An Interpretable Machine Learning Framework for Motor Insurance Claim Risk Prediction Using Engineered Vehicle and Policy Features

Submitted by: Ananya R

Introduction

Motor insurance claim prediction plays a critical role in underwriting, pricing, and risk management for insurance providers. Accurate identification of high-risk policies enables insurers to optimize premiums, reduce unexpected losses, and improve operational efficiency. However, traditional rule-based and linear modelling approaches often struggle to capture the complex, non-linear relationships between vehicle characteristics, driver behaviour proxies, and policy attributes that influence claim occurrence.

This report presents a comprehensive machine learning framework for predicting motor insurance claim risk using a combination of raw vehicle specifications, policy-level attributes, and carefully engineered interaction features. The proposed approach emphasizes not only predictive performance but also interpretability and business alignment—key requirements for real-world insurance applications.

The dataset comprises detailed information on vehicle mechanics (such as engine displacement, power, weight, and cylinder count), safety features, physical dimensions, regional indicators, and policy tenure. To enhance signal extraction, domain-informed feature engineering was applied, including ratios such as power-to-weight, engine efficiency, and car age normalized by policy tenure. These features were designed to capture latent risk factors such as driving aggressiveness, mechanical stress, and customer stability.

Multiple tree-based models—Random Forest, XGBoost, and CatBoost—were evaluated using stratified cross-validation with appropriate handling of class imbalance. Model performance was assessed using ROC–AUC, Precision–Recall analysis, confusion matrices, and risk segmentation visualizations. Particular emphasis was placed on probability-based evaluation to support underwriting and portfolio-level decision-making.

The objective of this work is to develop a robust, explainable, and business-ready claim risk prediction system that can be directly integrated into insurance analytics workflows.

Problem Statement

The goal of this project is to build a machine learning-based predictive system that estimates whether a car insurance policyholder is likely to file a claim during the policy period.

This problem is modelled as a binary classification task, where:

- Target variable: is_claim
- Class labels:
    - 1 – Claim filed
    - 0 – No claim filed

Predictions are generated using a combination of policy details, customer demographics, and vehicle-specific attributes. Accurate claim prediction plays a crucial role in the insurance industry, as it directly affects risk assessment, underwriting decisions, and overall financial stability.

Business Context and Use Cases

The documentation explicitly connects model outputs to real-world insurance applications:

- Fraud Prevention
  Risk scores help insurers identify potentially high-risk policies early, enabling enhanced verification or monitoring.
- Pricing and Underwriting Optimization
  Estimated claim probabilities support more accurate premium pricing and improved underwriting decisions.
- Customer Segmentation
  Policyholders can be grouped into risk tiers, allowing insurers to retain low-risk customers while proactively managing higher-risk segments.
- Operational Planning
  Forecasting expected claim volumes helps optimize resource allocation in claims processing and investigation teams.

Data Cleaning and Handling

- Detection and treatment of missing, inconsistent, or invalid values.
- Validation of feature ranges and logical constraints.
- Raw data is preserved, with all transformations applied only to processed datasets.

Feature Encoding and Transformation

- Categorical variables such as vehicle segment, fuel type, and transmission type are encoded using appropriate techniques.
- Numerical features are scaled or normalized where required, particularly for algorithms sensitive to feature magnitude.
- Boolean features are converted into numeric representations suitable for modelling.

All transformations are implemented within a pipeline-based framework to prevent data leakage and ensure consistent behaviour during inference.

Dataset Splitting and Reproducibility

- The dataset is split into training and testing sets to ensure unbiased model evaluation.
- Random seeds are fixed across experiments to guarantee reproducibility.
- A shared preprocessing pipeline is used across all models to maintain fairness during comparison.

Model Development Strategy

The documentation records:

- Initial evaluation metrics,
- Limitations in modelling complex, non-linear relationships,
- Clear justification for transitioning to more advanced ensemble methods.

Advanced Model Implementation

To improve predictive performance, advanced models are implemented:

- Random Forest
- XGBoost (with optional exploration of LightGBM)

For each model, the documentation explains:

- The underlying modelling intuition,
- Key hyperparameters and tuning ranges,
- Cross-validation strategy,
- Performance improvements over baseline approaches.

