From Data to Decision: A Predictive Approach to Managing Claim Risk

A Data-Driven Risk Assessment Study

Department of Quantitative Risk Analysis and Management

MSA8010: Data Programming

Instructor: Dr. Brian Albert Monroe

Submitted by:

Venkata Sai Prasad Kancharana

Georgia State University | J. Mack Robinson College of Business

A Practical Tool for Proactive Risk Management

The Challenge

2.4%

Claims represent a 'needle in a haystack' event across our 79,883 policies, making proactive identification difficult.

The Solution

We have developed predictive models that successfully identify high-risk policies, flagging up to **61% of potential** claims for early review.

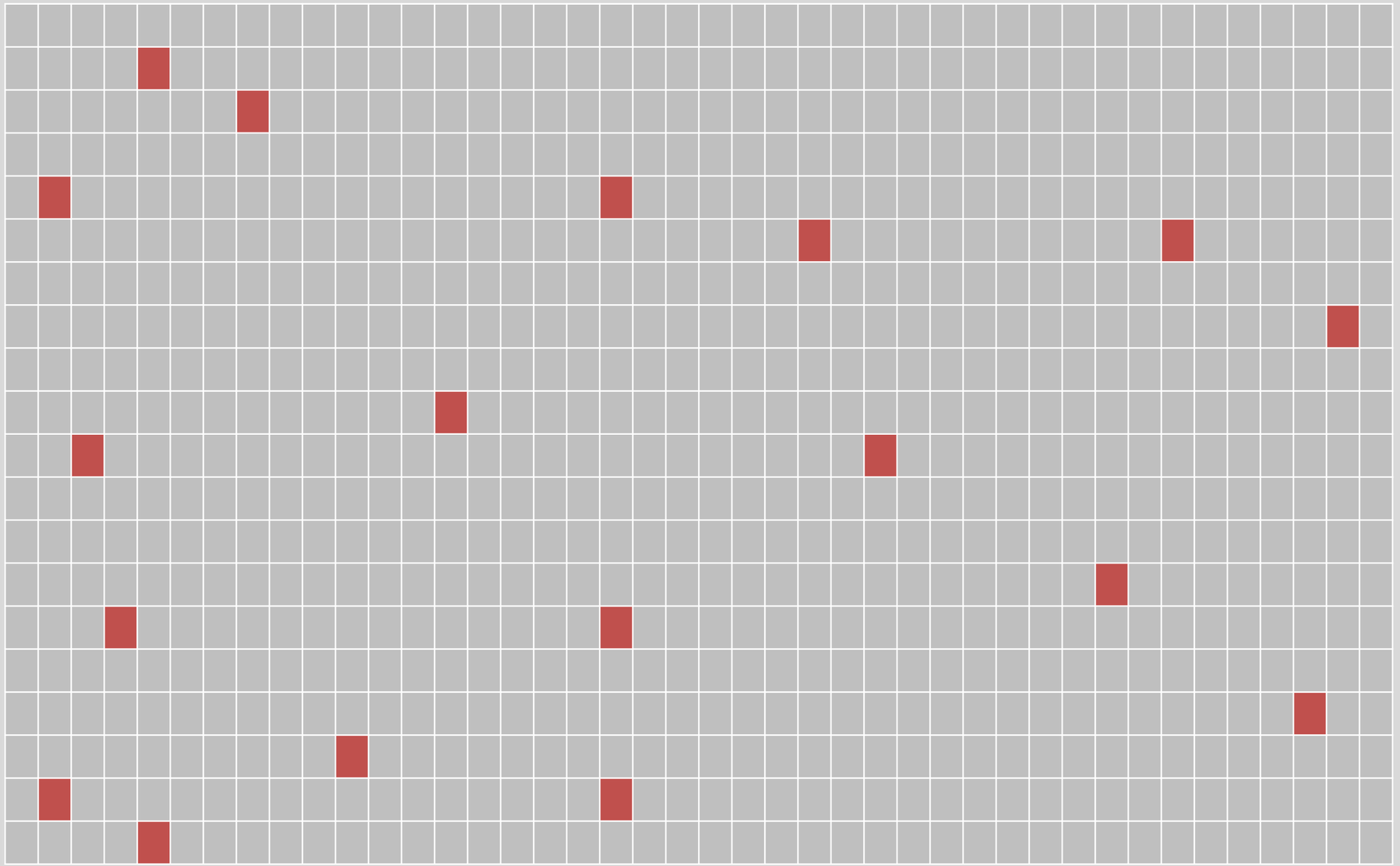
The Strategic Trade-off

The models are designed for **triage and prioritization,** not automation. **Low precision (~4-7%)** is an expected structural outcome of the low claim rate, resulting in a high number of false alerts.

