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 (external to this analysis) from a continuous dollar amount to binary ("high"/"low"). Four classification algorithms are implemented: L1-regularized (LASSO) 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

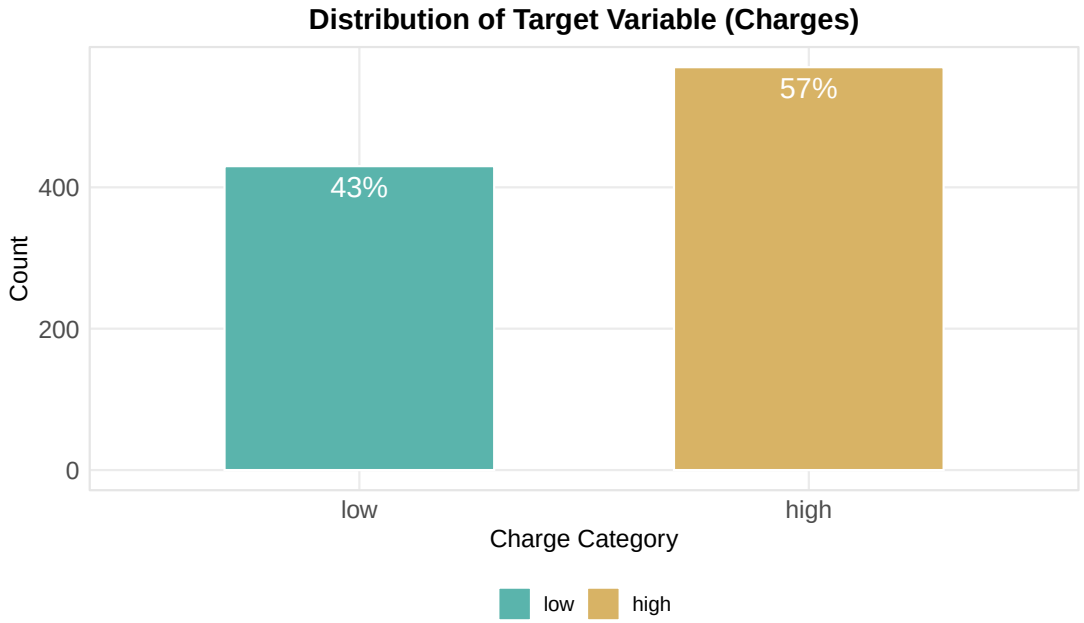
The model development approach comprises three phases:

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 the internal validation set with emphasis on the F1 score.
3. **External Validation Phase**: The best model is then applied to a separate dataset (`A2_testing.csv`).

This approach minimizes over-fitting risk and provides a more robust assessment of model generalizability.

2 Exploratory Data Analysis

2.1 Data Structure and Target Distribution



The target variable shows class imbalance with 43% "low" and 57% "high" 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

