Jurnal Ilmiah
Teknologi Sistem
Informasi

Jitsi

Vol. 6 (2025) No. 2

https://jurnal-itsi.org/index.php/jitsi; E-ISSN: 2722-4600; ISSN: 2722-4619
DOI: 10.62527/jitsi.6.2.443

Forecasting Next Year's Health Insurance Claims Using Machine Learning Models

Iyad S. Alkrunz[#], Ali H Abuzaid[*,**]

[#] *Department of Information Technology, Al Azhar University -Gaza, Palestine*
** Department of Mathematics, Al Azhar University -Gaza, Palestine*
[**] *Department of Statistics and Applied Probability, University of California, Santa Barbara, California, USA*
E-mail: eyadalkronz[at]gmail.com , alizaid33[at]yahoo.com

A B S T R A C T S

This study explores the transformative potential of big data analytics in the realm of health insurance risk management. Focusing on data sourced from Highmark Health from 2015 to 2018, the research aims to evaluate the efficacy of advanced data manipulation techniques and machine learning models in enhancing predictive accuracy. The analysis involves a comprehensive examination of Health Maintenance Organization (HMO) and Preferred Provider Organization (PPO) plans, with rigorous data preparation processes such as cleaning, aggregation, feature engineering, and outlier handling to ensure model suitability. Four distinct models were developed: an initial model utilizing raw data without outlier treatment, a model post-outlier treatment considering both HMO and PPO members, and models focusing exclusively on HMO and PPO members respectively. Results demonstrated significant improvements in predictive accuracy following outlier treatment, with Random Forest and Multivariate Adaptive Regression Splines showing superior performance. The Random Forest model achieved a Root Mean Square Error (RMSE) of 630.04 and an R-squared value of 0.757, underscoring its robust predictive capabilities. Similarly, the Multivariate Adaptive Regression Splines model exhibited strong fit with commendable metrics. The HMO-focused model yielded promising outcomes with a minimal RMSE of 675.85 and an R-squared value of 0.68. However, the PPO-focused model's suboptimal results highlight potential data quality issues and dataset limitations. This research underscores the critical role of integrating machine learning techniques in health insurance analytics, providing valuable insights for proactive risk management and decision-making, and enhancing efficiency and effectiveness within the industry.

Manuscript received Apr 27, 2025; revised Apr 30, 2025. accepted Mei 01, 2025 Date of publication Jun 30, 2025. International Journal, JITSI : Jurnal Ilmiah Teknologi Sistem Informasi licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Keywords / Kata Kunci — ***Big Data Analytics; Claims Forecasting; Health Insurance; Machine Learning Models; Outlier Handling***

CORRESPONDING AUTHOR

Ali H. Abuzaid
Department of Mathematics, Al Azhar University - Gaza, Palestine
Department of Statistics and Applied Probability, University of California, Santa Barbara, California, USA
Email: alizaid33[at]yahoo.com

1. INTRODUCTION

Big data refers to the vast volumes of structured and unstructured information that exceed the capabilities of traditional data processing methods [3]. Big data analytics involves using advanced algorithms, statistical techniques, and machine learning to extract actionable insights. These insights enable organizations to uncover hidden correlations, trends, and anomalies, enhancing decision-making, fostering innovation, and opening new

opportunities across industries. In the insurance sector, these capabilities drive operational efficiency, enhance customer experiences, optimize resource allocation, and support proactive risk management.

Insurance provides financial protection against risks through contractual agreements between insurers and policyholders [9]. Accurate risk evaluation is essential for determining premium structures and managing liabilities [4, 10]. The advent of big data has transformed traditional insurance operations, offering new ways to manage information and improve strategic decision-making [14]. Health insurance, a key segment, covers medical costs, including hospital care, physician services, prescriptions, preventive care, and mental health support [9].

This research explores the integration of big data analytics and machine learning within insurance risk management, addressing the limitations of traditional actuarial methods. Key objectives include:

- Applying machine learning to large datasets
- Enhancing decision-making with data visualization
- Developing and refining claim prediction models
- Ensuring data quality through preprocessing
- Identifying key predictive features
- Training and evaluating predictive models
- Optimizing model performance
- Analyzing different customer groups
- Providing recommendations for predictive modeling in health insurance

By leveraging big data, insurers can uncover hidden patterns and optimize resource allocation, leading to more efficient operations, reduced costs, and personalized customer services. Advanced analytics empowers insurers to anticipate risks, improve claim predictions, and enhance decision-making, ultimately transforming the industry.

2. LITERATURE REVIEW

2.1 Insurance and Risk Management

Insurance is a contract where an individual or entity (the insured) pays a premium to an insurance company (the insurer) in exchange for financial protection against potential risks or losses [9]. In healthcare, health insurance covers medical expenses, helping individuals manage healthcare costs. Health insurance is crucial as it provides financial protection and access to necessary medical services, medications, and treatments, thereby reducing the financial burden of medical expenses [9].

2.1.1 Health Insurance Plans

Health insurance plans vary, with two common types being Health Maintenance Organization (HMO) and Preferred Provider Organization (PPO) plans.

1. Health Maintenance Organization (HMO):
   HMOs provide coverage through a network of healthcare providers, requiring the insured to choose a primary care physician (PCP) from the network. The PCP coordinates all medical services, including referrals to specialists. PCP who manages healthcare needs and coordinates services, as well as in-network coverage primarily limited to in-network providers, with lower out-of-pocket costs. Additionally, there is an emphasis on preventive care and wellness programs.
2. Preferred Provider Organization (PPO):
   PPOs offer more flexibility, allowing the insured to seek medical services from both in-network and out-of-network providers. PPO plan offers greater flexibility in choosing providers but typically has higher deductibles, copayments, and coinsurance.