Our Recommendation

We recommend immediate pilot implementation in **Underwriting** (to create high-risk watchlists) and **Claims** (to triage investigation queues).

Our Core Challenge: Finding the 2.4% of Policies with Claims



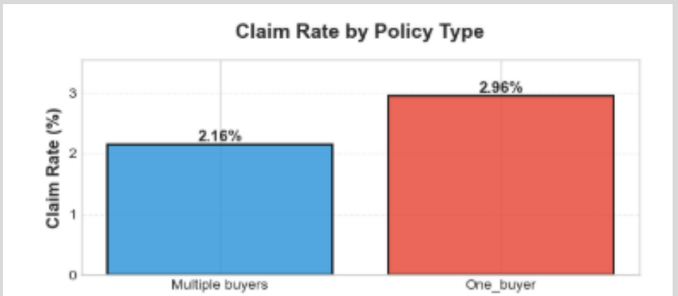
1,885

claims out of 79,883 total policies

2.4%

Overall Claim Frequency

One-Buyer Policies are ~37% Riskier



How Predictive Insights Create Value Across the Business

Underwriting Risk Segmentation

Identify high-risk policy characteristics ***before*** binding coverage to enable smarter pricing, documentation, and diligence.

Claim Investigation Triage

Rank incoming cases by their predicted risk, allowing claims handlers to focus their initial efforts where it matters most.

Portfolio Steering

Understand which seller industries, policy structures, and coverage types contribute disproportionately to claim risk, informing long-term portfolio strategy.

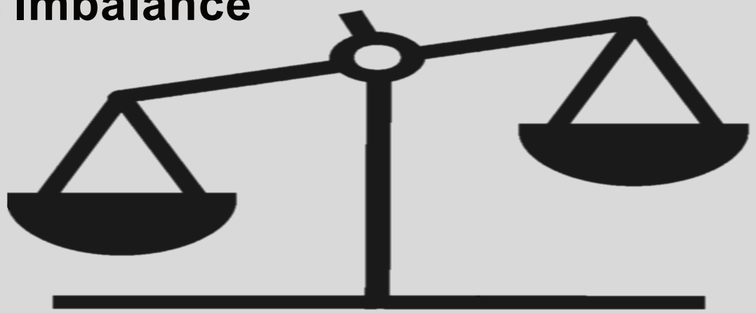
Our Approach: Building a System to Learn from Rare Events

The Goal

To build a binary classification model that **accurately predicts which policies will have a claim ('Has_Claim = 1').**

The Key Technical Challenge: Class Imbalance

"No Claim" - **62,398 (97.64%)** of training data.

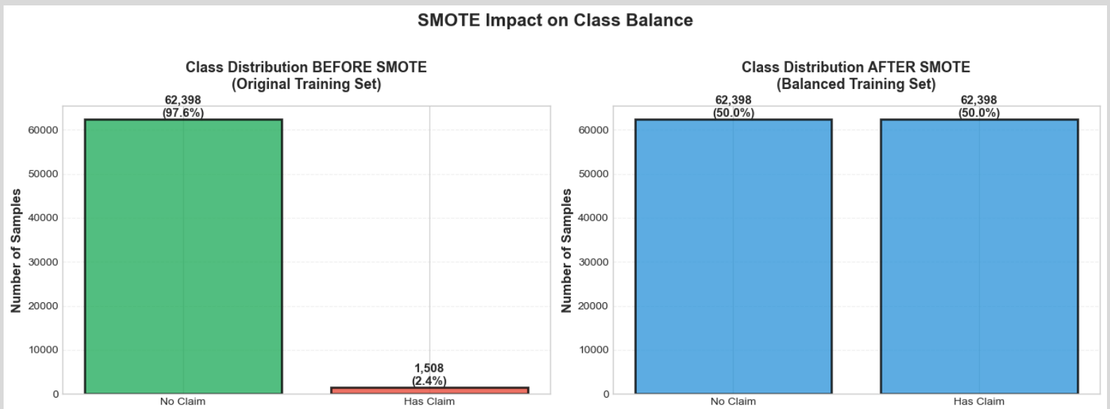


"Has Claim" - **1,508 (2.36%)** of training data.

