

The Outliers

NUS SDSS Mini-Datathon 2025

Erica, lian, Joanne, Kaylen, Kevan

Predicting Medical Insurance Charges & Analysing Key Factors



Problem Framing

- Rising medical costs make it vital for insurers to **predict individual expenses accurately**.
- **Age, BMI, smoking, and region** are key drivers of these charges.
- Many models overlook **interaction effects** (e.g., smoking × BMI) and **fairness** across gender and region.
- A **fair, interpretable model** enables insurers to design **equitable and sustainable premiums**.



Objectives

- **Predict** medical insurance charges using regression models (Linear Regression, Random Forest, xgBoost).
- **Identify key drivers** of medical costs through feature-importance and interaction analysis.
- **Evaluate model fairness** by comparing prediction residuals across **gender and region**.
- **Deliver insights** that are interpretable and actionable for both technical and policy stakeholders.

Method Flow:

Problem
Objective

Exploratory
Data Analysis

Data
Cleaning

Model Building
and Fine Tuning

Analysing Models'
Performance

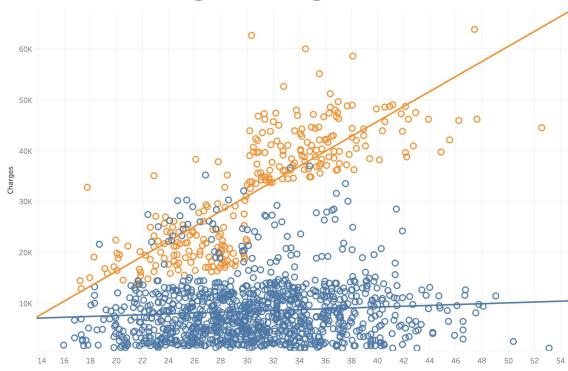
Feature Importance
& Fairness Analysis

Recommendations
& Challenges

EDA 1- Visualisation

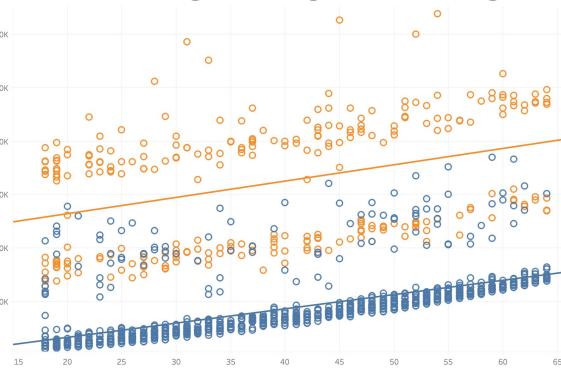
Smoker
no
yes

Charges Against BMI by Smoker



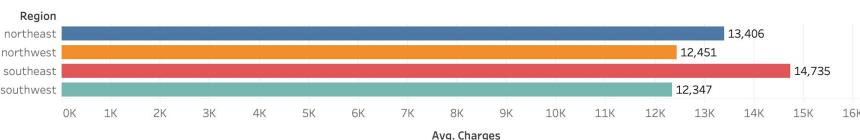
Medical charges rise with BMI, but the effect is far stronger among smokers, indicating that smoking amplifies the cost impact of higher BMI

Charges Against Age by Smoker



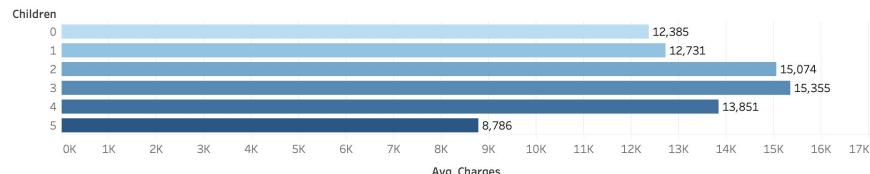
Charges increase with age for both groups, but smokers consistently incur much higher costs at every age.

