BA 305

Team 7

PREDICTING INSURANCE PRICES

Damian Grochowski, Ryan Whitehouse, Ole Nagorsen, Michael Lee, Reid Brock

Today's Agenda

1. INTRODUCTION
2. DATASET
3. MODEL
4. MODEL RESULTS
5. FINDINGS

1.INTRODUCTION:

WHY WE NEED AN INSURANCE PRICING MODEL

A Prevailing Issue in American Healthcare

U.S. MEDICAL INSURANCE HAS BECOME A SIZABLE AND FREQUENT EXPENSE

COMPLICATED BILLING SYSTEMS CONFUSE PATIENTS ON THEIR INSURANCE CHARGES

PATIENTS DO NOT KNOW WHETHER THEY ARE BEING OVERCHARGED OR NOT

Americans owe over

$220,000,000,000

in medical debt

Consumer Financial Protection Bureau. 1 Oct. 2024. CFPB Takes Aim at Double Billing and Inflated Charges in Medical Debt Collection. https://www.consumerfinance.gov/about-us/newsroom/cfpb-takes-aim-at-double-billing-and-inflated-charges-in-medical-debt-collection/

OUR GOAL

To accurately predict a patient's insurance costs given their personal profile in order to provide greater pricing clarity

2.OUR DATASET:

WHAT IT IS AND HOW WE USED IT

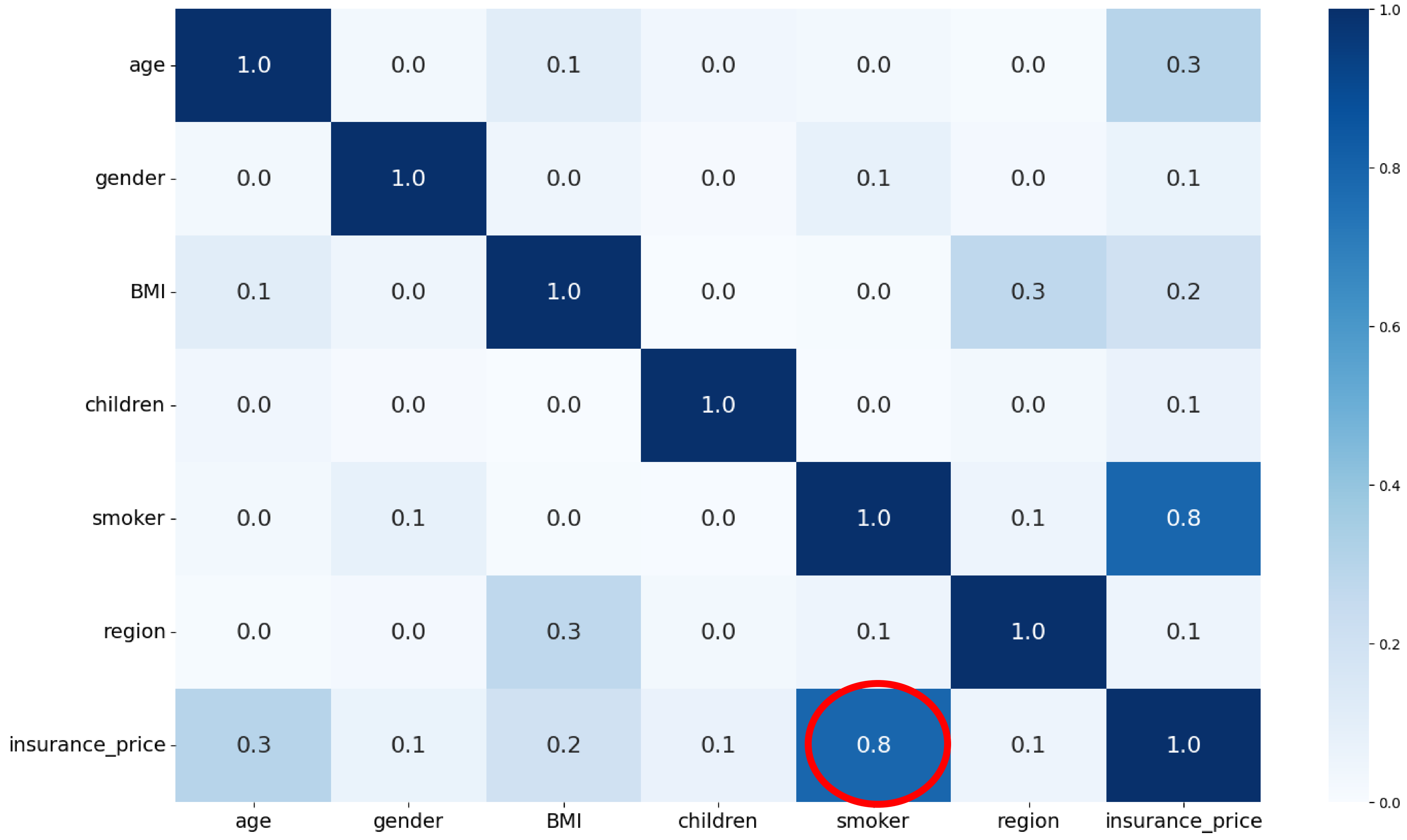
Final Dataset After Cleaning: 2,772 -> 2,761 Datapoints

Variable	Description	Values
Age	Age of the patient/beneficiary	18 - 64
Smoker	0 = non-smoker; 1 = smoker	0 or 1
Gender	1 = male; 2 = female	1 or 2
BMI	Body Mass Index	15 - 53
No. of Children	Number of possible dependents covered	0 - 5
Region	Geographic Region	Northwest (1), Northeast (2), Southwest (3), Southeast (4)
Insurance Price	Total Annual Medical Charges billed by Insurer	$1000 - $60,000+

