Data Science Healthcare project

Details of Team Members

Group Name: Health+	
Member 1: Sakib Mahmud	**Student Email:** sm1512633@qu.edu.qa **Personal Email:** sakib1263@hotmail.com **Official Email:** sakib.mahmud@qu.edu.qa
	Country: Qatar
	College + Company: Qatar University
	Specialization: Data Science
Member 2: Mohammad Odeh	**Personal Email:** odeh4893@gmail.com
	Country: United Arab Emirates (UAE)
	College/Company:
	Specialization: Data Science

Problem Description

Among the most critical challenges that pharmaceutical companies face in general, the most common one is the task of understanding the “Persistency of a Drug” as per the physician's prescription. To solve this problem “ABC Pharma” company approached an analytics company to automate this process of identification. ABC Pharma provided the company with their recorded data in an Excel file for analysis. The dataset contains four types of predictor parameters (or independent variables) for unique patients (each patient had a unique identifier or ID) such as Patient Demographics, Provider (Doctor/Nurse/Other Medical Staff) attributes, Clinical Factors, and Disease or Treatment Factors. On the other hand, the target or dependent variable for the dataset is the “Persistency Flag” for the patient i.e., whether the patients were persistent on their prescribed medicine(s) or not.

Among the patient demographics parameters, there were patients’ age, race, region, etc. Attributes of the physician who prepared the prescription or performed the task of observing the patient might be an important predictor, so it was included. The primary disease for which patients were treated in this case is Nontuberculous Mycobacterial (NTM). Various tests such as DEXA Scans are performed for NTM which produces metrics like T-Score. Clinical Factors like the outcomes of these tests during the Rx and the performance shift during the last 1 to 2 years were also accounted for, along with the Risk Segment of the patient and the possibility of prevalence of multi-risk among the patients. Other treatment factors such as comorbidity of patients for other diseases alongside NTM, Injectable Experience, and concomitancy of various drugs applied on the patient for NTM were also accounted for. All these parameters will be used to produce models using Machine Learning to correctly classify patients based on their “Persistency Flag”. Efforts will also need to be given to determine the most influential parameters (or class of parameter) for determining peoples’ choice on continuing the medicine.

Business Understanding

More than the healthcare perspective of this problem, it has an even more important business perspective. As discussed earlier, one of the challenges for all pharmaceutical companies is to understand the persistency of drugs as per the physician's prescription. The general trend of persistency of pharmaceutical products among a group of patients is downward, as depicted by the "Persistency Curve" in Figure 1. The pharmaceutical companies aim to determine the factors which affect the decline most so that they can address these issues properly and slow down the process (i.e., the smaller slope of the model). Pharmaceutical companies, healthcare organizations, and hospitals in the USA lose billions of dollars per year due to patients not being persistent in their prescribed medicines and/or treatments [1]. Based on the collected data by any pharmaceutical company or an Integrated Delivery Network (IDN), the most important factors behind the lower level of persistence among patients can be determined. The company or organization can address those issues properly to mitigate the process and over time they can resurvey to check for improvements.

