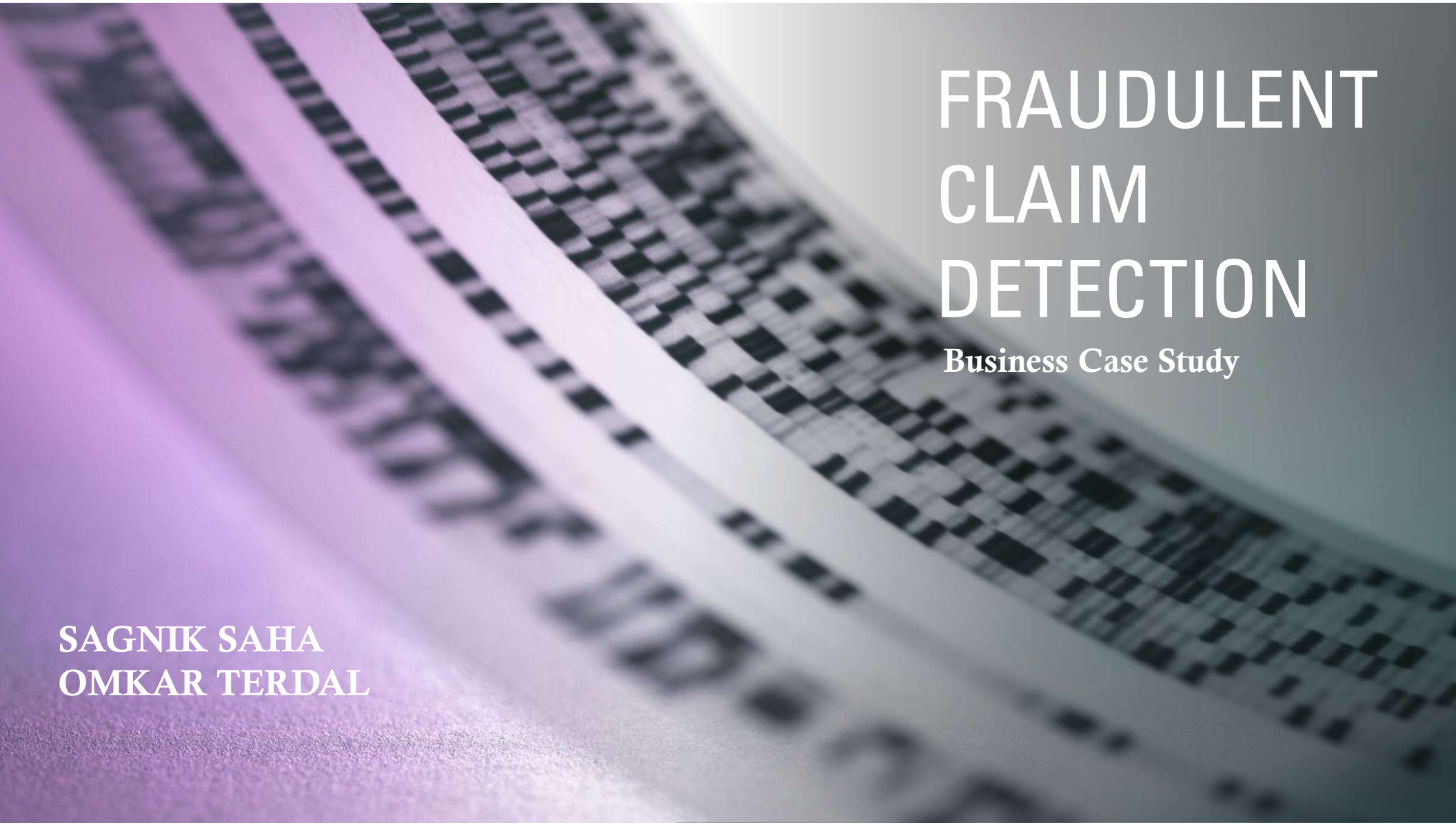


FRAUDULENT CLAIM DETECTION

Business Case Study

SAGNIK SAHA
OMKAR TERDAL

PROBLEM STATEMENT

Challenge Global Insure, a major player in the insurance industry, is experiencing substantial financial losses due to a high volume of fraudulent claims. The existing fraud detection system relies heavily on manual inspections, which are not only time-consuming and labor-intensive but also inefficient. As a result, many fraudulent claims are detected only after payouts have been made, limiting the company's ability to prevent losses and straining operational resources.

Objective To address this issue, Global Insure seeks to enhance its fraud detection capabilities by leveraging data-driven insights and advanced analytics. The goal is to implement an intelligent system that can accurately classify claims as fraudulent or legitimate at an early stage in the approval process. This proactive approach would help the company significantly reduce financial losses, improve the speed and accuracy of claims processing, and optimize overall efficiency in claims management

BUSINESS SUMMARY

- A data-driven approach to analyze historical claim records has revealed clear patterns associated with fraudulent behavior.
- Both logistic regression and random forest models demonstrated strong ability to identify fraudulent insurance claims, achieving validation accuracies above 80%. This enables early detection and intervention, reducing financial losses due to fraud.
- Features such as incident severity, insured hobbies, claim amounts (property, injury, vehicle), and customer/vehicle age were found to be highly predictive of fraud. Focusing on these variables can help streamline investigations and improve model interpretability.
- The models achieved a good balance between sensitivity (recall) and specificity, ensuring that most fraudulent claims are detected while minimizing false positives. This balance is crucial for operational efficiency and customer satisfaction.
- Logistic regression offers slightly better interpretability and similar performance compared to random forest, making it suitable for business environments where transparency is important. Random forest, however, can capture more complex patterns if needed.
- The analysis highlights the importance of continuous feature monitoring and periodic model updates. Integrating these models into claims processing can automate fraud detection, prioritize high-risk cases, and support data-driven decision-making for insurance operations.

