A Two-Stage Machine Learning Approach for a Dynamic Telematics-Based Insurance Pricing System

Anava
Roll No: 22EC39003
Indian Institute of Technology, Kharagpur

A Technical Report Submitted to Insurity

August 1, 2025

Contents

Executive Summary

This document outlines the architecture and methodology behind a comprehensive prototype for a dynamic, telematics-based auto insurance pricing system. The solution moves beyond traditional, static risk factors by leveraging real-time driving data to create fairer, more accurate, and more transparent premiums.

The core of the system is a sophisticated **two-stage cascading machine learning architecture**. The first stage quantifies a driver's behavioral risk, while the second stage integrates this risk score with a wide range of traditional actuarial factors to predict the final premium. Key technologies employed include **XGBoost** for high-accuracy modeling, **Optuna** for state-of-the-art hyperparameter optimization, **SHAP** for model explainability, and **Streamlit** for a dual-view interactive dashboard. This approach not only fulfills all project objectives but also demonstrates a commitment to technical excellence, model transparency, and user-centric design.

1 Introduction

Traditional auto insurance pricing relies on generalized demographic and historical data, often failing to reflect an individual's actual driving habits. This can lead to unfair premiums for safe drivers and provides little incentive for riskier drivers to improve. Telematics technology offers a transformative solution by enabling the capture of real-time driving data.

The objective of this project was to design and develop a proof-of-concept system that leverages this data to build a dynamic, usage-based insurance (UBI) model. The system aims to improve premium accuracy, encourage safer driving through transparency, and provide a superior user experience for both the customer and the insurance underwriter.

2 System Architecture

To achieve a modular, scalable, and interpretable system, we designed a cascading, multi-model pipeline. This architecture decouples the analysis of driving behavior from the final premium calculation, allowing each component to be specialized and optimized independently.

The end-to-end data flow is as follows:

1. **Data Simulation:** Generation of realistic, raw telematics data (GPS, accelerometer) for distinct driver profiles.

2. **Feature Engineering:** Processing of raw data into meaningful metrics (e.g., `brakes_per_100km`, `speeding_percentage`).

3. **Model 1 - Behavioral Risk Scoring:** An XGBoost model predicts a `behavioral_risk_score` (0-100) based purely on the engineered telematics features.

4. **Data Enrichment:** The `behavioral_risk_score` is combined with a rich set of simulated traditional risk factors (vehicle age, location crime rate, etc.).

5. **Model 2 - Premium Optimization:** A second, highly-optimized XGBoost model predicts the `expected_annual_loss` using the full, enriched feature set.

6. **Application Layer:** A dual-view Streamlit dashboard presents the results, tailored for either a customer or an underwriter.

