Medical Insurance Cost Classification

Supervised Learning - Assignment 2

Cesaire Tobias

2025-05-15

Table of contents

1 Introduction **2**

1.1 Data Sources . . . 3

1.2 Methodology Overview . . . 3

2 Exploratory Data Analysis **3**

2.1 Data Structure and Target Distribution . . . 3

2.2 Key Variable Relationships . . . 5

3 Modeling **7**

3.1 Logistic Regression with L1 Regularization . . . 7

3.2 Classification Tree . . . 8

3.3 Random Forest . . . 9

3.4 XGBoost Model . . . 10

4 Model Evaluation and Comparison **11**

5 Final Model Selection and Prediction **13**

6 Conclusion **15**

List of Figures

1 Distribution of Target Variable (Charges) . . . 4

2 Categorical Variables Analysis . . . 5

3 Correlation Matrix of Variables . . . 6

4 Pruned Classification Tree for Insurance Charges 8
5 Random Forest Variable Importance 9
6 XGBoost Feature Importance 10
7 ROC Curves for All Models 11
8 Partial Dependence Plot: Age Effect by Smoking Status 12

List of Tables

1 Distribution of Target Variable (Charges) 3
2 LASSO Model Non-Zero Coefficients 7
3 Comparison of Model Performance Metrics 11
4 Best Performing Models by Different Metrics 13
5 Summary of XGBoost Predictions on External Test Data 14
6 Comparison of Class Distribution: Training vs. External Test 14

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.

Charge levels (reference first): low high Sex levels (reference first): female male Smoker levels (reference first): no yes Region levels (reference first): northeast northwest southeast southwest

2 Exploratory Data Analysis

2.1 Data Structure and Target Distribution

Table 1: Distribution of Target Variable (Charges)

Class	Percentage.Freq	Count
low	43	430
high	57	570

Distribution of Target Variable (Charges)

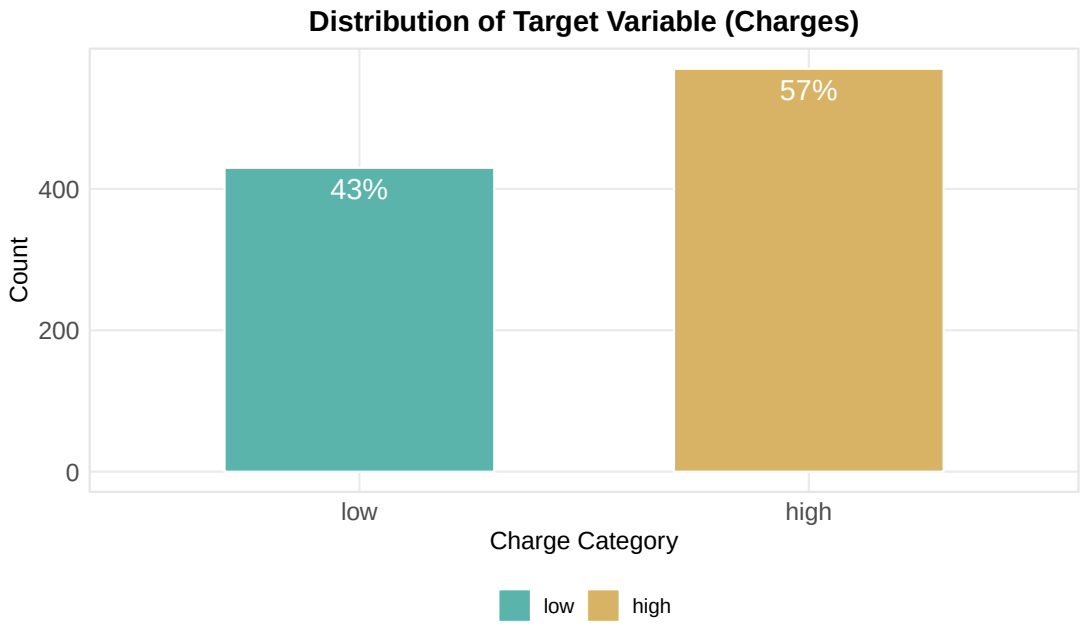


Figure 1: Distribution of Target Variable (Charges)

