Predicting Medical Insurance Costs

- Comparative Regression Analysis
- Otajon Yuldashev (473457)

Motivation & Problem

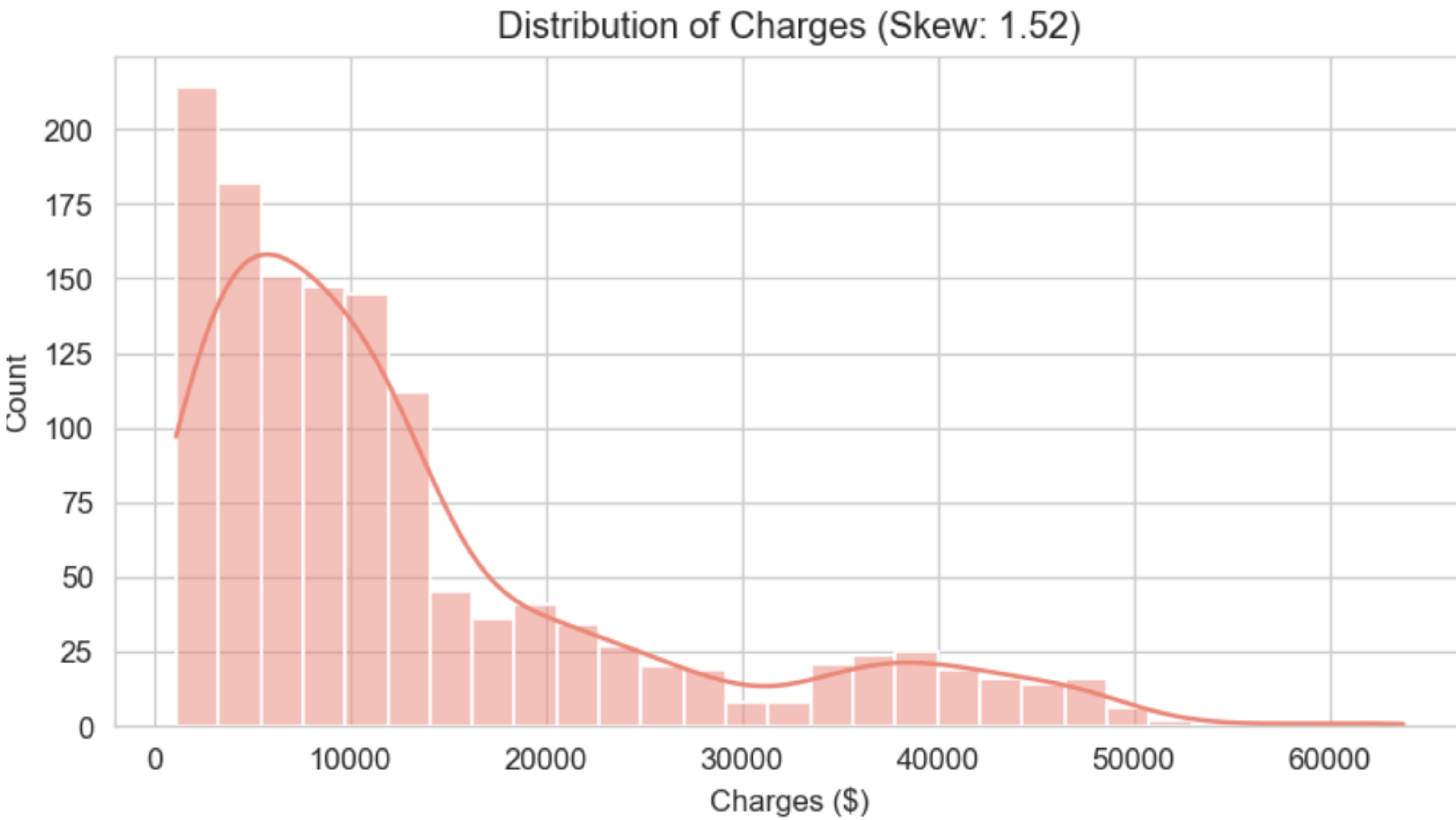
- Accurate pricing is critical for insurers.
- Underestimation → losses
- Overestimation → loss of clients
- Medical costs are non-linear and highly skewed.

