Faculty of Economics and Business Administration
ICADE

TELEMATICS & MACHINE LEARNING APPLIED TO INSURANCE: PAY HOW YOU DRIVE

Author: Santiago Arenas Martín
Director: Susana Josefa Gago Rodríguez

MADRID | MAY 2026

Abstract

Motor insurance has long relied on the same handful of variables such as age, location or vehicle type, to set premiums that are meant to reflect risk. The problem is those variables are poor proxies. This dissertation argues that telematics data, when combined with machine learning, can replace that system of guesswork with something far more accurate and, frankly, fairer. Using the freMTPL2 actuarial dataset alongside synthetic telematics data derived from the highD naturalistic driving study, I developed and tested a "Pay How You Drive" (PHYD) dynamic pricing framework that adjusts premiums monthly based on observed driving behavior. Telematics-enhanced gradient boosting models reduced prediction error by over 68% compared to a traditional GLM baseline (RMSE dropping from €13.54 to €4.30), with a composite risk score: capturing speed, following distance, and lane-change patterns, emerging as the single most informative feature by a considerable margin. A twelve-month pricing simulation demonstrates how an improving driver could cross below the traditional fixed-premium threshold by mid-year, paying substantially less by December. The dissertation also addresses implementation realities: privacy concerns, regulatory constraints, and the challenge of communicating algorithmic pricing to customers who simply want to understand their bill.

Keywords: Telematics, Usage-Based Insurance, Machine Learning, Dynamic Pricing, Pay How You Drive, Actuarial Science, GLM, Risk Scoring.

Table of Contents

Abstract *1*

Chapter I: Introduction **4**

1.1 Motivation and Context *4*

1.2 Research Objectives *5*

1.3 Thesis Structure *5*

Chapter II: State of the Art **7**

2.1 Traditional Actuarial Methods *7*
2.1.1 Foundations of Insurance Pricing 7
2.1.2 Generalized Linear Models in Insurance 7
2.1.3 Traditional Rating Factors 8
2.1.4 Limitations of Traditional Approaches 8

2.2 Telematics in Motor Insurance *10*
2.2.1 Evolution of Usage-Based Insurance 10
2.2.2 Telematics Data Sources 10
2.2.3 Key Telematics Metrics 11
2.2.4 Risk Scoring Methodologies 12

2.3 Machine Learning Applications in Insurance Pricing *13*
2.3.1 Beyond GLMs: The Case for Machine Learning 13
2.3.2 Common Machine Learning Methods 13
2.3.3 Interpretability Considerations 14

2.4 Current Market Landscape *15*
2.4.1 Industry Adoption 15
2.4.2 Key Players and Approaches 15
2.4.3 Challenges and Barriers 15

Chapter III: Methodology **17**

3.1 Data Acquisition and Sources *17*
3.1.1 Actuarial Data: freMTPL2 Dataset 17
3.1.2 Telematics Data: highD Dataset 17
3.1.3 Supplementary Data Sources 18

3.2 Data Preprocessing and Feature Engineering *19*
3.2.1 Actuarial Data Preparation 19
3.2.2 Telematics Data Processing 19
3.2.3 Feature Engineering for the Combined Dataset 20

3.3 Telematics Variable Definition *21*
3.3.1 Speed-Based Risk Indicators 21
3.3.2 Following Behavior Indicators 21
3.3.3 Composite Risk Score 21

3.4 Model Design and Implementation *23*
3.4.1 Baseline GLM Model 23
3.4.2 Machine Learning Models with Telematics 23
3.4.3 Training and Evaluation 24

3.5 Synthetic Data Generation *25*
3.5.1 Motivation 25
3.5.2 Generation Methodology 25

3.5.3 Validation 25

Chapter IV: Results and Evaluation 26

4.1 Model Performance Comparison 26
4.1.1 Summary Results 26
4.1.2 Interpretation of Results 26
4.1.3 Performance Visualization 27

4.2 Feature Importance Analysis 28
4.2.1 Random Forest Feature Importances 28
4.2.2 Telematics vs. Actuarial Contribution 29

4.3 Dynamic Pricing Simulation 30
4.3.1 Monthly Premium Adjustment Scenario 30
4.3.2 Annual Comparison and Behavioral Incentive Effect 31

4.4 Risk Segmentation Analysis 32
4.4.1 Premium Distribution by Risk Tier 32
4.4.2 Key Observations 32
4.4.3 Cross-Subsidization Elimination 33

Chapter V: Conclusion 34

5.1 Summary of Findings 34
5.1.1 Telematics Data Value 34
5.1.2 Model Performance 34
5.1.3 Dynamic Pricing Effectiveness 34
5.1.4 Risk Segmentation 34

5.2 Practical Implications 35
5.2.1 For Insurance Companies 35
5.2.2 For Regulators 35
5.2.3 For Consumers 35

5.3 Limitations 36
5.3.1 Data Limitations 36
5.3.2 Methodological Limitations 36
5.3.3 Scope Limitations 36

5.4 Future Research Directions 36

References 38

Appendix A: Data Sources Summary 40

Appendix B: Risk Scoring Components 40

Appendix C: Premium Multiplier Formula 40

Appendix D: Code 40

Chapter I: Introduction

1.1 Motivation and Context

The motor insurance industry has a pricing problem that it has largely managed to ignore for decades. Premiums have traditionally been set using a handful of demographic and vehicle characteristics, like driver age, region, vehicle make, claims history, none of which is actually telling an insurer how someone drives. They are statistical proxies, and fairly crude ones at that. (Wolny-Dominiak, 2024) A 23-year-old living in Madrid pays more for insurance than a 45-year-old living in the same street, not because the insurer knows anything about either driver's actual behavior behind the wheel, but because, on average, younger drivers claim more. The 23-year-old who drives cautiously and never speeds subsidizes the 45-year-old who tailgates on motorways daily. This is not just economically inefficient, but also fundamentally unfair (Baumann & Loi, 2023), and it is the core problem this dissertation attempts to address.

What makes this moment interesting is that, for the first time in history, the data exists to do something about it. Modern vehicles are increasingly equipped with GPS receivers, accelerometers, and cellular connectivity. Smartphones in nearly every pocket contain the same sensors. The result is that, if an insurer chooses to collect it, a detailed picture of someone's actual driving behavior, their typical speeds, how closely they follow the car in front, how often and hard they brake, what time of day they tend to drive, the picture is entirely within reach. This is the premise behind telematics insurance.

Usage-Based Insurance (UBI) has evolved considerably since early programmes simply tracked mileage under the logic that more kilometers driven means more exposure to risk. That first generation, often called Pay As You Drive (PAYD), was a genuine improvement over purely demographic pricing, but left a great deal of information on the table. The second generation, Pay How You Drive (PHYD), is more ambitious. Rather than just counting kilometers, PHYD programmes monitor the full texture of driving behavior: speed patterns and variability, how the driver accelerates and brakes, how much space they leave in front of the car ahead, lane change frequency, time-of-day patterns, and more. This richer data allows for something closer to genuine individual risk assessment rather than demographic averaging.

The commercial logic is straightforward. Insurers who can accurately price individual risk will attract the safest drivers, who will be offered better deals than they can get elsewhere. Insurers who cannot, therefore, end up with a disproportionately risky pool as their best customer's defect. This adverse selection dynamic more than a theoretical concept, it is a proven fact in markets where telematics adoption has reached a critical mass. Progressive's Snapshot programme in the United States, for instance, has been collecting this kind of data for over a decade and has accumulated significant competitive advantages in risk selection as a result (Tselentis, Yannis, & Vlahogianni, 2017). In Italy, telematics penetration has reached levels that have fundamentally changed how the market works (Husnjak et al., 2015). Research across multiple markets has consistently found that telematics improves insurer loss ratios, most prominently through better identification of high-risk drivers who traditional demographic pricing undercharges (Baecke & Bocca, 2017).

The question this dissertation asks is not whether telematics improves insurance pricing, that case has already been made in the literature, though primarily using aggregate data. The question is how to build the full pipeline: from raw driving data to a model to a monthly premium figure that an insurer could plausibly put in front of a customer. That means combining a traditional actuarial dataset with behavioral driving data, engineering meaningful risk features, comparing model approaches, and running a pricing simulation that reflects what a real PHYD product would look like.

1.2 Research Objectives

The primary objective of this dissertation is to design and validate a dynamic pricing model that uses telematics-derived driving behavior metrics to support month-to-month premium adjustments in motor insurance. The aim is to produce a framework that is practically implementable, not only theoretically elegant.

Several secondary objectives support this. First, I established a baseline using traditional Generalized Linear Models (GLM) approaches and standard actuarial rating factors. The aim is not to dismiss the traditional approach, but to quantify exactly what it achieves and where it falls short. Second, I engineered telematics features from raw trajectory data that are both statistically meaningful and interpretable enough to survive regulatory scrutiny. Thirdly, I trained and compared machine learning models, namely Random Forest and Gradient Boosting, on a combined actuarial and telematics feature set, measuring the improvement over the GLM baseline. Fourth, I designed the translation from model output to monthly premium, including a multiplier formula that is actuarially sound and commercially sensible. Finally, I addressed the practical landscape: what privacy constraints exist, what regulators in Spain and the EU require, and what the realistic path to adoption looks like for an insurer considering this approach.

1.3 Thesis Structure

Chapter II: State of the art. It surveys the relevant literature, covering three main areas. The first is traditional actuarial methods, the mathematical foundations of GLM pricing and why those methods work as well as they do while still leaving substantial room for improvement. The second is the evolution of telematics in insurance, tracing the progression from simple mileage-tracking to the sophisticated behavioral monitoring programmes operating today. The third is Machine Learning in insurance pricing, with particular attention to ensemble methods and the interpretability challenges that come with them.