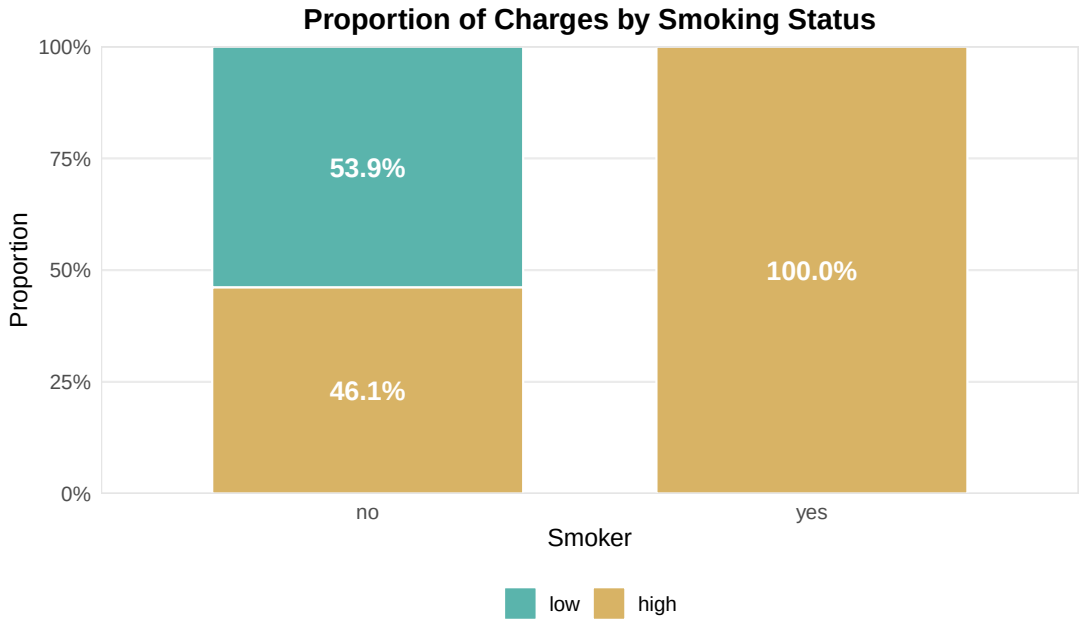


Figure 1: Proportion of Charges by Smoking Status

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 categorical predictor, with 100% of smokers classified as "high" charges compared to only 46.1% of non-smokers (as shown in Figure 1)
- **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

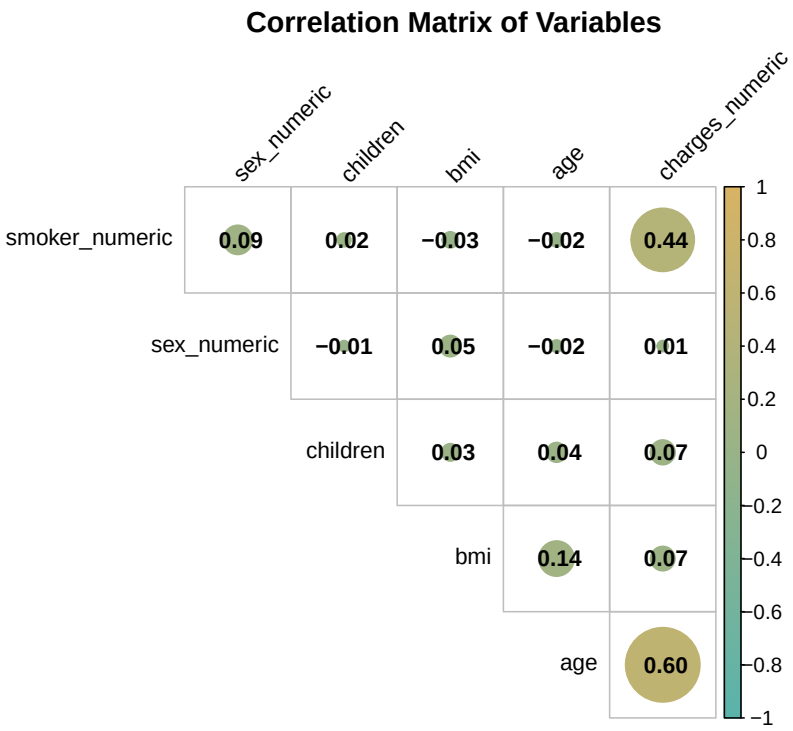


Figure 2: Correlation Matrix of Variables

Correlation analysis confirms:

- **Age** has the strongest correlation with high charges (0.6)
- **Smoking status** has the 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 all 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 1: Full Logistic Regression Model Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-8.4912	0.9602	-8.8432	0.0000	***
age	0.2181	0.0167	13.0769	0.0000	***
sexmale	-0.2805	0.2630	-1.0666	0.2862	
bmi	0.0077	0.0229	0.3378	0.7355	
children	0.3141	0.1120	2.8033	0.0051	**
smokeryes	22.3651	620.0146	0.0361	0.9712	
regionnorthwest	-1.1476	0.3856	-2.9765	0.0029	**
regionsoutheast	-1.4862	0.3945	-3.7674	0.0002	***
regionsouthwest	-1.4614	0.4016	-3.6392	0.0003	***

Note: Significance codes: *** p<0.001, ** p<0.01, * p<0.05

The full logistic regression model before regularization reveals several significant predictors. Most notably, smoking status exhibits an extremely large positive coefficient (22.365), indicating smokers have dramatically higher odds of being in the "high" charges category. Age also shows a significant positive effect (0.218), with each additional year increasing the log-odds of high charges. Regional variables demonstrate statistically significant negative coefficients compared to the northeast reference region, suggesting geographical variation in insurance costs. However, this unregularized model may be susceptible to overfitting, particularly with our relatively small dataset. To address this concern, we apply L1 regularization (LASSO) to perform variable selection and coefficient shrinkage simultaneously.

