The Economics of Health Insurance Coverage Levels in the U.S.: A Predictive Modelling of Policyholder Preferences

Abstract

This study examines the determinants influencing the selection of coverage levels of Basic, Standard, and Premium health insurance plans in private markets, using a synthetic dataset modeled on the insured U.S. population. Multinomial logistic regression and random forest models were employed to evaluate the impact of demographic, socioeconomic, lifestyle, and clinical variables. The findings reveal that insurance cost is the most decisive factor, with higher premiums steering consumers away from basic plans toward more comprehensive options. Older individuals, those with higher BMI, and those with more children were more likely to choose lower-tier coverage, likely due to financial constraints, while younger individuals preferred premium plans. Surprisingly, smokers and those with a history of heart disease often selected Basic coverage, suggesting cost-related underinsurance among high-risk groups. Other influencing factors included gender, exercise habits, region, and occupation. The random forest model validated these results with an accuracy of 80%. Overall, the study highlights that insurance choices are shaped by a complex interplay of affordability, perceived risk, and socioeconomic context, underscoring the need for personalized pricing, streamlined plan design, and targeted support tools to promote equitable and efficient plan selection.

Introduction

Health insurance is more than just a financial product. It is a fundamental component of well-being that protects individuals and households from the unpredictability of healthcare expenses while enabling access to timely, essential services. In the United States, where healthcare costs remain among the highest globally, insurance coverage often determines whether a person seeks preventive care, receives critical treatment, or falls into medical debt (Hoagland et al., 2024). It is not surprising, then, that insurance status has become a key social determinant of health, influencing outcomes across socioeconomic strata.

The U.S. health insurance landscape is bifurcated into public and private systems. While public programs such as Medicaid and Medicare offer fixed benefit packages based on eligibility, private insurance markets offer more flexibility, often in the form of vertically tiered plans, such as Basic, Standard, and Premium coverage levels (Marone & Sabety, 2022). These plans vary not only in cost but in risk exposure, deductibles, and service comprehensiveness. This vertical differentiation is designed to empower consumers to choose a coverage level aligned with their health risk and financial means (Fang & Kung, 2021; Yang et al., 2016). However, in practice, such freedom introduces complexity that many individuals are ill-equipped to navigate.

Research has shown that even in markets offering substantial choice, plan selection is rarely optimal. Consumers often struggle with understanding trade-offs, misjudging their future healthcare needs, or are swayed by behavioral biases such as loss aversion, framing effects, and inertia (Barker et al., 2021; Marone & Sabety, 2022). This mismatch between choice and actual needs, termed as mis-insurance, can result in both under-insurance and over-insurance, with profound implications for household financial security and health outcomes (Yang et al., 2016; Sun, 2020).

While the determinants of insurance enrollment have been widely studied, especially in public schemes, there is a surprising scarcity of research focused on the factors influencing the choice of coverage level in private markets. Studies from diverse contexts, including Ghana, Indonesia, and Kenya (Adjei-Mantey & Horioka, 2023; Sukartini et al., 2021; Yego et al., 2023) have identified income, education, marital status, and access to healthcare as key predictors of enrollment. However, these studies typically treat insurance as a binary decision (enroll or not enroll), overlooking the layered decision-making process required when choosing between competing coverage options.

The literature increasingly suggests that insurance choice is shaped by a combination of objective characteristics, such as age, body mass index (BMI), occupation, and chronic conditions, as well as subjective expectations, including anticipated utilization and perceived vulnerability (Barker et al., 2021; Hoagland et al., 2024). For instance, individuals with a history of smoking or heart disease may opt for more comprehensive plans, while younger, healthier adults may favor basic coverage with lower premiums (Sun, 2020). Moreover, recent findings show that administrative and structural barriers such as claim denials for preventive services are more common among low-income and minority groups, compounding the challenge of accessing appropriate coverage (Hoagland et al., 2024).

The objective of this study is to determine the factors that influence policyholders' preferences for specific coverage levels in private health insurance, namely, Basic, Standard, or Premium. It aims to determine how demographic, socioeconomic, behavioral, and health-related characteristics influence these preferences and whether predictive patterns can inform more responsive insurance design. To achieve this, the study utilizes a simulated dataset that reflects real-world consumer profiles. It analyzes how demographic factors (e.g., age and gender), lifestyle factors (e.g., smoking status and exercise habits), socioeconomic factors (e.g., occupation and region), and health-related factors (e.g., BMI and medical and family history) influence the likelihood of selecting each tier. The methodological approach combines logistic regression for interpretability with Random Forest classification to improve prediction accuracy and capture complex interactions among variables (Sun, 2020).

This dual-mode modeling framework enhances our understanding of who chooses what level of insurance and why, providing practical insights for insurers, regulators, and healthcare advocates. For insurers, the findings can inform the design of more personalized and equitable insurance products. For policymakers, the findings underscore the need for greater transparency, decision support tools, and targeted outreach to vulnerable populations. As (Marone & Sabety, 2022) argue that vertical choice without informed decision-making tools may widen disparities and erode the very welfare gains that insurance markets are meant to provide. This study makes a timely and policy-relevant contribution to the literature on health insurance design and consumer behavior. In an era where financial protection and access to healthcare are increasingly determined by the fine print of one’s coverage level, understanding the factors behind these choices is not only academically important but also socially urgent.

The scope of this study is limited to the U.S. private or commercial insurance landscape, utilizing synthetic, cross-sectional data that captures consumer-side characteristics but excludes insurer-

level variations such as benefit design, provider networks, and employer-based plan sponsorship. The findings may not be generalized to health systems with centralized or universal models, where institutional incentives differ markedly. While the dataset enables robust predictive modeling, it does not permit causal inference or account for dynamic behavior over time. Additionally, unobserved behavioral factors such as perceived value or information asymmetry limit the study's ability to capture the complexity of real-world decision-making fully. Nevertheless, the analysis yields valuable insights into the determinants of coverage level selection, providing a scalable framework for insurers seeking to optimize plan design and for policymakers aiming to address coverage disparities across demographic and clinical risk groups.

Literature Review

An emerging body of literature has focused on understanding the factors that influence national-level health insurance coverage. These studies have explored a diverse range of socioeconomic, demographic, and structural determinants that shape individuals' decisions to enroll in health insurance programs, as well as the broader implications for healthcare expenditure and equity. By examining country-specific contexts, researchers have provided valuable insights into the unique challenges and opportunities associated with expanding insurance coverage. The following section highlights key empirical contributions that have examined the dynamics of national health insurance in various countries, illustrating how individual behavior, policy design, and institutional frameworks interact to influence coverage outcomes.

The study by Adjei-Mantey & Horioka, (2023) investigated the factors influencing health insurance enrollment and healthcare spending in Ghana, drawing on micro-level data from Wave 7 of the Ghana Living Standards Survey (GLSS7). Their study focused particularly on the role of individual risk preferences and the availability of healthcare facilities within local communities. The findings revealed that individuals who are more risk-averse are significantly more likely to enroll in health insurance compared to their less risk-averse counterparts. Interestingly, the study also found that extremely poor households were more likely to be enrolled in health insurance, possibly due to their exemption from paying premiums under Ghana's health insurance scheme. Furthermore, the availability of health facilities within one's community was associated with a significant reduction in out-of-pocket healthcare expenditures, highlighting the importance of local access to care in managing health costs.

