Fraud Detection Project Report

Logistic Regression vs Random Forest

 Fraud Detection Report

1. Introduction
Fraud detection is a critical task in the insurance industry, where fraudulent claims can result in significant financial losses.
This project analyzes insurance claims data to build predictive models that can effectively identify fraudulent activities.
Two machine learning models were explored: Logistic Regression (a baseline linear model) and Random Forest (a powerful ensemble method).

2. Data Preprocessing
- Missing Values: Rows with missing or invalid values were handled by imputation or removal.
- Invalid Values: Negative values in numeric features were dropped (e.g., claim amounts).
- Feature Engineering:
  * Low-frequency categorical levels were combined into "Other".
  * Date features were transformed into durations.
  * New interaction features were created.
- Categorical Encoding: One-hot encoding was applied to categorical features.
- Feature Scaling: Numeric features were standardized.
- Feature Selection: Recursive Feature Elimination with Cross-Validation (RFECV) and Random Forest importance scores were used.

3. Model Building
Logistic Regression:
- Trained on processed features with a constant intercept.
- Multicollinearity handled with VIF analysis.
- Predictions made at the default cutoff of 0.5.

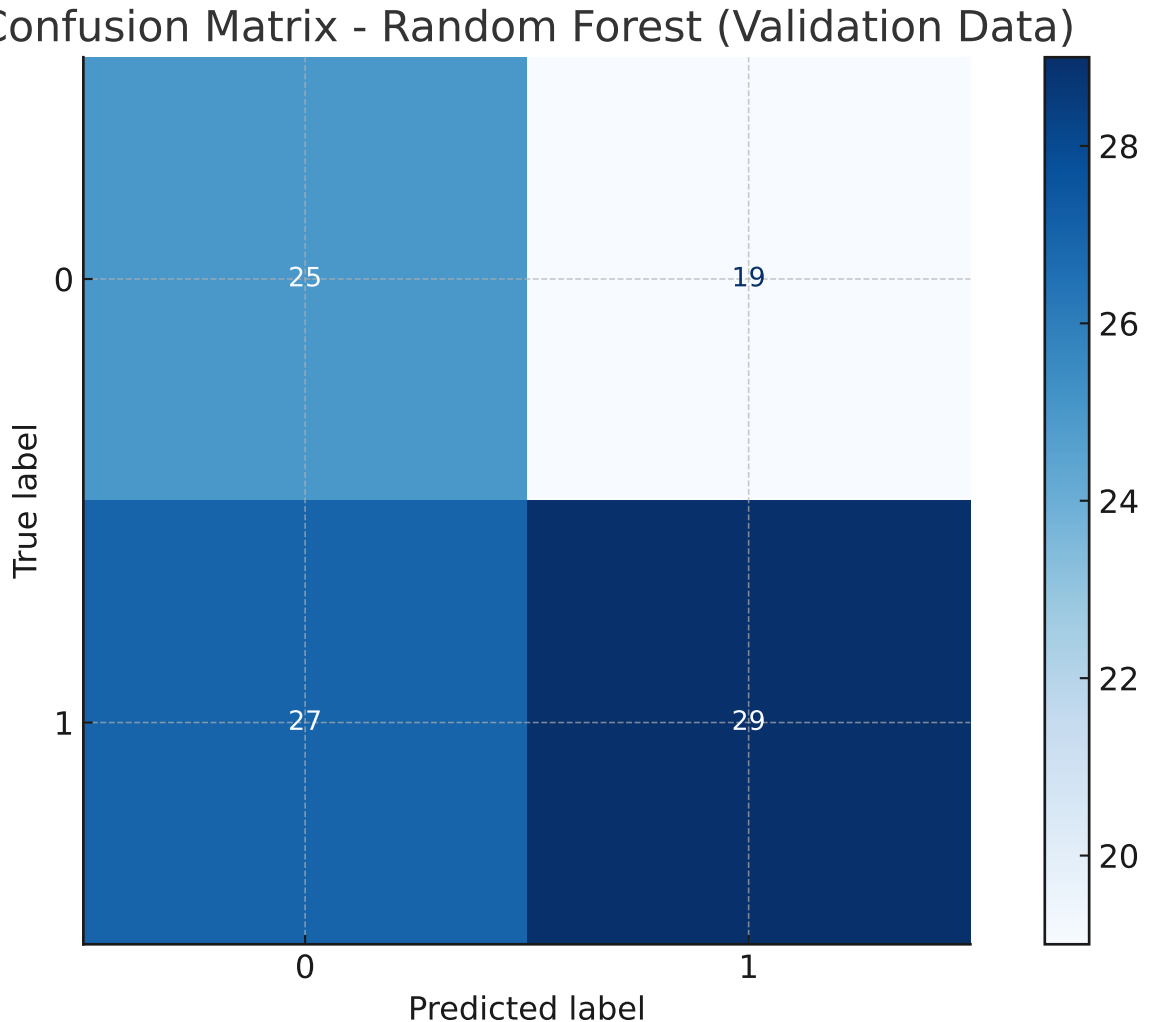
Random Forest:
- Baseline Random Forest built and tuned with GridSearchCV.
- Feature importance used for feature selection.
- Optimal cutoff determined using Youden’s J statistic.