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 regularization approach identifies key predictors while reducing overfitting.

The results show:

- **Smoking status** is the strongest predictor with a coefficient of 11.315
- **Regional variables** show moderate negative effects (with reference to northeast): southeast (-1.462), southwest (-1.437), and northwest (-1.127)
- **Children** has a stronger linear effect (0.31) than **age** (0.217)
- **BMI** shows a small but positive effect (0.007)

3.2 Classification Tree

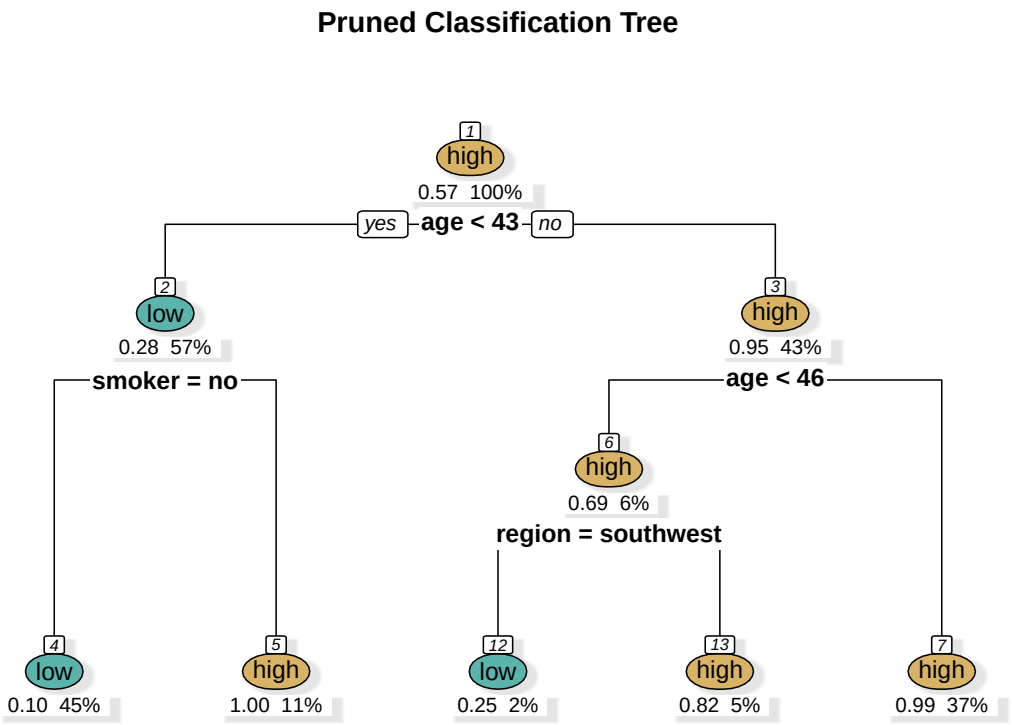


Figure 3: Pruned Classification Tree for Insurance Charges

The classification tree (Figure 3) reveals key decision rules:

1. **Age is the primary predictor**: The model first splits on age, with 43 being a critical threshold.
2. **For younger individuals ($<$ 43)**, smoking status is the most important factor:
    - Non-smokers are generally low risk (90% probability)

- Smokers are high risk with 100% certainty in this model

3. **For older individuals ($\geq$ 43)**:

   - Those 46 and older are almost certainly high risk (99% probability)
   - For the middle age band (43-45), geographical region matters:
     - Being in the southwest region is associated with lower risk
     - Other regions have a higher risk (82% probability)

This tree captures complex interaction effects among variables. The pathways reveal that while smoking is a very strong predictor, age is the primary decision factor in the tree model, with smoking becoming decisive for younger patients.