Hughes & Kaya, (2021) Investigated the long-run dynamics of healthcare expenditure, focusing on national health insurance coverage. Their findings revealed that the effects of increasing enrollment in Medicaid and Medicare on per capita expenditure are different. While Medicaid enrollment increases per capita expenditure, higher enrollment in Medicare brings about lower per capita expenditure.

In a recent study, Yego et al., (2023) harnessed the power of machine learning to uncover the key drivers influencing health insurance uptake in Kenya. The analysis identified poverty vulnerability, participation in social security schemes, income levels, educational attainment, and marital status as the most significant predictors of insurance enrollment. By revealing these patterns, the study highlights the urgent need to address affordability barriers and develop targeted, data-driven interventions that expand insurance coverage. These findings provide

valuable insights for policymakers seeking to accelerate progress toward Universal Health Coverage (UHC) and ensure equitable access to quality healthcare services for all Kenyans.

Sukartini et al., (2021) examined the key factors influencing enrollment in Indonesia's national health insurance program. Their study investigated a range of individual and household characteristics, including age, education level, wealth quintile, place of residence, number of living children, marital status, employment status, income, and existing insurance coverage. Their findings revealed that education, economic status, and demographic factors play a significant role in shaping individuals' likelihood of enrolling in the national health insurance scheme. These results underscore the importance of addressing social and economic disparities to promote participation and move closer to achieving universal health coverage in Indonesia.

While these previous studies provide valuable insights into the determinants of health insurance enrollment at the national level, their focus differs markedly from the specific issue of how individuals choose the level of coverage within health insurance plans offered by private health insurance entities. First, the studies primarily examine public or government-supported health insurance schemes such as Ghana's National Health Insurance Scheme (NHIS), Kenya's emerging UHC program, Indonesia's JKN program, and the U.S. Medicaid and Medicare systems. These programs often operate under universal or subsidized models where the main decision point is whether to enroll or not, especially for lower-income or vulnerable populations. Consequently, the drivers explored things such as poverty vulnerability, risk aversion, access to healthcare facilities, social protection participation, and demographic characteristics that are relevant to insurance uptake but not necessarily to the type or level of plan chosen. In contrast, the decision-making process in private health insurance markets involves a more nuanced and consumer-driven evaluation. Individuals must choose from a variety of plans offering different levels of coverage (e.g., Basic, Standard, Premium), each associated with varying costs, benefits, and risk-sharing arrangements. This adds complexity to the decision, as factors such as health expectations, risk tolerance, price sensitivity, benefit preferences, income elasticity, and perceived value become crucial in determining the level of insurance coverage chosen, not just whether to enroll or not.

Moreover, while national health insurance schemes often feature standardized or uniform benefit structures, private health insurance markets are highly fragmented, offering diverse options that require individuals to assess trade-offs between cost and coverage. As such, predicting coverage level choice requires a deeper understanding of consumer behavior, expectations of future health needs, and preferences for financial protection—factors that are typically under-explored in the public insurance enrollment literature. Therefore, the current study distinguishes itself by shifting the focus from insurance enrollment to the choice of coverage level within a commercial context. This distinction is crucial for informing insurers, policymakers, and healthcare market analysts on how to design and target products that better align with consumers' actual needs and expectations.

Diving into commercial health insurance, a significant portion of studies' attention has shifted to healthcare costs and insurance premium amounts. For example, Hanafy and Mahmoud (2021) found that individual characteristics, such as age, gender, and smoking habits, significantly impact the cost of premiums. Similarly, Terlizzi & Cohen (2022) highlighted that geographic

location plays a key role in determining insurance costs in the United States, with regions like the Southeast generally experiencing higher premiums than others. Bhardwaj et al., (2020) further emphasized that an individual's health status often has a stronger influence on insurance costs than the specific terms set by insurers. In another study, Sun (2020) used predictive analytics and personal attributes to show that the number of children and body mass index (BMI) are also strongly correlated with insurance expenses. Orji and Ukwandu (2024) deployed three regression-based machine learning models to explain the cost prediction of health insurance. The study revealed that age, chronic disease, and family health history were the most significant factors influencing the premium price. Yamada et al. (2014) also examine how the decision to purchase private insurance is influenced by household income, socio-demographic factors, and private health insurance factors. The study found that household income affects the purchase of health insurance.

While these studies provide valuable insights, they have primarily focused on predicting insurance costs using supervised machine learning models, often treating cost as a fixed outcome. However, one critical factor has been largely overlooked: the cost of insurance is not simply predetermined; it is closely tied to the level of coverage an individual chooses. In other words, the premium amount is often a reflection of the breadth and depth of the coverage selected. This study argues that understanding what drives individuals to choose different levels of insurance coverage is a crucial step in explaining variations in insurance costs. Therefore, the focus of this research shifts from directly forecasting premiums to identifying the key factors that influence coverage choices. By employing both mathematical modeling and machine learning techniques, this study aims to uncover the underlying variables that guide consumers' decisions regarding the scope of their health insurance plans.

Methods

The dataset for this study was sourced from Kaggle, providing a comprehensive foundation for analyzing predictions of health insurance coverage levels. An initial exploratory data analysis was conducted to assess the structure, distribution, and relationships within the dataset, ensuring its suitability for predictive modeling. The dataset was also scaled to provide standardisation for the model to analyse.

Model Framework

Following established methodologies (Gupta & Kanungo, 2022; Yego et al., 2023), we employed logistic regression to examine the predictive roles of key determinants influencing health insurance coverage levels. Logistic regression was chosen due to its proven effectiveness in modelling multi-class classification problems, where the dependent variable represents categorical outcomes. This model estimates the probability of selecting a particular coverage level within a range of 0 and 1 given a specified set of predictor variables. Additionally, the exponentiation of logistic regression coefficients allows for interpretation in terms of odds ratios, a feature that enhances the model's applicability in understanding the relative impact of independent variables (Hilbe, 2015). These characteristics have contributed to the widespread adoption of logistic regression in statistical and econometric analyses, reinforcing its suitability for this study.

