**Predictive Power-A Case Study on Regression Techniques**

Sai Kalyan Arutla

Student  
*Department of Computer Science and Engineering*, *Koneru Lakshmaiah Education Foundation,*

*Hyderabad-500075, Telangana, India.*

[saikalyan0126@gmail.comn](mailto:saikalyan0126@gmail.comn)

Koushik Reddy Katta

Student *Department of Computer Science and Engineering   
Koneru Lakshmaiah Education Foundation,*

*Hyderabad-500075, Telangana, India.*

[Koushikreddy.katta@gmail.com](mailto:Koushikreddy.katta@gmail.com)

Lalith Siddartha Macherla

Student  
*Department of Computer Science and Engineering   
Koneru Lakshmaiah Education Foundation,*

*Hyderabad-500075, Telangana, India.*

[*laliths.m9999@gmail.com*](mailto:laliths.m9999@gmail.com)

Dr. Sai Sudha Gadde

Assistant Professor *Department of Computer Science and Engineering   
Koneru Lakshmaiah Education Foundation,*

*Hyderabad-500075, Telangana, India.*saisudhagadde@gmail.com

***Abstract*- Clinical decision-making in healthcare is changing a lot. Predictions made by smart machines are becoming more common. Many tools that use machine learning have shown up in recent medical studies. These tools help predict different outcomes, like who might have serious health issues, including things like heart attacks or problems with the kidney .In this review, we take a closer look at the newest findings in this area that is Insurance. We’ll talk about how data is, how conclusions are made, and how models are checked for accuracy. All of this is important when we think about predicting outcomes based on information from previous Insurance records .But it’s not all perfect! We also point out some limits of the main assumptions in these models. Plus, there are plenty of chances for research to grow in the future. There’s so much to explore!**

*\*Keywords*—Predictive modelling, machine learning, insurance premiums, linear regression, healthcare, Insurance, Supervised Machine Learning, Regression Techniques .\*

# Introduction

In today's data-driven world, the insurance sector stands out as a major beneficiary of predictive modelling. Accurate insurance cost forecasting is essential for insurers and policyholders, enabling improved pricing strategies, risk evaluation, and customer satisfaction. Historically, calculating insurance premiums depended heavily on static statistical models. However, the rise of machine learning has led to the creation of more flexible and responsive models.

One of the most basic machine learning algorithms, linear regression, is commonly utilized to predict continuous variables. Its straightforward nature, ease of interpretation, and efficiency make it an excellent choice for modeling the connection between various customer characteristics and insurance premiums. This study aims to use linear regression to forecast insurance premiums based on significant features such as age, income, medical history, vehicle information, and other pertinent factors.

A screenshot of a graph

Description automatically generated

Table head of dataset

The goal of this research is to assess how well linear regression performs in predicting insurance costs. By conducting detailed analysis and experiments on real-world datasets, we strive to shed light on the model's effectiveness, the impact of feature selection, and the significance of data preparation. Furthermore, this paper examines the difficulties encountered when dealing with multicollinearity, outliers, and missing data, along with strategies to enhance prediction accuracy.

By exploring the predictive power of linear regression within the insurance sector, this research seeks to support more precise, data-centric processes for determining premiums, ultimately boosting the overall efficiency of the insurance industry.

# CONTEXT AND FRAMEWORK OF PREDICTIVE MODEL

First, In the insurance industry, accurately predicting premium costs based on personal attributes is vital for ensuring fair pricing and effective risk management. Just as clinical pathways in hospitals require an understanding of patient-specific data to predict outcomes, insurance premium prediction models rely heavily on key personal and behavioural factors. These factors include age, body mass index (BMI), smoking habits, number of children, and geographic region, which significantly influence the determination of premiums.

***Y = B0 + B1X***

**Equation-1**

* B0 is a Constant
* B1 is Regression Coefficient

As in healthcare, where patient monitoring systems continuously collect data to assess clinical risk, insurance companies leverage these factors to evaluate the likelihood of claims and associated costs. Age and BMI, for instance, are strongly correlated with health risks, while smoking habits further increase the probability of health complications. Each of these factors contributes to a model designed to predict insurance premiums, where higher risks are linked to higher premium costs[6].

