Predicting Health Insurance Costs

Written By: Leo Zheng, Eric Tysinger, Michael Whitfield, Nelson Paguada

Purdue University, Mitch Daniels School of Business

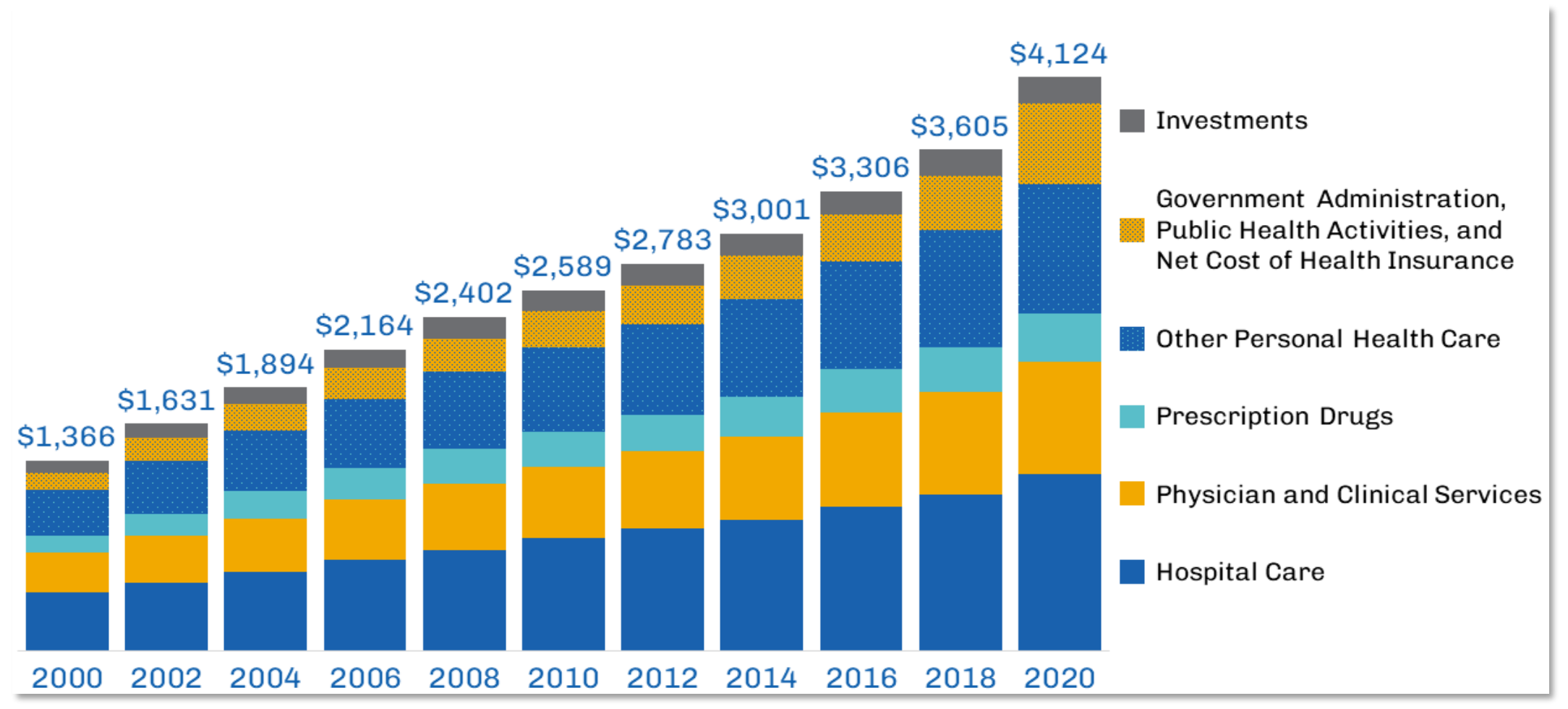
zheng702@purdue.edu; etysinge@purdue.edu; mlwhitefi@purdue.edu; npaguada@purdue.edu

PURDUE UNIVERSITY.

Mitch Daniels School of Business

BUSINESS PROBLEM

Rising healthcare spending, as highlighted by the American Medical Association, emphasizes the need for **transparency** and **accurate** cost predictions.

