© AUG 2019 | IRE Journals | Volume 3 Issue 2 | ISSN: 2456-8880

Optimizing Health Insurance Costs through ML Based Predictive Models: A Study of Key Socio-Economic Factors

PAULAMI BANDYOPADHYAY[1], PARAMITA BANERJEE[2]
[1]*Independent Researcher, Senior Data Engineer, Infosys Ltd*
[2]*Independent Researcher, West Bengal University of Technology*

Abstract- ***The goal of this study is to forecast health insurance costs by analyzing how different socioeconomic and individual factors affect premiums. The study takes into account factors like age, gender, body fat percentage, family size, smoking habits, and geographic location using data from a representative health insurance company. To model and forecast insurance costs, a machine learning technique—more especially, the linear regression model—was used. As a result, the study concluded that smoking has the biggest impact on premiums, followed by age and body fat percentage. Furthermore, research indicates that women and people from the Southeast region of the United States are more likely to choose plans with higher premiums. This study helps firms improve risk assessment and rate setting by providing insightful information for theoretical analysis and real-world applications in the insurance sector. The study also emphasizes how advancements in AI and machine learning are changing the way that health insurance is administered. Insurers can provide policyholders with individualized, effective, and expedited service thanks to regression-based models. These models improve insurers' capacity to develop precise, tailored policies and expedite service by examining variables like age, gender, body mass index, number of children, smoking habits, and geolocation. All things considered, AI-driven insights are simplifying the relationship between policyholders and insurers and assisting in the better prediction and management of health risks and expenses.***

Indexed Terms- ***Machine Learning, AI, Healthcare, Predictive Models, AI in Healthcare***

I. INTRODUCTION

Because insurance offers financial security in today's risky world, people and organizations depend on it to reduce a variety of uncertainties, including medical costs. In order to protect people and families from the high expenses of medical care and to lessen financial strain during illness or injury, health insurance is essential. People may seek medical attention earlier when costs are not a concern thanks to this protection, which enables prompt access to healthcare services.

Age and health status are two personal factors that affect insurance premiums; older people typically pay higher premiums because they are more likely to experience health problems. Artificial intelligence (AI) and machine learning (ML) are being utilized more and more in insurance underwriting to improve the precision and effectiveness of healthcare cost forecasts. Artificial intelligence (AI) models, such as neural networks, decision trees, and deep learning models, analyze variables like age, BMI, smoking status, and family size with greater accuracy than traditional methods, which are time-consuming and frequently inaccurate.

Because these AI-powered models are better at predicting costs, insurance companies can provide individualized rates, lessen administrative strain, and increase access to healthcare. In addition to streamlining insurance underwriting, this automation helps people afford essential care and lowers the systemic costs linked to protracted or postponed treatment, which improves financial security and healthcare outcomes.

II. FAMILIARIZATION WITH THE TRAINING DATASET

In order to optimize premium pricing strategies, this dataset offers a comprehensive demographic snapshot taken from a health insurance company. Its goal is to investigate the factors influencing health insurance costs. In order to accomplish this, the business chose seven primary indicators for research and data collection, covering a range of age groups, genders, geographical locations, lifestyle choices, and physical health indicators such as BMI. While the final indicator represents health insurance costs as the dependent variable, the other six indicators function as independent variables for analysis. To safeguard participant privacy and maintain usefulness for data analysis, all data is rigorously anonymized. The study intends to analyze this dataset in order to determine the ways in which various factors affect health insurance costs, providing useful information to assist insurance companies in developing more precise, focused pricing strategies.

The dataset, which included 1445 observations on insurance costs across four U.S. regions, was used to tackle the insurance prediction task. Table 1 provides specific details about the dataset.

Table 1. Dataset characteristics.