Dataset Overview

1,337 observations

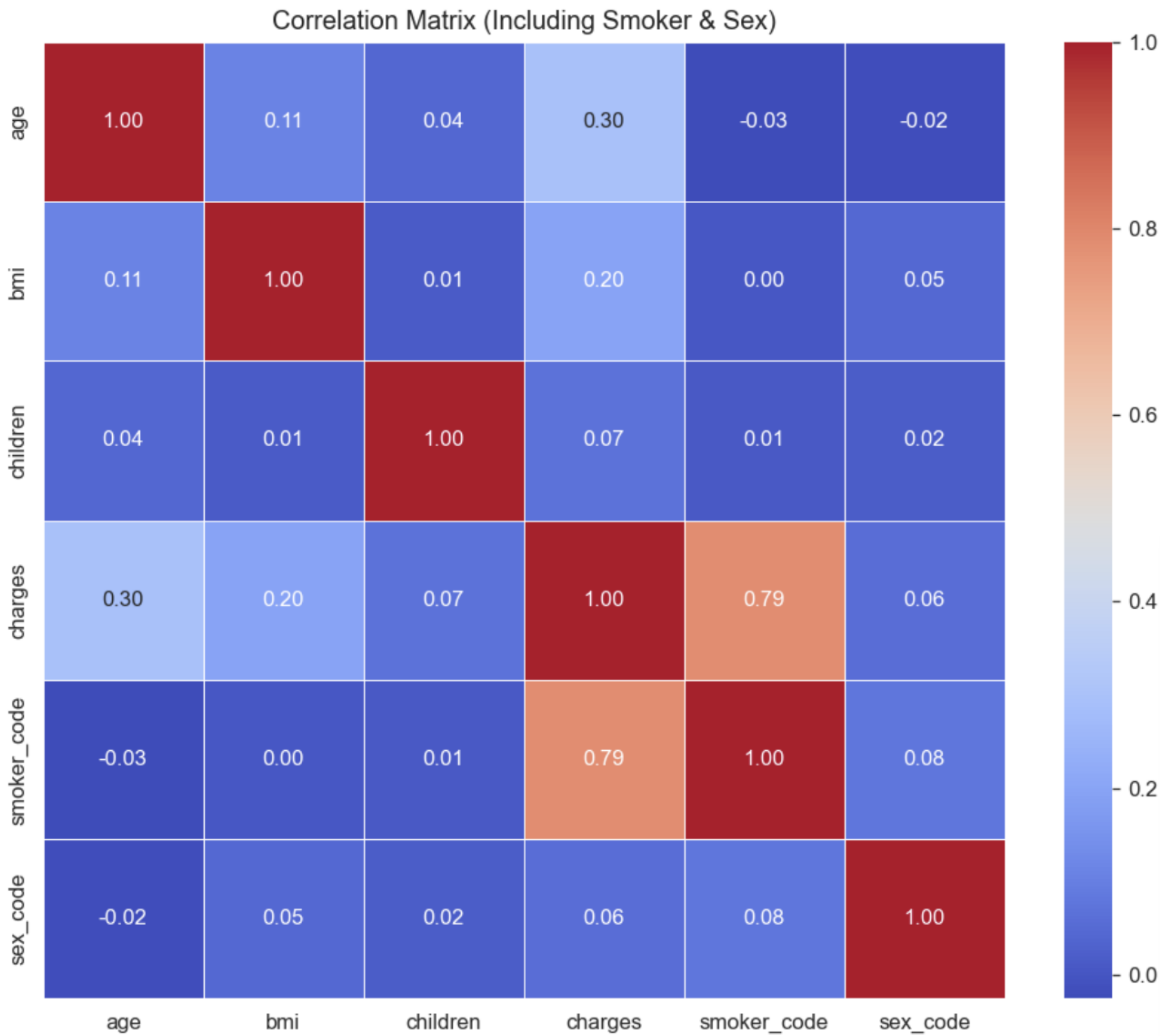
Features: age, sex, BMI, children, smoker, region