2.1.2 Claim Types

Healthcare services are classified into three main claim types [9]:

- Inpatient Claims: Services provided to patients admitted to a hospital or inpatient facility, involving continuous medical care and monitoring. These claims often use a bundled payment system.
- Outpatient Claims: Services provided without hospital admission, such as routine check-ups and minor surgeries, billed on a fee-for-service basis.
- Professional Claims: Services provided by individual healthcare professionals in various settings, billed by the provider or practice based on fees for specific services or procedures.

2.2 Studies on Utilizing AI in Health Insurance

This section reviews scholarly investigations into the integration of artificial intelligence (AI) within health insurance. Various empirical studies have explored AI applications in insurance pricing, fraud detection, and risk management, providing valuable insights into AI's potential to transform traditional practices in the health insurance industry.

2.2.1 Fraud Detection in Insurance

Rai et al. [10] proposed a fraud detection model for automobile insurance claims using machine learning classifiers and oversampling techniques. The study employed the MWMOTE algorithm to handle class imbalance and built SVM, DT, and RF models. The RF classifier demonstrated superior performance in terms of precision, recall, and F1-score. The study highlights the computational expense of MWMOTE and suggests future research on deep learning models and parallelizing the algorithm on GPUs.

Gupta et al. [5] introduced a Comprehensive Fraud Management (CFM) framework integrating actuarial techniques and AI for insurance fraud detection. The CFM framework includes prevention, detection, and further analysis stages, using statistical models and machine learning on motor insurance data to efficiently identify fraudulent activities.

Severino and Peng [14] examined machine learning models for fraud detection in property insurance using data from a major Brazilian insurance company. They found that the Random Forest model outperformed other methods, while the Deep Neural Network model excelled in recall. The study also identified a macro profile of fraudsters and suggested future research in spatial analysis and additional machine learning algorithms.

2.2.2 AI in Insurance Pricing

Maynard et al. [6] explored the role of AI in insurance pricing using simulated property data from three U.S. states. Neural network methods consistently outperformed traditional estimation approaches in predicting claim frequency, particularly when working with degraded data. The study introduced "risk boost polygons" to identify regions with elevated risk and discussed the practical and regulatory factors influencing the adoption of deep learning in insurance pricing.

Blier-Wong et al. [2] reviewed the application of machine learning techniques in pricing and reserving for property and casualty (P&C) insurance. They highlighted the limitations of traditional models based solely on structured data and emphasized the advantages of neural networks for incorporating novel data sources such as telematics and medical histories. The review addressed key challenges like model explainability, prediction uncertainty, and bias, offering solutions such as anticlassification techniques and analysis of prediction uncertainty.

2.2.3 Big Data Analytics and AI in Insurance

Senousy et al. [13] surveyed the benefits of big data analytics in enhancing insurance business models. The study outlined use cases such as determining policy premiums, fraud detection, customer insights, marketing, and financial optimization, while also raising concerns regarding privacy, data protection, and market competition.

Paruchuri [9] highlighted the critical role of machine learning in analyzing client entitlements and performance metrics within the insurance sector. The study advocated for the integration of machine learning technologies to improve client management and data handling processes.

Wu et al. [16] investigated the challenges and opportunities associated with healthcare big data in China’s commercial health insurance sector. The authors provided concrete action plans and strategic recommendations to improve the financial sustainability and operational efficiency of the industry through data-driven approaches.

2.2.4 AI and Decision Support in Insurance

Rawat et al. [11] employed machine learning models to analyze insurance claim data and predict claim status with high accuracy. The study utilized feature selection techniques to minimize dimensionality, which significantly enhanced model performance. The Random Forest classifier showed the highest predictive accuracy, emphasizing the value of effective feature selection.

Taha et al. [15] introduced a feature selection framework designed to improve the efficiency of machine learning algorithms in the insurance domain. Empirical evaluations confirmed that models using selected features outperformed those using full datasets, reinforcing the importance of eliminating redundant and irrelevant information in insurance analytics.

2.3 Gaps Identified

The review of existing literature reveals several gaps in the application of advanced technologies within the health insurance industry. Many studies have a narrow focus, emphasizing areas like fraud detection and risk assessment, while predictive modeling for customer claims remains underexplored. Additionally, much of the research relies on theoretical frameworks rather than empirical evidence, underscoring the need for practical studies to validate the effectiveness of these technologies. Challenges related to data quality, privacy, and security are often overlooked, despite being critical to successful implementation. Furthermore, ethical and social implications, such as potential bias or discrimination in algorithmic decisions, receive limited attention. Finally, most studies depend on secondary data, which may limit the real-world relevance of their findings, highlighting the need for research using primary data to enhance validity

3. RESEARCH METHODOLOGY

3.1. Data Source and Description

The dataset used in this study is sourced from Highmark Health, a prominent healthcare organization based in Pittsburgh, Pennsylvania. Highmark Health operates one of the largest integrated healthcare delivery and financing systems in the United States. The dataset, which is publicly available, spans from January 1, 2015, to December 31, 2018, and can be accessed through the following link:
https://github.com/zachcarlson/InsuranceDatabase.
The dataset comprises five primary tables:

- **Members Table**: This table contains 1,439 records with essential personal information, including gender, date of birth, first name, last name, address details (street, city, county, state, zip code), product type, client ID, and membership start and end dates.
- **Claims Table:** With 95,127 records, this table includes information on claims such as claim status, type, member ID, provider ID, service date, code type, services rendered, and paid claims. A comprehensive dictionary of this dataset, detailing all tables and column descriptions, will be provided with this proposal.
- **Member Conditions Table:** This table includes attributes related to members' health conditions, with fields such as IDNO, MemberID, product, client ID, gender, age range, member region, and indicators for specific conditions (e.g., diabetes, coronary artery disease (CAD), congestive heart failure (CHF), hypertension, chronic obstructive pulmonary disease (COPD)). It also records the month of incorporation (IncMonth), total member count (Member_Count), and co-morbidity count (CoMorbidity_Count).

