Predicting Insurance Costs

stackX

Presented by Fabian Lim, Ear Hong Wee, Yi Jie
Team Name: StackX

US Health Insurance Prediction: A Data-Driven Approach

Our Objective

Build a predictive model to estimate individual insurance charges based on demographic and lifestyle factors.
The goal is to enable fairer pricing, transparent cost assessment, and data-driven risk evaluation.

Our Problem

Insurance premiums are often determined using broad assumptions about age, BMI, or smoking.
This analysis explores how each factor truly influences cost, uncovering both key drivers and potential disparities in pricing.

Our Approach

The analysis identifies key cost drivers and tests model fairness across demographics.

Clean Complete & Representative Data

Dataset Characteristics

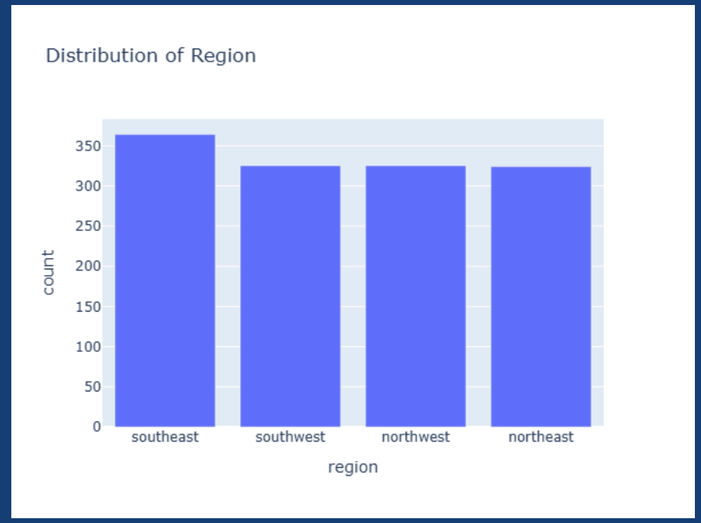
- 1,338 records, 7 columns (3 numerical, 4 categorical)
- Zero missing values – fully populated dataset
- All variables within logical, expected ranges

