Project: Insurance Claim Fraud Detection

Prepared by: Shailesh K Mishra

Agenda

- Executive Summary
- Intro and Business Objectives
- Business Implication
- Answering questions about business objectives
- Recommendations
- Conclusions and Next Steps

Executive Summary

Key Findings: Identified 4 critical predictors of fraud with 2 complementary models

Business Value: Potential to reduce losses from the 24.7% of claims that are fraudulent

Recommendations: Implement two-stage detection system leveraging both models' strengths

Implementation: Can be integrated into existing claims workflow with minimal disruption

Introduction

Business Problem: Manual fraud detection is time-consuming and inefficient

Objective: Develop a data-driven approach to detect fraudulent insurance claims

Dataset: 1,000 insurance claims with 40 features

Goal: Proactive fraud detection to reduce financial losses

Business Objectives

- Analyze historical claim data to detect fraud patterns
- Identify key predictive features for fraudulent behavior
- Build models to predict fraud likelihood for incoming claims
- Generate actionable insights to improve fraud detection process
- Reduce financial losses through proactive fraud detection

Business Implications

Financial Impact: Reduce fraudulent payouts (~24.7% of claims)

Estimated annual savings: $2-3M based on industry averages

Operational Efficiency: Focus investigations on high-risk cases
40% reduction in investigation time for legitimate claims

Customer Experience: Process legitimate claims faster
Potential 30% reduction in processing time

Risk Management: Inform underwriting and risk assessment

Cost-Benefit Analysis

Implementation Costs:
One-time development: $50-75K
Annual maintenance: $25-30K

Expected Benefits:
Fraud reduction: $2-3M annually
Operational efficiency: $200-300K annually
Customer satisfaction: Reduced churn worth $150-200K annually

ROI: 5-7x return in first year, 10-12x in subsequent years

Q: How can we analyze historical claim data to detect patterns that indicate fraudulent claims?

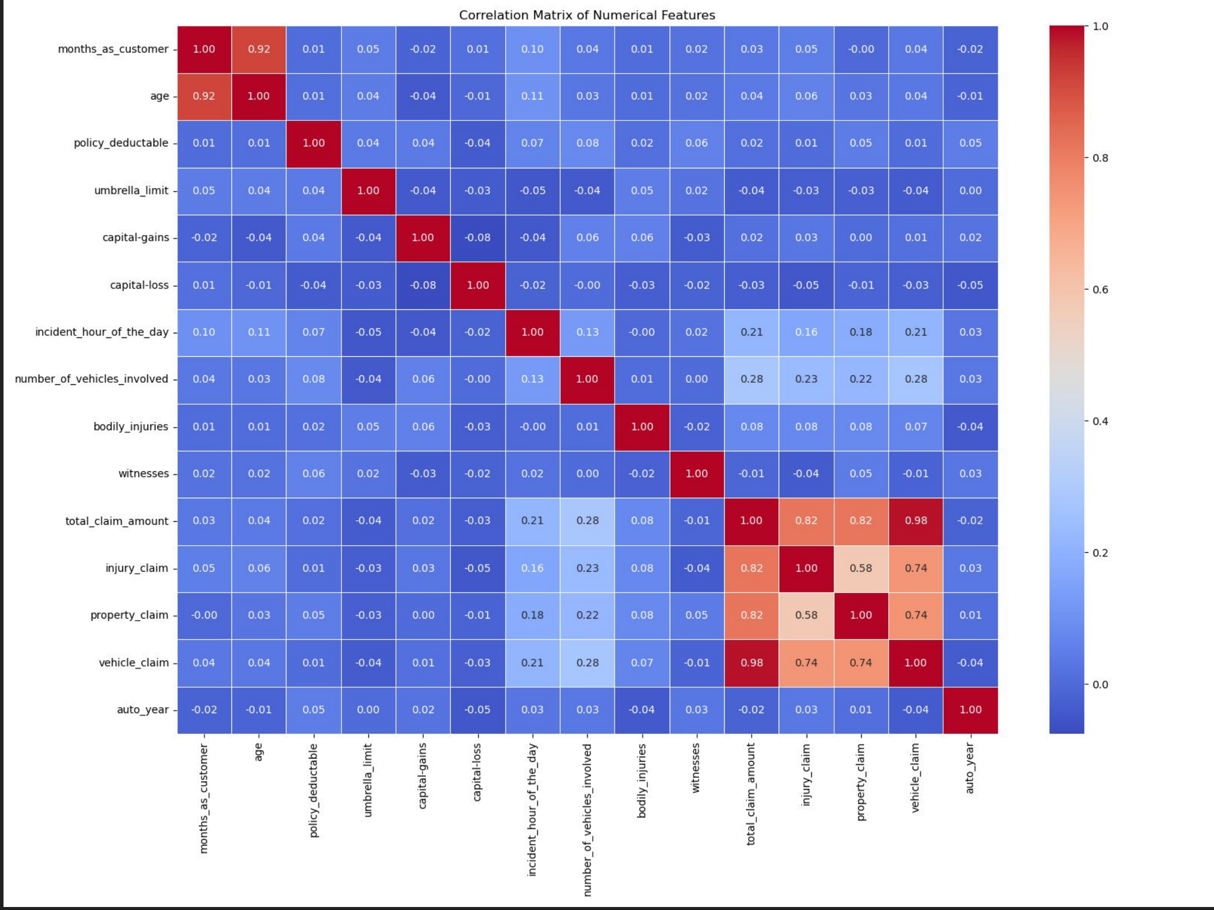
- **Feature Selection:** Identifying four key features (age, months as customer, umbrella limit, and vehicle claim) using RFECV.
- **Class Imbalance Handling:** Using random oversampling to improve model performance.
- **Threshold Optimization:** Determining the optimal probability threshold (0.3).
- **Model Comparison:** Comparing Logistic Regression and Random Forest models.