3.2 Description of the sample

In this section, we overview the dataset by analyzing members' personal information, including age and gender distributions, and examining claims data, such as claim types and their temporal trends. This foundational exploration sets the stage for a deeper understanding of the dataset's intricacies.

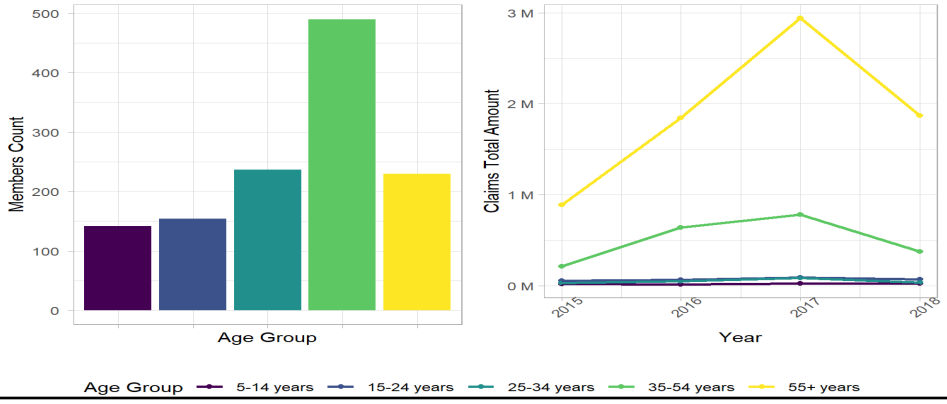

FIG 1. Total Claims Amount based on Age Group and year

Figure 1 displays the membership distribution across five age groups: 5-14, 15-24, 25-34, 35-54, and 55+. The bar chart reveals that the 35-54 years age group has the highest membership count, followed by the 25-34 years and 55+ years groups. The accompanying graph shows that the "55+ years" age group consistently incurs the highest total claims amounts over the years, with the 35-54 years group following in claims expenditure.

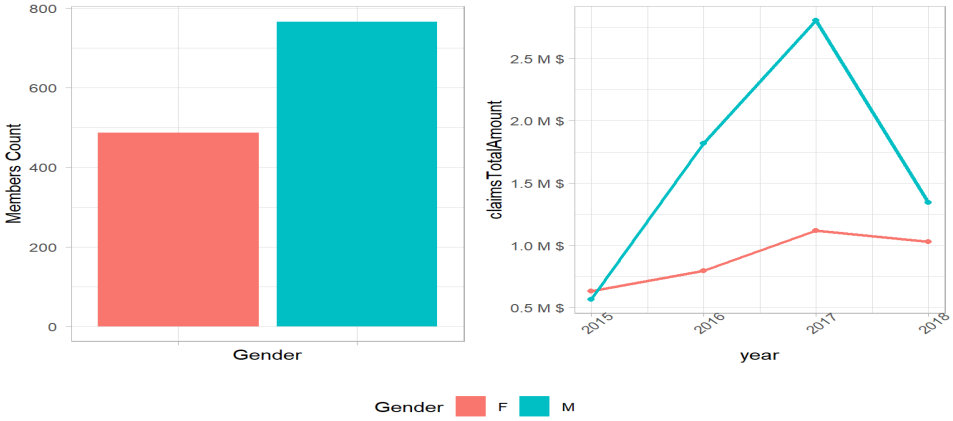

FIG 2. Total Claims Amount based on Gender and year

Figure 2 shows the gender distribution of members, with males comprising 61% (766 members) and females 39% (487 members). The accompanying chart indicates that males consistently have higher claim values compared to females across all years.

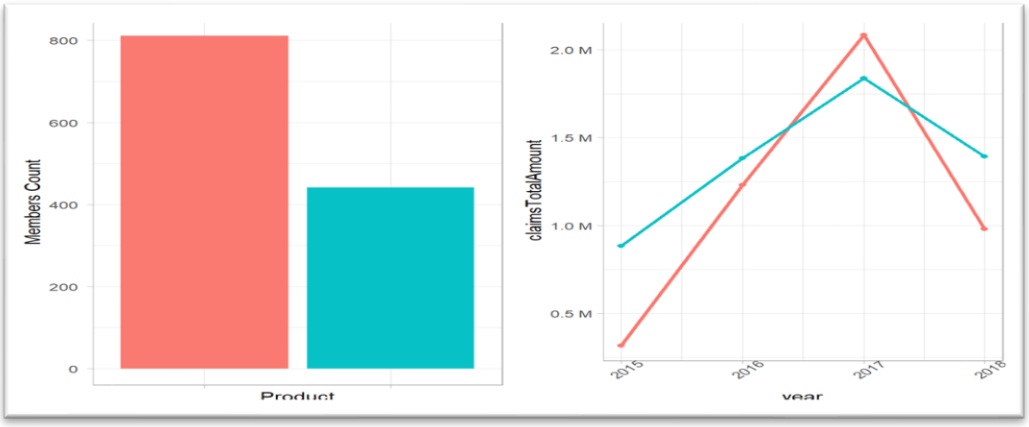

FIG 3. Total Claims Amount based on Product

Figure 3 shows that HMO members outnumber PPO members, with 811 versus 442. The right figure reveals that HMO claims amount steadily increased to about two million dollars in 2017 before declining to around one million dollars in 2018.