Average Charges by Region



Average charges are similar across all regions, suggesting that region has minimal impact on medical costs.

Average Charges by Number of Children



Average charges generally rise with the number of children up to three, before slightly declining, suggesting a modest, non-linear relationship.

EDA 2 – Data Preprocessing

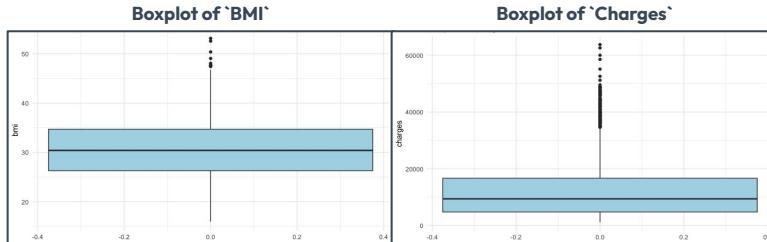
Initial Data Inspection

```
spc_tbl_ [1,338 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
$ age      : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
$ sex       : chr [1:1338] "female" "male" "male" "male" ...
$ bmi       : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
$ children: num [1:1338] 0 1 3 0 0 1 3 2 0 ...
$ smoker    : chr [1:1338] "yes" "no" "no" "no" ...
$ region   : chr [1:1338] "southwest" "southeast" "southeast" "northwest" ...
$ charges   : num [1:1338] 16885 1726 4449 21984 3867 ...
```

Checked for:

- Missing Values (None)
- Duplicates (None)
- Data Types (Numeric vs Categorical)

Outlier Detection



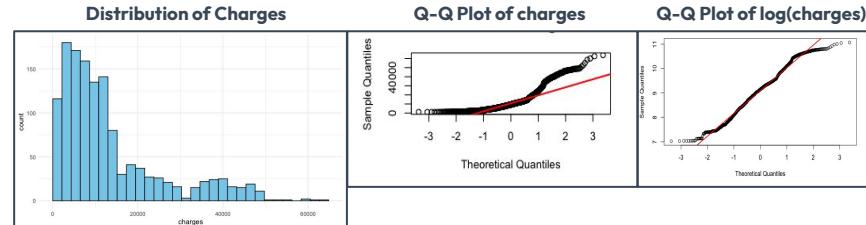
‘Bmi’ has 9 outliers while ‘Charges’ has 139 outliers. We do not omit them as the values represent individual with extreme obesity and real-high cost cases respectively.

Feature Encoding

Encoding converts categorical data into interpretable numeric form without introducing bias.

Variable	Method	Example
`sex`	Binary Encoding	Female → 0, Male → 1
`smoker`	Binary Encoding	No → 0, Yes → 1
`region`	One-hot Encoding	Northeast, Northwest, Southeast, Southwest → 4 dummy columns

Feature Transformation



The original ‘charges’ variable was highly right-skewed due to a few extremely high medical costs; applying a **log transformation** normalized the distribution, reduced heteroskedasticity, and improved the linearity and stability of relationships for regression-based and tree-based prediction models.

Model Development & Optimization

Linear Regression

Linear regression was chosen as a transparent and interpretable baseline to understand how demographic and lifestyle factors drive medical costs.

Model Used (AIC-based): charges_log ~ age * smoker_encoded + bmi * smoker_encoded + children + sex_encoded + region_northeast + region_northwest

Interactions Terms Used:

age * smoker_encoded:

reflects how the consequences of smoking often worsen with time.

BMI * smoker_encoded:

captures how two major health risk factors, obesity and smoking, can amplify each other's impact.

