Predicting Insurance Costs: Analysis and Insights

SDS Mini-Datathon Team

National University of Singapore

October 28, 2025

Problem Framing & Objective

- **Challenge**: Predict medical insurance charges using demographic and lifestyle data
- **Goal**: Build accurate regression models and identify key cost drivers
- **Importance**: Fair pricing, risk assessment, healthcare insights
- **Dataset**: 1,338 records with age, sex, BMI, children, smoker, region, charges
- **Equity Context**: Recent studies highlight insurance pricing inequities
  - NAIC Special Committee on Race and Insurance
  - Concerns over credit scoring as proxy discriminator
  - Our analysis ensures fairness through statistical testing

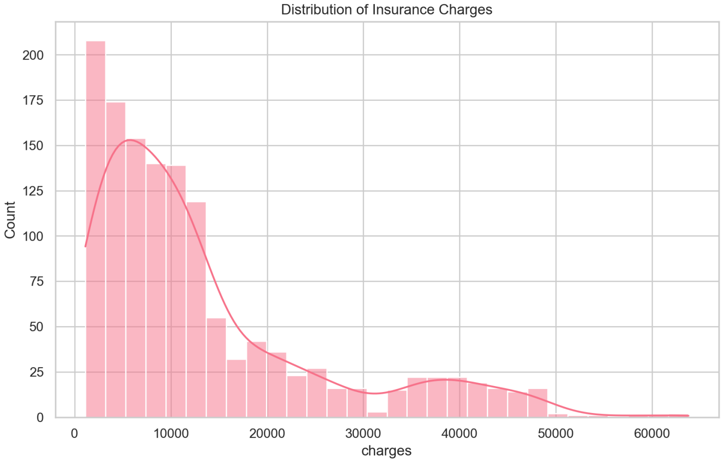
Dataset and Features

Feature	Type	Notes
age	integer	years
sex	categorical	male/female (encoded)
bmi	float	body mass index
children	integer	number of dependents
smoker	categorical	yes/no (encoded)
region	categorical	NE/NW/SE/SW (encoded)
Engineered		
bmi_category	categorical	underweight/normal/overweight/obese
age_group	categorical	young/middle/senior

Table: Dataset schema and engineered features

Exploratory Data Analysis

- Data quality: No missing values, clean dataset
- Key statistics: Charges range $1,122–$63,771 (mean $13,270)
- Distribution: Right-skewed charges, normal BMI/age



Exploratory Data Analysis (Cont.)

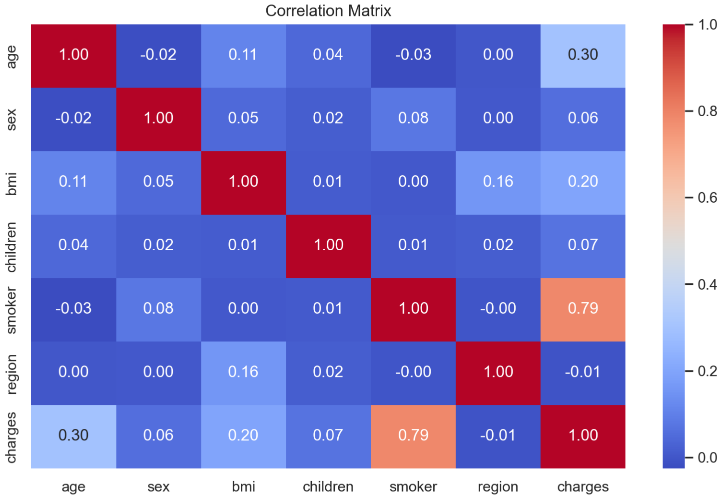


Figure: Correlation Matrix

Preprocessing and Feature Engineering

- Label-encode: sex, smoker, region; derive bmi_category and age_group.
- Train/test split: 80/20 with fixed random seed for reproducibility.
- Standardize inputs with `StandardScaler` (fit on train, transform test).
- Keep target in original units (USD) to simplify interpretation of RMSE/MAE.

