FRAUDULENT CLAIM DETECTION

ADVANCED ML CASE STUDY - ASSIGNMENT

BY,
ANISH DHONDI
ANKIT SHRIVASTAVA

PROBLEM STATEMENT

Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The company's current process for identifying fraudulent claims involves manual inspections, which is time-consuming and inefficient. Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts. Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process. This would minimise financial losses and optimise the overall claims handling process.

BUSINESS OBJECTIVE

Global Insure 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.

SUMMARY OF FINDINGS

🔍 What We Did

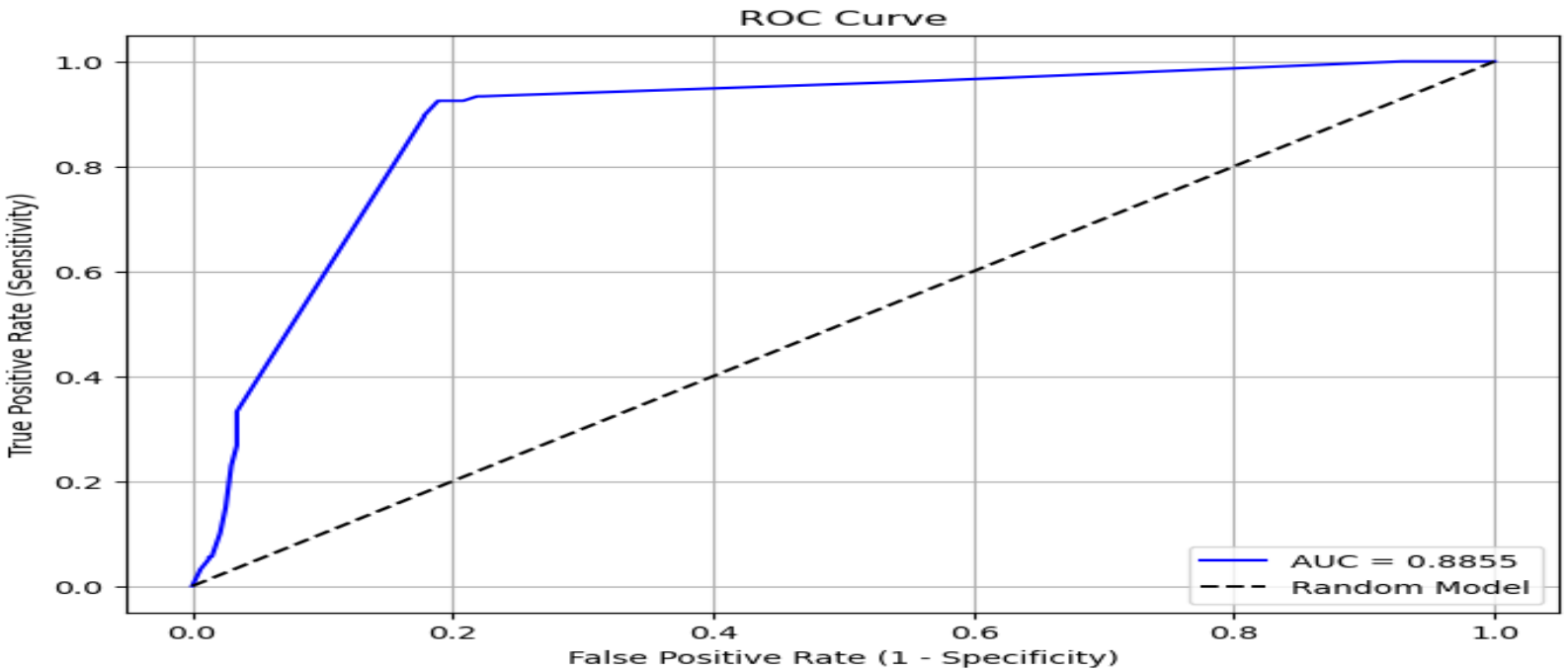
- Analyzed historical claim data from Global Insure.
- Performed data cleaning, EDA, and feature engineering.
- Built and compared Logistic Regression and Random Forest models.
- Tuned cutoffs and hyperparameters for optimal performance.