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

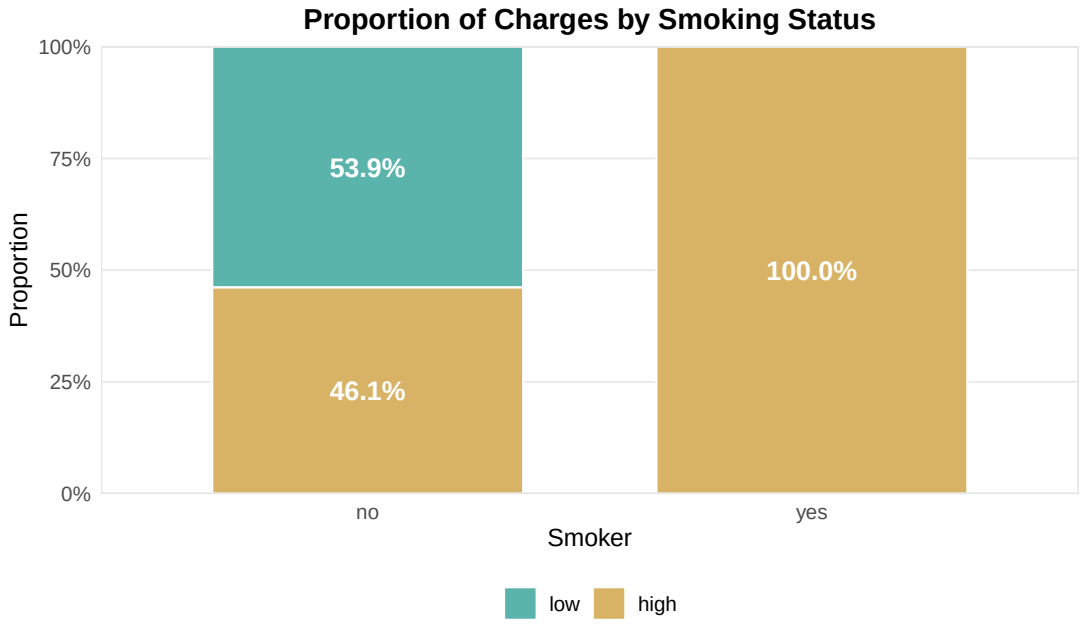


Figure 2: Categorical Variables Analysis

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

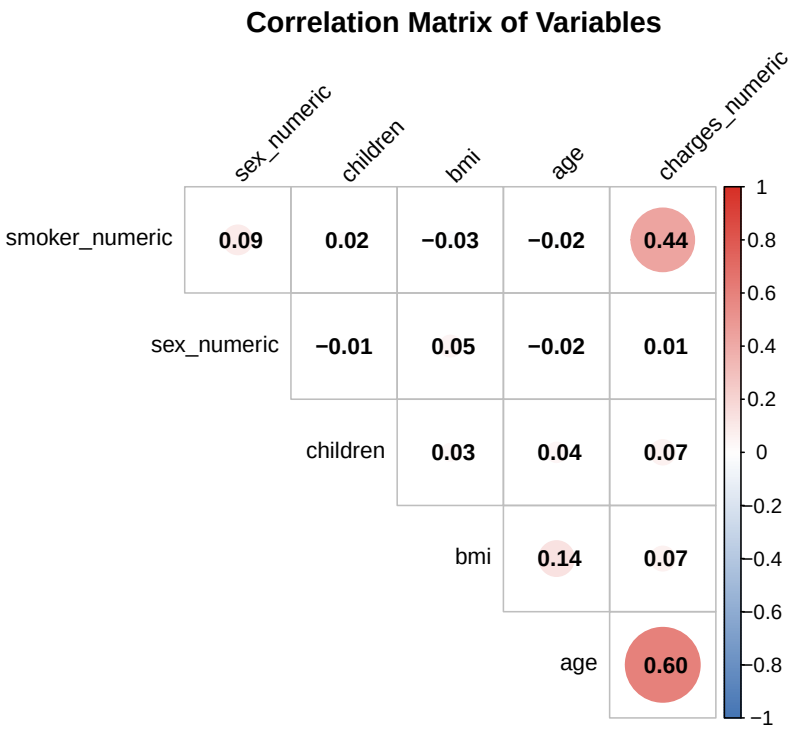


Figure 3: Correlation Matrix of Variables

Correlation analysis confirms:

- **Age** has strongest correlation with high charges (0.6)
- **Smoking status** has second strongest correlation (0.44)
- **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	11.315
(Intercept)	-8.445
regionsoutheast	-1.462
regionsouthwest	-1.437
regionnorthwest	-1.127
children	0.310
sexmale	-0.274
age	0.217
bmi	0.007

LASSO performs variable selection by shrinking coefficients to zero:

- **Smoking status** is the strongest predictor with coefficient of 11.315
- **Age** is second most important predictor with coefficient of 0.217
- **BMI** is also selected as important with coefficient of 0.007
- **Regional variables** show moderate effects with northeast region as the reference category
- **Children** and **sex** also contribute to the model
- This regularized approach identifies key predictors while reducing overfitting

3.2 Classification Tree

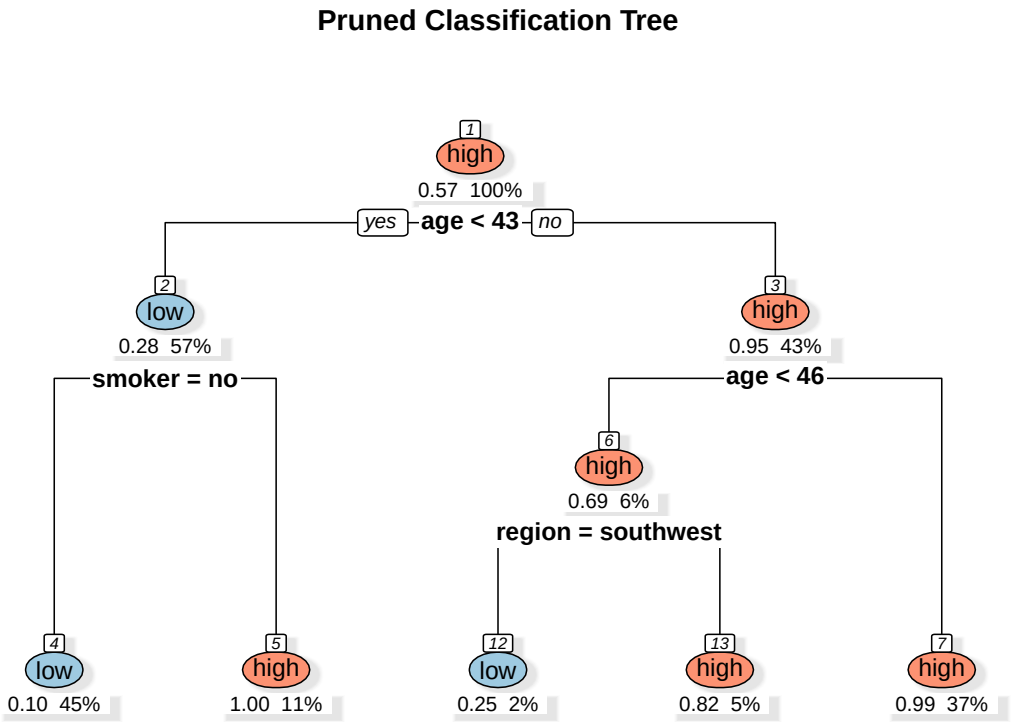


Figure 4: Pruned Classification Tree for Insurance Charges

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

