Week 8: Problem Understanding

Team Name:				
Name	Email	Country	College/Company	Specialization
Bao Khanh Nguyen	Nguyenkhanhbao8695@gmail.com	USA	American Energy Project	Data Science
Seyedeh Marzieh Hosseini	**shosseini@uni-potsdam.de**	Germany	University of Potsdam	Data Science
Guillermo Leija	leija.guillermo@gmail.com			Data Science
Zain Ul Haq	Zain Ul Haq	Germany	Universitat Rostock	Data Science

Project Life Cycle

Tasks	**08/11/2021 Week 0**	**15/11/2021 Week 1**	**22/11/2021 Week 2**	**29/11/2021 Week 3**	**6/12/2021 Week 4**
Week 7					
Week 8					
Week 9					
Week 10					
Week 11					
Week 12					

Problem Description

ABC is a pharmaceutical business that wants to know the persistency of a drug after a physician has prescribed it for a patient. This company has approached an analytics firm to automate the identifying procedure. This analytics firm has entrusted our team with the task of developing a solution to automate the persistence of a medicine for the client ABC.

One of the fundamental challenges in the pharmaceutical industry is to comprehend the persistency of drugs as per the physician's prescription. To address this problem “ABC Pharma Company” approached the Data Analytics Company to automate this process and identify the persistency of drugs among patients. ABC Pharma Company provided the recorded data with several attributes contained in an Excel file. The dataset contains four independent features such as for each patient (Unique patient ID), patient Demographics (Provider: Doctor/Medical Staff/ Nurse) attributes, Clinical Factors, and Treatment factors. In this dataset, the target variable (independent variable) is classified as a persistency flag for the patient that means understanding whether the patient is persistent with their medication or not.

The dataset contains a few demographic features such as Age, Race, Region, and Ethnicity. Attributes of the physician who prepared the prescription for the patient and this independent variable is supposed to be an important predictor. The disease which is considered in this project is Nontuberculous Mycobacterial (NTM). Numerous scans have been performed for NTM and the evaluation metric is used as T-score for this disease. Clinical factors have also been considered such as results of these tests during the Rx and the performance changes in the last one or two years, along with other factors of Risk segment, multiple risk factors, and tracking of risk segment after getting therapies. Other treatment factors have been considered such as comorbidity disease i.e. other diseases along with NTM and concomitancy of multiple drugs that have been used for curing the NTM disease. All of these features and attributes help in building up the automated Machine learning model to correctly identify the patients based on their “Persistency Flag”. Efforts have been invested to identify the most influential features which help in understanding the person’s choice for continuing the medicine.

Business Understanding:

According to WHO, non-persistent long-term medication for diseases such as hypertension and diabetics is a common problem that leads to health benefits and economic consequences such as waste of money and time, uncured disease [1]. The persistence of drugs is an open topic of research and researches is conducted to analyze what factors are contributing to the non-compliance behavior of patients towards drugs. As per the article on DTC perspectives, it is shown that most of the people in the US are not persistent with the drugs. Pharmaceutical companies, Medical organizations, and hospitals lose $100 billion per year along with 125,000 deaths per year due to the non-persistent behavior of patients [2].

Data Intake Report

Name: Health care- Data Science Specialization
Report date: 5 November 2021
Internship Batch: LISUM04
Version:1.0
Data intake by: Zain Ul Haq
Data intake reviewer: Bao Khanh Nguyen
Data storage location:

Tabular data details: https://github.com/Khanhbao8695/HealthCar_DS2021

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	xlsx
Size of the data	898KB

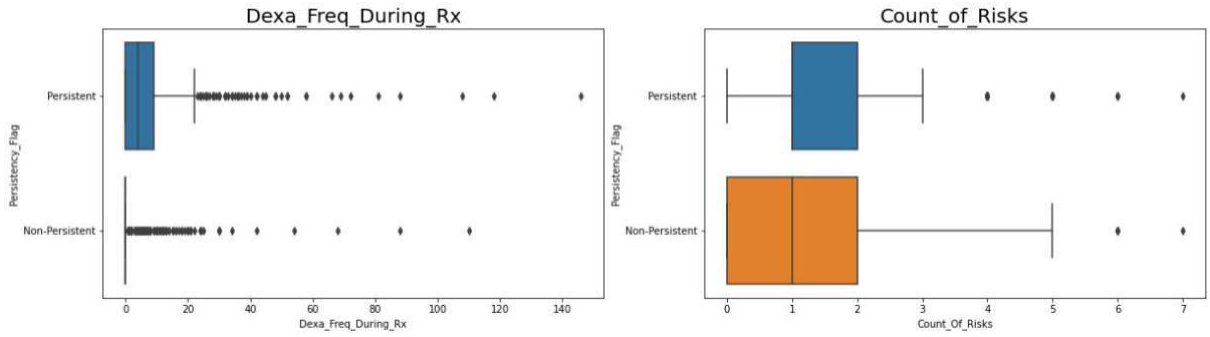
GitHub Repository:

Project Link: https://github.com/ZainUlHaq/Drug-Persistency-ML-Model

Data Types

There are 69 features in this dataset from more than 3424 inputs. The majority of the data types in this dataset are "object" types with more than 67 features and only 2 features are **“**inte64**”** data type.

Data Problems



For "Dexa Freq During Rx," a graph illustrates that this variable has a lot of skewness and Kurtosis (Platykurtic), which considers a lot of outliers. Furthermore, the data for "Count of

Risk" has a moderate skewness and is moderately kurtosis (Platykurtic), indicating that there are few outliers.

Data Transformation to resolve outliers

My first approach to deal with the skewness and outliers for these variables is using IQR Score. To remove outliers, this approach uses the IQR values calculated before. Anything outside of the range of (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR) is considered an outlier and should be eliminated.

For the Old Shape (3424,69) for both of Count of Risk" and "Dexa Freq During Rx" variable but after removed the outliers with this method, the new shape only (2964,69), which remove 460 data outside of the range of (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR).

Data Transformation with non-number data

For Persistent Flag column, there are two types of data: Persistent and Non-Persistent, but if we just leave with object types will be difficult for later training models so I will assign value 0 and

1 to Non-Persistent and Persistent in Persistent Flag column. In addition, most of the value in this dataset have Y and N value, which also difficult to train model, so I will change Y and N value to 1 and 0 as well.