Given the categorical nature of the dependent variable, we adopted a multi-class logistic regression approach as used by (El Kassimi et al., 2024) to differentiate between Basic, Standard, and Premium insurance coverage levels. We then defined the outcome variable $Y_i \epsilon \{0, 1, 2\}$, representing insurance coverage level, with 0 = Basic, 1 = Standard, and 2 = Premium. We estimated $X_i = [x_{i1}, x_{i2,} \ldots, x_{ip}]^T$ to be the vector of predictor variables (age, BMI, occupation, etc). The probability of an individual selecting a given coverage level is modeled as:

$$P(Y_i = k|X_i) = \frac{\exp(\beta_k^T X_i)}{\sum_{j=0}^{2} \exp(\beta_j^T X_i)}, \qquad k \ \epsilon \{0, 1, 2\} \quad \text{------------------------ (1)}$$

Where:

Y_i *is the chosen coverage level or class* (*Basic*, *Premium and Standard*)

X_i is the vector of independent variable (age, BMI, occupation, etc)

β_k is the coefficient vector associated with class (Basic, Premium and Standard)

So in computing $(\beta_j^T X_i)$, *we took the dot products of the coefficient and feature values*:

$$(\beta_j^T X_i) = \sum_{j=1}^{p} \beta_{ij} \cdot x_{ij} \quad -- (2)$$

This gave us a single scalar value to represent the linear predictor (logit) for each class.

Model estimation was performed using Python's *statsmodels* and *sklearn* libraries. The coefficients were interpreted as log odds, and their exponentiation yielded odds ratios, which quantified the effect of each predictor on the probability of selecting a given plan.

To validate the logistic regression results, we incorporated a random forest classification model, leveraging its ensemble learning capabilities to cross-check classification accuracy and assess potential improvements over logistic regression. The inclusion of random forest validation ensures that the result is robust, providing a comparative benchmark for evaluating the predictive performance of logistic regression in classifying health insurance coverage levels.

Introduction to the Dataset

The dataset contains 454,863 records with twelve features, including the predicted variable. The dataset also contains string and numerical data points. Features such as gender, region, smoker, medical history, etc are all categorical. These features are further explained in Table 1.

Table 1: Features and Description

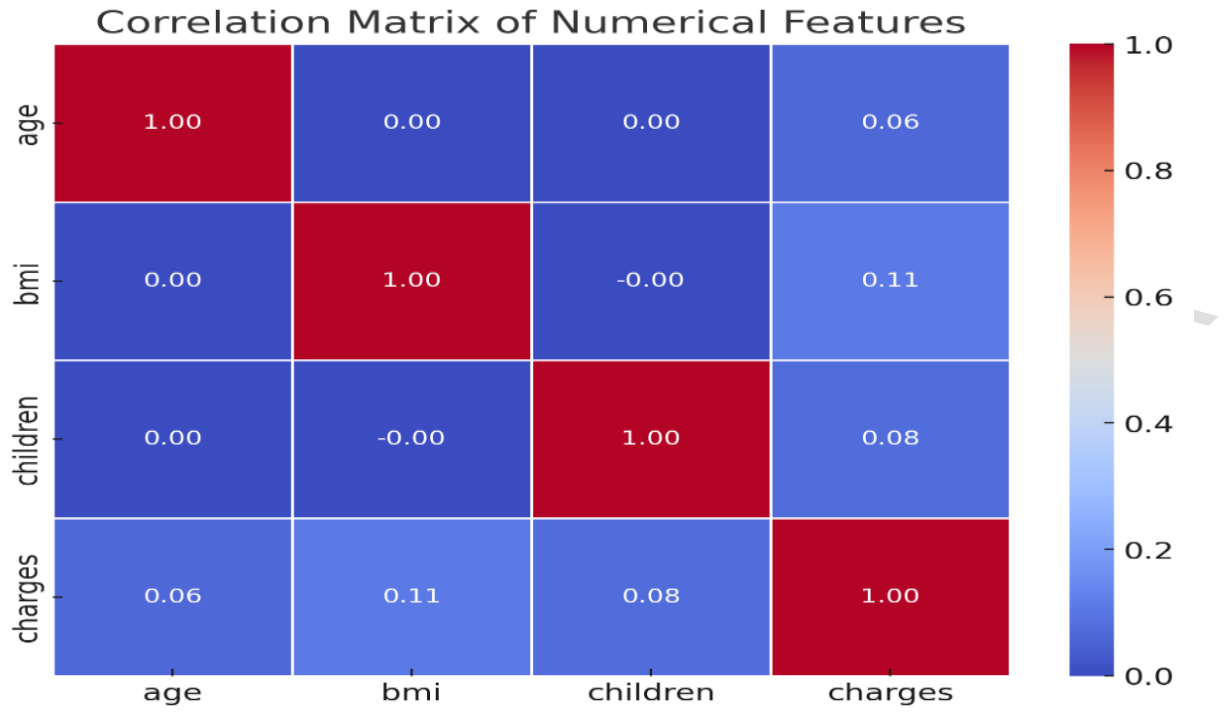
Features	**Description**
Age	Age of the insured individual
Gender	Gender of the individual (Male, Female)
Bmi	Body Mass Index (BMI) – measures body fat based on height & weight
Children	Number of dependent children covered under insurance
Smoker	Whether the individual smokes (Yes, No)
Region	Geographic region of the individual (Southeast, Northwest, etc.)
medical_history	Previous medical conditions (e.g., Diabetes, Hypertension, None)
family_medical_history	Family history of illnesses (High blood pressure, Diabetes, etc.)
exercise_frequency	How often the individual exercises (Never, Rarely, Occasionally, Frequently)
Occupation	Job type of the insured (Blue collar, White collar, Unemployed)
coverage_level	Type of insurance coverage (Basic, Standard, Premium)
Charges	Insurance cost

These features may influence the choice of insurance coverage taken by the individual insured.

Exploratory Data Analysis

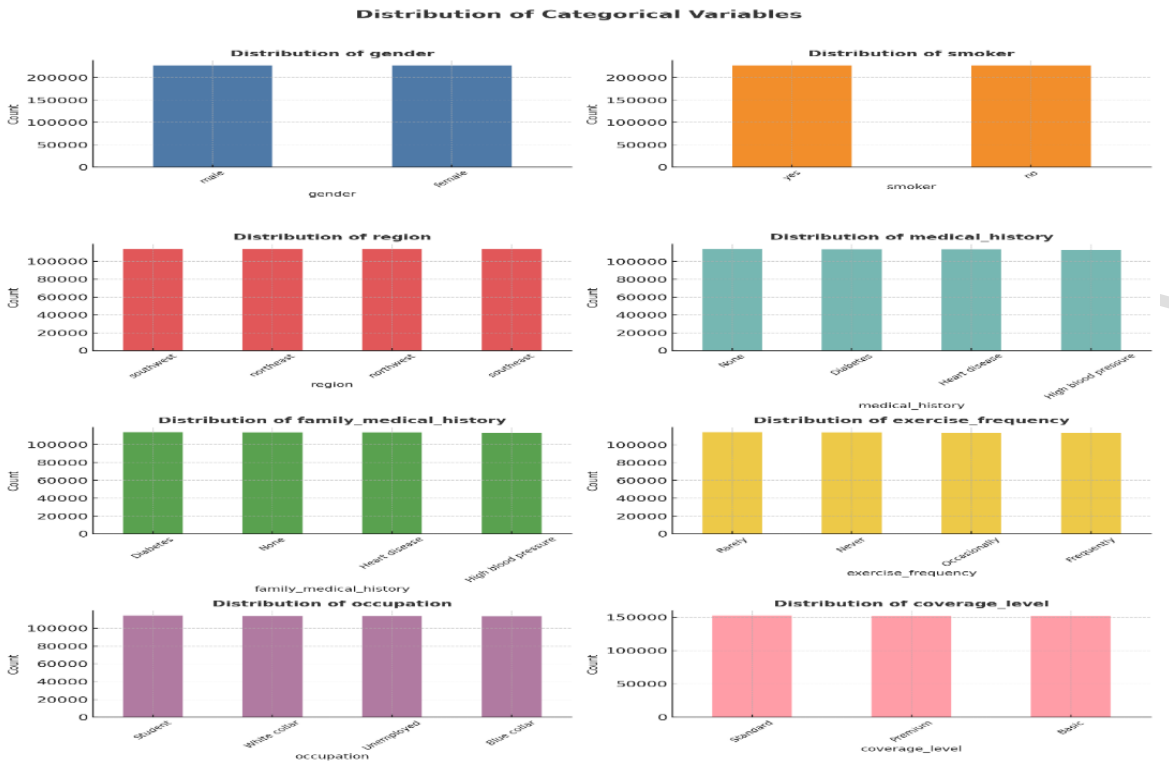
The dataset was quickly examined to identify any implicit patterns and anomalies within it. It was very prudent to check the relationships between some key features to identify their correlation (Bin Mahathir et al., 2025). This is shown by the Pearson correlation Heatmap in Figure 1 and the distribution of the categorical variables, as also shown in Figure 2.