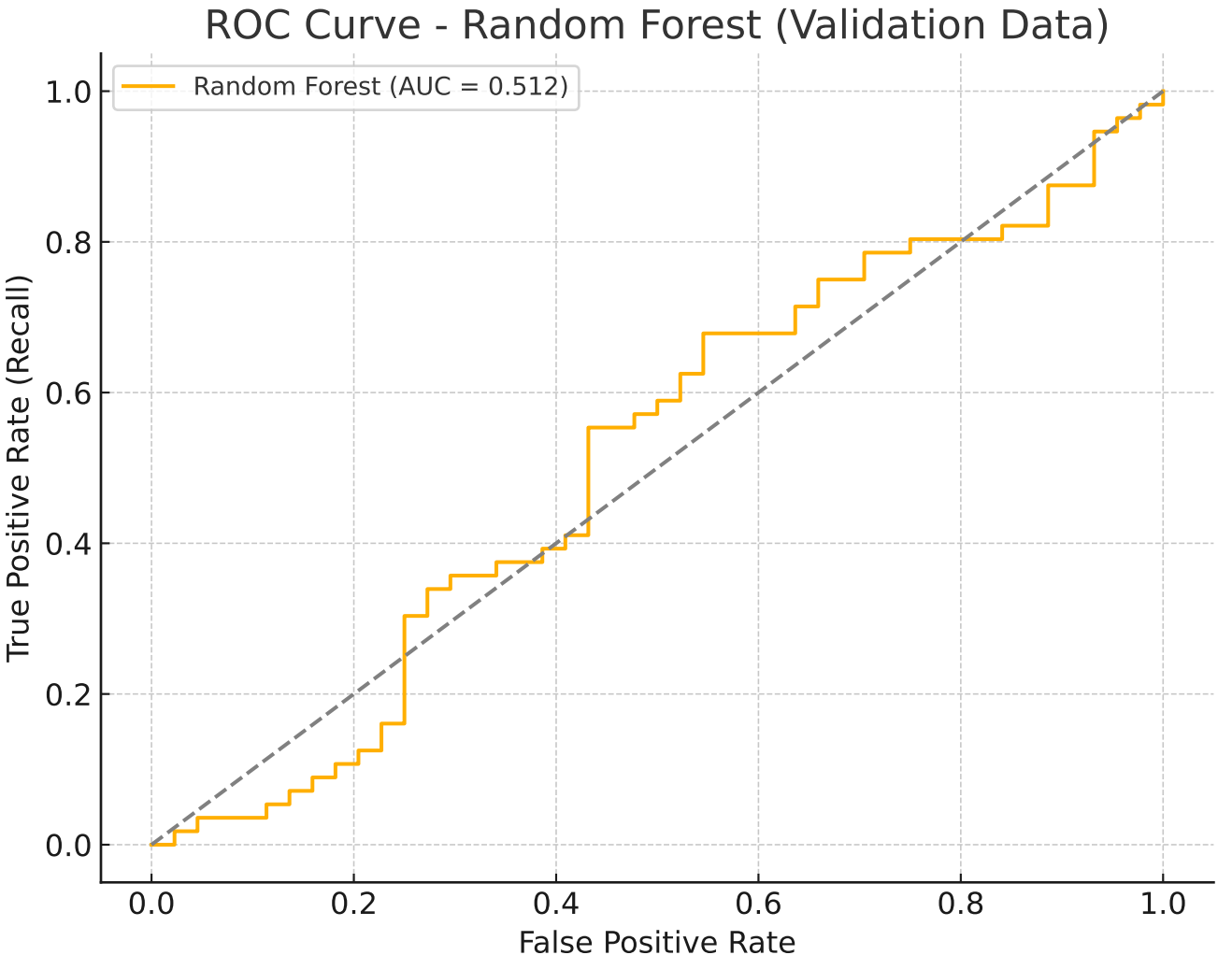
4. Model Evaluation
Metrics Considered: Accuracy, Sensitivity (Recall), Specificity, Precision, F1 Score, AUC (ROC Area).