For an insurance premium prediction project, the key variables involved are:

* Age: Numeric (Continuous)
* Gender: Categorical (Binary)
* Smoking Habit: Categorical (Binary)
* Children: Numeric (Discrete)
* BMI (Body Mass Index): Numeric (Continuous)
* Region: Categorical (Multiclass)
* Insurance Premium: Numeric (Target Variable)

# Exploratory data analysis

A graph with a bar graph

Description automatically generated with medium confidence

Fig- distribution of body mass index

From fig [1] we can extract that body mass index is distributed as a triangle. We can bring that mostly middle-aged people have higher body mass index. Like hospital-based models predicting patient deterioration, premium prediction models group individuals into risk categories based on their attributes. For example, smokers or individuals with higher BMI may be classified into higher-risk groups, similar to how patients are categorized based on the risk of clinical deterioration. [1] By inputting these features into a linear regression model, the risk is quantified in the form of a predicted insurance premium. Average age: 39.3 years, Median BMI: 30.1, Smoking proportion: 27% smokers, 73% non-smokers. As you can see fig [1] the BMI varies in various values.

A graph of a graph showing the amount of charge

Description automatically generated with medium confidence

Fig Distribution of Charges among smokers

Let’s talk about how insurance costs relate to whether someone smokes or not Look at the histogram! The bar graph fig[2] clearly shows the Distribution of charges among smokers is on higher side than non-smoker side. It shows how people fall into different premium groups. The red bars are for non-smokers, while the blue ones represent smokers. You can clearly see that most non-smokers are in the lower premium group, with many paying less than $10,000. On the flip side, smokers are more spread out. A lot of them pay between $20,000 and $40,000 or even more.

Now, there’s also a box plot here which gives us more details about the insurance charges. Non-smokers (the red box plot) have a tight bunch of premiums—most of them sit under $10,000 with just a few outliers. Smokers (the blue box plot), however, a much wider range. That means there's more variation in what they pay! Their median is higher too, with some paying up to $60,000! It really shows how smoking can affect insurance prices. Smokers end up spending more on average than those who don’t smoke.

A graph of different colored squares

Description automatically generated

Fig- Distribution of Smokers among Genders

In Fig[3],The count of smokers and nonsmokers among different genders were equally distributed. But non-smoker count is pretty high in both Male and Female. Let’s understand the chart with an example ,We have 100 males and 100 females then the results shows 70 of males and 70 of females have no smoking habit then 30 of males and 30 of females don’t smoke.

hospitals are getting better at watching patients in real-time and focusing on specific like ICU admissions. Meanwhile, insurance models are working hard [4]too! They refine how they assess risk by looking at past data and using smart techniques to predict what might happen next. The goal? To help insurers avoid losing money & keep prices fair for everyone who holds a policy. These models learn all the time from new data—kind of [3]like how medical systems improve by keeping an eye on patient results!

This approach ensures that premium predictions are both data-driven and contextually grounded, providing insurers with a more precise understanding of risk factors, while contributing to more personalized and equitable insurance offerings.

# Statistics

The statistical summary gives a friendly peek into four main things: age, kids, & insurance costs for total of 1,338 people., the average age is around 39.21 years. People in this group range from folks of 18 wise ones of 64.Next, we have BMI — [2]it averages at 30.66. This tells us many folks might be on the heavier side or even considered obese.

A table with numbers and numbers

Description automatically generated

Table Statistical Summary of table

The table gives a quick look at a dataset with 1,338 entries. This data covers four things: age, Body Mass Index (or BMI), number of kids, & charges.

First off, the average age of the people in this dataset is about 39.21 years. There's a standard deviation of 14.05, which means the ages spread out a bit—from 18 to 64 years old. Then we have the average BMI, which is around 30.66, with a standard deviation of 6.09. The BMI values go from 15.96 up to 53.13.

When it comes to kids, people in this group have about 1.09 children on average! The variation here shows a standard deviation of 1.21, meaning some folks have anywhere between 0 & 5 children.

Now, let’s talk about charges—this likely means health insurance costs. The mean charges are approximately $13,270.42! There’s quite a range here too—from $1,121.87 all the way to $63,770.43. Its standard deviation is $12,110.01.

Looking at the quartiles can tell us even more! For example, 25% of the data shows that ages are below 27 years old, BMI is at or below 26.30, most have no kids (0), and charges sit around $4,740.29.

The median values show half of the data below age 39 years with a BMI of 30.40, still mostly no children (0), & average charges of $9,382.03.

Lastly, the top quartile tells us that 75% of people are below age 51 years with a BMI of 3 4.69 and they usually have around 2 kids! Charges in this group are about $16,639.91..

### Correlation

In machine learning, we talk about correlation a lot. It's about how two relate to each other. When say there's a positive correlation, it means that if one thing goes up, the other one usually does too.

