Predicting Insurance Fraud with Machine Learning

This presentation outlines our comprehensive approach to building a robust fraud detection system for insurance claims. We will cover the project's objectives, methodology, key findings, and the performance of our predictive model.

Presentation by

Saloni Sharma
Shiva Shankar Iyer
Amitabh Kumar Biswas

Agenda

1

Project Overview

Understanding the critical need for fraud detection.

2

Business Objective

Defining the core goal: accurate claim classification.

3

Data & Approach

Exploring the dataset and challenges.

4

Exploratory Data Analysis (EDA)

Uncovering initial insights and relationships.

5

Predictive Modeling

Developing and evaluating the machine learning models.

6

Key Insights & Recommendations

Translating data into actionable strategies.

7

Results & Metrics

Detailed performance analysis of the model.

8

Conclusion & Next Steps

Summarizing impact and outlining future work.

Project Overview: Combating Insurance Fraud

Insurance fraud poses a significant threat, costing the industry billions annually and directly impacting policyholders through higher premiums. The financial and operational strain is immense, necessitating robust solutions.

Our goal is to leverage advanced machine learning to build a model that can **predict** fraudulent claims **before** they are fully processed. This proactive approach aims to minimize losses and streamline operations.

Business Objective: Data-Driven Claim Classification

Classify Claims

Accurately categorize each claim as **'Fraud'** or **'Legit'**.

Identify Patterns

Uncover key indicators and risk factors using historical data.

Enable Decision-Making

Facilitate faster, more informed operational decisions.

This objective ensures our solution directly addresses the need for efficiency and precision in fraud detection, moving beyond reactive measures to proactive prevention.

Data & Approach: Foundation of Our Model

Dataset Overview

We utilized a comprehensive historical insurance claims dataset, comprising various features relevant to claim processing.

- **Features:** Claim amounts, policyholder demographics, claim type (e.g., auto, property), incident severity, and more.
- **Target Variable:** Binary classification: 'Fraud' (Yes/No).

Challenges: Class Imbalance

A primary challenge was the significant class imbalance within the dataset:

- **75%** of claims were identified as non-fraudulent.
- **25%** of claims were confirmed as fraudulent.

This imbalance necessitated specific modeling techniques to avoid bias towards the majority class.

