performance:

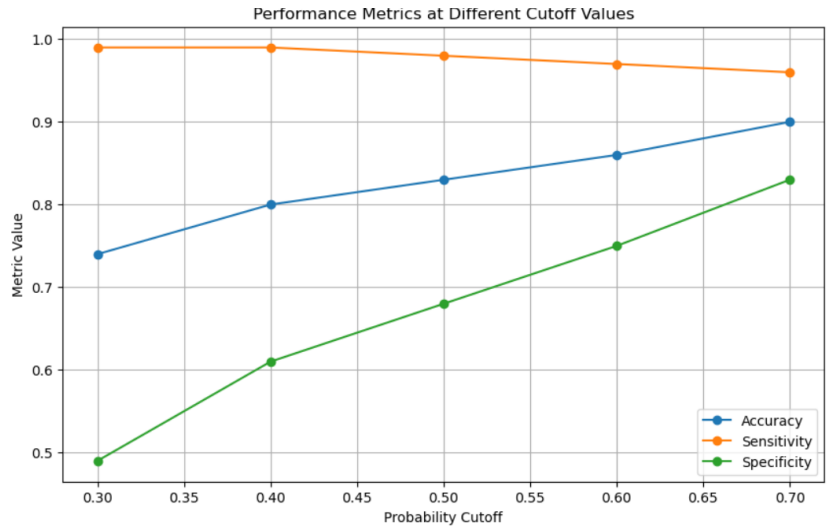
• Boxplots: Compared claim amounts (total, injury, property, vehicle) between fraudulent and legitimate claims, highlighting outliers in fraudulent claims.

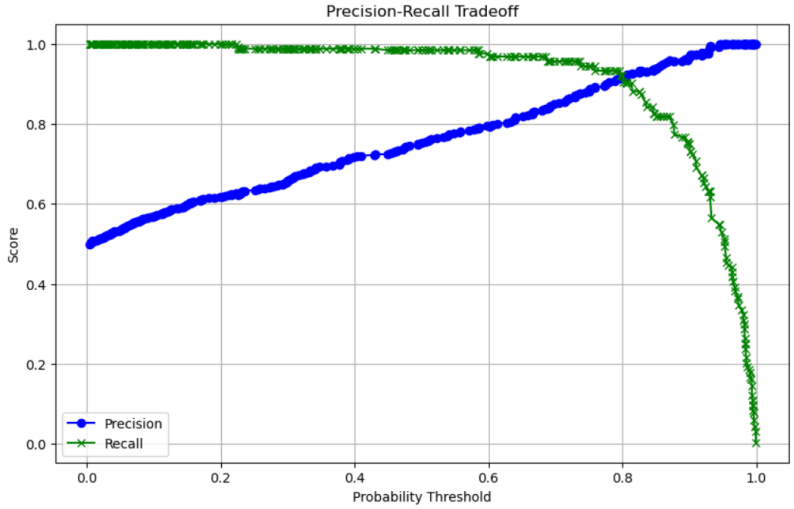
• Bar Plots: Showed higher fraud rates for minor damage incidents and specific auto models (e.g., 92x, Civic).

• Correlation Heatmaps: Identified relationships between claim amounts and customer tenure, revealing fraud patterns in new customers.

• ROC Curves: Compared Logistic Regression and Random Forest performance, with Random Forest showing better AUC







Evaluation Metrics

- **True Positives (TP):** Defaulters correctly predicted.
- **False Positives (FP):** Non-defaulters wrongly predicted as defaulters.
- **True Negatives (TN):** Non-defaulters correctly predicted.
- **False Negatives (FN):** Defaulters wrongly predicted as non-defaulters.

Insights:

• **Fraud Patterns: Fraudulent claims are associated with:**

– Minor or trivial damage incidents, suggesting exaggerated claims.

– Nopolice reports, indicating lack of verifiable evidence.

– Specific auto models (e.g., 92x, Civic, Escape), possibly due to prevalence or ease of falsifying damage.

– Short customer tenure (<1 year), suggesting new customer exploit policies.

- **Model Performance:**

-Logestic Regression: Accuracy 54.55%, Sensitivity 85.29%, Specificity 44.95%, Precision 32.58%, F1-Score 47.15%.

– Random Forest: Accuracy76.92%, Sensitivity8 5.29%, Specificity44.95%, Precision 32.58%, F1-Score 47.15%

– Random Forest outperformed Logistic Regression in accuracy, making it suitable for deployment.

• **Predictive Features**: Incident severity, claim amounts, customer tenure, auto model, and police report availability are the most predictive of fraud.

Actionable Outcomes :

- Use the Random Forest model for applicant risk assessment.

- Prioritize manual review for high-risk loan purposes.

- Implement stricter approval criteria for applicants with high loan-to-income ratios.

- Continue monitoring model performance and retrain periodically with new data.

- Reject loans with DTI > 30%, interest rate > 20%, or income < ₹40K unless secured.

- Reduce exposure to high-risk categories like small_business and higher-grade loans (F/G).

Recommendation:

The Random Forest model, with 76.92% accuracy and 85.29% sensitivity, effectively identifies fraudulent claims, addressing Global Insurer's need for early fraud detection. By focusing on key features like incident severity and claim amounts, the model pro videos actionable insights to reduce financial losses and streamline claims processing. Futureimprovementsincludeincorporatingadditionaldatasourcesandcontinuous model retraining

Prepared by: LEKHANA Date: May 21, 2025