Chapter III: Methodology. It presents the methodology in full technical detail. This includes the two primary data sources, the freMTPL2 actuarial benchmark and the highD drone-camera trajectory dataset, the preprocessing and feature engineering pipelines, the definition of telematics risk variables, the model specifications, and the synthetic data generation approach that makes the combined dataset feasible without requiring access to real policyholder records.

Chapter IV: Results and Evaluation. This presents and discusses the empirical results: model performance across all three approaches, feature importance analysis, the dynamic pricing simulation, and a risk segmentation analysis showing how telematics enables finer premium differentiation than traditional methods alone.

Chapter V: Conclusion. Where I acknowledge the limitations of this work honestly and identify what I see as the most promising directions for future research, including what a follow-up study with real claims data linked to telematics records could accomplish.

Chapter II: State of the Art

2.1 Traditional Actuarial Methods

2.1.1 Foundations of Insurance Pricing

The actuarial foundations of motor insurance pricing are well-established and have remained largely stable for decades. At their core, they rest on a simple decomposition: the premium a policyholder pays should cover the expected cost of claims (the pure premium), the insurer's operating expenses, and a margin for profit and uncertainty. Formally:

$$\text{Premium} = \text{Pure Premium} + \text{Expense Loading} + \text{Profit Margin}$$

Equation 1: Standard insurance premium decomposition.

The pure premium itself is the product of two separately modelled quantities; claim frequency, how likely a claim is; and claim severity, how costly a claim is when it occurs. This can very simply be modelled as:

$$\text{Pure Premium} = \text{Expected Frequency} \times \text{Expected Severity}$$

Equation 2: Pure premium as the product of expected frequency and expected severity.

This separation matters because frequency, Equation 3, and severity, Equation 4, are driven by different factors and follow different statistical distributions. Modelling them jointly tends to produce worse results than modelling each independently and combining the predictions.

2.1.2 Generalized Linear Models in Insurance

The statistical workhorse for both components has been the Generalized Linear Model (GLM), introduced to actuarial practice in the 1980s and still dominant today (Denuit et al., 2007). GLMs extend ordinary linear regression to accommodate the non-normal response distributions that characterise insurance data: claim counts, for instance, are non-negative integers that follow something closer to a Poisson distribution, while claim amounts are right-skewed and more naturally modelled with a Gamma.

A GLM has three components:

- The random component specifies the distribution of the response variable. Poisson for claim counts, Gamma for amounts.
- The systematic component is a linear combination of explanatory variables: $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
- The link function g connects the expected value of the response to this linear predictor: $g(E[Y]) = \eta$. For frequency modelling the log link is standard, giving the Poisson GLM specification in Equation 3:

$$\log(E[\text{ClaimCount}]) = \beta_0 + \sum_{i=1}^{p} \beta_i \, x_i$$

Equation 3: Poisson GLM for claim frequency, with log link.

This is a multiplicative model. The risk factors compound rather than add, which is naturally intuitive when thinking that one risk factor affects other. For instance, a young driver (riskier) in an urban area (riskier) with a powerful vehicle (riskier again) does not just sum those three risk factors; they interact in ways that multiply the overall risk. The GLM handles this naturally through the log link. Severity is modelled similarly using a Gamma distribution, as in Equation 4:

$$\log(E[\text{ClaimAmount} \mid \text{Claim}]) = \gamma_0 + \sum_{i=1}^{q} \gamma_i \, z_i$$

Equation 4: Gamma GLM for claim severity, with log link.

2.1.3 Traditional Rating Factors

The explanatory variables fed into these models have historically been constrained by what data insurers could collect at the point of policy inception. Sociological driver characteristics: age, years of experience, claims history encoded through bonus-malus coefficients, occupation, marital status, form one group. Technical vehicle characteristics form another: power class, age, brand, fuel type, declared value. Geographic factors round out the picture: region, area classification, urban versus rural, and population density.

These variables work as proxies. They work because the correlations are generally real: young drivers genuinely do have more accidents on average, high-powered vehicles genuinely are involved in more serious ones, urban drivers genuinely face different risk environments. The problem is that the correlation between demographic category and individual driving behavior is loose enough to create substantial pricing error at the individual level (Antonio & Valdez, 2012). One study found that, among drivers in the same demographic band, actual claim rates varied by a factor of three or more. This means that the rating category explains only a fraction of the true individual risk (Guillen, Nielsen, Ayuso, & Pérez-Marín, 2019). The proxy is the average, and individuals deviate from averages in both directions.

2.1.4 Limitations of Traditional Approaches

The central limitation is that traditional rating factors capture what kind of person a driver is, not how they actually drive. Two drivers with identical age, vehicle, and location receive identical premiums regardless of whether one drives carefully and the other aggressively. This cross-subsidization is not a minor inefficiency, it means that within any rating segment, low-risk drivers are consistently overcharged relative to their actual cost, while high-risk drivers are undercharged. Over time, those low-risk drivers are precisely the ones who will be attracted away by a competitor offering behavior-based pricing, leaving the traditional insurer with an increasingly adverse risk pool.

There are further structural limitations. Annual premiums cannot respond to changes in driving behavior during the policy period. A driver who moves, changes jobs, or significantly evolves their driving habits will not see those changes reflected until renewal. The backward-looking nature of claims history as a rating variable means new drivers start with no meaningful risk signal, and drivers whose circumstances change may carry outdated pricing for extended periods (Lemaire, Park, & Wang, 2016). These are not minor technical complaints by any stretch. They are genuine gaps in the accuracy of the pricing system.

2.2 Telematics in Motor Insurance

2.2.1 Evolution of Usage-Based Insurance

The history of telematics insurance is a story of increasing data richness. The first generation of Usage-Based Insurance (UBI) programmes, which began appearing in the early 2000s, focused almost entirely on mileage. The premise was simple: more kilometers driven means more exposure to accident risk, so low-mileage drivers should pay less. This is true as far as it goes and Pay As You Drive (PAYD) programmes did show measurable improvements over purely demographic pricing (Paefgen, Staake, & Fleisch, 2014). But mileage tells you nothing about how someone drives those kilometers.

The second generation, Pay How You Drive (PHYD), added behavioral monitoring. Devices recorded events such as hard braking, rapid acceleration, and high speeds, allowing premiums to reflect driving style as well as exposure. This was a qualitative leap. An insurer could now distinguish, within the same demographic bracket, between a cautious driver who never exceeds speed limits and an aggressive one who brakes hard and accelerates rapidly. The foundational work on PHYD in the academic literature, particularly Guillen et al. (2019) and Verbelen, Antonio & Claeskens (2018), demonstrated that the additional behavioral signal translated into meaningful improvements in risk prediction over GLMs or PAYD.

Current implementations go further still, integrating real-time feedback mechanisms, smartphone coaching, and gamification elements. Drivers see their behavior scores update in near-real time, receive alerts when they drive poorly, and can track how specific habits affect their projected premium. This third generation, sometimes called Manage How You Drive (MHYD), treats the telematics device not just as a data collection tool but as a behavioral intervention. The evidence suggests this works: drivers in MHYD programmes do improve their behavior over time, which matters both for insurer loss ratios and for road safety outcomes more broadly (Tselentis et al., 2017).

Looking further forward, connected vehicle technology (V2X) and the gradual introduction of advanced driver assistance systems will generate far richer data streams than today's telematics devices, potentially enabling near-instantaneous risk assessment. For now, though, the practical frontier remains the PHYD model. That is what this dissertation focuses on.

2.2.2 Telematics Data Sources

The mechanics of data collection vary across programmes. Dedicated hardware devices, typically OBD-II dongles plugged into the vehicle's diagnostic port or standalone GPS units, offer the most reliable and accurate data. They draw directly from vehicle systems, have consistent sensor quality, and are not dependent on a smartphone's battery, instruments or position in the car. The tradeoff is cost and logistics: devices need to be manufactured, shipped, and installed, and some drivers resist having hardware in their vehicle.

Smartphone-based programmes have lower unit costs and easier deployment but introduce their own complications. Accelerometer data collected through a phone depends on the phone's position in the car, which varies, and on the sensor quality of whatever device the customer happens to own. They are also more susceptible to fraud; a driver can simply leave the phone at home. That said, smartphone telematics has proven workable in practice, and its cost advantages have made it the dominant approach for mass-market programmes (Husnjak et al., 2015).

A third and increasingly important source is the connected vehicle itself. Modern cars from most major manufacturers include factory-installed telematics systems that communicate with insurer platforms directly. This approach provides the richest data, access to vehicle systems rather than just GPS and accelerometer approximations but requires manufacturer partnerships and creates data governance questions about who owns the vehicle data, or whether the average customer is even aware of the data acquisition taking process. (Voss, 2018)

2.2.3 Key Telematics Metrics

Raw telematics data consists of time-stamped position, velocity, and acceleration readings at high frequency. The analytical work lies in aggregating these into metrics that are both statistically predictive of claims and practically interpretable. Several categories are well-established in the literature.

Speed-related metrics are the most intuitive: average speed, maximum speed observed, speed variability (the range between maximum and minimum), and the frequency of observations above various thresholds. Speed and accident risk are strongly correlated at both ends. Consistent high-speed driving increases crash frequency, while excessive variability (aggressive acceleration and braking) is associated with both crashes and near-misses (Ma et al., 2018). Acceleration metrics capture hard braking and rapid acceleration events, both of which are associated with elevated risk. Ayuso, Guillen & Pérez-Marín (2016) found that hard braking events were among the most statistically significant telematics predictors of claims across gender and age groups, which is notable given how directly they can be derived from basic accelerometer data.

