Medical Insurance Cost Classification

Supervised Learning - Assignment 3

Cesaire Tobias

2025-05-30

Table of contents

1 Data Import and Preparation **2**

2 Introduction **2**

2.1 Data Sources 3

2.2 Methodology Overview 3

3 Exploratory Data Analysis **4**

3.1 Data Structure and Target Distribution 4

3.2 Variable Distributions and Relationships 4

3.3 Key Findings from EDA 6

4 Data Preprocessing **7**

5 Modeling **7**

5.1 Support Vector Machine (SVM) 7

5.1.1 SVM Model Specification and Justification 7

5.2 Neural Network 8

5.2.1 Neural Network Architecture Design 9

5.2.2 Neural Network Architecture and Justification 10

6 Model Evaluation and Comparison **11**

6.1 Model Performance Comparison 12

6.1.1 Key Observations 12

7 Feature Importance Analysis **13**

7.1 SVM Feature Importance . . . 13

7.2 Neural Network Feature Importance . . . 13

7.3 Feature Importance Comparison . . . 13

8 Prediction on Test Data **14**

9 Conclusion **15**

List of Figures

1 Distribution of Numerical Variables . . . 4

2 Relationships Between Features and Target Class . . . 5

3 Interaction Between Smoking Status and BMI . . . 6

4 Neural Network Architecture (Untuned) . . . 9

5 Neural Network Classification Performance . . . 10

6 ROC Curves Comparison . . . 11

7 Confusion Matrices . . . 12

8 Neural Network Feature Importance . . . 13

9 Predicted Probabilities by Age and Smoking Status . . . 14

List of Tables

1 SVM Hyperparameter Tuning Results . . . 7

2 Neural Network Hyperparameter Tuning Results . . . 9

3 Model Performance Comparison . . . 11

4 Model Performance Comparison . . . 12

5 Model Selection for F1 Score Optimization . . . 14

1 Data Import and Preparation

2 Introduction

This report applies supervised machine learning classification techniques to predict whether medical insurance costs will be high or low based on patient characteristics. Two models are developed for this analysis - Support Vector Machine (SVM) and Neural Network.

Key questions addressed:

- **Patients**: Which personal factors significantly increase the likelihood of high insurance charges?
- **Insurers**: How effectively can machine learning models classify high vs. low insurance costs?
- **Policymakers**: Which factors should be targeted to reduce high-cost insurance 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").

2.1 Data Sources

The data can be accessed from:

- **GitHub Repository**: sl-assignment3
- **Direct RData Link**: insurance_data_A2.RData

The dataset consists of 1000 observations with 0 missing values in the training data and 0 missing values in the test data, indicating complete datasets.

2.2 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`).

3 Exploratory Data Analysis

3.1 Data Structure and Target Distribution

The target variable shows class distribution with 43% "low" and 57% "high" charges. This relatively balanced distribution means that standard evaluation metrics like accuracy are appropriate, though we will still prioritize the F1 score as it balances precision and recall.