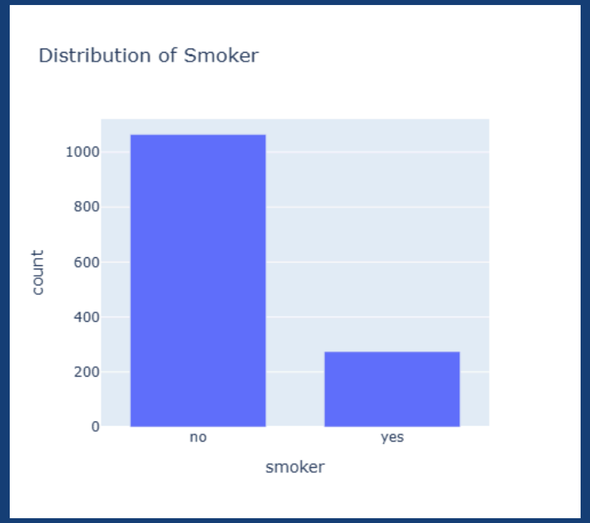
Sample Representativeness

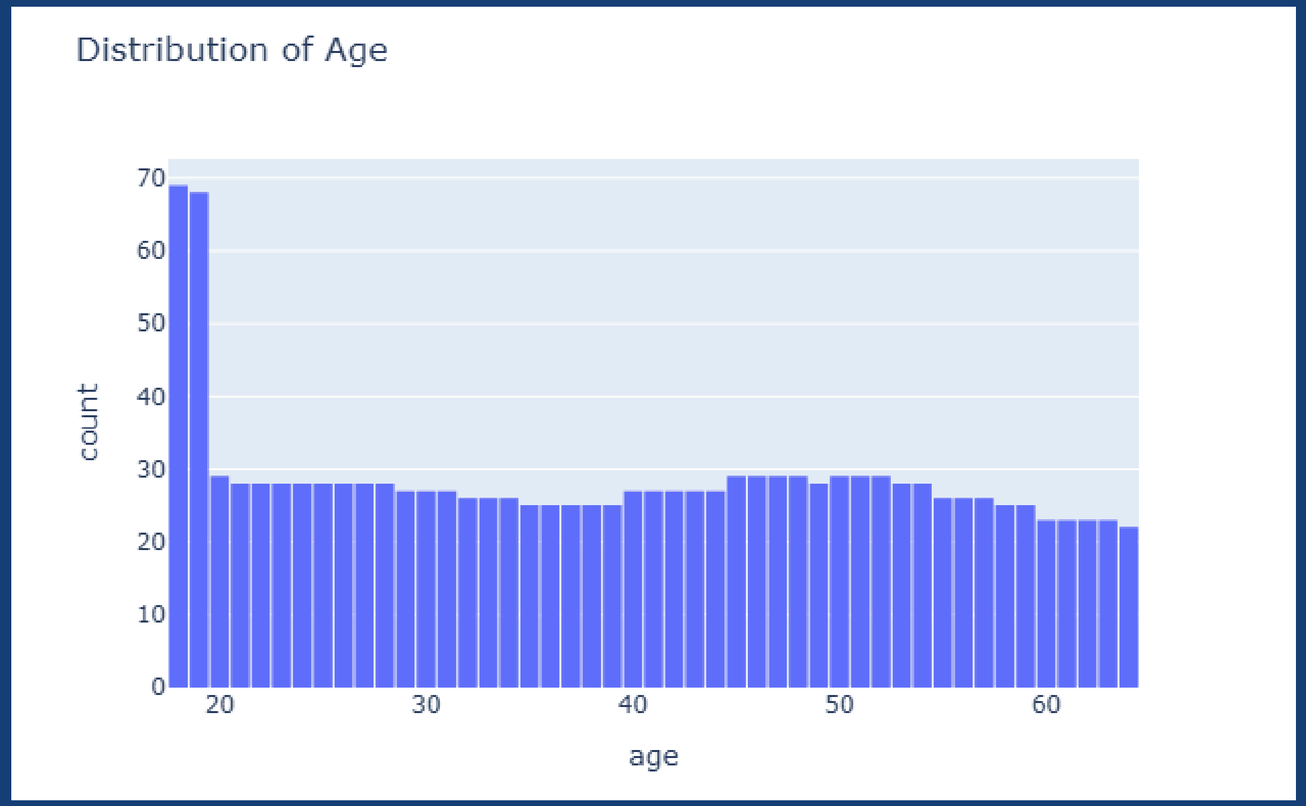
- **AGE:** Uniform distribution (18–64), spike at 18–19 (new insurance purchases)
- **GENDER:** Balanced (50.5% Male, 49.5% Female)
- **SMOKING RATE:** 20.5% (aligns with US national rate of 19.8%)
- **GEOGRAPHY:** Equal Representation across 4 regions

Data

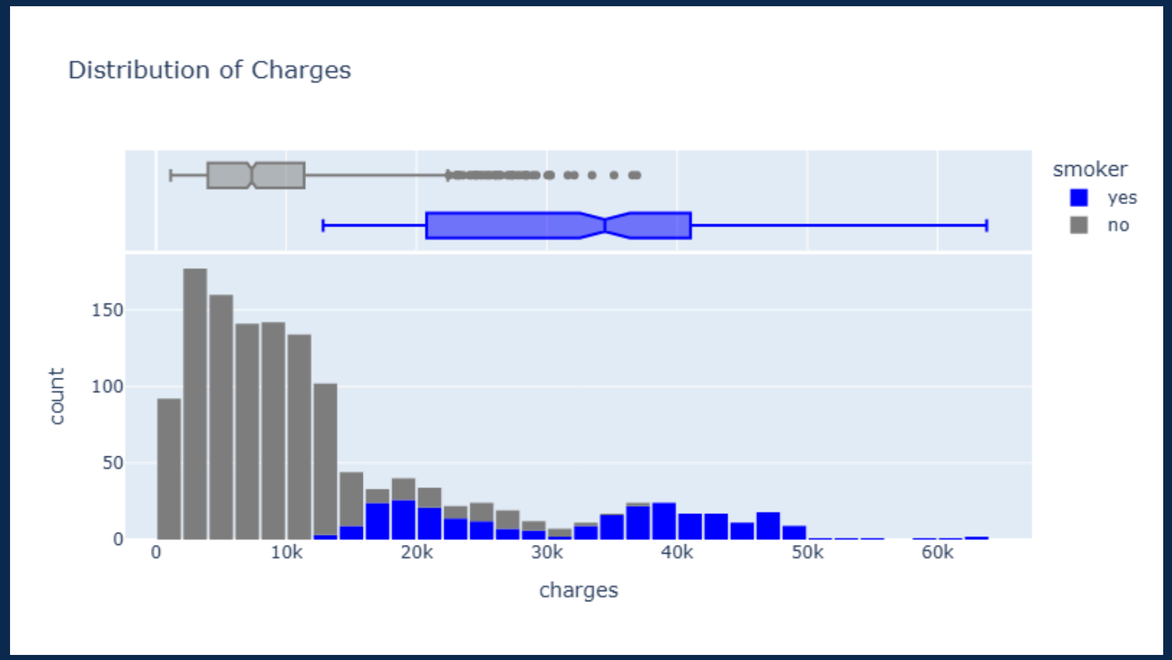
The dataset provided is clean, unbiased, analysis ready.