Data Preprocessing

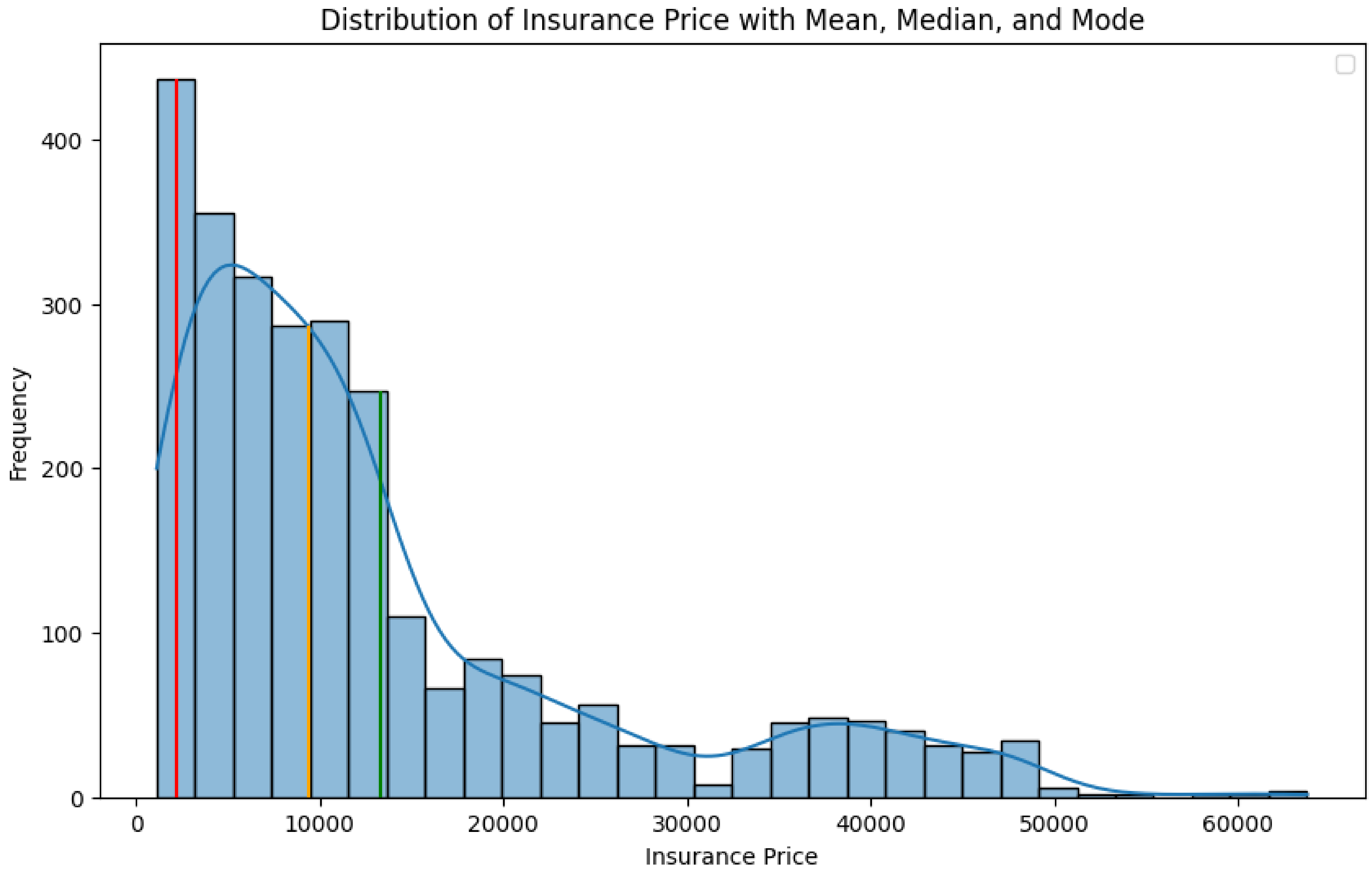
1. Removed rows with missing/unknown values (?)
2. Converted smoker and gender variables into numeric binary values
3. Applied One-Hot encoding to convert regions into numeric binary value
4. Verified data types

Initial Patterns in Our Data



Strong correlation between Smoking and Insurance Charges, followed by Age and BMI.