3.2 Variable Distributions and Relationships

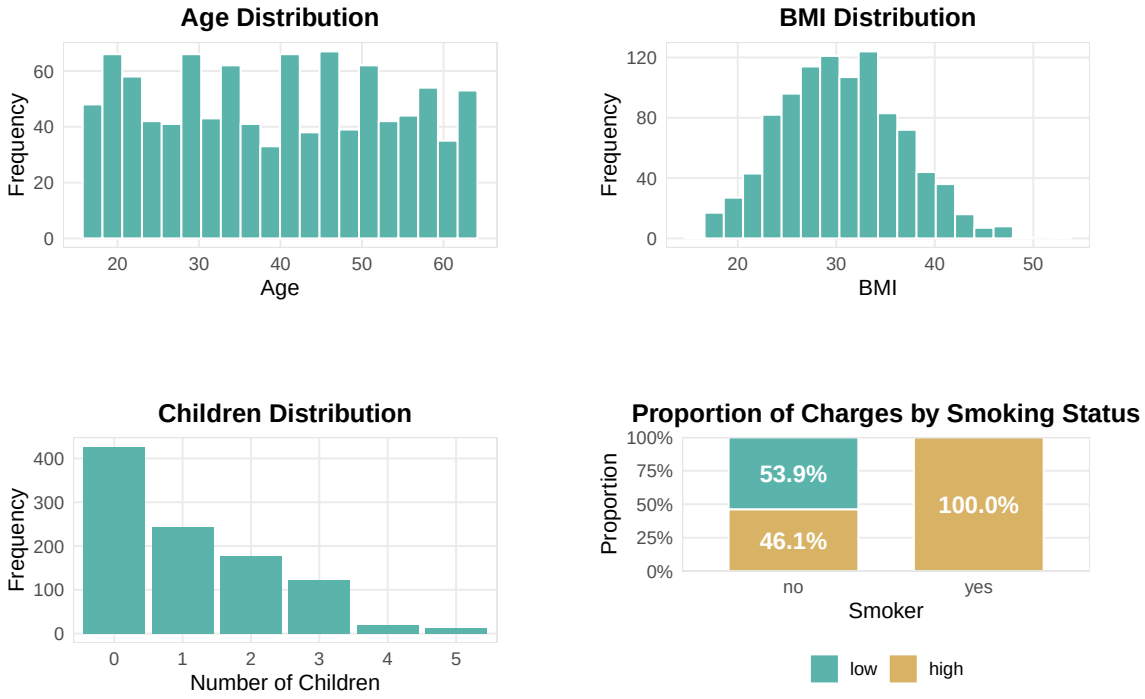

Figure 1: Distribution of Numerical Variables

- **Age**: Fairly uniform across the adult age range (18-65)
- **BMI**: Centered around 25-35, with most values in the overweight to obese range
- **Children**: Highly skewed, with most individuals having 0-2 children
- **Smoking Status**: Shows a dramatic relationship with charges - 100% of smokers are classified as "high" charges compared to only 46.1% of non-smokers

Correlation analysis indicates:

- **Age** has the strongest correlation with high charges (0.6)

- **Smoking status** has the second strongest correlation with high charges (0.44)
- **BMI** shows moderate positive correlation (0.07)
- **Children** and **Sex** show weaker correlations