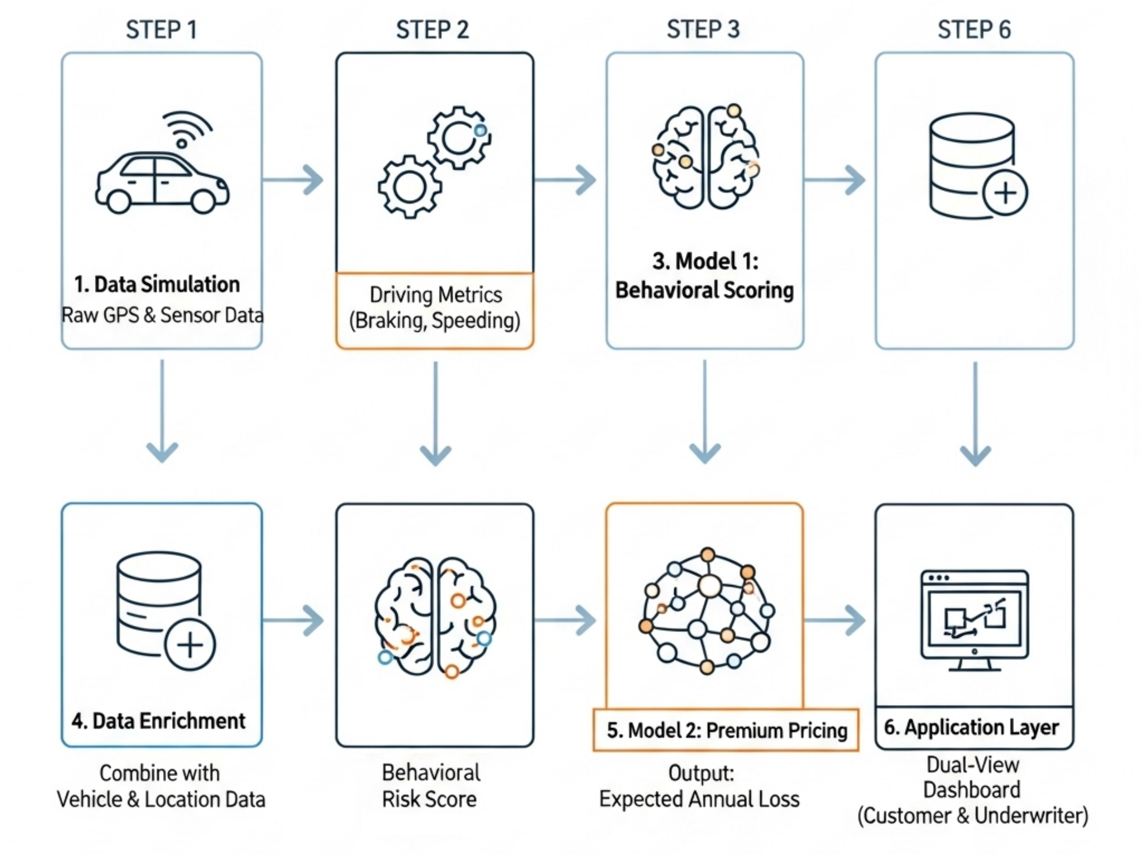

Figure 1: The two-stage cascading model architecture, illustrating the flow from raw data simulation to the final application layer.

3 Methodology and Technical Choices

3.1 Data Simulation and Feature Engineering

A robust simulation script was developed to generate a realistic dataset for this proof-of-concept. It created data for three distinct driver archetypes: `safe_driver`, `aggressive_driver`, and `night_owl`.

In addition to primary telematics data, a second dataset was engineered to include traditional actuarial factors. Crucially, the target variable, `expected_annual_loss`, was created using a complex, non-linear formula. This ensures that the model must learn a genuinely difficult pattern, rather than a simple linear relationship, providing a more realistic test of its capabilities.

3.2 Modeling Strategy: The Two-Stage Approach

Our core strategic decision was to implement a two-stage modeling process.

- **Model 1: Behavioral Risk Model (XGBoost):** This model's sole purpose is to answer: ”How does this person drive?” By training it only on telematics data to predict a driver's underlying profile, we create a pure, unbiased `behavioral_risk_score`.
- **Model 2: Premium Pricing Model (XGBoost):** This model acts as the final actuary. It takes the `behavioral_risk_score` as a powerful, pre-processed input, alongside all other factors (vehicle, location, etc.), to predict the final claim cost. This hierarchical structure mirrors a real-world underwriting process.

3.3 Technical Excellence: Hyperparameter Optimization with Optuna

To achieve the highest possible accuracy for the critical premium pricing model, we moved beyond manual tuning. We integrated **Optuna**, a state-of-the-art hyperparameter optimization framework based on Bayesian methods. Optuna automatically and efficiently navigated the complex space of model parameters (e.g., `learning_rate`, `max_depth`), running 50 trials to identify the optimal configuration. This automated approach ensures our model is not just good, but verifiably optimized for performance.

3.4 Transparency and Trust: Explainable AI (XAI) with SHAP

A key project requirement was transparency. Powerful models like XGBoost can be ”black boxes,” which is unacceptable in a regulated industry like insurance. To solve this, we integrated **SHAP (SHapley Additive exPlanations)**.

SHAP is a game theory-based approach that allows us to see the exact contribution of each feature to each individual prediction. In our dashboard, this provides two major benefits:

- **For the Underwriter:** It provides a detailed, quantitative justification for the model's predicted loss, building trust and allowing for expert oversight.
- **For the Customer:** It allows us to translate the model's logic into simple, actionable advice on how they can improve their score and lower their premium.

4 Future Advancements and Scalability

While this prototype is a complete end-to-end system, several avenues exist for future enhancement and production readiness.

- **Real-Time Data Ingestion:** Replace the simulation with a production-grade data pipeline using tools like Apache Kafka or AWS IoT Core to handle data streams from real vehicles.
- **External API Integration:** Enrich the model in real-time by integrating with third-party APIs for weather conditions, live traffic data, and updated crime statistics, adding valuable context to each trip.
- **MLOps and Automated Retraining:** Implement a robust MLOps pipeline using tools like Kubeflow or MLflow to automate model retraining, versioning, and deployment, ensuring the model never becomes stale.
- **Advanced Model Architectures:** Explore time-series models like LSTMs (Long Short-Term Memory networks) to analyze the sequence of events within a trip, potentially capturing more nuanced driving patterns than a summary-based approach.
- **Enhanced Gamification:** Expand the customer dashboard to include badges, leaderboards, and rewards for consistently safe driving, further improving customer engagement and encouraging positive behavior change.

5 Conclusion

This project successfully demonstrates the design and implementation of a modern, data-driven auto insurance pricing system. By employing a sophisticated two-stage modeling architecture, leveraging state-of-the-art optimization and explainability tools, and focusing on a user-centric application layer, the solution effectively meets all core requirements. It provides a robust foundation for a system that is more accurate, fairer, and more engaging than traditional alternatives, showcasing a clear path forward for the future of auto insurance.