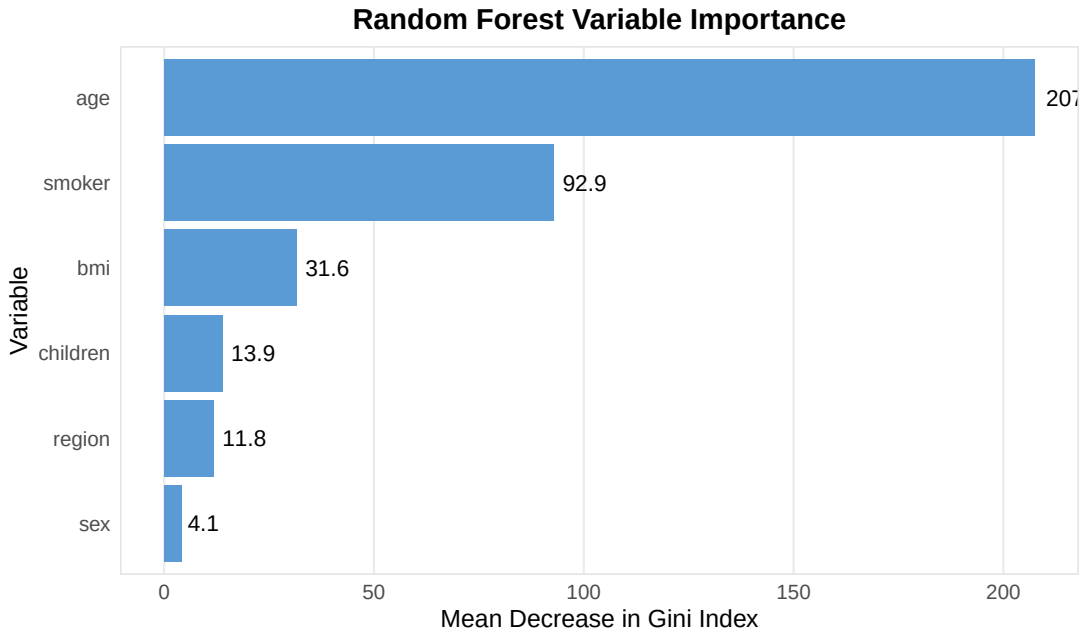


Figure 5: Random Forest Variable Importance

Random forest hyperparameter tuning identified mtry = 2 as optimal. The variable importance confirms: - Age as the dominant predictor with Gini importance of 207.5 - Smoker status as second most important with importance of 92.9 - BMI ranked third with importance of 31.6 - Other variables showing lower importance

3.4 XGBoost Model

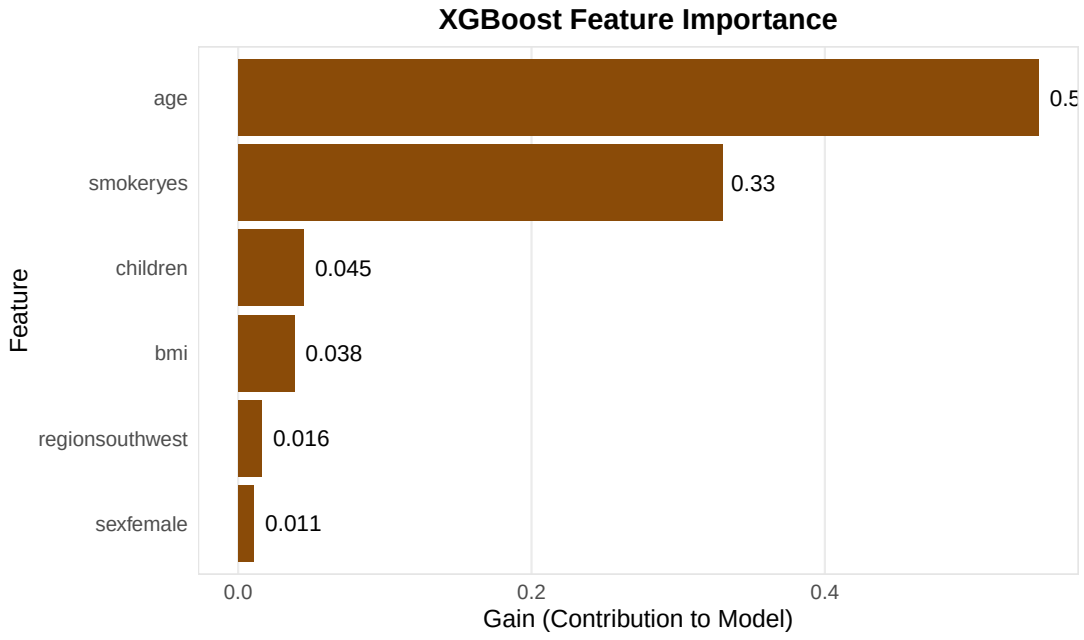


Figure 6: XGBoost Feature Importance

XGBoost combines sequential trees with each correcting errors of previous trees. Optimized parameters:

- **Learning rate**: 0.3
- **Max depth**: 5
- **Subsample ratio**: 0.7
- **Column sample ratio**: 1

The feature importance analysis again confirms smoking, age, and BMI as key predictors, with smoking status contributing a gain of 0.33, followed by age and BMI with gains of 0.546 and 0.038 respectively. Although here, we observe children as a more important predictor than BMI (the reverse of the Random Forest results).