Without adjustment, a model could achieve >97% accuracy by simply guessing 'No Claim' every time, making it useless for finding risk.

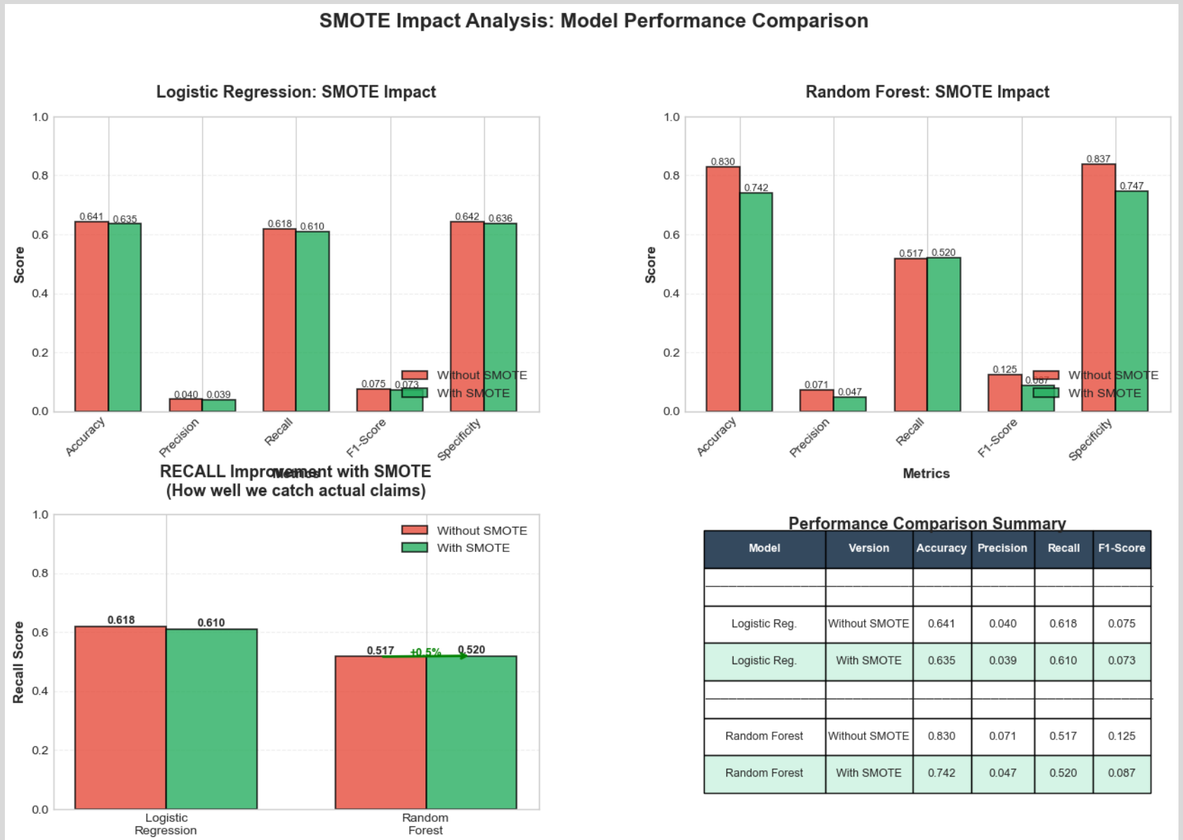
The Solution: SMOTE Resampling

We used the SMOTE technique to synthetically balance the training data, forcing the models to learn the patterns of the rare claim events.



