2025

Predictive Modeling of Medical Aid Charges Report

PDAN8411 PART 1

MAX WALSH

22 AUGUST 2025

Table of Contents

Objective

The objective of this analysis is to assist the medical aid scheme in predicting medical charges based on demographic and lifestyle data. To build a linear regression model that is transparent and interpretable to identify key cost drivers and improve forecasting accuracy for both new and existing clients (OpenAI, 2025). Ideally, the model is accurate in predicting costs, interpretable for actuarial teams, and scalable for quote calculations/risk profiling systems (OpenAI, 2025). Throughout the model's development, several key steps were taken, including exploring cost drivers through visual analytics, quantifying relationships through statistical modelling, validating model performance with real-world cost data, and delivering actionable insights (OpenAI, 2025). Building a linear regression model helped strike a balance between performance and transparency, enabling the business to accurately predict costs and understand the factors that drive them (OpenAI, 2025).

Dataset Overview

Source: Kaggle – “Medical Cost Personal Datasets” by Choi (Choi, 2018).

Size: 1,338 records, 7 features.

Target variables: charges.

Feature	Type	Description
age	Numerical	Age of the primary beneficiary
sex	Categorical	Male or female
bmi	Numerical	Body Mass Index
children	Numerical	Number of children/dependents
smoker	Categorical	Whether the person is a smoker
region	Categorical	Geographical region in the United States

Table 1: Dataset Features

Data Preparation

To ensure the dataset's suitability for linear regression modelling, the dataset underwent several critical preprocessing steps (Suzanne, 2023). Providing a dataset that ensures quality, consistency, and suitability of data (Suzanne, 2023).

The steps were taken as follows:

1. First, a check was made to ensure there were no null entries across any of the features (Suzanne, 2023).
2. Next, one-hot encoding was utilised to transform categorical values, sex, smoker, and region, into a numerical format (Suzanne, 2023). By creating new columns for the categorical variables and assigning binary values to each column, the categorical variables will be in numerical format, which is essential for machine learning algorithms based on mathematical equations (Suzanne, 2023). Linear regression modelling also assumes no multicollinearity, so 'drop_first=True' was used to remove the first category to avoid redundancy and multicollinearity (Suzanne, 2023).
3. We then corrected the skewness of the target variable charges. This variable was highly right-skewed, as there is a small number of individuals who have extremely high medical costs (OpenAI, 2025). As linear regression assumes normally distributed residuals, we applied a natural logarithm transformation to charges (Thrane, 2023). Log-transformed variable log_charges offered a normalised distribution and a stabilised variance (Thrane, 2023).
4. Features were then reviewed again to ensure they were in a numeric format post-encoding and transformation (Suzanne, 2023). Aligning them with scikit-learn's Linear Regression model requirements (Suzanne, 2023).
5. Finally, the dataset was split into train and test data. Following an 80/20 train-test split, respectively (Suzanne, 2023). This was done through random sampling to ensure the model could be accurately tested on unseen data (Suzanne, 2023).

These steps were crucial to ensure the data was clean, structured, and statistically ready for effective linear regression modelling (Suzanne, 2023). Whilst offering easy business interpretability for model output explanations to stakeholders, regulators, and end users (OpenAI, 2025).

Exploratory Data Analysis (EDA)

