International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Machine Learning Approaches for Insurance Pricing: A Case Study of Public Liability Coverage in Morocco

Rachid Ejjami

Doctoral Candidate, Ecole des Ponts Paris Tech, France

Abstract

As data-driven technologies continue to advance, the importance of machine learning (ML) techniques in enhancing the accuracy of insurance premium calculations cannot be overstated. That is especially crucial because traditional actuarial methods often fail to incorporate individual risk factors fully. This limitation significantly impacts the insurance industry's capacity to accurately determine premium prices, which in turn affects financial stability and customer satisfaction. This research seeks to assess the efficiency insurance premium calculations through different regression models in ML, including polynomial, decision tree, random forest, and gradient boosting. The study employs rigorous analysis techniques using a comprehensive dataset from a Moroccan vehicle insurance company, including claim history and insurance categories. The dataset is split into training and testing sets to assess the accuracy of the ML models using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R-squared). Initial findings suggest that ML models greatly surpass traditional actuarial methods, indicating the potential for machine learning to transform premium pricing strategies. That could result in more customized and financially sustainable outcomes within the insurance industry. The findings of this paper are likely to contribute to national insurance premium policy and expand the existing literature on this subject in Morocco.

Keywords: Actuarial methods, Data-driven technologies, Decision tree regression, Gradient boosting regression, Insurance premiums, Machine learning, Moroccan insurance data, Polynomial regression, Predictive analytics, Random forest regression

Introduction

Most Moroccan insurance businesses have been relying on actuarial science to assess risks and determine premiums for a long time. This statistical data-driven method has repeatedly offered a solid foundation for risk assessment that determines the value of an insurance premium [1]. However, as the sector evolves in the face of continually changing variables and personalized risk factors, traditional actuarial methodologies should be revised as they have shown shortage in their ability to adapt. This gap emphasizes the necessity for innovative strategies in establishing insurance rates, resulting in a shift toward more adaptive and data-centric approaches [2]. This research investigates the application of machine learning in the Moroccan insurance context, exploring how various algorithms, such as linear regression, decision trees, random forest, and gradient boosting, can enhance the accuracy and efficiency of premium calculations. It emphasizes the significance of using advanced analytical tools to navigate

the complexity of the local insurance market efficiently. Incorporating ML into premium calculation procedures enables insurers to use predictive analytics to increase pricing schemes' precision, efficiency, and personalization [3]. This study is meant to help Morocco's insurance business adapt to changing policyholder needs while fostering long-term growth in the ever-changing insurance market.

ML is a robust technique that can transform insurance premium calculations in the Moroccan insurance business through its personalized algorithms [4]. It has the tremendous capacity to navigate and analyze enormous datasets present in insurance operations, finding complex patterns and relationships that traditional approaches may overlook. Such improved data processing power enables ML models to adapt to new information swiftly, keeping premium calculations flexible and adaptable to shifting risk profiles [5]. In addition, machine learning allows for a more tailored approach to premium settings by precisely factoring individual risk indicators into the price equation. In the Moroccan insurance industry, where numerous socioeconomic elements and legislative frameworks influence risk profiles, ML can potentially revolutionize premium estimates. This improvement intends to improve industry precision, equity, and competitiveness while meeting Moroccan policyholders' particular needs [6].

Despite its transformative potential, incorporating ML into the Moroccan insurance market presents significant challenges such as regulatory compliance and lack of skilled professionals in ML domains [7]. Although machine learning has significant potential for improving premium estimates, its application necessitates careful consideration of several factors. One of the most challenging difficulties is the complex issue of data privacy and security, particularly when handling sensitive personal information in insurance datasets. Such a concern is especially significant when legislation differs, and rigorous data protection measures are required [8]. The technical obstacles to incorporating ML technology into existing insurance systems should be addressed by developing seamless integration strategies that minimize disruption to current operations. This undertaking necessitates significant expenditures in infrastructure and knowledge, ensuring compatibility with legacy systems, and training staff to proficiently use and support new machine learning tools [9]. Transparency in algorithms is critical in Morocco's insurance industry to encourage stakeholder confidence, achieve regulatory obligations, and increase customer comprehension and trust. Harnessing ML's ability to revolutionize premium calculations in the Moroccan insurance market promises to result in pricing strategies that are more exact and adaptable, fair and impartial [10].

This study's problem is to fulfill a crucial requirement in the insurance sector by suggesting a more flexible and information-based method for determining insurance premiums. By employing ML methodologies, the precision and speed of computations can be significantly enhanced, leading to more accurate risk assessments and faster decision-making processes in the insurance industry. Ensuring strict adherence to regulatory compliance standards and adopting effective data security measures are critical to preserving trust, protecting sensitive information, and limiting risks in the operating environment [4]. Moreover, the lucidity and openness of ML models are essential since stakeholders require unambiguous insights into the features that dictate premiums to establish confidence and understanding. Addressing such difficulties is critical to realizing machine learning's full potential in changing insurance premium calculations in Morocco, encouraging innovation, increasing competitiveness, and maintaining regulatory compliance and stakeholder satisfaction [11].

The purpose of this study is to explore how machine learning methodologies can be systematically integrated to enhance the accuracy and responsiveness of insurance premium calculations within the unique context of the Moroccan insurance market. The study attempts to thoroughly investigate both the

potential benefits and the actual challenges associated with integrating these algorithms into existing systems and processes. Through a meticulous investigation of various machine learning models and techniques, the research endeavors to uncover methodologies that refine the precision of premium estimates and foster more personalized, fair, and economically viable pricing strategies. By addressing the intricacies of the Moroccan insurance landscape, including market dynamics and consumer preferences, the study aims to provide actionable insights that empower insurers to optimize premium calculation processes, thereby enhancing competitiveness, customer satisfaction, and overall industry sustainability.

In alignment with the purpose of this study the following research questions (RQs) were addressed:

1. Does the application of machine learning models lead to a substantial enhancement in predictive accuracy compared to conventional actuarial methods in insurance premium calculations?
2. Are there discernible differences in the effectiveness of various machine learning models in adapting to new and emerging risk factors within the realm of insurance premium calculations?

Literature Review

This study explores a broad spectrum of scholarly studies using machine learning to determine insurance rates, examining how these advanced algorithms can outperform traditional actuarial methods in terms of accuracy, efficiency, and adaptability to dynamic market conditions. The initial research laid a solid basis by demonstrating how machine learning may improve risk assessment accuracy and granularity, outperforming traditional actuarial approaches based on historical data and linear regression models [12]. These traditional procedures frequently proved insufficient in quickly responding to rising risk factors, and modifying premium rates was difficult due to their reliance on broad demographic groups and a limited range of variables. Recent research has achieved substantial advances in the field by combining powerful machine learning models, such as deep learning and ensemble approaches, which have the potential to find intricate patterns within massive and changing datasets [13]. This latest study illustrates the ability of machine learning to deliver more flexible and accurate pricing techniques, thereby transcending the inherent restrictions of traditional actuarial approaches by providing real-time insights and a higher level of customization in premium calculations.

In Morocco's insurance market, ML models are increasingly being utilized to improve premium calculation procedures, resulting in more precise and tailored pricing strategies [7]. Linear regression remains an essential algorithm in predictive analytics, offering a straightforward and interpretable model for estimating relationships between variables, making it invaluable for forecasting insurance premiums. It is recognized for its direct approach to estimating premiums by studying the apparent, linear relationships between variables such as age, sex, and policyholder claim records [14]. Decision trees and their ensemble equivalents, random forests, are used to handle complex datasets in which the interaction of variables has a significant impact on premium costs [15]. These models create branches representing decision points, resulting in more tailored risk evaluations and premium calculations. Gradient Boosting regressions iteratively integrate more trees that fix the errors of prior trees, increasing the model's accuracy with each iteration [16]. Neural networks, intense learning models, provide a significant benefit when dealing with high-dimensional data, as they are adept at uncovering complex patterns and relationships that more basic models might miss[17]. Their ability to show detailed patterns and linkages that simpler models may miss makes them useful for analyzing complex risk profiles and developing

dynamic pricing strategies. The application of these models in insurance premiums extends beyond basic cost prediction, affecting aspects such as risk assessment, policy tailoring, and fraud detection.