First, for **Feature Selection**: If two features are really similar (or correlated), having both might not be needed. You could get rid of one! This can make the model easier without messing things up too much.

A screenshot of a graph

Description automatically generated

Table Correlation table

The image shows a correlation matrix. It looks at four things: age, Body Mass Index (BMI), number of kids, & charges. Each number in the matrix tells us how two variables relate to each other. The values go from -1 to 1.

You’ll notice that the numbers on the diagonal are all 1. That’s because they show how a variable compares to itself. Now, if we talk about age, it has a good connection with charges (0.299). This means older people might have higher charges. But when it comes to age and BMI, the connection is weak (0.109). Age and the number of children? That’s even weaker (0.042).

Now, let’s look at BMI. It has a moderate link with charges (0.198) and a tiny bit with children (0.013). The number of kids, on the other hand, has a little positive relation to charges (0.068).

Next is **Model Interpretation**: Correlation gives us a peek into how different inputs and outputs connect. It helps us see which features could be more important in making predictions.

r = n ∑ X Y − ∑ X ∑ Y ( n ∑ X 2 − ( ∑ X ) 2 ) ⋅ ( n ∑ Y 2 − ( ∑ Y ) 2 )

**Equation-2**

# Linear regression

Linear regression is a basic machine learning method. It helps predict continuous numbers based on one more input features. In simple terms, it looks at how dependent variable (the thing we're trying to predict)[1] relates to one or more independent variables (the inputs). It does this by fitting a straight line to the data we see[6].

Now, when we talk about multiple linear regression, things get a bit bigger. The equation takes into account multiple independent variables, too—each with its own special number called a coefficient[4].



**Equation-3**

This method thinks there’s a straight relationship between the dependent and independent variables. So, if you change the inputs, you should see proportional changes in the output. Also, it assumes that any errors that come up follow a normal pattern. They should have constant variance and be independent from one another.

In multiple linear regression, the equation gets a bit more complicated. It includes independent variables, each with its own special called a coefficient.

This type of regression thinks that there’s a straight-line relationship between what you’re measuring (the dependent variable) and the things that affect it (the independent variables). So, when one of these input variables changes, the output does too. They change in a way matches up nicely. Also, it believes that errors (or mistakes) are spread out normally, have a steady range (that’s what we call homoscedasticity), & act independently.

People use this method for all sorts of things like guessing how much houses will cost, figuring out insurance rates, and looking at other numbers that can be measured. Even though it seems simple, linear regression can work very well in lots of real-life situations. It shines when the links between the variables are straight or almost straight!

A graph showing the average value of a number of individuals

Description automatically generated with medium confidence

Fig- Scatter plot

A regression line on a scatter plot is a straight line that best fits the data points and shows the relationship between two variables. The regression line is also called the "line of best fit”.As we can see in fig4 there is a straight line with a slope of y = mx+c which represents LinearRegression best fit line through Scatter points.

A screenshot of a table

Description automatically generated

Table Actual vs Predicted values

## MODEL EVALUTION – METRICS

In linear regression, there are many ways to check how well a model works. important measure is R-squared (²). This number shows how much of the change in one (the dependent variable) can be explained by other things (the independent variables). It goes from 0 to 1. If R² is close to 1, it means the model explains a lot. If it’s closer to 0, there’s not much explanation happening.

**R2 = 1 − sum of squared regression**

Equation-4Then there's Adjusted R-squared. This one tweaks the R-squared number a bit by looking at how many predictors are in the model. It helps when you want to compare models that use different numbers of predictors. Sometimes, if you add more predictors but they don’t help much, Adjusted R-squared can even go down!

**Adjusted R2 = 1- [(n – 1)(1 – R2)/ (n – k – 1)]**

Equation-5

Another important metric is Mean Absolute Error (MAE). It looks at the average of how far off the predictions are from the real values. This measurement uses the same units as what you're trying to predict, which makes it easy to understand. The lower the MAE, the better your model is doing!

**MAE = (1/n) Σi=1n |yi – ypredi|**

Equation-6

On its side, we have Mean Squared Error (MSE). This one averages the squared differences between what was predicted and what actually happened. It gives bigger mistakes more weight than MAE does. Again, lower MSE means a more accurate model!

**MSE = (1/n) Σi=1n (yi – ypredi)2**

Equation-7

Root Mean Squared Error (RMSE) takes it a step further by finding the square root of MSE. Like MAE, it gives error in the same units as your target variable, too! It also puts a bit more focus on those big errors compared to MAE—so, again, lower RMSE means a better-performing model.

