NUS SDS Mini-Datathon

Insurance Price Prediction by machine learning

Babu Regan Eshwar (A0281192E)
Duan Yihe (A0276944M)
Lee Zaccary (A0273017L)

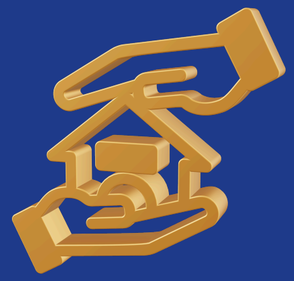
Problem Statement

Goal

Develop a machine learning model to accurately predict insurance charges based on demographic and lifestyle factors.

Key Question

Which variables influence insurance costs the most, and which model provides the best predictive performance?



Business Relevance

Our approach provides insurers and regulators with insights to ensure fair and data-driven premium decisions for all customers.

Exploratory Data Analysis

A dataset of 1338 entries was provided, the 7 features as follows

Feature	Type	Description
age	Numeric	Age of primary beneficiary
sex	Categorical	Male or female
bmi	Numeric	Body mass index
children	Numeric	Number of dependents
smoker	Categorical	Smoking status
region	Categorical	U.S. region (northwest, southeast, etc.)
charges	Numeric	Annual medical cost (target)

Fig. 1 Dataset Features, Types and Description for EDA

Observations of dataset:

- **No missing values.**
- **Charges range from $1k to $63k, charges are spread over a large range**
- **Identify the given data set is mixed with categorical and numerical data**

Exploratory Data Analysis

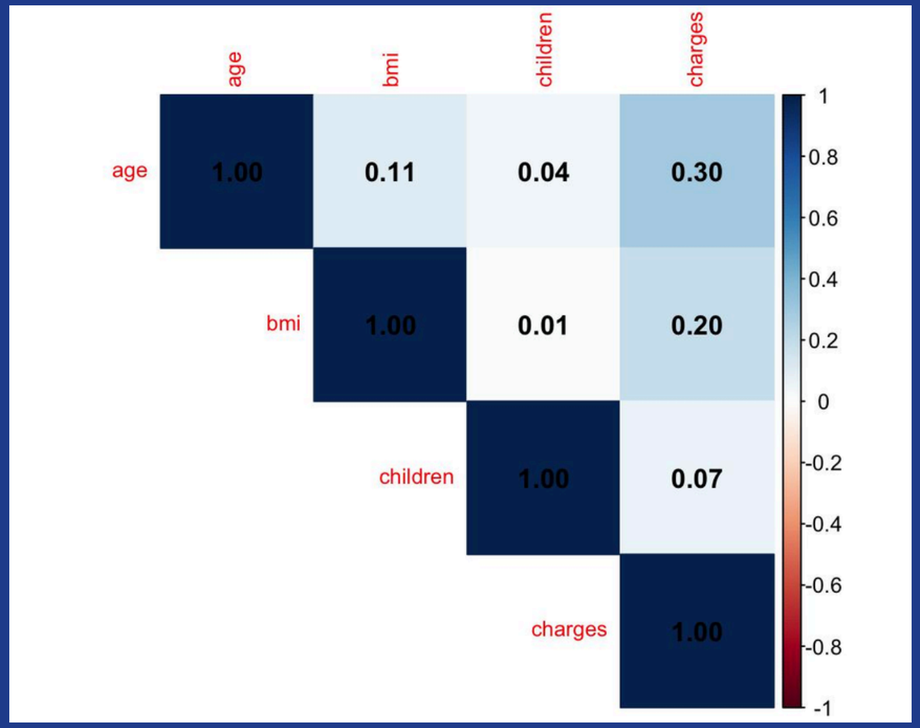


Fig. 2 Dataset Feature Heatmap

The correlation heatmap indicates that age and BMI have a positive relationship with insurance charges, whereas the number of children has a minimal effect. Age has the strongest correlation (r = 0.30), followed by BMI (r = 0.20). Low inter-feature correlations indicate minimal multicollinearity, supporting the use of multiple linear regression. Overall, age and BMI are the most influential numerical factors affecting charges.

Regression & modeling approach

	Model	RMSE	R2
1	Ridge	6280.304	0.7130288
2	Multiple Linear	6346.556	0.7078166

The RMSE represents the Root Mean Square Error. Both Ridge and Multiple Linear Regression models show similar performance.

The R^2 value or R value refers to the strength of linear relationship between predictors and the targets. Since the difference of R^2 value between these two models, the RMSE is a supplement to our analysis.

Since the predictor is unlikely to be perfectly linear to the target, therefore RMSE is more accurate matrix to evaluate our model performance.