3.3 Performance indicators

Performance indicators, also known as evaluation metrics, are essential tools in assessing the effectiveness and accuracy of predictive models, particularly in the field of machine learning and data analysis. These metrics help measure the performance of a model by comparing its predictions to the actual values in the dataset. For a set of size n observations, where (y_i) is the actual value of the target variable for the (i)th data point, and $(\hat{y}_i)$ is the predicted value of the target variable for the (i)th data point. Several commonly used performance indicators are outlined by Murphy [7].

1. Root Mean Squared Error (RMSE):
   The Root Mean Squared Error is a widely used metric to assess the accuracy of regression models. It measures the average squared difference between the predicted values and the actual values in the dataset. It penalizes larger errors more heavily, making it sensitive to outliers and given by:
   $$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

2. R-squared (R^2):
   R-squared, also known as the coefficient of determination, is a metric that evaluates how well the regression model explains the variability of the data. It provides a value between 0 and 1, where 0 indicates that the model does not explain any of the variability, and 1 indicates that the model perfectly fits the data. It is obtained by
   $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (2)$$
   where $(\bar{y})$ is the mean of the actual values of the target variable.

3. Mean Absolute Error (MAE):
   Mean Absolute Error is a metric used to evaluate the accuracy of regression models. It measures the average absolute difference between the predicted values and the actual values, and given by
   $$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$
   Unlike RMSE, MAE does not square the errors, making it more robust to outliers.

4. Root Mean Squared Logarithmic Error (RMSLE):
   Root Mean Squared Logarithmic Error is a variation of RMSE, but it applies a logarithmic transformation to both the predicted and actual values before calculating the error and given by
   $$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2} \quad (4)$$
   This metric is commonly used when the target variable has a wide range and includes both small and large values.

5. Explained Variance (Explained_Var):
   Explained Variance measures the proportion of the variance in the target variable that the model's predictions account for. It provides insights into how well the model captures the variability of the data, and obtained by :

$$\text{Explained\_Var} = 1 - \frac{\text{Var}(y-\hat{y})}{\text{Var}(y)} \quad (5)$$

   where $\text{Var}(y)$ represents the variance of y, and $\text{Var}[\hat{y} - y] = \frac{1}{n}\sum\left(error - mean(error)\right)^2$.

6. Standard Deviation of RMSE (RMSE_SD):
   The Standard Deviation of RMSE is a measure of the variability or spread of the RMSE values across different subsets of the data or multiple runs of the model. It provides information about the consistency of the model's performance.
   Mathematical Formula:

$$\text{RMSE\_SD} = \sqrt{\frac{\sum_{i=1}^{k}(RMSE_i - \text{Mean\_RMSE})^2}{k}} \quad (6)$$

   where k is the number of subsets or runs, $(\text{RMS}E_i)$ is the RMSE value for the (i)th subset or run, and (Mean_RMSE) is the mean of all RMSE values across subsets or runs.

In summary, these performance indicators are used to evaluate the performance of regression algorithms by quantifying their accuracy, fit, and variability. Each metric serves a specific purpose in assessing different aspects of the model's predictive capabilities. When comparing regression models, it is essential to consider multiple metrics to gain a comprehensive understanding of their performance.

3.4 Data Preprocessing

To prepare the Highmark Health dataset for analysis, several data pre-processing steps will be undertaken:

- Identifying and Removing Missing Data: Any missing values in the dataset will be detected and addressed to ensure completeness.
- Outliers Handling: Outliers, data points deviating significantly from the majority in a dataset, can profoundly impact machine learning model performance (Abuzaid & Alkronz, 2024). Our research underscores the significance of handling outliers and showcases Isolation Forests' effectiveness in enhancing machine learning models.
- Converting Categorical Variables: Categorical variables will be converted into numerical formats suitable for analysis.
- Normalizing Numerical Variables: Numerical variables will be normalized to a common range to facilitate comparison and modeling.
- Feature Selection: Relevant features will be selected to enhance the analysis by focusing on significant variables.

Due to the large size of the data, it would be challenging to process using traditional frameworks. Therefore, we will be utilizing the Spark framework for data processing.

3.5 Model Construction

In the modeling phase of this research, we aimed to predict the "next_year_amount" using a variety of predictive variables. The dataset was split into 80% for training and 20% for testing to evaluate the models' performance on unseen data accurately.