The rmse loss of this model is 6041.which means the model needs quiet of bit improvement where we are going to use a few methods for model improvement.

# Linear regression

Feature scaling and one-hot encoding are crucial preprocessing steps in preparing data for machine learning models, especially in algorithms that are sensitive to the scale and format of input data.

**Feature scaling** is a technique used to normalize or standardize the range of independent variables or features of data. It’s essential because many machine learning algorithms perform better or converge faster when features are on a similar scale. Two common methods for feature scaling are

**One-hot encoding** is a technique used to convert categorical variables into a numerical format that can be used by machine learning algorithms. Categorical variables are those that represent distinct categories or classes.

A screenshot of a cell phone

Description automatically generated

Table Distribution of weights

.

**In Table-5 different features have different weights in which Smoker code have the highest priority which is the key factor for predicting Insurance premium. Gender code have the Least preference.**

A screenshot of a computer

Description automatically generated

Table one hot encoding

Here, we standardize the weights based on how important they are. From our visuals, we see that age is a factor along with whether someone smokes. But here's the tricky part— in this model, age gets a smaller weight than BMI. That makes things confusing!

Well, it's because a column has a range of numbers, it can really mess the loss and take over the whole optimization process. Age ranges from 18 to 60, while smoking and gender only have two options: 0 or 1. So, to clear up this confusion, we go ahead & use standardization.

After the improvement the rmse error rate is 4662 which is better than previous value which means there is definite improvement in the model with the methods of one hot encoding and Feature Scaling..

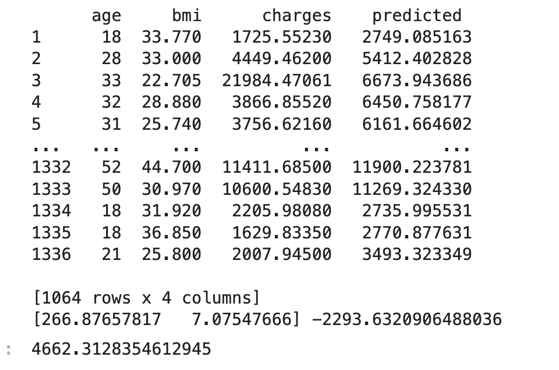


Table Actual vs Predicted

# Conclusion

This Case Study provides potential Understand of Linear Regression on Customer real time data. I Conclude that Linear Regression is a basic Supervised learning model which Gives us Results on linear Calculations. To improve model to more extend we can refer Algorithms like Decision Tree, Random Forest, XGBoost Algorithms and These Algorithms Are very complex than Linear Regression. Accurate Realtime data and better statistics could bring better improvement in the model. These kinds of Models are so helpful for both policyholders and insurers.

##### IV References

1. Bhardwaj, N.; Anand, R. Health Insurance Amount Prediction. *Int. J. Eng. Res.* **2020**, *9*, 1008–1011. [[**Google Scholar**](https://scholar.google.com/scholar_lookup?title=Health+Insurance+Amount+Prediction&author=Bhardwaj,+N.&author=Anand,+R.&publication_year=2020&journal=Int.+J.+Eng.+Res.&volume=9&pages=1008%E2%80%931011&doi=10.17577/IJERTV9IS050700)] [**[CrossRef](https://doi.org/10.17577/IJERTV9IS050700" \t "_blank)**]
2. Lui, E. Employer Health Insurance Premium Prediction. Available online: [**http://cs229.stanford.edu/proj2012/Lui-EmployerHealthInsurancePremiumPrediction.pdf**](http://cs229.stanford.edu/proj2012/Lui-EmployerHealthInsurancePremiumPrediction.pdf) (accessed on 17 May 2022).
3. F. Petropoulos, S. Makridakis, “Forecasting the novel coronavirus COVID-19”, Available: https://doi.org/10.1371/journal.pone.0231236, Published: March 31
4. W.C.Roda, M. B.Varughese, D. Han, M. Y. Lia, “Why is it difficult to accurately predict the COVID-19 epidemic?”, Infectious Disease Modelling, vol. 5, Available: https://doi.org/10.1016/j.idm.2020.03.001, pp. 271-281, (2020)
5. H. Berrer and I. e. a. Paterson, "Evaluation of machine-learning algorithm ranking advisors", Proceedings of the PKDD-2000 Workshop on DataMining Decision Support Meta-Learning and ILP, 2000.
6. F. Pedregosa and G. e. a. Varoquaux, "Scikit-learn: Machine learning in Python", vol. 12, pp. 2825-2830, 2011.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.