Key findings & visualisations

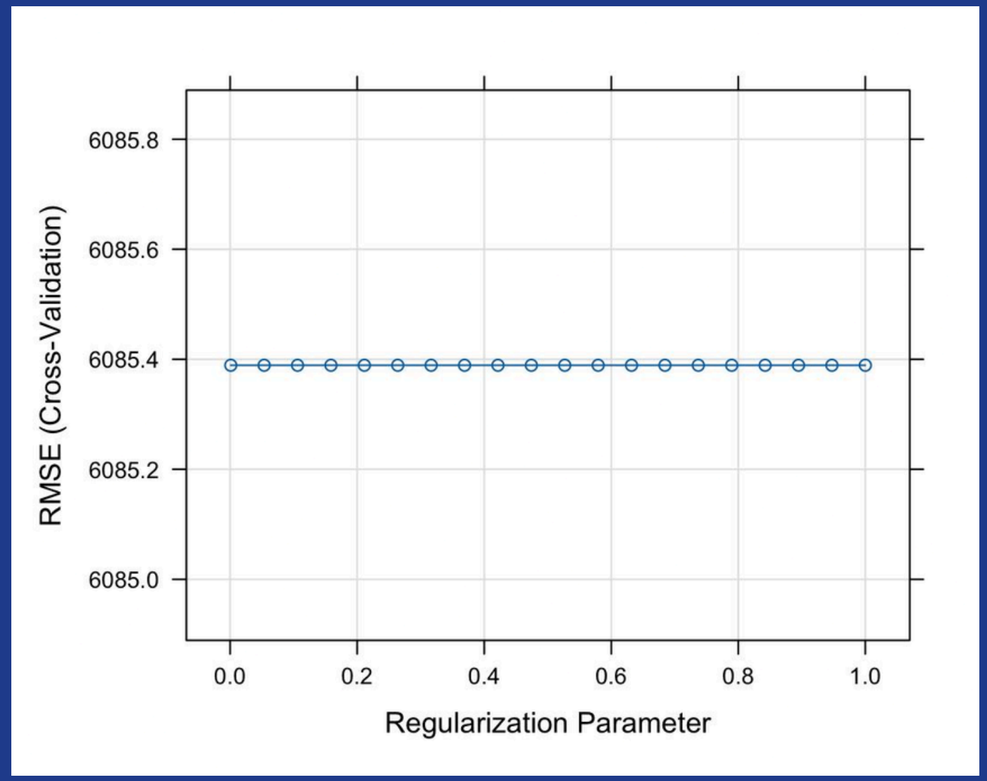
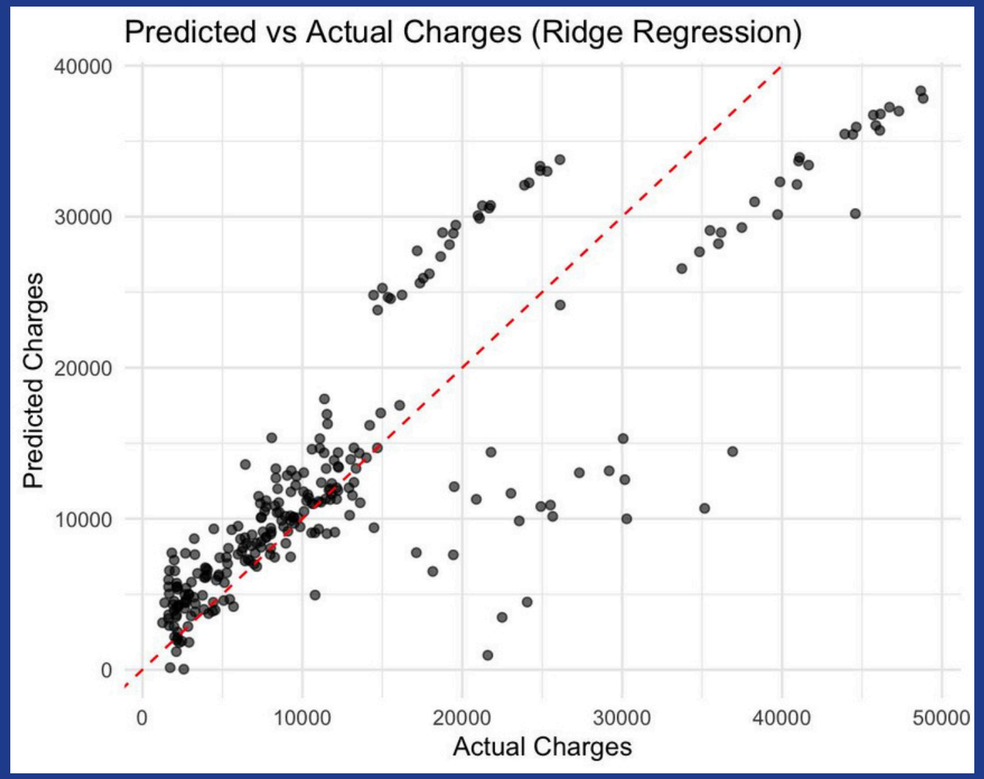


