Project Report on Predicting Hospital Readmission within 30 Days

Prathistha Pandey
Roll No: 2022CSB1105

Abstract

This report presents a machine learning model developed to predict the likelihood of a patient being readmitted to the hospital within 30 days of discharge. By analyzing patient records and applying various machine learning techniques, the goal is to identify high-risk cases and provide insights that can inform hospital management and preventive care. The project includes data preprocessing, model selection, evaluation, and analysis of results.

Contents

1 Introduction

Hospital readmissions within 30 days are a critical issue in healthcare, impacting patient outcomes and increasing healthcare costs. Accurately predicting which patients are likely to be readmitted can enable healthcare providers to allocate resources efficiently and implement preventive measures. This project leverages machine learning to predict readmission probability based on patient demographics, medical history, and other health indicators.

1.1 Objective

The primary objective of this project is to:

- Develop and evaluate machine learning models to predict 30-day readmission.
- Identify significant predictors of readmission to provide actionable insights for healthcare providers.

2 Data Description and Preprocessing

2.1 Data Collection

The dataset comprises hospital patient records, including demographics, medical history, comorbidities, and discharge details. The target variable is a binary indicator of readmission within 30 days.

2.2 Data Preprocessing

The following steps were performed to prepare the data for modeling:

- **Standardization:** Continuous variables, such as age and days in hospital, were standardized using `StandardScaler`.
- **Encoding Categorical Variables:** Categorical features, such as gender and primary diagnosis, were encoded using label encoding to convert them into numerical format.
- **Dimensionality Reduction:** Principal Component Analysis (PCA) was applied for visualization purposes, reducing features to two dimensions to facilitate decision boundary plotting.

3 Methodology

Multiple models were developed to evaluate their effectiveness in predicting readmission. The models included Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and a Neural Network model.

3.1 Model Training and Evaluation

Each model was trained on the full dataset, and a subset of the dataset was used for validation. The neural network was further trained in the PCA-transformed 2D space for visualization.

3.2 Model Descriptions

- **Logistic Regression:** A simple linear classifier optimized using gradient descent to minimize binary cross-entropy.
- **Gradient Boosting:** Gradient Boosting is a powerful machine learning algorithm used for both regression and classification tasks. It is based on the concept of building an ensemble of weak learners, typically decision trees, and optimizing their predictions to minimize a specified loss function.
- **Decision Tree:** A model based on decision rules derived from features, providing interpretable outputs for healthcare applications.
- **Random Forest:** An ensemble of decision trees to improve predictive performance and reduce overfitting.
- **Neural Network:** A deep learning model consisting of multiple layers with ReLU activation functions, trained to classify readmission risk.

4 Results and Analysis

4.1 Evaluation Metrics

The following metrics were used to evaluate model performance:

