Zachary Balgut Tan zb2362
IEOR 4572 Term Paper

# Predicting Risk Behavior in Insurance with Predictive Models

Zachary Balgut Tan

## Abstract

The insurance industry increasingly relies on data-driven approaches to improve risk assessment, pricing, and underwriting decisions. While traditional actuarial methods grounded in statistical modeling and historical experience remain central to insurance operations, their ability to capture complex, non-linear relationships in modern, high-dimensional data is limited. This study develops and evaluates a predictive modeling framework for classifying policyholder risk using historical policy and claims data. Key features—including customer demographics, policy characteristics, claim frequency, and premium information—are analyzed to identify behavioral patterns associated with elevated claim likelihood.

Multiple modeling techniques are compared, including logistic regression, decision trees, random forests, gradient-boosted trees (XGBoost), and a deep learning model. Model performance is evaluated using accuracy, precision, recall, and F1-score to assess both overall classification performance and the ability to identify higher-risk policyholders. The results indicate that ensemble methods outperform simpler models, with XGBoost achieving the strongest overall performance, while logistic regression remains competitive and offers superior interpretability. Deep learning models do not demonstrate a clear performance advantage in this setting, likely reflecting the structured nature and limited dimensionality of the dataset.

Overall, the findings demonstrate that machine learning models—particularly ensemble-based approaches—can meaningfully enhance risk segmentation in insurance applications. At the same time, the results highlight the importance of balancing predictive accuracy with interpretability and practical deployment considerations. The proposed framework supports more equitable pricing, improved underwriting efficiency, and proactive risk management, contributing to the responsible adoption of advanced analytics within the insurance industry.

Zachary Balgut Tan zb2362
IEOR 4572 Term Paper

# 1. Introduction and Motivation

Risk assessment is central to the insurance business model. Insurers need to estimate the likelihood and severity of future claims to price policies correctly, allocate capital efficiently, and maintain long-term solvency. Historically, this has been done using actuarial techniques that depend on statistical assumptions, historical experiences, and relatively simple predictive frameworks. While these methods have worked well for decades, they often struggle to capture complex, nonlinear relationships in modern insurance data.

Recently, improvements in data access, computing power, and machine learning have opened up new ways to enhance traditional actuarial methods. Other industries, such as finance, e-commerce, and technology, have quickly adopted predictive analytics to improve customer segmentation, pricing, and risk management. In contrast, the insurance sector has been slower to embrace these tools because of regulatory restrictions, outdated systems, and a cautious risk culture.

At the same time, insurers face increasing competition and shifting customer expectations. More detailed pricing, fairer risk classification, and the early identification of high-risk behavior are now essential for maintaining profitability while ensuring compliance with regulations and fostering customer trust. As a result, there is growing interest in adding data science techniques to actuarial models to reveal hidden patterns in policyholder behavior.

This paper aims to develop a predictive modeling framework that uses historical policy and claims data to classify insurance customers based on their risk behavior. Instead of replacing actuarial methods, the goal is to show how machine learning models can complement traditional approaches by improving predictive accuracy and interpretability. The main research question of this study is: Can machine learning techniques enhance the identification and segmentation of high-risk insurance policyholders compared to conventional statistical models?

# 2. Industry and Literature Context

The actuarial foundations of insurance pricing are built on probability theory, credibility theory, and generalized linear models (GLMs) (McCullagh & Nelder, 1989; Frees, 2010). Logistic regression and Poisson regression have long been used to model claim frequency and severity due to their interpretability and alignment with insurance loss distributions. These methods allow actuaries to incorporate rating variables such as age, location, and policy characteristics while maintaining transparency and regulatory acceptance.

However, a growing body of academic and industry literature has highlighted the limitations of linear and parametric models when dealing with high-dimensional and nonlinear data. Studies have shown that decision trees and ensemble methods—such as random forests and gradient boosting machines—can outperform traditional GLMs in predictive accuracy, particularly when interactions between variables are complex or difficult to specify a priori (Wuthrich, 2018; Henckaerts et al., 2018).

Recent research in insurance analytics has explored the use of machine learning for fraud detection, lapse prediction, telematics-based auto insurance pricing, and claims severity modeling (Boodhun & Jayabalan, 2018). These studies consistently find that ensemble models offer improved classification performance, though often at the cost of reduced interpretability. This tradeoff has become a central consideration in insurance applications, where regulatory transparency and fairness are critical.