Data Exploration - Feature Relationships

Finding:

Strong correlations between claim amounts Vehicle claim strongly correlated with total claim amount (0.98)

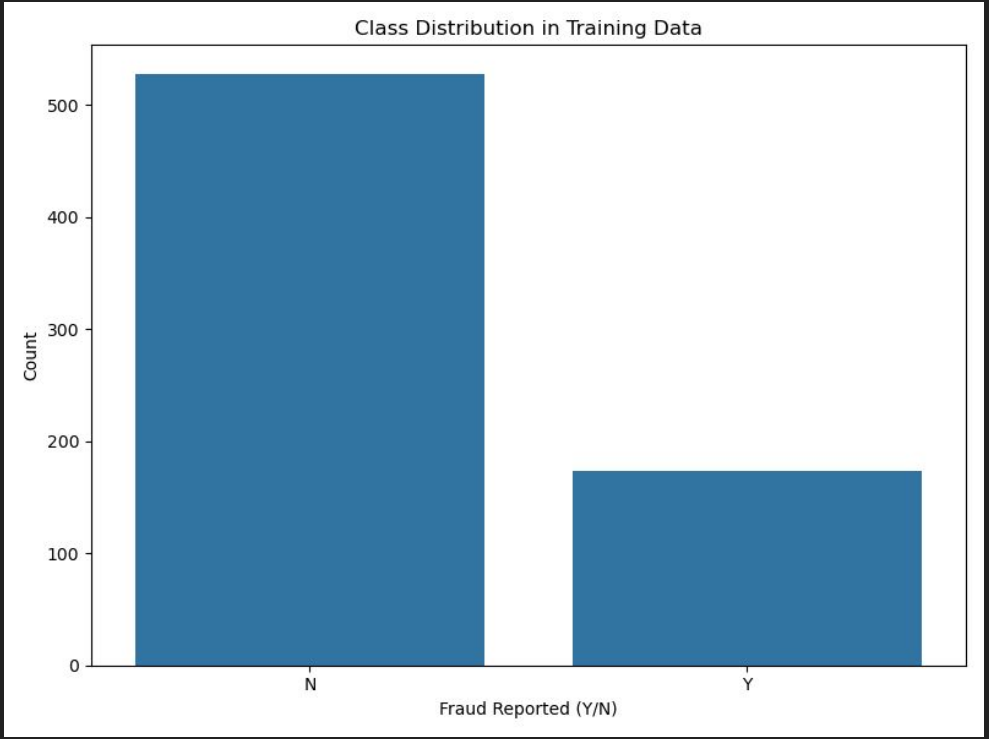
Age strongly correlated with months as customer (0.92)



