Project Report

October 6, 2024

Team Member's Details

- **Group Name**: Sachin Subramanian and Nishi Gandhi
- **Name**: Nishi Gandhi
- **Email**: gandhi.nis@northeastern.edu
- **Country**: USA
- **College/Company**: Northeastern University
- **Specialization**: Data Science

Problem Description

This project focuses on predicting **patient persistence or medical outcomes** using a variety of clinical and demographic features. The goal is to apply a **Random Forest Classifier** to categorize patient outcomes based on historical data. Persistence refers to whether a patient adheres to a treatment protocol over a given period.

The primary objective is to develop a machine learning model that can accurately predict whether patients will persist with their prescribed treatments, based on input features like medication frequency, clinical exam codes, and other medical encounters.

Business Understanding

Healthcare providers rely heavily on patients adhering to treatment plans to ensure effective care. Non-persistence leads to adverse health outcomes, increased costs for patients and healthcare providers, and inefficiencies in resource allocation.

By predicting patient persistence, healthcare organizations can:

- **Improve patient engagement**: Early intervention for at-risk patients.
- **Optimize resource allocation**: Focus on patients more likely to drop out.
- **Enhance decision-making**: Personalize care strategies based on predicted patient behavior.

Stakeholders:

- Healthcare Providers (Hospitals, Clinics)
- Pharmaceutical Companies
- Insurance Companies

Project Lifecycle

Stage 1: Data Collection & Preparation (Completed)

- Collected patient records, medication history, and encounter data.
- Performed preprocessing steps: handling missing values, encoding categorical variables, and scaling numerical features.
- **Deadline**: 1 week

Stage 2: Model Training (Completed)

- Model training using a Random Forest Classifier.
- Model was validated using accuracy, precision, recall, and F1-score.
- **Deadline**: 1 week

Stage 3: Model Evaluation (Completed)

- Evaluated using metrics: accuracy, precision, recall, F1-score.
- Feature importance was analyzed to identify critical factors for prediction.
- **Deadline**: 2 days

Stage 4: Business Integration (Pending)

- Model integration into healthcare systems to improve patient interventions.
- **Deadline**: 2 weeks

Data Intake Report

Source of Data

The data for this project is obtained from historical clinical records and demographic datasets. It includes various features related to patient treatments, encounters, and medical history.

Key Features

The dataset contains 3,605 features. Some of the key features include:

- **Dexa_Freq_During_Rx**: Frequency of Dexa scans during treatment.
- **Comorb_Encounter_For_Immunization_Y**: Indicator for immunization-related clinical encounters.
- **Comorb_Encntr_For_General_Exam_W_O_Complaint**: General exam encounters without patient complaints.
- **Ptid**: Unique patient identifiers.
- **Persistency_Flag**: The target variable, representing whether a patient is persistent with treatments.

Data Preprocessing

Several preprocessing steps were performed:

- **Handling Missing Values**: Missing data was imputed or removed.
- **Categorical Encoding**: Categorical features were transformed using one-hot encoding.
- **Feature Scaling**: Numerical features were scaled uniformly.

Target Variable

The target variable is the **Persistency_Flag**, indicating whether a patient remains persistent with treatment.

Data Volume

The dataset consists of approximately 3,605 features. Each row represents a patient's medical data and treatment history.

Github Repo Link

https://github.com/Sachinsub0/DG$_{H}ealthcare_{P}roject/blob/main/DG_{P}roject.ipynb$