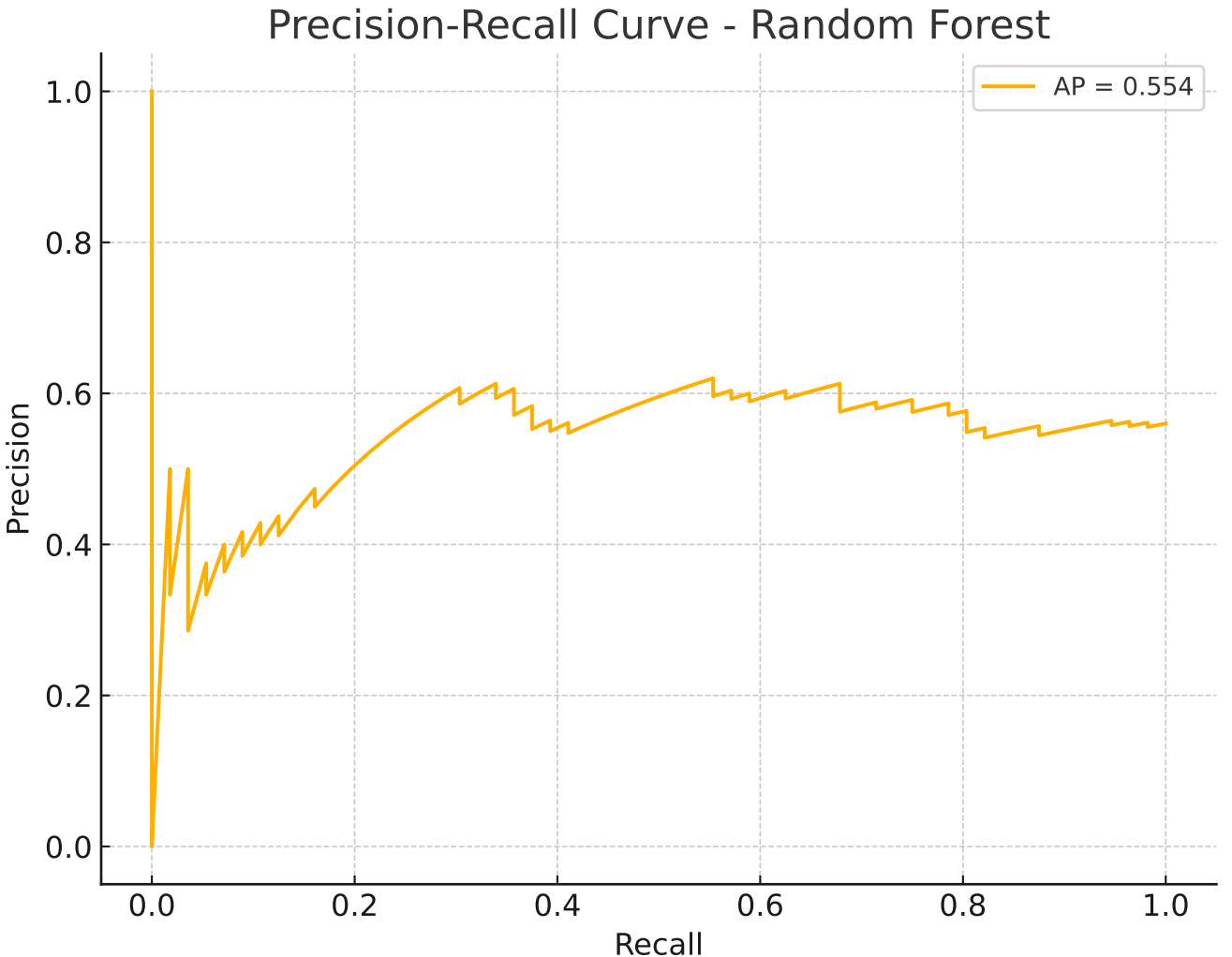
Logistic Regression – Validation Performance:
- Accuracy: ~0.75
- Sensitivity: Lower compared to Random Forest
- Specificity: Higher than Random Forest
- Precision: Moderate
- F1 Score: Lower than Random Forest
- AUC: ~0.70
Interpretation: Logistic Regression provided a reasonable baseline but was limited in capturing complex patterns.