Figure1:



The correlation matrix analysis reveals that the numerical variables (age, BMI, number of children, and charges) exhibit either weak or no significant correlation with one another. Age and BMI (0.00), age and number of children (0.00), and BMI and number of children (-0.00) show no relationship, indicating their independence within the dataset. The correlation between age and insurance charges (r = 0.06) and BMI and charges (r = 0.11) is weak, suggesting that these factors alone do not significantly influence insurance costs. Additionally, the correlation between the number of children and charges (0.08) suggests that having more dependents does not substantially increase premiums. The predictors are thus uncorrelated.

Figure 2:



In exploring the dataset, the researchers analyzed the distribution of categorical variables to grasp their potential influence on the level of insurance coverage predictions. The dataset presents a well-balanced representation across various categories, including gender, smoking status, region, medical history, family medical history, exercise frequency, and occupation, providing a solid foundation for predictive modeling. Key factors, including medical history, smoking status, and exercise frequency, are expected to be significant predictors since they affect health risk perceptions and insurance plan choices. Individuals with chronic conditions or a family history of health issues may prefer higher-tier plans, while those leading active lifestyles might opt for lower coverage options. Differences in occupation are also crucial, as job type and income levels affect insurance decisions. The balanced distribution of these elements reduces bias, enhancing the reliability of predictive analytics in examining insurance plan selection patterns. We followed (Bin Mahathir et al., 2025) all the categorical variables with encoding or one-hot encoding to make them usable for multi-class logistic regression analysis in Python.

Results

This study aims to understand what factors influence a person's decision when choosing between different health insurance plans—Basic, Standard, or Premium. The following section shares the key findings from the analysis.

Logistic Regression

The multi-class logistic regression model was employed to examine the relationship between individual characteristics and the likelihood of selecting among three levels of health insurance coverage: Basic, Standard, and Premium. Each coefficient in the model represents the change in the log-odds of selecting a particular insurance plan associated with a one-unit increase in the

predictor variable, holding all other variables constant. Positive coefficients indicate an increased likelihood of choosing the corresponding plan, while negative coefficients suggest a decreased likelihood. Multi-class logistic regression was performed using python and the results (log odds) are shown in table 2.

Table 2: Logistic Regression Coefficients for Each Feature and Coverage Level

Class	Basic	Premium	Standard
Age	0.602216	-0.705189	0.103055
BMI	0.989462	-1.16238	0.1792012
Children	0.742598	-0.868248	-0.130256
Charges	-9.643573	1.271737	0.628719
gender_male	1.091435	-1.275383	-0.011398
smoker_yes	5.453232	-6.376655	0.630859
region_northwest	-0.661678	0.771533	-0.012886
region_southeast	-0.469499	0.550141	-0.323588
region_southwest	-0.76568	-0.895935	-0.167727
medical_history Heart disease	3.786682	-4.415401	-0.246641
medical_history_High blood pressure	-0.003715	0.015114	-0.162728
family _medical_history Heart disease	3.785142	-4.416001	-0.242903
family_medical_history_High blood pressure	-0.005993	0.018879	0.075046
exercise_frequency_Never	-1.903519	2.227107	0.102973
exercise_frequency_Occasionally	-0.942253	1.10998	0.172918
exercise_frequency Rarely	-1.429307	1.675947	0.125651
occupation_Student	-0.953065	1.115793	-1.628164
occupation_Unemployed	-1.420966	1.663869	0.183948
Occupation: White collar	0.474178	-0.549224	0.923423

The age of the individual was found to have a positive influence on the selection of Basic and Standard plans, with coefficients of 0.6022 and 0.1030, respectively. In contrast, the coefficient for Premium coverage was −0.7052, indicating that younger individuals are more likely to opt for Premium plans, while older individuals may prefer more affordable options. Similarly, body mass index (BMI) exhibited a positive association with Basic coverage (0.9895), a modest positive relationship with Standard (0.1792), and a negative association with Premium (−1.1624). This suggests that individuals with higher BMIs may opt for lower-tier plans, potentially due to concerns about affordability or a perceived limited value in comprehensive coverage.

The number of children a person has also influenced insurance selection. A positive coefficient for Basic (0.7426) and Standard (0.1257) plans suggests that individuals with dependents tend to prefer lower- or mid-tier plans, while the negative coefficient for Premium (−0.8682) implies a reduced likelihood of selecting high-cost plans. Charges, a proxy for healthcare utilization and costs, had the most pronounced effect. The Basic plan showed a significantly negative coefficient (−9.6436), while the Premium (1.2717) and Standard (0.6287) plans had positive coefficients.

This indicates that individuals incurring higher healthcare expenses are more likely to select plans with greater coverage benefits.

Gender also played a role, with males more likely to choose Basic (1.0914) and Standard (0.1839) plans and less likely to choose Premium (−1.2754). This may reflect differing health-seeking behaviors or financial considerations between genders. Smoking status was one of the most influential predictors. The coefficients for smokers selecting Basic plans, Premium plans, and Standard plans were 5.4532, -6.3767, and 0.6309, respectively. This suggests that smokers are highly likely to opt for Basic coverage and strongly avoid Premium plans, possibly due to higher costs or limited access caused by health-related underwriting.

