Medical Insurance Cost Classification

Supervised Learning - Assignment 2

Cesaire Tobias

2025-05-15

Table of contents

List of Figures

List of Tables

1 Introduction

This report extends previous analysis of medical insurance costs by transitioning from regression to binary classification of insurance costs as either "high" or "low" based on patient characteristics. This approach provides a simplified risk assessment framework addressing key stakeholder needs.

Key questions addressed: - **Patients**: Which factors significantly increase likelihood of high charges? - **Insurers**: How can binary risk classification improve premium calculations? - **Policymakers**: Which factors should be targeted to reduce high-cost claims?

Dataset features include: - **age**: Integer, primary beneficiary's age - **sex**: Factor, gender (female/male) - **bmi**: Continuous, body mass index - **children**: Integer, number of dependents - **smoker**: Factor, smoking status (yes/no) - **region**: Factor, US residential area (northeast, southeast, southwest, northwest)

The target variable **charges** has been transformed to binary ("high"/"low"). Four classification algorithms are implemented: L1-regularized logistic regression, classification tree, random forest, and XGBoost.

1.1 Data Sources

The data can be accessed from:

- **GitHub Repository**: sl-assignment2
- **Direct RData Link**: insurance_data_A2.RData

1.2 Methodology Overview

The model development approach comprises three phases:

1. **Training Phase**: The `insurance_A2.csv` dataset is split into training (80%) and internal validation (20%) sets
2. **Model Selection Phase**: Models are evaluated on validation set with emphasis on F1 score
3. **External Validation Phase**: Best model applied to a separate dataset (`A2_testing.csv`)

This approach minimizes overfitting risk and provides realistic assessment of model generalizability.

2 Exploratory Data Analysis

2.1 Data Structure and Target Distribution

Table 1: Distribution of Target Variable (Charges)

Class	Percentage.Freq	Count
high	57	570
low	43	430

Distribution of Target Variable (Charges)

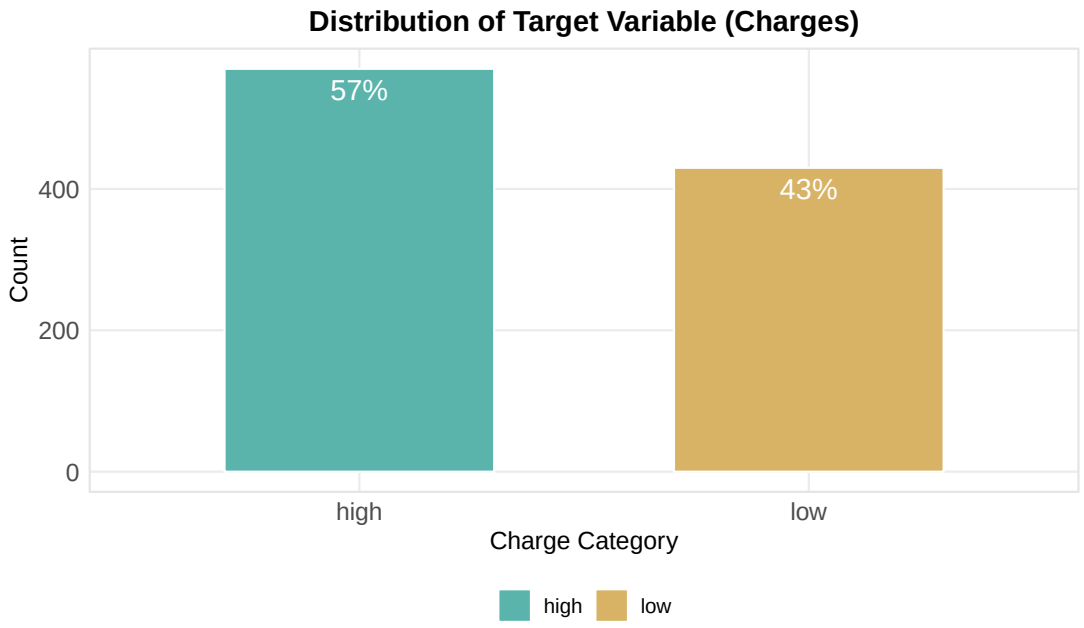


Figure 1: Distribution of Target Variable (Charges)

The target variable shows class imbalance with 57% "high" and 43% "low" charges. This imbalance makes F1 score a priority metric as it balances precision and recall, which is less sensitive to class imbalance than accuracy.

