Health Analytics

Linear Regression

Name: Kwanda Mazibuko

Student number: 1077167

Faculty: Science

Table of Contents

[Introduction 3](#_Toc192685869)

[Methods 4](#_Toc192685870)

[Results 6](#_Toc192685871)

[Discussion 6](#_Toc192685872)

[Conclusion 6](#_Toc192685873)

[References 7](#_Toc192685874)

ChatGPT was used in building the model, to find the correct metric to use to measure the performance as well feature scaling. Including structuring the references to conform to the report requirements.

## Introduction

Insurance pricing is a not a straightforward process. It is often founded on client risk assessment, actuarial science, and data analytics. As it is a business on its own, as an insurer you want to ensure that the premiums sent are reasonable such that the business is able to pay for claims, administrative costs, and still make profit. Simultaneously, they must be low enough to remain competitive and attractive to customers.

Insurers have traditionally relied on risk-based pricing methods, where individuals are segmented into various groups like age, health status or lifestyle habits. Those that seem to be on a higher risk are charged higher amount. A real-life example is that individuals who smoke pay more for health or life insurance due to their likelihood of having chronic diseases or dying prematurely.

In recent years, insurers have considered the use predictive modelling and customer segmentation to improve pricing models. This allows for wider segmentation techniques like behavioural or demographic differences to influence pricing.

The aim is to build a machine learning (ML) model to predict insurance charges based on a number of variables such as age, body mass index (BMI), smoking status, and region influence insurance costs.

The insurance dataset has variables that are considered in real life insurance pricing process. Age, as older individuals have higher health risks and medical costs. Sex, gender can have an impact in health outcomes and life expectancy (women tend to take less risks than men). Body Mass Index (BMI) this takes into account someone’s height and mass which are strongly related to conditions like diabetes and heart disease. Smoking status is one of the most important variables, as smokers are more prone to have chronic health issues, leading to higher insurance charges. Region looks at the geographic differences in access to healthcare and pricing thereof.

## Methods

The **insurance dataset** commonly used in regression tutorials (often titled insurance.csv) is a **synthetic dataset**. That means it wasn’t collected from real patients or insurance companies, but rather **generated to simulate realistic patterns** found in actual insurance data. Here's what we know about its background:

* 📍 **Source**: The dataset was originally published by [Dataquest](https://www.dataquest.io/blog/predicting-insurance-costs-with-linear-regression/) as part of a machine learning tutorial. It was designed to help learners practice regression modeling in Python.
* 🧪 **Simulation-Based**: The values were crafted to reflect plausible relationships between demographic factors and insurance charges. For example:
  + Smokers tend to have higher charges.
  + BMI and age correlate with increased costs.
  + Regional differences are included to mimic geographic pricing variations.
* 🔐 **Privacy-Conscious**: Because it’s synthetic, there are no privacy concerns or HIPAA restrictions. This makes it ideal for public use in education and competitions.
* 📊 **Structure**: It contains 1,338 rows and 7 columns, with a mix of categorical and numerical features — just enough complexity to be useful, but not overwhelming for beginners.

Explain what analysis you chose to use and why it was suitable for your dataset

. Lack of Real-World Complexity

Synthetic data often simplifies relationships between variables. It may not capture rare events, edge cases, or the messy correlations found in real-world datasets — which can lead to overly optimistic model performance.

2. Limited Generalizability

Models trained exclusively on synthetic data may perform poorly when applied to real-world data. This is because synthetic data lacks the noise, anomalies, and diversity that real data contains.

3. Risk of Bias Amplification

If the synthetic data is generated using biased assumptions or flawed source data, those biases can be baked into the dataset and amplified during modeling.

4. False Sense of Security

Because synthetic data is clean and well-structured, it can give the illusion that your model is robust. But when deployed in real-world scenarios, it may fail to handle unexpected inputs or data drift.

5. Missing Contextual Nuance

Real-world data often includes subtle patterns influenced by cultural, economic, or behavioral factors. Synthetic datasets may overlook these nuances, limiting the depth of insights your model can uncover.

6. Validation Challenges

It’s harder to validate models trained on synthetic data because there’s no “ground truth.” You can’t compare predictions to actual outcomes, which makes performance metrics less meaningful.

** Strong Baseline Performance**  
Linear regression is widely used as a **baseline model** due to its simplicity and transparency. It helps identify which features (like age, BMI, or smoking status) have the strongest linear relationships with insurance charges.