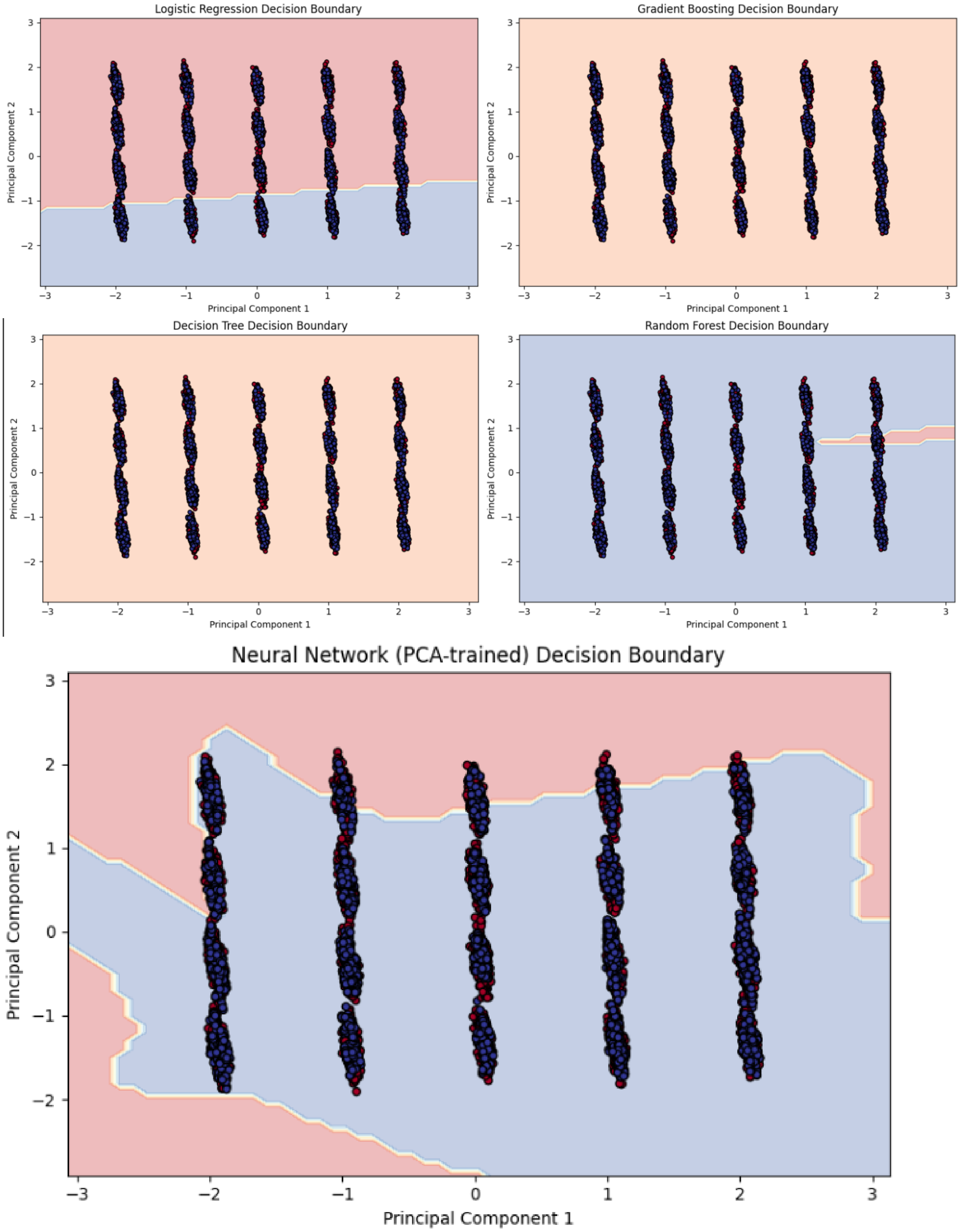
- **Accuracy:** The proportion of correct predictions over the total predictions.

4.2 Model Performance

- **Logistic Regression:** Achieved an accuracy of 66.3%, showing good interpretability but limited flexibility.
- **Gradient Boosting:** The model achieved an accuracy of 75.9% and was effective for separable data but computationally intensive.
- **Decision Tree and Random Forest:** The Decision Tree achieved 76.9% accuracy, while the Random Forest model performed significantly better, achieving an accuracy of 80.6% due to reduced overfitting.
- **Neural Network:** Trained in PCA-transformed space for visualization and achieved an accuracy of 61.4%. It showed the highest flexibility and could capture complex patterns.

4.3 Decision Boundary Visualization

To visualize the decision boundaries of each model, PCA was used to reduce the dataset to two dimensions. The resulting decision boundaries showed how each model separates readmitted patients from non-readmitted ones.



Logistic Regression Decision Boundary

Gradient Boosting Decision Boundary

Decision Tree Decision Boundary

Random Forest Decision Boundary

Neural Network (PCA-trained) Decision Boundary

Figure 1: Decision Boundaries of Various Models in PCA-Transformed Space

5 Discussion

The Random Forest Classifier outperformed other models, with the Logistic Regression model showing the best overall performance but requiring more computational resources.

5.1 Significant Features

The following features were identified as significant predictors of readmission:

- Age
- Number of procedures
- Comorbidity score

These factors indicate that patients with higher comorbidity scores and more procedures are at increased risk of readmission, which aligns with existing medical research.

6 Conclusion

This project demonstrates that machine learning models, particularly ensemble methods and neural networks, are effective for predicting hospital readmission within 30 days. The findings provide valuable insights into which patients are most at risk and can assist healthcare providers in implementing targeted interventions to reduce readmission rates.

6.1 Future Work

Future work could involve exploring additional features or external data sources, such as patient lifestyle or socioeconomic factors, to improve model accuracy. Enhancing the neural network architecture could also further improve predictive performance.