Policy Risk Predictor

Bryan Guin (bg478), Cornelius Schramm (cs2538), Jannik Obenhoff (jo456)

January 26, 2025

Abstract

The insurance industry faces persistent challenges in accurately predicting claims and providing interpretable explanations for these predictions. This study proposes an AI framework that integrates classical machine learning (ML) techniques with large language models (LLMs) to predict insurance claims and generate human-readable explanations. Using the Kaggle Car Insurance Claims dataset, we address key challenges, including class imbalance and high-cardinality features. LLMs are employed to transform unstructured claim descriptions into structured representations and to provide natural language explanations. Experimental results demonstrate the efficacy of combining LLMs and classical ML models in enhancing predictive performance and explainability, offering a robust and interpretable solution for decision-making in insurance.

1 Introduction

The insurance industry plays a critical role in managing financial risk, but it faces challenges in balancing predictive accuracy with interpretability. Reliable claim predictions enable insurers to optimize risk assessment, pricing strategies, and fraud detection, while transparent explanations foster trust among stakeholders. Traditional machine learning models are well-suited for structured data but often lack the capability to process unstructured inputs or provide human-readable justifications. On the other hand, large language models (LLMs) excel in handling unstructured data and generating natural language outputs but require integration with structured prediction frameworks for end-to-end solutions.

This project explores an innovative AI system that combines the strengths of classical ML and LLMs to predict insurance claims and explain the predictions. The system processes user-provided natural language descriptions of claims, transforms them into structured feature vectors using LLMs, and predicts the claim likelihood using classical ML models. In addition, the system generates natural language explanations, improving transparency and usability for both insurers and policyholders.

Our work focuses on the Kaggle-provided Car Insurance Claims dataset, aiming to address challenges such as significant class imbalance, high-cardinality features, and the need for clear and interpretable results. Through a systematic approach involving data preprocessing, feature engineering, and model evaluation, we benchmark our system's performance against industry standards like CatBoost, highlighting areas of success and identifying opportunities for improvement.

2 Background

Machine learning techniques, including gradient boosting machines and neural networks, excel in predictive tasks but often function as "black boxes," limiting interpretability. Explainable AI (XAI) methods aim to address this limitation by offering insights into model decisions, though they are typically applied post hoc rather than being integrated into the model pipeline. Large language models (LLMs), such as ChatGPT, have advanced natural language understanding, making them valuable for extracting features from unstructured text. However, LLMs are resource-intensive and less suited to structured data tasks. By integrating LLMs with classical ML techniques, we aim to create a system that combines the interpretability of traditional methods with the natural language understanding of LLMs.

Our dataset exemplifies real-world challenges faced in insurance analytics. It includes rich features such as policyholder demographics, vehicle specifications, and safety indicators, but exhibits significant class imbalance, with non-claims heavily outweighing claims. Addressing this imbalance is crucial for building fair and effective models. By leveraging techniques like SMOTE and class weighting, along with explainability-focused LLM integration, our project tackles these challenges to create a system that is both effective and user-centric.

3 Method

Insurance claim prediction typically involves developing classification models that differentiate between two classes: "Claim" and "No Claim." In highly imbalanced settings, most traditional metrics can present a skewed view of model quality. For this reason, our methodological focus is on:

1. **Class Imbalance Mitigation**: Using techniques like class weighting and synthetic oversampling (SMOTE).
2. **Minority-Class Metrics**: Prioritizing metrics that reveal the true performance on the minority class (Claim), rather than relying on aggregate metrics like Weighted F1. While Weighted F1 is commonly used in imbalanced scenarios, it can still be dominated by the majority class performance. We therefore consider the F1 score of the minority class (Claim) itself to better capture whether the model is identifying claims effectively.
3. **Model Comparison**: Testing baseline models (Logistic Regression, Random Forest, CatBoost) with both weighting and SMOTE, and then extending to more advanced methods (XGBoost, hyperparameter-tuned CatBoost, and a stacked ensemble) to identify improvements in minority-class F1.