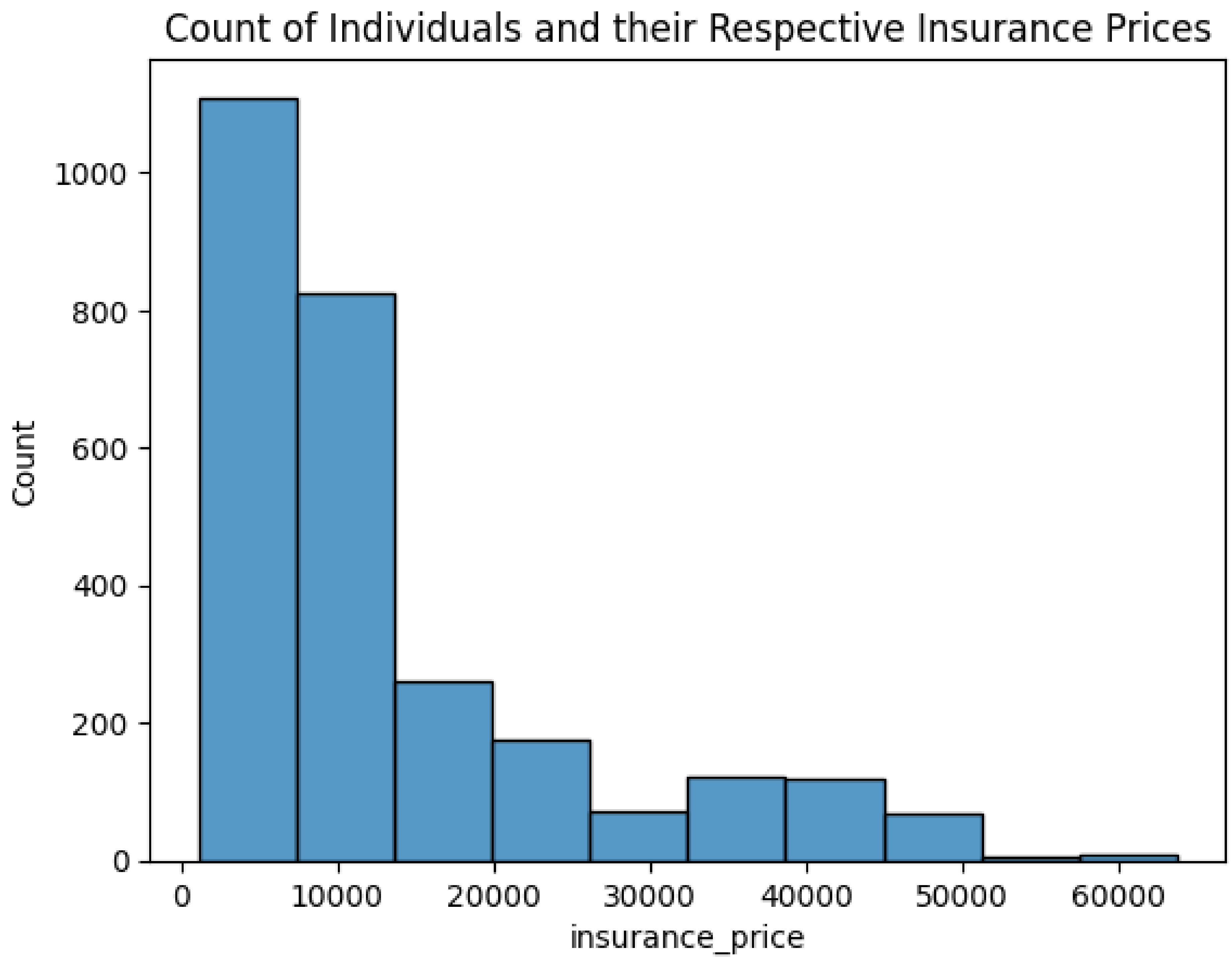
Our Dataset’s Averages



Key:
Mean: 13274.76
Median: 9377.90
Approx. Mode: 2166.02

- Mean is pulled right/upward by extreme cases
- Median reflect typical costs
- Mode shows most common prices are low
- Insurance Prices are not evenly distributed

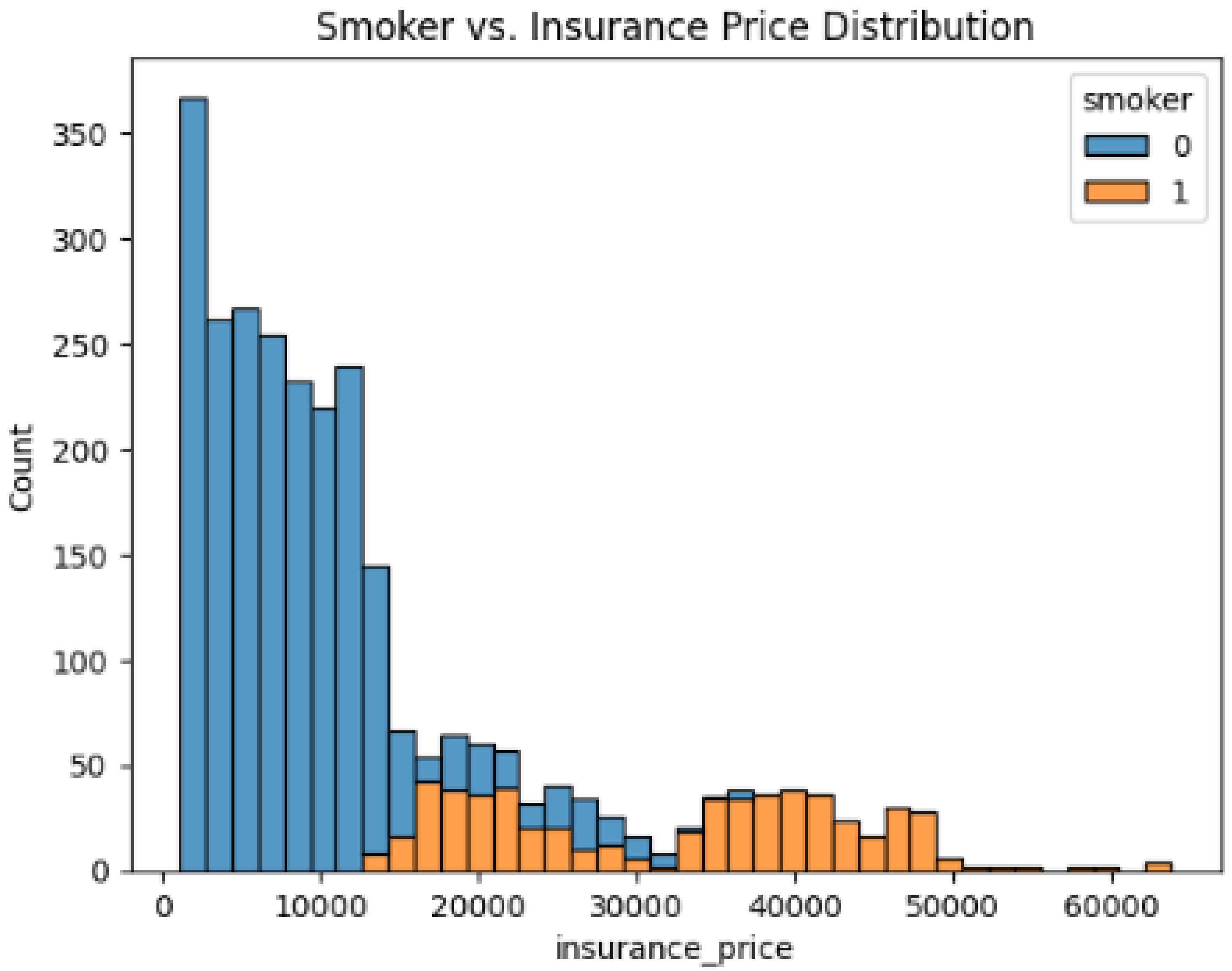
Addressing Extreme Insurance Price Cases