Regression Coefficients

```
## Call:  
## lmformula = charges_log ~ age + smoker_encoded + bmi + children +  
##   sex_encoded + region_northeast + region_northwest + age:smoker_encoded +  
##   smoker_encoded:bmi, data = train)  
##  
## Residuals:  
##    Min      1Q  Median      3Q     Max  
## -0.67388 -0.15080 -0.06398 -0.02302  2.37075  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.9447295  0.0814500 85.264 < 2e-16 ***  
## age         0.0412632  0.0009006 42.079 < 2e-16 ***  
## smoker_encoded 1.2036784  0.1707498  7.049 3.24e-12 ***  
## bmi        0.0003952  0.0023211  0.170 0.864837  
## children   0.1090631  0.0101854 10.788 < 2e-16 ***  
## sex_encoded 0.0863229  0.0244499  3.531 0.000432 ***  
## region_northeast 0.1181796  0.0307292  3.846 0.000127 ***  
## region_northwest 0.0825317  0.0301953  2.733 0.0006375 **  
## age:smoker_encoded -0.0339625  0.0021742 -15.621 < 2e-16 ***  
## smoker_encoded:bmi  0.0539159  0.0049723 10.843 < 2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 0.3961 on 1059 degrees of freedom  
## Multiple R-squared:  0.8109, Adjusted R-squared:  0.8092  
## F-statistic: 504.4 on 9 and 1059 DF, p-value: < 2.2e-16
```

Model Performance

Training Set Performance:

RMSE: 0.3942
MAE: 0.2151
R²: 0.8109

Test Set Performance:

RMSE: 0.346
MAE: 0.2053
R²: 0.8719

Multiplicative Error (\pm): 41.33 %

Strengths vs Limitations

Strengths	Limitations
Simple, interpretable coefficients	Cannot capture non-linear patterns
Captures key effects (e.g., smoker \times BMI)	Sensitive to outliers
High R ² (0.81) — strong explanatory power	Assumes constant variance (violated slightly)

Conclusion: Linear Regression is a strong baseline model (R² = 0.81), but residual patterns suggest potential gains from non-linear ensemble models.

Model Development & Optimization

Our goal is to predict **individual medical insurance charges** based on demographic and lifestyle attributed (age, BMI, smoking status, region) from a dataset of **1,338 records**. These features exhibit nonlinear relationships and interactions which simpler linear models struggle to capture.

Random Forest Regressor

Model Overview: Random Forest is an ensemble of decision trees that averages multiple predictions to reduce variance and improve generalisation. Each tree is trained on a bootstrapped subset of data and a random subset of features.

Why Random Forest fits the problem:

- Captures nonlinear dependencies between health and lifestyle factors effectively
- Reduces overfitting in a small dataset (1,338 rows) through averaging across multiple trees
- Naturally handles both numerical and categorical variables, minimising preprocessing

XGBoost Regressor

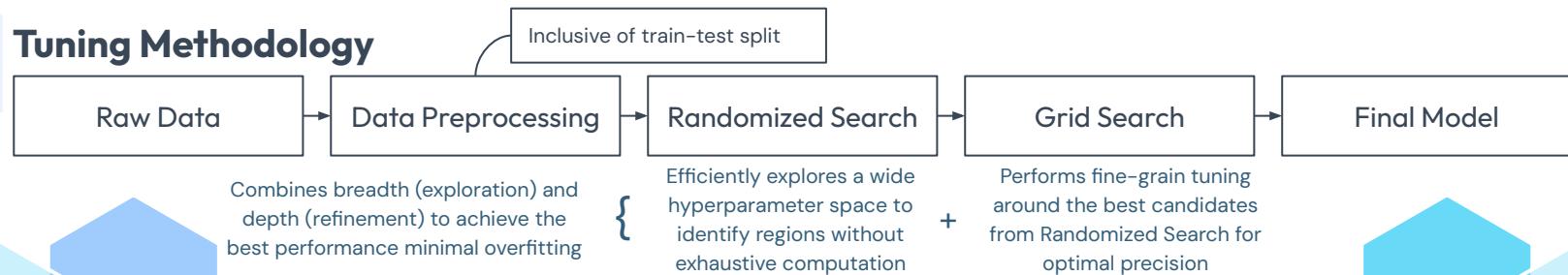
Model Overview: XGBoost (eXtreme Gradient Boosting) is an optimised learning algorithm that builds numerous smaller decision trees sequentially, where each new tree focuses on correcting the errors of the previous ones. It is designed for speed and scalability, with regularisation to prevent overfitting.