Despite these advances, many insurers remain cautious about deploying machine learning models in production environments. Concerns around explainability, bias, and governance have slowed adoption (NAIC, 2020). As a result, there is a need for applied studies that compare traditional and machine learning approaches using realistic insurance datasets, while emphasizing interpretability and practical relevance.

This project contributes to the existing literature by evaluating multiple modeling techniques within a single, unified framework. By directly comparing logistic regression, decision trees, and ensemble methods on the same dataset, the analysis provides insight into the strengths and limitations of each approach from both a predictive and business perspective.

# 3. Data Description and Sources

For this analysis, we leveraged the publicly available **Prudential Life Insurance Assessment dataset** from the Kaggle competition platform, which was curated to support predictive modeling in insurance risk assessment (Prudential Life Insurance Assessment, 2025). The dataset consists of a rich collection of policyholder characteristics, demographic information, and outcome labels that reflect claim severity and risk indicators, making it well-suited for comparing traditional statistical models with modern machine learning approaches. By using a real-world dataset with diverse feature types and real insurance outcomes, this project demonstrates how advanced analytics can be applied to practical underwriting and pricing challenges while providing a benchmark for evaluating model performance across a range of techniques.

Key variables in the dataset include demographic characteristics such as age and gender, policy attributes such as coverage type and premium amount, and claims-related metrics including claim frequency and historical loss indicators. The target variable for modeling is a binary indicator representing whether a policyholder exhibits high-risk behavior, defined as filing claims above a specified threshold during the observation period.

While the dataset provides a rich view of policyholder behavior, it is not without limitations. Certain behavioral factors—such as driving habits or lifestyle characteristics—are not directly observed. Additionally, the data reflects historical conditions and may not fully capture future shifts in customer behavior or market dynamics. These limitations are discussed further in a later section.

# 4. Data Preprocessing and Feature Engineering

Prior to model development, substantial preprocessing and feature engineering were conducted to ensure data quality, reduce dimensionality, and align the dataset with the assumptions of the selected modeling techniques. Insurance datasets often contain missing values, highly correlated features, and a mix of numerical and categorical variables, all of which require careful treatment to avoid biased or unstable model estimates.

## 4.1 Variable Selection and Initial Filtering

The raw dataset contained a wide range of policy, demographic, medical, and insurance history variables. From this initial set, a subset of features was selected based on relevance to underwriting risk assessment and data completeness. Variables retained included policy attributes (e.g., product information), demographic characteristics (age, height, weight, body mass index), employment information, medical history indicators, insurance history variables, and family history features. The target variable, representing underwriting response, was preserved for later transformation into a binary risk classification.

Observations with missing values in critical employment-related variables were removed, as these records lacked sufficient information to reliably assess risk. This step reduced noise in the dataset while maintaining a large enough sample size for robust model training.

## 4.2 Handling Missing Values

Missing data was addressed using a combination of deletion and imputation strategies, chosen based on both statistical considerations and domain knowledge. For variables where missingness was minimal but information content was essential—such as employment-related features—rows with missing values were removed entirely.

For medical and family history variables, missing values were more prevalent and potentially informative. Rather than discarding a large portion of the dataset, missing family history variables were imputed with zeros and aggregated under the assumption that missing entries likely corresponded to an absence of reported family history rather than an unknown condition. Similarly, missing values in one medical history variable were imputed using information from a related medical history indicator, leveraging internal consistency within the dataset to preserve signal while limiting distortion.

## 4.3 Feature Aggregation and Dimensionality Reduction

To improve model stability and interpretability, several related variables were combined into composite features. Multiple insurance history indicators were aggregated into a single insurance history score, summarizing a policyholder's past insurance behavior into a unified metric. This aggregation reduces multicollinearity while capturing the cumulative effect of prior insurance outcomes on current risk.

Family history variables were similarly aggregated into a single family history score by summing across multiple family-related indicators. This approach preserves the overall influence of family medical background on risk classification while simplifying the feature space and reducing sparsity.

After aggregation, the original component variables were removed from the dataset to prevent redundancy and overrepresentation of closely related information.