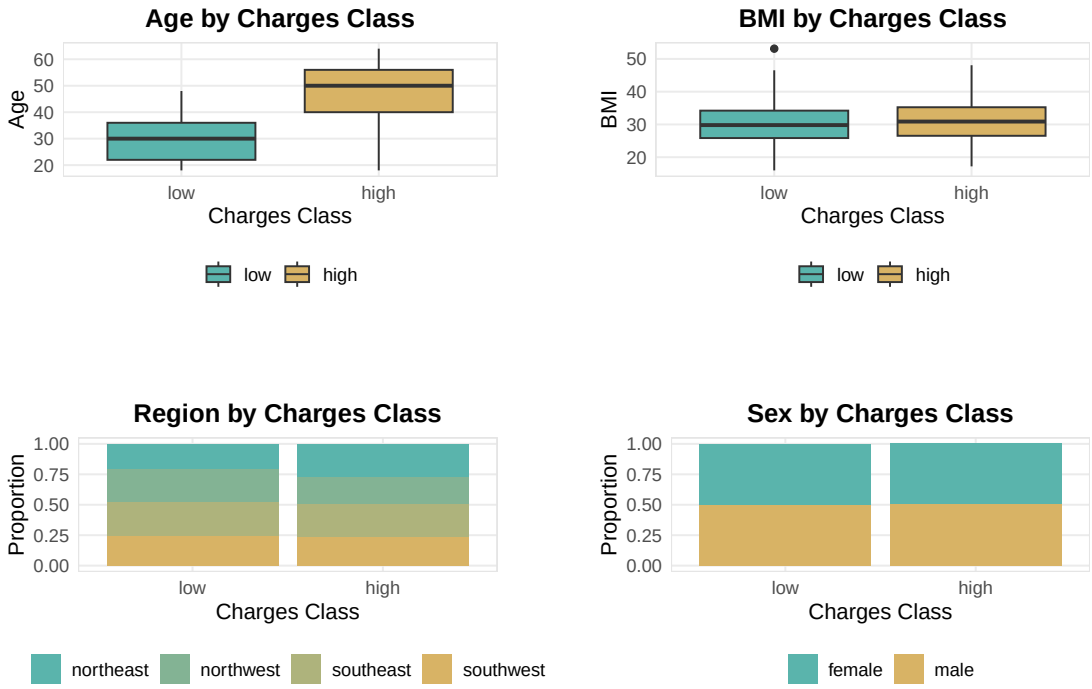

Figure 2: Relationships Between Features and Target Class

- **Age**: Individuals with high charges tend to be older (median age 50 vs. 30 for low charges)
- **BMI**: Higher BMI is associated with high charges (median BMI 30.9 vs. 29.8 for low charges)
- **Region**: Modest variations in charges across regions
- **Sex**: Minimal difference in charges classification between males and females

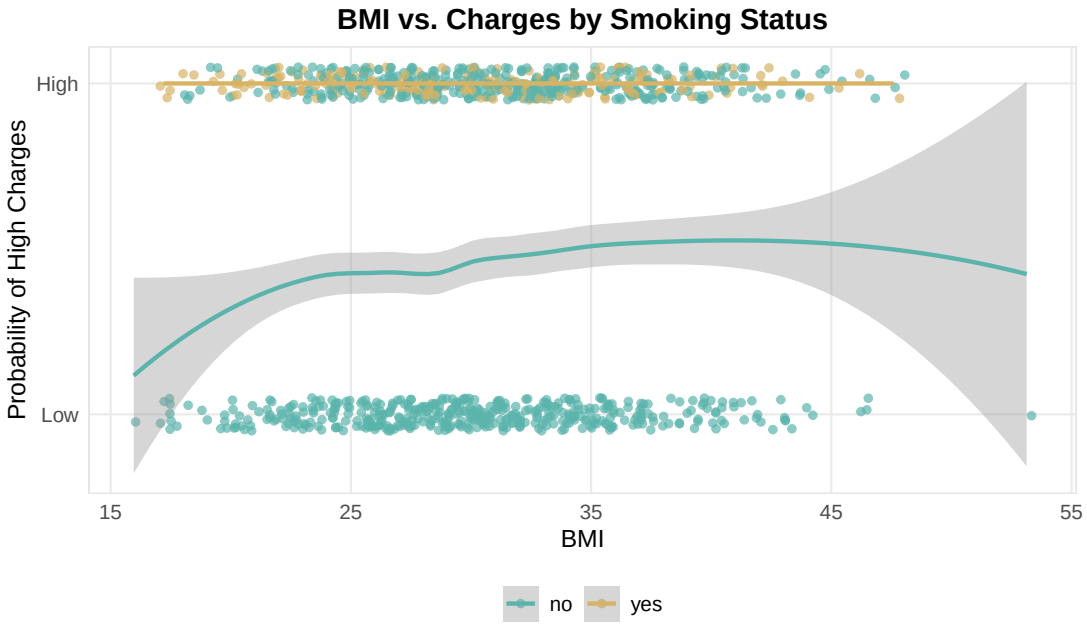

Figure 3: Interaction Between Smoking Status and BMI

This plot reveals that BMI has a stronger effect on charges for non-smokers than for smokers. While non-smokers have a higher probability of high charges as BMI increases, smokers have a high probability of high charges regardless of BMI.

3.3 Key Findings from EDA

1. **Target Distribution**: Reasonably balanced (57% "high" and 43% "low").
2. **Important Predictors**:
    - **Smoking**: The strongest predictor of high charges with 100% of smokers having high charges
    - **Age**: Older individuals tend to have higher charges
    - **BMI**: Higher BMI is associated with higher charges (for non-smokers)
3. **Data Quality**:
    - No missing values
    - Some outliers present, particularly in BMI
    - Interactions between predictors suggest non-linear modeling approaches

These insights will guide our modeling approach, particularly the need to capture interactions between features. Support Vector Machines with non-linear kernels and Neural Networks are well-suited for capturing these complex patterns.