SUMMARY OF FINDINGS

📈 What We Observed

• Fraud cases are fewer but costlier.
Claims with high severity and specific hobbies showed higher fraud likelihood.
• Certain vehicle models and occupations are linked with fraud.
• Key numeric features like *total_claim_amount*, *injury_ratio*, and *property_claim* were highly predictive.
• **Random Forest** initially overfit; **Logistic Regression** offered stable generalization.

Model Evaluation – ROC Curve



Observation: The ROC Curve illustrates strong classification ability with high AUC. This suggests the model is capable of distinguishing between fraudulent and legitimate claims effectively.

SUMMARY OF FINDINGS

🏁 Final Outcome

- **Selected Model:** Logistic Regression with 0.4 cutoff.
- **Recall:** 87.5% (captures more fraud).
- **F1 Score:** 74.1% (balanced performance).
- **Deployment-Ready:** Interpretable, scalable, and business-aligned.

Metric	Logistic Regression (cutoff=0.4)	Random Forest (tuned)
Accuracy	86.80%	83.20%
Precision	64.30%	66.30%
Recall	87.50%	73.60%
Specificity	82.60%	86.60%
F1 Score	74.10%	69.70%

CONFUSION MATRIX – LOGISTIC REGRESSION (CUTOFF = 0.4)

7.3.5 Calculate the accuracy [1 Mark]

[80]:
```
# Check the accuracy now
from sklearn.metrics import accuracy_score
final_accuracy = accuracy_score(predict_df['Actual'], predict_df['Final_Prediction'])
print(f"Final Model Accuracy (at cutoff = 0.4) is {final_accuracy:.4f}")
```

```
Final Model Accuracy (at cutoff = 0.4) is 0.8680
```

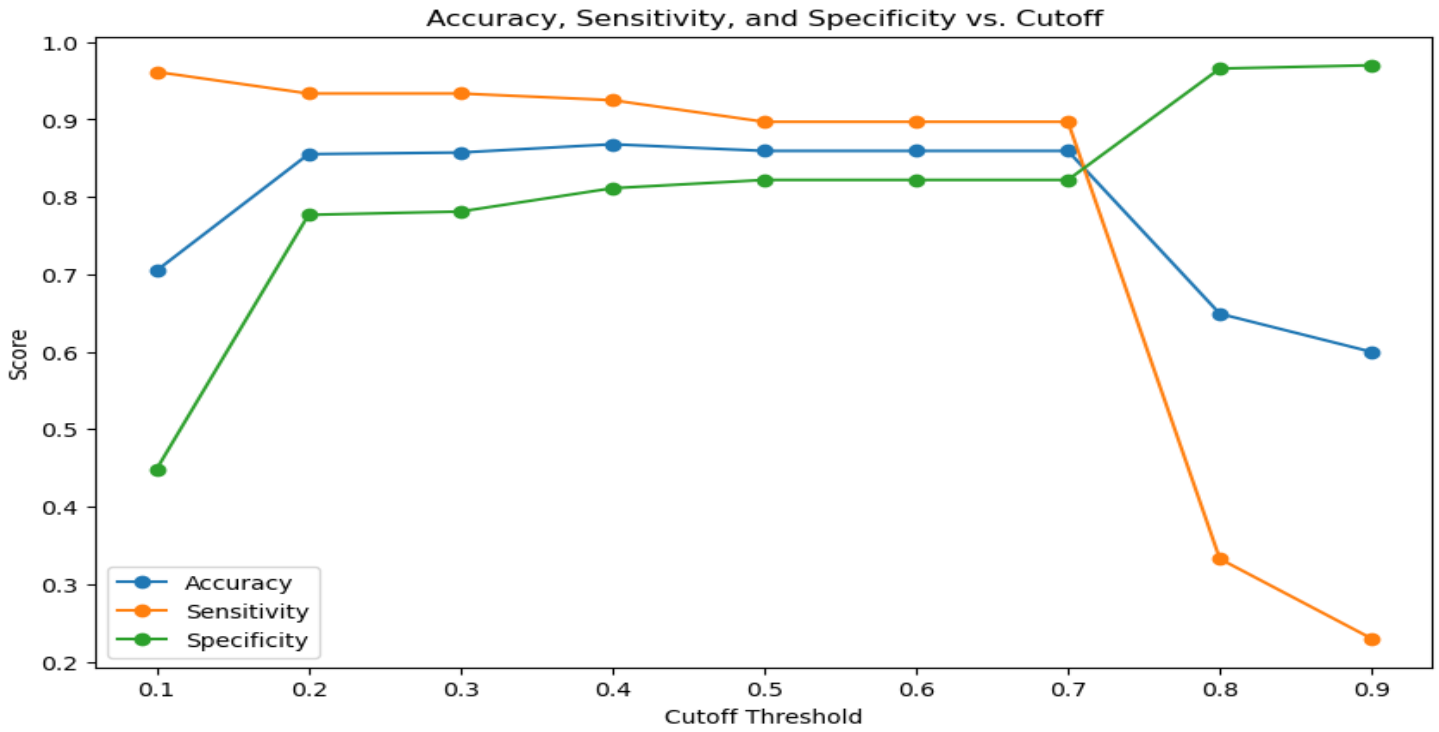
7.3.6 Create confusion matrix [1 Mark]