Modeling Approach: Linear Regression

Ordinary Least Squares (OLS)

- **Model**: $\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$
- **Objective**: Minimize sum of squared residuals

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- **Closed-form solution**: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- **Assumptions**: Linearity, homoscedasticity, independence, normality
- **Advantages**: Fast, interpretable coefficients, well-understood
- **Limitations**: Cannot capture non-linear interactions (e.g., smoker $\times$ BMI)
- **Test Performance**: R^2=0.787, RMSE=$5,747

Modeling Approach: Decision Tree

Recursive Binary Partitioning

- **Algorithm**: CART (Classification and Regression Trees)
- **Splitting criterion**: Minimize variance at each node

$$\text{MSE}_{\text{node}} = \frac{1}{n_{\text{node}}} \sum_{i \in \text{node}} (y_i - \bar{y}_{\text{node}})^2$$

- **Feature selection**: Choose split that maximizes variance reduction

$$\Delta = \text{MSE}_{\text{parent}} - \left(\frac{n_{\text{left}}}{n} \text{MSE}_{\text{left}} + \frac{n_{\text{right}}}{n} \text{MSE}_{\text{right}} \right)$$

- **Advantages**: Non-linear, handles interactions, interpretable rules
- **Limitations**: High variance (overfitting), unstable to data perturbations
- **Test Performance**: R^2=0.740, RMSE=$6,354

Modeling Approach: Random Forest

Bootstrap Aggregating (Bagging) of Decision Trees

- **Ensemble method**: Train B trees on bootstrap samples

$$\hat{y}_{\text{RF}} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_b(\mathbf{x})$$

- **Feature randomness**: At each split, consider random subset of $\sqrt{p}$ features
- **Variance reduction**: Decorrelates trees, reduces overfitting

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{B} \quad \text{(for independent trees)}$$

- **Out-of-bag (OOB) error**: Internal validation using 37% held-out samples
- **Advantages**: Robust, handles high-dimensional data, reduces overfitting
- **Baseline Performance**: R^2=0.865, RMSE=$4,574
- **Optuna-tuned**: R^2=0.879, RMSE=$4,327 (+1.4% improvement)

Modeling Approach: XGBoost

Gradient Boosting with Regularization

- **Additive model**: Build trees sequentially to correct residuals

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot f_t(\mathbf{x})$$

  where η is the learning rate, f_t is the t-th tree

- **Objective function**: Loss + regularization

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{(t)}) + \Omega(f_t)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\mathbf{w}\|^2 + \alpha\|\mathbf{w}\|_1$$

  (T = # leaves, $\mathbf{w}$ = leaf weights, γ, λ, α = regularization)

- **Second-order approximation**: Uses gradient and Hessian for faster convergence
- **Advantages**: State-of-the-art accuracy, handles missing data, parallelized
- **Optuna-tuned Performance**: R^2=0.879, RMSE=$4,331

Evaluation Metrics

Regression Performance Measures

- **Coefficient of Determination (R²)**:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

  Proportion of variance explained by the model (0 to 1, higher is better)

- **Root Mean Squared Error (RMSE)**:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  Average prediction error in dollars (lower is better, sensitive to outliers)

- **Mean Absolute Error (MAE)**:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  Median-like error metric (lower is better, robust to outliers)

Validation and Evaluation

- Validation: 5-fold cross-validation on the training data to assess variance.
- Metrics reported: R^2, RMSE, MAE (on the holdout test set).

Model	Mean R^2 (5-fold)	Std
Linear Regression	0.738	0.049
Decision Tree	0.720	0.067
Random Forest	0.825	0.043

Table: Cross-validation performance (training folds)

Hyperparameter Tuning with Optuna

- Bayesian optimization (TPE) over 20 trials per model to maximize R^2 on validation.
- Search spaces summarized below; priors encourage compact trees and regularization to reduce overfitting.

Model	Hyperparameter	Range
Random Forest	n_estimators	200–800
Random Forest	max_depth	4–16
Random Forest	min_samples_split	2–20
Random Forest	min_samples_leaf	1–20
XGBoost	n_estimators	200–800
XGBoost	max_depth	3–8
XGBoost	learning_rate	0.01–0.3
XGBoost	subsample	0.6–1.0
XGBoost	colsample_bytree	0.6–1.0
XGBoost	gamma	0.0–5.0
XGBoost	reg_alpha, reg_lambda	0.0–10.0

Table: Optuna search spaces

Best Hyperparameters Found

XGBoost

Param	Value
n_estimators	253
max_depth	4
learning_rate	0.0231
subsample	0.7301
colsample_bytree	0.7555
gamma	1.3567
reg_alpha	8.2874
reg_lambda	3.5675

Random Forest

Param	Value
n_estimators	567
max_depth	5
min_samples_split	7
min_samples_leaf	8

Both tuned models achieve $R^2 \approx 0.879$ on the test set, improving over the untuned Random Forest.