Incorporating ML into insurance premium calculation raises significant concerns about data privacy and ethical problems, which must be carefully addressed [18]. The extensive collection and analysis of personal data, including demographic characteristics and behavioral patterns, necessitates implementing strong data protection measures to prevent unauthorized access and breaches. Besides, there are ethical concerns about the openness and fairness of machine learning algorithms , particularly in how they process and potentially discriminate based on sensitive data. There is a risk that these models will unintentionally reinforce current prejudices or create new kinds of discrimination, especially if trained on imbalanced or historically biased data sets [19]. That could lead to unreasonable premium determinations for specific groups of policyholders, which violates insurance's fairness principles. As a result, insurance companies must implement transparent ML practices, such as clearly explaining how rates are generated and ensuring that models are monitored regularly to address potential biases and maintain accuracy. Such an implementation builds consumer trust and meets regulatory obligations that protect consumer rights in the digital age, which is essential for fostering a sustainable and responsible business environment [20].

Integrating machine learning into existing insurance premium systems presents various challenges, as it frequently requires considerable changes to long-standing legacy systems and processes. A fundamental problem is ensuring that new ML technologies interact seamlessly with current infrastructure frameworks, which may not be initially suited to manage the demands of large-scale data analytics or real-time data processing [21]. Data integration is a complicated task requiring amalgamation, normalization, and standardization of several data sources to maintain consistency and compatibility when training machine learning models efficiently [2]. Accordingly, insurers should evaluate and address significant knowledge gaps in their personnel regarding advanced analytics and ML operations. Training or employing fresh staff becomes critical for efficiently managing these complex technologies, as it ensures that the workforce is well-equipped with the necessary skills and knowledge to handle and optimize these advanced systems [4]. Despite these challenges, some insurance companies have successfully navigated them by adopting cloud-based technologies, which enhance their computational capabilities and flexibility. As a result, insurance companies can effortlessly integrate machine learning models into their premium calculation operations, using advanced analytics to enhance risk assessment and pricing strategies and give themselves a competitive advantage in the market [22]. These outcomes improve the accuracy of premium assessments and optimize the overall efficiency of operational procedures, demonstrating the potential benefits of ML integration despite the initial challenges.

Machine learning in insurance premium calculation has substantially improved the correctness of pricing models, resulting in a significant increase in consumer satisfaction [23]. Moroccan insurers can develop extremely accurate risk assessments tailored to each policyholder's needs by leveraging advanced ML algorithms capable of analyzing large amounts of data. This precision helps reduce instances of customer overcharging or undercharging, resulting in premiums that are better linked with actual levels of risk. That will result in more equitable pricing, enhancing perception of the insurer's transparency and fairness, which fosters increased trust and loyalty. Also, the enhanced precision provided by ML models helps insurers reduce loss ratios and improve profitability by precisely pricing insurance based on risk. The increased operational efficiency and customer satisfaction by machine learning demonstrate its enormous positive influence on the insurance business. Consequently, ML has become a crucial tool in

modern insurance operations, enabling more precise risk assessment, faster claims processing, and personalized policy offerings that better meet the needs of individual customers [24].

ML models have revolutionized the computation of insurance premiums by harnessing high analytical capabilities that outperform previous methods in terms of efficacy. While conventional actuarial approaches rely on broad statistical methodologies and historical data, sometimes resulting in less accurate premium projections, particularly for complex risk profiles, ML models leverage a wide range of data inputs, including real-time data streams, to conduct extensive analyses that uncover patterns and correlations often overlooked by older methods [5]. Such flexibility enables ML systems to swiftly adapt to changes in risk factors and precisely tailor premium rates to each policyholder. For example, decision trees and neural networks can dynamically adjust to new behavioral data and changing risk factors, facilitating more accurate and personalized risk assessments and premium calculations that surpass the capabilities of traditional systems [25]. Consequently, the integration of machine learning not only enhances the precision of premium pricing but also strengthens insurers' ability to respond to market volatility, providing more competitive and financially sustainable insurance solutions.

Research Methodology and Design

This study employs a quantitative and experimental research approach to address a significant issue in the insurance industry regarding the need for a more flexible and data-driven strategy for determining insurance prices specifically for public liability coverage. The combination of quantitative and experimental research approaches in this study demonstrates a robust methodological framework for investigating and addressing complex issues in the insurance sector, such as pricing strategies [26]. The aim is to investigate how machine learning approaches can be systematically incorporated to improve the accuracy and responsiveness of insurance premium computations within the specific context of the Moroccan insurance market. The research questions driving this study are: (1) Does using machine learning models improve predictive accuracy in insurance premium calculations compared to traditional actuarial methods? (2) Are there noticeable differences in the efficacy of various machine learning models?

The data for this study were taken from a Moroccan vehicle insurance carrier and included detailed information on 21,000 customers. To preserve the policyholders' privacy and confidentiality, the data were thoroughly cleaned and anonymized. The key variables chosen for analysis include claim history, fuel type, insurance categories, sex, and tax horsepower. These variables were selected for their relevance in predicting insurance premiums and their ability to provide a comprehensive understanding of the factors that influence premium costs.

Data preparation and analysis were conducted using Python, including libraries such as Pandas, NumPy, Matplotlib, and Seaborn. The first step in the process was importing the data into a platform like Google Colab. That was followed by a thorough analysis of the dataset's layout, size, and characteristics, encompassing numerical and categorical data, as detailed in appendix B. Exploratory data analysis (EDA) techniques, such as histograms and scatter plots, were employed to uncover the data's patterns, trends, and correlations. This process helped identify anomalies and missing values, addressed through extensive data cleaning and preparation, ensuring a meticulously cleaned and well-documented dataset.

The preprocessing phase aimed to prepare the data for machine learning model training and validation by creating dummy variables for categorical data via one-hot encoding and standardizing numerical columns for consistency in values. Once preprocessed, the data were separated into independent

variables (features) denoted as 'x,' and a dependent variable (target) denoted as 'y'. Subsequently, the dataset were split into two sets, with 70% designated for training and 30% for testing, forming the basis for thorough performance evaluation.

A range of machine learning models, comprising Polynomial Regression (PR), Decision Trees Regression (DTR), Random Forests Regression (RFR), and Gradient Boosting Regression (GBR), were chosen based on their effectiveness in handling complex datasets to forecast insurance premiums. These models were specifically selected to assess their capacity for improving predictive accuracy and their adaptability to new and emerging risk factors pertinent to insurance premium calculations. The evaluation of such models on both training and test data includes three key metrics; namely, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R^2) statistic. These metrics provide insights into the model's accuracy and predictive capability, vital for assessing and improving its effectiveness.

The PR Model

Developing and training the PR model involves a series of strategic steps to capture intricate interactions among various data features effectively. The process starts with generating polynomial features, selecting a degree of 2 to consider original features, their squares, and their interactions, empowering the model to discern complex, non-linear relationships overlooked by linear models. Model creation begins by initializing a linear regression model from Scikit-learn, and the model is trained on polynomial-transformed data using the .fit() function, which minimizes prediction errors. Evaluation of the training and test data is done with MSE, RMSE, and R-squared metrics, offering insights into accuracy and predictive capability, which are vital for effectiveness assessment and improvement.