[81]:
```
# Create the confusion matrix once again
conf_matrix = confusion_matrix(predict_df['Actual'], predict_df['Final_Prediction'])
print(conf_matrix)
```

```
[[378  88]
 [ 35 431]]
```

Observation: The confusion matrix shows the model captures a high number of actual frauds (high recall of 86.8%), although with some false positives — a desirable tradeoff for fraud detection tasks.

MODEL METRICS VS. CUTOFF



Observation: This plot helps visualize the trade-off between sensitivity and specificity at various thresholds, supporting the decision to use a 0.4 cutoff for optimal fraud detection balance.

RECOMMENDATIONS

- **Deploy the Logistic Regression model** with a cutoff of **0.4** to prioritize fraud detection with high recall (87.5%).
- Use the model to **flag high-risk claims** for manual review, reducing financial loss.
- **Monitor false positives** and adjust the cutoff as needed to balance precision and recall.
- Integrate key predictors like **incident_severity**, **insured_hobbies**, and **claim ratios** into business rules.
- **Retrain the model periodically** using new claim data to maintain accuracy and relevance.

BUSINESS IMPLICATIONS

- **Reduced Fraudulent Payouts**

Leveraged early detection and automated flagging of suspicious claims to minimize financial losses from fraud.

- **Increased Operational Efficiency**

Focused manual reviews on high-risk claims, streamlining processes and reducing unnecessary workload.