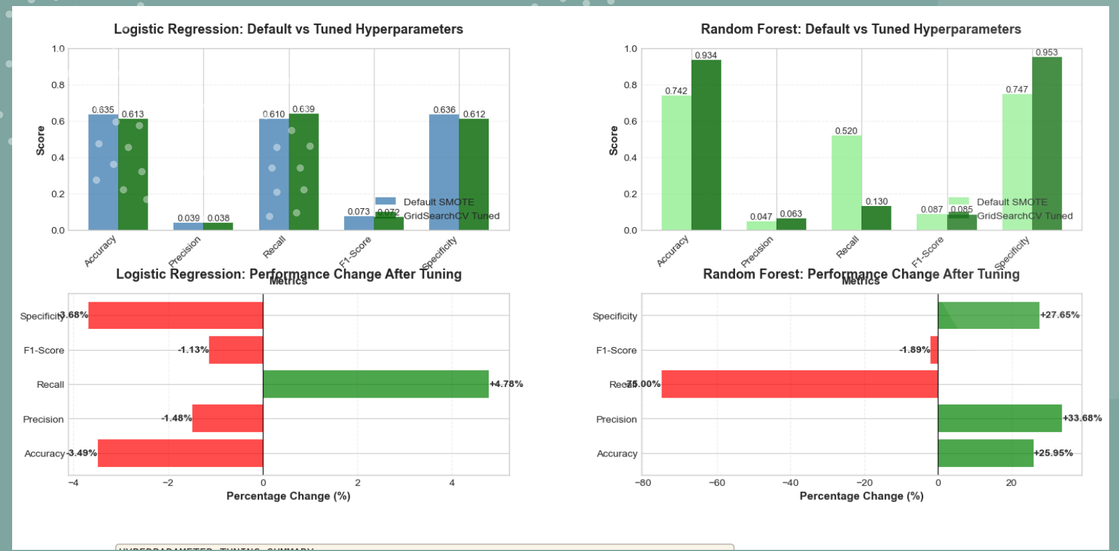
Model Performance is a Strategic Choice

We have two tools, each optimized for a different business objective.



	Logistic Regression	Random Forest
Recall	**61.8%** Catches nearly 2 out of every 3 claims	**51.7%** Stills catches nearly half claims
Key Trade-off (Precision)	**4.0%** (Generates more false alerts)	**4.7%** (Generates fewer false alerts compared to Regression)
Best Use Case	Claims triage where the top priority is to ensure no high-risk claim goes unnoticed.	Underwriting review or claims triage in a resource-constrained environment where minimizing false positives is critical.

A Warning: Optimizing for Accuracy Can Mask Significant Risk



Standard hyperparameter tuning aimed at improving overall accuracy makes the model useless for its primary business function.

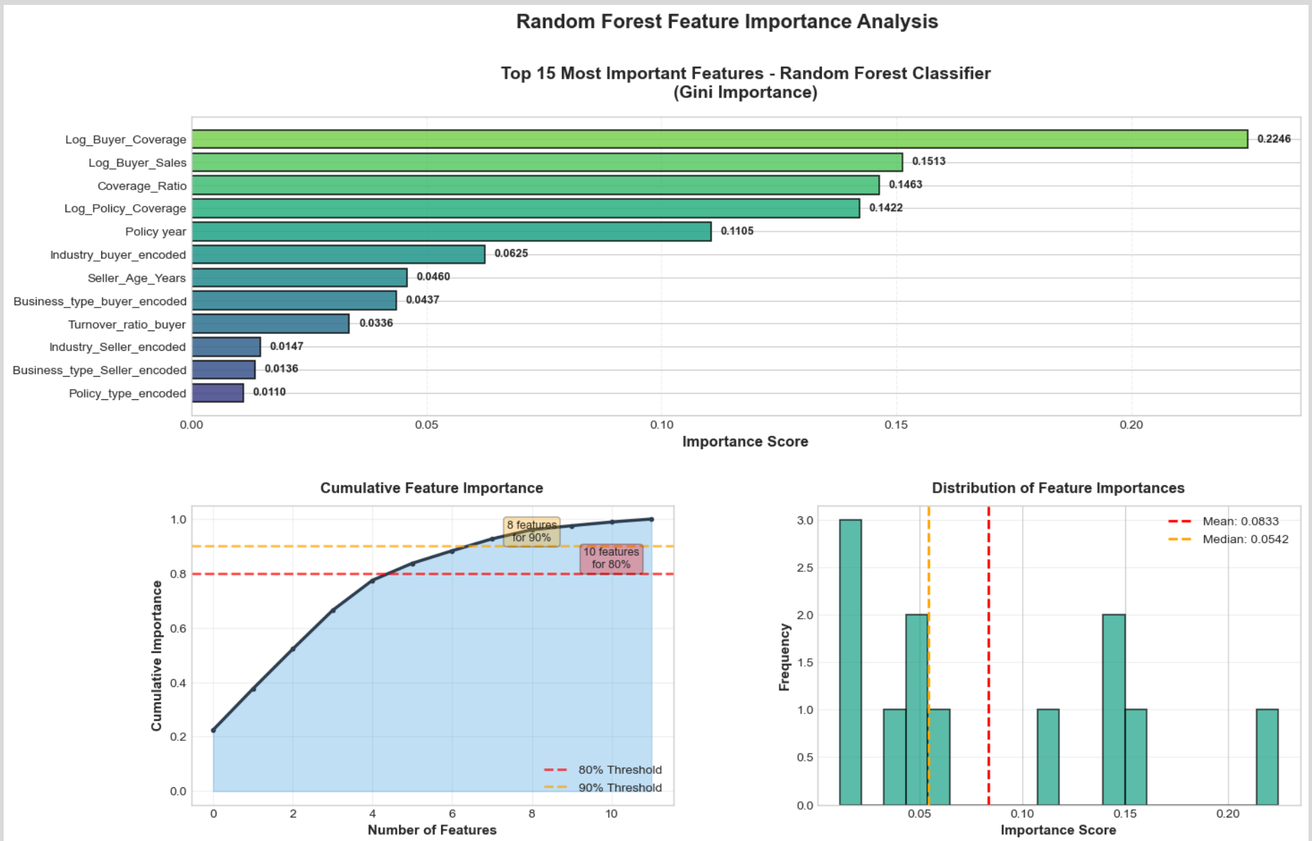
```
Logistic Regression - Best Parameters:
  • C: 0.001
  • max_iter: 1000
  • penalty: l1
  • solver: liblinear
  • Best CV F1-Score: 0.6420

Random Forest - Best Parameters:
  • bootstrap: True
  • max_depth: None
  • max_features: sqrt
  • min_samples_leaf: 1
  • min_samples_split: 2
  • n_estimators: 200
  • Best CV F1-Score: 0.9635

KEY INSIGHTS FOR STAKEHOLDERS:
• GridSearchCV systematically tested multiple parameter combinations using cross-validation
• Logistic Regression: Tested 24 combinations
• Random Forest: Tested 24 combinations
• Tuning optimized models for F1-Score, balancing precision and recall
• Average performance change - Logistic Regression: -1.00%
• Average performance change - Random Forest: +2.08%
```