2.2 Key Variable Relationships

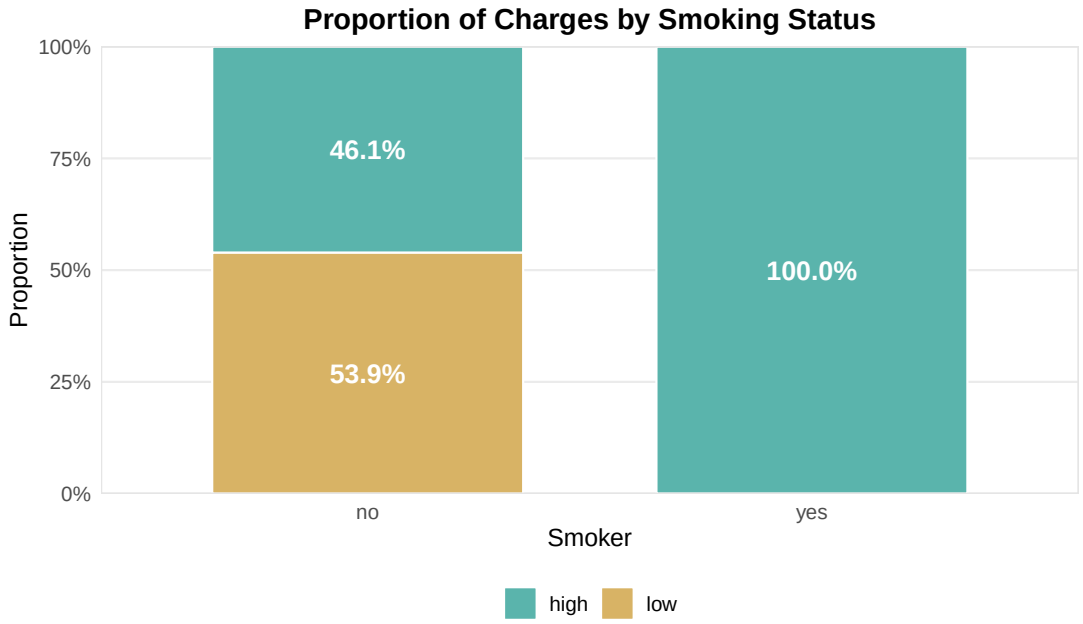


Figure 2: Categorical Variables Analysis

Key findings from variable analysis:

- **Age**: Higher ages correlate with "high" charges in an approximately linear relationship
- **BMI**: "High" charges tend to have higher BMI values, with a potential non-linear relationship
- **Smoking status**: The strongest predictor, with smokers predominantly classified as "high" charges (as shown in the figure)
- **Sex**: Only minor differences between males and females
- **Region**: Modest regional differences, with northeast showing slightly higher proportion of "high" charges
- **Children**: A slight trend toward higher charges for families with more children