3.3 Random Forest

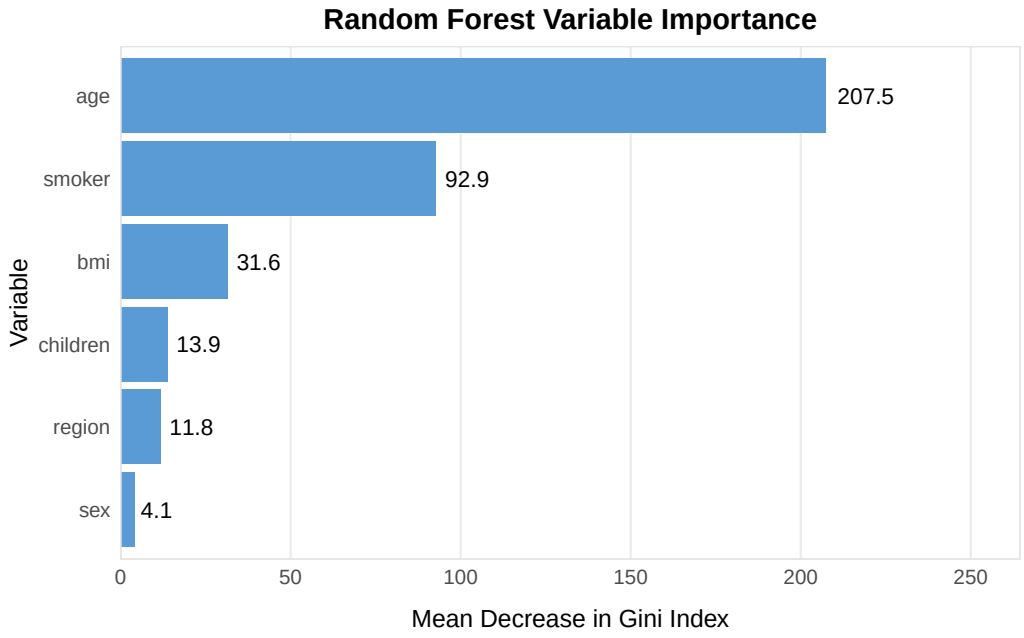


Figure 4: Random Forest Variable Importance

mtry parameter selection rationale:

- With 6 predictor variables, we tested mtry values of 2, 3, and 4, which represent approximately 1/3, 1/2, and 2/3 of the total features.

- This range spans from the default classification setting (sqrt(p) 2.45 for p=6) to a more aggressive feature sampling approach.
- Lower mtry values (e.g., 2) create more diverse trees with less correlation, potentially better for capturing non-linear patterns.
- Higher mtry values (e.g., 4) provide more candidate variables at each split, potentially capturing more complex relationships.
- Through 10-fold cross-validation, mtry = 2 minimized out-of-bag error rate (0.0612), balancing between tree diversity and predictive power

ntree parameter justification:

- We selected 500 trees as this provides sufficient ensemble size to stabilize predictions, balancing computational efficiency with model stability and accuracy.

Random forest hyperparameter tuning identified mtry = 2 as optimal. The variable importance pattern reveals a different perspective than the LASSO model:

- **Age** is ranked as the dominant predictor with Gini importance of 207.5
- **Smoker status** is second most important with importance of 92.9
- **BMI** ranked third with importance of 31.6
- Other variables showing lower importance

This variable importance ranking differs from the LASSO model's ranking, suggesting that when non-linear relationships are considered, age emerges as more important than smoking status across the entire dataset, even though smoking creates a more dramatic binary separation.