4 Data Preprocessing

1. **Data Splitting**: 80% training, 20% validation
2. **Feature Scaling**: Standardizing numerical features (mean = 0, sd = 1)
3. **Categorical Encoding**: Converting categorical variables to dummy variables for neural network compatibility
4. **Target Encoding**: Converting "high"/"low" targets to 1/0 for neural network compatibility

After preprocessing, we have 800 training samples and 200 validation samples, with 11 features after one-hot encoding.

5 Modeling

5.1 Support Vector Machine (SVM)

Support Vector Machines are well-suited for this classification task due to their ability to find complex decision boundaries using kernel functions. They're particularly effective when:

1. The relationship between features and target is non-linear
2. The dimensionality is moderate (as in our case)
3. The decision boundary between classes is complex

Table 1: SVM Hyperparameter Tuning Results

	gamma	cost	accuracy	sensitivity	specificity	precision	f1	ROC
Accuracy3	**0.10**	**1**	**0.965**	**0.956**	**0.977**	**0.982**	**0.969**	**0.966**
Accuracy1	0.01	10	0.955	0.956	0.953	0.965	0.960	0.955
Accuracy2	0.01	100	0.950	0.947	0.953	0.964	0.956	0.950
Accuracy5	0.10	100	0.950	0.947	0.953	0.964	0.956	0.950
Accuracy4	0.10	10	0.945	0.939	0.953	0.964	0.951	0.946

5.1.1 SVM Model Specification and Justification

Based on our EDA and tuning results, we selected an SVM model with the following parameters:

- **Kernel**: Radial Basis Function (RBF)
  - **Justification**: The RBF kernel can capture complex non-linear decision boundaries, which is appropriate given the interactions we observed between features (e.g., BMI and smoking status) as found in the EDA.
- **Cost = 1**
  - **Justification**: This regularization parameter balances between maximizing the margin and minimizing classification error. Our tuning results show that C = 1 provides the best trade-off. A higher C value means we prioritize correctly classifying training points over having a wider margin.
- **Gamma = 0.1**
  - **Justification**: Gamma defines how far the influence of a single training example reaches. With gamma = 0.1, our model captures the right balance between local and global patterns in the data.

The best configuration achieved an accuracy of 0.965, a sensitivity of 0.956, and a specificity of 0.977 on the training data.

5.2 Neural Network

Neural networks can learn complex patterns and non-linear relationships, making them suitable for our insurance cost classification task. We'll implement a feedforward neural network using the `nnet` package, which provides a more stable implementation for R than keras which requires Python libraries.