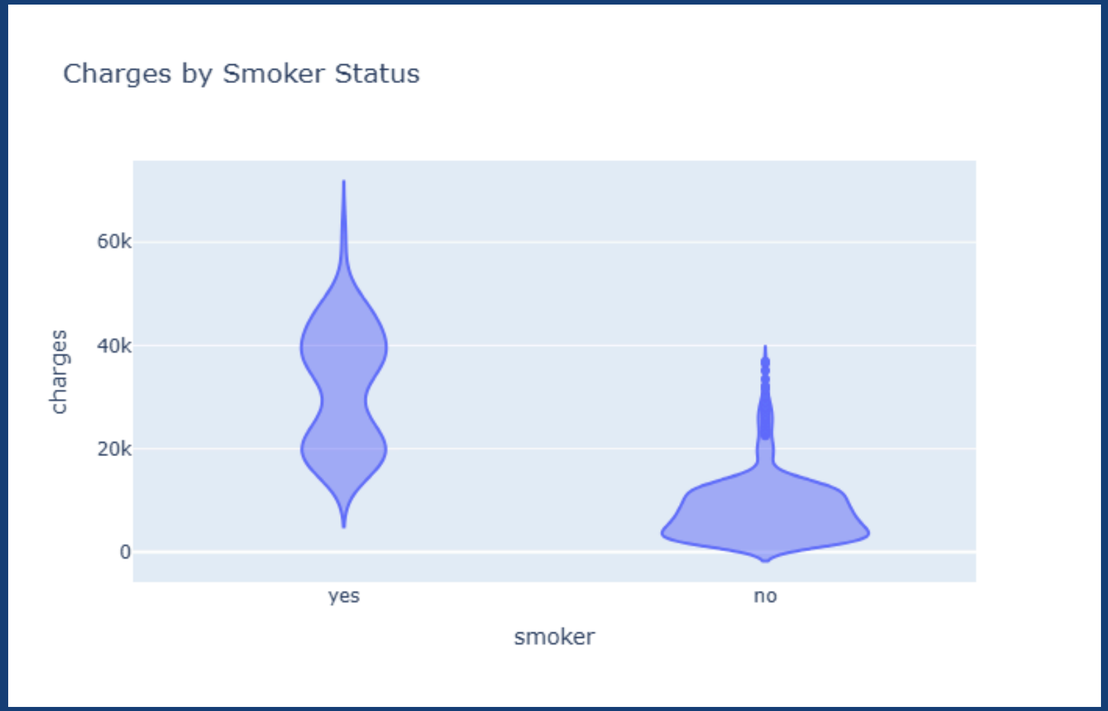


Distribution of Insurance Charges



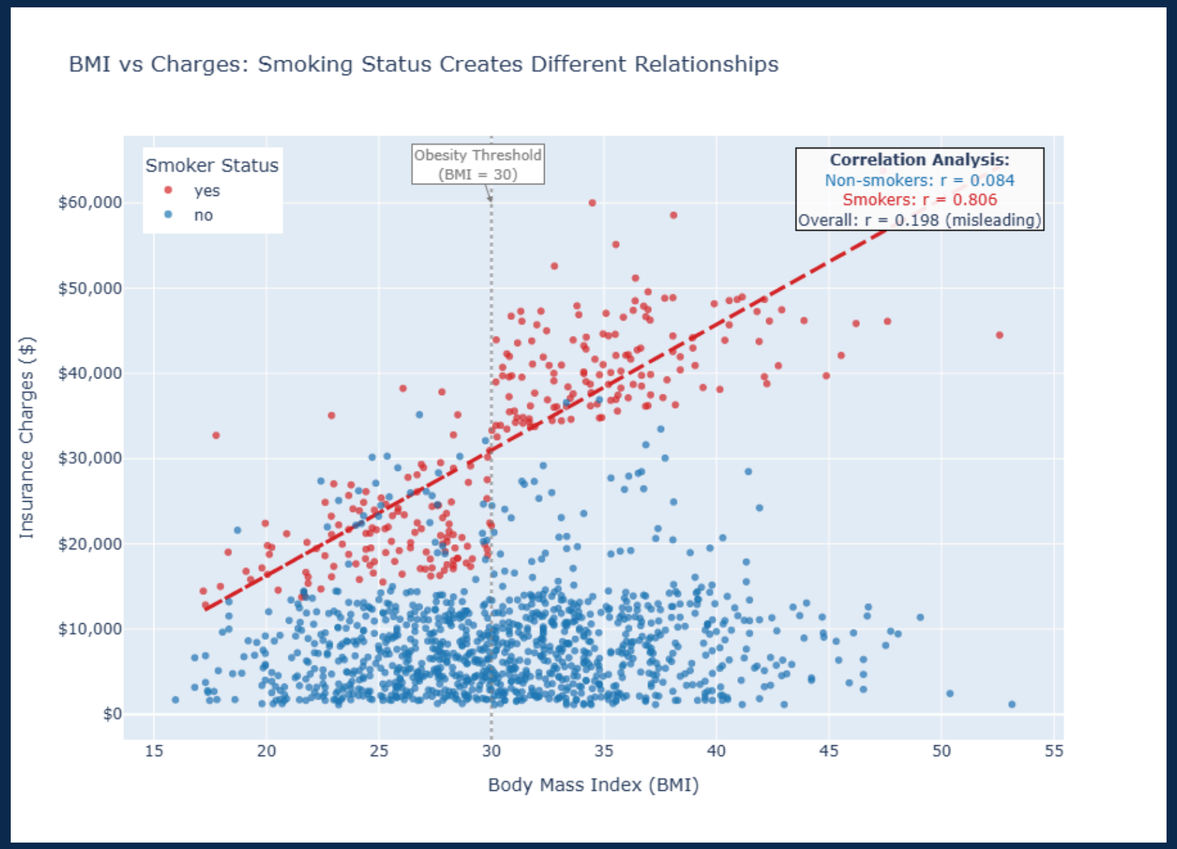
The bimodal distribution of charges for smokers suggests two sub-groups. To test the hypothesis that this is driven by BMI, smokers will be split into two groups: those above the median charge and those below within smokers. The mean BMI for each group will then be calculated and compared.