Why XGBoost fits the problem:

- Is able to automatically model nonlinear relationships and feature interactions without manual feature engineering
- Regularisation and robustness of enables XGBoost to model the dataset while avoiding underfitting or overfitting.

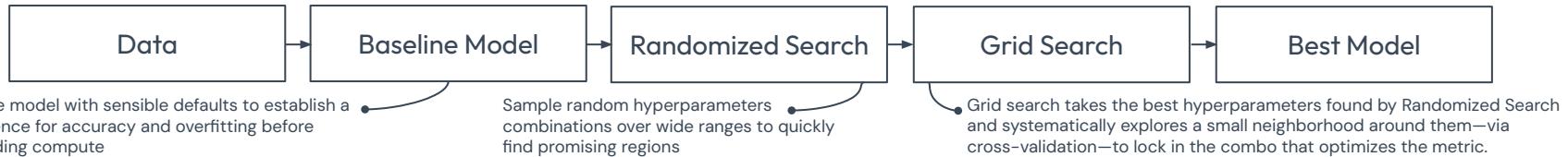
These reasons make Random Forest and XGBoost an ideal balance between accuracy, robustness, and interpretability for our dataset.

Tuning Methodology



Tuning Methodology

A Repeated 5-Fold Cross-Validation (5 splits × 2 repeats) was used to ensure the model's performance is robust and consistent across different data partitions.



Scoring choice: Optimised neg RMSE

Data pipeline note: Train (80%)/test(20%) split first; all tuning done on **log(charges)** to stabilise variance; no leakage from test set

Categorical features are encoded up front; interactions (smoker x BMI, smoker x age) included

Objective: Minimise Test RMSE on **log(charges)** to penalise large dollar errors while stabilising variance (higher test R²); keep **Train-Test gap small** (generalisation)

Random Forest Regressor

RandomizedSearchCV (broad):

- **Capacity controls:** max_depth, min_samples_split, min_samples_leaf → limit tree size and prevent overfit.
- **Variance controls:** max_features, bootstrap, max_samples → inject randomness and reduce variance.
- **Stability:** n_estimators → more trees = smoother predictions (diminishing returns past ~200)

GridSearchCV (local refine):

3-point neighbourhood around each best Randomized Search knob (± 1 or ± 2 for integers; ± 100 for n_estimators) to lock in an optimal bias-variance mix with far fewer fits.

Best Parameters:

```
{'bootstrap': True, 'max_depth': 10, 'max_features': 0.6, 'min_samples_leaf': 8, 'min_samples_split': 11, 'n_estimators': 200}
```

XGBoost Regressor

Parameters:

The following parameters were chosen to control the model's complexity, learning speed, and regularisation, thereby balancing underfitting and overfitting. (To capture nonlinear patterns without memorising noise)

- **n_estimators**: The number of boosting rounds (trees)
- **max_depth**: Maximum depth of each tree
- **learning_rate**: Shrinks feature weight to make each step more conservative
- **subsample/colsample_bytree**: Respectively the fraction of training data/features randomly sampled for each tree
- **min_child_weight**: Minimum number of instances needed to be in each node
- **gamma**: A regularisation term where higher values lead to more a conservative algorithm

Best Parameters:

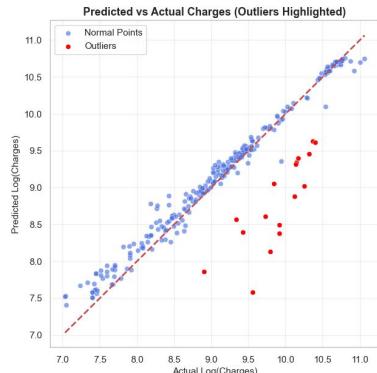
n_estimators: 100
max_depth: 3
learning_rate: 0.1

subsample: 1.0
colsample_bytree: 0.8

min_child_weight: 3
gamma: 0.2

Tuning Insights & Results

Random Forest Regressor



Blue points mostly lie close to the line → the tuned Random Forest model fits the bulk of cases well.