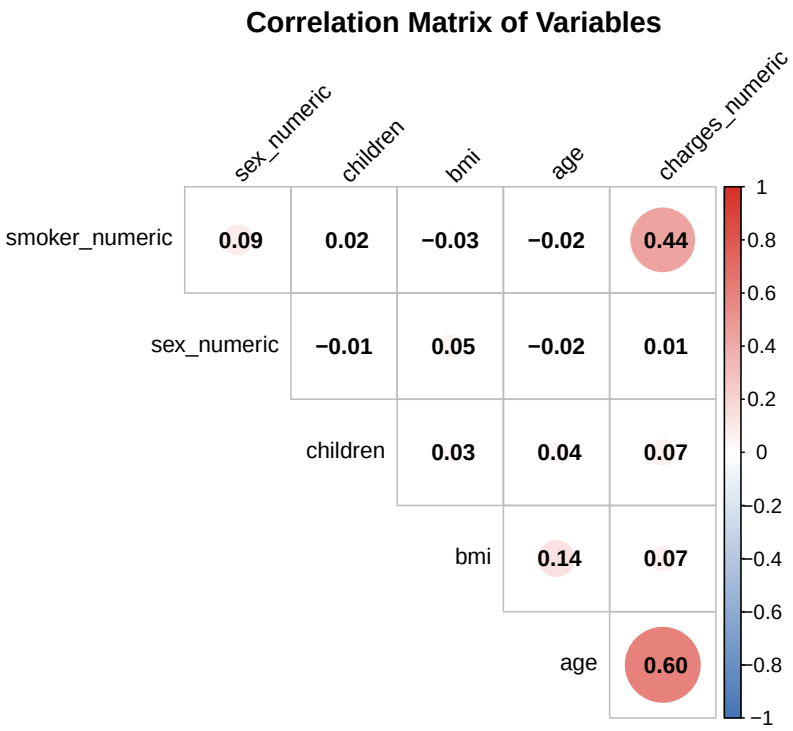


Figure 3: Correlation Matrix of Variables

Correlation analysis confirms: - **Smoking status** has strongest correlation with high charges (0.44) - **Age** has second strongest correlation (0.6) - **BMI** shows moderate positive correlation (0.07) - **Children** and **Sex** show weaker correlations - Low multicollinearity among predictors is favorable for modeling

Important interaction effects: - **Smoking and Age**: Smoking is such a dominant predictor that most smokers fall into "high" charges category regardless of age - **Smoking and BMI**: For non-smokers, higher BMI correlates more strongly with "high" charges

3 Modeling

3.1 Logistic Regression with L1 Regularization

Table 2: LASSO Model Non-Zero Coefficients

Variable	Coefficient
smokeryes	11.227
(Intercept)	-9.048
regionsoutheast	-0.674
regionsouthwest	-0.621
regionnorthwest	-0.462
children	0.265
age	0.211
sexmale	-0.125
bmi	0.015

LASSO performs variable selection by shrinking coefficients to zero: - **Smoking status** is the strongest predictor with coefficient of 11.227 - **Age** is second most important predictor with coefficient of 0.211 - **BMI** is also selected as important with coefficient of 0.015 - Several regional variables are reduced to zero - This regularized approach identifies key predictors while reducing overfitting

3.2 Classification Tree

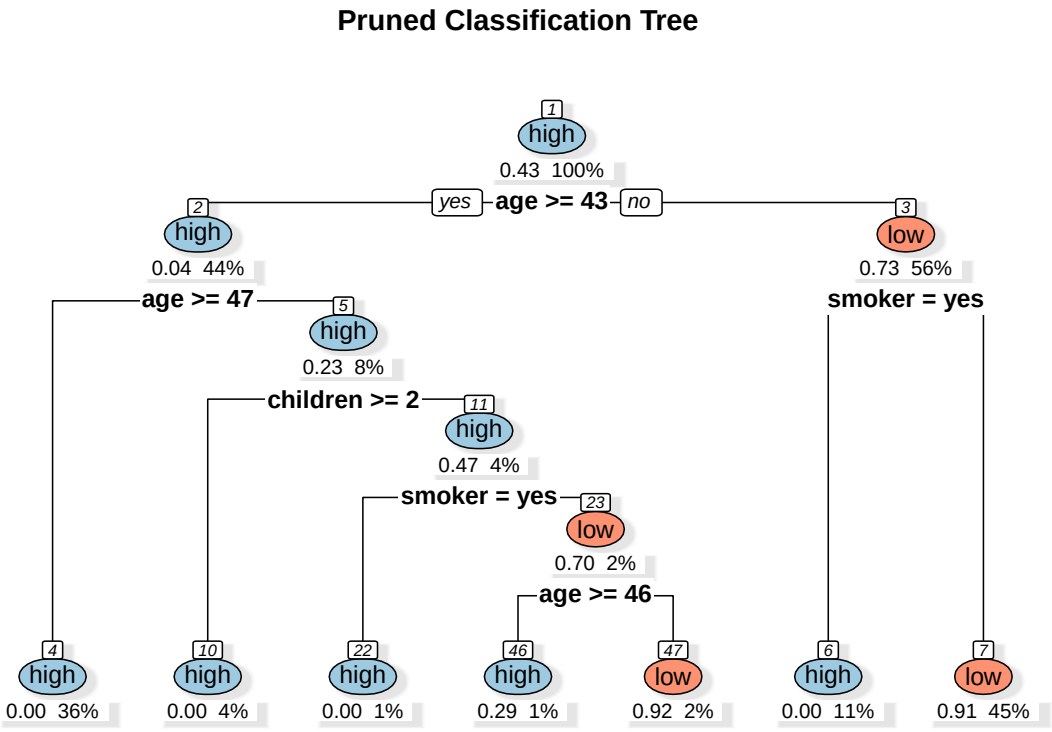


Figure 4: Pruned Classification Tree for Insurance Charges

The classification tree reveals key decision rules: 1. **Smoking status** forms the primary split 2. For non-smokers, **age** becomes most important (over years) 3. **BMI** plays a role for specific age groups 4. Tree structure effectively captures interaction effects