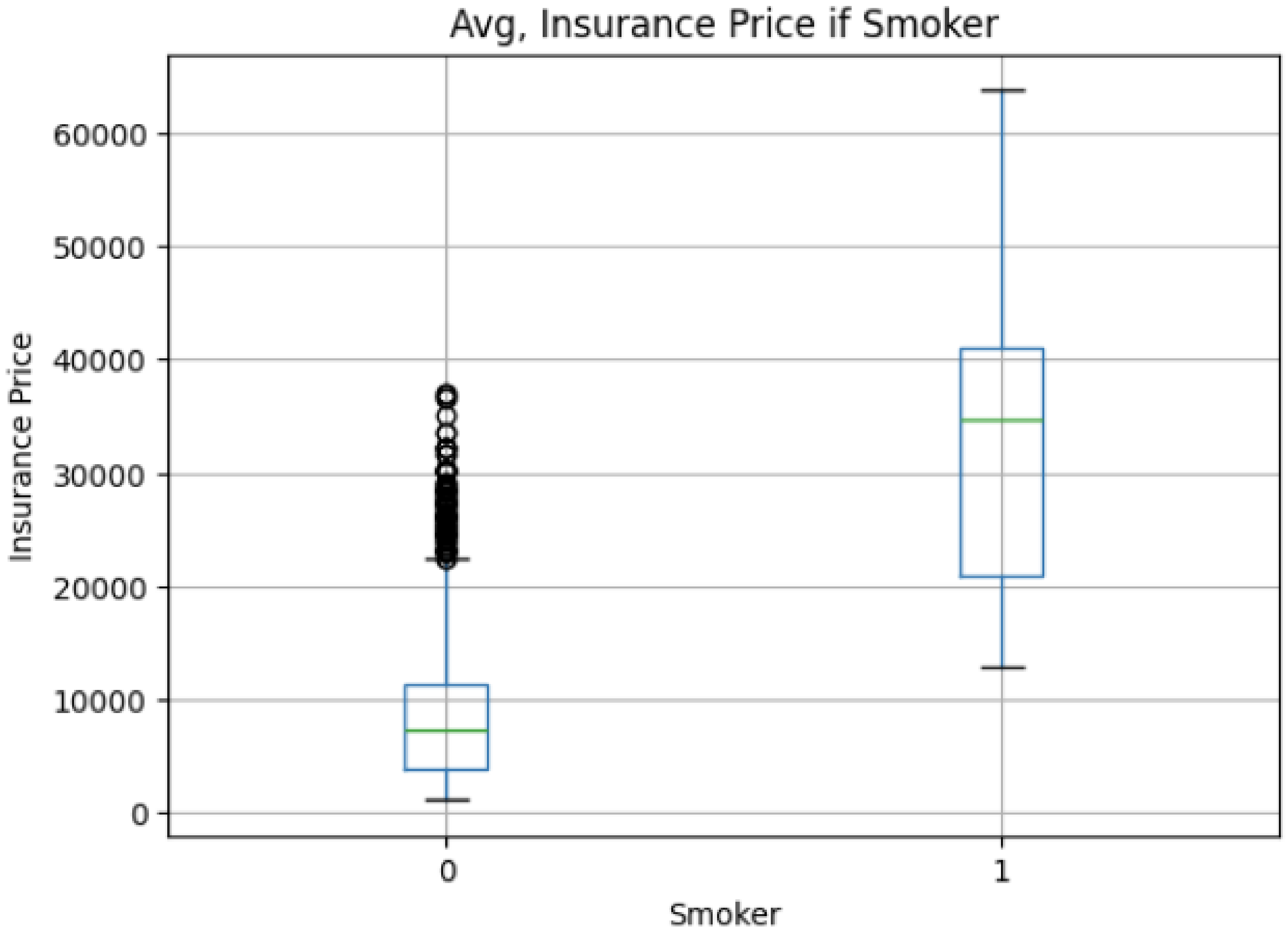


- 14 users experience costs over **$50,000**
- ICU stays can exceed **$4,300/day**
- Removing cases would distort the true distribution

"Critical Care Statistics," Society of Critical Care Medicine (SCCM), accessed December 1, 2025, https://www.sccm.org/communications/critical-care-statistics.