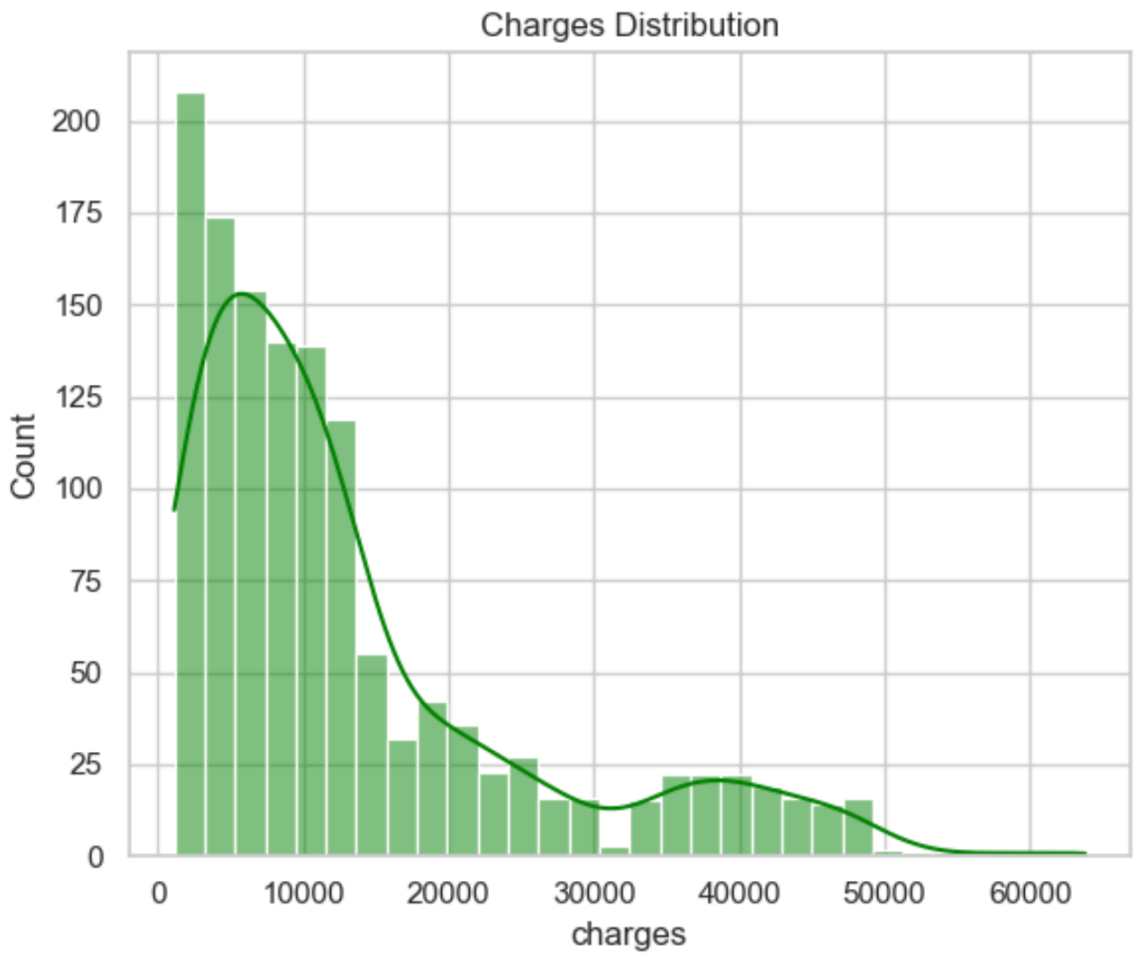

Figure 1: Histogram of Medical Charges Before Transformation

As seen in Figure 1, the target variable charges is seen to have a strong right-skewed distribution. This figure visually shows the motivation to apply a log transformation to normalise the target variable and stabilise variance (OpenAI, 2025).

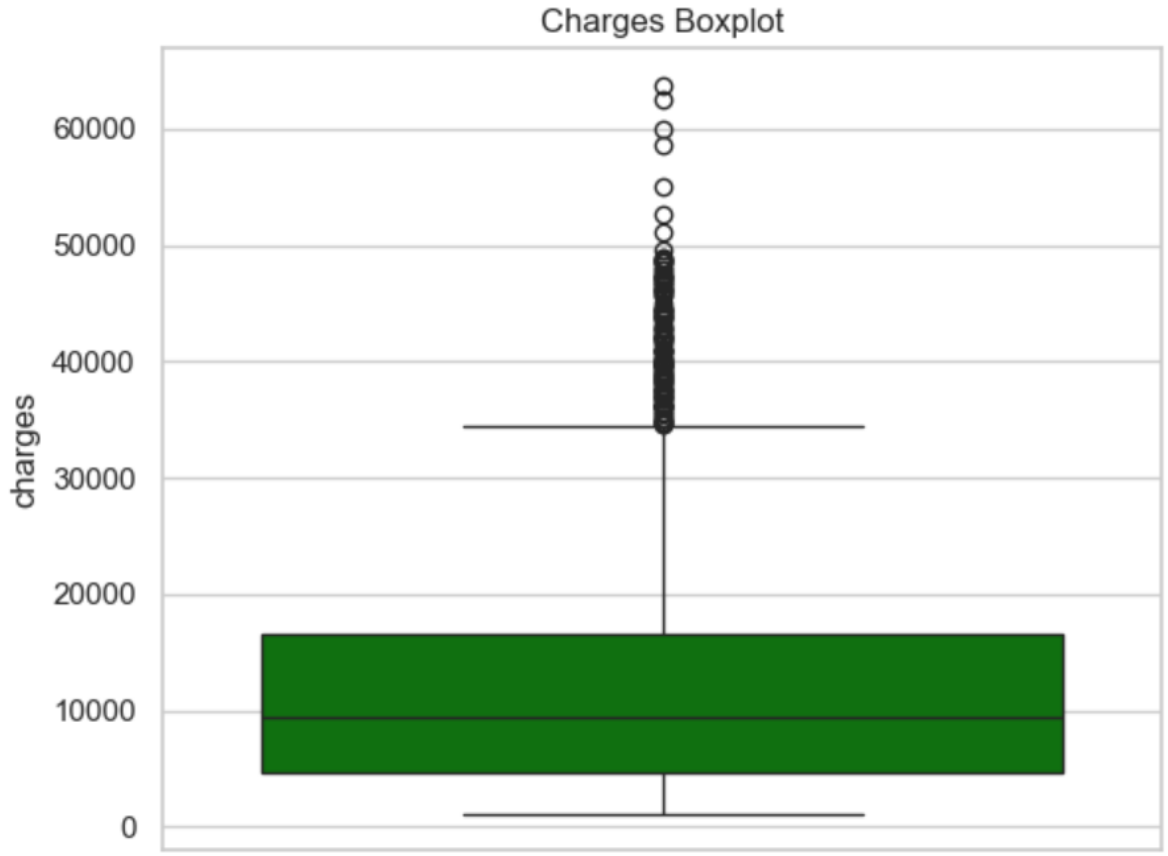

Figure 2: Boxplot of Medical Charges

Figure 2 shows that charges distribution shows extreme outliers with many data points above R40,000. This is additional support to log-transform the target variable, as variation is visible (OpenAI, 2025).