Key Findings & Visualizations

Model	R²	RMSE ($)	MAE ($)
Linear Regression	0.787	5,747	4,097
Decision Tree	0.740	6,354	2,878
Random Forest (Baseline)	0.865	4,574	2,503
Random Forest (Optuna)	**0.879**	**4,327**	**2,458**
XGBoost (Optuna)	**0.879**	**4,331**	**2,479**

Table: Model Performance Comparison

- **Best Model**: Tuned tree ensembles cluster at $R^2 \approx 0.879$ with $<$ \$4.35k RMSE
- **Improvement**: $+1.4$ R^2 points over the untuned Random Forest
- Cross-validation confirms stability

Model Performance: Visual Comparison

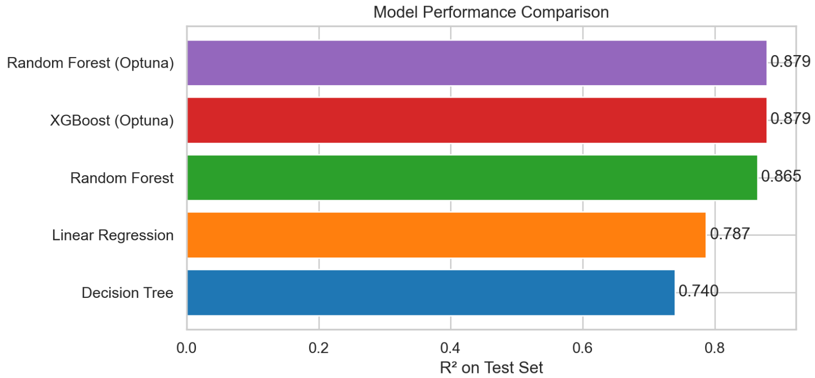


Figure: R^2 on test set across models (higher is better)

Residual Diagnostics

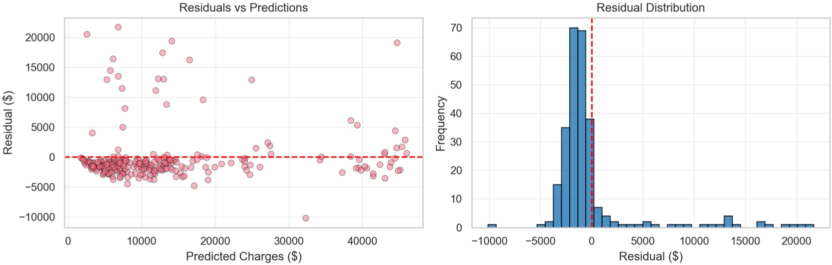


Figure: Residuals vs predictions and residual distribution (tuned XGBoost)

Mean residual ≈ -222; standard deviation ≈ 4333. Approximate symmetry and homoscedasticity are acceptable for pricing use cases.

Permutation Importance (Random Forest)