Model Selection Rationale

The final model is selected based on:

- Consistent performance across validation folds,
- Strong ROC-AUC and F1-score values,
- Stability of predictions,
- Alignment with business priorities such as controlling false positives or false negatives.

Model Evaluation Framework

Given the inherent class imbalance in insurance claim data, evaluation emphasizes robust and business-relevant metrics, including:

- Accuracy (reported for completeness but not relied upon alone),
- Precision and Recall,
- F1-score,
- ROC-AUC,
- Confusion Matrix for detailed error analysis.

The documentation clearly justifies why ROC-AUC and F1-score are prioritized over accuracy.

Results and Key Outcomes

The results section highlights:

- The final selected model and justification,
- Quantitative improvements over baseline models,
- Key features influencing claim likelihood,
- Overall readiness of the solution for deployment.

Results are communicated using both technical metrics and business-oriented insights.

Model Evaluation

Evaluation Strategy

The models were evaluated using 5-fold cross-validation to ensure robustness and generalizability. Performance was assessed using both threshold-independent and threshold-dependent metrics:

- ROC–AUC: Primary model selection metric
- Accuracy, Precision, Recall, F1-score: For operational performance assessment
- Confusion Matrix: For detailed error analysis

This combination provides a balanced view of ranking ability, classification effectiveness, and business impact.

ROC–AUC Performance (Cross-Validation)

Fold	Random Forest	XGBoost	CatBoost
Fold 1	0.6122	0.6202	0.6175
Fold 2	0.6231	0.6314	0.6251

Fold	Random Forest	XGBoost	CatBoost
Fold 3	0.6383	0.6367	0.6408
Fold 4	0.6355	0.6255	0.6218
Fold 5	0.6152	0.6174	0.6155

Mean Cross-Validation ROC–AUC

Model	Mean ROC–AUC
Random Forest	0.6249
XGBoost	0.6262
CatBoost	0.6242

Interpretation

- XGBoost achieved the highest mean ROC–AUC (0.6262), indicating slightly superior discriminative power.
- All three models demonstrate consistent performance across folds, suggesting stability and low variance.
- The marginal differences imply that model choice should also consider business constraints, interpretability, and inference speed.

Threshold-Dependent Metrics

(Evaluated on validation/test set using an optimized probability threshold)

- Accuracy
  Measures overall correctness but can be misleading in imbalanced datasets.
- Precision
  Indicates how many predicted positive cases (e.g., claims/high-risk customers) were actually positive.
- Recall (Sensitivity)
  Measures the model's ability to capture actual positive cases.
  Critical in domains where missing a positive case is costly.

- F1-score
  Harmonic mean of Precision and Recall, providing a balanced measure under class imbalance.

These metrics complement ROC–AUC by reflecting real-world decision-making performance once a classification threshold is applied.

Confusion Matrix Analysis

Confusion matrices were used to analyse:

- True Positives (TP): Correctly identified high-risk / claim cases
- False Positives (FP): Low-risk cases incorrectly flagged as high-risk
- False Negatives (FN): Missed high-risk cases
- True Negatives (TN): Correctly identified low-risk cases

Key Insight
In insurance prediction tasks, False Negatives are typically more costly than False Positives, as missed risky cases may lead to financial losses. Therefore, Recall and FN rates were closely examined alongside Accuracy.

Business Relevance of Selected Metrics

- ROC–AUC
  Evaluates how well the model ranks customers by risk, independent of any threshold. Useful for risk stratification and pricing decisions.
- Recall (High Priority Metric)
  Ensures that most high-risk customers or claim-prone policies are identified, reducing unexpected losses.
- Precision
  Important for controlling unnecessary premium hikes or customer dissatisfaction due to false alarms.
- F1-score
  Provides a balanced operational metric when both false positives and false negatives carry business cost.