Fig. 3 RMSE vs. Regularization Parameter

Multiple Linear Regression was eventually selected. As seen in the figure above, since the RMSE remains nearly identical across different regularization strengths, there is no evidence that Ridge regularization improves model performance. Therefore, the simpler multiple linear regression model is preferred.

Key findings & visualisations



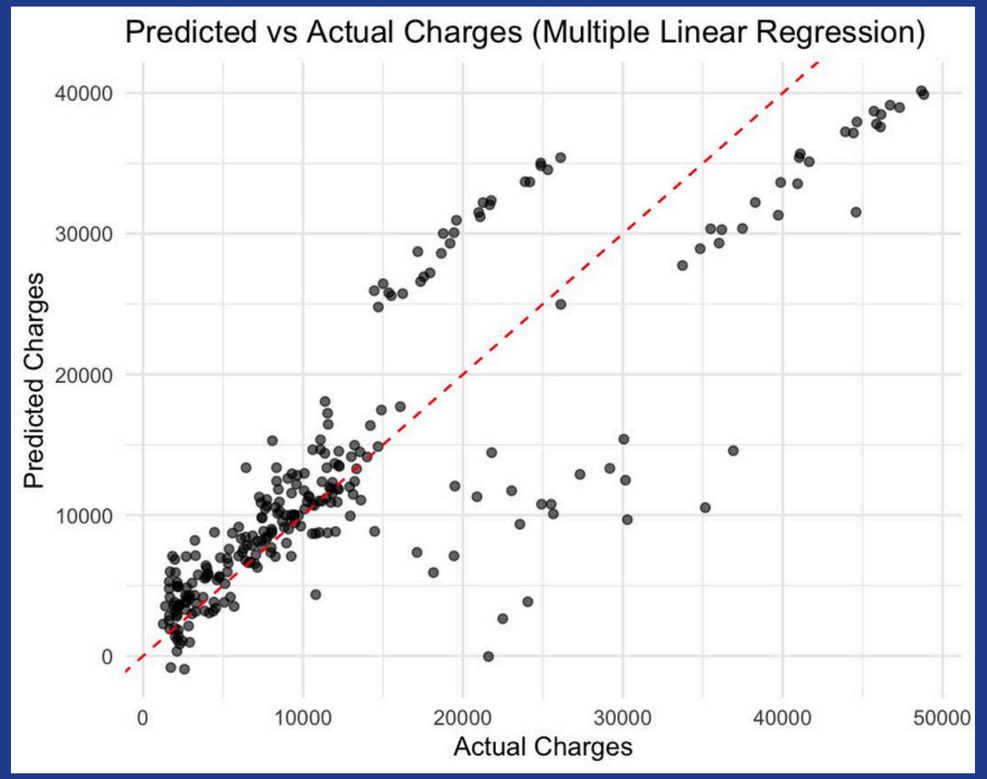


Fig 6. & 7. Predicted vs. Actual Charges for Ridge Regression (L) and Multiple Linear Regression (R)

Red Diagonal Line represents a perfect prediction. The closer the dot is to the Line, the more accurate our prediction is.

Key Findings & Visualisations

	Predictor	VIF	Df	GVIF^(1/(2*Df))	Collinearity_Level	F value
1	age	1.016822	1	1.008376	Low	477.023920
2	bmi	1.106630	1	1.051965	Low	148.206388
3	children	1.004011	1	1.002003	Low	15.551926
4	smoker	1.012074	1	1.006019	Low	3359.202387
5	region	1.098893	3	1.015841	Low	2.115253
6	sex	1.008900	1	1.004440	Low	0.155553

1. low correlations among factors, since collinearity level is low and Variance Inflation Factor is 1.
2. Smoker is the most dominant factor to a higher insurance charge with the highest F value. A higher F value indicates that a predictor explains a greater proportion of the variation in insurance charges.

Feature Impact & Fairness Analysis

The predictors sex and region exert minimal influence on insurance charges, as indicated by their low F values (0.16 and 2.12) and VIFs close to 1, signifying negligible explanatory power and low collinearity.