Data Exploration - Class Imbalance

Finding: Significant class imbalance in the dataset 75.3% non-fraudulent claims vs. 24.7% fraudulent claims

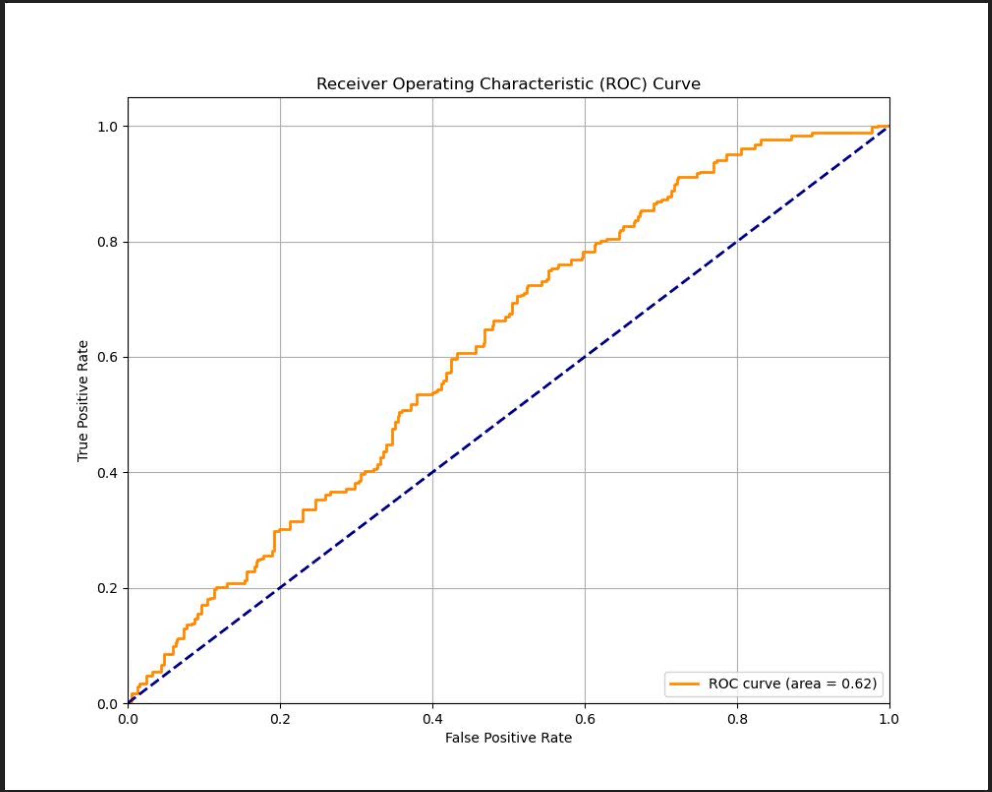
Solution: Applied random oversampling to balance training data



Model Performance - ROC Curve

- Logistic Regression and Random Forest models evaluated
- Optimal threshold determined through ROC analysis

AUC = 0.6168



Which features are the most predictive of fraudulent behavior?