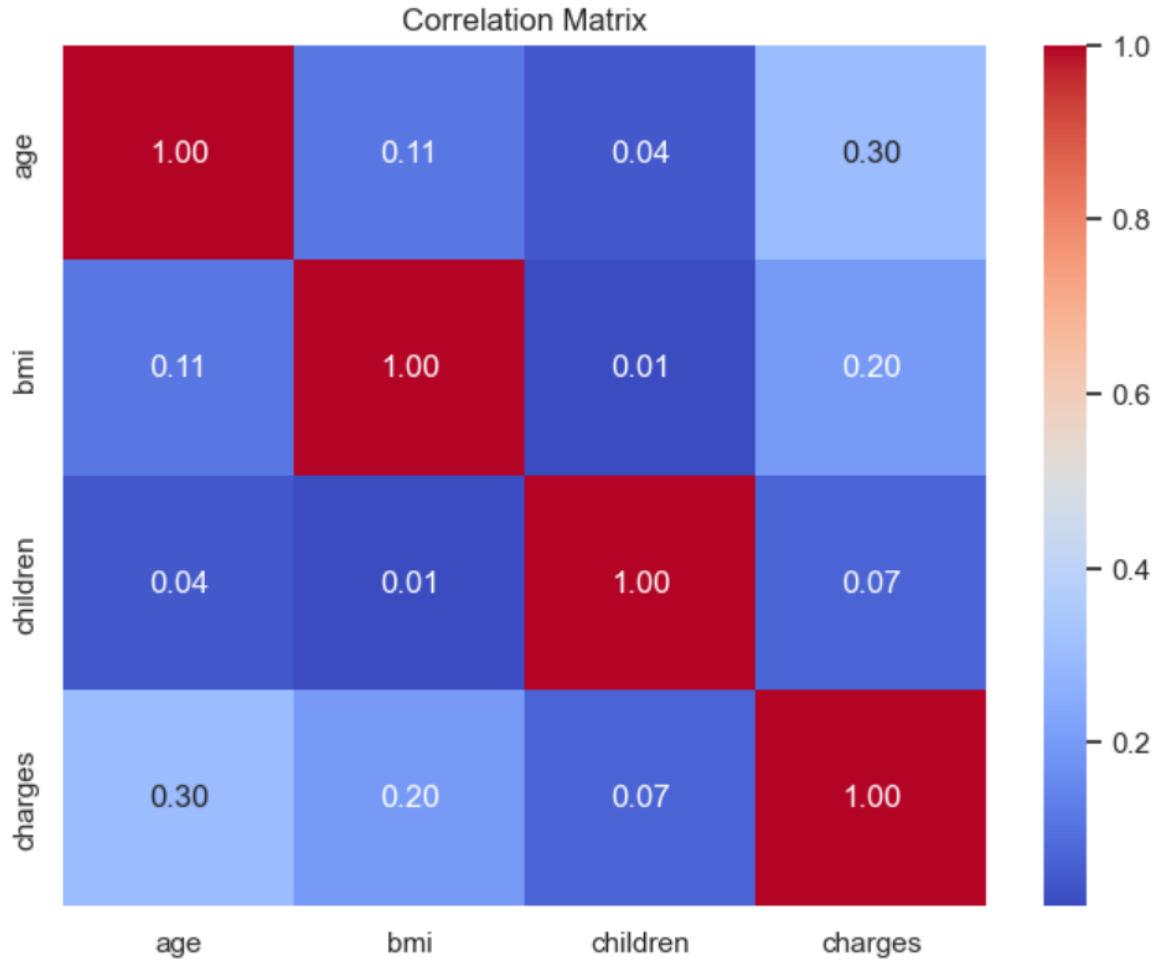

Figure 3: Correlation Heatmap

Figure 3 matrix indicates the strength of linear correlation between key numerical variables (Wagavkar, 2024). We can see numerical correlation values between age, bmi, children, and charges. Correlation values range from -1 to 1. A -1 value indicates a strong negative relationship, a 0 value indicates no relationship, and a +1 value indicates a strong positive relationship (Wagavkar, 2024). The strongest correlation with charges is age, with a correlation value of 0.30, followed by bmi with a correlation value of 0.20. Justifying these variables' inclusion in the model. The variable children has almost no correlation with charges, with a correlation value of 0.07. However, the variable children was retained to observe potential interaction effects (OpenAI, 2025).