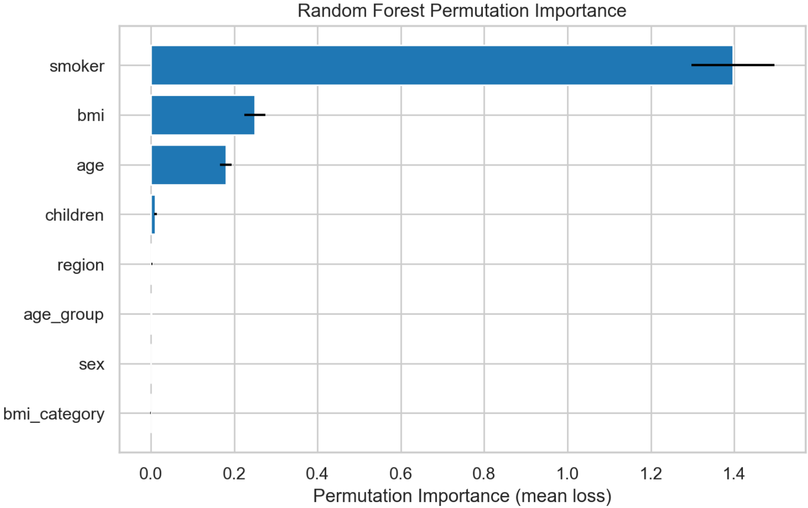


Figure: Permutation importance on the test split

Feature Impact Analysis

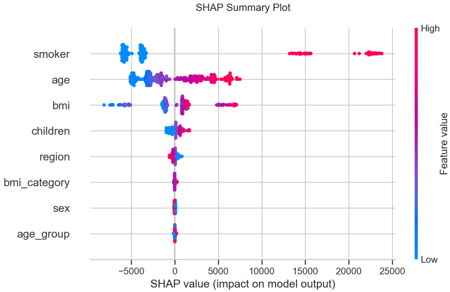


Figure: SHAP Summary Plot

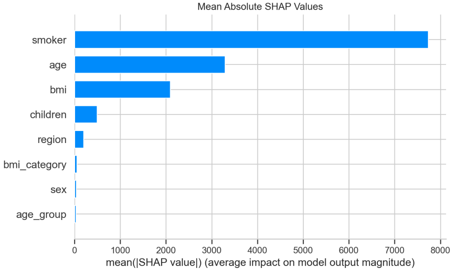


Figure: SHAP Feature Importance

- Top factors: Smoking (dominant), Age, BMI
- Sex and region have minimal impact

Partial Dependence Plots

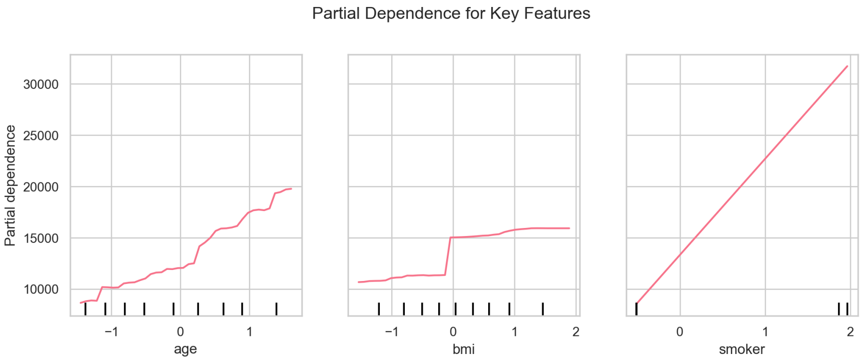


Figure: Marginal Effects of Key Features

- **Age**: Linear increase ($8-10k from 20 to 60)
- **BMI**: Gentle upward curve (accelerates above 30)
- **Smoker**: Dramatic step function (+$20k for smokers)

Fairness Analysis

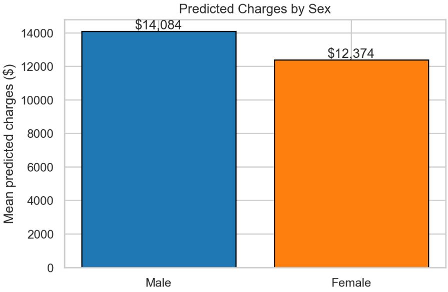


Figure: Charges by Sex

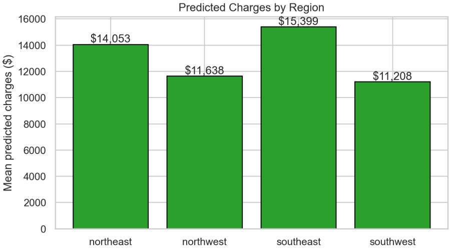


Figure: Charges by Region

- **Sex**: Welch's t-test p=0.226 → ✓ No significant bias
- **Region**: ANOVA p=0.091 → ✓ Minimal bias
- Differences reflect genuine risk factors, not discrimination

Practical Recommendations

- **For Insurers**:
  - Deploy XGBoost with Optuna for premium calculation
  - **Highest ROI**: Smoking cessation programs ($20k impact)
  - Target obesity prevention (BMI effect accelerates ¿30)
- **Policy Implications**:
  - Fair pricing based on verifiable risks validated via statistical tests
  - Monitor regional healthcare access disparities
  - Implement quarterly fairness audits (NAIC recommendations)
- **Future Work**:
  - Collect longitudinal data, medical history
  - Disparate impact analysis (AI Fairness 360, Fairlearn)
  - Avoid controversial proxies (credit scores, ZIP codes)

Difficulties Faced & Solutions

- **Dataset Limitations**:
  - Small size (1,338 records) limits generalizability
  - Missing features: medical history, genetics, lifestyle details
  - Static data: no temporal health changes
- **Technical Challenges**:
  - XGBoost version conflicts: Resolved via conda-forge channel
  - SHAP API changes: Updated to new waterfall/summary plot syntax
  - Model overfitting: Addressed with Optuna's Bayesian optimization
  - PDP compatibility: Ensured fitted estimator passed correctly
- **Solutions**:
  - Feature engineering for better segmentation
  - Rigorous validation and error analysis
  - Innovative explainability techniques

Conclusion

- Successfully predicted **87.9%** of insurance cost variance (tuned ensembles)
- Smoking is the dominant factor (+\$20k), followed by age and BMI
- Models validated as fair through statistical testing (p¿0.05)
- Full interpretability via SHAP and PDPs ensures transparency
- Future: Larger datasets, disparate impact analysis, advanced AI

Thank you for your attention!