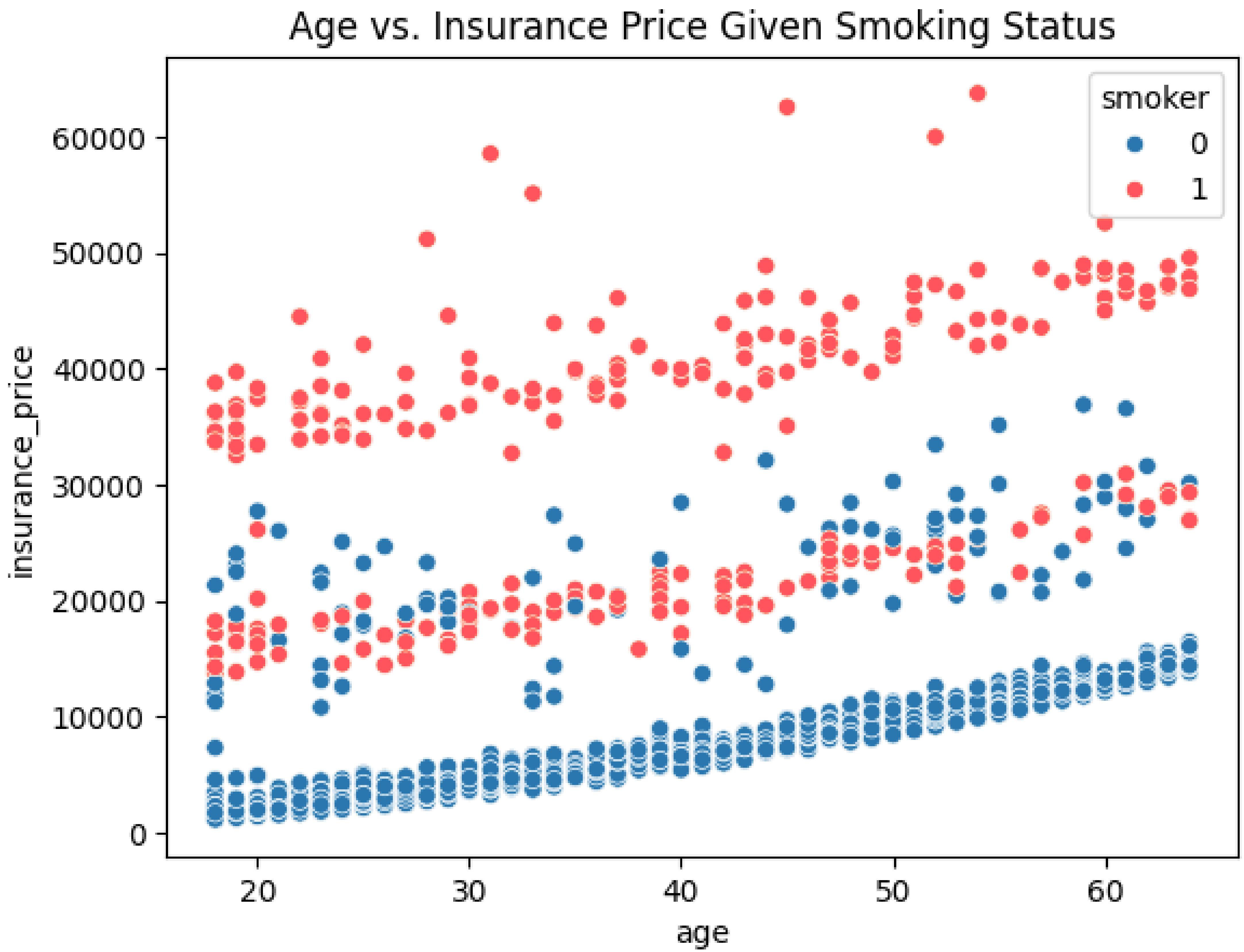
Initial Patterns in Our Data





Smokers incur higher and more variable costs compared to non-smokers, with extreme high cases pushing annual costs.

Abnormal Trends In Price



3.MODEL:

EXPLORATION AND POTENTIAL MODELS

Our Approach

Step 1:
Establish a baseline performance benchmark for all subsequent models

Step 2:
Introduce more complex, flexible models to capture relationships

Step 3:
Evaluate the models using same training and test split, tune hyperparameters of models, when feasible, and apply 5-fold cross validation

Step 4:
Comparison using MAE, RMSE and R-squared

Step 5:
Building and interpreting our two stage prediction algorithm

Comparison Criteria

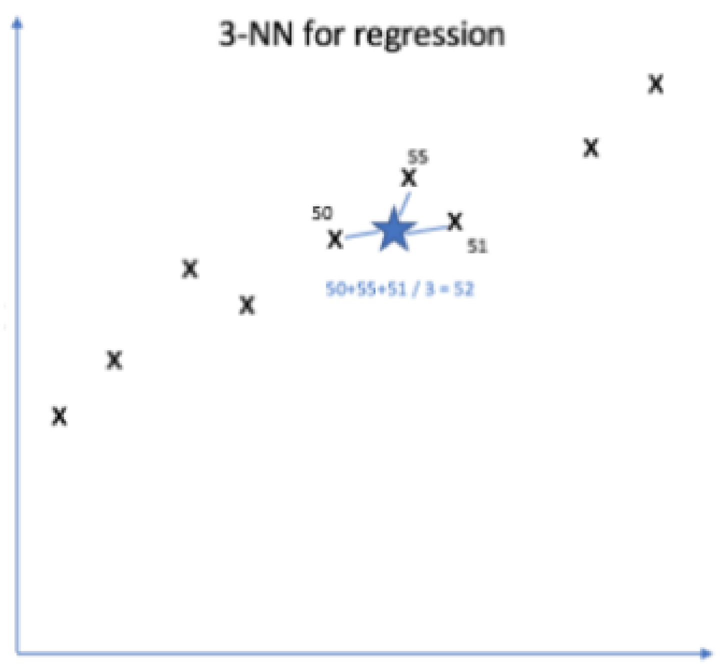
MAE	RMSE	R^2
Measures absolute average difference between predictions and values. Simple measure of average error.	Measures square root of average squared error. Penalizes large errors.	Analyzes variability of model around its mean. AKA. how well does the model predict actual values?