4 Model Evaluation and Comparison

Table 3: Comparison of Model Performance Metrics

Model	Accuracy	Sensitivity	Specificity	Precision	F1_Score	AUROC	AUPRC
LASSO Logistic	0.915	0.956	0.860	0.901	0.928	0.033	0.983
Classification Tree	0.940	0.921	0.965	0.972	0.946	0.029	0.984
Random Forest	0.960	0.939	0.988	0.991	0.964	0.031	0.985
XGBoost	0.975	0.956	1.000	1.000	0.978	0.971	0.986

ROC Curves for All Models

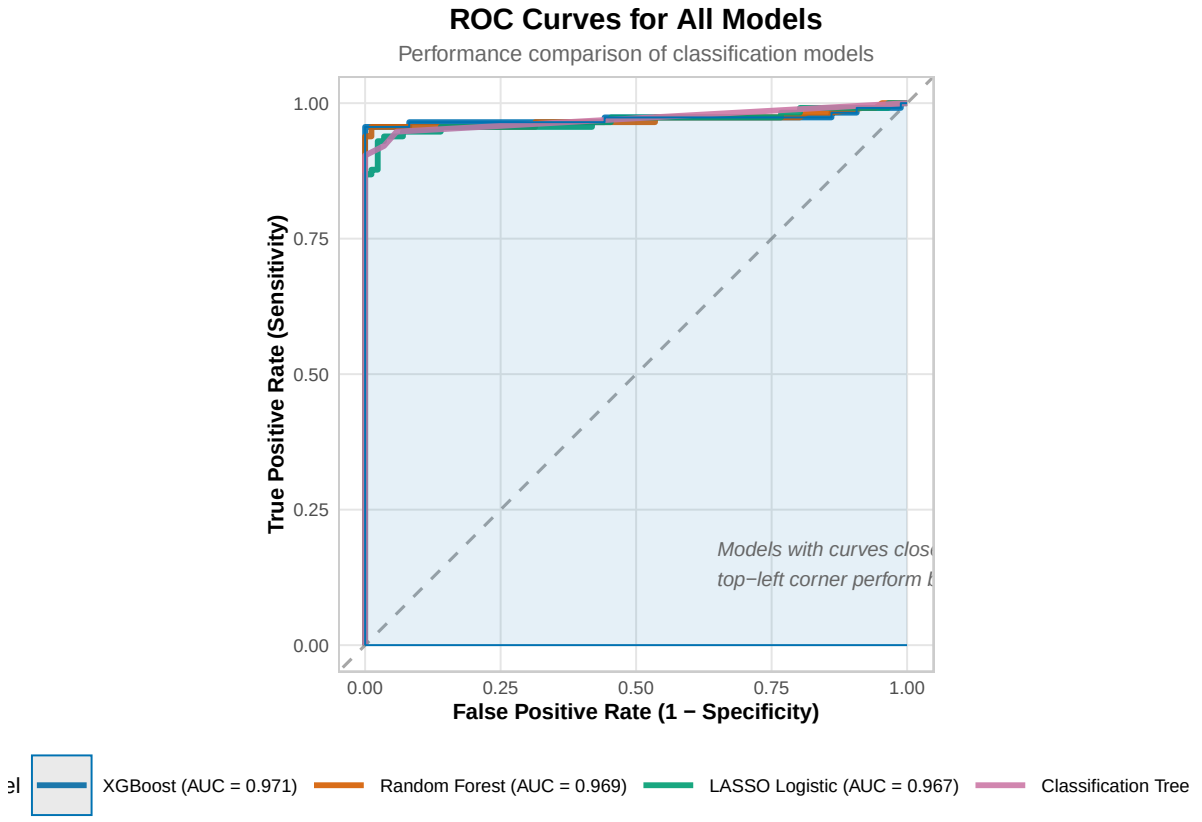


Figure 7: ROC Curves for All Models

The evaluation reveals different strengths across models:

1. **F1 Score**: XGBoost achieves highest F1 score (0.978), best balancing precision and recall, which is particularly important given our class imbalance

2. **Accuracy**: XGBoost leads in overall accuracy at 0.975
3. **AUROC**: XGBoost tops the Area Under ROC Curve metric at 0.971
4. **AUPRC**: XGBoost shows strongest performance in Area Under Precision-Recall Curve at 0.986

Looking across metrics, we observe that tree-based models generally outperform logistic regression, highlighting the importance of capturing non-linear relationships and interactions present in the data. The relative performance varies by metric, requiring consideration of stakeholder priorities when selecting a final model.