MODEL EVALUATION

Logistic Regression

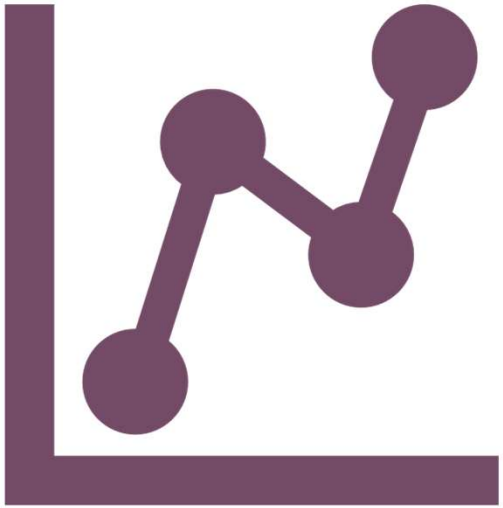
Cut off	Train data	Test data
0.58	Model Accuracy: 0.89 Sensitivity (Recall): 0.9 Specificity: 0.87 Precision: 0.87 F1 Score: 0.89	Model Accuracy: 0.84 Sensitivity (Recall): 0.78 Specificity: 0.85 Precision: 0.64 F1 Score: 0.7

Random Forest

Best estimator	Train data	Test data
max_depth=15, max_features=5, min_samples_leaf=10, min_samples_split=20, n_estimators=15	OOB Score: 0.85 Model Accuracy: 0.88 Sensitivity (Recall): 0.9 Specificity: 0.86 Precision: 0.87 F1 Score: 0.88	OOB Score: 0.82 Model Accuracy: 0.82 Sensitivity (Recall): 0.74 Specificity: 0.82 Precision: 0.58 F1 Score: 0.65

KEY QUESTIONS & INSIGHTS

1. **How can we analyze historical claim data to detect patterns that indicate fraudulent claims?**
   - Performing target likelihood analysis on categorical and numerical features.
   - Identifying anomalous behaviors using domain knowledge and statistical thresholds.
   - Leveraging Exploratory Data Analysis (EDA) and Visualization to find plot distributions of claim amounts, claim types, and other features for fraudulent vs.
   - non-fraudulent claims.
   - Use machine learning models (e.g., logistic regression, random forest) to predict fraud signals in combinations of features and get the important features which plays as a key indicator of fraud detection. historical claim data to detect patterns that indicate fraudulent claims?
2. **Which features are most predictive of fraudulent behaviour ?**
   - incident_severity
   - insured_hobbies
   - months_as_customer
   - vehicle_claim, property_claim, injury_claim
   - vehicle_age
   - vechile model and type (auto_make & auto_model)



KEY QUESTIONS & INSIGHTS

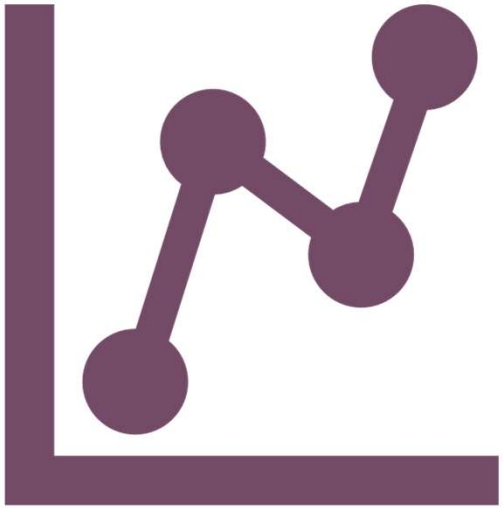
3. **Can we predict the likelihood of fraud for an incoming claim?**

   Yes — using models like Logistic Regression and Random Forest:

   - Claims can be scored in real-time based on fraud probability.
   - A cutoff threshold (e.g. 0.58) can be applied to flag high-risk claims for further review.
   - This enables automated early detection and reduces dependency on slow, manual inspections.

4. **What insights can be drawn from the model to improve fraud detection?**

   - Prioritize high-impact features to improve model performance and interpretability.
   - Remove low-importance features to simplify the model without sacrificing accuracy.



RECOMMENDATIONS & BUSINESS IMPLICATIONS

Recommendations

- Deploy the prediction model in the claims processing workflow to automatically flag high-risk claims for further investigation.
- Focus on feature importance to help investigators for enhanced fraud screening.
- Regularly retrain and validate models with new data to adapt to evolving fraud patterns.
- Monitor model performance metrics (accuracy, recall, precision, F1-score) and adjust thresholds as needed.
- Improve data collection for features with missing or ambiguous values (e.g., property_damage, police_report_available).
- Develop real-time dashboards for fraud monitoring and reporting.

Business implications

- Early detection and intervention can significantly reduce financial losses due to fraudulent claims.
- Automated flagging allows investigators to focus on the most suspicious cases, optimizing resource allocation.
- Robust fraud detection supports compliance with insurance regulations and audit requirements.
- Scalable and proactive fraud strategy that evolves with new data.

THANK YOU