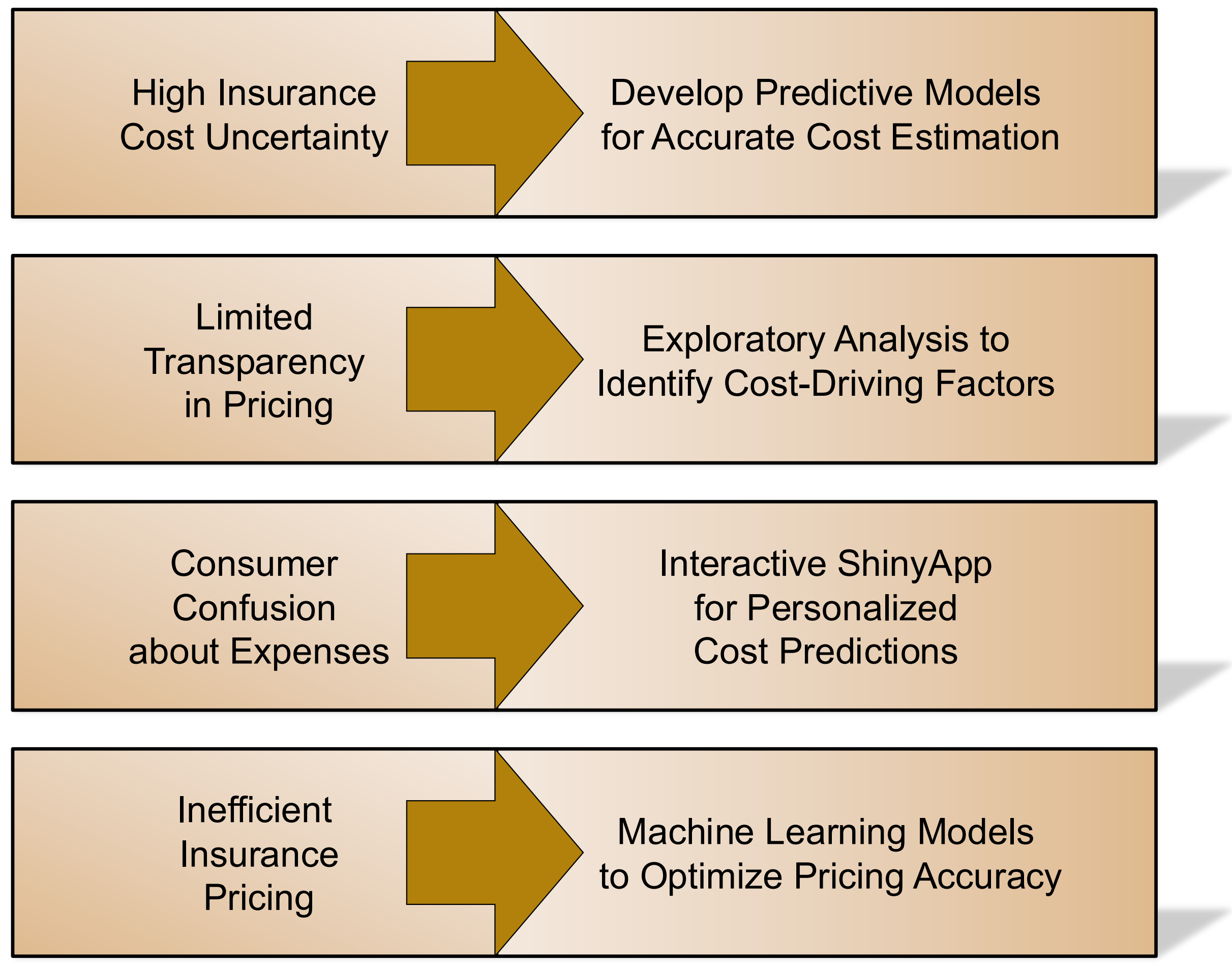


To address this, we developed a predictive analytics Application that estimates health insurance expenses for individuals and families.

This interactive tool provides consumers with **clear insights** into potential costs and helps insurers effectively price their plans.