What the Model Taught Us: The Key Drivers of Claim Risk

- **High Policy & Buyer Coverage** Log_Policy_Coverage (+0.5310) and Log_Buyer_Coverage (+0.4410) are the strongest signals.
- **High Coverage-to-Sales Ratio** Coverage_Ratio (+0.5068) indicates potential over-insurance.
- **Seller Characteristics**
  Specific seller business types and industries show higher baseline risk.
- **Established Seller Age**
  Older, more established sellers are less risky (Seller_Age_Years -0.1936).
- **High Buyer Sales**
  Financially strong buyers are a positive signal (Log_Buyer_Sales -0.2005).
- **Policy Year**
  Policies written in 2016 were slightly less risky than those in 2015 (Policy year -0.0481).



Action Plan: Recommendations for Underwriting

Implement a High-Coverage Watchlist

What

Automatically flag policies with top-decile risk **scores** for enhanced manual review before binding.

Why

The model's strongest predictors of risk are 'Policy Coverage','Buyer Coverage', and the 'Coverage-to-Sales Ratio'. This focuses review efforts on the highest-leverage policies.

Segment and Price by Policy Structure

What

Treat one-buyer policies as a distinct, higher-risk segment. Consider adjusted pricing, stricter documentation requirements, or lower capacity limits.

Why

One-buyer policies have a demonstrably **higher claim frequency (2.96%)** compared to multi-buyer policies **(2.16%).**

Action Plan: Recommendations for Claims Operations

Triage and Prioritize, Do Not Automate Decisions

What

Use the **model's risk score** to **rank** the daily queue of incoming claim notifications.

Do **not use the model** for automated claim denials.

Why

Low precision is an expected and unavoidable feature given the **2.4% base claim rate**. The model's value is in providing a **powerful signal for prioritization**, not a definitive judgment.

Choose the Right Tool for the Job

What

Deploy the appropriate model based on the team's current priority.

- If the goal is... **"Maximize Claim Detection"**: Use the **Logistic Regression** model (Recall:::! 61%).
- If the goal is... **"Reduce Investigator Workload"**: Use the **Random Forest** model (Fewer False Positives).

Roadmap for Governance and Continuous Improvement

Evolve Performance Metrics Beyond Accuracy

Shift primary model tracking from simple Accuracy to **Precision-Recall curves** and **PR-AUC.** These metrics are more appropriate for imbalanced datasets and align better with business value.