Predictive Variables: We considered 12 predictor variables to build our models:

Table 1 Model Predictors description

Variable	Description
Gender	Male or Female
Age_group	less than 5, 5-14, 15-24, 25-34, 35-54, 55+ years
Product	HMO, PPO
Duration_months	Members total duration in insurance company in months
Has_Disease	Is member having any disease?
Claim_Amount	Total claim amount in base year
I_ClaimType_Claim_Amount_total	Total Inpatient Claims amount in base year
O_ClaimType_Claim_Amount_total	Total Outpatient Claims amount in base year
P_ClaimType_Claim_Amount_total	Total Professional Claims amount in base year
DRG_Code_type_Claim_Amount_total	Total DRG Claims amount in base year
HCPC_Code_type_Claim_Amount_total	Total HCPC Claims amount in base year
REVCD_Code_type_Claim_Amount_total	Total REVCD Claims amount in base year

Modeling Techniques and R Packages:

To develop robust and accurate predictive models, we utilized the R package "caret," which offers a wide range of machine learning algorithms and tools. Specifically, we employed the "train" and "resamples" methods provided by the "caret" package

Algorithms Applied:

The predictive models were constructed using the following machine learning algorithms: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Partial Least Squares, Support Vector Machines (Linear), K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting Machines, Multivariate Adaptive Regression Splines, and Bayesian Generalized Linear Models.

3.6 Model Building

Four models were constructed during this phase:

- Initial Model: (All Data before Outlier Treatment)
  This model was trained on the complete dataset, including all members (PPO and HMO) before excluding outliers.
- First Model (All Data after Outlier Treatment)
  Trained on the entire dataset, this model included all members—PPO and HMO—after excluding outliers.
- Second Model (HMO Members Only)
  This model focused solely on HMO members, using only their data for training and testing.
- Third Model (PPO Members Only)
  Developed specifically for PPO members, this model used their data exclusively for training and testing.

Separate models for PPO and HMO members were built to capture potential variations in predictive relationships unique to each group. Various machine learning algorithms were applied to predict the "next_year_amount" using relevant predictor variables, with the "caret" package facilitating model building and performance evaluation. Constructing distinct models for all members, HMO members, and PPO members allowed us to tailor predictions more effectively, potentially leading to more accurate and meaningful insights. The outcomes from these models contribute significantly to our research findings and provide valuable input for informed decision-making within the domain of interest.

3.7 Models Evaluation

In the model evaluation phase, we rigorously assessed the performance of the predictive models developed in the previous step using 20% of the data set aside for this purpose. To measure the models' effectiveness in predicting the "next_year_amount," we employed several performance indicators: Root Mean Squared Error (RMSE), R-squared (R^2), Mean Absolute Error (MAE), RMSE Standard Deviation (RMSE_SD), Explained Variance (Explained_Var), and Root Mean Squared Logarithmic Error (RMSLE).

4. RESULTS AND DISCUSSION

This subsection presents the findings from four models, designed to assess the impact of outlier handling and membership type on our results. The first model uses the raw dataset without outlier treatment, serving as a baseline. The second model incorporates data after handling outliers to evaluate the effect of data cleaning. The third and fourth models focus on specific membership subsets: PPO (Preferred Provider Organization) and HMO (Health Maintenance Organization) members. These models offer insights into unique patterns within the data, enhancing our understanding of the research objectives.

Initial Model: All Data without Outlier Treatment

We developed an initial predictive model using a dataset of 565 observations and 12 predictor variables, including all members' data without outlier handling. This approach aimed to assess the raw impact of the data, capturing potential extreme values and their influence on the predictors. By evaluating the model's performance on unfiltered data, we gained insights into its behaviour without preprocessing. Future iterations will apply outlier handling techniques to refine and improve predictive accuracy.

The evaluation of various algorithms, as shown in Table 2, indicates suboptimal performance. Low R^2, values and negative explained variance suggest that the model error exceeds the variance of the dependent variable, reflecting poor model performance. This underperformance is primarily due to the absence of outlier handling during preprocessing. Outliers introduce noise and affect parameter estimation, reducing predictive accuracy and impairing the model's ability to generalize to new data.

Table 2 Initial Model Results: Raw Data without Outlier Treatment

Model	RMSE	R^2	MAE	RMSE_SD	Explained_Var	RMSLE
Linear Regression	14005.25	0.006	4840.122	14046.24	-0.047	NA
Ridge Regression	14048.16	0.006	4845.218	14089.60	-0.053	NA
Lasso Regression	13672.88	0.003	4768.995	13714.49	0.002	1.866
Elastic Net	13759.62	0.006	4571.727	13803.30	-0.010	1.696
Partial Least Squares	**13723.07**	**0.011**	4354.789	13766.16	-0.005	1.622
Support Vector Machines	13769.50	0.004	**3709.822**	**13720.97**	-0.012	1.090
K-Nearest Neighbors	14667.05	0.000	4693.684	14709.85	-0.148	0.987
Decision Tree	13692.00	NA	4846.069	13732.19	0.000	1.896
Random Forest	14851.40	0.001	4735.759	14898.60	-0.177	1.383
Gradient Boosting Machines	13902.00	0.013	4348.662	13945.95	-0.031	1.495
Multivariate Adaptive Regression Splines	14040.84	0.006	4381.456	14084.13	-0.052	1.412
Bayesian Generalized Linear Models	14005.23	0.006	4839.904	14046.22	-0.047	NA

First Model: All Data After Outlier Treatment

We employed Isolation Forests within our machine learning models to effectively detect and manage outliers. As a result, 17 observations were identified as outliers and excluded from the analysis. After this adjustment, the dataset was reduced to 548 observations, and the models were reevaluated accordingly, showing significant improvements in predictive accuracy. Linear regression-based algorithms, such as Linear Regression and Ridge Regression, performed well, with low RMSE values of 698.78 and 698.82, and high R-squared values, capturing around 70% of the variance. Lasso Regression and Elastic Net also achieved respectable results, with RMSE values of 763.07 and 714.57, and R-squared values close to 70%.