Regional differences were also evident in the plan choice. Living in the northwest or southeast regions reduced the likelihood of selecting Basic coverage (−0.6617 and −0.4695, respectively), but increased the odds for Premium plans (0.7715 and 0.5501, respectively). These differences may reflect regional variations in healthcare markets, insurance offerings, or socioeconomic conditions. Individuals with a personal history of heart disease were more likely to select Basic coverage (3.7867) and less likely to opt for Premium (−4.4154) or Standard (−0.2466). A similar pattern was observed for those with a family history of heart disease, who also showed a strong positive coefficient for Basic (3.7851) and negative associations with Premium (−4.4160) and Standard (−0.2429). These results, although counterintuitive, may indicate financial limitations among higher-risk individuals or a lack of awareness regarding the benefits of more comprehensive coverage.

Exercise frequency also revealed insightful trends. Individuals who never exercised were less likely to select Basic coverage (−1.9035) and more likely to opt for Premium (2.2271). Additionally, occasional and rare exercisers had higher likelihoods of selecting Premium (1.1999 and 1.6759, respectively). This may suggest that those who perceive themselves at greater health risk—due to lower physical activity gravitate toward higher-tier coverage. Conversely, those with healthier lifestyles might feel less need for expensive plans.

Occupation was another important determinant. Students and unemployed individuals had negative coefficients for both Premium and Standard plans, and positive associations with Basic, suggesting a preference for the most expensive option. For example, being unemployed was associated with −1.4210 for Basic and 1.6639 for Premium. Meanwhile, white-collar professionals were more likely to choose Standard coverage (0.9234), perhaps seeking a balance between affordability and benefit comprehensiveness. They also had a modest positive association with Basic (0.4742) and a negative one with Premium (−0.5492), indicating a general preference for mid-range or minimal plans.

In summary, the results highlight multidimensional factors influencing insurance plan selection. Financial capacity, as reflected in charges and occupation, along with health behaviors such as smoking and exercise, play a critical role in determining the choice of insurance coverage. Individuals with higher healthcare costs and risk indicators tend to favor Premium plans, while those with financial constraints or higher-risk lifestyles often settle for Basic plans. These findings provide important implications for insurers and policymakers aiming to align health plan offerings with population needs and promote equitable access to health coverage. These

results also suggest that policy interventions, such as cost subsidies or personalized premium structures, may be necessary to ensure that high-risk individuals can access appropriate insurance coverage.

Results from Machine Learning: Logistic Regression

The study also performs logistic regression using a machine learning approach to check the consistency of the results. The logistic regression metrics are shown in table 2

Table 3: Logistic Regression Metrics

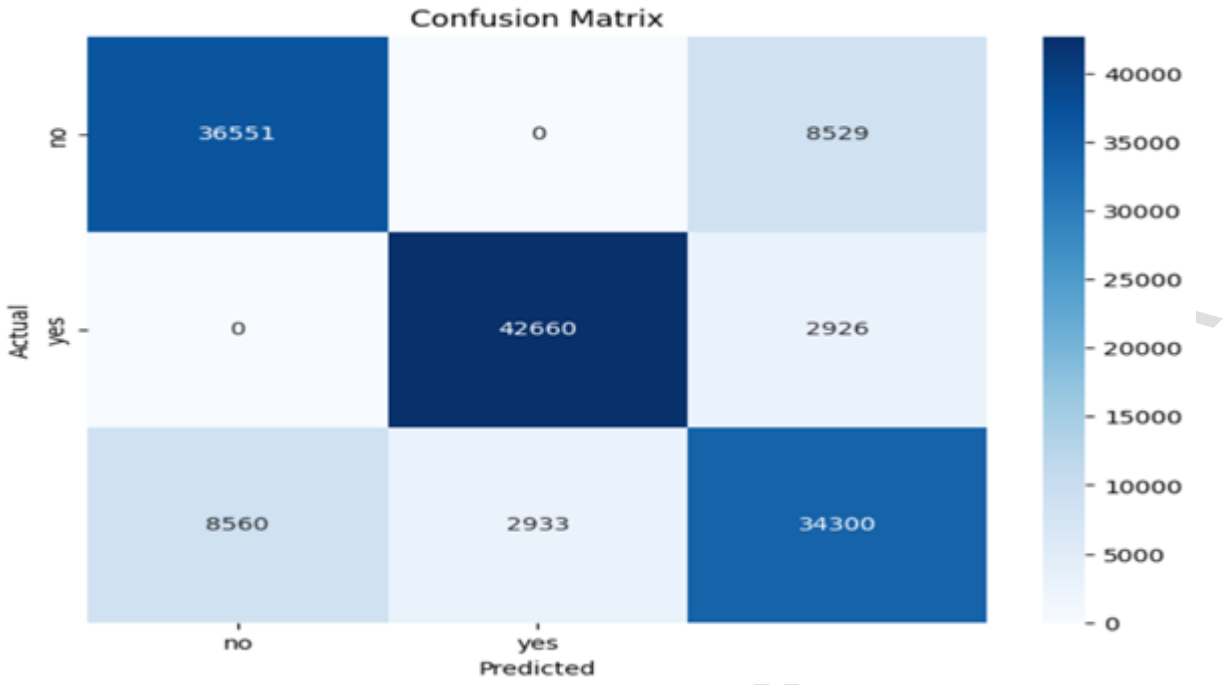
Class	Precision	Recall	F1-Score
Basic	0.81	0.81	0.81
Premium	0.94	0.94	0.94
Standard	0.75	0.75	0.75
Accuracy: 0.83			
Macro Avg	0.83	0.83	0.83
Weighted Avg.	0.83	0.83	0.83

Accuracy gives the percentage of classifications that were correctly made. A perfect model has an accuracy of 1 or 100%. From Table 2, the logistic regression model achieved an overall accuracy of 83%, demonstrating a strong ability to classify insurance coverage levels (Basic, Standard, and Premium). Relying on accuracy alone for a conclusive decision may be misleading. This is because it does not provide enough information to evaluate model performance. To address this, the classification report provides other key performance indicators, including precision, recall, and F1-score, to assess the model's effectiveness across different coverage categories. Precision measures the model's ability to correctly classify the level of coverage that we care most about in this study. The model exhibited high performance in predicting Premium coverage, with a precision and recall of 0.94, indicating that most Premium classifications were correct, and nearly all actual Premium cases were identified. Both basic and standard coverage also have precision scores of 81% and 75% respectively. In showing the percentages of true outcomes that were correctly classified as being true, basic and standard health insurance coverage levels scored 83% as recall. Premium and Basic categories also performed well, achieving an F1-score of 94% and 81%, suggesting a reliable classification of individuals opting for Premium and Basic coverages. However, the Standard category had the lowest F1-score 75%, indicating higher misclassification rates, possibly due to feature overlap with the Basic and Premium categories. The balanced class distribution (approximately 30,000 instances per category) ensures that the model's performance is not skewed by class imbalance. The macro and weighted average F1-score 83% confirm that the model maintains consistency across all categories. These findings highlight the predictive capability of logistic regression in insurance coverage classification.

Confusion Metrics for Logistic Regression

The confusion matrix provides a detailed evaluation of the logistic regression model's classification performance in predicting insurance coverage levels. Figure 2 shows the confusion metrics for the logistic results.