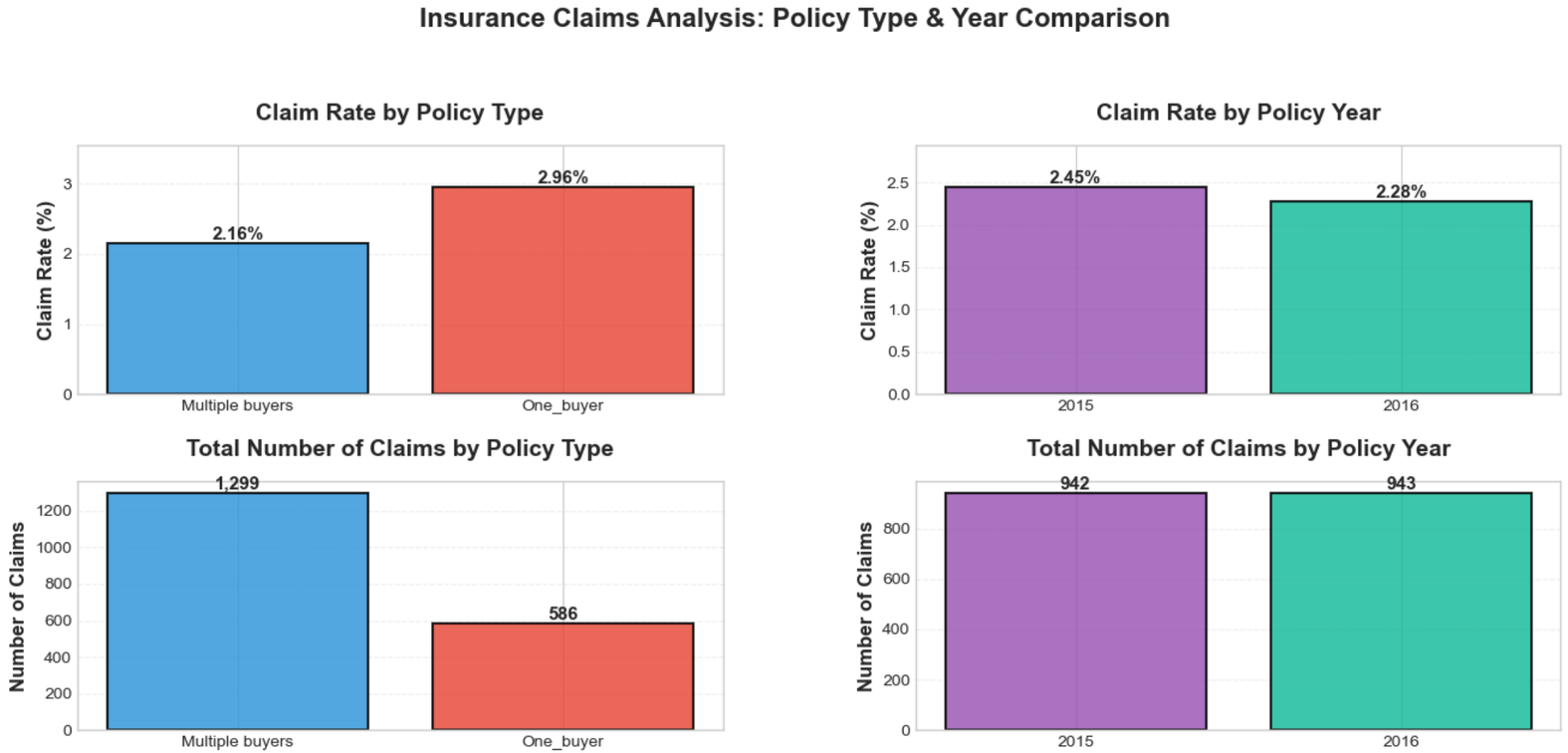
Introduce Cost-Based Evaluation

Work with business units to define the financial cost of a **missed claim** versus the operational cost of investigating a **false alarm.** Use this to find the optimal risk score threshold.

Monitor for Performance Degradation

Continuously validate model performance across key segments (e.g., by seller industry, policy year) to identify any hidden performance cliffs or model drift over time.