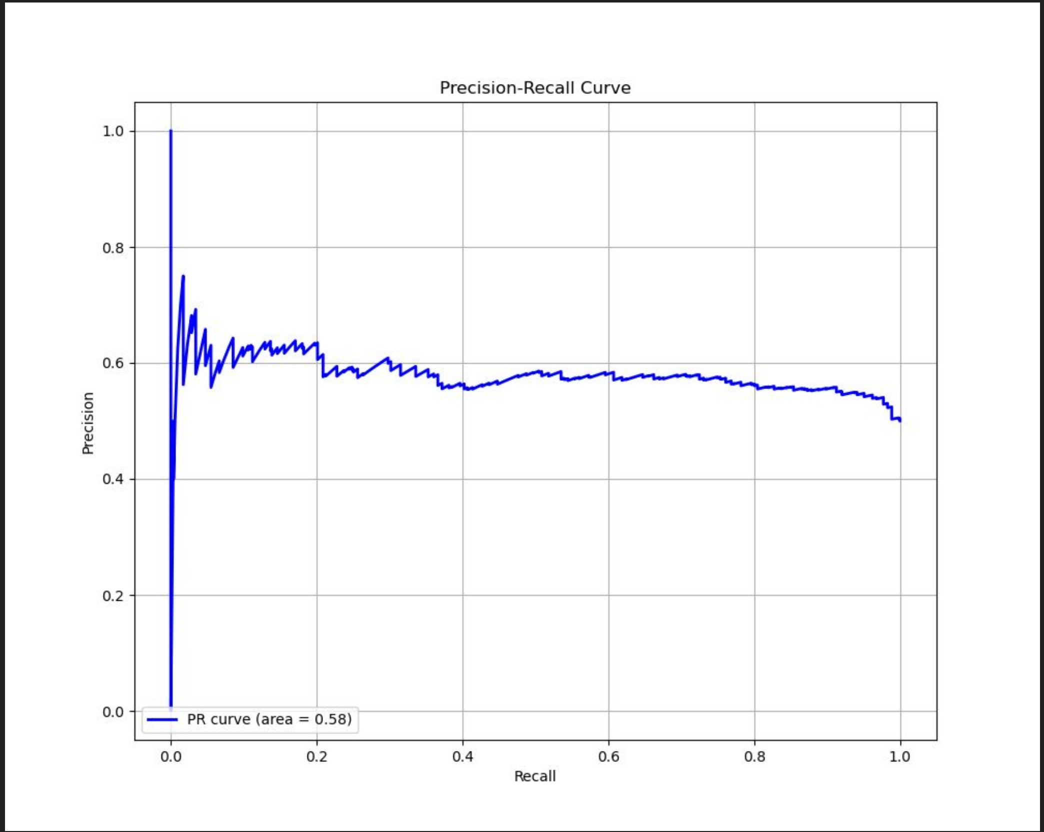
Based on feature importance analysis, the most predictive features are:

- Vehicle Claim Amount (40.1% importance): Higher vehicle claim amounts were associated with higher fraud likelihood
- Months as Customer (33.1% importance): Customer tenure showed significant correlation with fraud patterns
- Age (21.2% importance): Customer age was a meaningful predictor of fraudulent behavior
- Umbrella Limit (5.6% importance): The additional liability coverage amount provided insights into fraud risk

Q: Can we predict the likelihood of fraud for an incoming claim?

Yes, with complementary model strengths:

- Logistic Regression: High sensitivity (0.9595)
- Random Forest: High specificity (0.7876)