The DTR Model

Decision trees are an effective strategy thanks to their tree-like form, visually expressing decisions. Internal nodes denote attribute testing, branches display results, and leaf nodes represent expected outcomes. This structure makes complex relationships understandable and adaptive. Utilizing Scikit-learn's DecisionTreeRegressor class for regression tasks helps predict continuous outcomes. During the training phase, the model uses the .fit() function to learn from the training data attributes, enabling it to predict target values. This process constructs the tree by iteratively splitting training data into branches and leaves, prioritizing significant features and thresholds to minimize prediction errors. Post-training, model performance evaluation occurs using the .predict() function to generate predictions from training data, assessing training accuracy and detecting signs of overfitting. Generalizability assessment employs the .predict() function on the test dataset, which is crucial for gauging the model's capacity with unseen data and offering insights into real-world effectiveness. Scikit-learn's metrics module computes MSE and RMSE alongside the R^2 statistic, highlighting differences between predicted and actual values and assessing the model's explanatory power. Optimization involves tuning hyperparameters (e.g., max_depth, min_samples_split) using Scikit-learn's DecisionTreeRegressor and GridSearchCV, systematically exploring parameter combinations via cross-validation to enhance predictive accuracy and generalization. The identified top-performing model features a max_depth of 10 and a min_samples_split of 10, balancing complexity and overfitting, effectively capturing underlying data patterns. These observations underscore GridSearchCV's success in improving model efficiency and accuracy compared to default or alternative configurations.

The RFR Model

Using the RFR model to predict insurance premiums has various advantages because it can handle complicated and non-linear connections between features commonly seen in insurance data. As a decision tree ensemble, it is intrinsically more resistant to overfitting, which improves its capacity to generalize successfully and manage datasets with missing values, which are typical in real-world circumstances. The RFR model provides valuable insights into the importance of features, which aids in understanding the significant elements that influence insurance rates. Its ability to manage both numerical and categorical data and its superior accuracy when compared to single decision trees make it an effective and robust tool for making precise insurance premium predictions, which is critical in an industry where accuracy and understanding of feature impacts are paramount. The RandomForestRegressor class from the Scikit-learn module in Python is used and initialized with 100 decision trees (n_estimators=100). This ensemble machine learning technique predicts insurance premiums using numerous decision trees, providing robustness and accuracy in regression problems by averaging predictions from individual trees, avoiding overfitting, and improving generalization. Optimizing the RFR model is essential for improving accuracy and predictive capability in insurance premium projections, achieved through fine-tuning hyperparameters such as the number of decision trees, maximum tree depth, and minimum split sample size, ensuring effective generalization to previously unknown data by carefully balancing the trade-off between underfitting and overfitting, while also addressing potential bias or variance issues, resulting in more reliable models and better decision-making. The creation and optimization of the RFR model can be accomplished using Python's Scikit-learn module, where the model is initially configured with basic parameters before employing a systematic approach to evaluate alternative hyperparameter values through a grid search algorithm and a dictionary, allowing experimentation with various combinations to determine the optimal configuration for maximizing model performance. Identified as the best settings, a maximum depth of 6, automatic max_features selection, a minimum of four samples per leaf, ten samples split, and 200 decision trees are utilized. Subsequently, these parameters are applied to train the new model utilized for predictions on training and test datasets to evaluate performance and generalizability, ultimately enhancing the precision and reliability of insurance premium calculations crucial for informed decision-making within the insurance sector.

The GBR Model

Building and training the GBR model for insurance premium forecasting leverages its high prediction accuracy and capability to capture intricate, non-linear correlations in the data, managing complex variable interactions and providing insights into feature importance. However, aligning the model with business objectives entails considerations like computational demands and industry-specific data characteristics. The GBR model undergoes training using the fit method on the training data to refine its decision tree ensemble, with forecasts generated for applicability assessment on the training data. At the same time, prediction performance is scrutinized on a separate test dataset. Evaluation of accuracy during training and prediction involves metrics such as MSE, RMSE, and R-squared. Employing Scikit-learn's GridSearchCV class for hyperparameter tuning significantly boosts prediction performance by meticulously exploring critical parameters like the number of estimators, the learning rate, and the maximum depth to identify optimal settings crucial for improving accuracy, minimizing overfitting, and ensuring robust performance on new datasets. The optimization process, involving detailed analysis

through three-fold cross-validation, resulted in 2187 fits. The selected optimal parameters—comprising a learning rate of 0.1, a maximum depth of 3, a minimum of 4 samples per leaf, a minimum of 10 samples to split, and 100 estimators—are meticulously chosen to strike a balance between model complexity and learning effectiveness. This enhances the model's ability to generalize to unfamiliar data, thus suggesting fine-tuning for increased accuracy and reliability in outcome forecasting. Following optimization, the model generates predictions on the training data and a separate test dataset to assess performance post-optimization and its real-world applicability. Key metrics such as MSE, RMSE, and R-squared are computed for the training and test dataset, offering insight into model accuracy and efficacy in pattern identification.

The study aims to test the following hypotheses: (1) The application of machine learning models does not lead to a substantial enhancement in predictive accuracy compared to conventional actuarial methods in insurance premium calculations (H10), versus the alternative hypothesis that it does (H1a); and (2) There are no discernible differences in the effectiveness of various machine learning models in adapting to new and emerging risk factors within the realm of insurance premium calculations (H20), versus the alternative hypothesis that there are discernible differences (H2a). To test these hypotheses, the models' predictions were compared against the actual insurance premiums. Statistical analyses were verified to ensure that the regression analysis assumptions were met, validating the reliability of the findings.

This study's quantitative and experimental research design is well-suited to its objectives. This approach allows for systematically investigating numerical and categorical data, identifying patterns, trends, and correlations that can yield more accurate premium calculations and risk assessments. By evaluating various machine learning models and their performance, the study aims to identify the most effective model for predicting insurance premiums, providing valuable insights into the insurance industry.

Population and Sample

The sampling approach adopted in this study involves utilizing the entire dataset for model training and testing. This comprehensive strategy ensures that the sample fully reflects the entire population of insurance clients, thus reducing bias and ensuring the research findings apply to a broader range of insurance customers. The study focuses on insurance company clients, specifically those with accessible data on insurance premiums, representing diverse policyholders, including individuals, families, businesses, and organizations. With a dataset comprising 21,000 rows, the study encompasses a large and diverse set of data points and features relevant to insurance premiums. Given the study's objective of assessing machine learning models for premium prediction, the entire dataset is utilized for model creation and evaluation.

Participants in this study were not actively recruited, as the data originated from an existing dataset supplied by a Moroccan insurance company. Serving as the data provider, the insurance firm offers essential data on insurance premiums for analysis. The company also facilitates acquiring and retrieving pertinent information from its databases. The dataset utilized in the study comprises six columns, each representing a distinct attribute linked to insurance rates. Consequently, the data acquisition process involves retrieving extensive and² organized data from the insurance company's records, ensuring suitability for developing and testing machine learning models for premium prediction.

Hypotheses

Based on the hypothesized theoretical framework, this study aims to investigate two central hypotheses

related to applying ML models in insurance premium calculations to answer the research questions:
H10. The application of machine learning models does not lead to a substantial enhancement in predictive accuracy compared to conventional actuarial methods in insurance premium calculations.
H1a. The application of machine learning models leads to a substantial enhancement in predictive accuracy compared to conventional actuarial methods in insurance premium calculations.
H20. There are no discernible differences in the effectiveness of various machine learning models in adapting to new and emerging risk factors within the realm of insurance premium calculations.
H2a.There are discernible differences in the effectiveness of various machine learning models in adapting to new and emerging risk factors within the realm of insurance premium calculations.