Smoking Effect: The Dominant Cost Driver



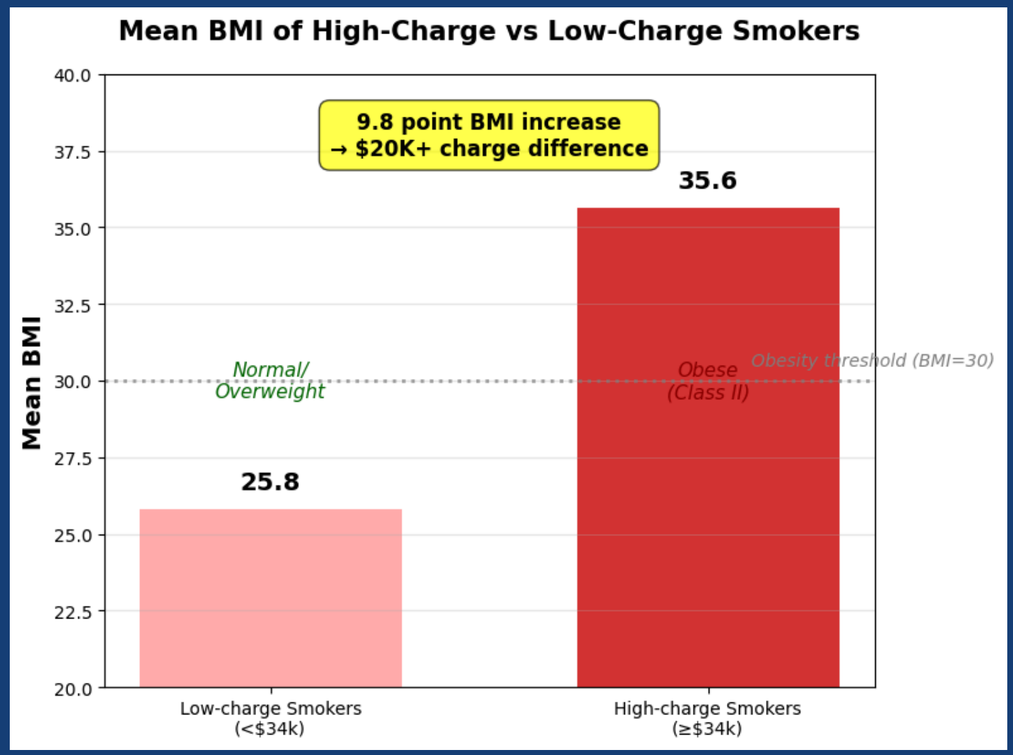
- Non-smokers: Median ≈ $8.4K, tightly concentrated (<$17K).
- Smokers: Median ≈ $34.5K, wide spread ($15K–$63K).
- Welch's t-test: $p < 0.001$, Cohen's $d = 2.57$ → exceptionally large effect.
- Smoking alone explains most of the cost variation observed in the dataset.