## 4.4 Target Variable Construction

The original underwriting response variable was transformed into a binary risk classification to support supervised classification models. Response categories associated with lower underwriting concern were grouped into a "not risky" class, while higher response categories were classified as "risky." This binary transformation aligns with practical underwriting decisions, where policyholders are often segmented into broad risk tiers rather than granular ordinal categories.

The resulting target variable was encoded numerically to facilitate model training, with 0 representing lower-risk policyholders and 1 representing higher-risk policyholders. Class distribution analysis indicated a relatively balanced dataset, reducing concerns related to severe class imbalance.

## 4.5 Feature Scaling and Normalization

To ensure comparability across features and improve model convergence, numerical variables were normalized using min–max scaling. This transformation maps all continuous variables to a common range between 0 and 1, preventing features with larger numeric ranges—such as weight or insurance history—from disproportionately influencing model estimates.

Normalization is particularly important for distance-sensitive models and optimization-based methods, including logistic regression and neural networks. Applying consistent scaling across training and validation sets ensures stable and interpretable model performance.

## 4.6 Encoding of Categorical Variables

The dataset included a categorical product information variable that captures differences in policy structure and coverage characteristics. This variable was encoded using one-hot encoding, with one category dropped to avoid perfect multicollinearity. The encoded features were then concatenated with the normalized numerical variables to form the final modeling dataset.

This encoding approach allows tree-based and linear models alike to incorporate categorical information without imposing artificial ordinal relationships between categories.

### 4.7 Train–Validation Split

Following preprocessing, the finalized dataset was divided into training and validation subsets using an 80–20 split. The training set was used for model estimation, while the validation set was reserved for out-of-sample performance evaluation. This separation ensures that reported results reflect generalization ability rather than in-sample fit.

## 5. Modeling Framework and Algorithm Selection

Following data preprocessing and feature engineering, multiple supervised learning models were implemented to classify insurance policyholders into high-risk and low-risk categories. The modeling strategy was designed to balance predictive performance, interpretability, and practical applicability within an insurance underwriting context. Rather than relying on a single algorithm, a range of models with varying complexity was evaluated to assess tradeoffs between transparency and accuracy.

### 5.1 Baseline Model: Logistic Regression

Logistic regression was selected as the baseline model due to its long-standing role in actuarial science and insurance analytics. The model estimates the probability that a policyholder belongs to the high-risk category as a function of the explanatory variables, using a logit link to constrain predictions between zero and one. Coefficients can be interpreted as the marginal effect of each feature on the log-odds of being classified as risky, making the model particularly attractive from a regulatory and governance perspective.

Despite its simplicity, logistic regression provides a strong benchmark against which more complex models can be evaluated. Regularization and an increased iteration limit were applied to ensure numerical stability and convergence given the dimensionality introduced by one-hot encoding. Model performance was evaluated on the validation set using accuracy, precision, recall, and F1-score, with particular attention paid to recall due to the asymmetric cost of misclassifying high-risk policyholders.

## 5.2 Decision Tree Classifier

To capture nonlinear relationships and feature interactions that logistic regression may fail to identify, a decision tree classifier was implemented. Decision trees partition the feature space into a series of hierarchical decision rules, allowing the model to learn threshold-based relationships in the data. This structure closely mirrors human decision-making processes and aligns well with underwriting logic.

While decision trees offer strong interpretability, they are prone to overfitting, particularly in high-dimensional datasets. To assess their standalone effectiveness, an unconstrained decision tree was trained and evaluated. The resulting performance highlighted the model's ability to capture nonlinear patterns but also demonstrated reduced generalization relative to ensemble approaches.

## 5.3 Ensemble Learning: Random Forest

To address the instability and variance of single decision trees, a random forest classifier was employed. Random forests aggregate predictions from multiple decision trees trained on bootstrap samples of the data, with random feature selection at each split. This ensemble approach reduces variance while preserving the ability to model complex interactions.

Random forests are particularly well-suited for insurance data, where relationships between risk drivers are often nonlinear and interdependent. Feature importance metrics derived from the model provide insight into which variables contribute most to risk classification, supporting interpretability and model validation. The random forest model demonstrated improved performance across most evaluation metrics relative to both logistic regression and the standalone decision tree.