3.4 XGBoost Model

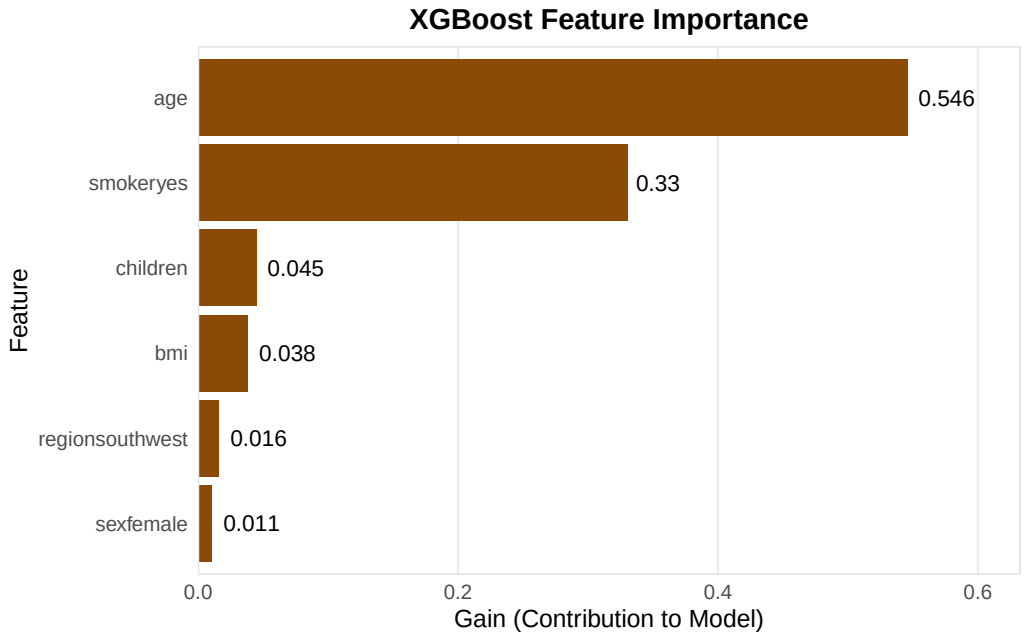


Figure 5: XGBoost Feature Importance

XGBoost combines sequential trees with each correcting errors of previous trees. The optimized parameters from our tuning process are:

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

The feature importance analysis shows that age has the highest gain (0.546), followed by smoking status (0.33). Similar to the Random Forest model, XGBoost ranks age as more important than smoking status in terms of overall contribution to model performance, though the two models differ in their ranking of secondary predictors. In the XGBoost model, children has a gain of 0.045, placing it higher than BMI with a gain of 0.038.

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	0.967	0.983
Classification Tree	0.940	0.921	0.965	0.972	0.946	0.971	0.984
Random Forest	0.960	0.939	0.988	0.991	0.964	0.969	0.985
XGBoost	0.975	0.956	1.000	1.000	0.978	0.971	0.986