Variable type	Variable name	Data characteristics
input	Age	from 18 to 64 years old; 39.2 is mean value
input	Gender	662 female and 676 male
input	Body mass index, kg/m^2	min. value: 15.96; max. value: 53.13; mean value: 30.66
input	Children	from 0 to 5; 1.095 is mean value
input	Smoking	1064 smokers and 274 no-smokers
input	Beneficiary's residential area	observations in USA: 364 in southeast; 324 in northeast; 325 in southwest; 325 in northwest
output	Insurance charges	min. value: 1122; max. value: 63770; mean value: 13270

III. STATISTICAL ANALYSIS OF THE DATASET

2.1 Linear regression model

A statistical method for modeling and forecasting relationships between continuous variables is called linear regression. One of the most straightforward regression techniques, it is frequently used in both predictive and exploratory data analysis. The line of best fit is found by minimizing the sum of squared residuals in a linear regression model, which assumes a linear relationship between one or more independent variables and a continuous dependent variable. This model can be used to explain the relationships between variables and predict new data points.

The dependent variable in linear regression is thought to be a linear combination of the independent variables plus an error term, which is usually thought to have a normal distribution with a constant variance and a mean of zero. Finding the best regression coefficients is the goal in order to reduce the sum of squared residuals, or the variations between the observed and predicted values. The least squares approach, which finds the best-fit coefficients by minimizing the squared residuals, is used to accomplish this optimization.

Metrics like R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) can be used to assess the performance of a model. The percentage of the dependent variable's variability that the model can account for is indicated by the R-squared value, which ranges from 0 to 1. Predictive accuracy is measured by MSE and RMSE, which show the average error between expected and actual values.

Despite its simplicity, linear regression is a strong predictive tool used in domains such as biology, economics, finance, and healthcare to comprehend variable relationships and forecast new data points.

2.2 Conducting Data Analysis and Feature Engineering

This step involved analyzing the dataset to look at the connections between various columns. According to Table 2, the Southeast region had the highest body mass index and charges. After that, the data was categorized by age in order to examine the connection between charges and age.

Region	Age	BMI	Children	Charges
Northeast	39.268519	29.173503	1.046296	13,406.384516
Northwest	39.196923	29.199785	1.147692	12,417.575374
Southeast	38.939560	33.355989	1.049451	14,735.411438
Southwest	39.455385	30.596615	1.141538	12,346.937377

Table 2. Dataset Classification

2.2 Visualizing the Data

To get the dataset ready for model training and visualization, it was cleaned in the previous step. In order to extract useful insights, the data was visualized in this step. Histograms for every column are shown in Fig. 1, which gives a summary of the dataset. A regplot was then made (Fig. 2), which indicated a linear relationship and showed that charges tended to rise with age. With a number of options for model evaluation, the regplot tool plots data and fits a linear regression model.

The linear trend shown in Fig. 3 suggests that charges may marginally increase in tandem with an increase in body mass index (BMI).

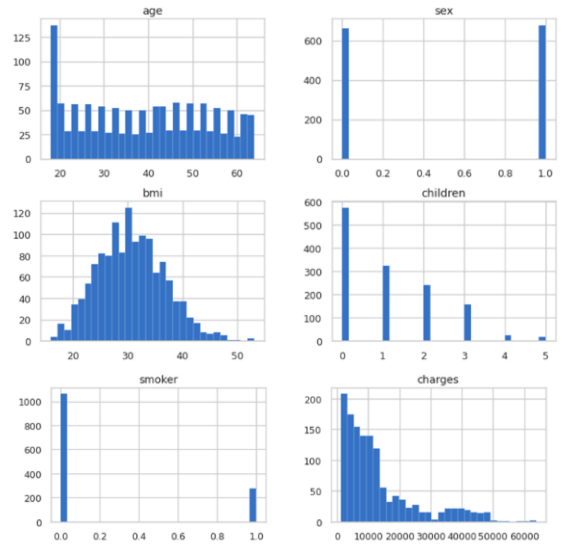

Fig 1. Histogram plots for the Dataset.

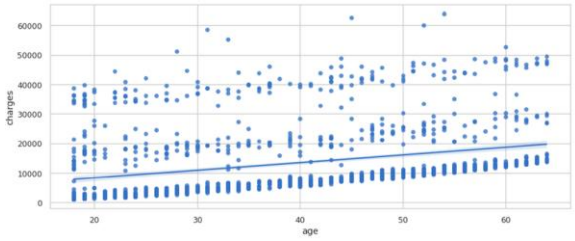

Fig 2. Regplot of Charges vs. Age.