Figure 2: Confusion Metrics for Logistic Regression

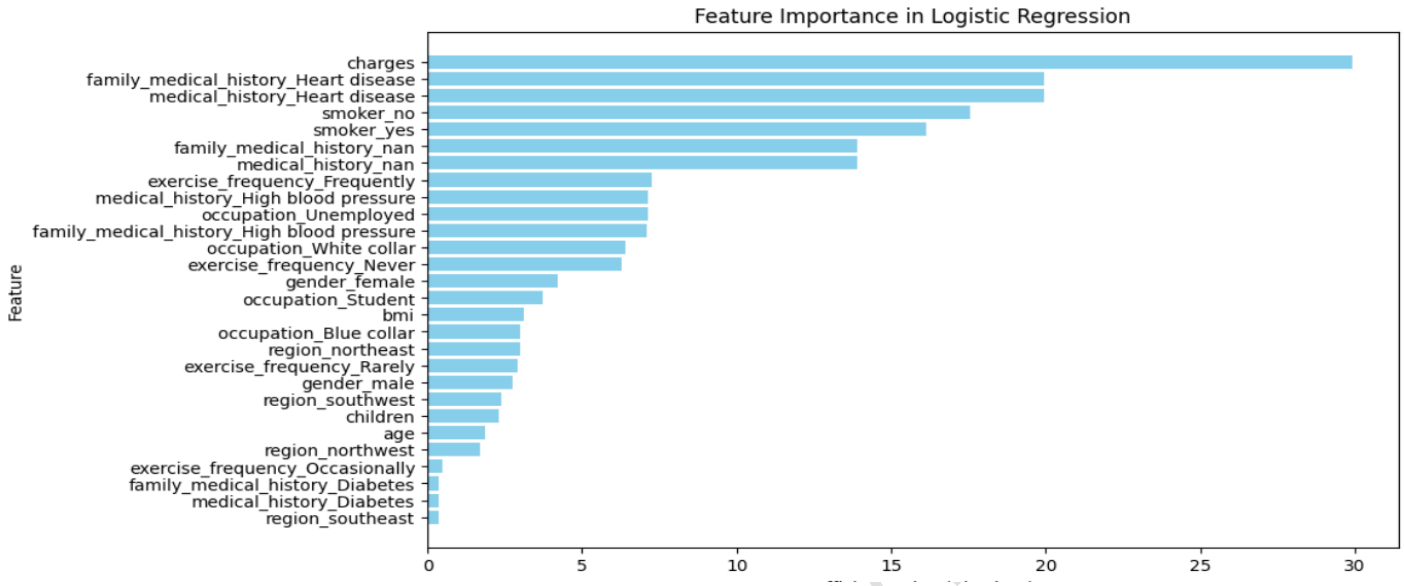


The results indicate that the model correctly classified most cases, with 36,551 instances accurately identified as "No" (Basic or Standard coverage), 42,660 instances correctly classified as "Yes" (Premium coverage), and 34,300 instances correctly predicted as "Yes" (Standard coverage). These values demonstrate the model's ability to distinguish between different insurance categories effectively. However, some misclassification patterns were observed. Specifically, 8,529 instances were incorrectly classified as Premium or Standard when they belonged to the Basic category, while 8,560 instances were misclassified as Basic or Standard when they should have been classified as Premium. These errors suggest that Standard coverage shares overlapping characteristics with both Basic and Premium plans, making it more difficult to differentiate. Additionally, the model exhibits zero false positives in the middle category, suggesting stronger predictive performance in classifying Premium coverage plans.

Feature Importance for Logistic Regression

The importance of each feature in predicting the level of coverage is shown in figure 3

Figure 3:



The feature importance analysis reveals that insurance charges (29.93) are the most influential determinant of coverage selection, highlighting the critical role of cost sensitivity in individuals' decision-making. Higher charges significantly decrease the likelihood of selecting Premium plans, reinforcing financial constraints as a primary factor in coverage choices. Medical history, particularly a personal (19.92) or family history (19.97) of heart disease, strongly influences insurance selection, as individuals with chronic cardiovascular conditions tend to opt for higher-tier plans to mitigate potential healthcare costs. Similarly, smoking status (17.57) plays a crucial role, with smokers showing a stronger preference for comprehensive coverage due to elevated health risks and increased medical expenses. While high blood pressure (7.12, personal; 7.38, family history) remains relevant, it has a lower impact than heart disease, suggesting that policyholders differentiate between chronic conditions based on perceived severity and long-term financial burden.

In addition to health-related factors, employment status and lifestyle choices also contribute to coverage selection. Those who engage in frequent exercise (7.11) tend to opt for lower-tier plans, possibly perceiving themselves as healthier and requiring fewer medical interventions. Occupational status further differentiates coverage preferences, with white-collar workers (6.93) more likely to select higher-tier insurance, while unemployed individuals (7.11) predominantly opt for Basic coverage, reflecting financial constraints. In contrast, demographic factors such as BMI (3.74), age (2.79), and number of children (2.38) show relatively lower predictive importance, indicating that coverage choices are primarily driven by health risks and financial capacity rather than standalone demographic attributes. Furthermore, regional differences (Southwest: 2.18, Northwest: 1.47, Southeast: 0.36) exhibit minimal impact on coverage selection, suggesting that geographic variations in healthcare costs and accessibility do not significantly influence insurance preferences. Surprisingly, diabetes (0.37, personal; 0.38, family history) has a low contribution, implying that its impact on insurance decisions is likely moderated by other factors such as pre-existing conditions and overall financial stability.

These findings emphasize that insurance selection is driven by a combination of financial constraints, health risk perception, and socioeconomic status. While cost remains the dominant

factor, individuals with severe chronic conditions, particularly heart disease and smoking-related risks, are more inclined to opt for higher-tier plans. Additionally, occupational status and lifestyle behaviors suggest that insurers could benefit from customizing policy structures to different socioeconomic segments.

Validation of the Results from Logistic Regression with Random Forest

The study follows (Yego et al., 2023) to adopt another classification model called random forest to validate the results from the multi-class logistic regression. The results from the random forest are shown below:

Table 4: Classification Report of Random Forest

Class	Precision	Recall	F1-Score
Basic	0.79	0.80	0.80
Premium	0.90	0.92	0.91
Standard	0.72	0.69	0.70
Accuracy: 0.80			
Macro Avg	0.80	0.80	0.80
Weighted Avg.	0.80	0.80	0.80

The classification report provides key performance indicators, including precision, recall, and F1-score, for evaluating the model's ability to classify insurance coverage levels (Basic, Standard, and Premium). These results serve as a validation benchmark for the logistic regression model, facilitating a comparative assessment of classification accuracy.