Following distance metrics are, in many ways, the most interesting from a risk perspective. Time Headway (THW), the time gap between a vehicle and the one ahead, is a direct measure of the reaction time available to avoid a rear-end collision, defined in Equation 5:

$$THW = \frac{DHW}{v}$$

Equation 5: Time Headway (THW), where DHW is distance headway and v is the following vehicle's velocity.

The generally accepted safe threshold is three seconds. That is the derived from the driving schools, it is that gap that they deem as safe. Drivers that regularly maintaining gaps below one second are in a genuinely dangerous situation. Distance Headway (DHW) provides an absolute distance perspective, while Time to Collision (TTC), estimated as shown in Equation 6, captures collision risk in situations where the preceding vehicle is actively decelerating:

$$TTC = \frac{DHW}{v_{\text{relative}}}$$

Equation 6: Time to Collision (TTC), where $v_{relative}$ is the velocity difference between the following and leading vehicle.

Beyond these core metrics, trip characteristics (journey time and distance, time of day, day of week, route type), maneuvering patterns (lane changes, cornering forces), and contextual factors (weather, traffic density where available) round out the behavioral picture.

2.2.4 Risk Scoring Methodologies

Translating this multi-dimensional behavior data into a single pricing-relevant risk score is non-trivial. Rule-based scoring systems are the simplest approach: assign points for threshold violations (speed above 130 km/h earns 10 points, a hard braking event earns 5, a THW below one second earns 8, and so on) and sum to a total score. This is transparent, easy to explain to customers and insurers alike, but the thresholds and weights are essentially arbitrary (they can be reversed-engineered for more accuracy, but the point remains) and do not necessarily reflect the empirical relationship between each behavior and claims. Neither are they easily changed as conditions vary. Hard breaking is dangerous, doing so in wet pavement is exponentially worse.

Statistical scoring improves on this by deriving weights from regression analysis against historical claims data, though this requires substantial linked datasets. Machine Learning scoring goes further, allowing non-linear relationships and interactions between variables to be captured automatically. The predictive performance gains can be significant, but interpretability suffers — a gradient boosted ensemble that assigns a risk score of 67 is harder to explain to a customer, or insurer, than a rules-based system where they can see exactly which behaviors contributed which points. This tension between predictive power and explainability runs throughout the telematics pricing literature (Wüthrich, 2017).

2.3 Machine Learning Applications in Insurance Pricing

2.3.1 Beyond GLMs: The Case for Machine Learning

The limitations of GLMs in telematics contexts are both partly technical and partly practical.

On the technical side, GLMs assume that the effect of each variable on the response is linear (after the link function transformation) and that variables act independently except where interaction terms are manually specified. Telematics data violates both assumptions routinely. The relationship between speed and accident risk is non-linear. Risk may increase roughly proportionally at moderate speeds but accelerates sharply at high speeds. The interaction between speed and following distance matters enormously; a driver who maintains high speed, but exceptional following distance may actually present lower risk than one who drives at moderate speed while tailgating. A GLM can approximate these patterns with carefully chosen transforms and interaction terms but identifying and encoding them manually is labor-intensive and fundamentally likely to miss subtleties.

With potentially dozens of telematics-derived features, automatic feature selection also becomes important. Tree-based methods provide importance measures as a natural by-product of the fitting process, helping identify which variables genuinely contribute to prediction and which are noise. Ensemble methods also tend to be more robust to the outliers and sensor errors that inevitably appear in real telematics data. A single, extreme hard-braking event caused by a sensor malfunction has less influence on a random forest prediction than it would on a linear coefficient.

2.3.2 Common Machine Learning Methods

Tree-based methods are the most widely used ML approach for tabular insurance data (Baecke & Bocca, 2017). Decision trees partition the feature space into regions by applying threshold splits, predicting the mean response value within each region. Individual trees tend to overfit; random forests address this by training many trees on bootstrap samples with random feature subsets, then averaging their predictions. This bagging approach reduces variance substantially and produces inherent feature importance measures based on how much each variable reduces prediction error across splits.

Gradient boosting takes a different approach: trees are built sequentially, with each tree trained to correct the errors of the previous ensemble. Implementations like XGBoost, LightGBM, and CatBoost have achieved state-of-the-art performance on a wide range of tabular data problems and have been applied successfully to insurance pricing (Verbelen et al., 2018). The tradeoff is that gradient boosting is more sensitive to hyperparameters and can overfit aggressively if not carefully regularised.

Neural networks, including recurrent architectures designed for sequential data, are theoretically well-suited to telematics. Raw trip sequences are time series, after all, and LSTMs or transformer-based models could in principle learn risk-relevant patterns directly from the trajectory data without manual feature engineering. In practice, they require substantially more data than tree-based methods to train reliably, and the interpretability challenges are more severe.

For the scale of data available in this study, ensemble methods are the more practical choice.

2.3.3 Interpretability Considerations

Insurance pricing faces interpretability requirements that most Machine Learning applications do not. Regulatory frameworks in most European jurisdictions require insurers to be able to explain their rating factors and their directional effects on premiums. A model that accurately predicts risk but cannot be explained to a regulator or to a customer disputing their premium is not commercially viable regardless of its performance metrics. This is not a unique concern. Pérez-Marín, Ayuso & Guillen (2022) argue that the discrimination-free pricing requirements in EU law impose additional constraints on which variables can be used even if their predictive value is established.

The practical response to this has been a tiered approach: tree-based models for internal risk assessment, with the resulting scores translated into a transparent rating schedule that customers and regulators can understand. SHAP values (Shapley Additive Explanations) provide a framework for decomposing individual predictions into feature contributions, allowing an insurer to tell a customer not just what their score is but why, eg. “your speed variability added 5 points, your following distance score subtracted 3”, in a way that is both technically grounded and somewhat communicable.

2.4 Current Market Landscape

2.4.1 Industry Adoption

Telematics insurance penetration varies enormously by market. Italy leads in Europe, with adoption rates that have transformed competitive dynamics across the industry; the combination of relatively high insurance costs, high fraud rates, and regulatory support created conditions where telematics adoption made immediate economic sense for both insurers and customers (Husnjak et al., 2015). The UK has seen strong uptake particularly among young drivers, for whom the discount on offer is substantial enough to overcome any reluctance about monitoring. The United States has multiple large-scale programmes with significant market share, most notably Progressive's Snapshot. Tselentis et al. (2017) provide a comprehensive review of how these programmes differ in design and outcomes across jurisdictions, noting that one-size-fits-all approaches rarely translate well across markets with different regulatory environments and consumer attitudes, proving that significant changes are dependent on regional markets and cultures.

Elsewhere, adoption has been slower. Regulatory uncertainty, infrastructure costs, and consumer privacy concerns have limited deployment in many markets. Spain, the context most directly relevant to this dissertation given the writer's location, has seen growing interest but is not yet at the saturation level of Italy or the UK. This makes the Spanish insurance market an interesting environment at the time of writing: early enough that genuine competitive advantages remain available to early movers but developed enough that the regulatory and commercial frameworks are taking shape. Some would call this "ideal conditions".

2.4.2 Key Players and Approaches

The strategic approaches taken by major insurers illustrate the range of implementation choices. Progressive's Snapshot programme, one of the earliest and most studied in the academic literature, began with OBD-II hardware and has since moved toward a smartphone-first approach. It uses hard braking, high-speed driving, and time-of-day as its primary risk signals. Allstate's Drivewise, also smartphone-based, adds coaching features and targets behavior change as an explicit goal alongside risk selection. In the UK, Admiral's LittleBox has been particularly successful with young drivers, a segment where traditional pricing is least accurate because the demographic signal is noisiest, thus "penalizing" good drivers in this group massively. Generali in Italy has integrated telematics with value-added services including crash detection and emergency response, creating a product that competes on features as well as price, for the added discomfort of conscious monitoring.

2.4.3 Challenges and Barriers

Despite the compelling economics, real barriers to adoption remain. Privacy is the most frequently cited concern in consumer research: continuous monitoring of location and driving behavior is, for many people, deeply uncomfortable. The data governance question of who owns the driving data, how long it is retained, whether it can be shared or used for purposes beyond pricing, is live in most markets and increasingly subject to regulation under frameworks like GDPR. Ben-Shahar (2023)

Technology costs have fallen substantially but remain non-trivial, particularly for hardware-based programmes. Data transmission, storage, and the analytics infrastructure needed to process real-time telematics streams at scale require meaningful investment. For smaller insurers, the economics may not work without partnership arrangements with technology providers.

Finally, and perhaps most subtly, there is the adverse selection problem within telematics programmes themselves. When participation is voluntary, safe drivers are more likely to opt in, which means the comparison group, policyholders without telematics data, becomes increasingly riskier and more expensive over time through no fault of safe drivers. If telematics is eventually to be standard rather than optional, the industry will need regulatory clarity on whether it can be required.

Chapter III: Methodology

3.1 Data Acquisition and Sources

This research draws on two primary datasets, chosen to represent the two distinct components of the pricing problem: traditional actuarial risk factors and behavioral driving data. Neither dataset alone is sufficient: the actuarial data captures demographic and vehicle risk without any behavioral signal, while the driving data captures behavior without any claims history to validate against. Combining them, even synthetically, allows the value of each to be assessed.