ROC Curves for All Models

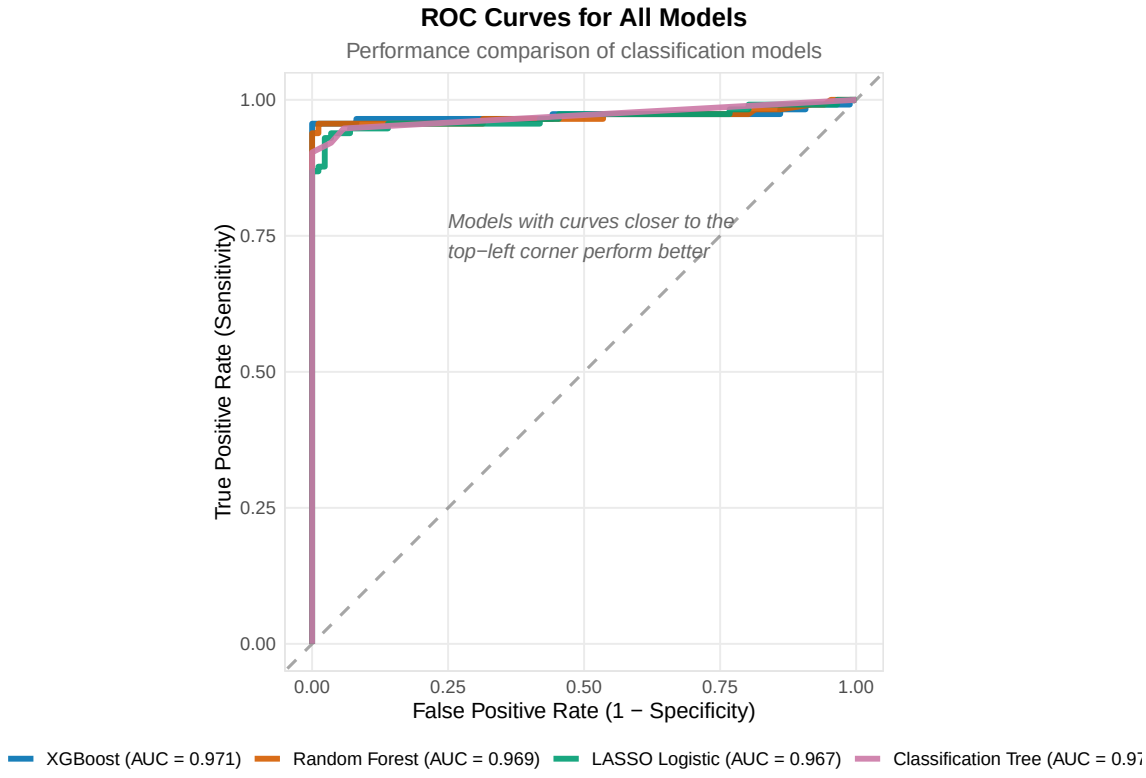


Figure 6: ROC Curves for All Models

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 and Classification Tree both perform strongly in Area Under ROC Curve metrics (around 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.

The following analysis provides insight into why different models rank variables differently. While smoking creates a perfect binary separation for one group of patients, age provides more discriminative power across the entire dataset, particularly for non-smokers. Tree-based models can capture this complex relationship better than linear models like logistic regression.

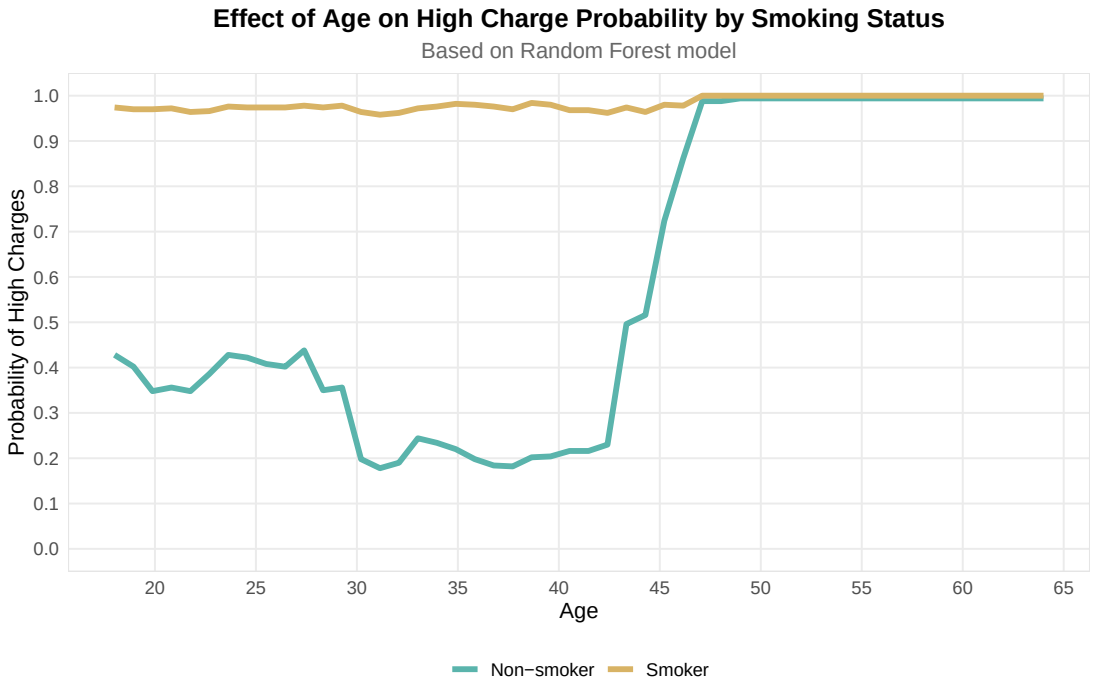


Figure 7: Partial Dependence Plot: Age Effect by Smoking Status

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. Smokers consistently maintain a high probability (near 1.0) regardless of age.