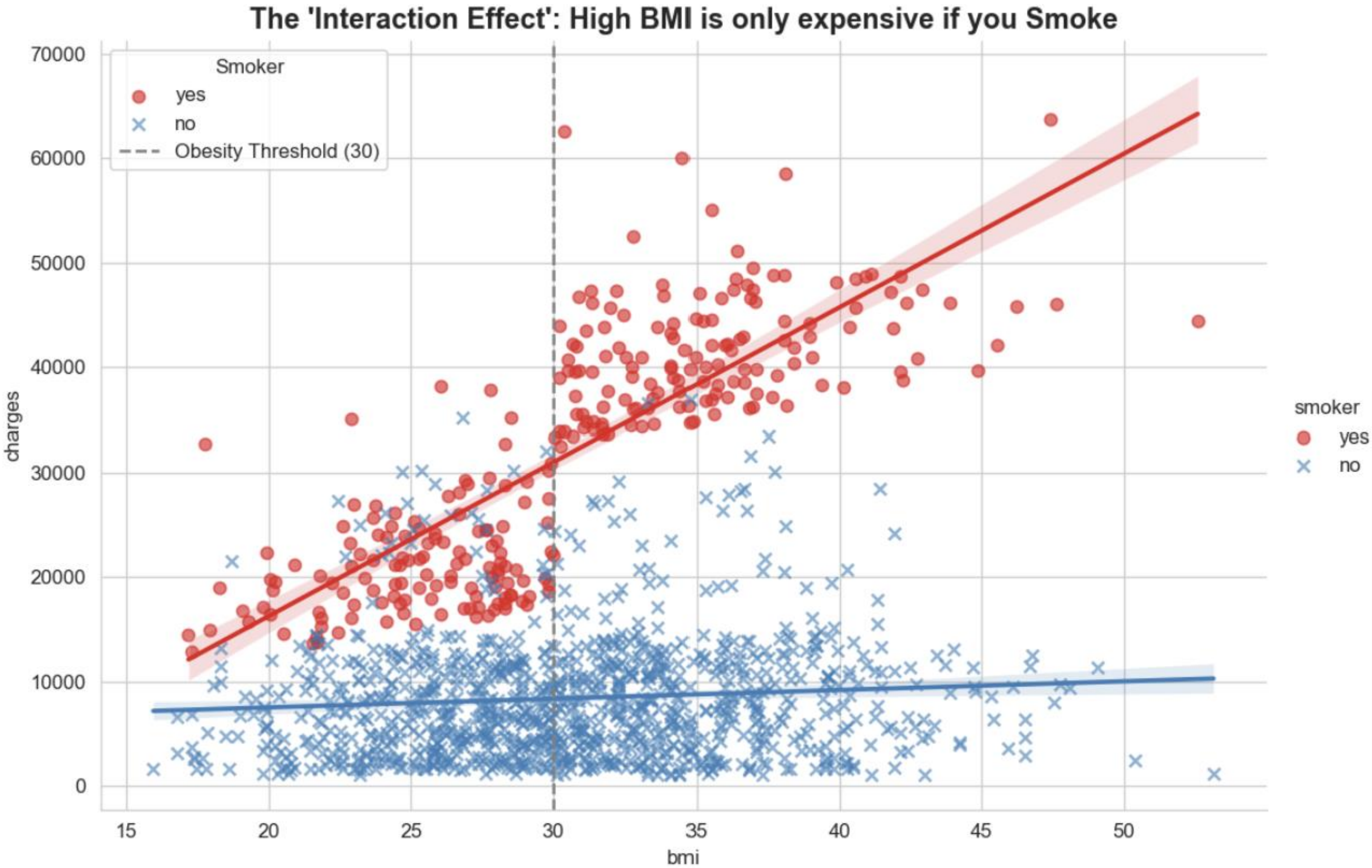
Target: medical charges ($)