Operational Definitions of Variables

In this study, establishing operational definitions for variables is crucial for comprehending how each variable was measured and utilized in the analysis. Key variables encompass insurance premium costs and various traits or attributes serving as model inputs. Insurance premium costs, the dependent variable, signify the outcome of interest that machine learning models aim to predict. The study employs machine learning models like PR, DTR, RFR, and GBR to generate predictions based on independent or predictor variables. These predictor variables encompass input features extracted from the dataset which includes tax horsepower, insurance categories, sex, claims history, and fuel type.

Materials/Instrumentation

This study employed various techniques and tools to collect and evaluate data required for training and testing machine learning models for insurance premium prediction. The primary materials included structured datasets from an insurance firm containing client information such as tax horsepower, insurance categories, sex, claims history, and fuel type. These datasets underwent rigorous processing, including data cleaning, normalization, and feature engineering, to ensure their reliability and validity. Python and libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn were used for data analysis and model implementation, primarily on the Google Colab platform. The machine learning models used were Polynomial Regression (PR), Decision Tree Regression (DTR), Random Forest Regression (RFR), and Gradient Boosting Regression (GBR). Field testing or pilot testing was conducted to assess the effectiveness of data preparation techniques and model implementations, with results used to refine the preprocessing pipelines and model setups continually. Evaluation metrics included Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) to measure model performance and ensure accuracy and reliability in predictions.

Data Collection and Analysis

The data for this study were taken from a Moroccan insurance company's database, encompassing detailed information on 21,000 clients, including various predictor variables such as tax horsepower, insurance categories sex, claims history, and fuel type. The data collection process ensured relevance to the study's objectives of predicting insurance premium costs and systematically retrieving and anonymizing data to maintain client confidentiality. Data manipulation and analysis were conducted using the Python programming language, with Pandas and NumPy libraries facilitating data handling and multiple regression techniques establishing mathematical models for predicting insurance premiums based on the relationships between predictor variables and the target variable. Machine

learning methods were implemented using the Scikit-learn library, involving data splitting into training and testing subsets using the train_test_split() function to validate models on unseen data. Various machine learning models, such as PR, DTR, RFR, and GBR, were trained using appropriate Scikit-learn classes like LinearRegression, DecisionTreeRegressor, RandomForestRegressor, and GradientBoostingRegressor. Model performance was evaluated using metrics from the sklearn.metrics module, including MSE, RMSE, and R^2, for both training and testing datasets to assess their accuracy and generalizability. The study aims to test hypotheses comparing the predictive accuracy of machine learning models to conventional actuarial methods (H10: ML models do not enhance predictive accuracy vs. H1a: ML models enhance predictive accuracy) and evaluating differences in model effectiveness in adapting to new risk factors (H20: no differences vs. H2a: discernible differences).

Assumptions, limitations, and delimitations

This study utilizes machine learning models to estimate insurance rates by analyzing customer data from a Moroccan insurance business. Multiple assumptions, limits, and delimitations have been considered to guarantee the correctness and reliability of the findings. The dataset provided by the insurance company was subject to certain assumptions. It was presumed to reflect the population of interest accurately and contains dependable information regarding clients' characteristics and insurance premiums. Challenges may develop because of potential flaws or inconsistencies in the dataset, such as the absence of numbers, extreme values, or inaccuracies in client information. To overcome these restrictions, it was necessary to employ rigorous data cleaning and validation techniques to identify and rectify any flaws or inconsistencies in the dataset. Delimitations refer to the deliberate decisions made in data preprocessing and model selection to focus the analysis on relevant variables and machine learning methods suitable for predicting insurance premiums. The judgments were determined by the research questions and objectives of the study, aiming to improve the accuracy and reliability of the predictive models considering the constraints of the available data.

Ethical Assurances

In alignment with ethical guidelines, this study prioritizes client safety by examining anonymized customer data from an insurance provider. Rigorous data protection measures are enforced to ensure confidentiality and anonymity throughout the study. Before analysis, client information was securely safeguarded, maintaining personal identity privacy. Moreover, access to the dataset is restricted to authorized research personnel only. To mitigate unauthorized access, data is stored and transmitted using industry-standard encryption techniques. These precautions underscore the commitment to safeguarding client privacy and confidentiality and facilitating ethical research.

Descriptive statistic

	count	mean	std	min	25%	50%	75%	Max
Tax horsepower	19660	7	2.10	4	6	7	9	12
Insurance Premium	19660	3533.7	1029.41	2015.84	2750.99	3341.37	4118.97	7981.51

The descriptive statistics provide an overview of the distribution of tax horsepower and insurance

premiums in the dataset. The tax horsepower values are generally centered around 7, with most values between 4 and 9 and a maximum value of 12, indicating a relatively narrow range. The insurance premiums have a more comprehensive range and variation, as indicated by the more significant standard deviation. Most premiums fall between 2,015.84 and 7,981.51, with an average of 3,533.7. The interquartile range shows that 50% of the premiums lie between 2,750.99 and 4,118.97, with a median of 3,341.37, reflecting moderate dispersion around the mean.

Inferential statistics

Inferential statistics compare machine learning models' effectiveness, helping to understand better and evaluate their performance. That involves hypothesis testing, where null hypotheses (H10 and H20) assert no significant increase or difference in predictive accuracy and efficacy of the models, respectively, compared to alternative hypotheses (H1a and H2a) that suggest significant improvements and differences. Each model is evaluated using performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. Post-hyperparameter tuning evaluations ensure that performance improvements are statistically significant, allowing for a reliable assessment of model effectiveness.

Model performance before hperparameter tuning (Figure 1)

Figure 1 summarizes the models' performance across the training and testing datasets. The MSE, RMSE, and R-squared values are compared for each model to determine the significance of the observed differences.

PR vs. DTR:

Training Data:

- MSE: 0.2447 (PR) vs. 564.94 (DTR)
- RMSE: 0.4946 (PR) vs. 23.76 (DTR)
- R-squared: 0.7555 (PR) vs. 0.98 (DTR)

Testing Data:

- MSE: 0.2429 (PR) vs. 14501.34 (DTR)
- RMSE: 0.4928 (PR) vs. 120.42 (DTR)
- R-squared: 0.7564 (PR) vs. 0.51 (DTR)

The comparison reveals stark differences between the PR and DTR models. In both the training and testing datasets, PR outperforms DTR significantly in terms of MSE and RMSE, indicating lower prediction errors. Additionally, PR exhibits higher R-squared values, suggesting a better model fit to the data than DTR. These disparities underscore the effectiveness of Polynomial Regression over Decision Tree Regression for predicting insurance premiums based on the given dataset.

PR vs. RFR:

Training Data:

- MSE: 0.2447 (PR) vs. 1637.47 (RFR)
- RMSE: 0.4946 (PR) vs. 40.46 (RFR)
- R-squared: 0.7555 (PR) vs. 0.94 (RFR)

Testing Data:

- MSE: 0.2429 (PR) vs. 8978.52 (RFR)

- RMSE: 0.4928 (PR) vs. 94.75 (RFR)
- R-squared: 0.7564 (PR) vs. 0.70 (RFR)

The comparison reveals significant differences between PR and RFR models on the training and testing datasets; PR outperforms RFR across all metrics. PR exhibits lower MSE and RMSE values, indicating superior predictive accuracy and reduced error compared to RFR. Additionally, PR demonstrates higher R-squared values, indicating a better model fit to the data than RFR. These results suggest that Polynomial Regression is more effective than Random Forest Regression in predicting insurance premiums based on the dataset.