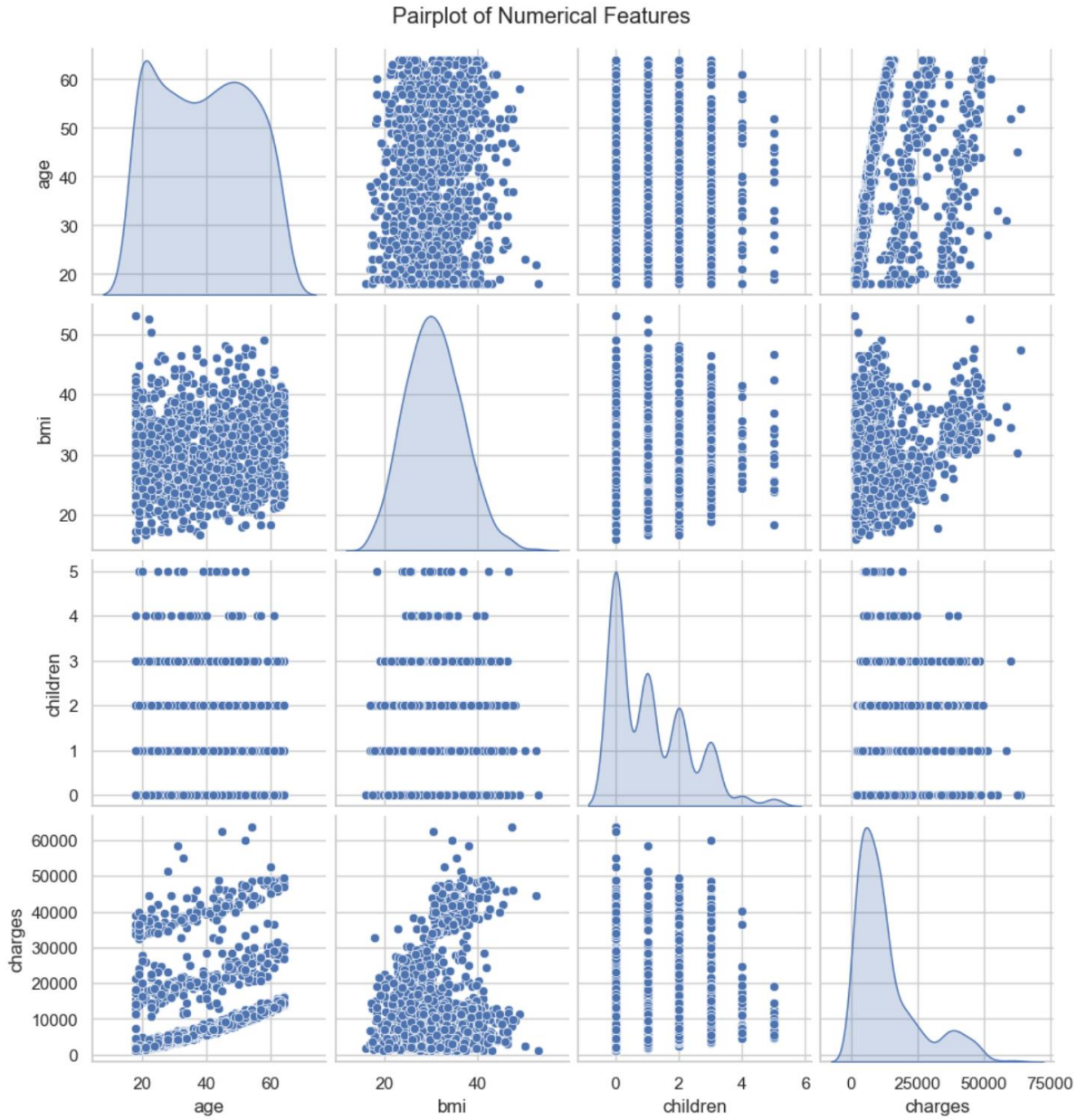

Figure 4: Pairplot of Numerical Features

Figure 4 shows the distribution of pairwise relationships among the numerical variables: age, bmi, children, and charges. The variable charges seems to increase with variable age (OpenAI, 2025). The variable bmi seems to correlate with higher costs in some cases (OpenAI, 2025). The seen distribution of charges clearly shows the need for log transformation to normalise the distribution.

Feature Engineering

Variance Inflation Factor (VIF) was applied to measure and ensure no multicollinearity (Shrestha, 2020). VIF shows how much a feature is linearly correlated to other features (Shrestha, 2020). As stated, linear regression assumes no multicollinearity; a VIF value close to 1 indicates no multicollinearity (Shrestha, 2020).

	Feature	VIF
0	const	35.527488
1	age	1.016822
2	bmi	1.106630
3	children	1.004011
4	sex_male	1.008900
5	smoker_yes	1.012074
6	region_northwest	1.518823
7	region_southeast	1.652230
8	region_southwest	1.529411

Table 2: Variance Inflation Factor (VIF) for Each Feature

As seen in Table 2, all VIF values are below 2.0. Showing low multicollinearity in all features (Shrestha, 2020). Promoting the use of linear regression for this dataset (Shrestha, 2020). Whilst also showing the independence of the predictor variables in the model (Shrestha, 2020).

In addition to VIF, one-hot encoding and skew correction were applied as feature engineering to improve the dataset's compatibility with linear regression:

- One-hot encoding was applied to categorical variables sex, smoker, and region to transform them into a machine-readable numerical format (Katya, 2023). To prevent multicollinearity, the first category in each variable was dropped (OpenAI, 2025).
- The target variable charges was heavily skewed to the right, as seen in Figure 1, so to normalise this target variable and stabilise variance, a log transformation was applied,

giving log_charges (OpenAI, 2025). Enhancing the model's ability to find linear patterns (Katya, 2023).

These feature engineering steps, VIF, one-hot encoding, and skew correction, provide the dataset compatibility with linear regression's assumptions and offer interpretable and accurate results (Katya, 2023).

Model Training

The model selected is a linear regression model. The library utilised to implement the model is scikit-learn. Scikit-learn is a powerful Python library for machine learning (Suzanne, 2023). The model was trained on a random 80/20 train-test split (Suzanne, 2023). The target variable is log_charges, a log-transformed version of the variable ‘changes’, necessary to correct the strong right-skew seen in Figure 1 and stabilise variance (OpenAI, 2025). The model was trained on all 8 encoded and engineered features, including one-hot encoded and categorical features (OpenAI, 2025). Linear regression was the selected model because the target variable, charges, is continuous and approximately linear with several features, including age, bmi, and smoker (OpenAI, 2025). The Exploratory Data Analysis processed confirmed this model is suitable, as seen in Figure 3 and Figure 4. Additionally, the transparent and interpretable provision of a linear regression model is suitable for this case, as the clients would value understanding how lifestyle factors influence medical charges (OpenAI, 2025). More so, the client's business needs to provide predictable, scalable premium estimates that are efficiently met with this model (OpenAI, 2025). The clients required an accurate, yet explainable, scalable, and efficient model, all of which this linear regression model is. The model is fast to train and robust when multicollinearity is controlled, which was done through the Variance Inflation Factor (VIF) analysis (Suzanne, 2023). As seen in Table 2, all features were seen to have a VIF value below 2.0, indicating suitability for inclusion within this statistically robust model (Shrestha, 2020).