Among non-linear models, Random Forest performed the best, yielding the lowest RMSE of 630.04 and a high R-squared value of 0.757, outperforming Gradient Boosting Machines. While Multivariate Adaptive Regression Splines also achieved a low RMSE of 645.56 and a strong R-squared value of 0.75, models like Decision Tree and Gradient Boosting Machines exhibited higher RMSEs, indicating less accurate predictions.

Table 3 First Model Results: All Data After Outlier Treatment

Model	RMSE	R2	MAE	RMSE_SD	Explained_Var	RMSLE
Linear Regression	698.783	0.701	486.679	701.818	0.701	NA
Ridge Regression	698.818	0.702	483.716	701.854	0.701	NA
Lasso Regression	763.073	0.699	488.301	766.104	0.643	0.562
Elastic Net	714.573	0.700	480.901	717.575	0.687	0.489
Partial Least Squares	700.439	0.700	476.327	703.450	0.700	0.387
Support Vector Machines (Linear)	809.641	0.702	465.882	796.556	0.599	0.211
K-Nearest Neighbors	747.214	0.700	494.277	746.282	0.658	0.442
Decision Tree	842.659	0.569	501.484	846.324	0.565	0.594
Random Forest	**630.043**	**0.757**	**440.153**	**632.541**	**0.757**	0.545
Gradient Boosting Machines	861.929	0.548	534.615	865.038	0.545	0.463
Multivariate Adaptive Regression Splines	645.557	0.752	456.816	648.164	0.745	0.636
Bayesian Generalized Linear Models	698.758	0.701	486.603	701.793	0.701	NA

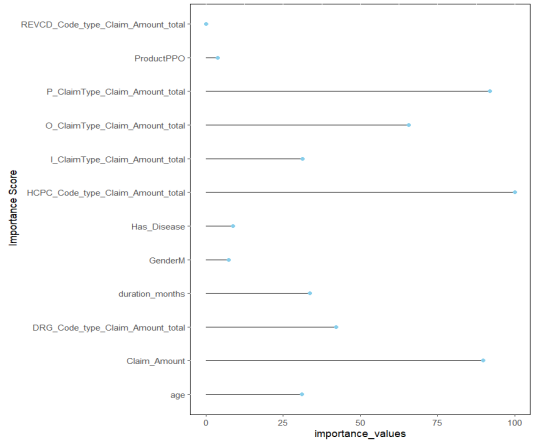

FIG 4. First Model, Variable Importance for Random Forest model

The variable importance plot for the Random Forest model offers insights into the predictors' influence. HCPC_claim_amount_total emerged as the most impactful variable, with the highest bar height, followed by P_Claim_Amount_total, O_Claim_Amount_total, and others with decreasing influence. Notably, the absence of a bar for REVCD_claim_Amount_total indicates its negligible impact on predictions.

In conclusion, handling outliers improved the performance of the regression and machine learning models, with linear regression-based algorithms and Random Forest performing particularly well in predicting the target variable. These findings highlight the importance of outlier handling in enhancing model accuracy and reliability, enabling more meaningful insights and better-informed decision-making.

Second Model (HMO Members Only):

The second model focused exclusively on HMO members, using 273 observations from this subset for training and testing. The goal was to evaluate how the predictive models perform with this limited dataset. Among the algorithms, Multivariate Adaptive Regression Splines achieved the best results, with the lowest RMSE of 675.85 and the highest R-squared value of 0.68, indicating a good fit and explaining 68% of the variance. Support Vector Machines (Linear) and Random Forest also performed well, with Random Forest achieving an RMSE of 693.27 and an R-squared value of 0.67, demonstrating strong predictive accuracy.

However, not all models performed equally well. Linear Regression, Gradient Boosting Machines, and Decision Tree showed fewer promising results, with relatively higher RMSE values and lower R-squared values, suggesting less accurate predictions and weaker fit to the data.

Table 4 Second Model Results: (HMO Members Only)

Model	RMSE	R2	MAE	RMSE_SD	Explained_Var	RMSLE
Linear Regression	887.527	0.510	619.723	876.837	0.436	NA
Ridge Regression	785.180	0.618	580.960	775.636	0.558	NA
Lasso Regression	782.551	0.600	577.538	775.173	0.561	NA
Elastic Net	728.466	0.636	534.356	723.264	0.620	0.610
Partial Least Squares	736.896	0.635	531.033	730.260	0.611	0.642
Support Vector Machines (Linear)	752.320	0.664	482.204	747.097	0.594	0.272
K-Nearest Neighbors	827.830	0.595	510.265	831.155	0.509	0.545
Decision Tree	829.520	0.636	521.038	827.948	0.507	0.617
Random Forest	693.274	0.672	479.706	699.975	0.656	0.691
Gradient Boosting Machines	888.558	0.475	564.119	895.582	0.434	NA
Multivariate Adaptive Regression Splines	**675.846**	**0.683**	475.936	682.006	0.673	0.701
Bayesian Generalized Linear Models	887.388	0.510	619.574	876.696	0.436	NA

In conclusion, the second model, focused on HMO members with 273 observations, showed that algorithms like Multivariate Adaptive Regression Splines, Random Forest, and Support Vector Machines (Linear) performed best, offering higher predictive accuracy. In contrast, models such as Linear Regression, Gradient Boosting Machines, and Decision Tree delivered weaker results. These findings emphasize the importance of tailoring models to specific data subsets and highlight the role of algorithm selection in achieving optimal predictive performance.