The overall accuracy of the model is 80%, which is slightly lower than the 83% accuracy observed in the logistic regression model. Similarly, the macro and weighted average F1-scores are 80%, reflecting balanced classification across all coverage categories but showing a marginal decrease compared to logistic regression (83%). Examining the class-specific F1-scores reveals that Premium coverage maintains high classification performance (F1 = 91%), slightly lower than the logistic regression model's 94%, suggesting that Premium policyholders exhibit distinct characteristics that the model effectively captures. In contrast, Standard coverage exhibits the lowest F1-score (70%) and recall (69%), indicating challenges in differentiating this class from Basic and Premium plans. This decline from 75% in logistic regression suggests that Standard Plan policyholders share overlapping characteristics with other groups, leading to increased misclassification rates. The classification performance for Basic coverage remains stable (F1 = 80%), showing a minor decline from logistic regression (81%), further affirming the consistency of model predictions in this category.

These findings suggest that while the model effectively classifies Premium policyholders, its performance in distinguishing Standard coverage remains a key limitation, mirroring the logistic regression model's challenges. The overall classification decline compared to logistic regression indicates that logistic regression remains a slightly stronger model for this dataset.

Discussion

The main objective of this study is to identify the predictive factors driving the preference of the level of health insurance coverage in the United States. It offers new empirical evidence on the

determinants of coverage level selection in private health insurance markets, highlighting how socio-economic, demographic, health-related, and behavioral factors shape consumer preferences among Basic, Standard, and Premium plans. The results underscore that insurance plan choice is influenced not solely by clinical need or actuarial risk but by a complex set of personal expectations, affordability constraints, and behavioral heuristics.

Age emerged as a significant factor, with a strong positive association with Basic plan selection (0.6022) and a significant negative coefficient for Premium (−0.7052). This indicates that older individuals tend to select lower-tier coverage, likely driven by affordability concerns or risk-averse behavior in the context of fixed incomes. This result is consistent with prior literature (e.g., Barkeret al., 2021.) suggesting that health expectations may not always align with comprehensive plan selection. Conversely, younger individuals showed a greater tendency toward Premium coverage, possibly due to employment-linked benefits or forward-looking risk perceptions.

BMI followed a similar pattern. Individuals with higher BMI levels were more likely to choose Basic plans (0.9895) and showed a significant negative association with Premium (−1.1624). This suggests that affordability or perceived discrimination may discourage individuals with higher health risks from selecting more comprehensive coverage, even when medically indicated, a pattern also observed by (Fang& Kung, 2021; Sun, 2020).

The number of dependent children significantly influenced plan choice. Individuals with more children were more likely to opt for Basic (0.7426) and Standard coverage (0.1257) and less likely to select Premium (−0.8682), aligning with the findings of (Marone & Sabety, 2022), who observed that family budgeting dynamics often lead to more conservative plan selection.

One of the most striking results was the role of insurance cost, proxied in the model by the charges variable. The coefficient for charges was strongly negative for Basic (−9.6436) and positive for both Premium (1.2717) and Standard plans (0.6287). This indicates that as insurance costs increase, individuals are more likely to opt for higher-tier coverage and less likely to select Basic coverage. This behavior may reflect a rational consumer assessment of value-for-money in Premium plans: those paying more expect or require more benefits. However, the steep negative coefficient for Basic suggests that individuals who face higher plan prices may either be priced out of low-value plans or redirected toward employer-sponsored Premium offerings. Unlike many prior studies that use premiums as exogenous determinants of enrollment, this analysis treats plan cost as an endogenous signal of coverage generosity, consistent with the economic framing in (Handel et al., 2020).

Gender and smoking status were also significant behavioral predictors. Males showed a strong preference for Basic plans (1.0914) and avoidance of Premium (−1.2754), consistent with findings from Lenhart (2019), who documented gender differences in health-seeking behavior and risk tolerance. Smokers, meanwhile, showed a highly pronounced preference for Basic plans (5.4532) and an equally strong aversion to Premium coverage (−6.3767). This suggests that smokers may avoid higher-cost plans due to perceived discrimination in underwriting or a belief that comprehensive coverage may not serve their needs. These patterns are echoed in (Hoagland et al., 2024) where socially marginalized health behaviors were correlated with underinsurance.

Regional variables also showed meaningful heterogeneity. Individuals in the Northwest and Southeast were less likely to choose Basic coverage (−0.6617, −0.4695) and more likely to opt for Premium plans (0.7715, 0.5501). This regional variation is in line with findings by (Holahanet al., 2024) who demonstrated how regional pricing and competition influence access to and preference for higher-tier insurance products.

Perhaps most concerning is the inverse relationship between medical history and plan comprehensiveness. Individuals with a personal or family history of heart disease were significantly more likely to choose Basic coverage (3.7866) and less likely to select Premium (-4.4154) or Standard (-0.2466). This suggests that even those with clear health risks may self-select into underinsurance, potentially due to affordability barriers or information asymmetries. Similar underinsurance behavior among high-risk populations has been documented by (Fang & Kung, 2021) and (Samek & Sydnor, 2020) raising critical concerns about the equity of vertical choice systems.

Exercise frequency also exhibited predictive power. Those who never exercised were less likely to choose Basic plans (−1.9035) and more likely to opt for Premium coverage (2.2271), possibly reflecting increased perceived vulnerability. Individuals who exercised occasionally or rarely also showed positive associations with Premium coverage. These results echo findings by (Barker et al., 2021.) who reported that self-rated health risk perceptions significantly influence coverage decisions.

Finally, occupational status emerged as a proxy for income and socioeconomic capacity. Students and unemployed individuals were significantly more likely to choose Basic coverage and avoid Premium plans, as evidenced by negative coefficients for Basic (−0.9531, −1.4210) and large positive coefficients for Premium (1.1158, 1.6639). White-collar professionals, in contrast, showed a preference for Standard plans (0.9234), suggesting a deliberate balancing of benefits and affordability. These findings support the arguments by (Lenhart, 2019) and (Samek & Sydnor, 2020) that the plan choice is strongly conditioned by income, employment, and benefit design.

Overall, the results of this study emphasize that health insurance plan selection is deeply shaped by behavioral and economic constraints. Contrary to the assumption that consumers act as perfectly informed, utility-maximizing agents, the evidence suggests that plan choice reflects a combination of perceived risk, financial burden, and systemic limitations. High-risk individuals may be under-insured not because they fail to recognize their needs, but because the cost of adequate coverage is beyond their reach, or the value proposition is unclear.

These insights have significant policy implications. Ensuring vertical choice in insurance markets must go beyond offering multiple plans—it must include adequate subsidies, transparent communication, personalized recommendation tools, and simplification of benefits to improve plan alignment. For insurers, the findings suggest that incorporating behavioral data and socio-demographic profiling into plan design and marketing strategies could improve product uptake and consumer satisfaction while minimizing risk segmentation.

By unpacking the behavioral dynamics behind tiered plan selection, this study contributes to a more comprehensive understanding of consumer behavior in private insurance markets. It moves beyond cost prediction to explore the motivations and constraints that influence how individuals choose the level of protection that best aligns with their perceived needs and financial realities.