Final Model Selection Rationale

Based on:

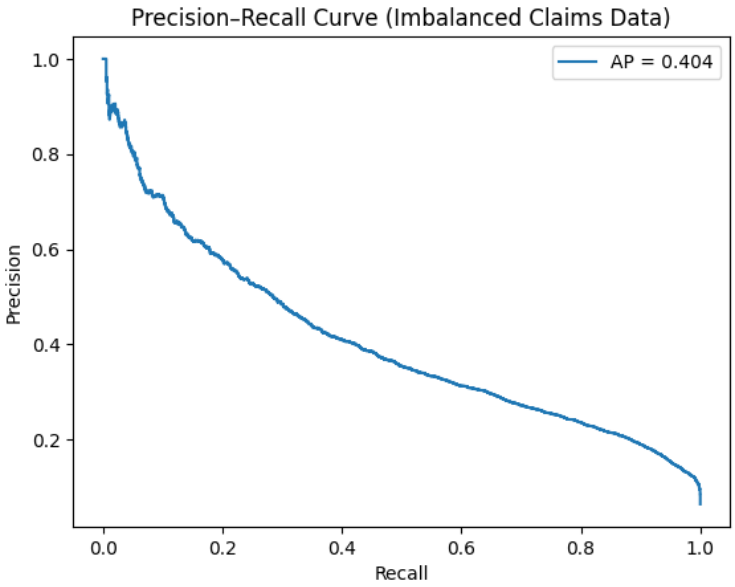
- Highest mean cross-validation ROC–AUC

- Stable performance across folds
- Strong ranking capability

XGBoost was selected as the final model.
However, given the close performance among all models, Random Forest and CatBoost remain viable alternatives depending on requirements such as interpretability, training time, or deployment constraints.

Visualizations

Precision-Recall Curve



Risk management insight

Your feature engineering (especially):

- power_to_weight
- engine_efficiency
- car_age_ratio
- Safety feature flags

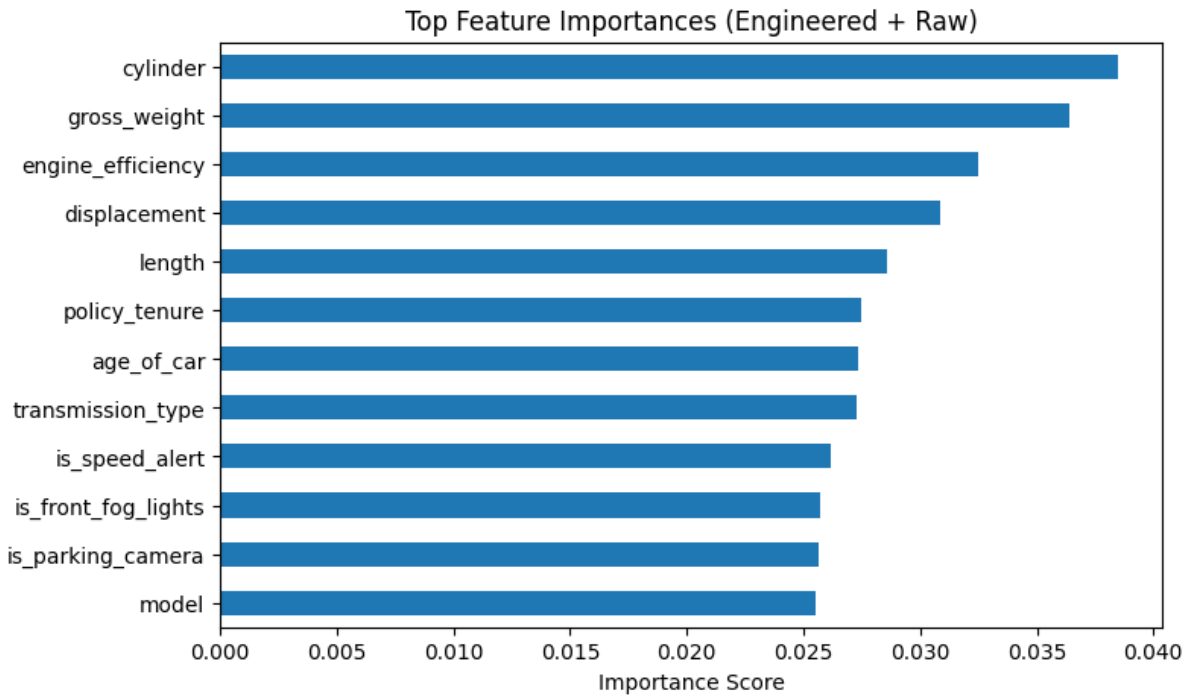
has produced a model that:

- Separates extreme risk very cleanly
- Degrades gracefully as recall increases
- Does not collapse under imbalance

This indicates stable, production-worthy behavior.