Interaction Effect – BMI x Smoking



Regression lines show that smoking amplifies BMI's impact on charges. Smokers have a strong positive correlation (r = 0.81), while non-smokers show a negligible one (r = 0.08). The overall correlation (r = 0.20) is misleading when smoker status is ignored.

BMI: The Key Driver of Cost Differences Among Smokers



High-charge smokers (≥ $34K) have a mean BMI of 35.6 (obese) versus 25.8 (normal) for low-charge smokers (< $34K).

A 9.8-point BMI increase corresponds to a $20K+ jump in insurance charges, highlighting a compounding BMI × Smoking effect.

Baseline Model

Linear Regression With All Features

Model Specifications

- Algorithm: Multiple Linear Regression
- Features: All 6 original features with sex, smoker, and region Binary/One-hot encoded + bmi x smoker interaction term.
- Train/Test Split: 80/20, stratified by smoker status
- Model Validation: Performed 5-fold cross-validation for more reliable measure of models performance

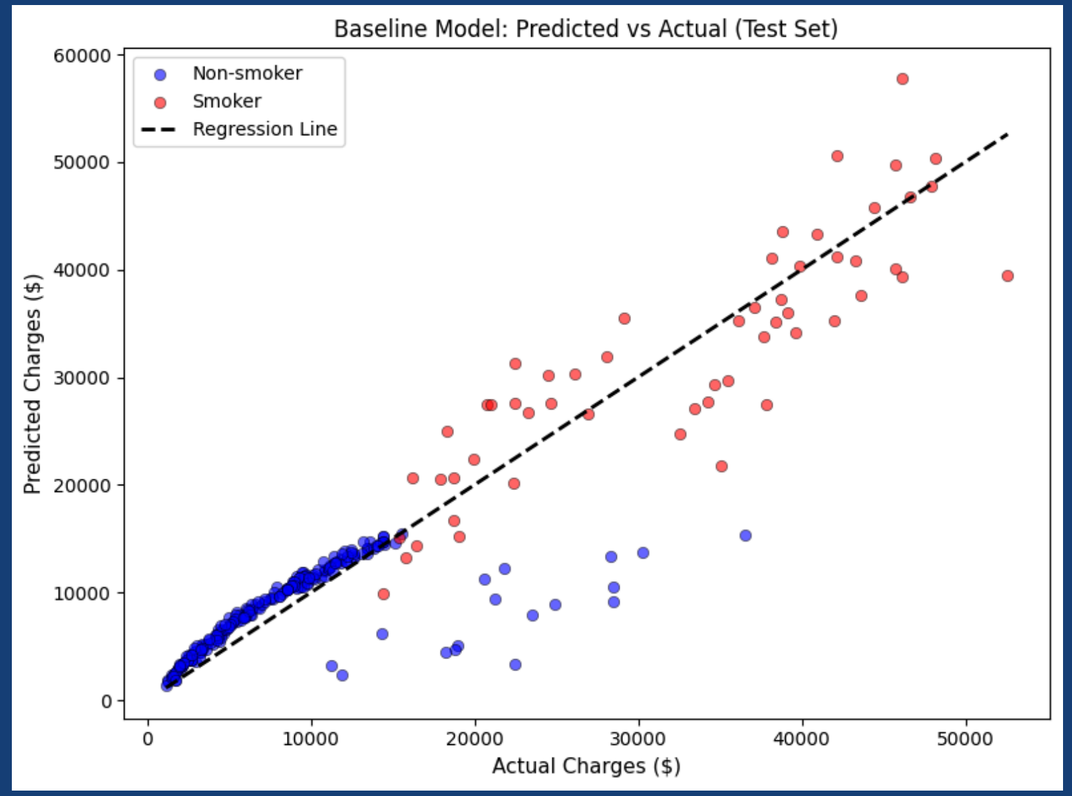
```
# Select features and target
feature_cols = ['age', 'bmi', 'children', 'smoker_encoded', 'sex_encoded',
                'region_northwest', 'region_southeast', 'region_southwest',
                'bmi_smoker_interaction']
```

Included Features

Performance metrics (cross validated on test set)

- R-squared (R^2): 0.837 → explains 83.7% of the variation in insurance charges
- Root Mean Square Error (RMSE): $4,850 → on average, model's prediction is about $4,850 different from actual charge.
- Normalised RMSE (RMSE / Mean Charges): 36.55% → average error is 36.55% of the mean charge, which provides contextualise the error's size

Residual Plot – Predicted vs. Actual



Key Findings

- The model is very accurate for non-smokers with charges <$15,000, as these blue points are tight along the line.
- The model is unreliable for smokers, with predictions (red points) widely scattered, indicating large errors.
- The model systematically under-predicts costs for non-smokers who have high charges (over $15,000).