Model Evaluation

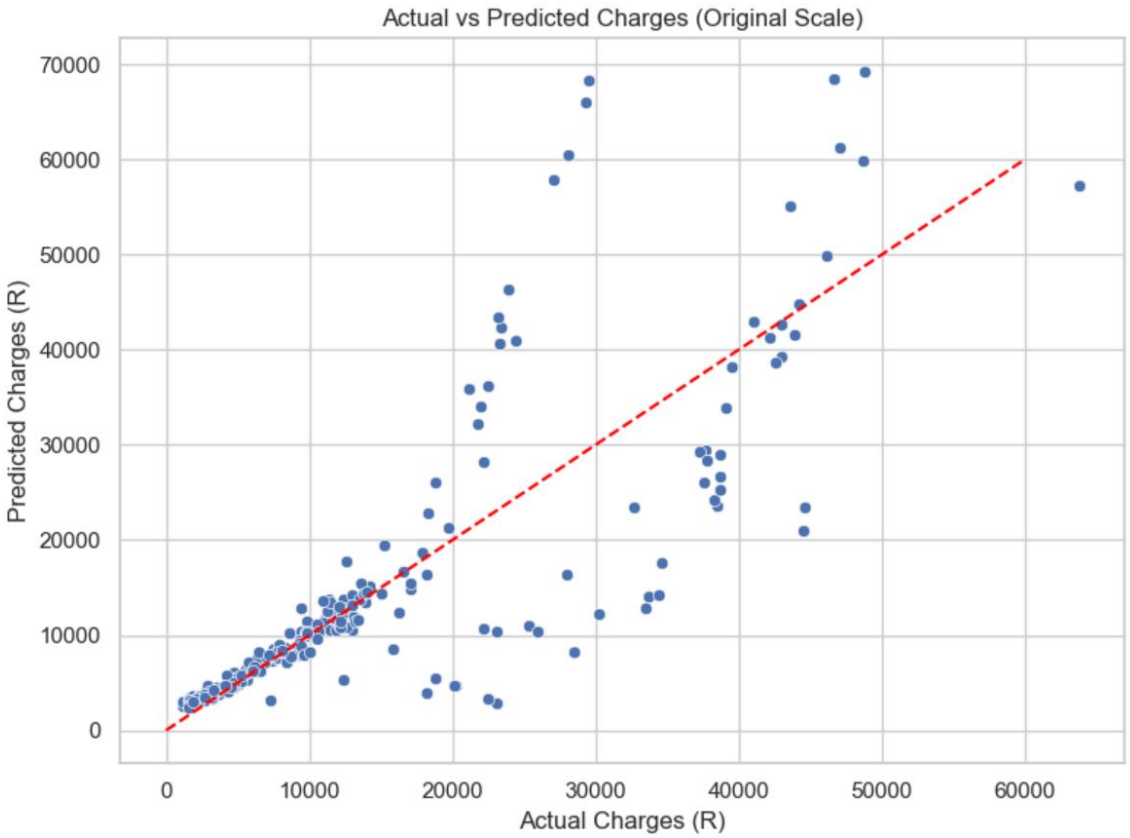

Figure 5: Scatterplot of Actual vs Predicted Charges

As seen in Figure 5, the plot shows a strong correlation between predicted and actual values (Ault, et al., 2025). This scatterplot is a visual diagnostic of the model's performance in monetary terms (OpenAI, 2025). Many points are clustered around the red line. The model is effectively capturing the underlying relationship between predictors and medical charges, as seen by the upward trend (Ault, et al., 2025). Over and under predictions are seen in the point deviations above and below the red line (Ault, et al., 2025). Below R30,0000, predicted charges follow a similar distribution to actual charges. As expected, higher charge cases, being a rarer occurrence, diverge from the red line, indicating outliers (OpenAI, 2025). Confirming the model is performing well for realistic cases (OpenAI, 2025).