Conversely, variables such as age, BMI, and smoking status show a stronger impact on insurance charges, with significantly higher F values (477.02, 148.21, and 3,359.20). These factors are primary cost drivers because they are directly related to individual health risk and lifestyle choices.

From a fairness standpoint, the limited effect of demographic factors like sex and region suggests the model remains primarily relying on medically and behaviorally relevant predictors rather than personal and more detailed attributes.

Practical recommendations

The model predicts lower insurance charges well, but higher-cost predictions show two precise but not accurate clusters, one above and one below the predicted line, indicating it may not fully capture variability due to unobserved factors. There are outliers between $ 20,000 and $ 37,000 (actual charges).

To address this, robust regression techniques or separate models for low- and high-cost groups could be used, additional relevant features could be incorporated, and prediction intervals could be reported to better account for uncertainty in high-cost predictions.

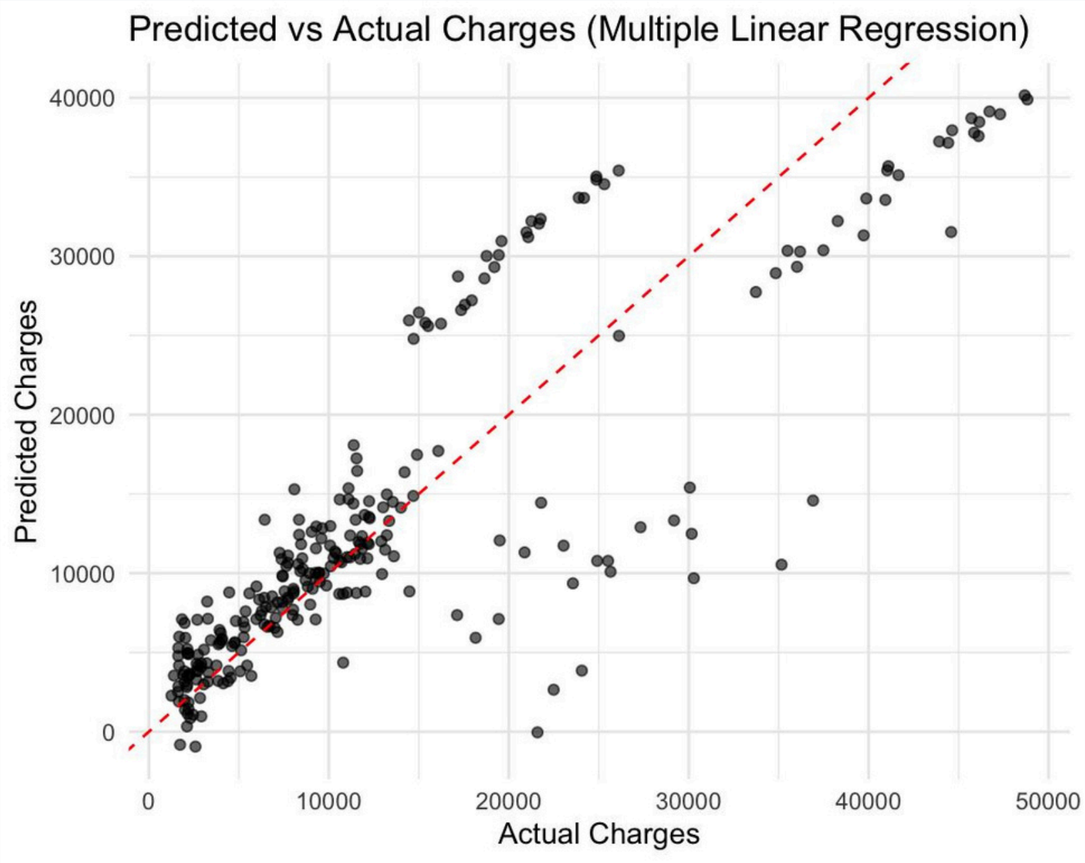


Fig 7. Predicted vs. Actual Charges for Multiple Linear Regression

Difficulties faced & Solutions

- Challenge 1: Concerned that the model's accuracy estimate might be insufficient with the usual 5-fold cross-validation.
- Solution: We used 10-fold cross-validation to obtain a more reliable performance estimate.

- Challenge 2: The dataset contains a mix of numerical and categorical variables, which some predictive models cannot handle directly.
- Solution: We transformed categorical variables into an appropriate numerical format (one-hot encoding) so they could be used in the models.

Appendix

[Link to our Github Repository.]