Final Model

Linear Regression, SIGNIFICANTLY BETTER

Model Enhancements

1. Performed Backward Elimination to remove least significant predictors ('bmi' and 'region_northwest'), simplifying the model
2. Engineered a binary 'obese' variable (BMI >= 30) → Medical data shows that a BMI of 30 is the clinical threshold for obesity, which is associated with increased health risks and costs.
3. Created 'obese_smoker' – interaction term between the new 'obese' feature and 'smoker' → allows the model to capture the compounded (and much higher) impact on charges for individuals who are both obese and smokers, rather than just adding their separate effects.

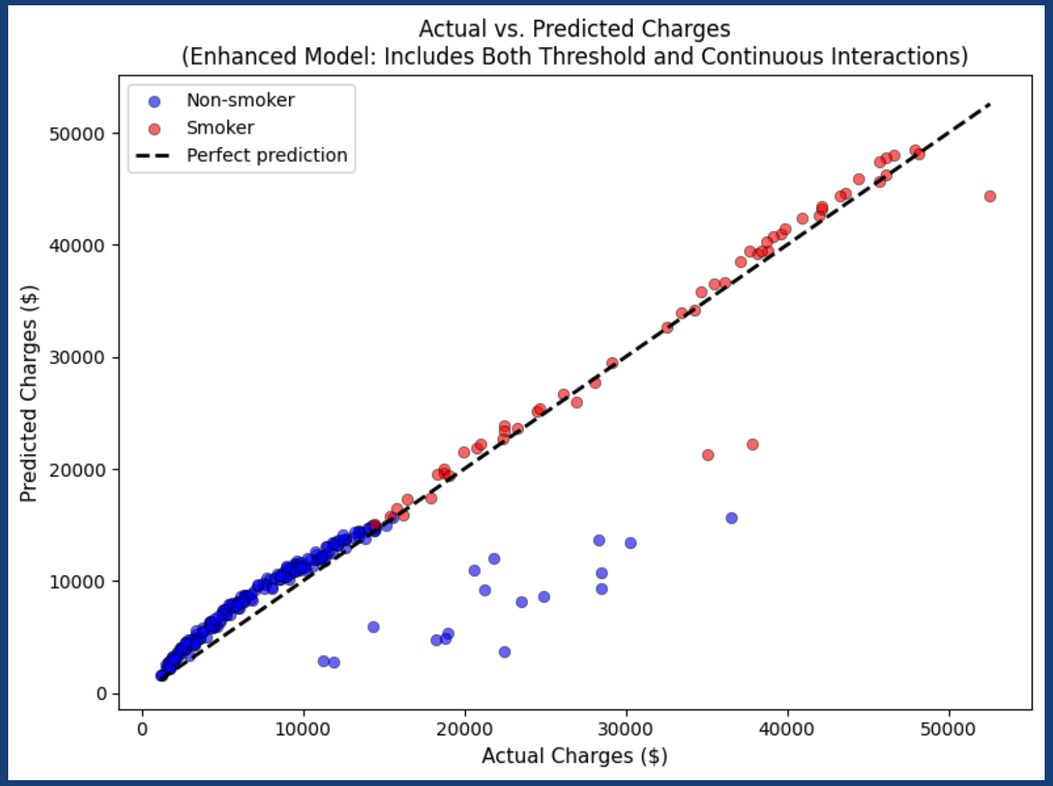
```
# create dummy variable and interaction term
df['obese'] = (df['bmi'] >= 30).astype(int)
df['obese_smoker'] = df['obese'] * df['smoker_encoded']

# include threshold interaction term 'obese_smoker', exclude continuous interaction term 'bmi_smoker_interaction'
feature_cols = ['age', 'bmi', 'children', 'smoker_encoded', 'sex_encoded',
                'region_southeast', 'region_southwest', 'obese_smoker', 'bmi_smoker_interaction']
```

Included Features

Performance Improvements (on test set)

- R-squared (R^2): 0.863 → Increased from 0.837. New model explains 2.6% more of the variation in charges.
- Root Mean Square Error (RMSE): $4,422 → Decreased from $4,850. The average prediction error was reduced by $428.
- Normalized RMSE: 33.32% → Decreased from 36.55%. new model's error is a smaller % of the mean charge, indicating better overall performance

Residual Plot – Predicted vs. Actual



Key Findings

- The model is now highly accurate and consistent for smokers → red points clustered tightly along the perfect prediction line – a major improvement from the baseline.
- The model remains very accurate for non-smokers with charges under $15,000 (the blue points on the line).
- Persistent Under-prediction: The model still has a clear weakness: it systematically under-predicts costs for a specific group of high-charge non-smokers (the blue points scattered below the line).