Evaluating on minority-class F1, along with precision and recall for the Claim class, is crucial. High Claim recall ensures fewer missed claims (a key objective in risk management), and improved Claim precision reduces unnecessary resource allocations due to false positives. The minority-class F1 is a single metric that balances these two aspects specifically for the claim class.

4 Experimental Analysis

4.1 Data and Preprocessing

The initial dataset consists of 58,592 samples and 41 features capturing policy, vehicle, and customer attributes. This data is split into training (41,014 samples), development (8,789 samples), and test (8,789 samples) sets. The training data exhibits significant imbalance (No Claim: 38,390; Claim: 2,624), which is mitigated using SMOTE to balance the classes (No Claim: 38,390; Claim: 38,390).

4.2 Models Evaluated

We assess six configurations that address imbalance via class weighting or SMOTE:

1. Logistic Regression (Weighted)
2. Random Forest (Weighted)
3. CatBoost (Weighted)
4. Logistic Regression + SMOTE
5. Random Forest + SMOTE
6. CatBoost + SMOTE

We further experiment with an XGBoost model, a hyperparameter-tuned CatBoost model, and a stacked ensemble that combines Random Forest, XGBoost, and CatBoost using Logistic Regression as a meta-learner.

4.3 Metric Selection

We primarily track the F1 score of the minority class (Claim). This directly measures how effectively the model balances precision and recall for the class of interest. We also consider:

- **Precision (Claim)**: Reducing false positives is critical to avoid wasted resources.
- **Recall (Claim)**: Minimizing missed claims is crucial for accurate risk assessment.
- **F1 Score (Claim)**: Harmonic mean of the precision and recall
- **ROC AUC**: Provides an overall sense of discriminative ability across both classes, useful as a secondary comparison metric.

4.4 Results

Table 1 summarizes the Test set performance of the evaluated models, focusing on the minority-class F1 score and related metrics.

Model	Sampling/Wtg.	Claim F1	Cl. Prec.	Cl. Rec.	ROC AUC
Logistic Regression (Weighted)	Weighted	0.16	0.10	0.52	0.65
Random Forest (Weighted)	Weighted	0.15	0.10	0.32	0.64
CatBoost (Weighted)	Weighted	**0.18**	0.12	0.37	0.65
Logistic Regression (SMOTE)	SMOTE	0.16	0.10	0.41	0.65
Random Forest (SMOTE)	SMOTE	0.14	0.09	0.34	0.64
CatBoost (SMOTE)	SMOTE	0.17	0.10	0.51	0.65
XGBoost (Weighted)	Weighted	0.16	0.09	0.62	0.65
CatBoost (Tuned)	Weighted	0.16	0.10	0.36	0.65
Stacked Ensemble (Weighted)	Weighted	0.17	0.09	**0.71**	**0.67**