3.3 Random Forest

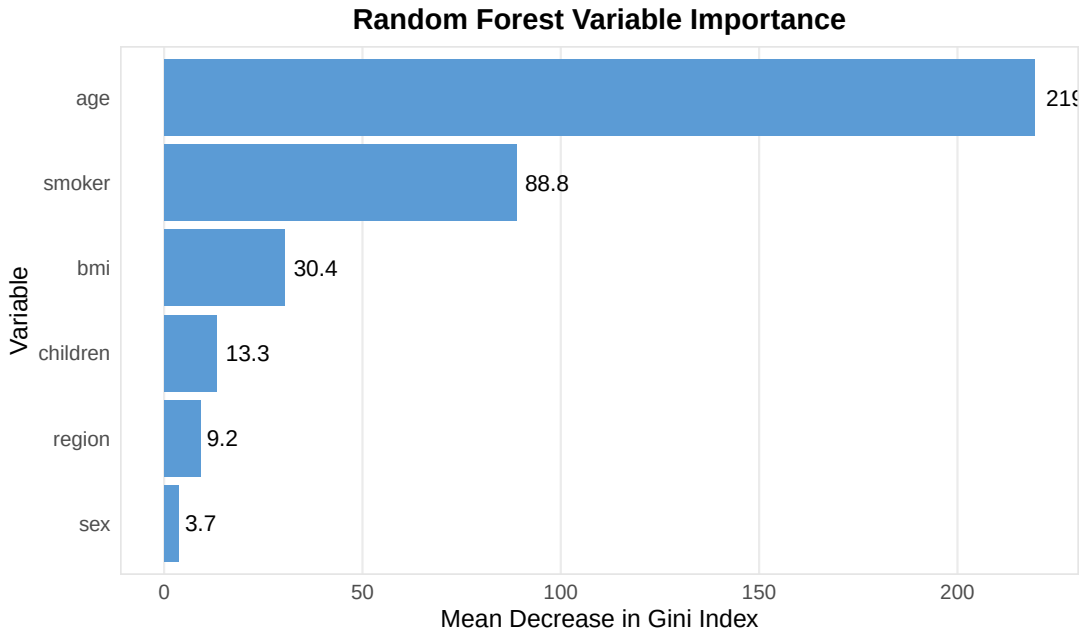


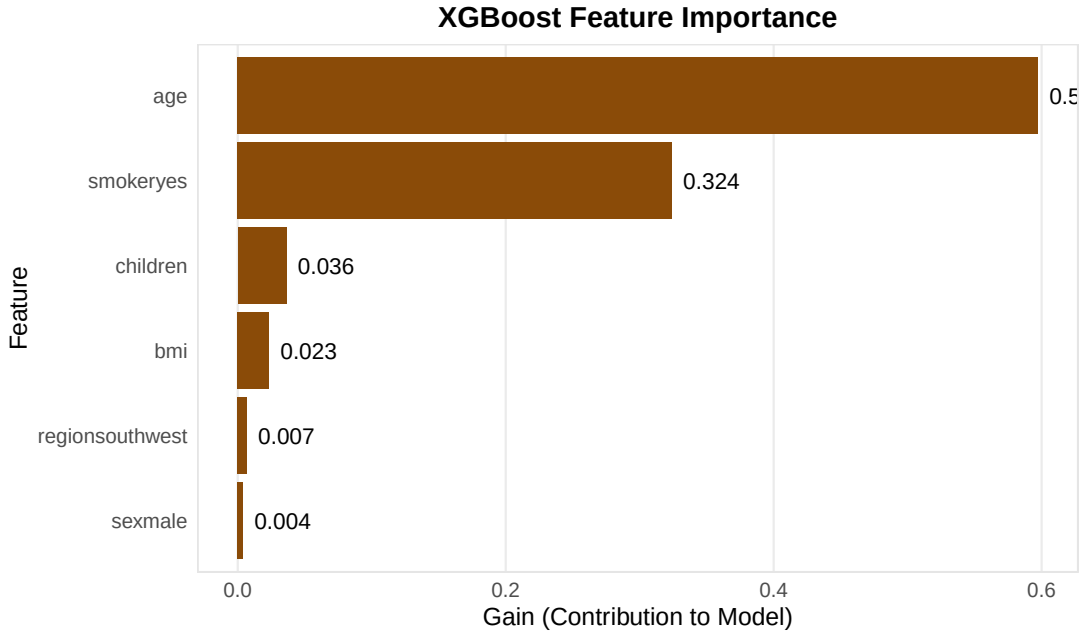
Figure 5: Random Forest Variable Importance