ANALYTICAL APPROACH

Our project translates the broader business challenge of healthcare cost unpredictability into specific analytical tasks. We structured our analytical approach into distinct steps, clearly linking each business issue to a measurable, data-driven solution:



High Insurance Cost Uncertainty → Develop Predictive Models for Accurate Cost Estimation

Limited Transparency in Pricing → Exploratory Analysis to Identify Cost-Driving Factors

Consumer Confusion about Expenses → Interactive ShinyApp for Personalized Cost Predictions

Inefficient Insurance Pricing → Machine Learning Models to Optimize Pricing Accuracy

In order to prepare the dataset for analysis, we first eliminated duplicates, converted categorical variables to factors, and looked for missing values. To examine the data and show the connections between characteristics like age, BMI, smoking status, and region with insurance charges, we utilized R's dplyr and ggplot2 packages.

We constructed a multiple linear regression model to comprehend the impact of these variables on costs. This made it possible for us to account for certain factors and measure the effects of others. We discovered that the biggest factors influencing costs were BMI and smoking.

We created a Shiny app that lets users enter personal data and get an instant insurance bill prediction, along with visual comparisons to national averages, in order to make the results actionable.

DATASET AND METHOD

We got our dataset from Kaggle. This dataset contains 1338 rows, and 7 columns which consists of variables such as age, sex, BMI, children, smoker (yes or no), region (northeast, northwest, southeast, and southwest), and charges.

We utilized this dataset to create a multiple linear regression model to find out which independent variables were the most significant and had the biggest impact on health insurance costs.

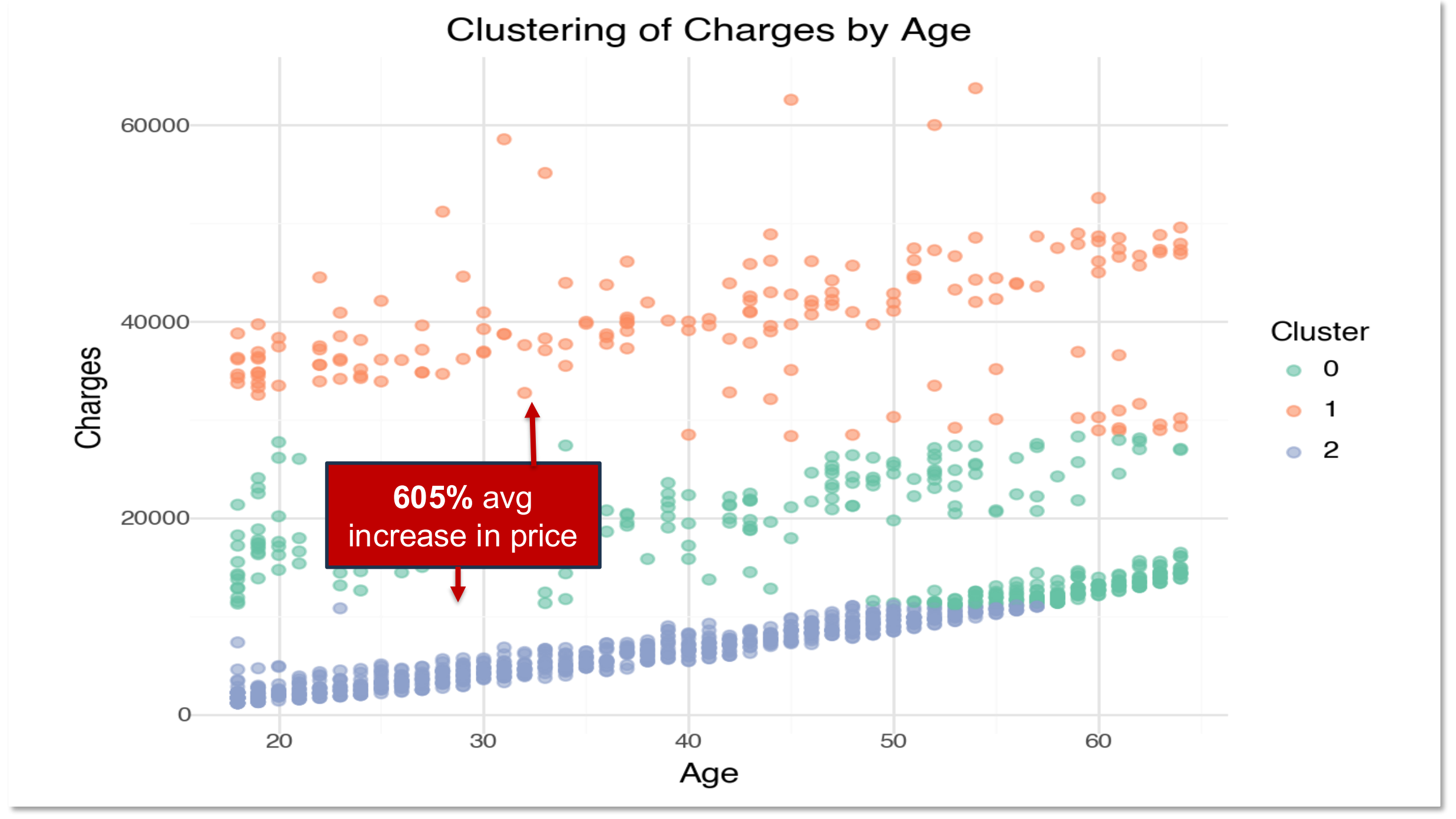
```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11987.42     979.21 -12.242  < 2e-16 ***
age                256.88      11.91  21.577  < 2e-16 ***
bmi                338.73      28.57  11.856  < 2e-16 ***
children           473.86     137.83   3.438 0.000604 ***
smokeryes        23835.21     412.04  57.847  < 2e-16 ***
regionnorthwest   -348.25     476.66  -0.731 0.465152
regionsoutheast  -1034.63     478.71  -2.161 0.030852 *
regionsouthwest   -959.42     477.95  -2.007 0.044914 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7494
F-statistic: 571.8 on 7 and 1329 DF,  p-value: < 2.2e-16
```