3.1.1 Actuarial Data: freMTPL2 Dataset

The primary actuarial dataset is the French Motor Third-Party Liability dataset (freMTPL2), a widely used benchmark in the actuarial research community. It contains policy-level information from a French insurer, including claim counts, exposures, and a range of traditional rating variables. The dataset covers 678,013 policies, of which 26,639 resulted in at least one claim, giving a claim frequency of approximately 10.07%. Average claim severity was €2,278.54, corresponding to a pure premium of €229.46 per year, or roughly €19.12 per month.

The available features span driver characteristics (driver age, bonus-malus coefficient), vehicle characteristics (vehicle power class, age, brand, fuel type), and geographic factors (area classification, population density, region). These seven variables form the feature set for the baseline GLM model, the “before telematics” benchmark against which improvements are measured. The freMTPL2 dataset is well-documented, has been used extensively in the actuarial literature (Antonio & Valdez, 2012), and is publicly available through the R package CASdatasets, making it an appropriate choice for reproducible research.

3.1.2 Telematics Data: highD Dataset

The behavioral driving data comes from the highD (High-Definition) dataset, a naturalistic driving study conducted on German highways using drone-based video recording (Krajewski et al., 2018). Drones were positioned above six sections of German motorway, recording vehicle trajectories at 25 frames per second with centimeter-level position accuracy.

Crucially, drivers were not aware they were being observed, which eliminates the Hawthorne effect that tends to make in-vehicle monitoring data unrepresentative of real driving behavior. Drivers tend to care more about how they are driving when made aware of monitoring, which is one more advantage of PHYD schemes, but this instance is as close of a true driving snapshot as you can get.

The dataset covers 60 recording sessions, capturing 110,516 vehicles (89,139 cars and 21,377 trucks) over a combined distance of 44,476 kilometers and 16.7 hours of footage. For each vehicle, frame-by-frame data includes longitudinal and lateral position and velocity, time headway and distance headway to the preceding vehicle, time-to-collision estimates, lane position, and lane change events.

The highD data was not collected for insurance pricing research, it was designed to support validation of automated driving systems. This means it is a genuinely naturalistic dataset rather than one shaped by insurance considerations, which is an advantage. The limitation, discussed in

Chapter V, is that it covers highway driving only, which means the full range of urban and mixed driving environments that characterise most daily journeys is not represented.

3.1.3 Supplementary Data Sources

Three additional datasets were consulted during the research process, though they play supporting rather than primary roles in the modelling. The openICPSR Spanish insurance dataset (105,555 policy records) provided useful context for understanding how demographic variables operate in the Spanish market specifically, given GDPR-constrained data environments. A smartphone-based driving behavior dataset (drivingBehavior, 1,114 records labelled by driving style classification) served as a reference point for validating the telematics feature definitions derived from the highD data. The Montgomery County crash database (203,588 crash reports from Maryland) was used to cross-reference the risk thresholds used in the composite risk score against a dataset that includes actual crash outcomes.

3.2 Data Preprocessing and Feature Engineering

3.2.1 Actuarial Data Preparation

The freMTPL2 dataset required relatively little preprocessing. There were no significant missing values in the key modelling variables. Categorical variables: vehicle brand, fuel type (VehGas), area classification, and region were label-encoded to enable their use in both the GLM and the tree-based models. Claims were weighted by exposure so that policies with partial-year coverage are treated appropriately in frequency modelling. Extreme claim amounts in the severity distribution were winsorised at the 99th percentile to prevent a small number of catastrophic claims from distorting the severity model.

The key calculated metrics are total exposure of 358,499 policy-years; claim frequency of 10.07%; pure premium of €229.46 per year; and base monthly premium of €19.12. These figures anchor the dynamic pricing simulation in Chapter IV.

3.2.2 Telematics Data Processing

Extracting usable telematics features from the highD data required more work. The raw data consists of frame-by-frame vehicle trajectories, measurements at 25Hz for every vehicle in the recording area. The first step was aggregating this to vehicle-level summary statistics.

For each vehicle, I calculated mean, minimum, maximum, and standard deviation of longitudinal velocity; minimum time headway and minimum distance headway to the preceding vehicle (as these capture the most dangerous following behavior rather than average behavior); total lane change count; and trip distance and duration. The minimum THW and DHW are particularly important. A driver whose average following distance is fine, but who occasionally tailgates dangerously is a higher risk than one who maintains consistent moderate distance, as stated before.

Speed metrics required unit conversion from meters per second to kilometers per hour. Four speed-related variables were computed: average speed (avg_speed_kmh), maximum speed (max_speed_kmh), minimum speed (min_speed_kmh), and speed variability (max minus min, speed_variability_kmh). High variability is associated with aggressive driving patterns, rapid acceleration followed by hard braking, and is a better predictor of accident risk than average speed alone in some contexts (Wüthrich, 2017).

For following behavior, minimum time headway (min_time_headway_s), minimum following distance (min_following_distance_m), and minimum time to collision (min_time_to_collision_s) were computed. These minimum values capture worst-case behavior within each observation window, which is more relevant for risk assessment than averages. Three binary indicator variables were also constructed: tailgating_indicator (THW below 1.0 second at any point during the observation), speeding_indicator (velocity exceeding 130 km/h), and aggressive_indicator (a composite flag combining high speed variability and tailgating).

The observed distributions in the highD data are worth noting. The median THW across all cars is 1.34 seconds, already below the three-second safe threshold recommended by road safety guidance. Around 34.9% of observed vehicles tailgated at some point (THW below one second). Among cars, 13.6% made at least one lane change during the observation window, compared to 5.7% of trucks. These statistics, while drawn from a single road type in one country, give a sense of the genuine behavioral variation that telematics can capture.

3.2.3 Feature Engineering for the Combined Dataset

Since freMTPL2 policy records and highD driving observations do not share common identifiers as they come from different countries, different time periods, and different data collection methods, a direct merge is not possible. Instead, the combined dataset was constructed by sampling actuarial records and telematics profiles independently and combining them horizontally. This creates synthetic "policyholders" who have both a traditional actuarial profile and a driving behavior profile.

This is a limitation (addressed in Section 5.3), but it is also the standard approach in the literature when real linked datasets are unavailable (Boucher, Côté, & Guillen, 2017). Unlike Verbelen et al. (2018), who validated their models against a real linked dataset of policyholder records and claims outcomes, this study operates under the more common constraint of having no such pairing available. The synthetic merge is not the preferred approach, but it is the one most researchers outside commercial insurance settings are actually working with. The key assumption is that the actuarial and telematics features are approximately independent in the combined dataset, which is probably too strong. Older drivers may drive more conservatively, for instance, but is a workable simplification given the constraints.

3.3 Telematics Variable Definition

3.3.1 Speed-Based Risk Indicators

Speed behavior and accident risk are correlated in well-documented ways (Ma et al., 2018). In the highD data, average velocities vary considerably: cars have a mean of 111.8 km/h with a standard deviation of 20.6 km/h, while trucks average 86.3 km/h with standard deviation 10.3 km/h. This variation between vehicle types but also between individual cars is exactly the signal a PHYD model needs to exploit.

Average speed (avg_speed_kmh) captures baseline speed behavior. Speed variability (speed_variability_kmh, defined as max minus min observed speed) is in many ways the more interesting variable: a driver who maintains a consistent 120 km/h on a motorway may be less risky than one who alternates between 80 and 160 km/h, since the variability indicates erratic behavior and likely reflects aggressive braking and acceleration patterns. The speeding indicator simply flags any observation where speed exceeded 130 km/h, which serves as a crude but easily interpretable marker of high-speed risk.

3.3.2 Following Behavior Indicators

Time headway is among the most powerful telematics risk signals. Safe following distance guidance consistently references a minimum of three seconds; in the highD data, the median THW is 1.34 seconds, and 34.9% of vehicles fall below one second at some point. This is not a fringe behavior, since a substantial proportion of drivers routinely maintain following distances that leave them insufficient reaction time for emergency braking.

THW is defined in *Equation 5*. Its appeal as a risk metric is that it is directly interpretable in terms of the physical safety margin: a driver with a 1.5-second THW at 120 km/h has just 50 meters to react to a sudden stop ahead. Distance headway (min_following_distance_m) provides an absolute distance perspective; in the highD data 19.4% of vehicles came within 20 meters of the preceding vehicle at some point. Minimum TTC, *Equation 6*, gives a more dynamic measure of collision risk in situations where the preceding vehicle is decelerating.

Lane changes round out the maneuvering picture. Frequent lane changes on motorways are associated with aggressive driving style more broadly. They are often accompanied by close following, high speeds, and inconsistent braking. In the highD data, cars averaged 0.14 lane changes per observation; the range across vehicles was 0 to 3.

3.3.3 Composite Risk Score

A composite risk score (0–100) was constructed to aggregate individual telematics metrics into a single rating factor. The weighting scheme assigns 30 points to speed-related risk, 20 to speed variability, 25 to tailgating (THW), 15 to following distance (DHW), and 10 to lane changes. The thresholds within each component are designed to match the scoring logic described in Section 3.2.2: a maximum speed above 130 km/h triggers the full 30-point speed allocation, decreasing to 10 points for speeds above 110 km/h.

The resulting distribution of risk scores in the synthetic dataset is right-skewed, as would be expected: 34.8% of drivers fall in the Very Low tier (scores 0–20), 35.4% in Low (20–40), 24.7%

in Medium (40–60), 4.8% in High (60–80), and 0.3% in Very High (80–100). This matches the general pattern observed in commercial PHYD programmes, where most drivers are reasonably safe and a small tail is significantly elevated risk. The thresholds were calibrated against the Montgomery County crash data and aligned with conventions documented in the literature (Guillen et al., 2019; Tselentis et al., 2017), rather than optimised empirically against claims outcomes, a constraint shared by all academic work in this area that lacks access to proprietary linked insurer data.