## 5.4 Gradient Boosting: XGBoost

Gradient boosting methods represent a more sophisticated ensemble approach by sequentially training models to correct the errors of prior iterations. In this study, an XGBoost classifier was implemented due to its strong performance in structured tabular data applications.

XGBoost optimizes a differentiable loss function using gradient descent while incorporating regularization to prevent overfitting. This makes it particularly effective for classification tasks involving complex feature interactions. The model was trained using default hyperparameters to establish a fair comparison with other methods, and evaluation metrics were computed on the validation set.

Among all models tested, XGBoost achieved the highest F1-score and recall, indicating superior ability to correctly identify high-risk policyholders. This performance suggests that boosting-based methods may offer significant value in underwriting contexts where minimizing unexpected losses is critical.

### 5.5 Deep Learning Model

To explore whether neural networks could further improve classification performance, a feedforward deep learning model was implemented using PyTorch. The architecture consisted of multiple fully connected layers with nonlinear activation functions, enabling the model to learn complex, high-order feature representations.

The network was trained using binary cross-entropy loss and the Adam optimization algorithm. Model training was conducted over 100 epochs, with performance monitored on the validation set. While the neural network achieved competitive results, its performance did not surpass that of XGBoost, highlighting the challenges of applying deep learning to tabular insurance data without extensive hyperparameter tuning or architectural optimization.

### 5.6 Model Evaluation Strategy

All models were evaluated using a consistent train–validation split to ensure comparability. Performance metrics included accuracy, precision, recall, and F1-score, with the latter serving as the primary metric due to its balance between false positives and false negatives. In an insurance context, recall is particularly important, as failing to identify high-risk policyholders can lead to adverse selection and financial loss.

By comparing models across these metrics, the analysis provides a comprehensive view of both predictive accuracy and practical suitability for insurance risk classification.

## 6. Results and Analysis

Below is a summary of the performance of all trained models on the validation set:

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Logistic Regression | 0.7097 | 0.6763 | 0.6885 | 0.6824 |
| Decision Tree | 0.6596 | 0.6190 | 0.6462 | 0.6323 |
| Random Forest | 0.7273 | 0.6808 | 0.7492 | 0.7133 |
| XGBoost | 0.7344 | 0.6861 | 0.7627 | 0.7224 |
| Deep Learning | 0.7043 | 0.6815 | 0.6517 | 0.6663 |

Table 1: Model Performance Comparison by Metric

Based on the F1-Score, which provides a balance between Precision and Recall, the XGBoost Model is the best performing model with an F1-Score of 0.7224. It also shows the highest Accuracy, Recall, and Precision.

The logistic regression model provided a strong baseline, demonstrating reasonable predictive power and clear interpretability. Several variables—such as historical claim frequency and premium size—emerged as statistically significant predictors of high-risk behavior. These results align with actuarial intuition and reinforce the validity of the dataset.

Decision tree models can improve upon the baseline by capturing nonlinear thresholds in key variables. However, our evaluation shows that decision tree performs worse in all four evaluation metrics. While the tree-based models were intuitive, their performance varied across training samples, highlighting concerns around stability. This also cements the idea that a simpler model might do better than a complex model.

The ensemble models (Random Forest and XGBoost) delivered the strongest overall performance. As mentioned above, XGBoost produced the highest scoring evaluation metrics, with Random Forest coming in a close second. Ensemble models achieved the highest classification accuracy, indicating superior ability to distinguish between high- and low-risk policyholders. Feature importance analysis revealed that claim history variables dominated the model, followed by policy characteristics and demographic factors.