Third Model (PPO Members Only):

The third model focused on PPO members, using 275 observations from this subset for training and testing. The results revealed varied predictive performance across algorithms. Gradient Boosting Machines achieved the best outcome, with the lowest RMSE of 668.4 and an R-squared value of 0.22, indicating strong predictive capability and explaining 22% of the variance. K-Nearest Neighbors also performed well, with an RMSE of 673.52 and an R-squared value of 0.20, effectively capturing meaningful patterns. Random Forest and Support Vector Machines (Linear) showed moderate accuracy with low RMSE values, while Decision Tree and Multivariate Adaptive Regression Splines performed poorly, with higher RMSE values and weaker fits to the data.

Table 5 Third Model Results: (PPO Members Only):

Model	RMSE	R2	MAE	RMSE_SD	Explained_Var	RMSLE
Linear Regression	718.539	0.179	477.619	723.512	0.048	0.291
Ridge Regression	732.759	0.177	484.029	737.205	0.010	0.182
Lasso Regression	718.619	0.177	475.793	723.166	0.048	0.279
Elastic Net	710.362	0.166	477.120	715.587	0.069	0.385
Partial Least Squares	732.401	0.153	487.556	737.376	0.011	0.330
Support Vector Machines (Linear)	709.689	0.168	450.982	696.062	0.071	0.091
K-Nearest Neighbors	673.521	0.196	426.837	671.479	0.163	0.311
Decision Tree	1069.236	0.068	608.433	1073.006	-1.109	0.644
Random Forest	691.323	0.186	448.956	696.716	0.119	0.447
Gradient Boosting Machines	**668.396**	**0.217**	420.860	671.608	0.176	0.423
Multivariate Adaptive Regression Splines	716.043	0.068	477.691	721.691	0.054	0.613
Bayesian Generalized Linear Models	718.546	0.179	477.648	723.520	0.048	0.292

The third model's results highlight the importance of tailoring predictive models to specific data subsets, such as PPO members. The findings demonstrate that some algorithms, like Gradient Boosting Machines and K-Nearest Neighbors, are better suited for predicting outcomes within this group. This underscores the need for careful algorithm selection and dataset curation to build reliable models that offer meaningful insights for specific populations.

5. DISCUSSION

5.1. Interpretation of Results

The four models were developed to analyze health outcomes for different groups of members, with varying data preprocessing techniques. The initial model without outliers handling showed unsatisfactory performance across all algorithms, as presented in Table 2, emphasizing the criticality of addressing outliers in the data preprocessing phase to improve predictive accuracy.

After applying outlier handling methods, substantial improvements were evident in the initial model's predictive capabilities. Random Forest stood out as a promising algorithm, exhibiting a decreased RMSE of 630.043 and an increased R-squared value of 0.757, as indicated in Table 3, highlighting its enhanced predictive accuracy and strong fit to the data. Additionally, Multivariate Adaptive Regression Splines demonstrated commendable performance in this phase, yielding an RMSE of 645.56 and an R-squared value of 0.75, further affirming its efficacy in predictive modeling.

In Table 4, the performance of the second model, which exclusively targets HMO members, revealed varying degrees of effectiveness across different algorithms. Notably, Multivariate Adaptive Regression Splines, Random Forest, Support Vector Machines (Linear), and Elastic Net demonstrated superior predictive accuracy. Conversely, models such as Linear Regression, Bayesian Generalized Linear Models, and Gradient Boosting Machines showed fewer promising results. This underscores the significance of customizing models to suit specific data subsets, thereby enhancing predictive capabilities.

Table 5 presents the evaluation findings of the third model, crafted for PPO members. It elucidates the superior performance exhibited by Gradient Boosting Machines and K-Nearest Neighbors, surpassing other algorithms in predictive accuracy, Nevertheless, it's crucial to acknowledge that, the overall effectiveness of all models remained unsatisfactory.

Insights from these models underscore the significance of targeted data analysis for specific member groups and the crucial role of proper data preprocessing. Implementing outliers handling and focusing on specific subsets of members can lead to more reliable and insightful predictions, enabling better decision-making in health insurance. It is essential to leverage the strengths of different algorithms while considering the unique characteristics of each member group to optimize predictive performance and derive meaningful insights from the data.

5.2. Limitations and Ethical Considerations

While this thesis followed ethical standards, the results should be interpreted with caution, considering the limitations of the data and methodology. Further research across different populations is necessary to validate findings and enhance generalizability.

- Data Bias: The dataset may contain inherent biases, such as under- or over-representation of certain groups, affecting model performance and generalizability.
- Data Quality: Errors, missing values, or inconsistencies in the data can impact the reliability of results.
- Model Assumptions: Deviations from assumptions about data relationships may affect performance and the validity of conclusions.
- Limited Generalizability: Results are specific to the studied community and may not directly apply to other populations with different demographics or socioeconomic factors.
- Data Changes: Future updates or changes in data could alter model predictions and performance.
- Model Complexity: Simpler algorithms were selected based on the research scope, though more complex models could yield different results.
- External Factors: Other variables not included in the analysis might influence predictions.

6. CONCLUSION AND FUTURE WORK

6.1. Conclusion

The research draws upon health insurance data spanning from 2015 to 2018, sourced from Highmark Health, encompassing diverse tables including Members (consisting of 1439 records) delineating personal and insurance particulars, and Claims (comprising 95127 records) elucidating claim-related details. A series of data manipulations were undertaken to prepare the data for modeling, involving data cleaning, aggregation, feature engineering, and outlier handling to ensure its suitability for modeling purposes.