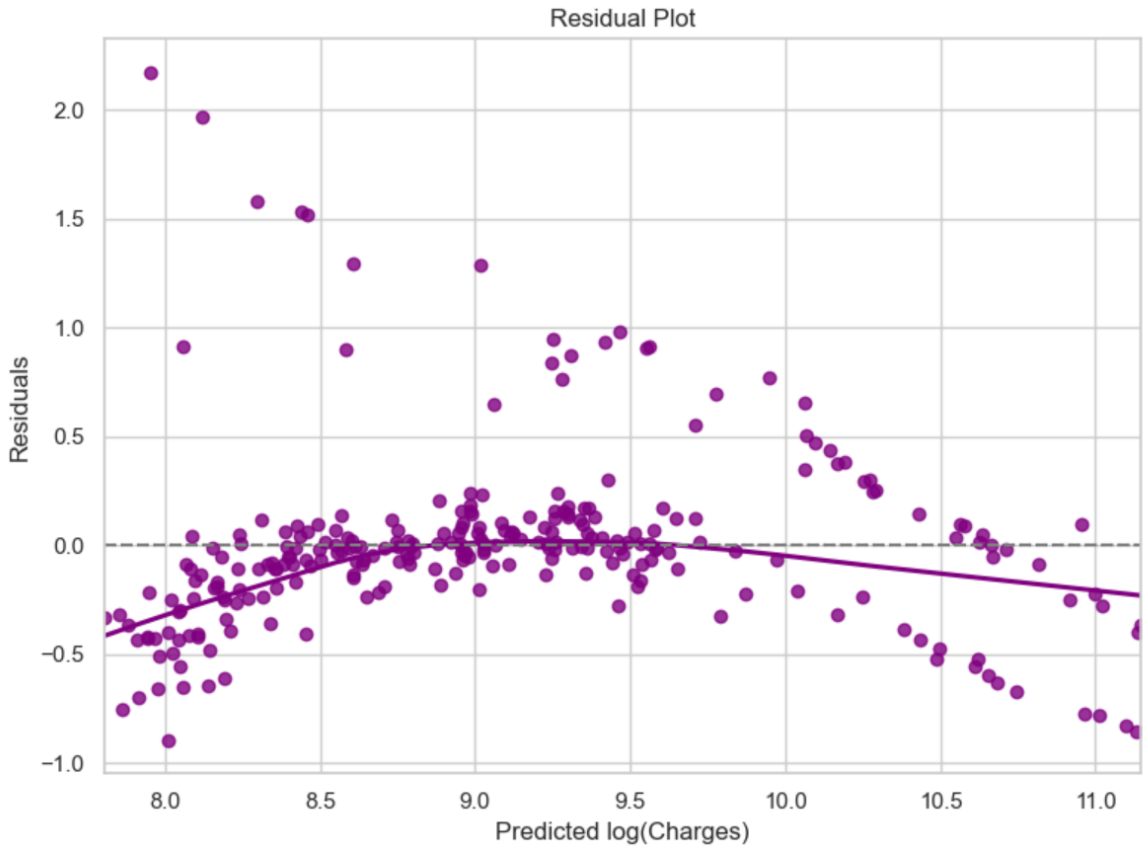

Figure 6: Residual Plot on Log-transformed Scale

Figure 6 depicts the residual plot on the log-transformed scale of the variable charges. Around the horizontal zero line, the residual appears randomly distributed with no distinct pattern, indicating an adequate capture of the variability in the data (GeeksForGeeks, 2024). The variance of model error remains constant across predicted values according to the lack of a distinct pattern and, therefore, no major violation of linear regression assumptions (GeeksForGeeks, 2024). Additionally, the lack of structure and curvature around the zero line provides confidence in the model's stability (GeeksForGeeks, 2024). Extreme value deviations are expected in real-world cost predictions (OpenAI, 2025).

Evaluation results on the full model on the log-transformed scale:

- R^2 Score: 0.8047
- Mean Absolute Error (MAE): 0.2697
- Root Mean Squared Error (RMSE): 0.419

This model explains that about 80% of the variance in 'log_charges', suggesting strong predictive performance on the log-transformed scale (OpenAI, 2025). This R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variable (Farshad, 2024). The mean absolute error of 0.2697 indicates that, on average, predictions are close to actual values in the normalised space (Agrawal, 2025). Mean absolute error measures the average magnitude of errors in predictions (Farshad, 2024). The root mean squared error value of 0.419 shows that the model is relatively stable, whilst not being penalised excessively by large errors (Agrawal, 2025). The root mean squared error gives an easily interpretable measure of average error size (Farshad, 2024). We can deduce that when predicting transformed medical charges, the model is accurate and reliable (OpenAI, 2025).

Model Retraining

The model was retrained using a reduced feature set of smoker_yes, age, and bmi, which were the strongest predictors identified during the correlation and EDA phase, as seen in Figure 7 below.