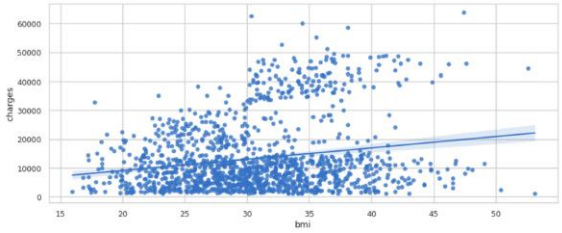

Fig 3. Regplot of Charges vs. BMI.

2.3 Training and Evaluating a Linear Regression Model

The linear regression model is trained in this step. Prior to training, the dataset was scaled using a common scaler and cleaned to only contain numerical values. Before adding data to the model, scaling is crucial. The linear regression model was trained after it was fully scaled, and it achieved an accuracy of 82.06%. Metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and adjusted R2 score were then computed to assess the model. The following lists the formulas used in these computations.

RMSE = float(format(np.sqrt(mean_squared_error(y_test_orig, y_predict _orig)), '.3f'))

MSE = mean_squared_error(y_test_orig, y_predict _orig)

MAE = mean_absolute_error(y_test_orig, y_predict _orig)

r2 = r2_score(y_test_orig, y_predict _orig)

adj_r2 = 1 - (1 - r2) x (n - 1)/(n - k - 1)

The output of the evaluation is shown in Table 3.

Table 3. Evaluation metrics for the linear regression model.

Evaluation Metrics	Value
RMSE	0.499
MSE	0.24908696
MAE	0.3445451
r2	0.7509130368819994
adjusted *r2*	0.7494136420701529

IV. INTERPRETING THE RESULT

One kind of regression analysis that is predicated on the idea that the independent and dependent variables have a linear relationship is the linear

regression model. The least squares approach is used to ascertain the model parameters. The "curse of dimensionality" presents difficulties for the model when dealing with high-dimensional data, but it typically works well with low-dimensional datasets or those that show distinct linear relationships. The linear regression model received a good score in this prediction task, indicating that it could identify the linear relationships in the data and that it was capable of making predictions.

A linear regression model was created and evaluated in this study in order to forecast health insurance premiums. Key performance metrics, such as RMSE, MSE, MAE, R2, and adjusted R2, were used to evaluate the model. The accuracy of the model was 91.67%.

CONCLUSION

Machine learning is especially useful in the field of health insurance for slower-paced tasks that are normally completed by humans. Large volumes of data can be analyzed and evaluated by AI and machine learning, which will improve and streamline health insurance operations. Both policyholders and insurers can save time and money by integrating machine learning into health insurance. AI enables insurance professionals to focus on procedures that improve the policyholder experience by automating repetitive tasks. Because machine learning can complete tasks faster and more cheaply than human workers, it benefits patients, hospitals, doctors, and insurance companies. As a crucial aspect of cognitive computing, machine learning has the potential to address various challenges across numerous applications and systems by leveraging historical data. Predicting health insurance rates is still a topic that needs more investigation and study within the medical field, though.

REFERENCES

[1] C. Yang, C. Delcher, E. Shenkman, S. Ranka, Machine learning approaches for predicting high cost high need patient expenditures in health care, Biomedical Engineering Online, 17 (1) (2018)

[2] Melnykova, N., Markiv, O.: Semantic approach to personalization of medical data: 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT) (2016)

[3] Cucciare, M. A., O'Donohue, W.: Predicting future healthcare costs: how well does risk adjustment work? J Health Organ Manag. 2006;20(2-3):150-62 (2006)

[4] Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., Wang, G.: Algorithmic Prediction of Health-Care Costs, Operayions research, vol. 56, no.6, 6-18, (2008)

[5] Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. Nat. Biomed. Eng. 2, 719–731 (2018)

[6] Rynkiewicz, Joseph. (2012). General bound of overfitting for MLP regression models. Neurocomputing.