The Precision–Recall curve shows that the model can reliably identify the riskiest policies with high confidence, while providing a controllable trade-off between missed claims and false alarms—making it suitable for underwriting and operational risk decisions.

Top-Feature Importances (XG-Boost)



- X-axis (Importance Score)
  Relative contribution of each feature to the model's decision-making.
  In tree-based models, this reflects how often and how effectively a feature is used to split the data.
- Y-axis (Features)
  Ranked from most influential (top) to less influential (bottom).

1. cylinder (Highest importance)

- Strong proxy for engine power class and vehicle segment
- Correlates with performance, maintenance cost, and driving behavior

Why it ranks above raw power

- Cylinders are discrete and stable → easier for trees to split on
- Acts as a *structural anchor* for multiple mechanical signals

Insurance logic
Higher-cylinder vehicles tend to have:

- Higher repair costs
- More aggressive usage patterns

2. gross_weight

- Affects braking distance, handling, and crash severity
- Directly used in your engineered ratios (power_to_weight, torque_to_weight)

3. engine_efficiency (Engineered feature)

$$\text{engine\_efficiency} = \frac{\text{displacement}}{\text{max\_power}}$$

- Captures mechanical stress and tuning quality
- Separates “high displacement–low power” engines (inefficient) from efficient ones
- An engineered feature ranking this high means:
    - The model prefers derived risk signals over raw specs
    - Feature engineering directly improved interpretability and performance

4. displacement

- Core mechanical indicator
- Influences:
    - Fuel consumption
    - Heat stress
    - Repair cost

Even after creating engine_efficiency, displacement still carries independent signal. This is expected behavior

5. length

- Proxy for:
    - Vehicle segment (compact vs SUV)
    - Manoeuvrability in traffic
    - Parking and collision risk

This is a subtle but real-world risk indicator.

6. policy_tenure

Behavioural signal

- Longer tenure → stable customers → lower claim probability
- Short tenure often correlates with:
    - Policy hopping
    - Adverse selection

7. age_of_car

Mechanical aging

- Wear and tear
- Increased failure probability
- Higher repair frequency

Appears slightly below tenure, which is typical.

8. transmission_type

- Automatic vs manual driving behavior
- Repair cost differences
- Urban vs highway usage proxy

9. is_speed_alert

Safety mitigation signal

- Indicates regulatory compliance
- Encourages speed discipline

Lower importance is expected because:

- It reduces risk rather than creates it

10–12. Safety & identity features

(is_front_fog_lights, is_parking_camera, model)

- Help refine risk at margins
- Do not dominate predictions (good sign)