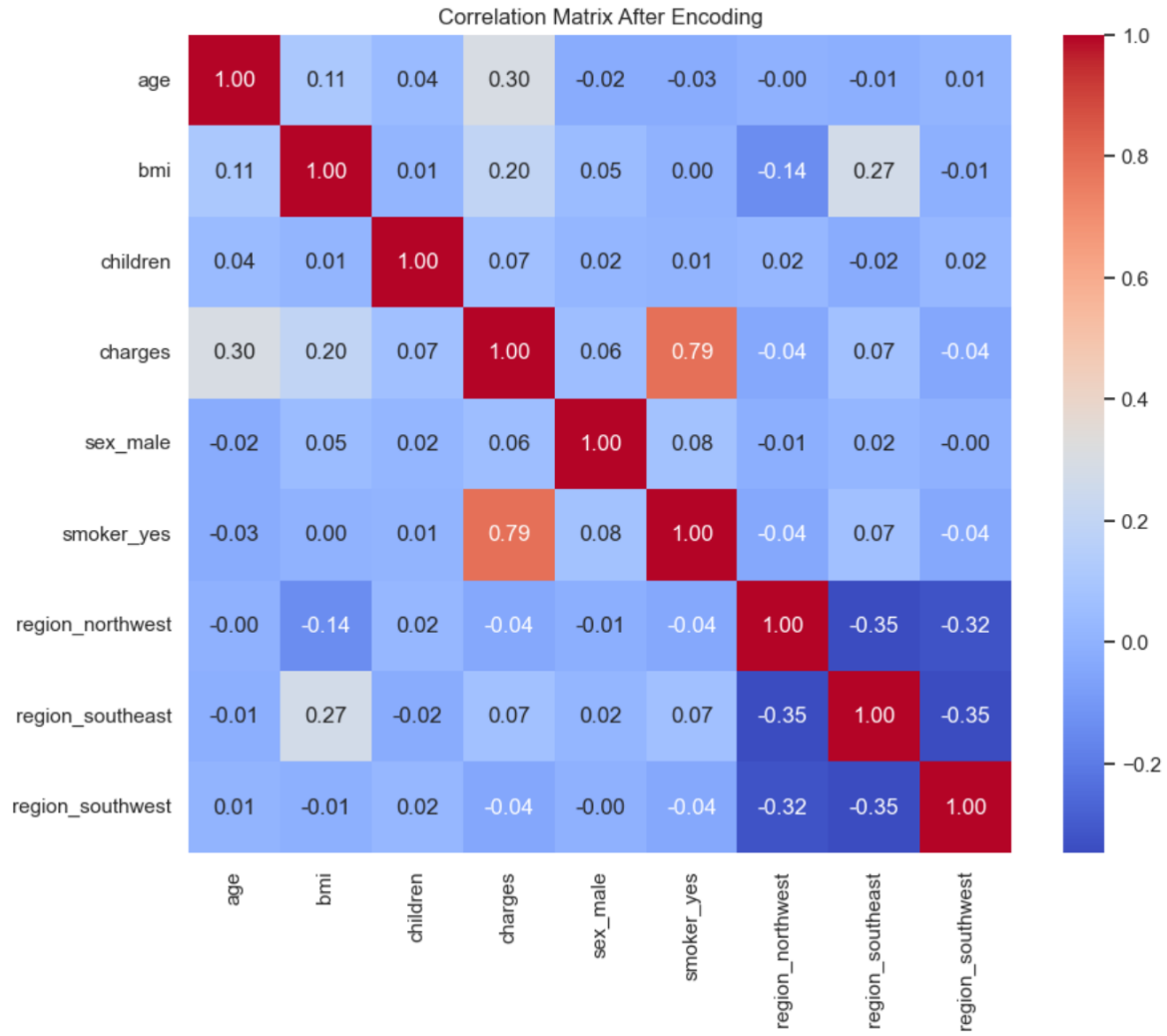

Figure 7: Correlation Matrix After Encoding

The correlation matrix seen in Figure 7 indicates the correlation between all encoded predictor variables and the target charges (OpenAI, 2025). The strongest predictor variables with a positive correlation to charges are smoker_yes with a value of 0.79, age with a value of 0.30, and bmi with a value of 0.20 (OpenAI, 2025). Hence, their selection in the simplified slim model, given their predictive strength and business relevance (OpenAI, 2025).

Evaluation results on the slim model on the log-transformed scale:

- R^2 Score: 0.7705
- Mean Absolute Error (MAE): 0.2988
- Root Mean Squared Error (RMSE): 0.4543

Although the slim model only utilised three variables, the model performed almost as well as the full model according to our evaluation metrics. Much of the predictive power is concentrated in these three features: smoker_yes, age, and bmi, as seen by the high R^2 score and the reasonably low mean absolute error and root mean squared error (Agrawal, 2025). This suggests that the full model still offers higher accuracy and includes a broader context, but the slim model shouldn't be neglected and offers a simple, faster, and more interpretable alternative with little performance trade-off (OpenAI, 2025). The comparison of the full vs slim model confirms the robustness of our feature selection process and indicates flexibility in deploying either model based on the application (OpenAI, 2025).

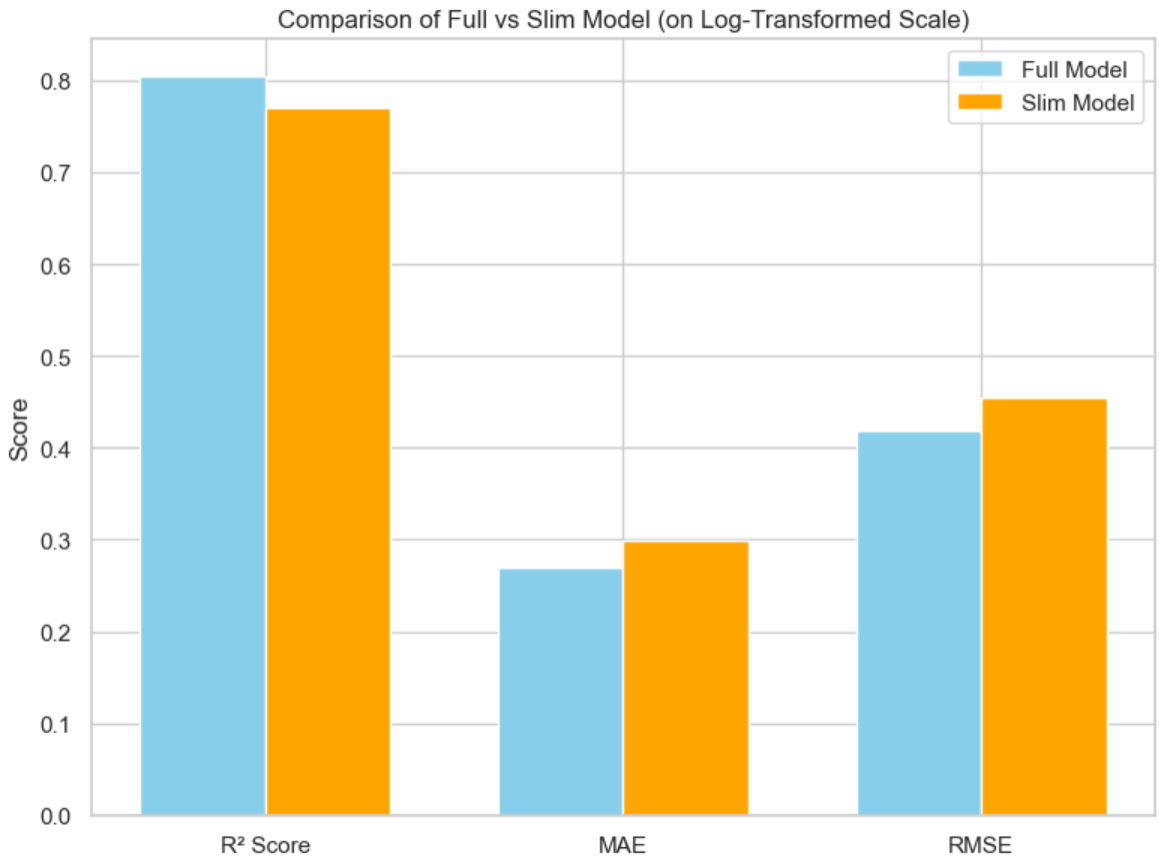

Figure 8: Comparison of Full vs Slim Models on Log-Transformed Scale

The bar chart seen in Figure 8 compares the performance metrics of the full model, which utilises all features, and the slim model, which only utilises three features (smoker_yes, age, and bmi). Although the full model achieved slightly better accuracy with its high R^2 score and low mean absolute and root mean squared error, the slim model performed competitively (OpenAI, 2025). Aiding the model's application in circumstances where simplicity and interpretability are necessary, and showing that these three features drive most of the cost variation (OpenAI, 2025).