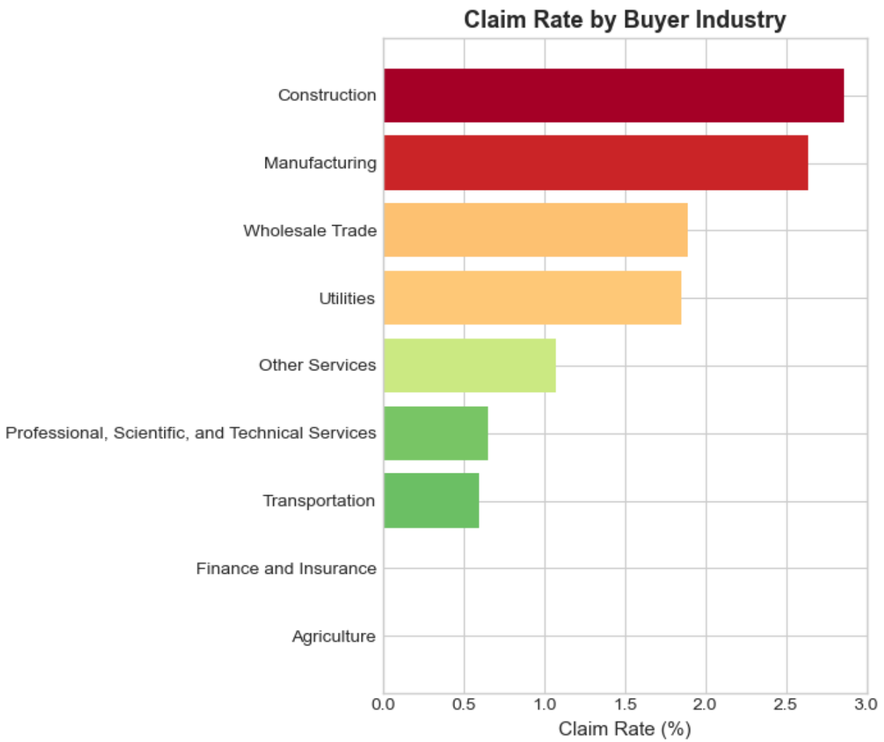
Appendix A: Portfolio & Composition

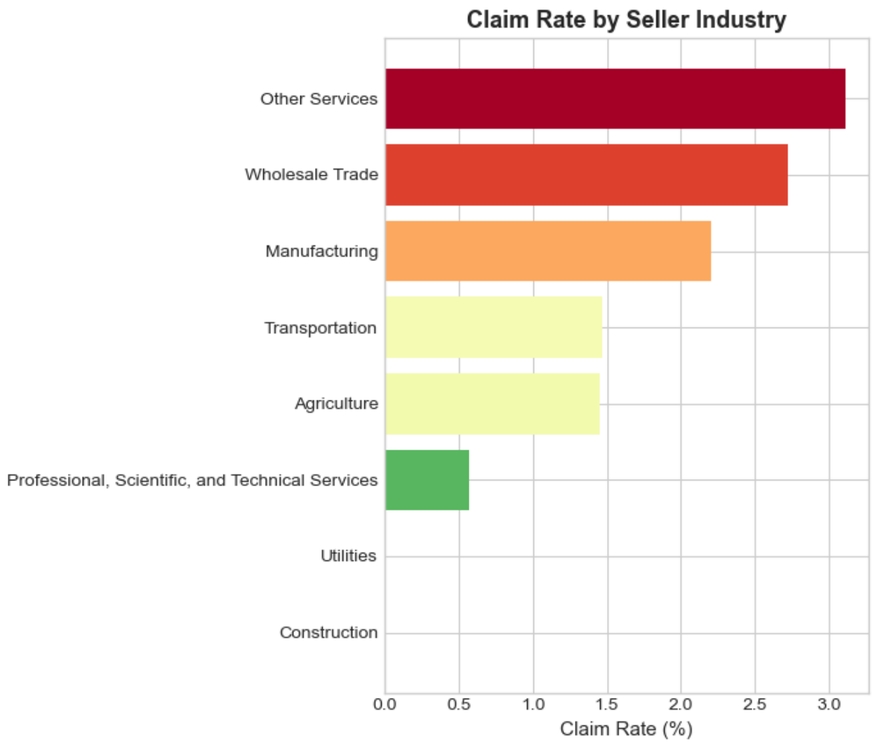


Claims Summary Table

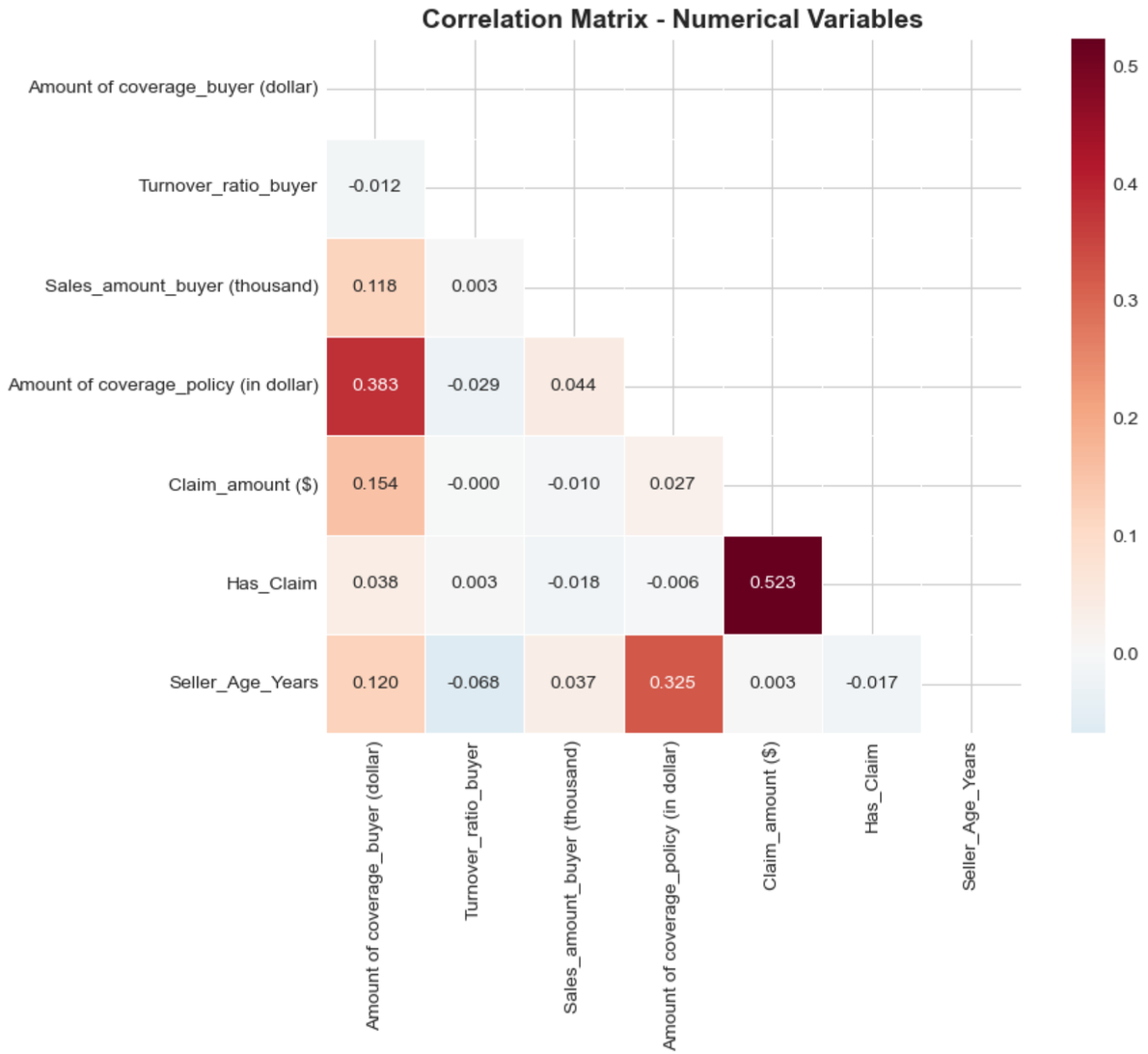
	Total Policies	Claims Filed	Claim Rate	Total Claim Amount
Multiple buyers	60,056	1,299	2.16%	$97,925,667.93
One_buyer	19,827	586	2.96%	$30,402,258.85
Year 2015	38,455	942	2.45%	$64,356,042.51
Year 2016	41,428	943	2.28%	$63,971,884.27

Appendix A: Portfolio & Composition





Correlation Matrix



Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression (No SMOTE)	0.641	0.040	**0.618**	0.075
Logistic Regression (With SMOTE)	0.635	0.039	0.610	0.073
Random Forest (No SMOTE)	0.830	**0.071**	0.517	**0.125**
Random Forest (With SMOTE)	0.742	0.047	0.520	0.087

Confusion Matrix: Logistic Regression (No SMOTE)

TP:230	FP: 5,683
FN: 147	TN: 9,917

Confusion Matrix: Random Forest (With SMOTE)

TP: 196	FP: 3,948
FN: 181	TN: 11,652