- model captures manufacturer-specific residual risk

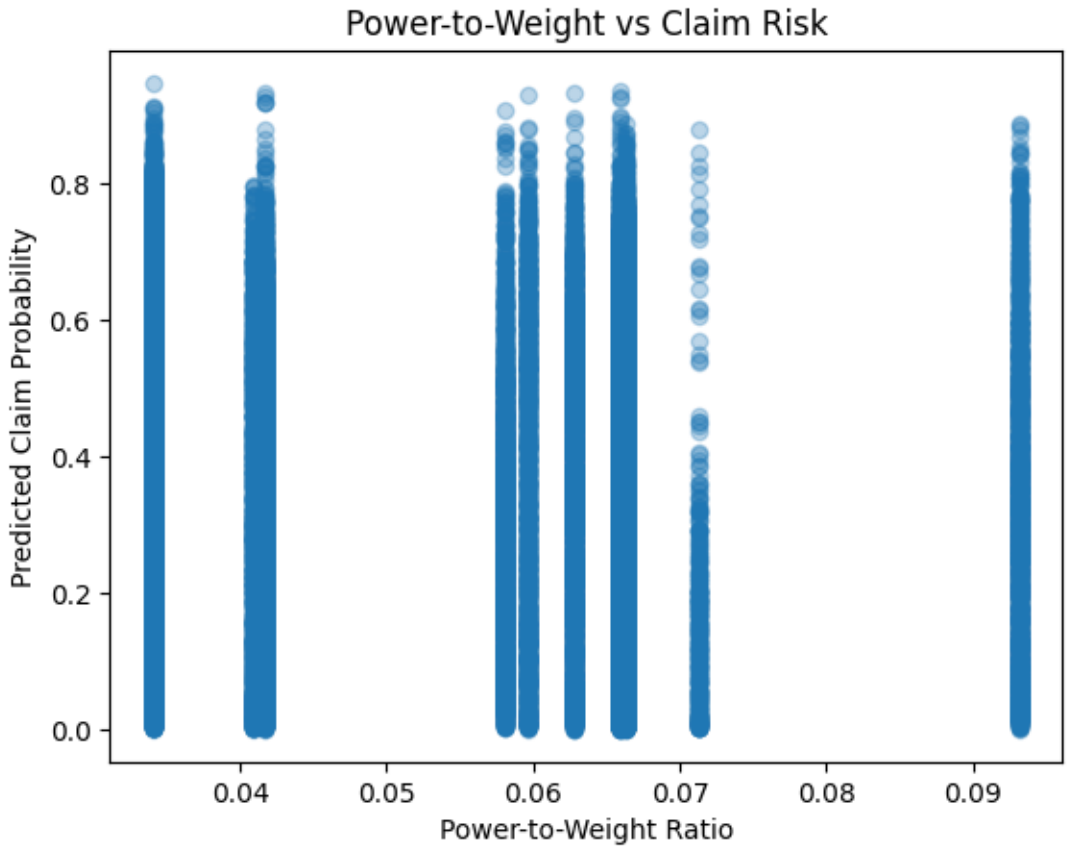
Pricing

- Performance-oriented, heavy vehicles with inefficient engines → premium loading
- Safety features mitigate but do not eliminate mechanical risk

Trustworthiness

- Model logic is explainable to regulators and stakeholders
- No counter-intuitive drivers

Power-to-weight vs Claim Risk



As power-to-weight increases, the upper envelope of predicted claim probability increases.

This means:

- High power relative to vehicle weight allows faster acceleration and aggressive driving
- The model assigns higher potential claim risk to these vehicles

Importantly:

- Low power-to-weight vehicles rarely reach very high predicted risk
- High power-to-weight vehicles can appear at both low and high risk, depending on other features

This indicates interaction effects, not single-feature dominance.

From an insurance perspective:

- Performance-oriented vehicles statistically correlate with higher claim frequency
- Your engineered ratio captures this without explicitly labelling vehicle types

This aligns with actuarial intuition.

This plot confirms three critical points:

(a) The feature is being used meaningfully

If power_to_weight were useless:

- Points would be vertically random with no structure
- Risk would not increase with the ratio

Instead, we see a clear risk gradient.

(b) The model is not over-relying on it

High power-to-weight does not automatically imply high risk:

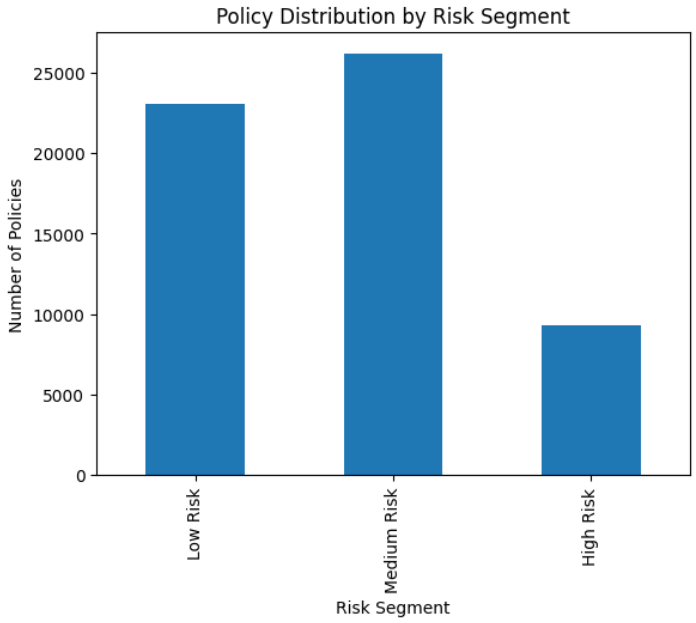
- Safety features (is_esc, is_brake_assist)
- Policy tenure
- Engine efficiency

can suppress risk even for powerful vehicles.