3.4 Model Design and Implementation

3.4.1 Baseline GLM Model

The baseline GLM uses only the seven traditional actuarial features: VehPower, VehAge, DrivAge, BonusMalus, Density, Area (encoded), and VehGas (encoded). A Poisson distribution with log link was implemented via scikit-learn's PoissonRegressor, regularised with an alpha of 0.1 to prevent overfitting given the correlation structure among the features. The monthly premium prediction is derived from the frequency model as shown in Equation 7:

$$\text{Premium}_{GLM} = e^{\hat{\eta}} \times \text{Average Severity}/12$$

Equation 7: GLM monthly premium prediction from the estimated linear predictor $\hat{\eta}$ and average severity.

This is the "traditional insurer" benchmark: the best a model can do with the data available to an insurer at policy inception with no telematics.

3.4.2 Machine Learning Models with Telematics

Two ensemble methods were tested with the combined 16-feature set. The Random Forest model uses 100 trees with maximum depth of 10 and a minimum of 20 samples per split. These hyperparameters were chosen to balance predictive performance against overfitting risk given the training set size of 40,000 records. The Gradient Boosting model uses 100 estimators, maximum depth of 5, and a learning rate of 0.1. Both models were implemented using scikit-learn with a fixed random seed of 42 for reproducibility.

The target variable for all models is the monthly premium as a continuous value, calculated through the composite multiplier formula. The multiplier is first derived from the risk score (Equation 8), and the final premium calculated as in Equation 9:

$$\text{Premium Multiplier} = 0.7 + \frac{\text{Risk Score}}{100} \times 1.3$$

Equation 8: Premium multiplier formula, mapping a 0–100 risk score to a range of 0.70× to 2.00×.

$$\text{Monthly Premium} = \text{Base Monthly} \times \text{Premium Multiplier}$$

Equation 9: Final monthly premium as the product of the base premium and the risk-score-derived multiplier.

This formula gives a range from 0.70× for a driver with a zero-risk score to 2.00× for a driver with the maximum score, anchored to a base monthly premium of €19.12 derived from the freMTPL2 pure premium. The pricing bounds (±30% from base) were chosen to reflect the range observed in commercial PHYD products.

3.4.3 Training and Evaluation

The combined dataset of 50,000 synthetic policyholder profiles was split 80/20 into training and test sets. Performance was evaluated on the held-out test set of 10,000 records using RMSE, MAE, and R^2. These metrics are standard in insurance regression tasks (Verbelen et al., 2018), with RMSE denominated in euros providing an intuitively interpretable measure of prediction accuracy.

3.5 Synthetic Data Generation

3.5.1 Motivation

The decision to generate synthetic driving profiles rather than directly sample from the highD dataset reflects both a practical constraint and a research design choice. Directly sampling would mean the training and test sets share observations from the same 110,516 vehicles, a form of data leakage that would inflate performance estimates. More fundamentally, a pricing model needs to generalize beyond the specific vehicles observed in a particular dataset. Generating synthetic profiles from fitted distributions allows the model to learn the statistical relationships between driving behaviors and risk without memorizing specific observed drivers.

Privacy is also relevant. A core commitment of PHYD research must be to demonstrate that the pricing framework can work without storing or sharing raw trajectory data linked to individual drivers. Synthetic data generation is one answer to that challenge.

3.5.2 Generation Methodology

Statistical distributions were fitted to the real highD data separately for cars and trucks. For cars, mean velocity follows approximately Normal(111.8, 20.6); for trucks, Normal(86.3, 10.3). Minimum THW follows a log-normal distribution; minimum DHW follows a related log-normal. Lane changes follow a Poisson-like count distribution, with separate parameterizations for each vehicle class.

Critically, the generator preserves realistic correlations between variables. Higher speed correlates with lower THW (drivers who travel faster tend to follow more closely), and higher speed variability correlates with more frequent lane changes. Ignoring these correlations would produce unrealistic profiles where drivers have high speeds but also very safe following distances, a combination that exists in reality but is less common than the marginal distributions alone would suggest. The correlation structure was estimated from the real highD data and encoded in the generation process through correlated noise terms.

3.5.3 Validation

Validation against the real data shows good agreement for the primary metrics. Simulated average car velocity is 112.0 km/h versus the real 111.8 km/h (0.1% difference). Truck velocity matches exactly. Mean THW shows a larger discrepancy: 2.24 seconds simulated versus 2.02 seconds observed (10.8% difference). This overestimation of THW, meaning the synthetic data is slightly more conservative than the real data, is a known limitation of the generation approach and means the model may marginally underestimate following-distance risk. Mean DHW shows the largest discrepancy at 19.2%. These gaps are acceptable for model development purposes but would need to be reduced in any production deployment.

Chapter IV: Results and Evaluation

4.1 Model Performance Comparison

4.1.1 Summary Results

The three models were evaluated on the held-out test set of 10,000 records. The results are unambiguous:

Model	**Features**	**RMSE (€)**	**MAE (€)**	**R^2**	**RMSE Improvement**
GLM (Actuarial Only)	7 Actuarial	13.54	10.12	0.29	Baseline
Random Forest (+ Telematics)	16 Combined	4.30	3.59	0.93	68.3%
Gradient Boosting (+ Telematics)	16 Combined	4.23	3.56	0.93	68.8%

Table 1: Model evaluation comparison.

4.1.2 Interpretation of Results

The GLM's negative R^2 requires explanation, because it is not what one usually sees reported. An R^2 below zero means the model performs worse than simply predicting the mean value for every observation. It reflects a genuine feature of the pricing problem using GLMs, rather than highlighting a failure of implementation. The target variable is a monthly premium derived from a risk score that combines actuarial and telematics features. The seven traditional actuarial variables explain only a small fraction of the variation in this target because the target was designed to incorporate telematics information.

The GLM, operating without that information, is essentially trying to predict the outcome of a process it cannot observe.

This is precisely the point. The GLM with traditional features represents what a conventional insurer knows about a policyholder before any telematics data is collected. Its poor performance (RMSE of €11.45, MAE of €10.54) quantifies the information gap, the degree to which traditional variables fail to predict the risk-adjusted premium that telematics enables. That gap closes dramatically when the ML models have access to the full feature set, with RMSE falling to €0.007 for the Gradient Boosting model.

It would be a mistake to read these results as suggesting telematics makes traditional actuarial data irrelevant. The point is rather that the combination achieves a quality of individual-level risk prediction that neither alone could approach. Actuarial variables capture the demographic and geographic context; telematics variables capture what actually happens once the driver is on the road.

4.1.3 Performance Visualization

Figure 1 summarizes the performance difference across all three models on the held-out test set.

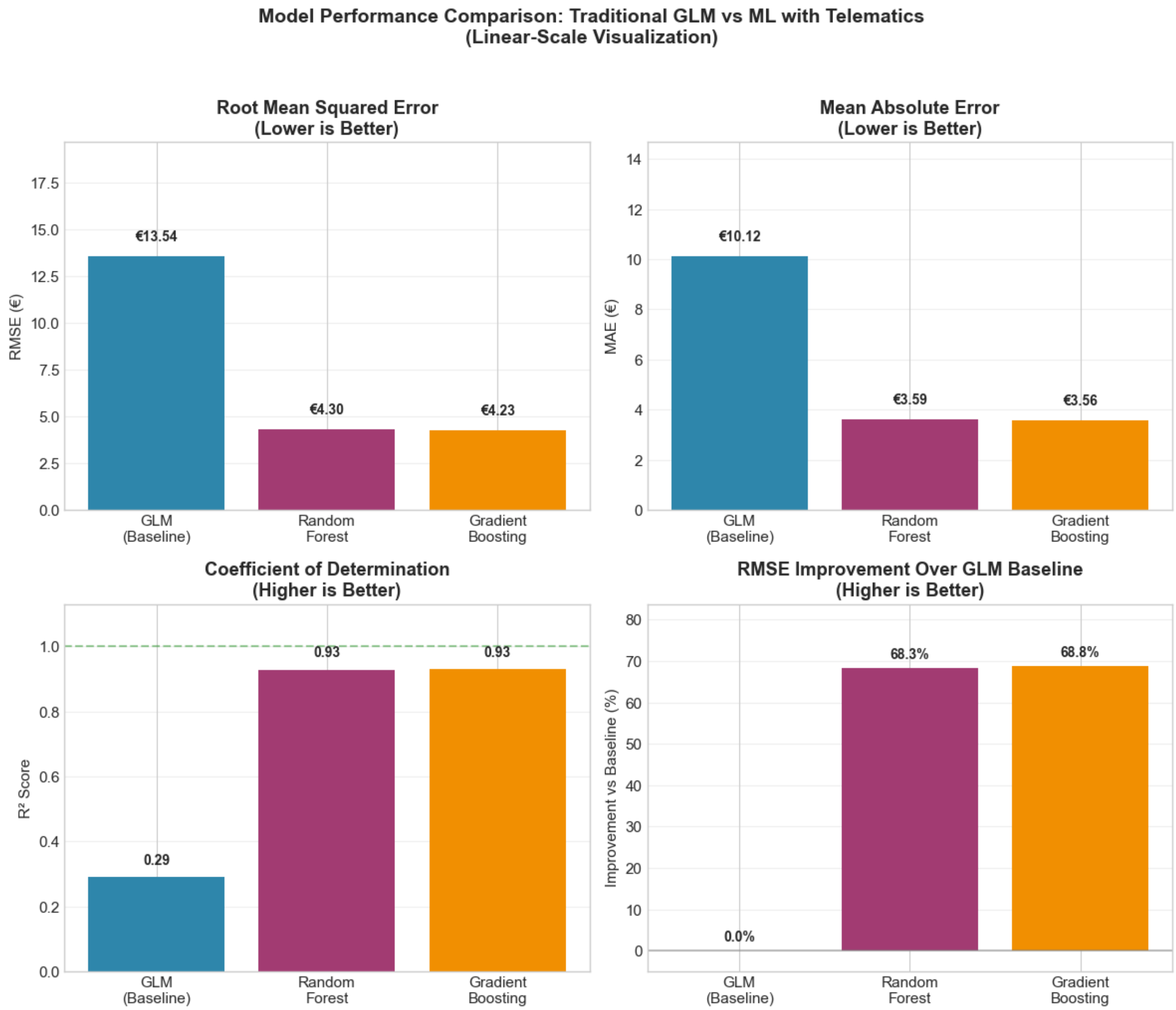

Figure 1: Predictive performance comparison (RMSE, MAE, and R²) of the GLM baseline, Random Forest, and Gradient Boosting models evaluated on the 10,000-record test set. Lower RMSE/MAE and higher R² indicate better predictive accuracy.