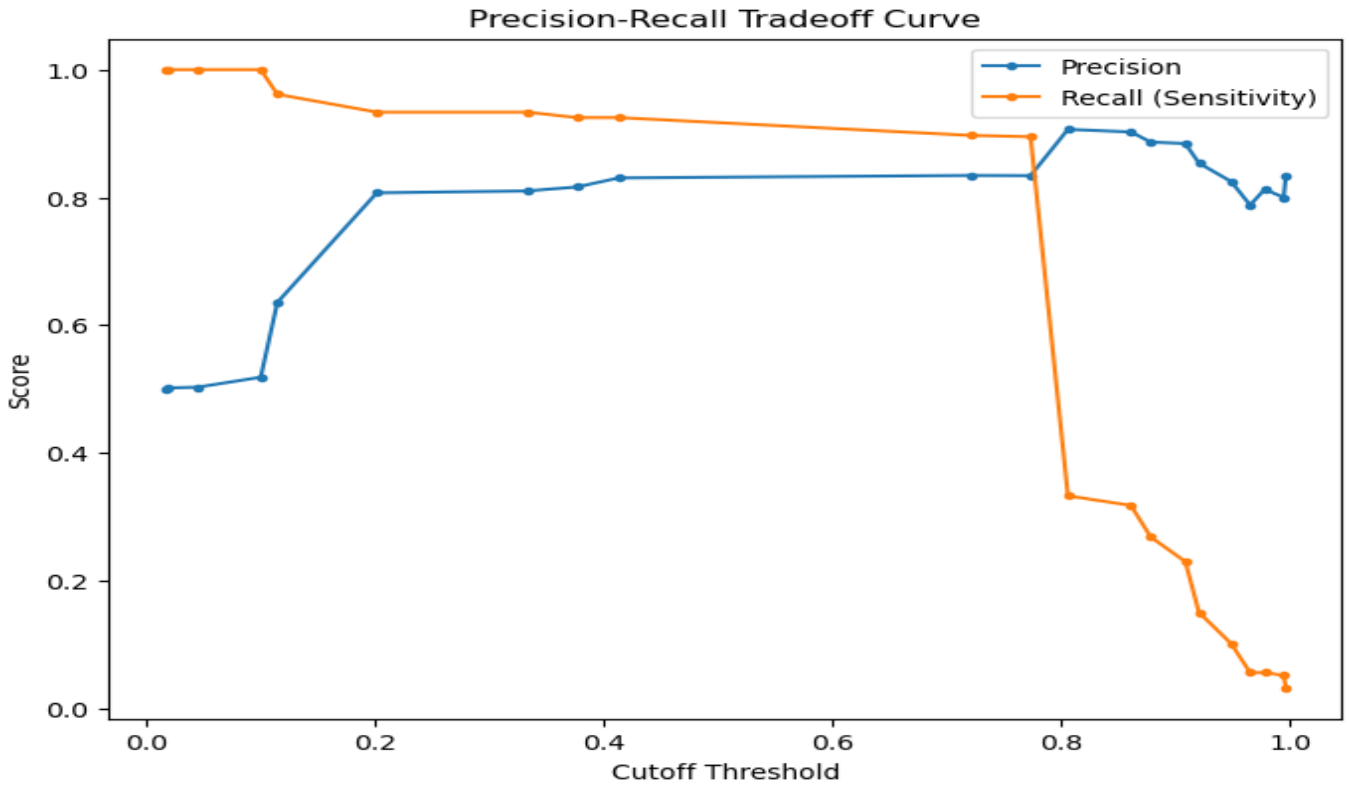
- **Improved Customer Experience**

Fast-tracked low-risk claims for quicker approvals and payouts, enhancing trust and satisfaction.

- **Data-Driven Risk Management at Scale**

Used predictive analytics to assign risk scores, enabling informed and consistent decisions across large claim volumes.

Precision vs. Recall Tradeoff



Observation: The precision-recall curve indicates strong recall performance even at lower precision, making it suitable for minimizing false negatives in fraud detection.

ANSWERS TO THE QUESTIONS

- How can we analyse historical claim data to detect patterns that indicate fraudulent claims?

We used **Logistic Regression** and **Random Forest** to analyze labeled historical claim data.

Key steps included:

- Selecting **important features** that contribute most to fraud detection.
- Applying **resampling techniques** to address class imbalance and ensure fair model training.
- Measuring performance using **recall, precision, and F1 score** to assess fraud detection accuracy.

By comparing both models' responses to past claims, we identified **features** which correlate with fraudulent activity. This helped refine fraud detection

- Which features are most predictive of fraudulent behaviour?

  From model training and importance analysis of the Logistic Regression model, the most predictive features included:

  - insured_hobbies_chess
  - insured_hobbies_cross-fit
  - insured_hobbies_video-games
  - incident_severity_Minor Damage
  - incident_severity_Total Loss
  - incident_severity_Trivial Damage
  - Auto_model_92x

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

Yes -

- Model Training:
  Logistic Regression and Random Forest were trained on historical claim data.
  Both models assign a fraud probability to incoming claims.

- Importance of Recall:
  In fraud detection, recall is crucial to minimize missed fraud cases.
  We preferred higher recall even if it meant a few false positives.

- Model Comparison:
  Logistic Regression (cutoff = 0.4) gave 87.5% recall.
  Random Forest had 73.6% recall but slightly higher precision.

- Final Choice – Logistic Regression:
  LR balances precision and recall with F1 score of 74.1%.
  It's simpler, interpretable, and ideal for deployment.

- What insights can be drawn from the model that can help in improving the fraud detection process?

  - Logistic Regression works better for fraud detection since it catches more fraudulent cases with higher recall.

  - Random Forest is more precise, but it tends to miss fraud cases, making it less useful in situations where missing fraud is risky.

  - Adjusting the cutoff from 0.5 to 0.4 for Logistic Regression improves fraud detection by striking a balance between recall and specificity.

  - The most important features identified can be used to improve screening and fraud detection systems.

THANK YOU