INSIGHTS

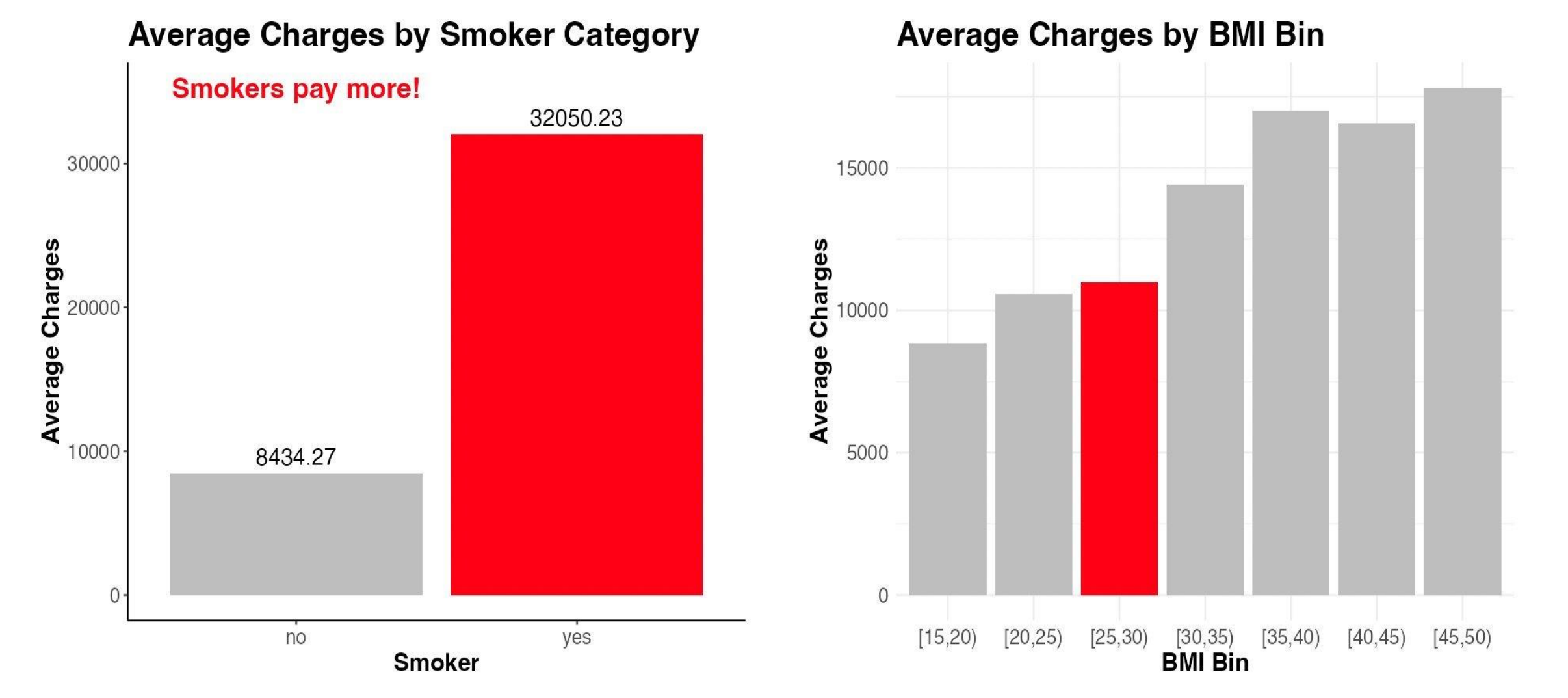
Our exploratory K-Means clustering (k = 3) on age versus annual charges revealed three clear groups—young policyholders with low costs, a middle-aged cohort with moderate premiums, and an older, high-cost segment—highlighting how age interacts non-linearly with charges. Building on this insight, we engineered a three-tier risk_level feature to capture compounding health risks:

Low Risk: *non-smoker & BMI < 30*
High Risk: *either a smoker or BMI ≥ 30*
Super High Risk: *smoker & BMI ≥ 30*



1. The “Super High Risk” group maps to the top-cost K-Means cluster, confirming the combined effect of smoking and obesity on premiums.
2. Adding risk_level, region, children count (peaking at two–three), and other demographics to our regression improved non-linear cost modeling and interpretability in the Shiny app.

VISUALS



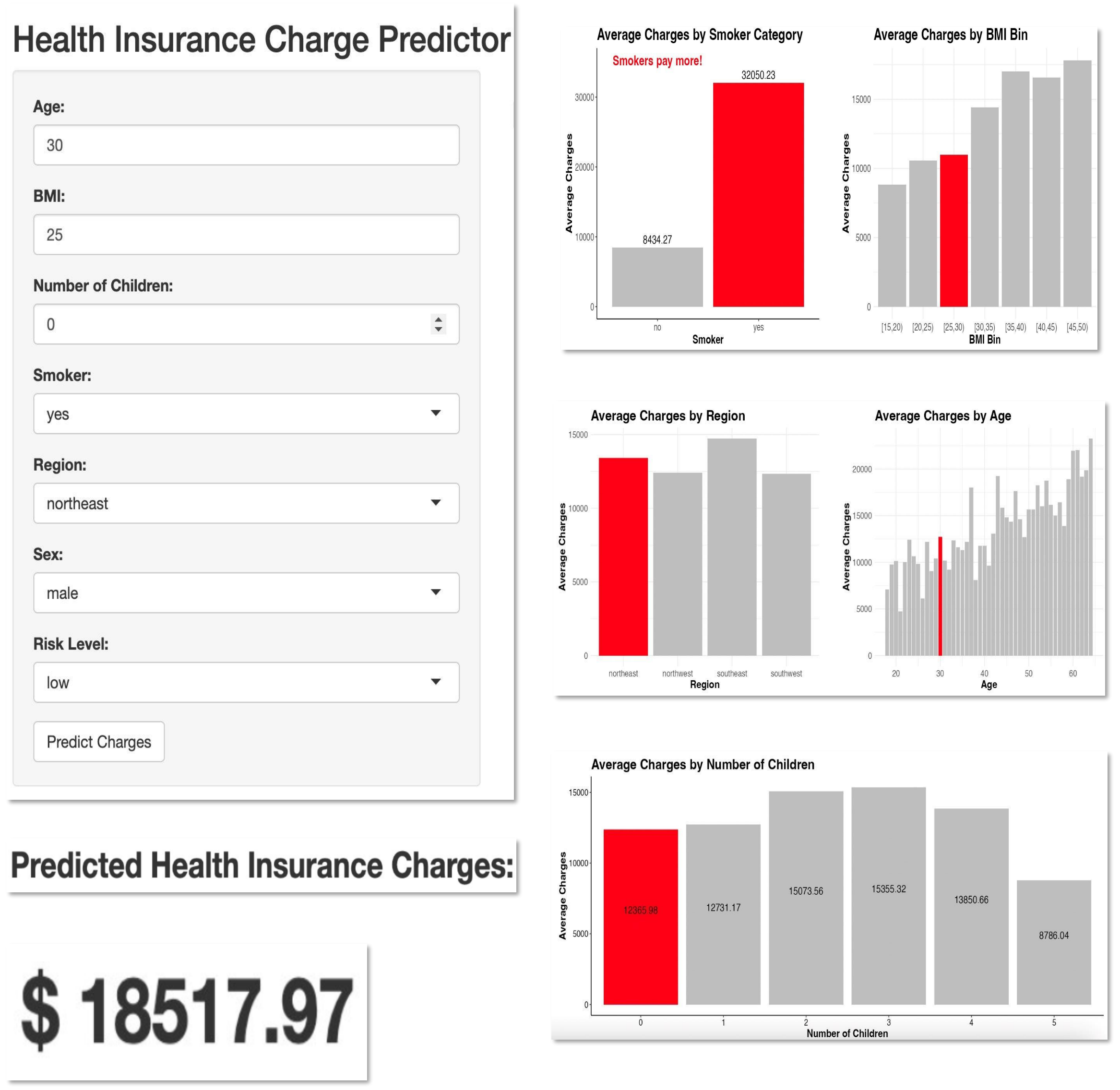
* **Note:** *Smoking is a key variable that drastically increases price*