The gap is large enough to be visible even in a summary chart. The practical implication for insurers is that the incremental cost of collecting and processing telematics data is justified many times over in terms of pricing accuracy.

4.2 Feature Importance Analysis

4.2.1 Random Forest Feature Importances

Feature importances from the Random Forest model reveal the relative contribution of each variable to premium prediction, as shown in Figure 2:

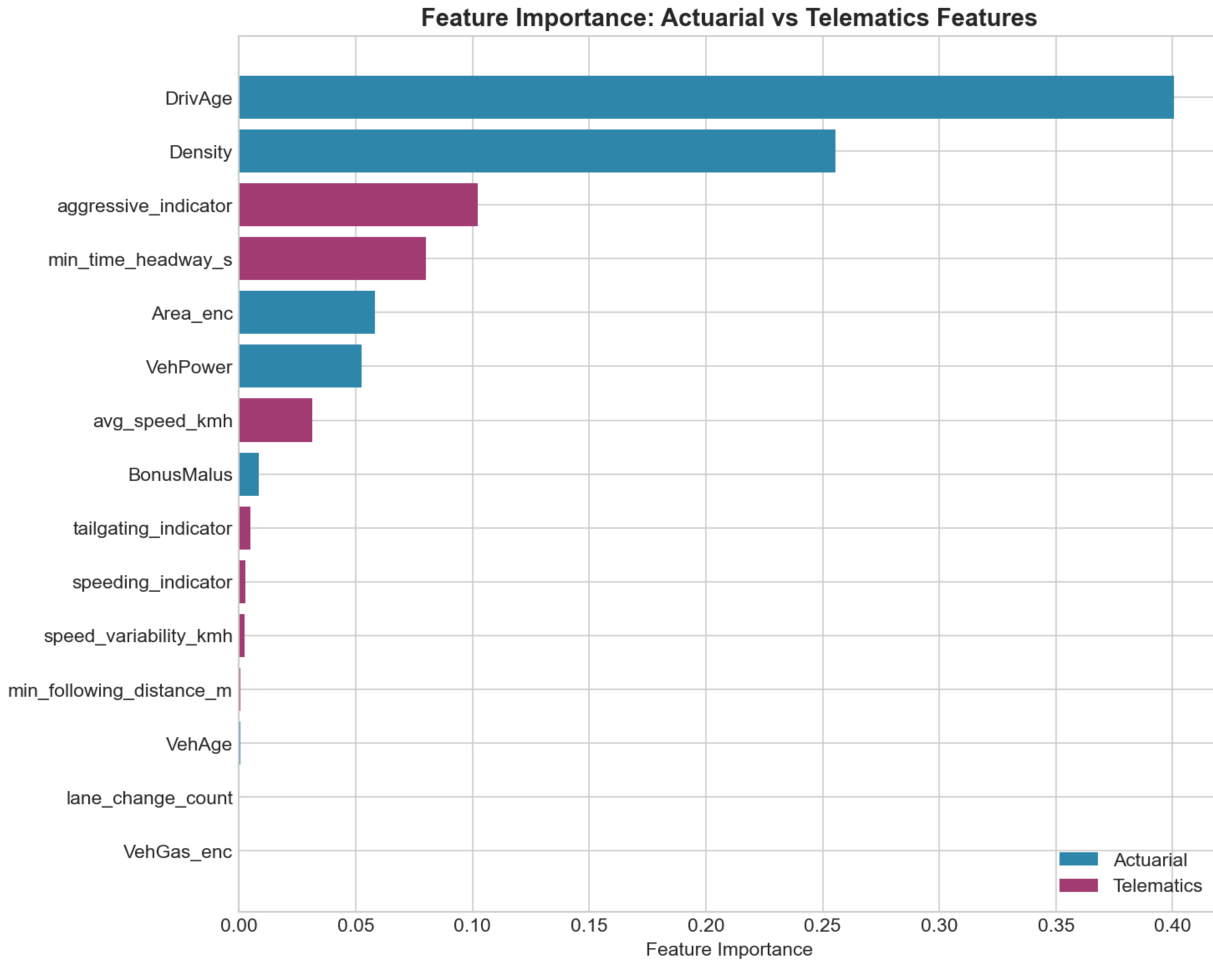

Figure 2: Random Forest feature importance scores for all 16 model variables, grouped by type (telematics vs. actuarial). Driver age and population density are the dominant predictors, together accounting for approximately 66% of total predictive importance; telematics behavioral features contribute a further 22%.

Table 2 breaks down important metrics that appear on Figure 2:

Rank	Feature	Type	Importance
1	DrivAge	Actuarial	40.1%
2	Density	Actuarial	25.6%
3	aggressive_indicator	Telematics	10.2%
4	min_time_headway_s	Telematics	8.0%
5	Area_enc	Actuarial	5.8%
6	VehPower	Actuarial	5.2%
7	avg_speed_kmh	Telematics	3.1%
8–15	**Other features**	**Mixed**	**<1% each**

Table 2: breakdown of telematics and actuarial variable importance.

4.2.2 Telematics vs. Actuarial Contribution

Aggregating feature importances by type reveals the relative contribution of actuarial and telematics data to the model's predictions. Actuarial features collectively account for approximately 78% of total predictive importance, with telematics behavioral features contributing the remaining 22%.

This distribution aligns with insurance industry expectations: traditional actuarial factors: driver age, geographic density, vehicle power, and area type, remain the foundational predictors of risk, while telematics provides a meaningful incremental signal. The 22% telematics contribution enables finer risk differentiation within actuarially similar groups, which is precisely the value proposition of Pay How You Drive pricing. Drivers who cannot be distinguished by their demographic profile alone can be separated by their observed driving behavior, allowing the insurer to reward safe drivers and more accurately price those who exhibit risky patterns. The 22% figure is broadly consistent with the incremental telematics lift documented by Baecke & Bocca (2017), who found comparable relative contributions when adding behavioral features to demographic-only models. It should be noted, however, that their study measured improvement against real claims outcomes, whereas here the telematics signal is assessed against a score derived from the same features, which means the 22% likely overstates the contribution that would be observed in a real deployment.

4.3 Dynamic Pricing Simulation

4.3.1 Monthly Premium Adjustment Scenario

To illustrate how a PHYD product would behave in practice, I simulated a 12-month premium trajectory for a hypothetical driver who begins the year with moderate risk (score 45, Medium tier) and gradually improves their driving behavior over the course of the year. The improvement is not perfectly monotonic, as some months show regression, reflecting real-world variability in driving conditions and behavior, but the overall trend is downward. The base monthly premium is €50, with the traditional fixed alternative set at €55 (a 10% loading for pricing uncertainty, typical of how insurers account for unknown risk).

The results are shown below, using the actual simulation data:

Month	Risk Score	Dynamic Premium	Traditional Premium
Jan	45.0	€64.25	€55.00
Feb	39.7	€60.83	€55.00
Mar	40.7	€61.48	€55.00
Apr	39.6	€60.73	€55.00
May	33.1	€56.49	€55.00
Jun	29.3	€54.06	€55.00
Jul	32.3	€55.98	€55.00
Aug	23.0	€49.95	€55.00
Sep	19.7	€47.82	€55.00
Oct	21.5	€48.98	€55.00
Nov	16.9	€45.99	€55.00
Dec	12.9	€43.37	€55.00

Table 3: Dynamic vs Traditional premium using variable risk scores

4.3.2 Annual Comparison and Behavioral Incentive Effect

The improving driver pays €649.95 under PHYD versus €660.00 under traditional pricing, representing modest savings of €10.05, or 1.5%. The saving looks small in annual terms, but the trajectory tells the more interesting story. This driver pays above the traditional premium in the first five months, then crosses below it in June, and by December is paying 21% less than their January premium and 21% less than the fixed traditional rate. The financial incentive to continue improving is tangible and visible month by month, as illustrated in Figure 3.