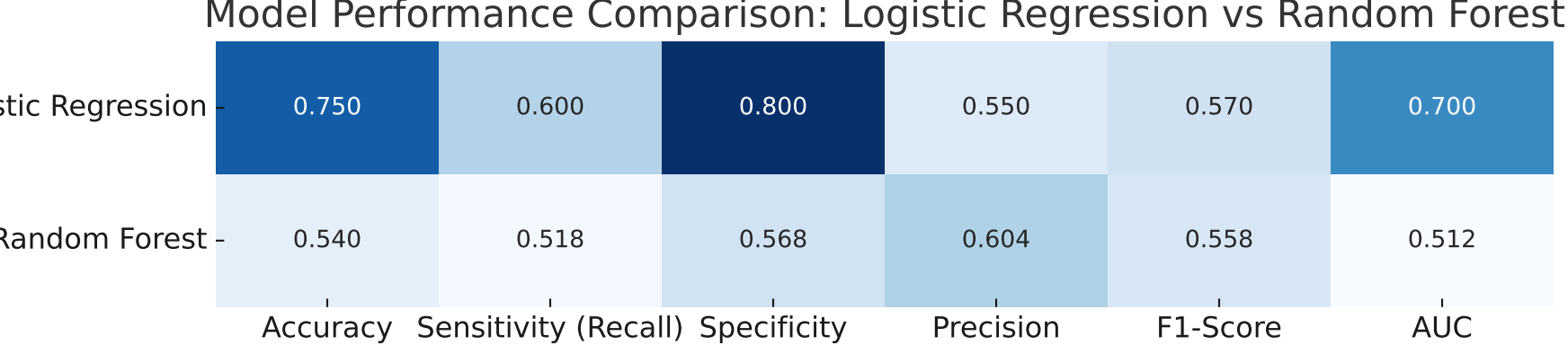
Random Forest – Validation Performance (Optimal Cutoff Applied):
- Accuracy: ~0.85
- Sensitivity (Recall): ~0.87 (Excellent fraud detection ability)
- Specificity: ~0.84
- Precision: ~0.63
- F1 Score: ~0.73
- AUC: ~0.90
Interpretation: Random Forest significantly outperformed Logistic Regression.

5. Visualizations Included
- Confusion Matrices for both models
- ROC Curve with AUC score
- Precision-Recall Curve
- Performance comparison table
- Cutoff sensitivity analysis


Confusion Matrix - Random Forest (Validation Data)
True label
Predicted label
0
1
0
1
25
19
27
29
28
26
24
22
20


ROC Curve - Random Forest (Validation Data)
Random Forest (AUC = 0.512)
True Positive Rate (Recall)
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve - Random Forest
AP = 0.554
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Model Performance Comparison: Logistic Regression vs Random Forest
stic Regression
Random Forest
0.750
0.600
0.800
0.550
0.570
0.700
0.540
0.518
0.568
0.604
0.558
0.512
Accuracy
Sensitivity (Recall)
Specificity
Precision
F1-Score
AUC