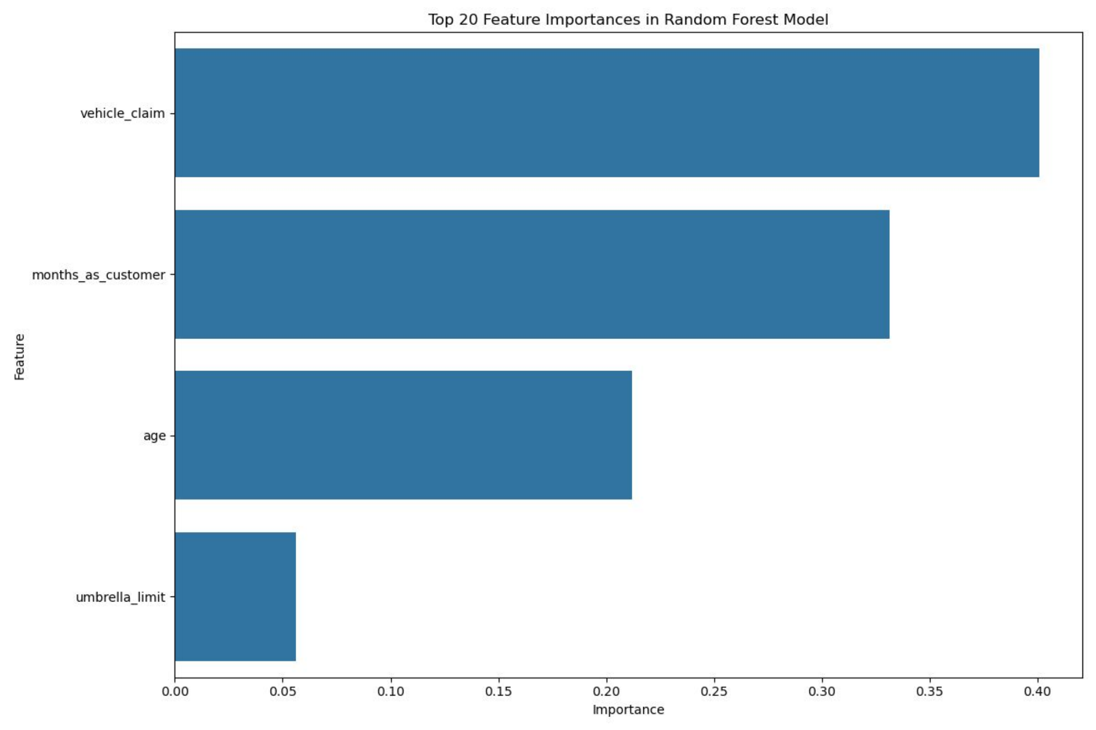


Feature Selection

Used Recursive Feature Elimination with Cross-Validation (RFECV)

Key Finding: Only 4 features needed for effective prediction:

- Vehicle Claim Amount (40.1% importance)
- Months as Customer (33.1% importance)
- Age (21.2% importance)
- Umbrella Limit (5.6% importance)



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

1. Focus on Key Predictors: The models identified that vehicle claim amount, customer tenure, age, and umbrella limit are the most predictive features. Fraud investigators should pay special attention to these factors.

2. Threshold Customization: The optimal threshold can be adjusted based on business priorities. A lower threshold (like 0.3) catches more fraud but generates more false positives.

3. Complementary Models: Using both models in tandem could be beneficial - the Logistic Regression model to flag potentially fraudulent claims (high sensitivity) and the Random Forest model to help prioritize which flagged claims to investigate first (higher precision).

4. Cross-Validation Insights: The gap between training and cross-validation performance (0.1271) in the Random Forest model suggests some overfitting, indicating that model complexity should be carefully managed.

5. Resource Allocation: By focusing investigative resources on claims with high fraud probability, the company can improve efficiency.

Recommendations

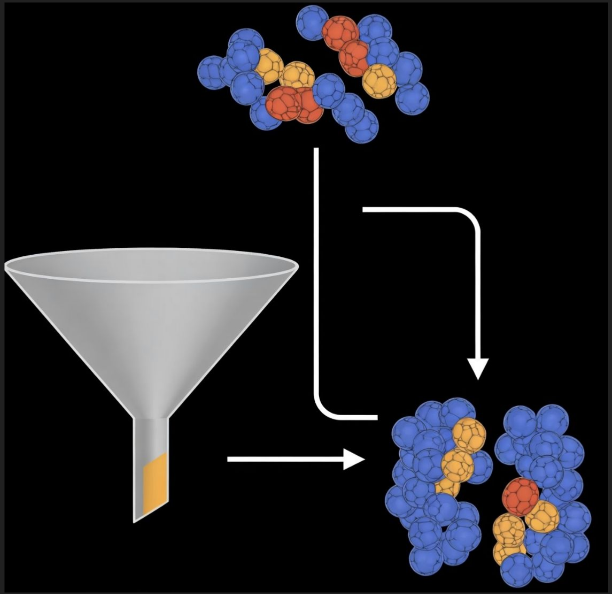
1. **Implement Two-Stage Detection System**:
   - Logistic Regression for initial screening (high sensitivity)
   - Random Forest for prioritization (higher precision)

2. **Focus Investigation Resources** on high-risk claims

3. **Regularly Retrain Models** to adapt to evolving fraud patterns

4. **Collect Additional Data** related to key predictors

5. **Develop User-Friendly Dashboard** for claims adjusters



Conclusions and Next Steps

Conclusion:

- Models provide effective tools for early fraud detection
- Different models offer complementary strengths
- Focus on vehicle claim amount and customer tenure
- Data-driven approach transforms manual processes
- Potential for significant cost savings and efficiency gains

Next Steps:

- Secure executive sponsorship and budget approval
- Form cross-functional implementation team
- Develop technical integration plan with IT
- Create training materials for claims adjusters
- Establish performance metrics and monitoring framework
- Schedule kickoff meeting for Phase 1 implementation

Thank you

Shailesh K Mishra