Red points (flagged outliers) cluster at high actual costs ($\approx 9.5\text{--}11 \log$) and lie below the line → the model systematically under-predicts very expensive cases (large positive residuals).

This tail asymmetry is consistent with the metrics (RMSE > MAE) and our fairness read (slide 9) where certain subgroups (e.g., Region-1 / males) show heavier tails.

Metric	Before Tuning	After Tuning
Train R ²	0.971	0.867
Test R ²	0.845	0.882
Train RMSE	0.154	0.331
Test RMSE	0.379	0.330
Train MAE	0.080	0.170
Test MAE	0.194	0.179

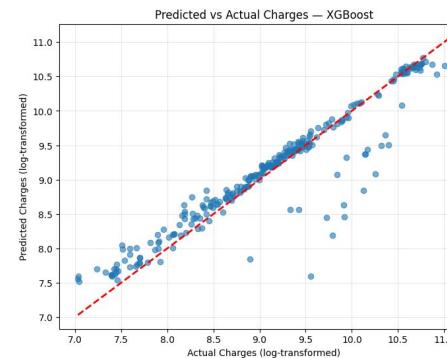
Before tuning: Very high Train R² (0.971) vs lower Test R² (0.845) ⇒ clear overfitting; the model memorized the train set.

After tuning: Train R² drops to 0.867 while Test R² rises to 0.882 ⇒ the gap narrows and generalization improves.

Error metrics: Test RMSE improves (0.379 → 0.330) and Test MAE improves (0.194 → 0.179); Train RMSE/MAE increase (0.154 → 0.331, 0.080 → 0.170) indicate stronger regularization rather than memorization.

Bottom line: The tuned Random Forest trades a bit of train fit for better, more stable test performance, which is what we want.

XGBoost Regressor



Majority of points cluster around the line of prediction, indicating strong agreement between predicted and actual *charges_log*. The consistency of error is good, as shown by the symmetricality of points around the line.

However, some outliers fall below the line, indicating the model under-predicts cases with high charges.

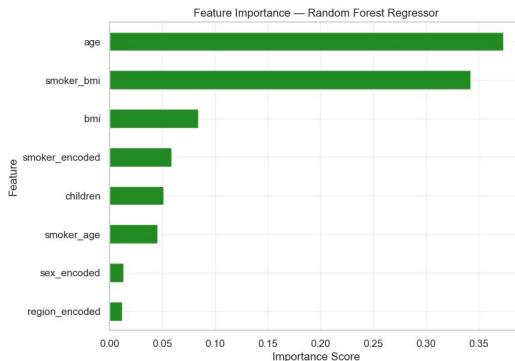
Metric	Initial Model	Tuned Model
0	RMSE	0.448239
1	R2	0.783535
2	MAE	0.231135

From the initial to tuned model, all 3 metrics show improvement (+0.101 R², -0.121 RMSE, -0.0478 MAE), indicating GridSearchCV was effective in tuning hyperparameters to better fit XGBoost to the data.

Metric	Train Value	Test Value
0	RMSE	0.350303
1	R2	0.850673
2	MAE	0.181422

Generally, the tuned model performs slightly better on unseen test data than train data except in MAE (+0.0338 R², -0.0228 RMSE, +0.00191 MAE). The estimated 1% increase in MAE is negligible, indicating that the model is likely not overfit and generalises well to unseen data.

