Comprehensive Analysis of Diabetic Readmissions

Milestone: Exploration of Candidate Data Mining Models, and Select the Final Model or Models

Group 18

Mallika Gaikwad
Divya Babulal Shah

gaikwad.mal@northeastern.edu
shah.divyab@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%
Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Mallika Gaikwad
Signature of Student 2: Divya Babulal Shah

Submission Date: 4th March 2024

Exploration of Candidate Data Mining Models, and Select the Final Model or Models

Baseline Model - Logistic Regression:

Logistic Regression is a statistical model used for binary classification. It predicts the probability of an instance belonging to a particular class, often utilized in healthcare scenarios. In the context of our problem—predicting diabetes-related readmission within 30 days—Logistic Regression applies the logistic sigmoid function to the weighted sum of input features, providing interpretable predictions. Its simplicity and interpretability allow us to understand the impact of individual features on the likelihood of readmission. Logistic Regression serves as a reference point against which more complex models are compared.

Advantages:

- Implementing logistic regression is straightforward, interpretable, and highly efficient for training.
- Model coefficients can be interpreted as reliable indicators of feature importance.
- It demonstrates good accuracy on simple datasets and performs effectively when the dataset exhibits linear separability.

Disadvantages:

- Logistic Regression's primary limitation lies in assuming linearity between the dependent variable and the independent variables.
- Capturing complex relationships proves challenging with logistic regression; more robust and compact algorithms like Neural Networks can surpass its performance easily.

Decision Trees:

Decision Trees are non-linear models that recursively split the dataset based on feature values to create a tree-like structure. Each leaf node represents a class label, making it an intuitive model for capturing complex decision rules. For predicting diabetes-related readmission within 30 days, Decision Trees can provide clear decision paths based on relevant patient features.

Advantages:

- Decision trees are relatively robust to outliers and missing data. They can handle missing values and still make predictions based on the available information.
- Decision trees are easy to understand and interpret, even for non-technical people. This makes them a great tool for explaining complex models to stakeholders.
- Decision trees can handle non-linear relationships between features and target variables, making them a great choice for datasets with complex relationships.

Disadvantages:

- Decision trees are prone to overfitting, especially when the tree is deep and complex. This can result in poor generalization performance on unseen data.

- Decision trees can be unstable, meaning that small changes in the data can result in different trees. This makes them less suitable for datasets with high variability.
- Decision trees are not well suited for high-dimensional data, as the number of splits required to split the data becomes very large.

Random Forest:

Random Forest is an ensemble model composed of multiple Decision Trees. It addresses overfitting by combining predictions from individual trees, providing improved generalization. In the context of predicting readmission within 30 days, Random Forest aggregates insights from various decision trees, contributing to robust predictions.

Advantages:

- Random forest can model complex, non-linear relationships between features and the target variable.
- Random forest reduces overfitting problem in decision trees and helps to improve the accuracy.
- Random forest can handle large datasets with high dimensionality, making it a popular choice in many industries.

Disadvantages:

- Random forest can be computationally expensive, particularly when working with large datasets. It requires a lot of memory, which can be a constraint when working with limited resources.
- Random Forests are not easily interpretable. They provide feature importance but it does not provide complete visibility into the coefficients as linear regression.

Support Vector Machines (SVM):

Support Vector Machines are powerful models that aim to find a hyperplane to separate different classes. They can handle non-linear relationships using kernel functions. In predicting diabetes-related readmission, SVMs seek to identify a decision boundary that maximizes the margin between patients who are readmitted and those who are not within the 30-day window.

Advantages:

- Effective in high-dimensional spaces: SVM performs well even in cases where the number of dimensions is greater than the number of samples. This makes it suitable for tasks involving text classification, image recognition, and gene expression analysis.
- Robust against overfitting: SVM maximizes the margin between classes, which helps in generalizing well to unseen data. By controlling the regularization parameter (C), SVM can be tuned to prevent overfitting.

Disadvantages:

- Training an SVM model can be computationally expensive, especially for large datasets. The time complexity of SVM training is approximately O(n^2) to O(n^3), where n is the number of samples.
- SVM performance heavily depends on the choice of hyperparameters such as the regularization parameter (C) and kernel parameters. Finding the optimal parameters often requires extensive experimentation and computational resources.
- SVMs typically produce black-box models, making it challenging to interpret the underlying decision-making process. Understanding the contribution of individual features to the model's output is not straightforward.

Gradient Boosting Models (XGBoost):

XGBoost is a gradient boosting algorithm that builds a predictive model by combining weak learners sequentially. It minimizes the loss function, enhancing overall model performance. For our problem of predicting readmission within 30 days, XGBoost can adapt to the complexity of the data, correcting errors from previous models to improve accuracy.

Gradient Boosting Models, particularly implementations like XGBoost (Extreme Gradient Boosting), have gained popularity for their effectiveness in various machine learning tasks. Here are some advantages and disadvantages of XGBoost:

Advantages:

- XGBoost often produces highly accurate predictions compared to other algorithms. It sequentially builds trees to correct the errors of the previous trees, leading to a strong predictive model.
- XGBoost has built-in capabilities to handle missing values in the dataset. It automatically learns the best direction to go when a value is missing, eliminating the need for imputation techniques.

Disadvantages:

- While XGBoost has regularization techniques to mitigate overfitting, it can still be overfit on noisy data or datasets with high dimensionality. Proper tuning of hyperparameters is essential to avoid overfitting.
- Training an XGBoost model can be computationally expensive, especially for large datasets and complex models. However, advancements like distributed computing and GPU acceleration can alleviate this issue to some extent.
- While XGBoost can handle missing values internally, it may still require preprocessing steps such as encoding categorical variables and scaling numerical features before training.

Neural Networks:

Neural Networks are a class of models inspired by the human brain's structure. They consist of interconnected nodes organized into layers, capable of learning complex patterns in data. In healthcare applications, Neural Networks can capture intricate relationships between various factors, making them suitable for predicting diabetes-related readmission within 30 days.

Advantages:

- Neural networks are capable of learning highly complex and non-linear relationships between input and output variables. With many neurons, layers, and connections.
- They can model intricate patterns and make accurate predictions in various domains, including image recognition, natural language processing, and speech recognition.
- Neural networks can automatically learn relevant features from raw data, eliminating the need for manual feature engineering. Through successive layers of neurons, neural networks can extract hierarchical representations of data, capturing both low-level and high-level features. This feature learning capability makes neural networks well-suited for tasks where handcrafted features may be difficult to define or time-consuming to engineer.

Disadvantages:

- Deep neural networks are inherently complex models with millions of parameters. Understanding the inner workings of a neural network and interpreting its decisions can be challenging, especially for deep architectures with multiple hidden layers. This lack of interpretability limits the trust and transparency of neural network models, particularly in safety-critical applications where interpretability is essential.
- Training deep neural networks often requires large amounts of labeled data and significant computational resources. Training deep architectures with many layers and parameters can be computationally expensive and time-consuming, requiring specialized hardware accelerators and substantial memory resources.
- Neural networks are prone to overfitting, especially when dealing with small datasets or when the model capacity is too high relative to the complexity of the data. Overfitting occurs when the model learns to memorize the training data rather than capturing generalizable patterns.

In summary, each model—Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Gradient Boosting (XGBoost), and Neural Networks—offers a unique approach to predicting diabetes-related readmission within a 30-day window, varying in complexity and interpretability. The choice of the model depends on the trade-off between understanding the decision process and achieving high predictive accuracy in our healthcare application.