Table 1: Minority-Class Performance on Test Set

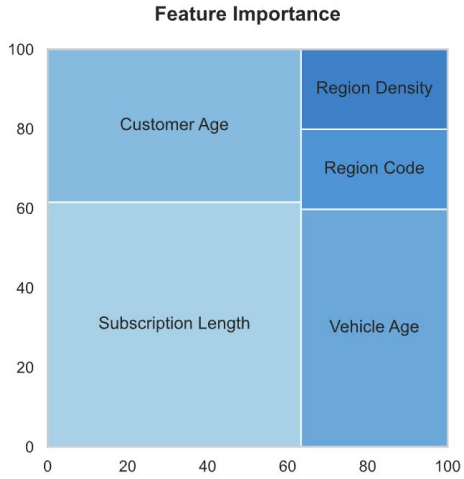

Figure 1: Feature Importance

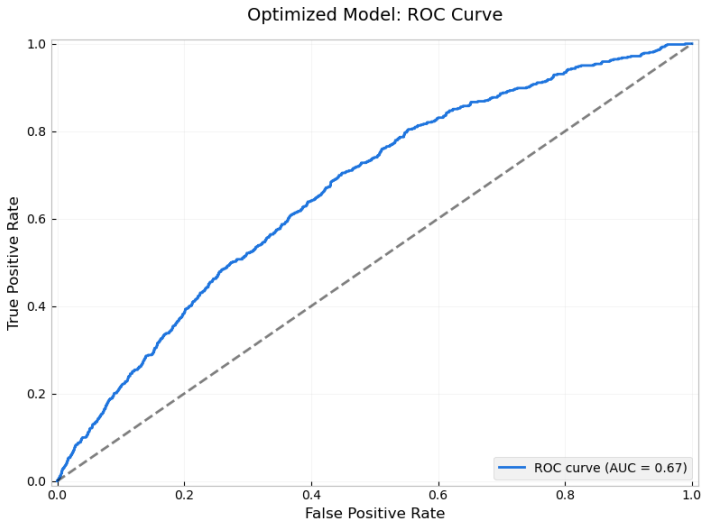

Figure 2: ROC-AUC curve for best performing (stacked ensemble model)

- The standard CatBoost (Weighted) model achieves one of the higher Claim F1 scores (0.18), balancing both precision and recall reasonably well.
- SMOTE variants sometimes increase recall at the cost of precision, leading to only modest improvements in Claim F1.
- The stacked ensemble achieves a very high Claim recall (0.71), but extremely low precision (0.09) leads to a more modest Claim F1 (0.17).
- Purely focusing on the minority-class F1 reveals that class weighting strategies can help models like CatBoost outperform their SMOTE counterparts for the claim detection task.
- We also applied L1-Regularization to our logistic regression model

4.5 Hyperparameter Tuning and Ensemble Models

A RandomizedSearch was used to optimize CatBoost parameters. While tuned CatBoost achieved a similar ROC AUC, it did not substantially improve the minority-class F1 compared to the baseline weighted CatBoost model.

The stacked ensemble combined predictions from multiple models. Despite its success in dramatically boosting Claim recall, its poor Claim precision limited the gain in Claim F1, highlighting that simply raising recall without sufficient precision does not necessarily lead to meaningful improvements for the minority class.

By focusing on the Claim (minority) class F1 score, the analysis shifts toward a more business-relevant assessment. CatBoost with class weighting provides a strong baseline for identifying claims with a reasonable balance of precision and recall. Although some models show improvements in recall, without corresponding gains in precision, their Claim F1 remains modest. Overall, focusing on the minority-class F1 metric ensures that both precision and recall are considered for the class that matters most in this insurance claims context.

5 Discussion and Prior Work

Despite experimentation with various algorithms and feature sets, we were unable to significantly boost the performance to a level. While it is possible that further optimization efforts or different model classes could improve the performance, it is possible that the data simply does not contain enough differentiated signal to effectively classify the minority class. One finding that corroborates this conjecture is shown in Figure 3 which clearly shows how similar the distributions were between the two classes even for the features which the models found most helpful for their classification.