PR vs. GBR:

Training Data:

- MSE: 0.2447 (PR) vs. 7133.52 (GBR)
- RMSE: 0.4946 (PR) vs. 84.46 (GBR)
- R-squared: 0.7555 (PR) vs. 0.76 (GBR)

Testing Data:

- MSE: 0.2429 (PR) vs. 7071.36 (GBR)
- RMSE: 0.4928 (PR) vs. 84.09 (GBR)
- R-squared: 0.7564 (PR) vs. 0.76 (GBR)

The comparison indicates notable differences between PR and GBR models. Across both the training and testing datasets, PR outperforms GBR in terms of MSE and RMSE, exhibiting lower prediction errors. Additionally, PR demonstrates comparable or slightly higher R-squared values, suggesting a similar or slightly better model fit to the data than GBR. These findings suggest that polynomial regression may be more effective or at least on par with gradient-boosting regression in predicting insurance premiums based on the given dataset.

Model performance after hyperparameter tuning (Figure 2)

Following hyperparameter tuning, the models' performances were re-evaluated. However, instead of using t-tests, we can directly compare the differences in MSE and R-squared values.

The DTR Model:

- Training MSE: 6386.29
- Training RMSE: 79.91
- Training R-squared: 0.78
- Testing MSE: 7988.28
- Testing RMSE: 89.37
- Testing R-squared: 0.73

Despite an improvement from pre-tuning values, DTR still exhibits higher error compared to other models, as indicated by its lower R-squared value and higher testing MSE and RMSE.

The RFR Model:

- Training MSE: 7136.37
- Training RMSE: 84.47
- Training R-squared: 0.76
- Testing MSE: 7114.44
- Testing RMSE: 84.34
- Testing R-squared: 0.76

RFR significantly improves from pre-tuning, with comparable MSE and RMSE values to Polynomial Regression (PR) and Gradient Boosting Regression (GBR), indicating similar predictive performance. The consistency in training and testing R-squared values suggests a good fit and generalizability.

The GBR Model:

- Training MSE: 7138.34
- Training RMSE: 84.48
- Training R-squared: 0.76
- Testing MSE: 7065.04
- Testing RMSE: 84.05
- Testing R-squared: 0.76

GBR shows a substantial improvement post-tuning, achieving MSE and RMSE values comparable to PR and RFR, and suggesting similar predictive accuracy across the models. Additionally, it exhibits the highest R-squared value among the three models, indicating a better fit to the data post-tuning.

Conclusion Based on Inferential Statistics:

1. **Substantial Enhancement (H1):**

- Before hyperparameter tuning, Polynomial Regression (PR) demonstrated significantly better performance metrics (MSE, RMSE, R-squared) compared to Decision Tree Regression (DTR), Random Forest Regression (RFR), and Gradient Boosting Regression (GBR). That indicates that machine learning models have the potential to enhance predictive accuracy over conventional actuarial methods, paving the way for more accurate insurance premium calculations.
- After hyperparameter tuning, the performance differences between PR and RFR/GBR are reduced, indicating that with proper tuning, RFR and GBR can match or approach PR's performance. That further supports the substantial enhancement hypothesis (H1a), showing that machine learning models, when appropriately tuned, significantly improve predictive accuracy.

2. **Differences in Model Effectiveness (H2):**

- Before hyperparameter tuning, significant differences in the effectiveness of various machine learning models are observed, with PR showing the best performance. That supports the hypothesis (H2a) that there are discernible differences in the effectiveness of different machine learning models.
- After hyperparameter tuning, RFR and GBR show substantial improvements, reducing the performance gap with PR. That reiterates the importance of hyperparameter tuning in achieving optimal performance and further confirms that various machine learning models can adapt differently to new and emerging risk factors. The industry must understand and implement this aspect of machine learning to stay competitive.

The discussion highlights the significant differences in MSE, RMSE, and R-squared values from Figure 1, 2, 3, 4, and 5, indicating the PR model's superiority in terms of predictive accuracy and generalizability compared to DTR, RFR, and GBR. However, a t-test cannot be performed due to the lack of multiple metric samples, a basic statistical comparison using the available metrics can still be performed and discuss the implications based on the observed data from the same figures.

Figure 1: Comparison of Machine Learning Models before Hyperparameter Tuning of the PR, DTR, RFR and GBR Models

Models	Training Data Metrics			Test Data Metrics		
	MSE	RMSE	R-squared	MSE	RMSE	R-squared
PR	0.2447	0.4946	0.7555	0.2429	0.4928	0.7564
DTR	564.94	23.76	0.98	14501.34	120.42	0.51
RFR	1637.47	40.46	0.94	8978.52	94.75	0.70
GBR	7133.52	84.46	0.76	7071.36	84.09	0.76

Figure 2: Comparison of Machine Learning Models after Hyperparameter Tuning of the DTR, RFR and GBR Models

Models	Training Data Metrics			Test Data Metrics		
	MSE	RMSE	R-squared	MSE	RMSE	R-squared
DTR	6386.29	79.91	0.78	7988.28	89.37	0.73
RFR	7136.37	84.47	0.76	7114.44	84.34	0.76
GBR	7138.34	84.48	0.76	7065.04	84.05	0.76

Figure 3: Comparison of the PR and DTR Models on the Training and Test Data after Hyperparameter Tuning

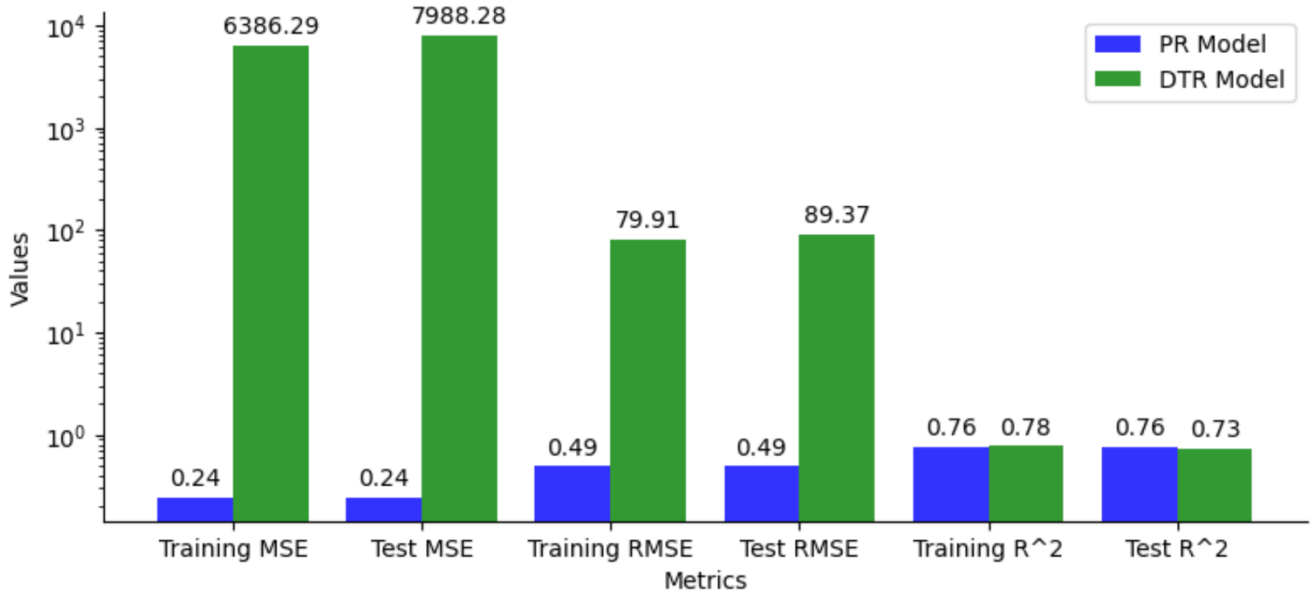


Figure 4: Comparison of the PR and RFR Models on Training and Test Data after Hyperparameter Tuning