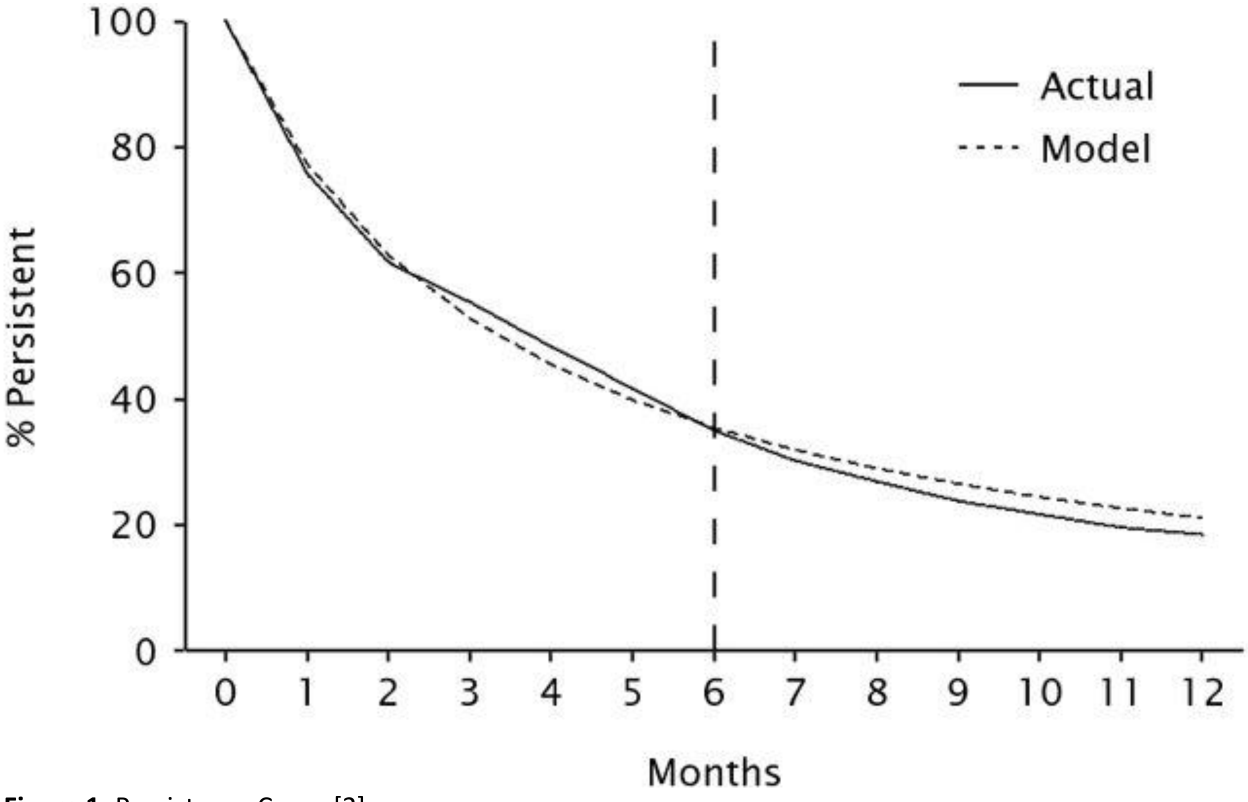

Figure 1: Persistency Curve [2]

Project Lifecycle along with Deadline

The entire project along with all requirements is due for submission by the 15th of May 2021, along with some weekly submissions or updates. The project has been broken into several sub-tasks to smooth and timely progression, which are as follows:

- Problem Understanding.
- Data Understanding.
- Data Cleaning and Feature engineering.
- Model Development.
- Model Selection.
- Model Evaluation.
- Report evaluation metrics such as Confusion Matrix and its derivatives (Accuracy, Precision, Recall, and f1-scores).
- Report ROC-AUC.
- Deploy the model.
- Explain the challenges and future improvements.

Data Intake Report

Name: Persistency of a Drug
Report Date: 25 April 2021
Internship Batch: LISP01
Version: 1.0
Data Intake: Mohammad Odeh and Sakib Mahmud
Data Intake Reviewer:
Data Storage Location: https://github.com/m-odeh/Persistency-of-a-drug

Tabular Data Details: Healthcare_dataset.csv	
Total number of Observations	3424
Total number of File(s)	1
Total number of Features (Independent Variables or Predictors)	68
Base format of the File	.csv
Size of the dataset	898 KB

Proposed Approach:

- Perform Exploratory Data Analysis (EDA) on the data set and visualize various characteristics of the data looking at it from different angles.
- Data pre-processing and cleansing such as getting rid of null values, removing unnecessary columns, check for outliers and unrepresentative data, so on and so forth.
- Transform data such as creating Dummy Variables from categorical data to perform Machine Learning (ML) and other regression-based statistical analysis.
- Build, Test, and Evaluate different ML models based on the features that are correlated.
- Apply dimensional reduction techniques such as Autoencoders to create compact models to represent the whole dataset while preserving the performance.

Assumptions:

- The data follows Normal Distribution.
- Patients' history data were recorded accurately without any errors in testing or examination.

GitHub Repo Link

https://github.com/m-odeh/Persistency-of-a-drug

References

[1] https://decisionresourcesgroup.com/solutions/integrated-delivery-networks-and-their-growing-influence-on-regional-healthcare-in-the-us/

[2] https://www.researchgate.net/publication/5055576_How_to_Project_Patient_Persistency