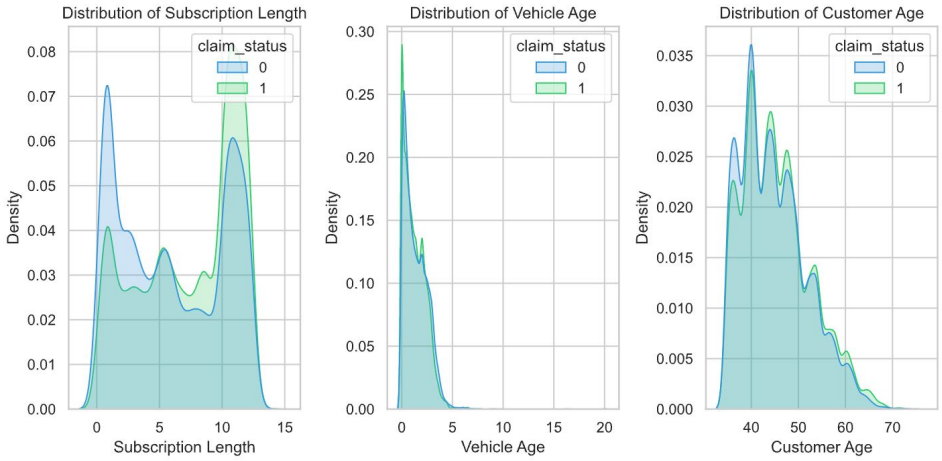

Figure 3: Distribution of Important Features Align closely between the 2 classes

Our findings echo the broader challenges in imbalanced classification, where attempts to boost minority- class recall often reduce precision. Although weighted CatBoost improved on simpler methods, additional techniques like SMOTE, hyperparameter tuning, and stacking offered only modest gains. These limited improvements hint that the data may simply lack the necessary discriminative signals. As Figure 3 shows, even the most informative features exhibit nearly identical distributions between classes, suggesting that more complex models or additional techniques may not yield substantial boosts.

In reviewing related work, we noted that several public Kaggle notebooks achieved promising results. However, upon closer inspection, these often relied on oversampling before data splitting, inadvertently introducing information leakage and inflating performance estimates.

Prior research in insurance analytics has similarly noted the difficulty of achieving fair, interpretable predictions. Our work complements these efforts by integrating LLM-driven text features and explanation generation with structured ML models. Rather than treating language and numeric data in isolation, we show that a hybrid approach can enrich predictive insights, building on existing literature that emphasizes explainability and trust in AI-driven decisions.

6 Conclusion

This study demonstrates that combining classical ML with LLM-based feature extraction and explanation can improve insurance claim prediction while maintaining interpretability. Focusing on minority-class metrics clarifies model shortcomings, guiding future research toward more nuanced imbalance mitigation. Subsequent work might explore richer text embeddings, enhanced integration of LLM outputs, and domain-aware feature engineering. These directions aim to refine performance, fairness, and user trust in AI-driven insurance analytics.

We developed an initial prototype of our solution 4, integrating our best-performing algorithm with the latest GPT-4o-mini model for reasoning and explanation generation. The reasoning component leverages the most important features (as shown in Figure 1) to articulate why a specific prediction was made. For example, the system might explain, "We predicted No Claim because the customer's age did not fall within the high-risk range identified in our data" or "We predicted Claim because the region density and region code matched patterns associated with previous claims."

This explainability not only enhances user trust in the predictions but also demonstrates the potential of combining classical machine learning with LLMs to provide actionable insights in insurance contexts. The prototype serves as a proof of concept, illustrating how explainable AI can bridge the gap between technical sophistication and user-centric design. Future iterations will aim to refine these explanations further, improving both their accuracy and intuitiveness.

Next possible steps would include, transforming the user provided information into structured feature vectors using LLMs to further enhance the predictions.

Appendix

Claim Predictor

Predicts the **Claim Likelihood** of a vehicle insurance claim!

Reference Data

	policy_id	subscription_length	vehicle_age	customer_age	region_code	region_density	segment
0	POL045360	9.3	1.2	41	C8	8,794	C2
1	POL016745	8.2	1.8	35	C2	27,003	C1
2	POL007194	9.5	0.2	44	C8	8,794	C2
3	POL018146	5.2	0.4	44	C10	73,430	A
4	POL049011	10.1	1	56	C13	5,410	B2
5	POL053680	3.1	2	36	C7	6,112	B2
6	POL053943	4.5	2.4	38	C2	27,003	C2
7	POL002857	10.7	2	56	C2	27,003	B2
8	POL028225	10.7	0.6	55	C5	34,738	B1
9	POL047631	0.3	2.4	45	C3	4,076	B2

Claim Description

Type `Claim description` below:

Customer born April 7 1998, has a 2.5 year old petrol car with a 12 month subscription.

Extract

Extracted Information

Subscription Length: `12`

Vehicle Age: `2.5`

Customer Age: `26`

Fuel Type: `petrol`

Prediction

Claim Likelihood

75.0%

↑ High Risk

Figure 4: POC Webapp built with Streamlit. Link to demo video