2. **Age Effect**: For non-smokers, age shows a more complex, non-linear relationship with the probability of high charges:

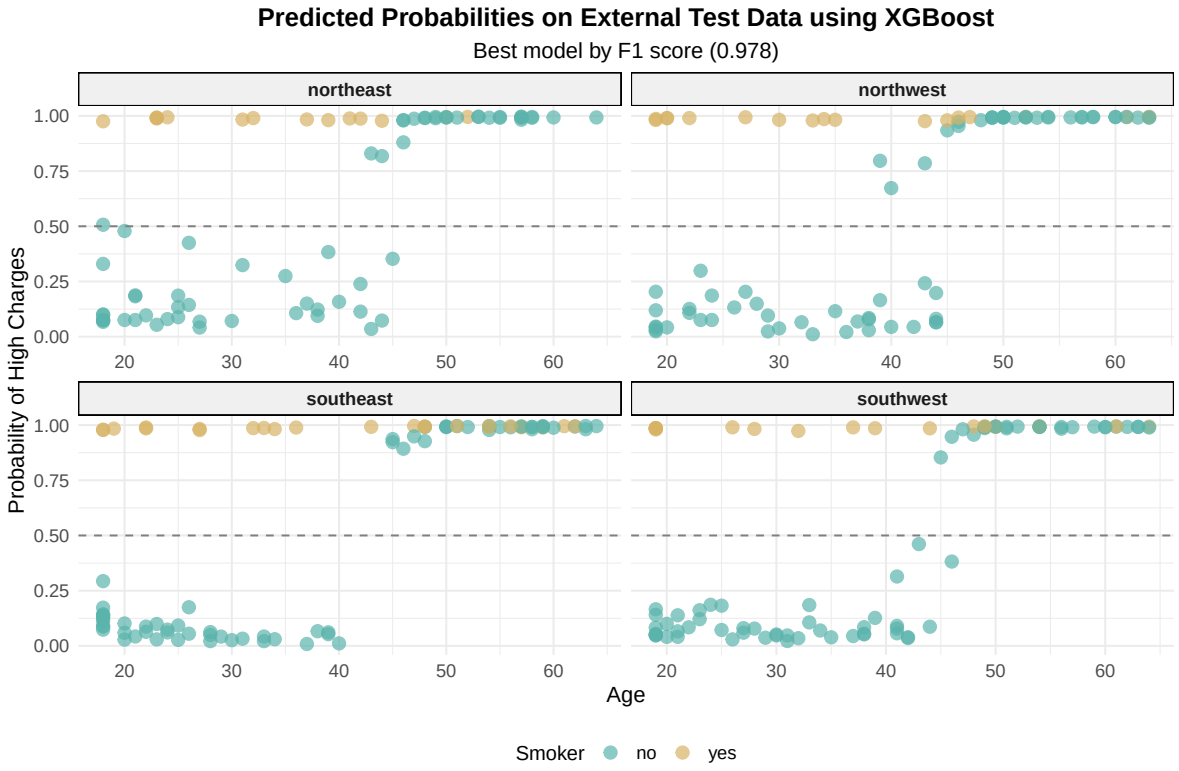
   - Ages 20-40: Moderate probability with some variability
   - Ages 40-45: Sharp increase in probability
   - Ages 45+: Nearly all classified as high charges

3. **Age-Smoking Interaction**: The visualization demonstrates how the Random Forest model captures these different effects without requiring explicit interaction terms. The effect of age is much more pronounced for non-smokers than for smokers (for whom the baseline probability is already high).

5 Final Model Selection and Prediction

Based on our 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**: The model appropriately weighs key predictors (age, smoking status, BMI) in line with our exploratory findings.
5. **Practical Implementation**: This model provides a good balance between predictive power and implementation complexity.



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.

6 Conclusion

This study developed and evaluated four classification models for predicting insurance charge categories (high/low). Key findings include:

1. **Model Performance**: XGBoost achieved the highest F1 score (0.978), though other models were only marginally poorer. The strong performance of tree-based models confirms the presence of complex, non-linear relationships in insurance cost factors.

2. **Key Determinants**: Different models rank predictors differently based on their underlying algorithms:

   - **Linear models** (LASSO): Smoking status emerges as the overwhelmingly dominant predictor
   - **Tree-based models** (Random Forest, XGBoost): Age is ranked as the most important variable, while still recognizing smoking as crucial

3. **Non-linear Relationships**: Important non-linear effects were identified, particularly for age, where risk increases dramatically around age 43-46, and 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.