Exploratory Data Analysis (EDA): Uncovering Insights

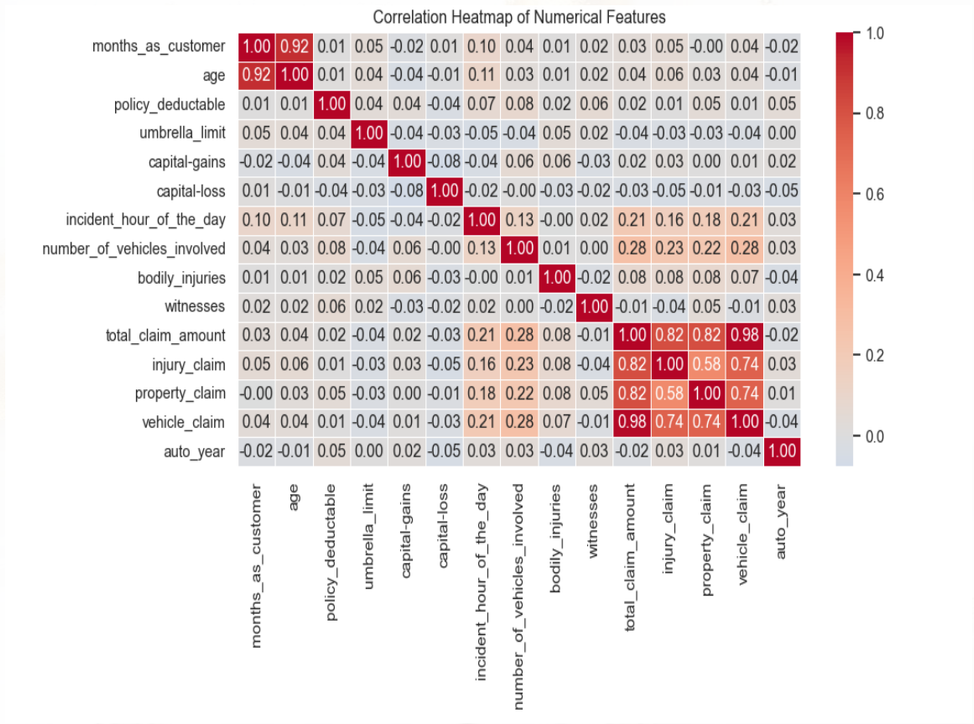
Methods Utilized

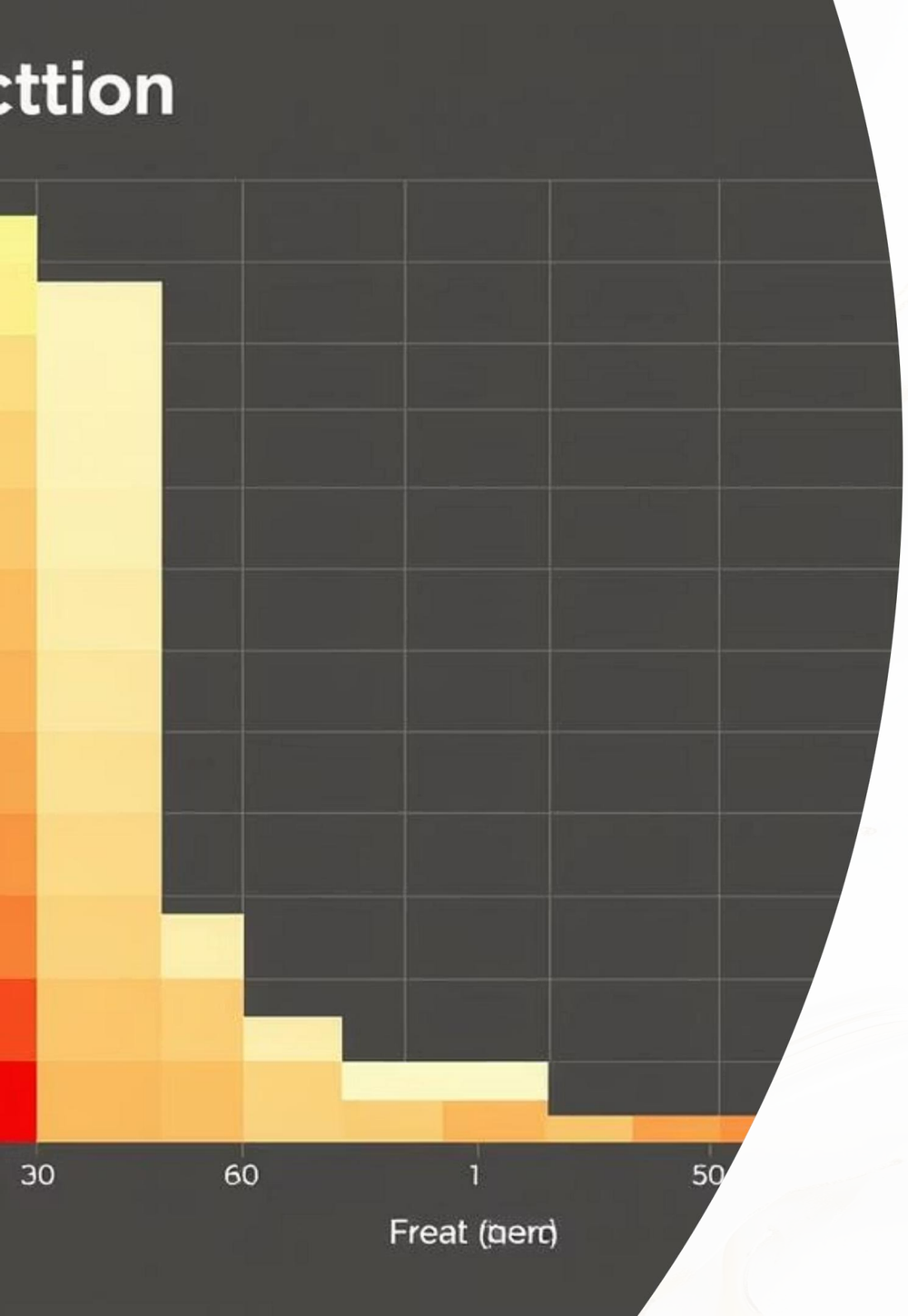
- **Visualizations:** Boxplots, countplots, and heatmaps to understand data distribution and relationships.
- **Correlation Analysis:** Identifying the strength and direction of relationships between variables.
- **Bivariate Relationships:** Examining interactions between pairs of variables.

