Predictive Analytics for Healthcare: Patient Readmission Prediction

Project Plan

Team Members:

Nithin Rajulapati: Focused on data preprocessing, exploratory data analysis, and model evaluation.

Lalitha Velagapudi: Focused on feature engineering, model development, and improving the performance of the model.

Aim:

Design a machine learning model able to predict readmissions but focused on diabetic patients in particular, so that this allows the healthcare provider to identify and be able to intervene with at-risk patients, improving care and curbing the chances of unnecessary readmissions.

Dataset

Utilizes a dataset of patient records with features that include demographics, details of hospital stay, lab procedures, diagnoses, and medications. A customized dataset that is able to predict and understand risks that have to do with readmission.

Technologies and Libraries

- Python: Primary programming language.
- pandas and numpy: For data manipulation and numerical computations.
- matplotlib and seaborn: For data visualization.
- scikit-learn: For traditional machine learning models.
- TensorFlow: Optional, for experimenting with deep learning models if the project scope extends.

Main Modules/Outlines

1. Data Preprocessing & Exploration: Clean up the dataset and conduct an initial analysis to get variable distributions and relationships.
2. Feature Engineering & Model Development: At this stage, done under the supervision of Lalitha, new features were produced to enhance the model's performance and several algorithms of machine learning were used for training models.

Handling Overfitting and Under-fitting

Overfitting: The team will be avoiding overfitting by applying data augmentation and using dropout layers in the neural network architecture; they will also apply early stopping during the training procedure. With traditional machine learning models, regularization methods like L1 and L2 regularization will also be looked for.

Under-fitting: At any rate, under-fitting usually requires a complex enough model to be able to capture the underlying patterns in the data adequately. This may involve, among other things, increasing model depth or experimentation with different architectures and hyper parameters.

Some Challenges and Solutions may occur:

Imbalanced Data: A challenge foreseen in the current project is handling imbalanced datasets, which are very common in medical datasets. In such cases, the team would adopt balancing techniques like SMOTE (Synthetic Minority Over-sampling Technique) or modify the class weights so that the model is not biased towards the majority class.

Feature Selection: Determining readmission prediction will be important for selecting only the relevant set of features. The team would be using models where feature importance is enabled, e.g., Random Forest, and try using approaches like Recursive Feature Elimination (RFE), so that they are able to choose and retain only relevant features.

Model interpretability: In the context of healthcare, interpretability of a model assumes paramount importance. We would like to use models that lend themselves well to a good balance of performance with interpretability, such that healthcare providers can understand and trust what the model predicts.

Evaluation Metrics: The project focuses on details with regard to the accuracy, precision, recall, and F1 score, with the ROC-AUC score capturing the model's performance focusing on the reduction of false negatives, which are particularly costly in a medical care set up.

Conclusion and Future Scope:

The model intervenes significantly with the intention of patient care, in that it helps to curb readmissions through targeted interventions. Future improvements may involve real-time deployment of the model through data feeds, more extensive deployment of the model for more conditions, and be available within the hospital management systems for live predictions.