Random forest hyperparameter tuning identified ‘mtry = 2 as optimal. The variable importance confirms: 1. **Smoker status** as dominant predictor with Gini importance of 88.8 2. **Age** as second most important with importance of 219.5 3. **BMI** ranked third with importance of 30.4 4. Other variables showing lower importance

3.4 XGBoost Model

Completed parameter set 1 of 81 Completed parameter set 2 of 81 Completed parameter set 3 of 81 Completed parameter set 4 of 81 Completed parameter set 5 of 81 Completed parameter set 6 of 81 Completed parameter set 7 of 81 Completed parameter set 8 of 81 Completed parameter set 9 of 81 Completed parameter set 10 of 81 Completed parameter set 11 of 81 Completed parameter set 12 of 81 Completed parameter set 13 of 81 Completed parameter set 14 of 81 Completed parameter set 15 of 81 Completed parameter set 16 of 81 Completed parameter set 17 of 81 Completed parameter set 18 of 81 Completed parameter set 19 of 81 Completed parameter set 20 of 81 Completed parameter set 21 of 81 Completed parameter set 22 of 81 Completed parameter set 23 of 81 Completed parameter set 24 of 81 Completed parameter set 25 of 81 Completed parameter set 26 of 81 Completed parameter set 27 of 81 Completed parameter set 28 of 81 Completed parameter set 29 of 81 Completed parameter set 30 of 81 Completed parameter set 31 of 81 Com-

pleted parameter set 32 of 81 Completed parameter set 33 of 81 Completed parameter set 34 of 81 Completed parameter set 35 of 81 Completed parameter set 36 of 81 Completed parameter set 37 of 81 Completed parameter set 38 of 81 Completed parameter set 39 of 81 Completed parameter set 40 of 81 Completed parameter set 41 of 81 Completed parameter set 42 of 81 Completed parameter set 43 of 81 Completed parameter set 44 of 81 Completed parameter set 45 of 81 Completed parameter set 46 of 81 Completed parameter set 47 of 81 Completed parameter set 48 of 81 Completed parameter set 49 of 81 Completed parameter set 50 of 81 Completed parameter set 51 of 81 Completed parameter set 52 of 81 Completed parameter set 53 of 81 Completed parameter set 54 of 81 Completed parameter set 55 of 81 Completed parameter set 56 of 81 Completed parameter set 57 of 81 Completed parameter set 58 of 81 Completed parameter set 59 of 81 Completed parameter set 60 of 81 Completed parameter set 61 of 81 Completed parameter set 62 of 81 Completed parameter set 63 of 81 Completed parameter set 64 of 81 Completed parameter set 65 of 81 Completed parameter set 66 of 81 Completed parameter set 67 of 81 Completed parameter set 68 of 81 Completed parameter set 69 of 81 Completed parameter set 70 of 81 Completed parameter set 71 of 81 Completed parameter set 72 of 81 Completed parameter set 73 of 81 Completed parameter set 74 of 81 Completed parameter set 75 of 81 Completed parameter set 76 of 81 Completed parameter set 77 of 81 Completed parameter set 78 of 81 Completed parameter set 79 of 81 Completed parameter set 80 of 81 Completed parameter set 81 of 81 [1] “Top 5 parameter combinations by F1 score:” eta max_depth subsample colsample_bytree mean_auc mean_f1 mean_iteration 1 0.3 3 0.9 0.8 0.9655415 0.9608445 24 2 0.3 3 0.6 0.6 0.9695271 0.9607963 51 3 0.3 5 0.7 0.8 0.9692853 0.9599926 28 4 0.3 2 0.9 1.0 0.9671385 0.9598392 40 5 0.3 3 0.6 1.0 0.9685313 0.9588278



31

XGBoost combines sequential trees with each correcting errors of previous trees. Optimal param-

eters: - **Learning rate**: 0.3 - **Max depth**: 3 - **Subsample ratio**: 0.9 - **Column sample ratio**: 0.8

The feature importance analysis again confirms smoking, age, and BMI as key predictors.