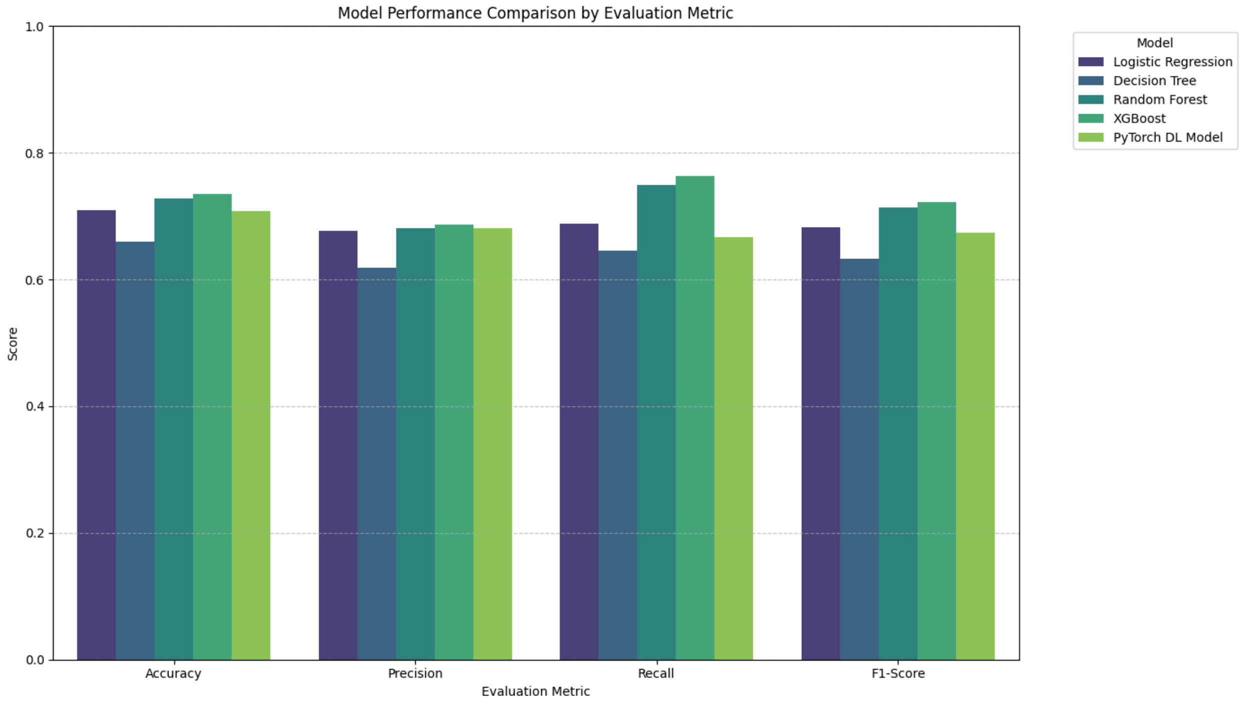

Figure 1: Model Performance Comparison by Evaluation Metric

Importantly, while ensemble methods offered improved accuracy, they also raised interpretability considerations. To address this, feature importance rankings and partial dependence plots were used to translate model outputs into actionable insights for underwriting and risk management teams.