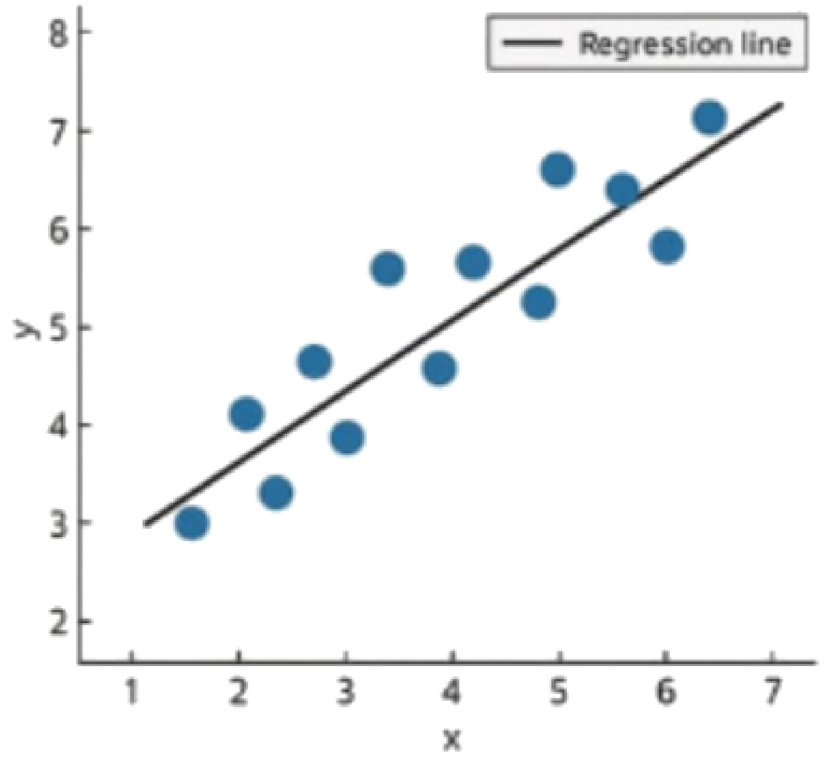
BEST MODEL = CONSIDERS ALL THREE

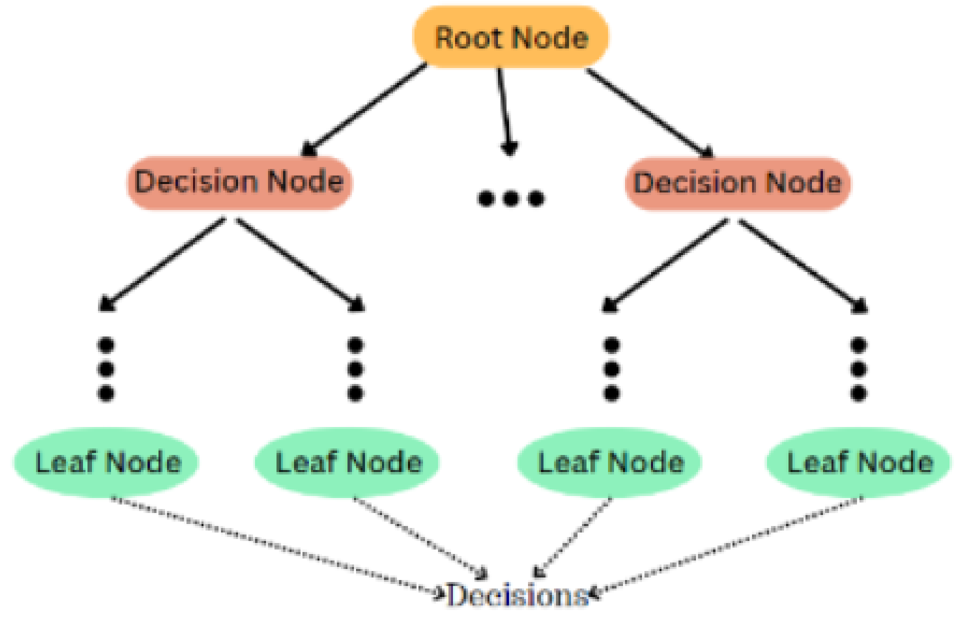
Baseline Model

Naive Rule	Calculate by mean
MAE	9132.437
RMSE	12156.556
R^2	0.000 (Based on mean)

Models In Consideration







Non-Linear	Linear	Tree-Based
KNN Regressor	Linear Regression	Decision Tree Regressor
	Polynomial Regression	Random Forest Regressor
	Ridge Regression	Gradient Boosting Regressor
	Lasso Regression	XGBoost Regressor
		Hist Gradient Boosting Regressor

4.MODEL RESULTS:

THE BEST MODEL AND WHAT IT TELLS US

Performance Overview

Model	MAE	RMSE	R^2
Baseline	9132.437	12156.556	0
KNN Regressor	1272.606	4189.186	0.879
Linear Regression	4181.771	6063.585	0.75
Polynomial Regression	2870.173	4766.909	0.845
Ridge Regression	4182.268	6063.584	0.75
Lasso Regression	4182.846	6063.102	0.75
Decision Tree Regressor	607.44	2857.938	0.944
Random Forest Regressor	1235.035	2567.749	0.955
Gradient Boosting Regressor	1273.326	2635.217	0.952
Hist Gradient Boosting Regressor	1528.816	2784.487	0.947
XGBoost Regressor	643.087	2234.407	0.966

Performance Overview: Best Model

MAE

1. Decision Tree Regressor
2. XGBoost Regressor*
3. Random Forest Regressor

XGBoost Regressor

MAE = 643.087
RMSE = 2234.407
R^2 = 0.966

RMSE

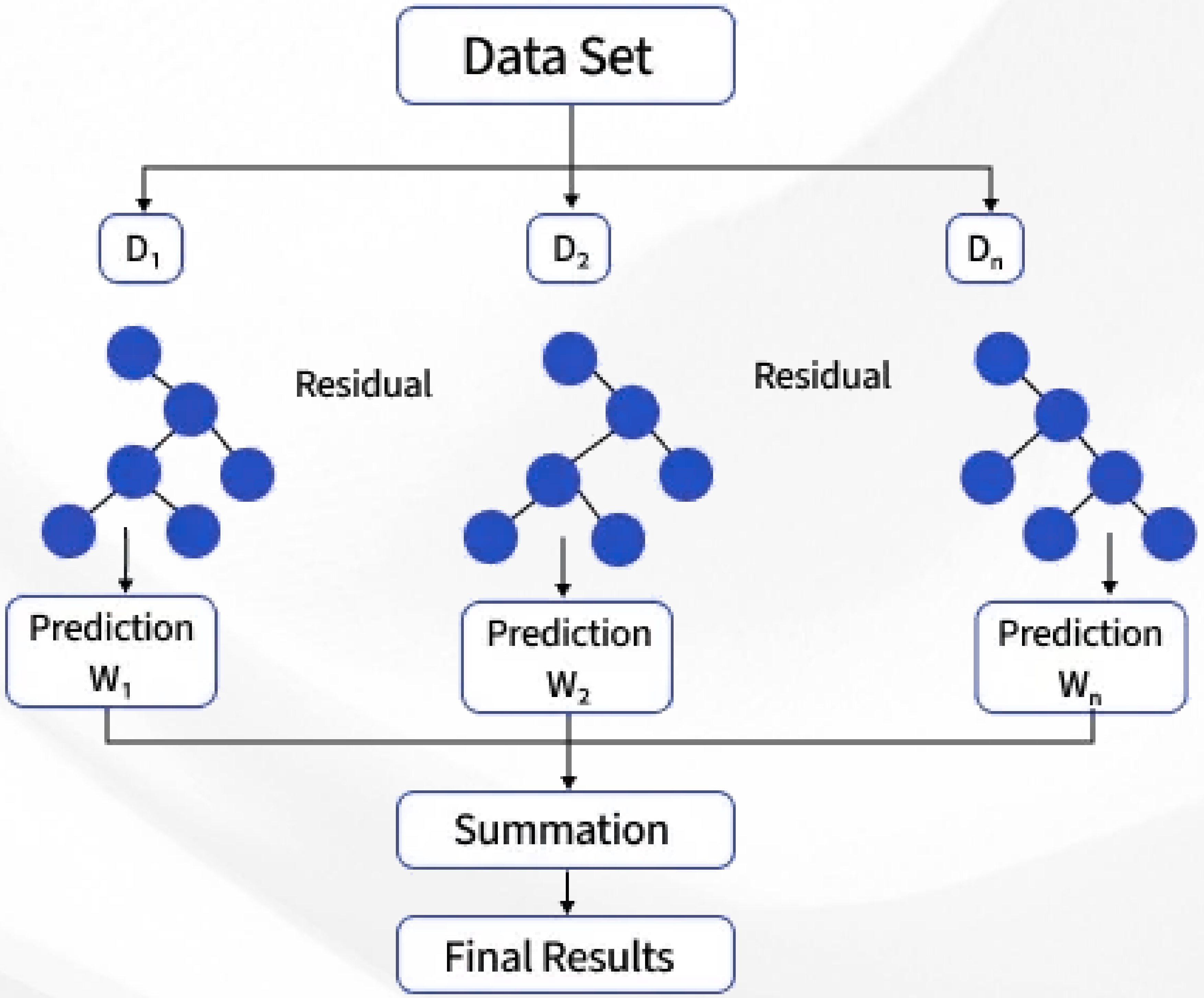
1. XGBoost Regressor*
2. Random Forest Regressor
3. Gradient Boosting Regressor