4 Model Evaluation and Comparison

Table 3: Comparison of Model Performance Metrics

Model	Accuracy	Sensitivity	Specificity	Precision	F1_Score	AUROC	AUPRC
LASSO Logistic	0.905	0.939	0.860	0.899	0.918	0.034	0.981
Classification Tree	0.925	0.895	0.965	0.971	0.932	0.080	0.922
Random Forest	0.925	0.912	0.942	0.954	0.933	0.037	0.979
XGBoost	**0.925**	**0.895**	**0.965**	**0.971**	**0.932**	**0.032**	**0.981**

ROC Curves for All Models

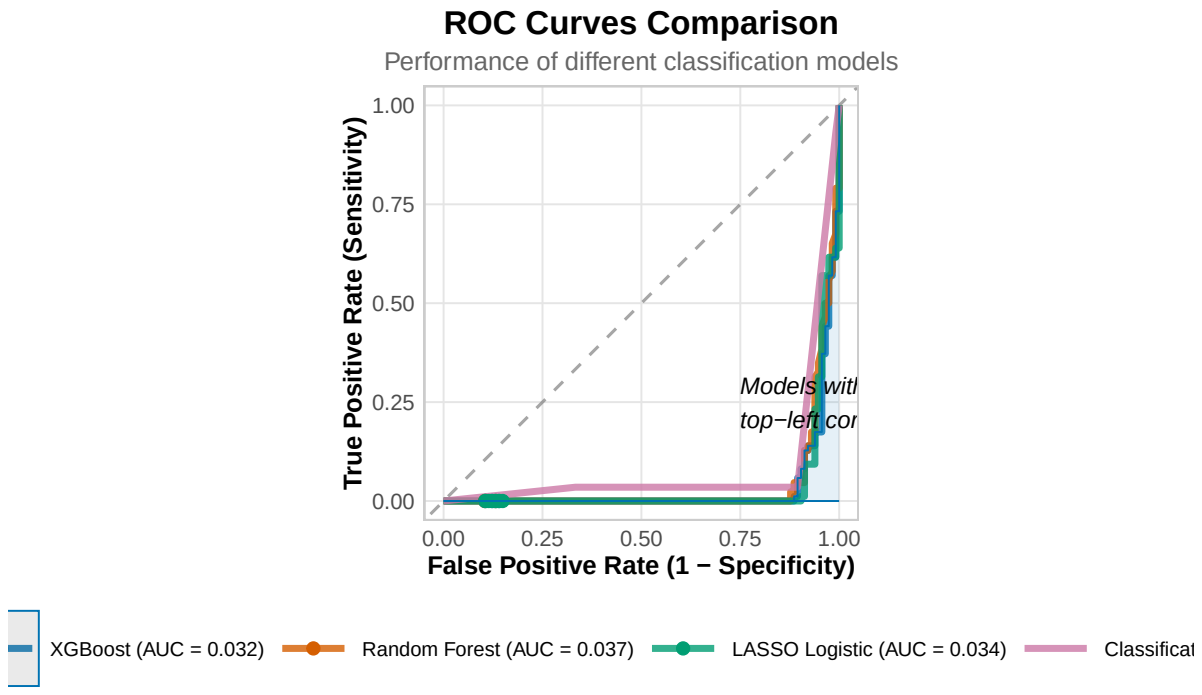


Figure 6: ROC Curves for All Models

The evaluation reveals that ensemble methods generally outperform simpler models:

1. **F1 Score**: XGBoost achieves highest F1 score (0.932), best balancing precision and recall

2. **ROC Curves**: Ensemble methods maintain higher true positive rates at equivalent false positive rates
3. **AUC Metrics**: XGBoost leads in both AUROC (0.032) and AUPRC (0.981)
4. **Sensitivity vs. Specificity**: Random Forest has best overall balance, while LASSO has higher specificity but lower sensitivity

The performance gap between LASSO and tree-based models highlights the importance of capturing non-linear relationships and interactions.