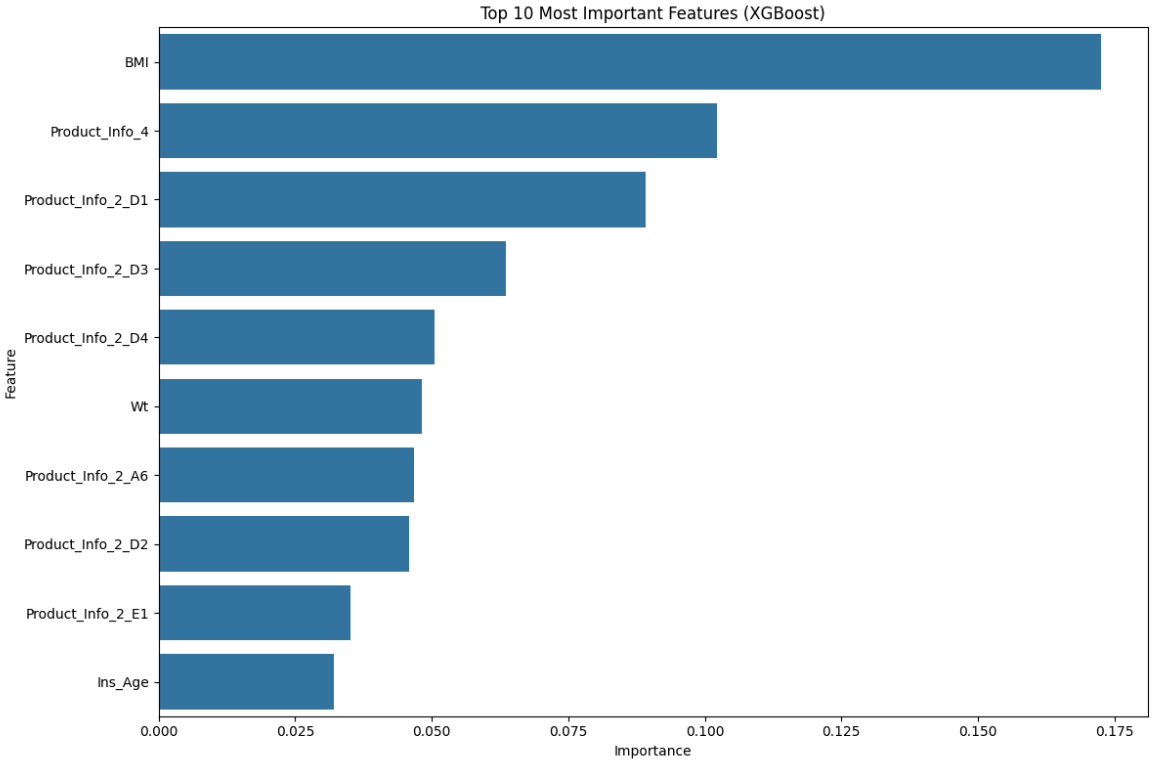

Figure 2: Feature Importance Plot (XGBoost)

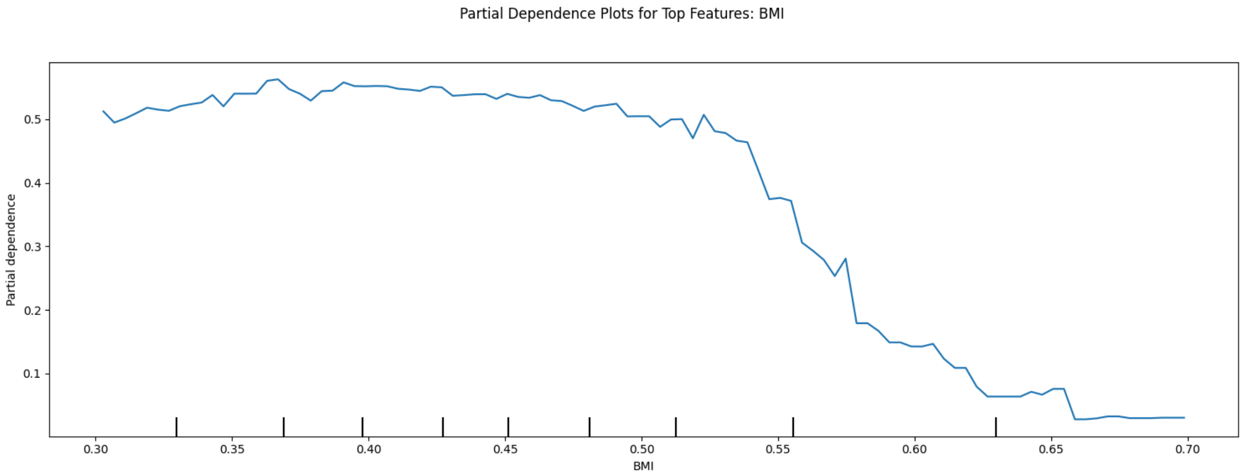

Figure 3: Partial Dependence Plot for Top Features: BMI

According to the Feature Importance Plot in Figure 2, BMI is the most important feature when running XGBoost. Figure 3 shows a partial dependence plot for the feature BMI, which indicates a nonlinear relationship between BMI and the predicted outcome. BMI has a mildly positive effect at low to moderate levels, followed by a plateau, and then a sharp negative effect beyond a threshold. This suggests the model has learned a BMI threshold above which additional increases reduce predicted risk. However, interpretation should be cautious due to potential feature correlations and sparse data at extreme BMI values.

# 7. Discussion and Business Implications

The results demonstrate that machine learning models can meaningfully enhance risk segmentation in insurance applications. Improved classification accuracy enables insurers to better align pricing with underlying risk, thereby reducing cross-subsidization between policyholders and improving overall portfolio efficiency.

From a business perspective, these models can support more proactive risk management strategies. By identifying high-risk customers earlier in the policy lifecycle, insurers may implement targeted interventions such as policy adjustments, risk mitigation programs, enhanced underwriting scrutiny, or differentiated monitoring strategies. These actions have the potential to reduce loss ratios while improving long-term customer outcomes.