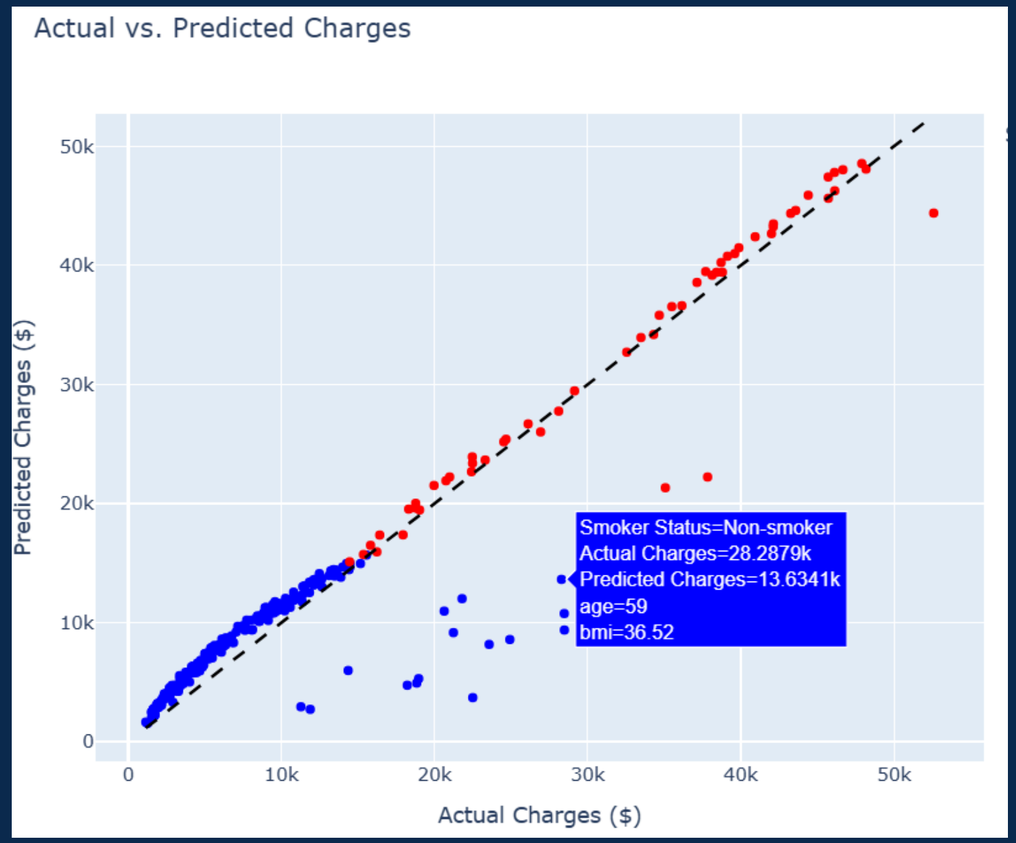
Key Insights

Insight 1 : Limitation of Dataset

- The model consistently under-predicts charges for a specific group of high-charges non-smokers.
- We investigated age and obesity as possible causes, but the data shows this is not the case → This group's obesity rate (52.5%) is nearly identical to the overall non-smoker obesity rate (52.8%).
- Key Insight: The model fails to predict these high costs because it is missing crucial features such as chronic diseases, or medication use.

Practical Recommendation

- To improve prediction accuracy, the model must include key features that identify specific health risks (like chronic conditions or medication use).



Insight 2 : Models Slight Overestimation

Based on the final models plot:

- Finding: The model has a slight, consistent tendency to over-predict charges (predictions are generally just above the perfect prediction line).

Practical Recommendation

- To correct this, the final predicted charges could be adjusted downwards by this small, uniform amount to improve practical accuracy.

Feature Impact

Rank	Feature	Std Coefficient	p-value	Interpretation
1	bmi_smoker_interaction	6,672.76	< 0.001	Continuous BMI effect for smokers (dominant)
2	obese_smoker	4,677.65	< 0.001	Threshold obesity effect for smokers
3	age	3,675.24	< 0.001	Linear age effect across all groups
4	children	604.72	< 0.001	Number of dependents covered
5	region_southwest	-481.99	0.001	Southwest region discount vs baseline (northeast)
6	sex_encoded	-364.13	0.009	Male discount vs female
7	region_southeast	-329.73	0.032	Southeast region discount vs baseline
8	smoker_encoded	-86.46	0.927	Smoker main effect (captured by interactions)
9	bmi	-22.78	0.888	BMI main effect (captured by interactions)

Key Takeaways

- Top 3 features (BMI–smoking interactions and age) account for vast majority of predictive power
- Smoker and BMI main effects are non-significant because their impact is fully captured through interaction terms
- Regional and demographic factors have minimal impact on costs

Fairness Analysis & Practical Recommendations

Model's Sex Coefficient

- Sex coefficient (–$761 for males) may reflect confounding:
  - Different healthcare utilisation patterns
  - Males unmeasured inherent health conditions
- Practical Recommendations
  a. Add Missing Features → The 'sex' coefficient is likely biased due to missing key features (like chronic conditions). Adding these features provide a more accurate model and reduce the confounding.