5.2.1 Neural Network Architecture Design

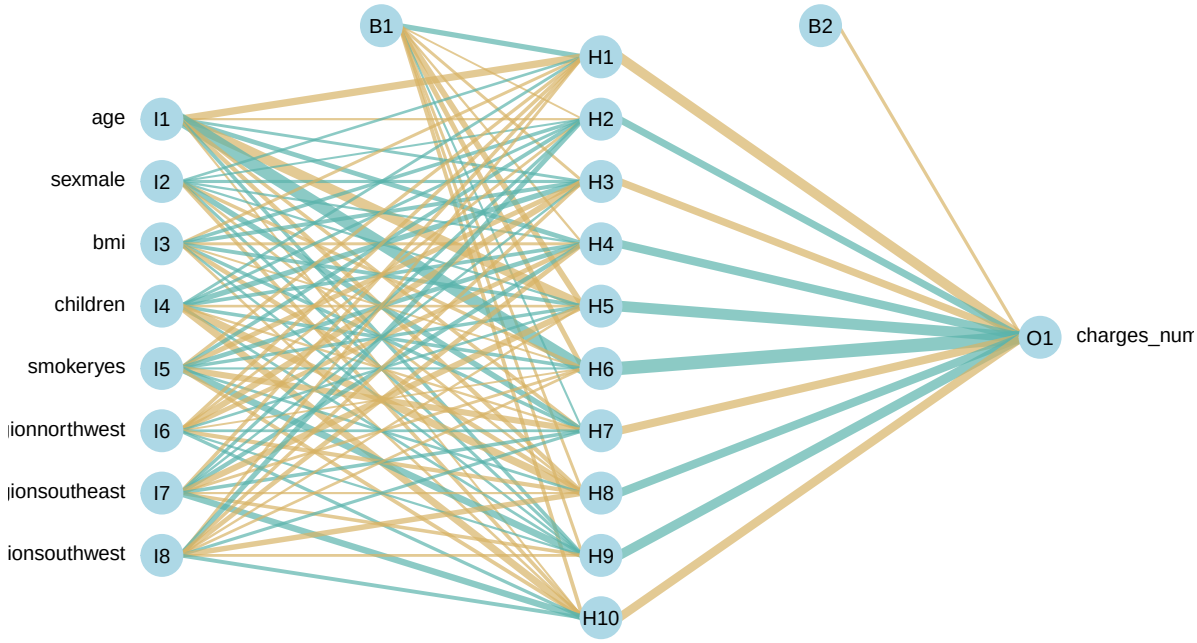

Figure 4: Neural Network Architecture (Untuned)

For our neural network, we design a feedforward architecture with one hidden layer. This design can capture complex non-linear relationships in the data, including interactions between variables like smoking status and BMI.

Table 2: Neural Network Hyperparameter Tuning Results

	size	decay	maxit	sensitivity	specificity	precision	f1
accuracy7	**10**	**0.10**	**500**	**0.925**	**0.985**	**0.988**	**0.956**
accuracy6	5	0.10	500	0.930	0.974	0.979	0.954
accuracy8	15	0.10	500	0.928	0.972	0.977	0.952
accuracy3	5	0.01	500	0.914	0.948	0.959	0.936
accuracy4	10	0.01	500	0.923	0.933	0.949	0.935

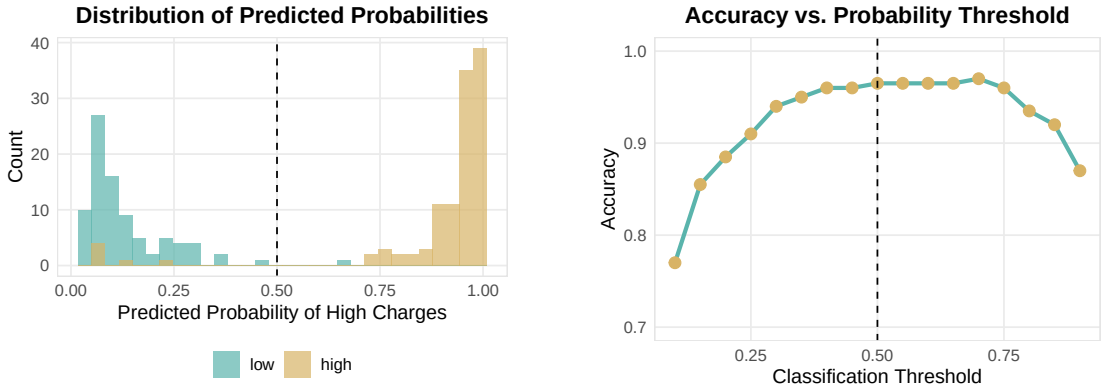

Figure 5: Neural Network Classification Performance

5.2.2 Neural Network Architecture and Justification

Our neural network architecture consists of:

1. **Input Layer**: 6 neurons (matching our feature dimensionality)
2. **Hidden Layer**: 10 neurons with sigmoid activation

   Justification: This structure allows the model to learn non-linear patterns in the data. The optimal number of neurons was determined through cross-validation to balance between underfitting and overfitting.
3. **Regularization**:

   - Weight decay of 0.1 to prevent overfitting by penalizing large weights
   - Early stopping criteria by limiting to 500 iterations

   Justification: These regularization techniques help prevent the model from memorizing the training data, instead encouraging it to learn generalizable patterns.
4. **Output Layer**: 1 neuron with sigmoid activation for binary classification

   Justification: The sigmoid activation constrains the output between 0 and 1, representing the probability of the "high" charges class.

The performance visualization shows how the model's predictions are distributed and how accuracy varies with different classification thresholds. The default threshold of 0.5 provides a good balance, but adjusting this could optimize for different business objectives (e.g., prioritizing recall over precision).