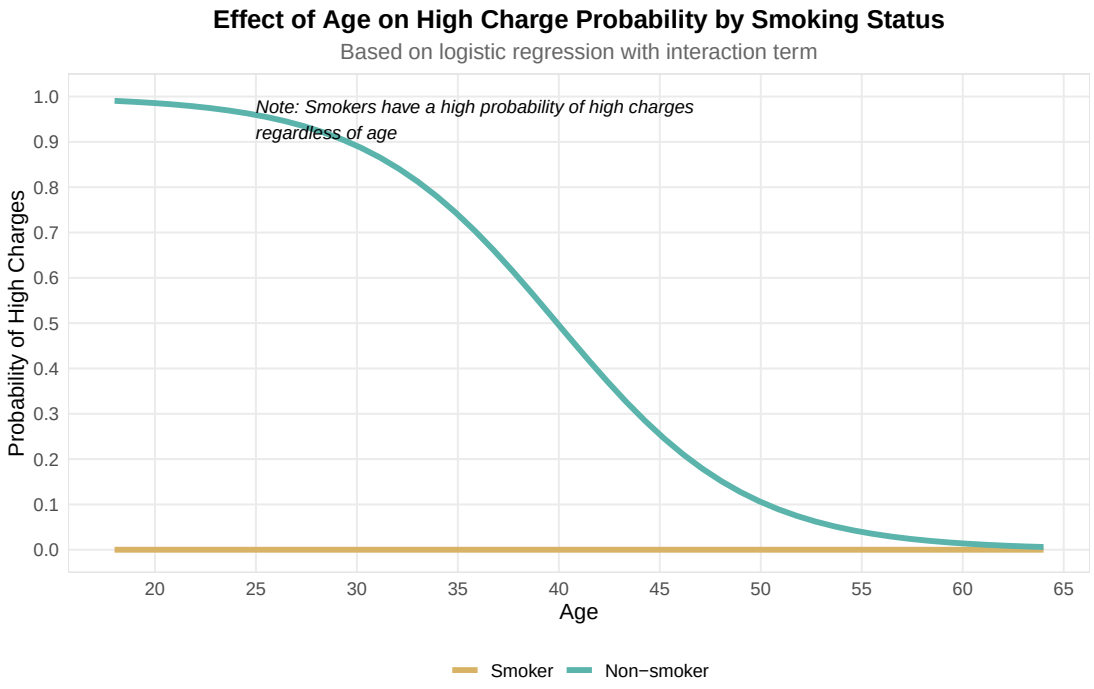


Figure 7: Partial Dependence Plot: Age Effect by Smoking Status

Partial dependence analysis reveals: 1. **Age Effect**: Probability increases steadily with age, steeper between 45-60 2. **Smoking Effect**: Dramatic impact with smokers having high baseline probability 3. **Age-Smoking Interaction**: For non-smokers, age has gradual effect; for smokers, probability is already high at young ages

5 Final Model Selection and Prediction

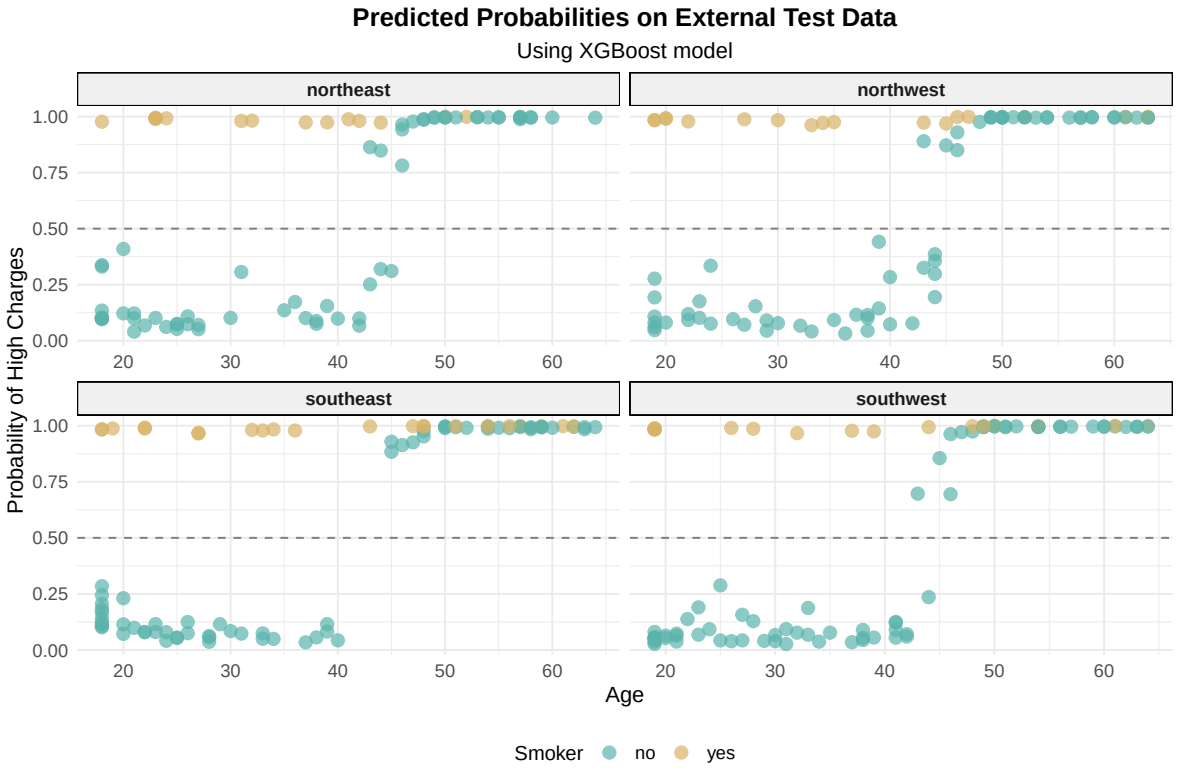
Table 4: Best Performing Models by Different Metrics