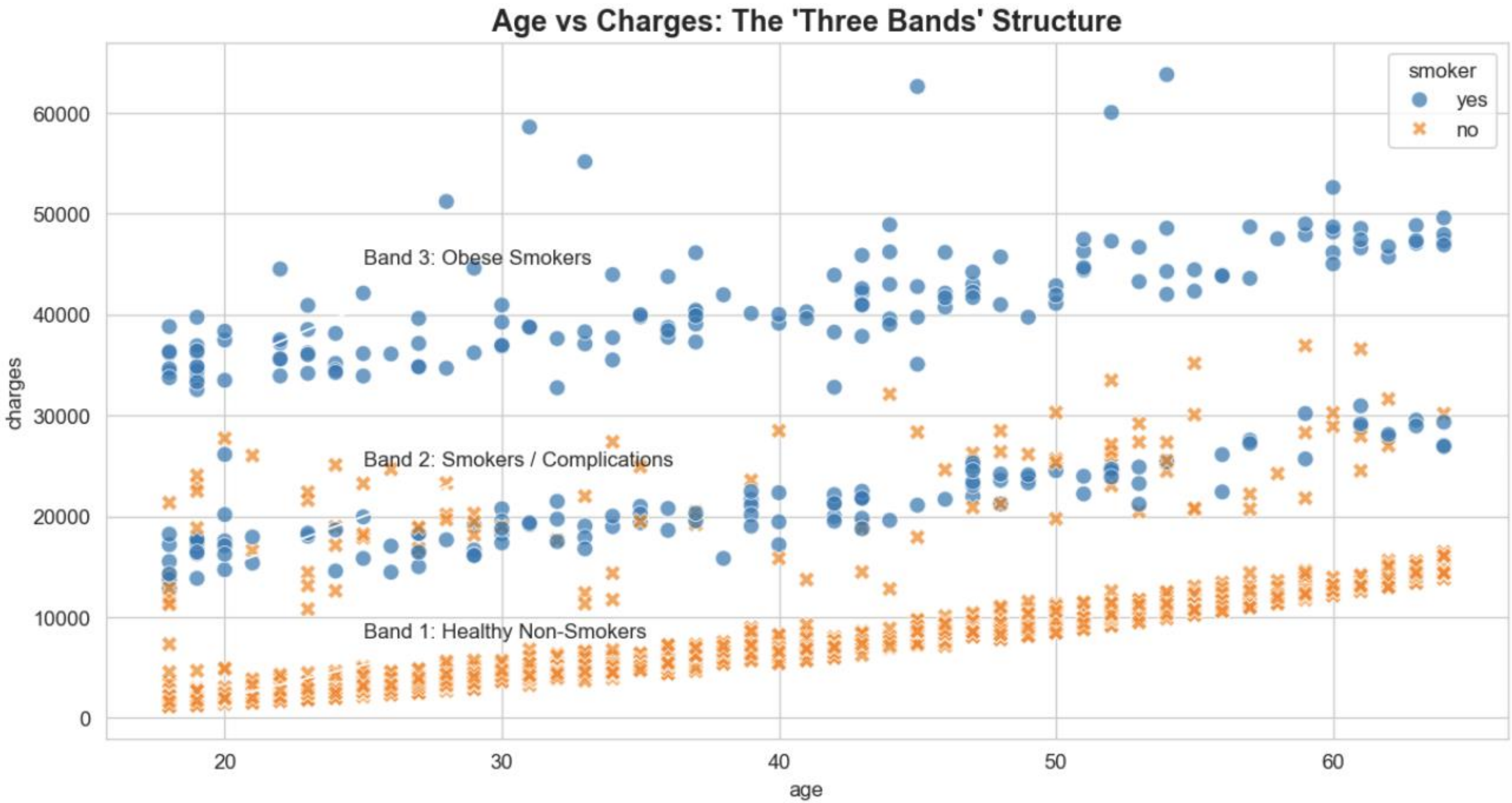

Distribution of Charges (Skew: 1.52)
Count
200
175
150
125
100
75
50
25
0
0
10000
20000
30000
40000
50000
60000
Charges ($)