Regional Disparities Detected

- Regional Coefficients
  - Southeast: –$741 (p=0.032, significant)
  - Southwest: –$1,128 (p=0.001, highly significant)
  - Northeast: Baseline reference
- Geographic pricing discrimination: Two scenarios possible:
  - Regional cost differences: Healthcare is genuinely cheaper in SE/SW (lower provider costs, cost of living)
  - Risk pool differences: Unmeasured health factors differ by region, but correlation ≠ causation
- Practical Recommendation
  - Investigate whether regional differences reflect cost or risk pool composition

Challenges & Solutions

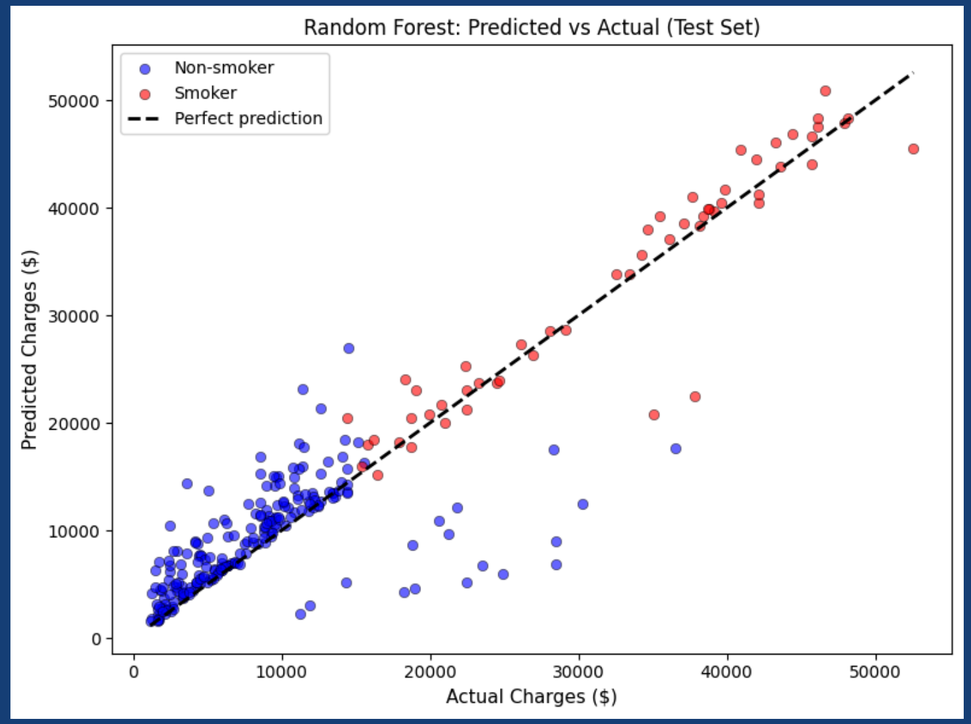
Challenge 1 : High-cost non-smokers remain unexplained

- Challenge: Model persistently underpredicts non-smokers with charges >$20k. Investigation showed these individuals aren't systematically more obese (52.5% vs 52.8% overall) or much older (45.1 vs 39.4 years average).
- Solution: Acknowledged dataset limitation—high costs likely driven by unmeasured factors (chronic conditions, surgeries, medications). Advanced models (Random Forest) also failed to improve predictions, confirming missing information rather than modeling deficiency.

Chellenge 2 : Feature Selection Complexity

- Challenge: Multiple candidate features (region, sex, children) showed weak univariate relationships but unclear multivariate importance.
- Solution: Used backward elimination via statsmodels, systematically removing non-significant features (BMI main effect p=0.921, region_northwest p=0.165). Standardised coefficients confirmed final feature ranking, ensuring only meaningful predictors retained.

Advanced Model Challenge: Random Forest Regressor



The Random Forest Regressor introduced significant complexity – sacrificing interpretability, computational efficiency, and regulatory transparency while delivering worse predictive performance than the simple Linear Regression model. This demonstrates that advanced algorithms cannot compensate for fundamental data limitations (missing medical history features) and that domain-driven feature engineering outperforms automated feature learning for this dataset. The Linear Regression model remains superior due to its interpretability, better accuracy, and alignment with business requirements for transparent insurance pricing.

THANK YOU

BowlOfBaifan/Insurance-Price-Predictor