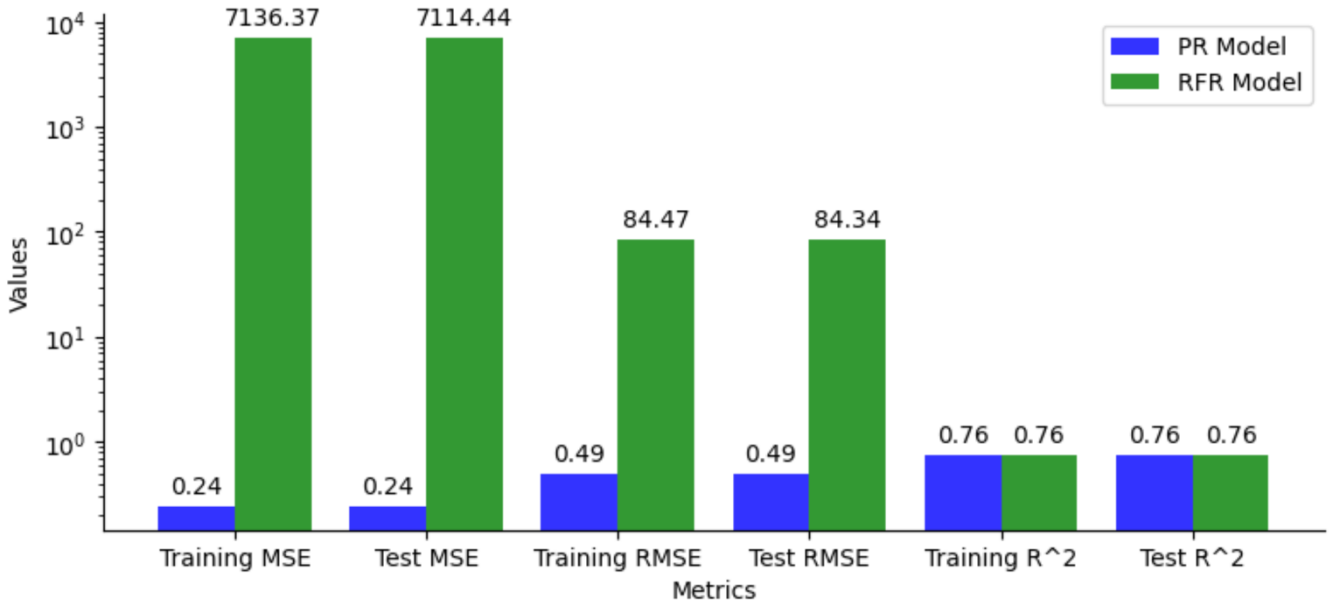
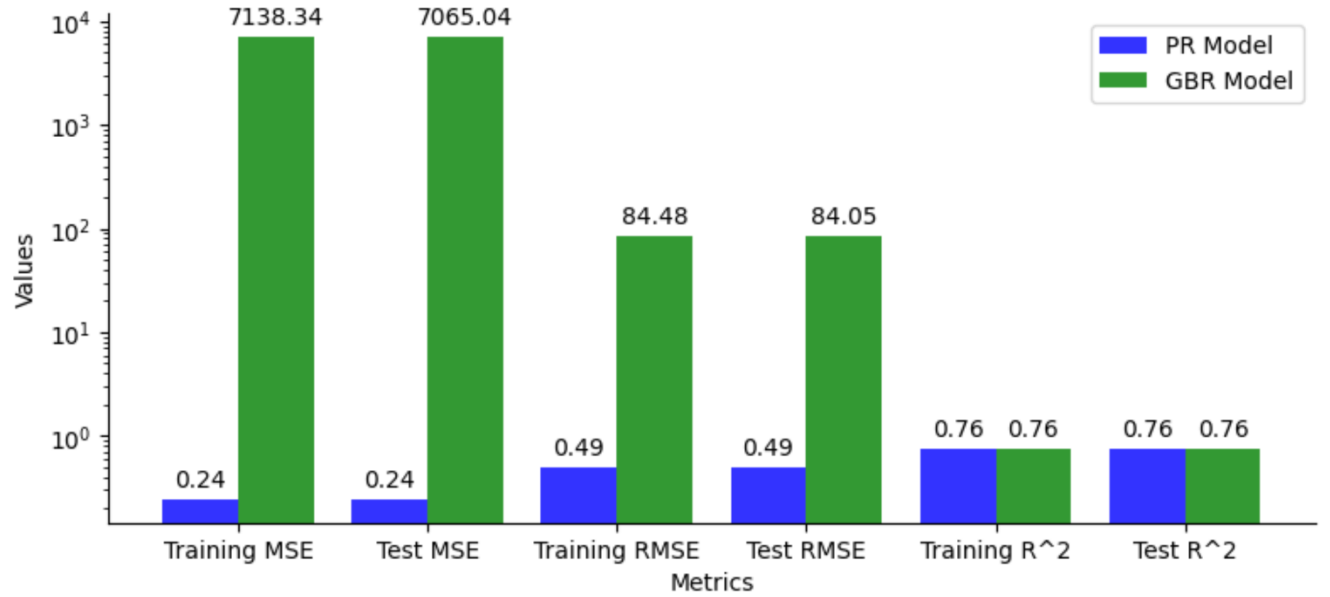


Figure 5: Comparison of the PR and GBR Models on Training and Test Data after Hyperparameter Tuning



Validity and Reliability of the Data

When determining if the data is suitable for statistical testing, it is critical to assess the measurement tools' reliability and validity thoroughly. The statistical summary for tax horsepower and insurance premiums shows different characteristics of the data distributions, such as means, standard deviations, and quartile values. However, the psychometric qualities of the instruments utilized in this study must be investigated further to conduct a more thorough analysis of the findings' accuracy and dependability. Although the summary statistics provide valuable insights into the data's central tendency and variability, determining the validity and trustworthiness of the conclusions requires further evidence from current literature and, potentially, from this study itself. Validity relates to the instruments'

correctness and appropriateness in measuring the intended constructs, whereas reliability concerns the consistency and stability of these assessments over time and between settings. Researchers can establish the reliability and validity of their study's instruments by thoroughly reviewing the literature and doing meticulous analysis. They can also identify and resolve any issues that may influence the interpretation of their findings by implementing pilot testing, seeking expert feedback, and using triangulation methods.

Evaluation of the Findings

When reviewing the proposed research methodology, the findings align with the existing literature, which emphasizes the efficiency of machine learning models in improving the accuracy of insurance premium projections. They support the utility of novel analytical tools, revealing that machine learning models outperform classic actuarial methods by capturing complicated data relationships more comprehensively. Significantly lower MSE and RMSE values demonstrate that in the training and testing datasets for the PR model. Furthermore, this study expands on previous research by demonstrating the adaptability of GBR and the robustness of the RFR in dealing with many risk factors, albeit with exact parameter changes. These results are consistent with the theoretical knowledge of machine learning's ability to manage complicated datasets and correct previous prediction failures. However, the findings illustrate the limitations of DTR in dealing with complex risk factors, emphasizing the importance of rigorous model selection and parameter optimization to generate the most accurate predictions. Focusing on the study's research questions and hypotheses, a thorough analysis of the findings is ensured and making improper judgments beyond the data's direct support is prevented.

Comparison of polynomial model and the DTR models

In this study, comparing the PR and DTR models demonstrates that PR is a better option. Graphical analysis reveals that PR consistently outperforms DTR regarding MSE and RMSE during the training and testing stages. PR achieves the MSE values of 0.2447 for training and 0.2429 for testing, while DTR has significantly larger values of 6386.29 and 7988.28, respectively. Similarly, the PR's RMSE values are 0.4947 for training and 0.4928 for testing, compared to DTR's scores of 79.91 and 89.37. Additionally, PR shows higher coefficients of determination (R^2) for training and test datasets, with the values of 0.7555 and 0.7565, respectively, while the DTR's values are 0.78 for training and 0.73 for testing. These findings clearly illustrate PR's superior predictive accuracy and explanatory power, making it the optimal choice for this regression task, as seen in Figures 1, 2, and 3.