Key Findings

- **High Claim Amounts:** Claims with unusually high financial values showed a strong correlation with fraudulent activity.
- **Severe Incidents:** Incidents classified as severe were more frequently associated with fraud.
- **Lack of Witnesses:** Claims reported without witnesses demonstrated a higher propensity for fraud.

Exploratory Data Analysis (EDA): Uncovering Insights





Important Features: Predictive Power

Our analysis identified several features as highly predictive of fraudulent claims. Understanding these drivers is crucial for both model development and operational focus.

1 **Financial Indicators**

Total claim amount, injury and property claim types.

2 **Incident Characteristics**

Incident severity, presence of witnesses and bodily injuries.

3 **Customer Profile & Behavior**

Insured hobbies, months as customer, incident hour, auto year.

Predictive Models & Evaluation

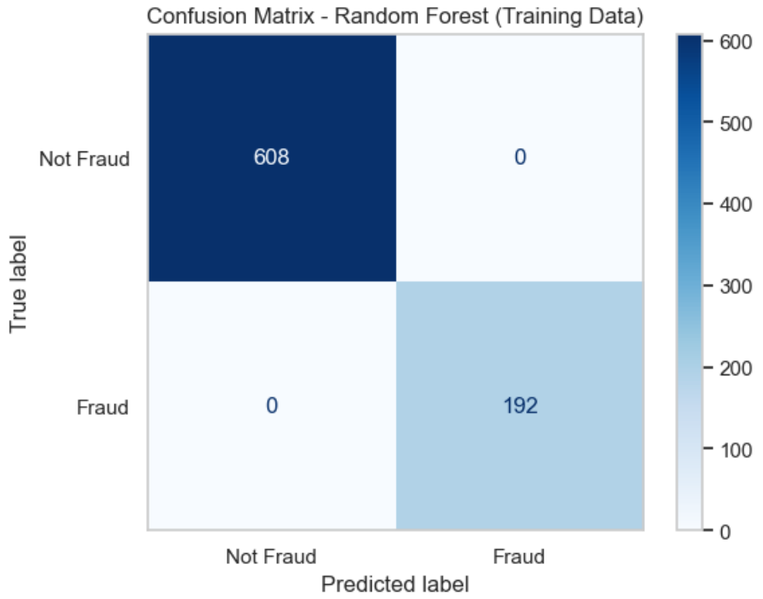
Models Evaluated

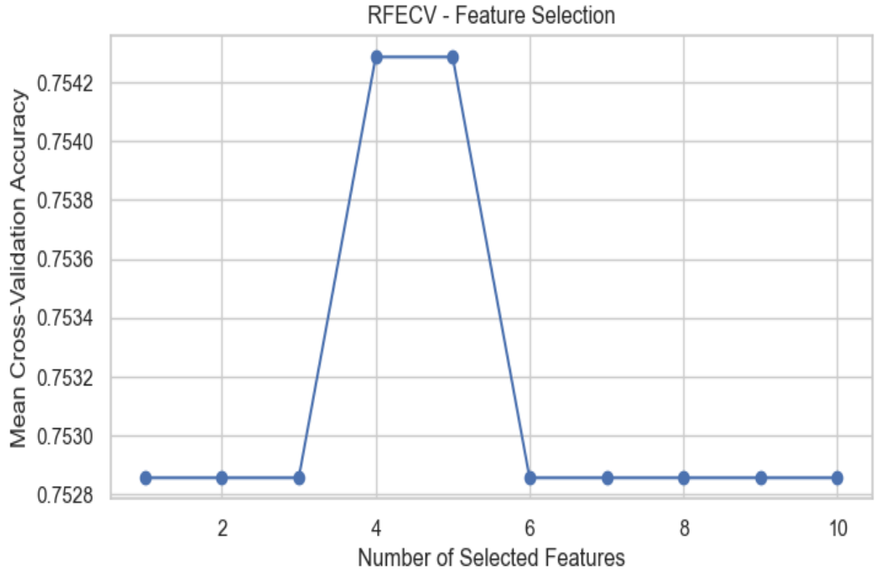
- **Logistic Regression:** Served as our robust baseline model.
- **Random Forest:** Utilized for its ability to handle complex relationships, with extensive hyperparameter tuning for optimal performance.

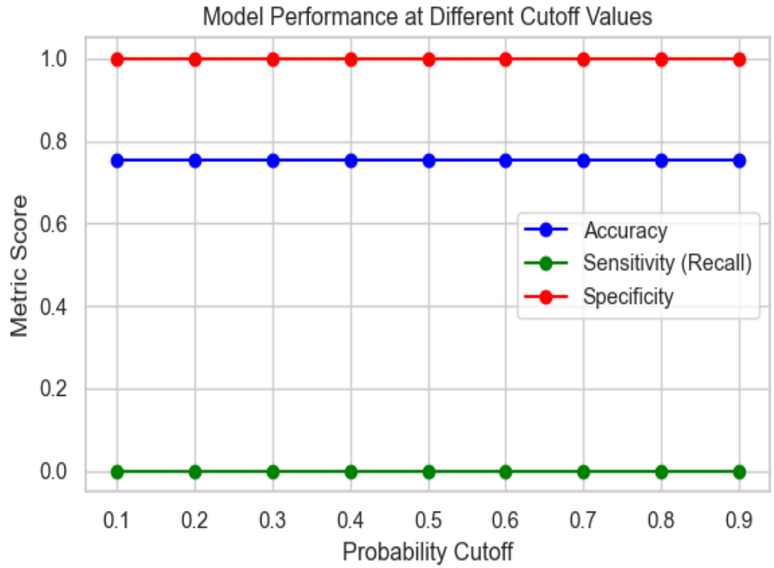
Evaluation Metrics & Threshold

- **Metrics:** ROC-AUC, Precision, Recall, and F1-Score were used to comprehensively assess model performance.