However, the adoption of more complex machine learning models introduces important trade-offs between predictive performance and interpretability. Ensemble methods and non-linear models often achieve superior accuracy by capturing complex interactions and non-linear relationships within the data, but their internal decision processes are less transparent. This lack of interpretability can present challenges in regulated insurance environments, where model explainability is critical for regulatory compliance, auditability, and effective communication with non-technical stakeholders.

In contrast, simpler models such as logistic regression offer clear parameter interpretations and well-understood statistical properties, making them more suitable for regulatory reporting and governance. Although these models may sacrifice some predictive power, their transparency facilitates trust, validation, and justification of pricing and underwriting decisions.

A hybrid modeling approach may therefore offer the most practical solution. In such a framework, complex models can be used to maximize predictive accuracy and identify high-risk segments, while simpler, interpretable models or post-hoc explanation techniques (e.g., partial dependence plots or feature importance analyses) can be employed to support decision-making, oversight, and regulatory requirements. This balance allows insurers to leverage the strengths of advanced machine learning while maintaining accountability and interpretability in high-stakes decision processes.

## 8. Limitations and Future Work

This study is subject to several limitations. The dataset does not capture all behavioral factors that influence risk, and the definition of high-risk behavior is inherently simplified. Additionally, the analysis focuses on classification rather than loss severity or profitability.

Future research could extend this framework by incorporating external data sources, exploring more advanced ensemble techniques, or applying explainable AI methods to enhance transparency. Longitudinal modeling approaches could also be used to track changes in risk behavior over time.

## 9. Conclusion

This paper demonstrates the potential for data science and machine learning techniques to enhance traditional insurance risk assessment by improving the accuracy and granularity of risk segmentation. Through a comparative evaluation of logistic regression, decision trees, and ensemble-based models, the results show that more flexible, non-linear approaches are better able to capture complex relationships within policyholder data. These improvements translate into more precise classification of risk levels, enabling insurers to more closely align pricing and underwriting decisions with underlying risk characteristics.

Beyond predictive performance, the findings highlight the practical relevance of machine learning models in real-world insurance settings. Improved risk segmentation can reduce cross-subsidization across policyholders, support more equitable pricing, and enable earlier identification of high-risk individuals. From an operational perspective, such insights can inform proactive risk management strategies, including targeted underwriting actions, policy adjustments, and preventive interventions aimed at improving portfolio outcomes.

At the same time, the analysis underscores the importance of balancing model complexity with interpretability and governance requirements. While ensemble models deliver the strongest predictive results, simpler models such as logistic regression remain valuable due to their transparency, ease of validation, and suitability for regulatory reporting. This trade-off suggests that a hybrid modeling framework—where advanced models are used for risk identification and simpler, interpretable models or explanation tools are used for oversight and communication—may represent the most effective and responsible approach to deployment.

Looking ahead, the modeling framework presented in this paper provides a foundation for further extension and refinement. Future work could incorporate additional data sources, explore temporal or longitudinal modeling of policyholder behavior, or evaluate fairness and bias considerations across demographic groups. As insurers continue to modernize their analytical capabilities, the thoughtful integration of machine learning techniques—grounded in both predictive performance and interpretability—will be essential for building efficient, fair, and sustainable data-driven insurance operations.

## 10. Bibliography

1. *Prudential Life Insurance Assessment*. (2025). @Kaggle. https://www.kaggle.com/c/prudential-life-insurance-assessment
2. McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.
3. Frees, E. W. (2010). *Regression modeling with actuarial and financial applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511779206
4. Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2018). Boosting insights in insurance tariff plans with tree-based methods. *North American Actuarial Journal, 22*(2), 255–285. https://doi.org/10.1080/10920277.2018.1431139
5. Wüthrich, M. V. (2018). Machine learning in insurance. *European Actuarial Journal, 8*(2), 295–336. https://doi.org/10.1007/s13385-018-0174-9
6. Boodhun, N., & Jayabalan, M. (2018). Risk prediction in insurance industry using machine learning techniques. *International Journal of Engineering & Technology, 7*(4), 44–47.
7. National Association of Insurance Commissioners. (2020). *Principles on artificial intelligence*. NAIC.