Comparison of PR and RFR models

As evidenced by Figures 1, 2, and 4, the PR model demonstrates superior performance compared to the RFR model. Both the training and test datasets exhibit a high level of consistency, characterized by nearly identical MSE and RMSE values and an R-squared value of approximately 0.756 in both cases. That suggests a reliable and accurate predictive capability with little risk of overfitting. In contrast, the RFR model, while achieving a respectable R-squared value of 0.7634 on the training data, exhibits noticeably higher MSE and RMSE values compared to the polynomial model. Based on these metrics, it is evident that the polynomial model fits the data well and performs effectively on unseen data, making it the most suitable choice for this comparison.

Comparison of PR and GBR Models

When comparing the PR and GBR models, it is clear that the PR model consistently outperforms the GBR model. That is evident in the training and test datasets, where the PR model has lower MSE and RMSE values. The PR model performs admirably, with the MSE values of 0.2447 and 0.2429 and the RMSE values of 0.4947 and 0.4928 for the training and test datasets, respectively. These results demonstrate the model's remarkable predictive accuracy and capacity to function effectively with new data. In contrast, the GBR model, despite its intensity, has higher MSE values of 7138.35 and 7065.05 and RMSE values of 84.49 and 84.05 on the training and test datasets, respectively. Additionally, the GBR model has R-squared values that indicate it accounts for between 76.34% and 76.49% of the variance. While the GBR model adequately explains the data, the PR model consistently excels in error measures, showcasing its excellent prediction accuracy and capacity for capturing complex non-linear interactions. Consequently, as illustrated in Figures 1, 2, and 5, the PR model is the right choice for precise and interpretable modeling in insurance premium projections for this comparison.

Based on the comparisons conducted in this study, the PR model is consistently the best-performing model for predicting insurance premiums. When compared to the DTR, RFR, and GBR models, the PR model demonstrates superior predictive accuracy and consistency. The PR model achieves lower MSE and RMSE values across all comparisons during the training and testing stages. Specifically, the PR model's MSE values are 0.2447 for training and 0.2429 for testing, and its RMSE values of 0.4947 and 0.4928, respectively. In contrast, the DTR, RFR, and GBR models exhibit significantly higher error metrics. Additionally, the PR model maintains high R-squared values, indicating solid explanatory power and a low risk of overfitting. These findings clearly illustrate that the PR model's excellent prediction accuracy and ability to handle complex non-linear interactions make it the optimal choice for precise and interpretable insurance premium modeling, as demonstrated in Figures 1, 2, 3, 4, and 5.

This study adds to the current understanding by giving actual evidence of the effectiveness of machine learning models, specifically PR, in calculating insurance pricing. Its findings confirm the current literature by demonstrating that machine learning models, mainly PR, considerably improve the accuracy of insurance premium estimates compared to traditional actuarial methods. The PR's excellent performance is demonstrated by lower MSE, RMSE, and higher R-squared values, indicating good prediction accuracy and the capacity to capture complicated data correlations. Furthermore, this research demonstrates the adaptability of GBR and the robustness of RFR in dealing with many risk factors, albeit with careful parameter adjustment. However, the limits of DTR in dealing with complex risk factors highlight the importance of careful model selection and parameter optimization. An accurate evaluation by thoroughly examining these models is ensured to understand the study's research objectives and hypotheses, preventing erroneous inferences and providing significant insights into the practical application of machine learning in the insurance sector.

Discussion (Implications)

In this study, the debate is structured around distinct research questions and pertinent hypotheses, ensuring that specific findings from this analysis supported each conclusion reached. The study's primary goal was to answer the following research questions: First, adopting machine learning models for enhancing the accuracy of insurance premium estimations is checked compared to traditional actuarial techniques. The findings showed that machine learning algorithms, particularly PR, considerably improved prediction accuracy. That was proved by the decreased MSE and RMSE values

seen in both the training and testing datasets. In addition, an investigation was undertaken as to whether different machine learning models are more or less effective in adjusting to novel and emerging risk factors. The analysis found substantial differences between the models. For example, the GBR and RFR models demonstrated their versatility by successfully managing complex datasets and correcting previous prediction failures. When confronted with complex risk factors, the Decision Tree Regression model struggled with overfitting, reducing its overall efficacy.

Recommendations for Practice

This study's findings provide valuable insights that may be used in practice and theory in insurance premium computation. Given the apparent benefits of machine learning models, particularly the Polynomial Regression model, insurance companies should consider integrating these advanced analytical approaches into their risk assessment frameworks. This model not only reduced prediction errors, as indicated by lower MSE and RMSE values, but it also proved the ability to capture significant variability in insurance premium data, as evidenced by high R-squared value. This integration has the potential to improve decision-making processes through data-driven approaches. It is consistent with the most recent advances in predictive analytics, which highlight the efficiency of machine learning in managing complex and multidimensional data sets. Nonetheless, it is critical to ensure that these findings are implemented with a thorough understanding of the specific settings and datasets unique to each insurance firm, avoiding drawing broad conclusions based on the results. This recommendation is supported by significant research demonstrating the effectiveness of machine learning in enhancing predictive modeling. However, it also warns against implementing it without consideration of underlying model assumptions and data characteristics.

Recommendations for Future Research

Expanding and improving the work on applying machine learning in insurance premium calculations opens up several options for future researchers to improve this study's framework and findings. An area worth looking into is the examination of various machine learning models or more sophisticated versions of the tested models, such as deep learning approaches. These techniques can detect more complex patterns and connections in data. This evolution is backed by this study’s discovery that Polynomial Regression, Random forest Regression, and Gradient Boosting Regression models had significant prediction accuracy. That shows that using more advanced models could lead to even better results. Moreover, investigating the incorporation of real-time data processing and analyzing how dynamic variables affect model performance should assist in overcoming one of the study's limitations—the dataset's static nature. Further research could examine how these models are used in different insurance policies to validate and increase the results' applicability. The next stage in this research is to undertake longitudinal studies to evaluate the performance of these machine-learning models over time. That would provide helpful information about their adaptability and long-term effectiveness in dealing with changing risk variables and market conditions. That would broaden the theoretical understanding of machine learning applications in actuarial science and provide practical insights for industry-wide predictive modeling technique development.

Conclusions

This study thoroughly explored how machine learning models can improve the accuracy of insurance premium computations. It focused on the issue of obsolete actuarial approaches, which frequently need to account for complex risk considerations. This study shows that models like Polynomial Regression, Random Forest, and Gradient Boosting Regression perform better and are more adaptable than traditional methodologies. These models are very successful at tackling emerging risk factors. The significant reduction in prediction errors and capacity to explain a significant percentage of the variance in premium data demonstrate the practical relevance and theoretical importance of incorporating advanced analytical models into insurance procedures. The fundamental outcome of the study is clear: adding machine learning models can improve accuracy, efficiency, and flexibility in insurance premium evaluations. That complements and expands on prior research, demonstrating that the area of actuarial science benefits considerably from applying these technologies. That helps bridge the gap between traditional methodologies and the needs of modern risk assessment. The findings verify theoretical advances in predictive analytics and provide a compelling case for their practical use, indicating a significant step forward in the development of risk management methods in the insurance business.