Feature Impact & Key Insights (RF & XGBoost)



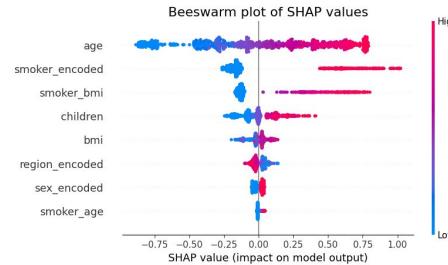
Top drivers: Age is the strongest predictor, followed closely by the smoker×BMI interaction term. These two features explain the bulk of predictive signal. BMI alone is a mid-tier driver.

This guides actionable recommendations on health risks and premium setting.

Secondary effects: Smoking status and children contribute modestly; the smoker×age interaction is smaller but non-zero.

Charges rise steeply with age and are especially high for smokers with higher BMI; reducing smoking and weight management likely has the greatest effect on reducing insurance charges

The Beeswarm plot of SHAP values shows the distribution of SHAP values for each feature across the given dataset, plotting both direction and value of the feature on insurance cost outcomes.



age: Age has a clear positive relationship with charges_log.

smoker_encoded: A distribution of high positive and less negative SHAP values. This demonstrates that smoking status is a binary trigger for a disproportionate increase in predicted insurance cost.

smoker_bmi: SHAP values have a similar distribution to smoker_encoded, indicating that BMI is more significant for a smoker's insurance cost than BMI alone.

children, bmi, region_encoded, sex_encoded, smoker_age: These features are clustered more tightly around 0.0, confirming that they have relatively less influence.

Final Model Chosen: Random Forest Regressor (RF)

Handles nonlinearity & interactions: Random Forest models interaction terms without manual feature engineering, unlike Linear Regression which assumes linear/additive relations.

Best generalization on our data size (~<1.5k rows): Tuned RF high Test R² and low Test RMSE with a small train–test gap, showing strong generalization and stability. XGBoost showed signs of overfitting and instability due to its higher sensitivity to hyperparameters on a smaller dataset.

Robust & explainable: RF is resistant to outliers unlike Linear Regression which can be heavily skewed by them..

In conclusion, Random Forest was selected as the final model because it provided the best trade-off between predictive accuracy, generalization, robustness and explainability, outperforming both Linear Regression (too simplistic) and XGBoost (too sensitive) for the dataset size and characteristics.

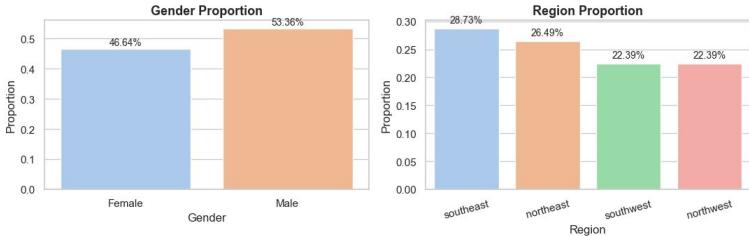
Fairness Analysis (Random Tree)

Scope

Protected Attributes (A): Gender (Female/Male) and Region(NE/NW/SE/SW)

Representation Check: Group proportions are balanced enough for comparison

Group Representation in Test Dataset



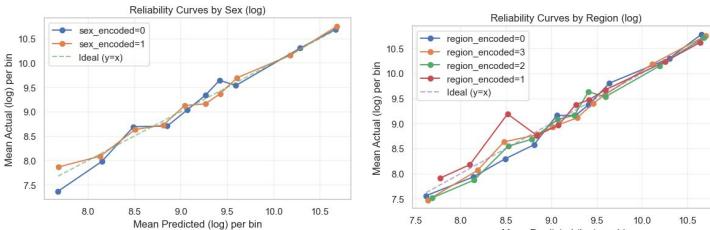
Framework

Independence (Statistical/Demographic Parity): predictions should not depend on (A)

Separation (Error-rate Parity): error behavior should not depend on (A) given the true outcome

Sufficiency (Calibration/Conditional use accuracy): for a given prediction score \hat{y} , the true outcome distribution should be similar across groups.