- **Optimal Threshold:** A threshold of **0.10** was chosen to align with the business goal of minimizing false negatives, ensuring genuine fraud cases are captured.

Predictive Models & Evaluation







Key Insights & Recommendations

1

High-Risk Indicators

- Claims with high amounts lacking sufficient supporting documents.
- Claims from new customers with limited history.
- Unusual hobbies or incident times that deviate from typical patterns.

2

Recommendations

- Assign a **risk score** to each claim to prioritize audits effectively.
- Focus investigative efforts on claims flagged by the **top features identified** by the model.
- Implement **automated alerts** for high-risk claims for immediate review.

Model Performance & Conclusion

Model Performance (Training Data)

- **Recall (Sensitivity):** 95.83%
- **Specificity:** 99.51%
- **Precision:** 98.40%
- **F1 Score:** 97.10%

The model demonstrates an excellent balance between catching fraudulent claims and minimizing false positives.

Conclusion & Next Steps

Our predictive model successfully detects insurance fraud with high precision and recall, significantly reducing reliance on manual reviews.

This solution promises improved operational efficiency and substantial cost savings. We are ready for deployment, with a plan for ongoing monitoring and tuning to maintain optimal performance.

Insurance_Claim_Fraud_Detection_Report_v1.pdf

Report attached for reference

Thank You.