SHINY APP

We replaced the original R-based linear model with a Python Random Forest regressor. The app now accepts **seven** inputs—Age, BMI, Children, Smoker, Region, Sex, and Risk Level**—where **Risk Level** (“low”, “medium”, “high”) is an ordinal encoding of combined smoking/BMI risk. When the user clicks **Predict Charges**, Shiny calls the Python model and displays:

1. **Predicted annual charge**
2. **Average Charges by Region** (highlights selected region)
3. **Average Charges by Age** (highlights selected age)
4. **Average Charges by Smoker Status** (highlights yes/no)
5. **Average Charges by BMI Bin** (5-point bins, highlights user's bin)
6. **Average Charges by Number of Children** (highlights selected count)

Each plot shows the full population distribution with the user's category in red, giving immediate context to their estimate.



Health Insurance Charge Predictor

Age: 30

BMI: 25

Number of Children: 0

Smoker: yes

Region: northeast

Sex: male

Risk Level: low

Predict Charges

Predicted Health Insurance Charges:

$ 18517.97

Note: *The figures below represent a 30 year-old smoker from the northeast who has no children and a BMI of 25*

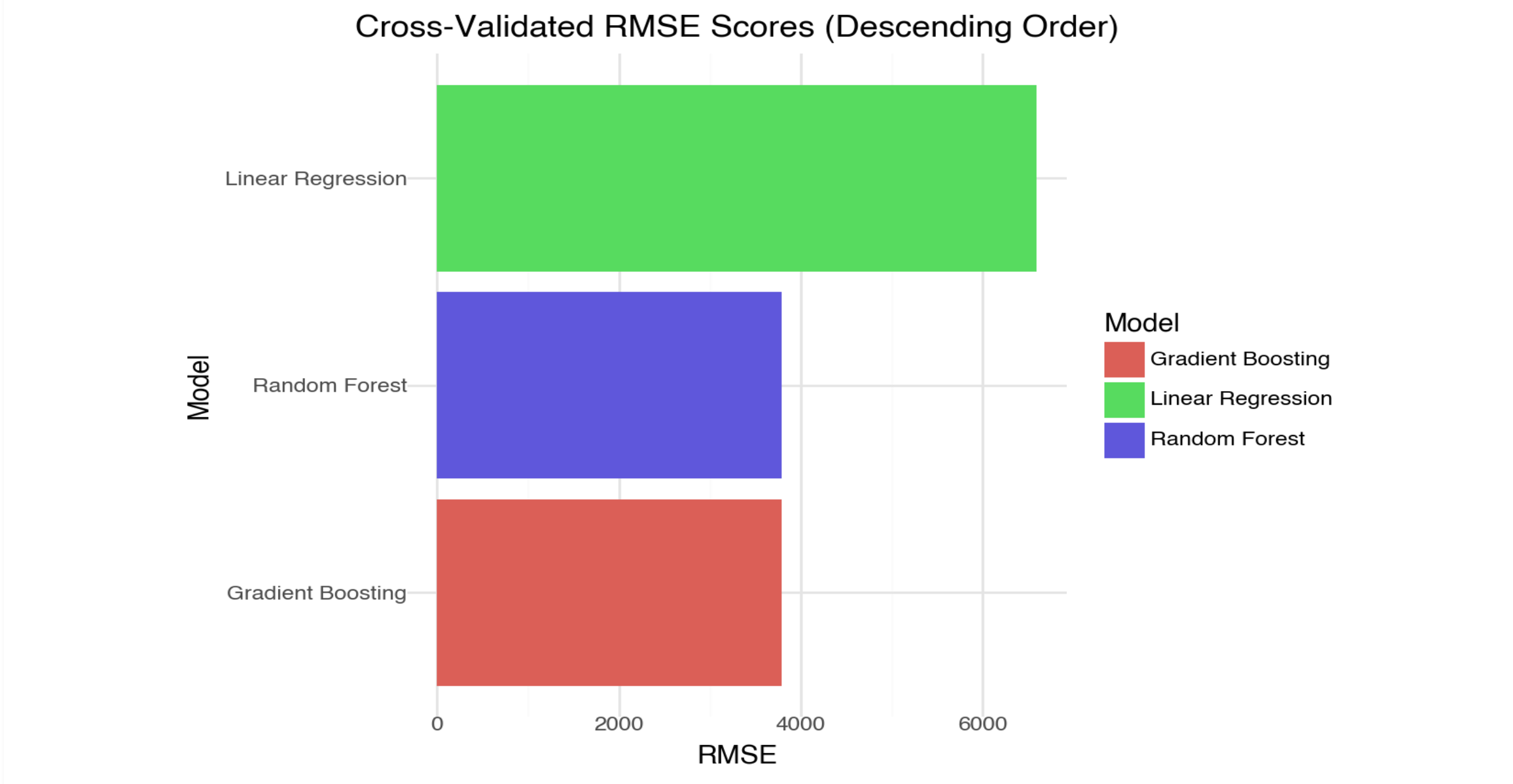
RESULTS

•**Baseline Linear Regression**: RMSE 5,796; R² 0.784

•**Key Effects**: +$3,615 per year of age; +$2,036 per BMI point; +$23,651 for smokers.

•**Non-linear Models (+ risk_level)**: RMSE ↓ > 40%

•**Cross-validated Linear Regression**: RMSE = 6,596.45, R² = 0.8090
•**Random Forest**: RMSE = 3,794.86, R² = 0.9360
•**Gradient Boosting**: RMSE = 3,793.51, R² = 0.9359



Note: *All three models agreed on the top seven predictors—**age**, **BMI**, **smoker_yes/no**, **region_southeast**, **children = 2**, and **risk_level***

CHALLENGES AND IMPLICATIONS

One difficulty we encountered was the dataset's small size and scope (1,338 rows), which might not adequately represent all real-world factors influencing insurance rates, like income level or pre-existing conditions.

Another challenge we encountered during this was the dataset's number of independent variables. Due to only having 5 independent variables that we could use, this dataset might have some shortcomings, and not accurately represent all of the factors that are usually included when it comes to determining health insurance costs.

In spite of this, our model offers insightful information. It draws attention to the financial consequences of health decisions for customers. The software can serve as a tool for risk assessment and education for insurers and policymakers, particularly when it comes to encouraging healthy behaviors like quitting smoking.

CONCLUSION

This experiment shows how healthcare cost transparency can be achieved through predictive modeling. We enable people to better understand and predict their health insurance costs by identifying the main cost factors and creating an intuitive Shiny app.

Using more sophisticated machine learning techniques and growing the dataset in the future may increase prediction accuracy and lead to wider applications.