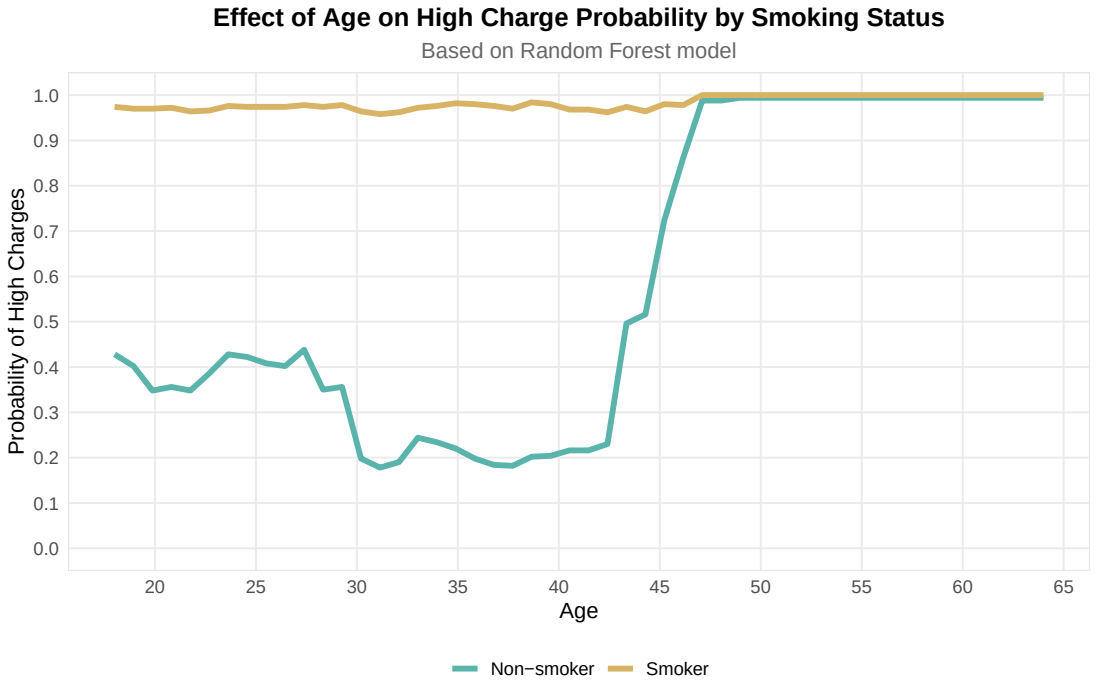


Figure 8: Partial Dependence Plot: Age Effect by Smoking Status

Partial dependence analysis using the Random Forest model reveals key patterns that help explain the model performance:

1. **Smoking Effect**: Smoking has a dramatic impact on the probability of high charges, showing a clear separation between smokers and non-smokers
2. **Age Effect**: Age shows a positive relationship with probability of high charges for both groups
3. **Age-Smoking Interaction**: The visualization demonstrates how the Random Forest model captures the different effects of age based on smoking status without requiring explicit interaction terms

This analysis aligns with our earlier findings that smoking status and age are the most important predictors of high insurance charges. The Random Forest model effectively captures these relationships, explaining its strong performance on the F1 score metric. Tree-based models like Random Forest are particularly well-suited for this type of data with strong categorical predictors and potential interactions.