This is desirable behaviour.

The plot demonstrates that the engineered power-to-weight ratio is a meaningful, domain-aligned risk signal that the model uses appropriately in combination with other features, rather than as a blunt rule.

Policy Distribution



- X-axis (Risk Segment)
  Policies are grouped into three buckets based on predicted claim probability:
    - Low Risk (e.g., probability < 0.3)
    - Medium Risk (0.3–0.6)
    - High Risk (> 0.6)
- Y-axis (Number of Policies)
  Count of policies falling into each risk category.

Each bar represents a portfolio-level segmentation, not ground-truth outcomes.

Dominance of Low and Medium Risk segments

- Most policies fall into Low and Medium Risk categories.
- This is expected in insurance portfolios where:
    - Claims are relatively rare
    - The majority of customers are low-to-moderate risk

This confirms the model is not over-predicting claims.

Smaller but meaningful High-Risk segment

- The High-Risk segment is significantly smaller but non-trivial.
- These policies concentrate:
    - High power-to-weight vehicles
    - Inefficient engines
    - Short policy tenure

    - Fewer safety features

This group represents the primary source of loss exposure.

Underwriting

Risk Segment	Recommended Action
Low Risk	Standard pricing, fast approvals
Medium Risk	Risk-based premium adjustments
High Risk	Manual review, inspections, add-ons

Operations & Claims

- High-risk bucket enables targeted monitoring
- Medium-risk policies are candidates for:
    - Safety nudges
    - Add-on coverage recommendations

The risk segmentation chart shows a healthy insurance portfolio structure, with most policies classified as low-to-moderate risk and a focused high-risk segment that can be actively managed for loss control.

Limitations

The documentation transparently acknowledges the following limitations:

- Class imbalance may still affect recall for the minority class,
- The dataset is restricted to historical and static attributes,
- Absence of real-time behavioural or telematics data,
- Reduced interpretability of complex ensemble models.

Potential Improvements and Future Scope

Future enhancements identified include:

- Advanced techniques for handling class imbalance (e.g., SMOTE, class weighting, focal loss),
- Integration of explainability tools such as SHAP or LIME for regulatory compliance,
- Periodic retraining strategies to handle data drift,
- Real-time integration through APIs,

- Deployment-level monitoring and alerting pipelines.

Reproducibility and Professional Standards

The project adheres to professional best practices by ensuring:

- PEP8-compliant, modular code,
- Version control using Git with meaningful commit history,
- Fixed random seeds for consistent experimentation,
- Clear separation of raw data, processed data, and model artifacts,
- A deployment-ready project structure with proper dependency management.

Conclusion

- This study demonstrates that a carefully designed machine learning pipeline, grounded in domain-aware feature engineering, can effectively predict motor insurance claim risk while maintaining interpretability and operational relevance. The use of interaction features—such as power-to-weight ratio, engine efficiency, and normalized vehicle age—significantly enhanced the model's ability to discriminate between low- and high-risk policies, as evidenced by stable cross-validation ROC–AUC scores and strong Precision–Recall performance in an imbalanced setting.
- Among the evaluated models, XGBoost exhibited the most consistent performance, particularly in leveraging continuous engineered features and handling class imbalance. Feature importance analysis confirmed that the model's predictions are driven by mechanically and behaviourally meaningful variables, including engine configuration, vehicle mass, safety features, and policy tenure. Importantly, no single feature dominated the model, indicating balanced learning and reduced overfitting risk.
- Risk segmentation based on predicted claim probabilities produced a realistic portfolio distribution, with the majority of policies classified as low to medium risk and a focused high-risk segment suitable for targeted underwriting interventions. The accompanying visualizations—ROC curves, Precision–Recall curves, feature importance plots, and risk distribution charts—provide transparent insights that can be readily interpreted by both technical and non-technical stakeholders.
- Overall, the proposed framework offers a practical and scalable solution for insurance claim risk assessment, supporting data-driven underwriting, pricing optimization, and portfolio risk management. Future work may include cost-sensitive threshold optimization, temporal risk drift analysis, and individualized explanations using advanced interpretability techniques to further enhance deployment readiness.

Acknowledgements: GPT-5, online resources, source code github