Key EDA Findings

- Smoking is dominant risk factor
- BMI and age amplify costs for smokers
- Sex, region, children → minimal impact
- Strong interaction effects present


Correlation Matrix (Including Smoker & Sex)
age
bmi
children
charges
smoker_code
sex_code
1.00 0.11 0.04 0.30 -0.03 -0.02
0.11 1.00 0.01 0.20 0.00 0.05
0.04 0.01 1.00 0.07 0.01 0.02
0.30 0.20 0.07 1.00 0.79 0.06
-0.03 0.00 0.01 0.79 1.00 0.08
-0.02 0.05 0.02 0.06 0.08 1.00
1.0
0.8
0.6
0.4
0.2
0.0


The 'Interaction Effect': High BMI is only expensive if you Smoke
Smoker
yes
no
Obesity Threshold (30)
charges
bmi
smoker
yes
no
0
10000
20000
30000
40000
50000
60000
70000
15
20
25
30
35
40
45
50


Age vs Charges: The 'Three Bands' Structure
smoker
yes
no
Band 3: Obese Smokers
Band 2: Smokers / Complications
Band 1: Healthy Non-Smokers
charges
age
0
10000
20000
30000
40000
50000
60000
20
30
40
50
60

Feature Engineering & Preprocessing

- Log-transform of target variable
- BMI × Smoker interaction term
- One-Hot Encoding for categorical features
- Stratified train-test split by smoker status

Models & Tuning

- Ridge Regression (linear baseline)
- Random (bagging)
- XGBoost (boosting)
- GridSearchCV used for hyperparameter tuning

Intermediate Evaluation (Log Scale)

- Random Forest RMSE: 0.2529, R^2: 0.9233
- XGBoost RMSE: 0.2579, R^2: 0.9202
- Ridge RMSE: 0.3480, R^2: 0.8547

Final Evaluation (Dollar Scale)