1. **Feature Interpretability**  
   Studies emphasize that linear regression allows for **clear interpretation of coefficients**, making it easier to understand how each variable contributes to cost. For example, smoking status and BMI often show the highest positive coefficients, indicating strong influence on charges.
2. **Limitations in Capturing Nonlinearity**  
   While effective for linear relationships, linear regression struggles with **nonlinear patterns**. Research comparing models found that techniques like **Support Vector Machines (SVM)** and **Random Forests** outperform linear regression in terms of accuracy and error metrics.
3. **Use in Educational and Exploratory Analysis**  
   Papers often use linear regression for **exploratory data analysis**, helping visualize correlations and build intuition before applying more complex models.
4. **Model Extensions**  
   Some studies extend basic linear regression to **Ridge** and **Lasso regression** to improve generalization and feature selection, especially when dealing with multicollinearity or overfitting.

** Dai (2024)**: Demonstrates that linear regression provides a strong baseline for predicting insurance charges, especially when smoking status and BMI are included as predictors.

1. **Jiang (2022)**: Shows that linear regression models can effectively estimate medical costs, but their accuracy is limited by the model’s inability to capture nonlinear relationships.
2. **Narayana et al. (2023)**: Compares multiple regression models and concludes that while linear regression is interpretable, ensemble methods like Random Forests yield better predictive performance.

## Results

Include both descriptive and simple inferential results

## Discussion

Discuss your analyses and how this relates to the outcome of interest. And recommendations of how these findings could assist in improving the healthcare system

## Conclusion

## References

* Christiano, P., 2023. Insurance Pricing in 2025: An In-Depth Look at How Insurers Price Policies and Emerging Technologies Transforming Pricing Models. ExpertBeacon. Available at: <https://expertbeacon.com/insurance-pricing/> [Accessed 24 Aug. 2025].
* Szczepaniak, A., 2024. Insurance Pricing 101: Understanding the Fundamentals of Rate-Making. ShapedThoughts. Available at: <https://shapedthoughts.io/insurance-pricing-101-understanding-the-fundamentals-of-rate-making/> [Accessed 24 Aug. 2025].
* Chowdhury, B., 2017. How Information Asymmetry Can Drive Up Insurance Prices. Harvard Business School Online. Available at: <https://online.hbs.edu/blog/post/how-information-asymmetry-can-drive-up-insurance-prices> [Accessed 24 Aug. 2025].
* Kot, I., 2021. The Pros and Cons of Synthetic Data. DATAVERSITY. Available at: <https://www.dataversity.net/the-pros-and-cons-of-synthetic-data> [Accessed 24 Aug. 2025].
* Syntheticus, 2023. The Benefits and Limitations of Generating Synthetic Data. Syntheticus Blog. Available at: <https://syntheticus.ai/blog/the-benefits-and-limitations-of-generating-synthetic-data> [Accessed 24 Aug. 2025].
* Burden, J., Chiodo, M., Grosse Ruse-Khan, H., Markschies, L., Müller, D., Ó hÉigeartaigh, S., Podszun, R. and Zech, H., 2025. Model Collapse and the Right to Uncontaminated Human-Generated Data. Harvard Journal of Law & Technology. Available at: <https://jolt.law.harvard.edu/digest/model-collapse-and-the-right-to-uncontaminated-human-generated-data> [Accessed 24 Aug. 2025].
* Dai, W., 2024. *Predicting insurance charges using linear regression models*. CONF-MPCS 2024 Workshop. Available at: <https://www.researchgate.net/publication/385547453_Predicting_insurance_charges_using_linear_regression_models/fulltext/6729d44eecbbde716b585764/Predicting-insurance-charges-using-linear-regression-models.pdf> [Accessed 24 Aug. 2025].
* Jiang, Z., 2022. *Build Linear Regression Models for Insurance Charges*. International Journal of Social Science and Economic Research, 7(2). Available at: <https://ijsser.org/2022files/ijsser_07__31.pdf> [Accessed 24 Aug. 2025].
* Narayana, K.L., Yogesh, and Kowshik, P., 2023. *Medical Insurance Premium Prediction Using Regression Models*. IJRTI. Available at: <https://ijrti.org/papers/IJRTI2304248.pdf> [Accessed 24 Aug. 2025].