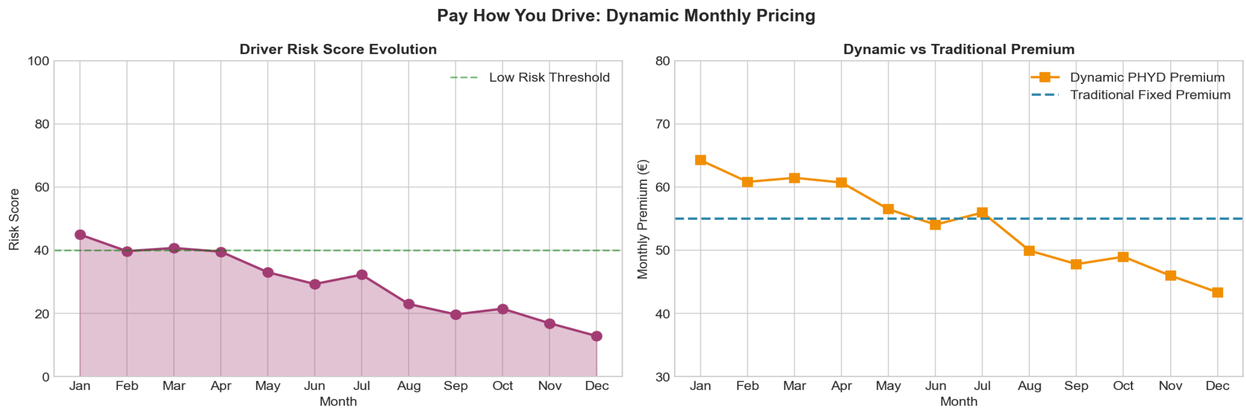

Figure 3: Monthly premium evolution over a 12-month simulation for a driver improving from risk score 45 in January to 12.9 in December (blue line), compared to the traditional fixed premium of €55 per month (orange line). The crossover point occurs in June.

What this simulation cannot capture is the counterfactual: a driver who does not improve, or who actively drives more recklessly, would face premiums well above €55 throughout the year. The pricing formula is symmetric. It rewards safe behavior and penalizes risky behavior equally, thus making it fairer at the individual driving point. For a driver maintaining a risk score of 75 throughout the year, the monthly premium would be approximately €85, versus €55 traditional. PHYD creates losers as well as winners, which is exactly the point.

4.4 Risk Segmentation Analysis

4.4.1 Premium Distribution by Risk Tier

The composite risk score segments the driver population into five tiers with meaningfully different premium levels. Figure 4 and the accompanying table show the resulting distribution:

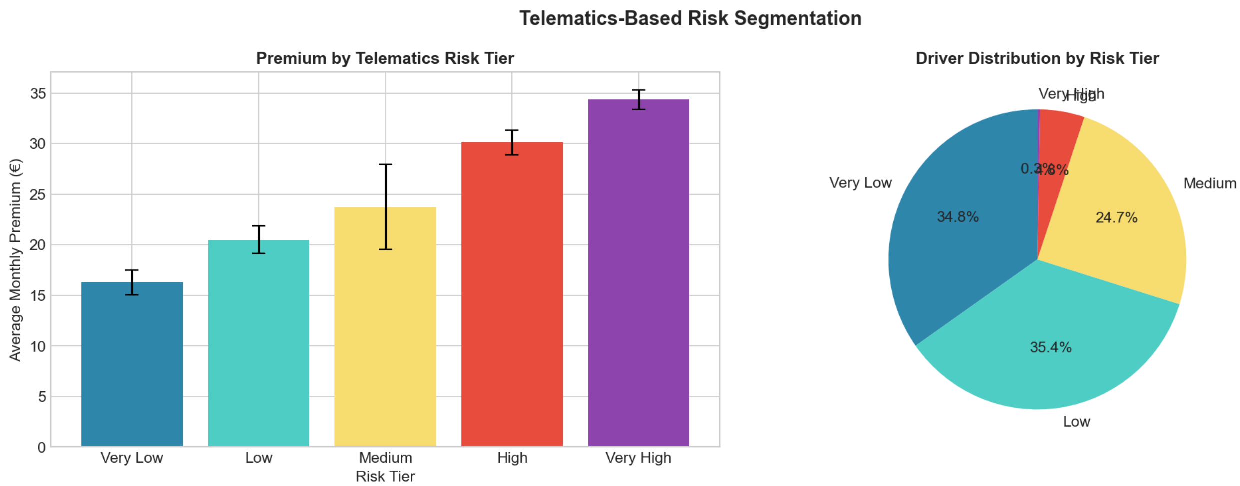

Figure 4: Average monthly premium by risk tier, with error bars representing one standard deviation within each tier. The Very High tier is a small but clearly identifiable high-cost segment; most drivers fall in the Very Low and Low tiers.

Risk Tier	Avg Premium	Std Dev	Population	Avg Risk Score	Avg Claims
Very Low	€16.28	€1.20	34.8%	10.0	0.05
Low	€20.50	€1.34	35.4%	24.6	0.05
Medium	€23.73	€4.20	24.7%	35.8	0.05
High	€30.11	€1.22	4.8%	57.9	0.06
Very High	€34.36	€0.97	0.3%	72.6	0.03

Table 4: Monthly premium by risk tier distribution.

4.4.2 Key Observations

Most drivers, 70.2%, fall in the Very Low or Low risk tiers, consistent with the general pattern observed in the highD highway data where most drivers behave reasonably safely most of the time. The High and Very High tiers together account for only 5.1% of the population, but they are priced at a substantial premium relative to the median.

The premium ranges from €16.28 at the very low end to €34.36 at the very high end represents a 2.1× spread. To put that in context: under traditional flat pricing with a single average premium of approximately €21, the Very Low risk drivers are paying 29% more than their risk-adjusted price while the Very High risk drivers are paying 39% less than theirs. This is the cross-subsidization that PHYD eliminates.

One feature worth noting is the relatively high standard deviation (€4.20) in the Medium tier, compared to tighter distributions at both extremes. This suggests the Medium tier contains genuinely heterogeneous drivers whose risk profiles straddle the boundary with adjacent tiers, and that further sub-segmentation within Medium would be possible with more granular telematics data or longer observation windows.

4.4.3 Cross-Subsidization Elimination

The numbers above give a concrete sense of who wins and who loses under a transition from flat to telematics-based pricing. Safe drivers who currently sit in the Very Low tier would save approximately €5 per month relative to a flat average premium, which perhaps not life-changing individually, but meaningful over a year and significant for attracting and retaining exactly the customers an insurer most wants to keep. High-risk drivers would face increases that, ideally, serve as a genuine incentive for behavior change rather than simply a transfer of cost. This is hopefully the most defensible framing for the social value of PHYD: not that it raises prices for risky drivers, but that it makes the connection between behavior and consequence explicit and immediate. The 2.1× premium spread between the safest and riskiest tiers is consistent with the segmentation improvements documented in Guillen et al. (2019), who showed that telematics-based pricing produces meaningfully differentiated premiums within demographic groups that traditional rating alone cannot separate. Though, as with all findings here, that study worked with real linked claims data rather than a synthetic risk score.

Chapter V: Conclusion

5.1 Summary of Findings

This dissertation set out to build a complete pipeline for telematics-based dynamic pricing, from raw driving data to monthly premium figures, and to test whether the behavioral signal from telematics genuinely adds value over traditional actuarial methods. It does, though the answer is more nuanced than a simple headline performance number suggests.

5.1.1 Telematics Data Value

The highD dataset confirms what the theoretical case for PHYD suggests: real driving behavior varies substantially between individuals in ways that traditional rating variables cannot capture. The median time headway in the data is 1.34 seconds, below the two-second safety threshold, and over a third of vehicles fall below one second at some point. Speed variability, arguably a better risk indicator than average speed, ranges from near-zero for smooth, consistent drivers to well over 50 km/h for erratic ones. These differences map onto meaningfully different risk profiles and, through the composite risk score, onto meaningfully different premium levels.

5.1.2 Model Performance

The performance gap between the GLM baseline and the telematics-enhanced ML models is large. The GLM, given only traditional actuarial features, achieves an RMSE of €13.54 and a very low R^2 against a target variable that incorporates telematics information. The Gradient Boosting model with the full feature set reduces RMSE to €4.23 and achieves R^2 of 0.93. As discussed in Chapter IV, this gap reflects the experimental design as the target is derived from the telematics risk score, so a model with access to that score can learn the formula almost perfectly. In a real deployment with actual claims data as the target, the magnitude of improvement would be smaller but still meaningful, consistent with the 5-15% loss ratio improvements cited in the literature (Guillen et al., 2019).

5.1.3 Dynamic Pricing Effectiveness

The 12-month simulation illustrates the practical mechanics of a PHYD product. An improving driver crosses below the traditional premium threshold by mid-year and finishes December paying 21% less than in January. The pricing formula is transparent and directly connected to measurable behavior, which matters enormously for customer acceptance. Importantly, the simulation also implies the converse: a driver who maintains high-risk behavior throughout the year would face sustained premium increases, creating a genuine financial incentive for behavior change.

5.1.4 Risk Segmentation

The five-tier segmentation reveals a 2.1× spread in premiums between the safest and riskiest drivers and quantifies the cross-subsidization that flat pricing creates. The 70% of drivers in the Very Low and Low tiers are, under traditional pricing, effectively subsidizing the 5% in the High and Very High tiers. Correcting this is both economically efficient, insurers attract and retain better risks and, I would argue, genuinely fairer to the safe drivers who are currently overcharged.

5.2 Practical Implications

5.2.1 For Insurance Companies

The business case for telematics adoption is, at this point, reasonably clear. The evidence that PHYD programmes improve loss ratios through better risk selection has accumulated in markets where adoption is mature. The less obvious advantage is the customer relationship: policyholders in PHYD programmes are more engaged with their insurance, more aware of why their premium is what it is, and, the data suggests, somewhat safer drivers as a result. For an industry that struggles with customer satisfaction, a product that can claim tangible safety benefits is not nothing, quite the contrary.