References

1. Yassine K, Abderrahim EA, Mostafa EH, Retakaful contributions model using machine learning techniques, J Islam Monet Econ Finance, 2023, 9(3), 511-32. doi: 10.21098/jimf.v9i3.16812
2. Bertsimas D, Orfanoudaki A, Algorithmic insurance [Internet], 2022 Dec 14 [cited 2024 May 2]. https://arxiv.org/abs/2106.00839
3. Eling M, Nuessle D, Staubli J, The impact of artificial intelligence along the insurance value chain and on the insurability of risks, Geneva Pap Risk Insur Issues Pract, 2022, 47, 205-41. doi: 10.1057/s41288-020-00201-7
4. Halima EH, Yassine T, Insurtech & blockchain: implementation of technology in insurance operations and its environmental impact, IOP Conf Ser Earth Environ Sci, 2022, 975, 012010. doi:10.1088/1755-1315/975/1/012010
5. Agashe HR, Bhangre PS, Karle AR, Kharde KS, Niphade AS, Insurance premium prediction and forecasting using machine learning, Int J Res Pub Rev, 2023, 4(5), 5459-5464. https://ijrpr.com/uploads/V4ISSUE5/IJRPR13523.pdf
6. Aissa H, Tarik A, Zeroual I, Yousef F, Using machine learning to predict outcomes of accident cases in Moroccan courts. Procedia Comput Sci, 2021, 184, 829-834. doi:10.1016/j.procs.2021.03.103
7. Hamdoun N, How machine learning is transforming the insurance sector: case of fraud detection in Morocco, Int J Appl Pattern Recognit, 2021, 6(4), 273-282. doi:10.1504/IJAPR.2021.118913
8. Oubibi M, Zhou Y, Oubibi A, Fute A, Saleem A, The challenges and opportunities for developing the use of data and artificial intelligence (AI) in North Africa: case of Morocco, In: Motahhir S, Bossoufi B, editors, Digital technologies and applications, ICDTA 2022, Lecture Notes in Networks and Systems, vol 455, Cham: Springer, 2022. doi :10.1007/978-3-031-02447-4_9
9. Zarifis A, Holland CP, Milne A, Evaluating the impact of AI on insurance: the four emerging AI- and data-driven business models, Emerald Open Res, 2023, 1(1). doi:10.1108/EOR-01-2023-0001
10. Marouane M, Mkik S, el menzhi K, The challenges of artificial intelligence for Moroccan companies?, 2021 Feb 13, 7, 2021.

11. Mostafa EH, Mohammed EH, Abderrahim EA, Minimization of value at risk of financial assets portfolio using genetic algorithms and neural networks, J Appl Finance Bank, 2016, 6, 1-3. https://api.semanticscholar.org/CorpusID:37862127
12. LAKHCHINI W, WAHABI R, EL KABBOURI M, Artificial intelligence & machine learning in finance: a literature review, Int J Account Finance Audit Manag Econ, 2022, 3(6-1), 437-455. doi:10.5281/zenodo.7454232
13. Jones KI, Sah S, The implementation of machine learning in the insurance industry with big data analytics, Int J Data Inform Intell Comput, 2023, 2(2), 21–38. doi:10.59461/ijdiic.v2i2.47
14. Zhang T, Prediction for insurance premiums based on random forest and multiple linear regression, BCP Bus Manag, 2023, 38, 2315-2321. doi:10.54691/bcpbm.v38i.4097
15. Hanafy M, Mahmoud OMA, Predict health insurance cost by using machine learning and DNN regression models, Int J Innov Technol Explor Eng (IJITEE), 2021, 10(3), 137-143. doi:10.35940/ijitee.C8364.0110321
16. Sprangers O, Schelter S, de Rijke M, Probabilistic gradient boosting machines for large-scale probabilistic regression, In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining [Internet], 2021 [cited 2024 May 2], p. 1510–20. http://arxiv.org/abs/2106.01682
17. Tomasini U, Wyart M, How deep networks learn sparse and hierarchical data: the sparse random hierarchy model [Internet], arXiv; 2024 [cited 2024 May 2]. http://arxiv.org/abs/2404.10727
18. Hamid Z, Ajmal S, Torshin I, Ethical Considerations in AI and Machine Learning, Cosmic Bull Bus Manag, 2023, 2(1). doi:10.13140/RG.2.2.15343.20648
19. Ferrara E, Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies, JMIR Preprints, 2023 Apr 21, 48399. doi:10.2196/preprints.48399
20. Stephen M, Potter K, Mohamed S, Ethical considerations in machine learning: balancing innovation and responsibility, Comput Sci, 2024.
21. Sahai R, Al-Ataby A, Assi S, Jayabalan M, Liatsis P, Loy C, et al, Insurance risk prediction using machine learning, In: Wah YB, Berry MW, Mohamed A, Al-Jumeily D, editors, Data science and emerging technologies, DaSET, 2022, Lecture notes on data engineering and communications technologies, vol 165, Singapore: Springer, 2023, doi:10.1007/978-981-99-0741-0_30
22. Lopez Garcia A, Tran V, Alic AS, Caballer M, Campos Plasencia I, Costantini A, et al, A cloud-based framework for machine learning workloads and applications, IEEE Access, 2020, 8, 18681-18692. doi:10.1109/ACCESS.2020.2964386
23. Sarker IH, Machine learning: algorithms, real-world applications and research directions. SN Comput Sci, 2021, 2, 160. doi:10.1007/s42979-021-00592-x
25. Eckert C, Neunsinger C, Osterrieder K, Managing customer satisfaction: digital applications for insurance companies, Geneva Pap Risk Insur Issues Pract, 2022, 47, 569–602. doi:10.1057/s41288-021-00257-z
26. Aytekin C, Neural networks are decision trees [Internet], arXiv, 2022 [cited 2024 May 3]. http://arxiv.org/abs/2210.05189
27. Sereyrath EM, Exploring experimental research: methodologies, designs, and applications across disciplines, SSRN Electronic Journal, March 2024. doi:10.2139/ssrn.4801767

28. Appendices

Appendix A: Key Metrics Used

Key metrics	Description
Mean Squared Error (MSE)	A low mean squared error (MSE) indicates that the model's predictions closely match the actual values, implying high accuracy. A significant mean squared error (MSE) implies that the model's predictions significantly differ from the actual values, implying a lack of precision.
Root Mean Squared Error (RMSE)	Lower RMSE values, like MSE, suggest that the model's predictions are more accurate as they are closer to the actual values. Higher RMSE values show that the model's predictions deviate from the actual values, implying a lower level of accuracy.
The coefficient of determination (R-squared)	A score of 0 implies that the model explains none of the variability in the target variable, whereas a value of 1 suggests that it explains all of it. Higher R-squared values show that the model is better fitted to the data, implying that the independent variables explain a more significant fraction of the variance in the dependent variable.

Appendix B: Characteristics of Insurance Policyholders

Features	Types	Values	
Tax horsepower	Numerical	Tax Horsepower Diesel	[6,7,9,12]
		Tax Horsepower Gasoline	[4, 5, 6, 7,..12]
Insurance premium	Numerical	From 2015.84 to 7981.51	
Insurance categories	Categorical	Tourism, Commerce < 3500 kg, Commerce > 3500 kg	
Sex	Categorical	Man, Woman	
Claim History	Categorical	No previous claims, Claim in the last year	
fuel type	Categorical	Diesel, Gasoline, Hybrid	

Appendix C: Standardization of Numerical Variables Using StandardScaler

	Tax horsepower	Insurance premium
0	-0.280530	0.458931
1	-1.228651	-0.521107

2	-0.280530	-0.973608
3	-0.754590	-0.428033
4	-0.280530	-0.951236
…	…	…
20995	-0.754590	-0.307580
20996	1.615712	-1.127722
20997	-0.754590	2.198070
20998	-1.702711	0.105728
20999	-0.754590	-1.158529