Sufficiency:



Reliability by Region (log): All regions track the diagonal closely; Region 1 deviates upward at lower predicted bins (under-prediction) but aligns well at mid-high ranges.

Reliability by Sex (log): Both sexes are well-calibrated overall; sex=0 is slightly below the diagonal at the lowest bins (mild over-prediction) and matches closely thereafter.

Independence:

Region leakage: `bmi` has the strongest mutual information (MI) with `region`, followed by `children` and then `age` (proxy risk)

Sex leakage: only a small MI for `children` with `sex`

After conducting a permutation test at 5% significance level ($B=500$), BMI's MI_obs (0.0745) exceeded the perm_95pct threshold (0.0614, $p=0.020$), hence the **dependence is statistically significant**.

We further conducted the same test for `age`, `children`, and `smoker`; their $MI_{obs} < perm_{95pct}$ ($p>0.05$), indicating **non-significant dependence**.

To tackle this:

We keep BMI (it's predictive), but add guardrails:

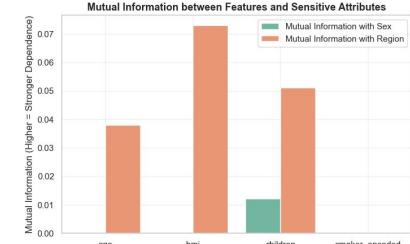
Audit by region: keep monitoring mean residuals, MAE/RMSE, and calibration.

Stratify CV by region to avoid region-specific overfitting.

Separation:

Errors are compared **by group** (mean residual, MAE, RMSE, RMSE/MAE). Parity indicates separation fairness; gaps point to groups with **systematic bias or heavier tails**

Model is roughly fair by sex on average errors but shows **heavier tails for males** (higher RMSE/MAE); Region-1 is under-predicted with the highest MAE/RMSE, requiring region-specific calibration/interaction terms.



feature	MI_obs	perm_95pct	p_value
age	0.008724	0.071042	0.416
bmi	0.074517	0.061379	0.020
children	0.035420	0.050272	0.334
smoker_encoded	0.001788	0.015994	0.802

Scope	Group	Overall Error Parity on log(charges) scale				
		n	Mean Residual (log)	MAE (log)	RMSE (log)	RMSE/MAE
Sex	0	125	-0.029	0.181	0.294	1.62
	1	143	0.011	0.177	0.359	2.03
Region	0	60	-0.037	0.136	0.245	1.80
	1	71	0.075	0.246	0.439	1.78
	2	77	-0.043	0.185	0.330	1.78
	3	60	-0.029	0.135	0.244	1.80
						14.5%

Practical Recommendations

Personalised Premium Pricing

- Design data-driven premium tiers reflecting individual health risks
 - Older smokers with higher BMI assigned higher premiums
 - Healthier non-smokers enjoy discounts
- Improve pricing fairness and sustainability

Preventive Health Interventions

- Insurers offer targeted wellness programmes to reduce long-term claims
 - Observed **smoking & BMI** are **top cost-drivers** (especially when combined)
 - Offer weight management, smoking cessation subsidies

Policy & Operation Insights

- Policymakers focus subsidies or public campaigns on high-risk demographic groups
- Hospitals and clinics use the model to anticipate patient cost burdens and design preventive outreach programmes

Difficulties Faced

Data Size Limitations

- Dataset only had 1,338 rows, limited model generalisation
- Small sample size makes models sensitive to outliers
- Less representative of real-world diversity

Feature Interactions

- Capturing interaction effects was challenging
- Required manual scatterplots between multiple variables
- May not be feasible in a larger dataset with more variables to consider

Appendix

Link to Code: https://github.com/joannekzx/predict_medical_insurance_charges.git