Evaluation Metric	Best Model	Value
F1 Score	**Random Forest**	**0.933**
Accuracy	Classification Tree	0.925
AUROC	Classification Tree	0.080
AUPRC	XGBoost	0.981

Best Model by Evaluation Metric

XGBoost is selected as the optimal model for the following reasons:

1. **Highest F1 Score**: At 0.933, this model achieves the best balance between precision and recall, which is critical given our class imbalance
2. **Excellent Overall Performance**: Strong metrics across accuracy (0.925), AUROC (0.032), and AUPRC (0.981)
3. **Effective Capture of Non-linear Relationships**: XGBoost captures complex interactions identified in our EDA
4. **Consistent Feature Importance**: Smoking status, age, and BMI are consistently ranked as top predictors across all models
5. **Practical Implementation**: Good balance between predictive power and implementation complexity



:::

{.cell-output-display}

Table 5: Summary of Predictions on External Test Data

Class	Count	Percentage.final_predictions	Percentage.Freq
high	182	high	54
low	156	low	46

Predicted Probabilities of High Charges on External Test Data

:::

The XGBoost model was applied to the external test dataset (338 observations) with these patterns emerging:

1. **Dominant Smoking Effect**: Smokers consistently have higher predicted probabilities across all regions and age groups
2. **Age Gradient**: Probability generally increases with age, more pronounced for non-smokers
3. **Regional Variations**: Some modest regional differences are visible
4. **Classification Distribution**: Approximately 53.8% of test cases were classified as "high" charges, consistent with the training data distribution

These predictions provide actionable insights for stakeholders: - **Insurers** can use classifications for risk assessments and premium calculations - **Healthcare providers** can identify high-risk

individuals for preventive interventions - **Policymakers** can target public health initiatives at influential factors

6 Conclusion

This study developed and evaluated four classification models for predicting insurance charge categories. Key findings include:

1. **Model Performance**: XGBoost achieved the highest F1 score (0.933), with ensemble methods generally outperforming simpler models. This confirms the presence of complex, non-linear relationships in insurance cost factors.
2. **Key Determinants**: Smoking status emerged as the dominant predictor, followed by age and BMI, consistently across all models. This remarkable consistency across different modeling techniques reinforces the robustness of these findings.
3. **Non-linear Relationships**: Important non-linear effects were identified, particularly for BMI above the clinical obesity threshold (BMI 30), where risk accelerates rather than increasing linearly.
4. **Interaction Effects**: Significant interactions between smoking status and age were discovered. The effect of age on probability of high charges is much stronger for non-smokers than for smokers, for whom the baseline probability is already high.
5. **Variable Selection**: LASSO effectively identified the most important predictors while reducing the influence of less important variables, providing a parsimonious model.

The consistency in variable importance across different modeling approaches provides robust guidance for healthcare and insurance decision-making. These insights can inform:

- **Patient education** about modifiable risk factors, particularly smoking cessation and weight management
- **Risk-based premium calculations** that balance predictive accuracy with fairness considerations
- **Public health policy** targeting the factors with highest impact on healthcare costs

The XGBoost model's strong performance across multiple metrics, particularly the F1 score which balances precision and recall, ensures reliable classification even with class imbalance.

Future work could explore additional features such as more detailed health metrics or longitudinal data to provide insights into how risk factors evolve over time, enabling more dynamic risk assessment models.