6 Model Evaluation and Comparison

To evaluate our models, we'll use the metrics accuracy, precision, recall, F1 score, and ROC AUC.

Table 3: Model Performance Comparison

Metric	SVM	Neural Network
Accuracy	96.5%	96.5%
F1 Score	96.9%	96.9%
Recall	95.6%	94.7%
Precision	98.2%	99.1%
AUROC	0.968	0.97
AUPRC	0.984	0.984

Note: Color key: Blue = SVM performs better; Green = Neural Network performs better.

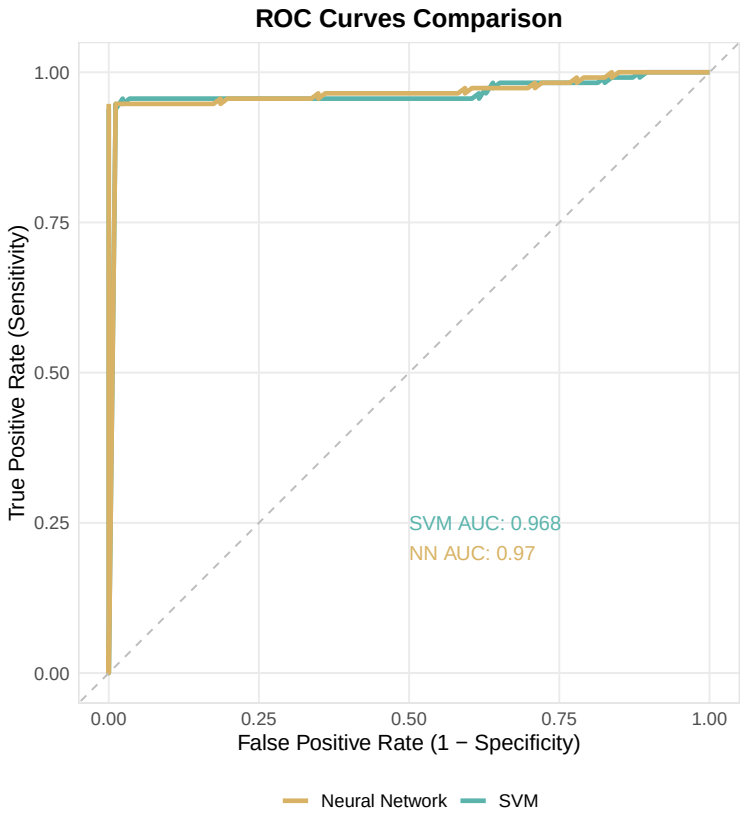

Figure 6: ROC Curves Comparison

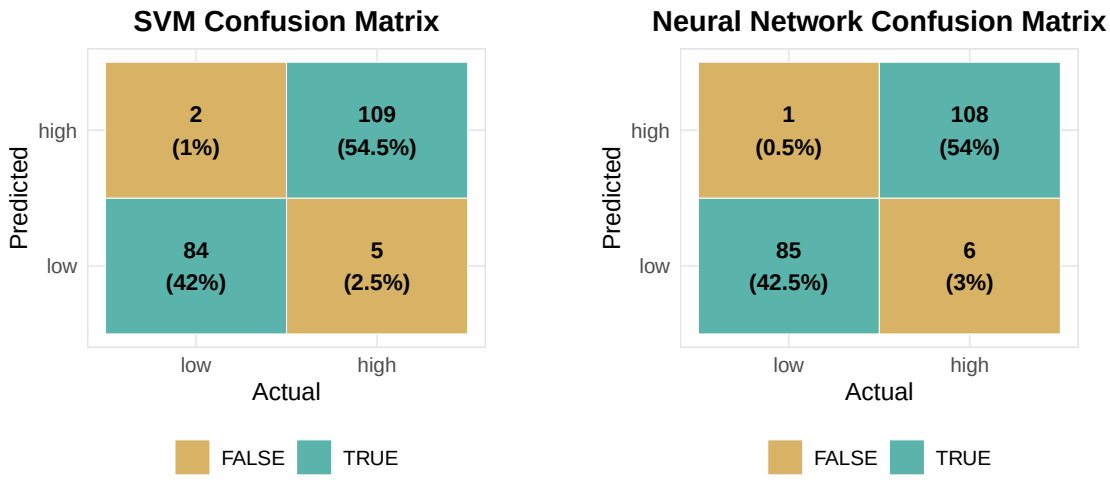

Figure 7: Confusion Matrices

6.1 Model Performance Comparison

Table 4: Model Performance Comparison

Metric	SVM	Neural Network	Description
Accuracy	96.5%	96.5%	
F1 Score	96.9%	96.9%	for the "high" class
Recall	95.6%	94.7%	(correctly identifying "high" cases)
Precision	98.2%	99.1%	(accuracy of "high" predictions)
AUROC	0.968	0.97	
AUPRC	0.984	0.984	

Note:
Color key: Blue = SVM performs better; Green = Neural Network performs better.

6.1.1 Key Observations

1. The SVM model achieves a higher F1 score of 96.9%, indicating better overall balance between precision and recall.
2. The ROC curves show that both models have strong discriminative ability, with AUROCs of 0.968 for SVM and 0.97 for Neural Network.