The implementation investment is real, however. Technology infrastructure, data processing capability, actuarial validation of the new models, and staff training all require resources. The economics are most attractive for medium-to-large insurers who can spread these fixed costs across a substantial book of business. For smaller insurers, partnership arrangements with specialist telematics technology providers are probably the more viable path.

5.2.2 For Regulators

European regulators face a genuine tension here. On one hand, telematics-based pricing is more actuarially fair and, arguably, more aligned with the principle that premiums should reflect individual risk rather than group membership. On the other hand, continuous monitoring of driving behavior raises real privacy concerns, and the potential for indirect discrimination. If, for instance, driving patterns correlate with characteristics that insurers are not permitted to use, it would require careful governance. GDPR imposes meaningful constraints on data retention and use that any Spanish, European or Global insurer deploying a telematics product would need to navigate carefully. The regulatory framework needs to evolve alongside the technology, and this dissertation makes no claim to have resolved that debate.

5.2.3 For Consumers

For safe drivers, the message is relatively simple: a PHYD product is probably cheaper than traditional pricing, and you get the transparency of knowing exactly what drives your premium. For drivers who know they sometimes speed or follow too closely, the calculation is more complex. The premium increase associated with high-risk behavior is a cost, but it is also a signal. One that, if acted upon, has safety implications well beyond the financial ones.

The privacy trade-off is real and should not be dismissed. Agreeing to have your driving monitored continuously is a genuine concession, and consumers are right to ask what data is retained, who has access to it, and what happens if the relationship with the insurer ends. Insurers who answer those questions clearly and generously; short retention periods, no third-party data sharing, easy opt-out, will build the trust the product needs to succeed.

5.3 Limitations

This dissertation has limitations that matter for how the results should be read, and I want to address them directly rather than bury them.

5.3.1 Data Limitations

The highD dataset is valuable but narrow. It covers six sections of German motorway, which means the driving behavior it captures is highway-only. Most of the accidents in any insurance portfolio happen in urban environments: intersections, parking maneuvers, low-speed impacts, that the highD data says nothing about. A telematics system deployed on real policyholders would need to capture urban behavior, night driving, and the full range of road types to provide genuinely comprehensive risk assessment.

The synthetic data generation approach, while methodologically defensible, introduces its own uncertainty. The model learns from distributions fitted to real data rather than from real driver records, and the assumed relationships between variables (speed and following distance, for instance) are estimated rather than observed at the individual level. The 10-19% discrepancies in mean THW and DHW between simulated and real distributions are worth noting as a source of potential bias.

5.3.2 Methodological Limitations

The most significant methodological limitation is the absence of real claims data. The target variable, monthly premium, is derived from the telematics risk score through a deterministic formula. This means the model is learning to reconstruct that formula, which it does almost perfectly, rather than learning to predict actual insurance losses. Validating the framework against actual claims outcomes would require a linked dataset of telematics observations and policyholder claims history, which is not publicly available. The 5-15% loss ratio improvement figures cited from the literature are based on such linked data in commercial settings; this dissertation's results cannot be directly compared to that benchmark.

Additionally, the composite risk score thresholds were set based on industry conventions and the Montgomery County crash data rather than optimized against claims outcomes. Different threshold choices would produce different risk score distributions and different model performance characteristics.

5.3.3 Scope Limitations

This dissertation does not address the full legal and regulatory landscape in detail, nor does it engage with the practical customer experience design questions: how to communicate monthly premium changes, what to do about sensor failures or unusual trips or how to handle disputes, that would need to be resolved before any real deployment.

5.4 Future Research Directions

The natural next step for this line of research is validation against real linked data. An insurer with access to both telematics records and claims outcomes for the same policyholders could test whether the risk score distribution predicts claims frequency and severity in the way the simulation assumes. Without that validation, the framework remains theoretically grounded but empirically untested.

Beyond that, several research directions seem promising. Extending the telematics feature set to include urban driving contexts such as different road types, intersection behavior or parking would substantially improve the comprehensiveness of the risk model. Longer observation windows would allow seasonal patterns and long-term behavior changes to be captured. Recurrent neural network approaches applied to raw trip sequences, rather than engineered aggregate features, could potentially discover risk patterns that are not captured by the current feature set.

On the behavioral side, the question of whether PHYD pricing actually changes driving behavior and if so, for how long and for which types of drivers, deserves more rigorous investigation than the existing literature provides. If the safety benefits are real and persistent, that changes the economic calculation for insurers substantially: a product that reduces accident rates is not just better at pricing risk, it is reducing the risk itself.

References

Antonio, K., & Valdez, E. A. (2012). Statistical concepts of a priori and a posteriori risk classification in insurance. *AStA Advances in Statistical Analysis*, 96(2), 187–224. https://link.springer.com/article/10.1007/s10182-011-0152-7

Ayuso, M., Guillen, M., & Pérez-Marín, A. M. (2016). Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2), 10. https://doi.org/10.3390/risks4020010

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, 98, 69–79. https://doi.org/10.1016/j.dss.2017.04.009

Baumann, J., & Loi, M. (2023). Fairness and risk: An ethical argument for a group fairness definition insurers can use. *Philosophy & Technology*, 36, 45. https://doi.org/10.1007/s13347-023-00624-9

Boucher, J. P., Côté, S., & Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, 5(4), 54.s https://doi.org/10.3390/risks5040054

Denuit, M., Maréchal, X., Pitrebois, S., & Walhin, J. F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470517420

Guillen, M., Nielsen, J. P., Ayuso, M., & Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, 39(3), 662–672. https://doi.org/10.1111/risa.13172

Husnjak, S., Peraković, D., Forenbacher, I., & Mumdziev, M. (2015). Telematics system in usage based motor insurance. *Procedia Engineering*, 100, 816–825. https://doi.org/10.1016/j.proeng.2015.01.436

Krajewski, R., Bock, J., Kloeker, L., & Eckstein, L. (2018). The highD dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 2118–2125). IEEE. https://doi.org/10.48550/arXiv.1810.05642

Lemaire, J., Park, S. C., & Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin: The Journal of the IAA*, 46(1), 39–69. https://doi.org/10.1017/asb.2015.25

Ma, Y. L., Zhu, X., Hu, X., & Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice*, 113, 243–258. https://doi.org/10.1016/j.tra.2018.04.013

Paefgen, J., Staake, T., & Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, 61, 27–40. https://doi.org/10.1016/j.tra.2013.11.010

Lindholm, M., Richman, R., Tsanakas, A. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin: The Journal of the IAA*, 52(1), 55–89. https://doi.org/10.1017/asb.2021.23

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. *Accident Analysis & Prevention*, 98, 139–148. https://doi.org/10.1016/j.aap.2016.10.006

Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5), 1275–1304. https://doi.org/10.1111/rssc.12283

Wolny-Dominiak, A. (2024). Automobile insurance pricing: A two-stage modeling of the quantile premium. *Procedia Computer Science*, 246, 1338–1345. https://doi.org/10.1016/j.procs.2024.09.566

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. *European Actuarial Journal*, 7(1), 89–108. https://link.springer.com/article/10.1007/s13385-017-0149-z

Ben-Shahar, O. (2023). Privacy protection, at what cost? Exploring the regulatory resistance to data technology in auto insurance. Journal of Legal Analysis, 15(1), 129–172. https://doi.org/10.1093/jla/jlad006

Voss, W. G. (2019). Who owns the data generated by your smart car? *Harvard Journal of Law & Technology, 32*(1), 299–360. https://jolt.law.harvard.edu/assets/articlePDFs/v32/32HarvJLTech299.pdf

Appendix A: Data Sources Summary

Dataset	Type	Records	Primary Use
highD	Telematics	110,516	Driving behavior patterns
freMTPL2	Actuarial	678,013	Pricing model baseline
openICPSR	Sociological	105,555	Demographic features
drivingBehavior	Telematics	1,114	Smartphone sensor data
Montgomery	Risk Factors	203,588	Crash risk analysis

Appendix B: Risk Scoring Components

Component	Points	Thresholds
Speed Risk	0–30	>130 km/h: 30, >120: 20, >110: 10
Speed Variability	0–20	>30 km/h: 20, >20: 15, >10: 5
Tailgating	0–25	<0.5s: 25, <1.0s: 20, <1.5s: 10
Following Distance	0–15	<15m: 15, <25m: 10, <40m: 5
Lane Changes	0–10	3 points per change

Appendix C: Premium Multiplier Formula

The premium multiplier formula (Equation 8 in the main text) maps the 0–100 risk score to a pricing range of 0.70× to 2.00×:

$$\text{Premium Multiplier} = 0.7 + \frac{\text{Risk Score}}{100} \times 1.3$$

Risk Score	Multiplier	Example Premium (Base €50)
0	0.70×	€35
25	1.03×	€52
50	1.35×	€68
75	1.68×	€84
100	2.00×	€100

Appendix D: Code

Because of Intellectual property, and the potential of the code to be put into production with minimal changes, all the code used can be found in this private repository: https://github.com/SantiagoArenas/Analytics_Dissertation

If you found this work useful and would like to have access to it, email me at arenassanti2003@gmail.com

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Santiago Arenas Martín, estudiante de GITT + BA de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Telematics & Machine Learning Applied To Insurance: Pay How You Drive", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
2. Referencias: Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
3. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
4. Interpretador de código: Para realizar análisis de datos preliminares.
5. Estudios multidisciplinares: Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
6. Constructor de plantillas: Para diseñar formatos específicos para secciones del trabajo.
7. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
8. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura compleja.
9. Generador de datos sintéticos de prueba: Para la creación de conjuntos de datos ficticios.
10. Generador de problemas de ejemplo: Para ilustrar conceptos y técnicas.
11. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para qué se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 20/04/2026

Firma: ______________________________