The findings underscore the paramount importance of outlier treatment, as discernible enhancements were observed in the results subsequent to their rectification. The initial model demonstrated the efficacy of machine learning in accurately forecasting claim amounts for the ensuing year, manifesting an interpretable Root Mean Square Error (RMSE) of 630.04 and a relatively high R-squared value of 0.757 employing the Random Forest algorithm.

The subsequent model, focusing solely on members enrolled in the HMO program, yielded commendable outcomes, with the most proficient model achieving a minimal RMSE of 675.85 and the highest R-squared value of 0.68 utilizing Multivariate Adaptive Regression Splines.

Conversely, the third model yielded suboptimal results, potentially attributed to the data quality pertaining to members of the PPO category and the inadequacy of data utilized in the modeling process, encompassing a mere 275 records. However, there remains scope for improvement by augmenting the dataset or employing additional data processing techniques.

Among the investigated models, the Random Forest model emerged as one of the most effective, exhibiting an RMSE of 630.04 and an R-squared value of 0.757 in the first model, and an RMSE of 693.274 and an R-squared value of 0.672 in the second model. Similarly, the Multivariate Adaptive Regression Splines model showcased commendable performance, achieving an RMSE of 645.56 and an R-squared value of 0.75 in the first model, and an RMSE of 675.85 and an R-squared value of 0.68 in the second model, indicative of a robust fit to the data.

Regarding influential variables, the HCPC claim amount_total emerged as the most prominent predictor, followed by variables such as P_Claim_Amount_total and O_Claim_Amount_total, each exerting a notable influence on the outcomes.

In conclusion, the research underscores the paramount importance of integrating cutting-edge artificial intelligence and machine learning techniques within the domain of health insurance. Through a comprehensive review of various predictive models, the research elucidates their pivotal role in accurately forecasting claim amounts, consequently enabling proactive risk mitigation strategies.

6.2. Future Work

Future research should generalize the developed models to various insurance datasets, assessing their robustness across sectors like life, car, and property insurance. Enhanced evaluation and rigorous validation are essential for improving predictive accuracy and reliability.

Additionally, methodologies can be expanded to tackle fraud detection with innovative algorithms, while integrating non-traditional data sources such as social media and IoT can deepen insights into customer behavior.

Lastly, it's vital to explore the ethical and regulatory challenges of advanced analytics in insurance, focusing on data privacy, fairness, and compliance to ensure responsible use of technology while maintaining industry trust.

REFERENSI

[1] Abuzaid, A., & Alkronz, E. (2024). A comparative study on univariate outlier winsorization methods in data science context. *Statistica Applicata - Italian Journal of Applied Statistics*, 36(1), 85–99.

[2] Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2020). Machine learning in P&C insurance: A review for pricing and reserving. *Risks*, 9(1), 4.

[3] Boobier, T. (2016). *Analytics for insurance: The real business of big data*. John Wiley & Sons.

[4] Dorfman, M. S. (1998). *Introduction to risk management and insurance* (6th ed.). Prentice Hall.

[5] Gupta, R., Mudigonda, S., Kandala, P., & Baruah, P. K. (2019). A framework for comprehensive fraud management using actuarial techniques. *International Journal of Scientific and Engineering Research*, 10, 780–791.

[6] Maynard, T., Bordon, A., Berry, J. B., Baxter, D. B., Skertic, W., Gotch, B. T., Shah, N. T., Wilkinson, A. N., Khare, S. H., & Jones, K. B. (2019). What role for AI in insurance pricing. *A Preprint*.

[7] Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

[8] National Association of Insurance Commissioners (NAIC). (2024). Health insurance. Retrieved March 5, 2024, from https://content.naic.org/consumer/health-insurance.htm

[9] Paruchuri, H. (2020). The impact of machine learning on the future of insurance industry. American Journal of Trade and Policy, 7(3), 85–90.

[10] Rai, N., Baruah, P. K., Mudigonda, S. S., & Kandala, P. K. (2018). Fraud detection supervised machine learning models for an automobile insurance. International Journal of Scientific and Engineering Research, 9(11), 473–479.

[11] Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. International Journal of Information Management Data Insights, 1(2), 100012.

[12] Rejda, G. E. (2005). Principles of risk management and insurance. Pearson Education India.

[13] Senousy, Y. M. B., Mohamed, N. E.-K., & Riad, A. (2018). Recent trends in big data analytics towards more enhanced insurance business models. International Journal of Computer Science and Information Security, 30111817, 39–45.

[14] Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. Machine Learning with Applications, 5, 100074.

[15] C. Chang Yu, I. R. A Hamid, Z. Abdullah, K. Kipli, and H. Amnur, “A Multi-tier Model and Filtering Approach to Detect Fake News Using Machine Learning Algorithms,” JOIV Int. J. Inform. Vis., vol. 8, no. 2, p. 643, May 2024, doi: 10.62527/joiv.8.2.2703.

[16] Taha, A., Cosgrave, B., & Mckeever, S. (2022). Using feature selection with machine learning for generation of insurance insights. Applied Sciences, 12(6), 3209.

[17] Wu, J., Qiao, J., Nicholas, S., Liu, Y., & Maitland, E. (2022). The challenge of healthcare big data to China’s commercial health insurance industry: Evaluation and recommendations. BMC Health Services Research, 22(1), 1189