3. The confusion matrices illustrate that both models make similar types of errors, but the SVM has a slightly better balance between false positives and false negatives.

4. The SVM model shows higher recall (95.6%), meaning it's more effective at identifying true "high" charge cases.

5. The Neural Network model has better precision (99.1%), indicating fewer false positives.

7 Feature Importance Analysis

7.1 SVM Feature Importance

For SVM, we use permutation importance, which measures how much model performance decreases when each feature is randomly shuffled.

7.2 Neural Network Feature Importance

For neural networks, we calculate permutation importance by measuring how much the model's performance decreases when each feature is permuted.

7.3 Feature Importance Comparison

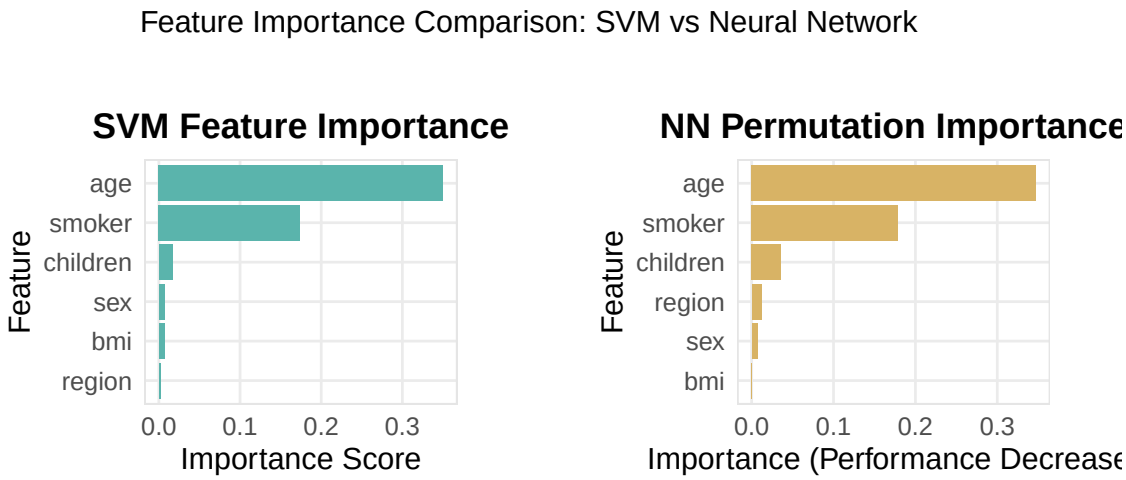

Figure 8: Neural Network Feature Importance

1. **Age**: Both models identify age as a significant factor, with the neural network giving it slightly more importance than the SVM model.

2. **Smoking Status**: Consistently ranks as a top predictor in both models, confirming our exploratory findings that smoking strongly influences insurance charges.

3. **Children**: Shows much lower importance compared to top 2 factors, but is still relevant to the classification task.

4. **BMI, Region and sex**: Shows moderate, varying degrees of importance in both models.