R^2

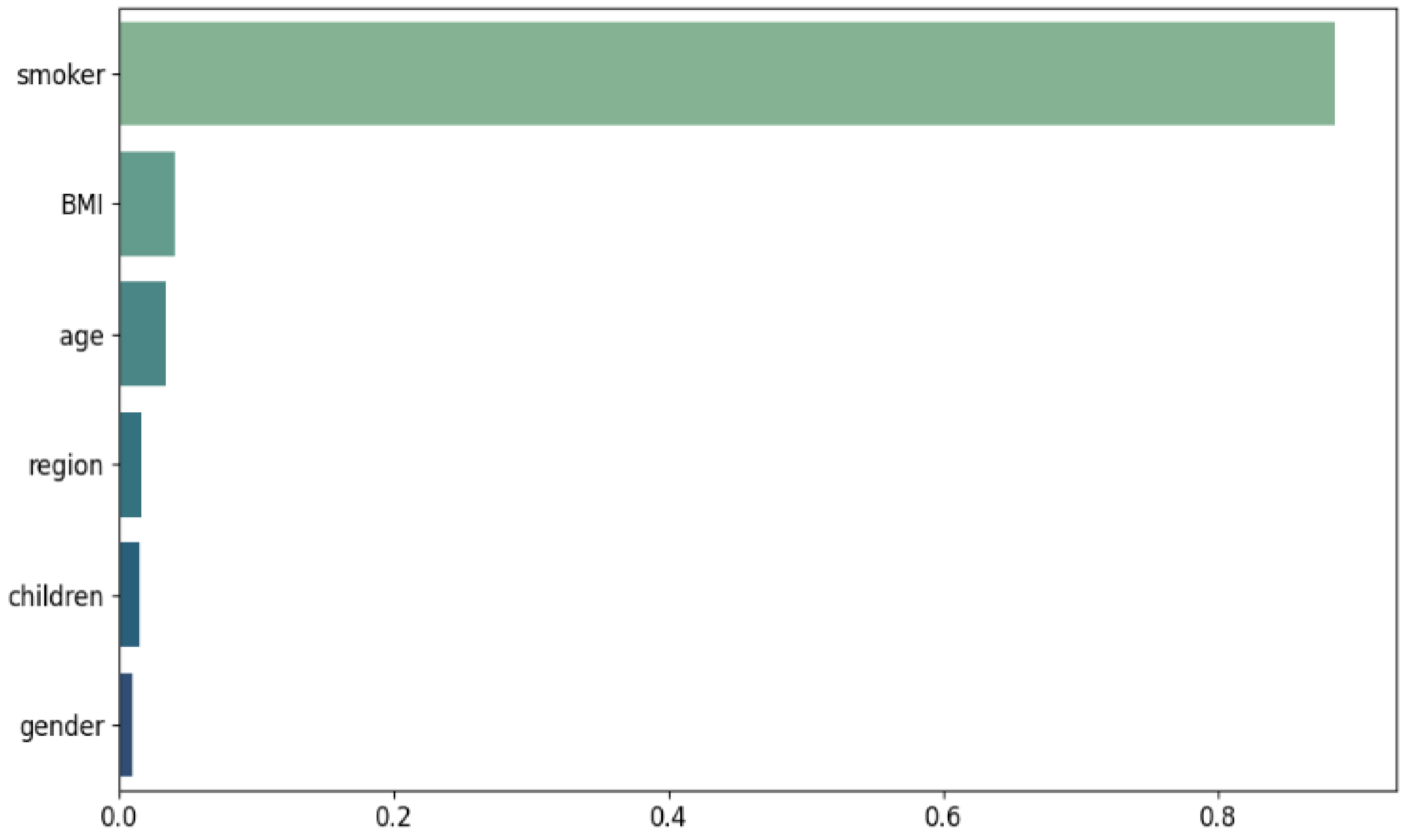
1. XGBoost Regressor*
2. Random Forest Regressor
3. Gradient Boosting Regressor

Model Interpretation: XGBoost

- An ensemble model that sequentially creates decision trees derived from the errors of previous trees
- Prunes trees backwards after reaching minimum depth to maximize gain and reduce chances of overfitting
- Then factors in the predictions of all models together for a final decision



Model Interpretation: XGBoost



Most Important Features

Smoker - 88.85%

BMI - 4.05 %

Age - 3.43 %

Least Important Features

Region - 1.585 %

Children 1.49 %

Gender - 0.941 %

Why is Smoker so High?

Real Life: Smokers usually have drastically higher medical costs and premiums.

Model: XGBoost learns and builds more trees that split on Smoker first.

Warner, David O., Benjamin J. Borah, John Moriarty, David R. Schroeder, Yan Shi, and Nilay D. Shah. "Smoking Status and Health Care Costs in the Perioperative Period: A Population-Based Study." JAMA Surgery 149, no. 3 (March 2014): 259–266. https://doi.org/10.1001/jamasurg.2013.5009

5.OUR FINDINGS:

APPLYING OUR MODEL AND WHAT PATIENTS CAN TAKE AWAY FROM THIS

Why Are Costs So High?