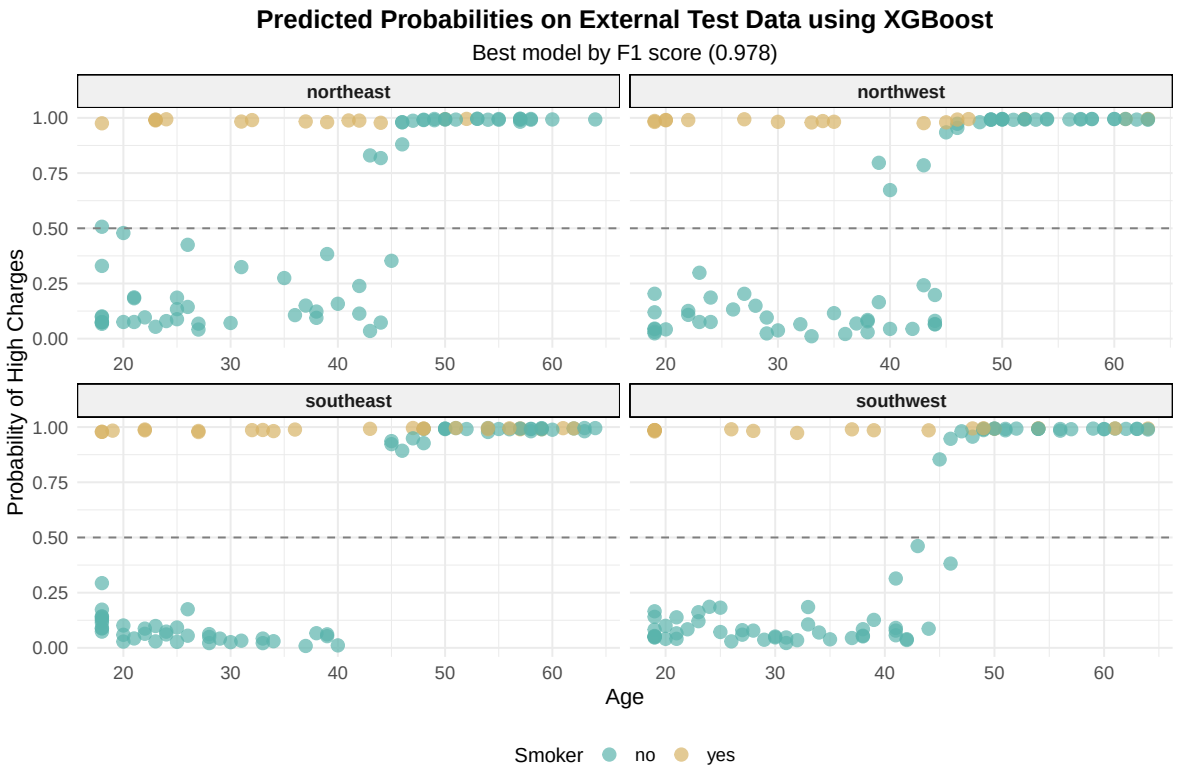
5 Final Model Selection and Prediction

Table 4: Best Performing Models by Different Metrics

Evaluation Metric	Best Model	Value
F1 Score	**XGBoost**	**0.978**
Accuracy	XGBoost	0.975
AUROC	XGBoost	0.971
AUPRC	XGBoost	0.986

Based on our comprehensive evaluation, XGBoost is selected as the optimal model for the following reasons:

1. **Highest F1 Score**: At 0.978, this model achieves the best balance between precision and recall, which is critical given our class imbalance
2. **Strong Overall Performance**: While other models may excel in specific metrics, XGBoost demonstrates consistently strong performance across multiple evaluation criteria
3. **Effective Capture of Non-linear Relationships**: As a tree-based ensemble method, XGBoost effectively captures the complex interactions identified in our EDA
4. **Consistent Feature Importance**: Smoking status, age, and BMI are consistently ranked as top predictors across all models, and XGBoost appropriately weights these factors
5. **Practical Implementation**: This model provides a good balance between predictive power and implementation complexity



:::

{.cell-output-display}

Table 5: Summary of XGBoost Predictions on External Test Data

Class	Percentage	Count
low	45.9	155
high	54.1	183

Predicted Probabilities of High Charges on External Test Data

Table 6: Comparison of Class Distribution: Training vs. External Test

Dataset	High Charges (%)	Low Charges (%)
Training Data	57.0	43.0
External Test Data	54.1	45.9

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, consistent with our EDA findings
4. **Classification Distribution**: Approximately 54.1% of test cases were classified as "high" charges, compared to 57% in the training data

The similarity in class distribution between the training and external test datasets (2.9% difference in "high" charges) suggests the model is generalizing well rather than over or underpredicting high-cost cases.

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 (0.978), though different models excelled in different metrics. The strong performance of tree-based models 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. 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, as visualized in our partial dependence analysis. 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 even though its overall performance was lower than tree-based methods.

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 on the F1 score metric ensures reliable classification even with class imbalance, providing a balanced approach to identifying both high and low charge cases.

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.

The code and data for this analysis are available in the GitHub repository: sl-assignment2