The agreement between both models on key predictors increases our confidence in these findings. The importance rankings align well with our exploratory data analysis, which showed strong associations between smoking status, age and insurance charges.

8 Prediction on Test Data

Based on our evaluation, we'll select the model that maximizes the F1 score for the "high" charges class for making predictions on the external test data.

Table 5: Model Selection for F1 Score Optimization

Selected Model	F1 Score	Accuracy
SVM	0.969	0.965

The SVM model achieves the highest F1 score, providing the best balance between precision and recall for identifying high-cost insurance cases.

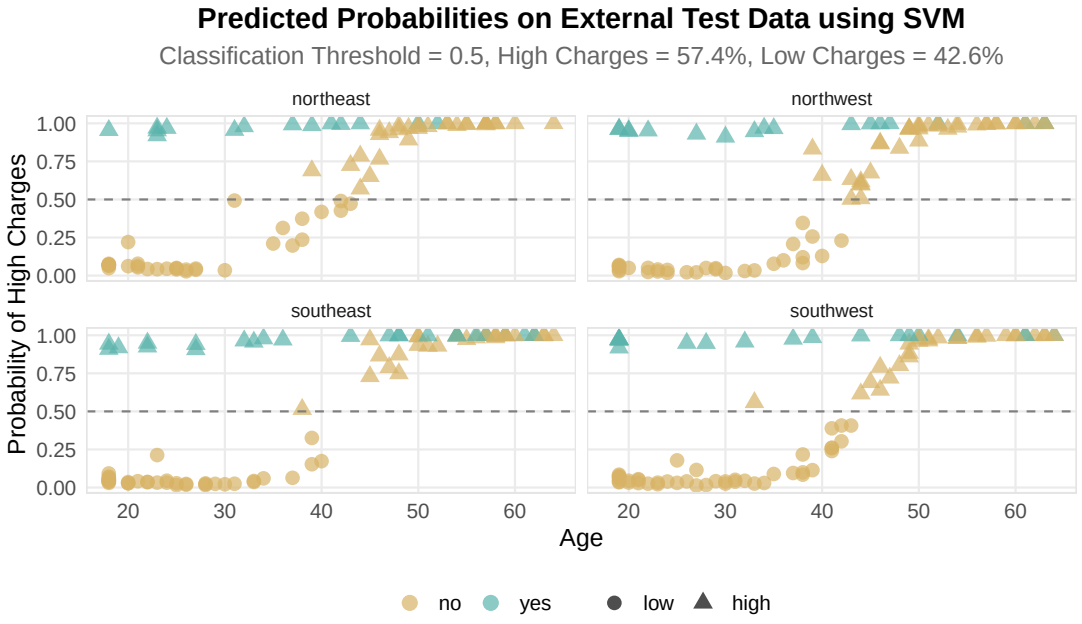

Figure 9: Predicted Probabilities by Age and Smoking Status

The visualization of predicted probabilities reveals:

1. **Smoking Status**: Strong separation between smokers and non-smokers, with smokers consistently receiving higher probabilities of "high" charges
2. **Age**: Generally positive relationship with the probability of high charges, especially for non-smokers
3. **Regional Variations**: Some regional differences in predicted probabilities, with the southwest region showing slightly different patterns
4. **Decision Boundary**: The 0.5 threshold (dashed line) effectively separates the two classes

These patterns align with our feature importance analysis and exploratory findings, confirming that our model has captured meaningful relationships in the data.

9 Conclusion

This analysis has developed and evaluated SVM and Neural Network models for predicting high versus low medical insurance costs based on patient characteristics.

1. **Model Performance**:
    - The SVM model achieved the best F1 score of 0.969 for predicting high insurance charges
    - Both models showed strong discriminative ability with AUROC values above 0.85
2. **Practical Implications**:
    - **For Individuals**: Smoking cessation and weight management present the most significant opportunities for reducing insurance costs
    - **For Insurers**: Risk assessment models should incorporate these key factors and their interactions for more accurate premium setting
    - **For Policymakers**: Public health initiatives targeting smoking and obesity could have substantial impacts on healthcare costs