Conclusion

This study investigated the determinants of health insurance coverage level selection—Basic, Standard, or Premium—within a private insurance context using both logistic regression and random forest classification models. The analysis revealed that consumer decisions are shaped by a multidimensional interplay of financial capacity, health risk perception, and socio-behavioral factors, with cost considerations emerging as the most salient driver of plan preference.

The logistic regression model demonstrated strong predictive performance (83% accuracy), particularly in classifying Premium policyholders (F1-score = 0.94), reinforcing the robustness of interpretable statistical models in insurance behavior prediction. Notably, insurance charges—serving as a proxy for premium cost—exerted the largest marginal effect on plan selection, significantly deterring uptake of higher-tier coverage. This underscores the centrality of affordability in shaping access to comprehensive protection, particularly among individuals facing economic constraints.

Health-related indicators also played a significant role. Smoking status and a history of heart disease were among the most influential predictors, supporting the hypothesis that perceived vulnerability prompts preference for richer coverage, albeit with some paradoxical evidence of underinsurance among high-risk individuals. Socioeconomic variables such as occupational status, exercise frequency, and region of residence also contributed meaningfully to the model, though with relatively lower weight compared to financial and clinical factors.

The random forest model, with an 80% overall accuracy, served as a robust validation tool, confirming model consistency while highlighting the relative difficulty in classifying Standard policyholders (F1-score = 0.70), who appear behaviorally and demographically intermediate between Basic and Premium enrollers. This finding points to potential ambiguity in mid-tier plan value perception and suggests an opportunity for insurers to clarify product differentiation in the market.

Taken together, the findings affirm that health insurance plan selection is far from a uniform or purely rational process. Rather, it reflects structural barriers, psychological heuristics, and economic realities that vary across population segments. For policymakers and insurers, this implies a critical need to enhance affordability, streamline coverage tiers, and design personalized, data-driven decision aids that help consumers select plans aligned with both their health needs and financial circumstances. Tailored subsidies, transparent pricing mechanisms, and simplified benefit designs may be particularly effective in mitigating underinsurance among vulnerable populations. Future research should further explore longitudinal shifts in plan preferences, behavioral responses to pricing changes, and the role of policy nudges in improving insurance match quality.

REFERENCES

Adjei-Mantey, K., & Horioka, C. Y. (2023). Determinants of health insurance enrollment and health expenditure in Ghana: an empirical analysis. *Review of Economics of the Household*, *21*(4), 1269–1288. https://doi.org/10.1007/s11150-022-09621-x

Barker, A. R., Maddox, K. E. J., Peters, E., Huang, K., & Politi, M. C. (2021). *Predicting Future Utilization Using Self-Reported Health and Health Conditions in a Longitudinal Cohort Study*. *58*, 1–9. https://doi.org/10.2307/27153299

Bhardwaj, N., Delhi, R. A., Akhilesh, I. D., & Gupta, D. (2020). *Health Insurance Amount Prediction*. https://economictimes.indiatimes.com/wealth/insure/what-you-need-to-

Bin Mahathir, A. A., Ee Shan, L., Bin Khairudin, A., Ting Xi, N., & Ul Amin, N. (2025). *Predictive Modelling of Healthcare Insurance Costs Using Machine Learning*. https://doi.org/10.20944/preprints202502.1873.v1

El Kassimi, M., El Badraoui, K., & Ouenniche, J. (2024). On the efficiency of U.S. community banks around the COVID-19 outbreak. *Applied Economics*. https://doi.org/10.1080/00036846.2024.2413421

Fang, H., & Kung, E. (2021). Why do life insurance policyholders lapse? The roles of income, health, and bequest motive shocks. *Journal of Risk and Insurance*, *88*(4), 937–970. https://doi.org/10.1111/jori.12332

Gupta, S., & Kanungo, R. P. (2022). Financial inclusion through digitalisation: Economic viability for the bottom of the pyramid (BOP) segment. *Journal of Business Research*, *148*, 262–276. https://doi.org/10.1016/j.jbusres.2022.04.070

Hanafy, M., & Mahmoud, O. M. A. (2021). Predicting Health Insurance Cost by using Machine Learning and DNN Regression Models. *International Journal of Innovative Technology and Exploring Engineering*, *10*(3), 137–143. https://doi.org/10.35940/ijitee.C8364.0110321

Handel, B. R., Kolstad, J. T., Minten, T., & Spinnewijn, J. (2020). *The Social Determinants of Choice Quality: Evidence from Health Insurance in the Netherlands*.

Hoagland, A., Yu, O., & Horný, M. (2024). Social Determinants of Health and Insurance Claim Denials for Preventive Care. *JAMA Network Open*, *7*(9), e2433316. https://doi.org/10.1001/jamanetworkopen.2024.33316

Holahan, J., Wengle, E., & Simpson, M. (2024). *Comparing Pricing and Competition in Small-Group Market and Individual Marketplaces*.

Hughes, P. (n.d.). *DETERMINANTS OF HEALTH CARE EXPENDITURE FOCUSING ON INSURANCE COVERAGE*.

Lenhart, O. (2019). *Pathways Between Minimum Wages and Health*. https://doi.org/10.2307/48730461

Marone, V. R., & Sabety, A. (2022). *American Economic Association When Should There Be Vertical Choice in Health Insurance Markets? 112*(1), 304–342. https://doi.org/10.2307/27105180

Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications*, *15*, 100516. https://doi.org/10.1016/j.mlwa.2023.100516

Research Project, M., & Sun, J. J. (2020). *Identification and Prediction of Factors Impact America Health Insurance Premium*.

Samek, A., & Sydnor, J. R. (2020). *NBER WORKING PAPER SERIES IMPACT OF CONSEQUENCE INFORMATION ON INSURANCE CHOICE*. http://www.nber.org/papers/w28003

Sukartini, T., Arifin, H., Kurniawati, Y., Pradipta, R. O., Nursalam, N., & Acob, J. R. U. (2021). Factors Associated with National Health Insurance Coverage in Indonesia. *F1000Research*, *10*, 563. https://doi.org/10.12688/f1000research.53672.1

Terlizzi, E. P., & Cohen, R. A. (2022). Geographic Variation in Health Insurance Coverage: United States, 2022. In *National Health Statistics Reports*. https://www.cdc.gov/nchs/products/index.htm.

Yamada, T., Yamada, T., Chen, C. C., & Zeng, W. (2014). Determinants of health insurance and hospitalization. *Cogent Economics and Finance*, *2*(1). https://doi.org/10.1080/23322039.2014.920271

Yang, S.-Y., Wang, C.-W., & Huang, H.-C. (2016). The Valuation of Lifetime Health Insurance Policies With Limited Coverage. *Source: The Journal of Risk and Insurance*, *83*(3), 777–800. https://doi.org/10.1111/jori.l2070

Yego, N. K. K., Nkurunziza, J., & Kasozi, J. (2023). Predicting health insurance uptake in Kenya using Random Forest: An analysis of socioeconomic and demographic factors. *PLoS ONE*, *18*(11 November). https://doi.org/10.1371/journal.pone.0294166