Recommendations

It is recommended that both the full and slim models be used as predictive tools in applicable scenarios to estimate the expected charges for new members. The full model should be considered for scenarios where in-depth analysis is required, and accuracy and comprehensiveness are crucial (OpenAI, 2025). The slim model should be considered for real-time quoting tools or scenarios that require high interpretability and speed (OpenAI, 2025). Additionally, focus wellness initiatives on smokers, individuals with high body mass index, and older members (OpenAI, 2025). These factors are seen to significantly impact medical expenses and show a clear opportunity for targeted cost reduction (OpenAI, 2025). This approach effectively allows the organisation to balance accuracy, transparency, and efficiency across various use cases.

Next Steps

To maximise the value of the predictive model, the following is recommended:

- Deploy the model through a user-friendly interface. Like a web-based dashboard, Excel plugin, or quoting API (Application Programming Interface) (OpenAI, 2025). This will allow internal teams to easily access the model and provide real-time charge calculations during onboarding and renewal (OpenAI, 2025).
- Continuously track key performance metrics such as R^2 score, mean absolute error and root mean squared error to monitor the model performance (Farshad, 2024). An annual model update based on new data is recommended to reflect changes in population health trends or cost structures (OpenAI, 2025).
- Implementing advanced non-linear techniques like decision trees, random forest trees, and gradient boosting is a sure way to expand the modelling approach and potentially unveil complex interactions between features and provide improved accuracy, especially for outlier cases (Kalusivalingam, et al., 2022).
- Incorporate more features in future model iterations to enhance model precision and value further (Wang, 2021). Features such as chronic condition indicators, income brackets, or lifestyle variables like alcohol use, exercise frequency, and diet can provide valuable insights to improve precision (Wang, 2021).
- Conduct audits periodically, confirming the model remains unbiased and aligns with the changing ethical standards in medical underwriting and pricing (Wang, 2021).

These steps provide a sustainable and maximised value use of the predictive model.

Disclosure of AI Use

Sections: Part 1.

Name of the tool used: ChatGPT4.

Purpose behind use: Outlines, summaries, Python analysis code, queries, evaluations, and suggestions.

Date used: 10/04/2025.

Link to chat: https://chatgpt.com/share/6800b44d-98b4-8004-a81b-7855cb0b5000

References

Agrawal, R., 2025. *Know The Best Evaluation Metrics for Your Regression Model.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/
[Accessed 21 April 2025].

Ault, D. S. V., Liao, D. S. N. & Musolino, L., 2025. *Principles of Data Science.* Houston: OpenStax.

Choi, M., 2018. *Kaggle - Medical Cost Personal Datasets.* [Online]
Available at: https://www.kaggle.com/datasets/mirichoi0218/insurance
[Accessed 01 April 2025].

Farshad, K., 2024. *Essential Regression Evaluation Metrics: MSE, RMSE, MAE, R^2, and Adjusted R^2.* [Online]
Available at: https://farshadabdulazeez.medium.com/essential-regression-evaluation-metrics-mse-rmse-mae-r%C2%B2-and-adjusted-r%C2%B2-0600daa1c03a#:~:text=MSE%20and%20RMSE%20are%20useful,t%20penalize%20for%20extra%20predictors.
[Accessed 18 April 2025].

GeeksForGeeks, 2024. *Residual Analysis.* [Online]
Available at: https://www.geeksforgeeks.org/residual-analysis/
[Accessed 21 April 2025].

Kalusivalingam, A. K., Sharma, A., Patel, N. & Singh, V., 2022. *Leveraging Random Forests and Gradient Boosting for Enhanced Predictive Analytics in Operational Efficiency.* [Online]
Available at: https://cognitivecomputingjournal.com/index.php/IJAIML-V1/article/view/72
[Accessed 18 April 2025].

Katya, E., 2023. *Exploring Feature Engineering Strategies for Improving Predictive Models in Data Science.* [Online]
Available at: https://technicaljournals.org/RJCSE/index.php/journal/article/view/88/84
[Accessed 18 April 2025].

OpenAI, 2025. *Open AI ChatGPT4.* [Online]
Available at: https://chatgpt.com/share/6800b44d-98b4-8004-a81b-7855cb0b5000
[Accessed 10 April 2025].

Shrestha, N., 2020. *Detecting Multicollinearity in Regression Analysis.* [Online]
Available at: https://www.researchgate.net/publication/342413955_Detecting_Multicollinearity_in_Regression_Analysis
[Accessed 17 April 2025].

Suzanne, 2023. *Data Pre-Processing for Linear Regression in Machine Learning.* [Online]
Available at: https://medium.com/@sds152/data-pre-processing-for-linear-regression-in-machine-learning-4b73ec48392a
[Accessed 17 April 2025].

Thrane, C., 2023. *The normality assumption in linear regression analysis — and why you most often can dispense with it.* [Online]
Available at: https://medium.com/@christerthrane/the-normality-assumption-in-linear-regression-

analysis-and-why-you-most-often-can-dispense-with-5cedbedb1cf4
[Accessed 17 April 2025].

Wagavkar, S., 2024. *Introduction to the Correlation Matrix.* [Online]
Available at: https://builtin.com/data-science/correlation-matrix#:~:text=For%20example%2C%20let's%20say%20you,the%20relationship%20is%20positively%20strong.
[Accessed 20 April 2025].

Wang, Y. P., 2021. *The Acturial Society of South Africa.* [Online]
Available at: https://www.actuarialsociety.org.za/convention/wp-content/uploads/2021/10/2021-ASSA-Wang-FIN-reduced.pdf
[Accessed 21 April 2025].