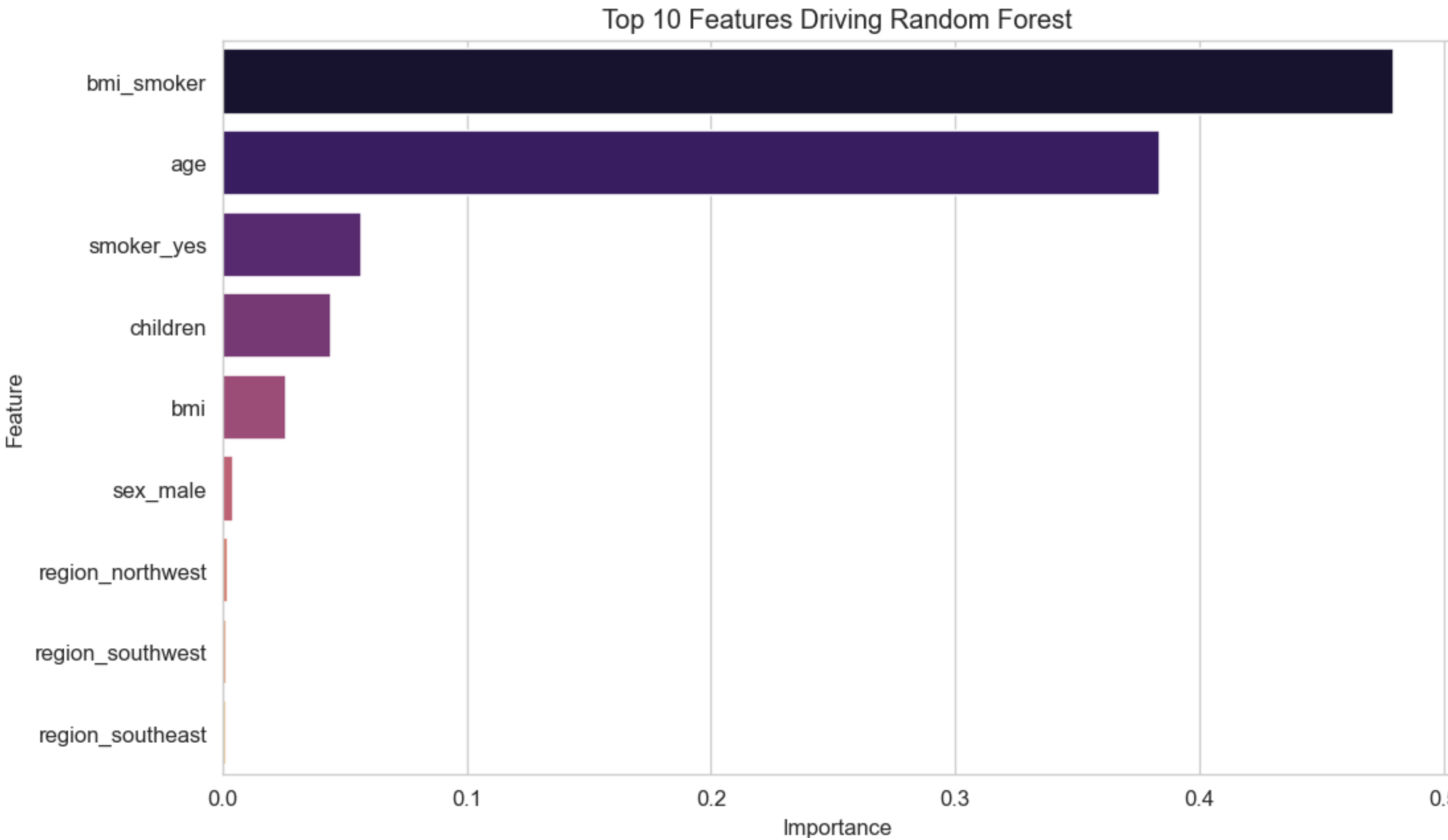
Random Forest
RMSE: $3,280 |
R^2: 0.925

XGBoost
RMSE: $3,333
R^2: 0.923

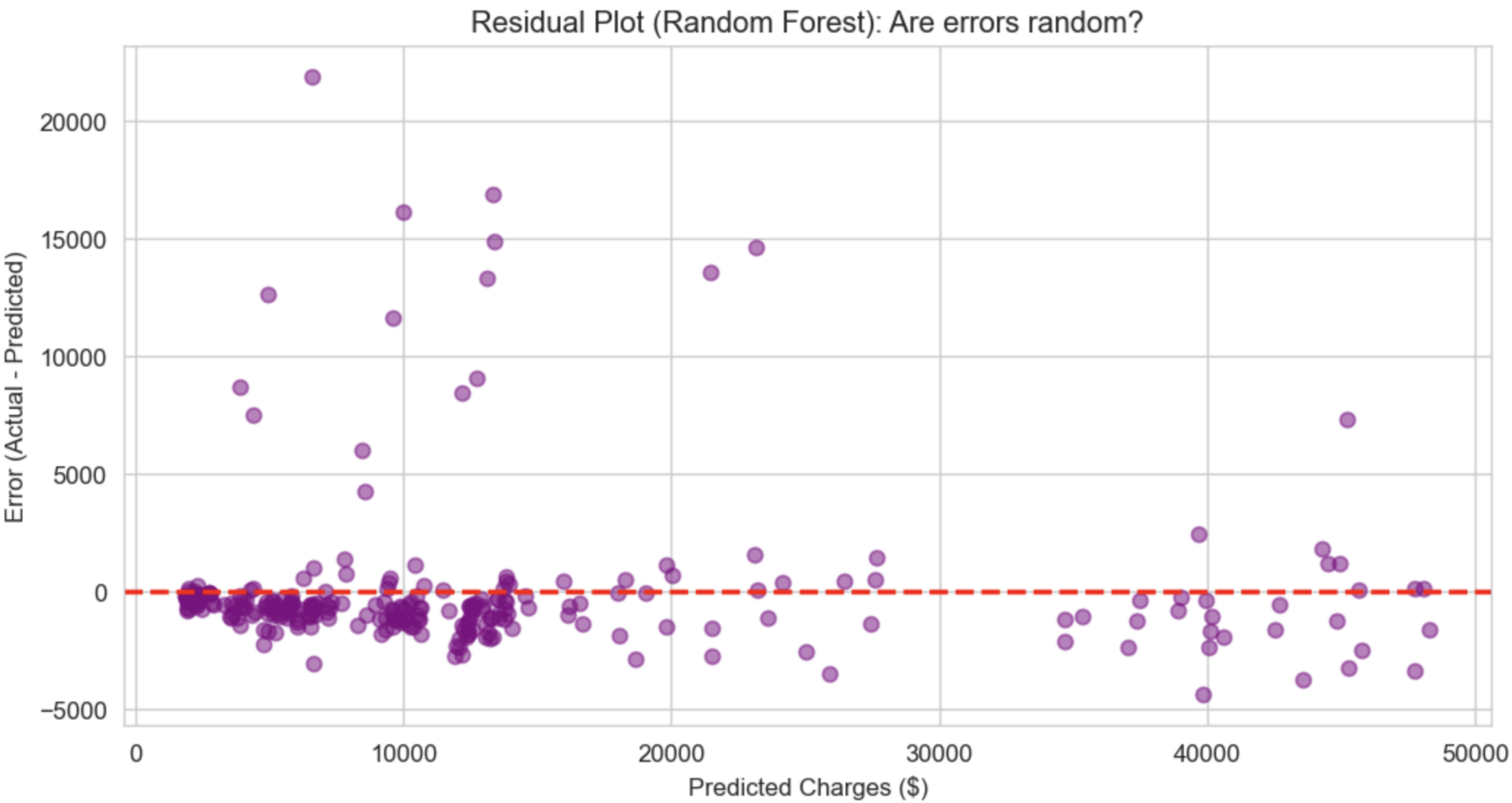
Ridge
RMSE: $7,710 |
R^2: 0.588

Model Interpretation

- Most important feature: BMI × Smoker
- Age is second most important
- Random Forest captures non-linear effects
- Residuals increase for high-cost cases



Top 10 Features Driving Random Forest
bmi_smoker
age
smoker_yes
children
bmi
sex_male
region_northwest
region_southwest
region_southeast
Feature
0.0
0.1
0.2
0.3
0.4
0.5
Importance


Residual Plot (Random Forest): Are errors random?
Error (Actual - Predicted)
20000
15000
10000
5000
0
−5000
0
10000
20000
30000
40000
50000
Predicted Charges ($)

Conclusion

Random Forest performs best overall

Strong interpretability and robustness

High-cost cases slightly underestimated

Future work: focus on extreme-risk clients

Credit Card Default Prediction

Comparative classification analysis

Problem & Motivation

- Binary classification problem
- Predict default next month (0/1)
- Important for credit risk management
- Dataset: 30,000 clients

Dataset Overview

- Demographics: age, sex, education, marriage
- Credit profile: credit limit
- Repayment behavior: PAY variables
- Financial history: bill and payment amounts
- Target imbalance (~22% default)

Exploratory Data Analysis

- Default class is underrepresented
- Credit limit and age are skewed
- Categorical variables are unevenly distributed
- Repayment status strongly associated with default

Models & Methodology

- Logistic Regression (baseline)
- Random Forest (bagging)
- Gradient Boosting (boosting)
- Train/test split with stratification
- Hyperparameter tuning using GridSearchCV
- Evaluation via precision, recall, F1-score

Model Performance Comparison

- Logistic Regression: low recall for default
- Random Forest: improved balance
- Gradient Boosting: **best F1-score**
- Gradient Boosting selected as final model.

Error Analysis

- Confusion matrix interpretation
- False negatives remain a challenge
- Repayment history is the strongest predictor
- Trade-off between recall and precision

Ethical Considerations

- Risk of bias from historical data
- Cost of false positives and false negatives
- Models should support, not replace, human decisions
- Importance of transparency and monitoring

Conclusions

Ensemble models outperform baseline

Gradient Boosting performs best

PAY variables dominate prediction