Why the Numbers Look High

- Dataset reflects total medical insurance costs billed to beneficiaries
- Explains why some individuals exceed **$50,000** in annual charges
- Average one-night stay at a U.S. hospital exceeds **$3,000**

Context From National Data

- Avg. employer-sponsored premium (2024): **$8,951** for single coverage
- Smoking-related illnesses add over **$100 Billion** in direct U.S. healthcare costs annually

Kaiser Family Foundation. 2024. *2024 Employer Health Benefits Survey*. Kaiser Family Foundation. https://www.kff.org/health-costs/2024-employer-health-benefits-survey/#3f3fc2dd-74dd-4cb6-9d1c-9c19ff972f6a

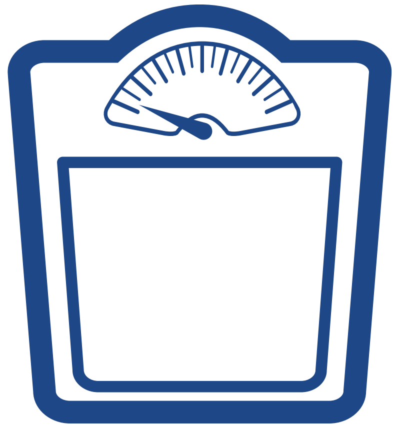
Warner, David O., Benjamin J. Borah, John Moriarty, David R. Schroeder, Yan Shi, and Nilay D. Shah. "Smoking Status and Health Care Costs in the Perioperative Period: A Population-Based Study." JAMA Surgery 149, no. 3 (March 2014): 259–266. https://doi.org/10.1001/jamasurg.2013.5009

Dominant Driver in Insurance Costs

Age

Smoking

Obesity



The single strongest determinant of insurance price is smoking

Sample_ID	Mock_Name	Original_Smoker_Status	Predicted_Non_Smoker_Price	Predicted_Smoker_Price	Price_Increase_If_Smoker
1362	Ole	0	**$8,965.83**	**$19,895.90**	**$10,930.07**
2543	Reid	0	**$4,530.18**	**$17,303.95**	**$12,773.77**
2229	Michael	0	**$11,731.44**	**$23,315.32**	**$11,583.88**
2048	Ryan	0	**$3,945.26**	**$33,153.18**	**$29,207.92**
446	Damian	0	**$4,685.57**	**$39,034.12**	**$34,348.55**

Practical Applications of Our Model

1. Consumers can input their own health characteristics to see an **approximate cost**.
2. If insurer quotes **differ** dramatically from model predictions, users can flag or **investigate discrepancies**.
3. Helps individuals **anticipate** future insurance **expenses**.

How This Will Help Consumers?

Clear Pricing Expectations

Transparency Into Cost Drivers

+

Actionable Health Insights

Better Decision-Making

Pricing Clarity

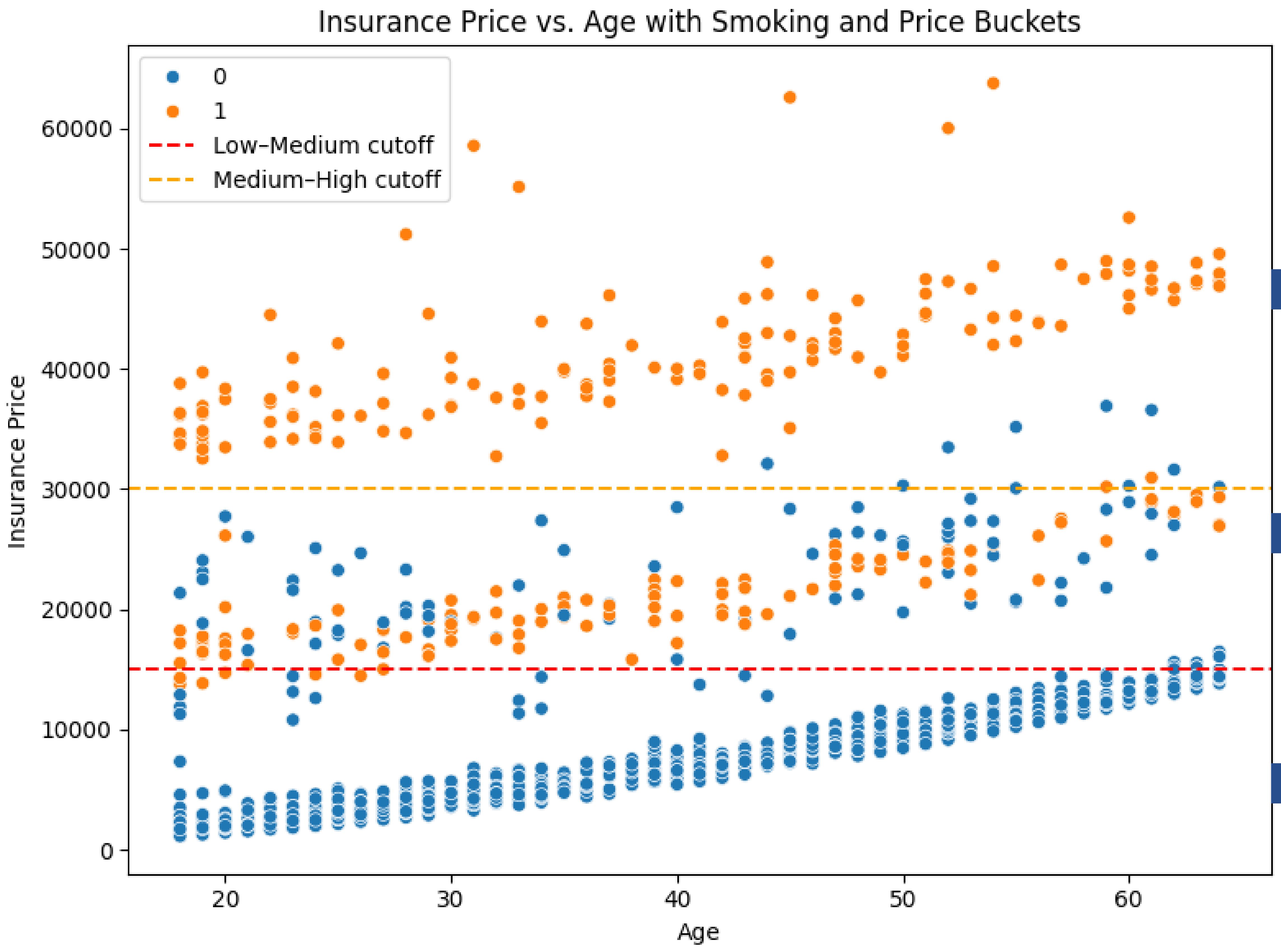
Health Impact

BONUS:

THE CUTTING ROOM FLOOR

(AND POTENTIAL IMPROVEMENTS)

Two Stage Approach Classification into Regression



Classification: Random Forest
Regression: XGBoost
MAE: 1115
R^2: 0.876

C: High Price Class
$30K+, 339 Entries

B: Medium Price Class
$15K-$30K, 397 Entries

A: Low Price Class
$0-$15K, 2,025 Entries

Thank You!

Any Questions?