Data Science Intern at Data Glacier

Project: Healthcare – Persistency of drug

Week 13

Table of contents

Problem Description 2
Business Understanding..2
Data Understanding..2
Data Cleaning and Transformation 6
Exploratory Data Analysis 8
Recommendations for modeling 10
Data Preprocessing..11
Modeling and Evaluation 11
GitHub Repo Link..12

Problem Description

The objective of this project is to **understand and predict the persistency of a drug prescribed by physicians**. Persistency, in this context, refers to whether a patient, based on his/her information, will follow the prescribed medication regimen over a certain period.

Identifying factors that influence persistency is crucial for pharmaceutical companies to improve patient outcomes, reduce healthcare costs, and enhance their product offerings.

Business Understanding

Persistency of medication is a critical factor in the effectiveness of treatment plans. High persistency rates generally correlate with better health outcomes, as patients are more likely to follow their treatment plans. Conversely, non-persistence can lead to worsening health conditions, increased hospitalizations, and higher overall healthcare costs.

Pharmaceutical company ABC Pharma is interested in identifying the key factors that influence whether patients persist with their medication. By leveraging machine learning to predict persistency, the company can develop strategies to improve adherence rates.

Data Understanding

Healthcare dataset has 3424 observations and 69 features. Our intention is to build a model that predicts if a given patient will persist on his/her treatment or not. Having this, out target is the “Persistency Flag” variable, which is a binary data having values True or False depending on the other features.

Besides individual identificators and the target variable, there are other 4 buckets:

- Demographics

- Provider attributes
- Clinical factors
- Disease an treatment factors

Bucket	**Variable**	**Variable Description**	**Information**
Unique Row Id	Ptid	Unique ID of each patient	Type: object Missing values: 0% Unique values: 3424
Target Variable	Persistency_Flag	Flag indicating if a patient was persistent or not	Type: object Missing values: 0% Unique values: 2 Values: ['Persistent', 'Non-Persistent'] Mode: 'Non-Persistent'
Demographics	Age_Bucket	Age of the patient during their therapy	Type: object Missing values: 0% Unique values: 4 Values: ['>75', '55-65', '65-75', '<55'] Mode: '>75' (42.03%)
	Race	Race of the patient from the patient table	Type: object Missing values: 2.83% as 'Other/Unknown' Unique values: 4 Values: ['Caucasian', 'Asian', 'Other/Unknown', 'African American'] Mode: 'Caucasian' (91.94%)
	Region	Region of the patient from the patient table	Type: object Missing values: 1.75% as 'Other/Unknown' Unique values: 5 Values: ['West', 'Midwest', 'South', 'Other/Unknown', 'Northeast'] Mode: 'Midwest' (40.39%)
	Ethnicity	Ethnicity of the patient from the patient table	Type: object Missing values: 2.66% as 'Unknown' Unique values: 3 Values: ['Not Hispanic', 'Hispanic', 'Unknown'] Mode: 'Not Hispanic' (94.48%)
	Gender	Gender of the patient from the patient table	Type: object Missing values: 0% Unique values: 2 Values: ['Male', 'Female'] Mode: 'Female' (94.43%)
	Idn_Indicator	Flag indicating patients mapped to IDN	Type: object Missing values: 0% Unique values: 2

			Values: ['Y', 'N'] Mode: 'Y' (74.68%)
Provider Attributes	Ntm_Speciality	Specialty of the HCP that prescribed the NTM Rx	Type: object Missing values: 9.05% as 'Unknown' Unique values: 36 Values: ['GENERAL PRACTITIONER', 'Unknown', 'ENDOCRINOLOGY', 'RHEUMATOLOGY', 'ONCOLOGY', 'PATHOLOGY', […], 'VASCULAR SURGERY', 'CARDIOLOGY', 'NUCLEAR MEDICINE'] Mode: 'GENERAL PRACTITIONER' (44.83%)
	Ntm_Specialist_Flag	Specialty flag of the HCP that prescribed the NTM Rx	Type: object Missing values: 0% Unique values: 2 Values: ['Others', 'Specialist'] Mode: 'Others' (58.79%)
	Ntm_Speciality_Bucket	Specialty bucket of the HCP that prescribed the NTM Rx	Type: object Missing values: 0% Unique values: 3 Values: ['OB/GYN/Others/PCP/Unknown', 'Endo/Onc/Uro', 'Rheum'] Mode: 'OB/GYN/Others/PCP/Unknown', (61.45%)
Clinical Factors	Tscore_Bucket_Prior_Ntm	T Score of the patient prior of the NTM Rx	Type: object Missing values: 0% Unique values: 2 Values: ['>-2.5', '<=-2.5']
	Tscore_Bucket_During_Rx	T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate)	Type: object Missing values: 43% as 'Unknown' Unique values: 3 Values: ['<=-2.5', 'Unknown', '>-2.5']
	Change_T_Score	Change in Tscore before starting with any therapy and after receiving therapy	Type: object Missing values: 43% as 'Unknown' Unique values: 4 Values: ['No change', 'Unknown', 'Worsened', 'Improved']
	Risk_Segment_Prior_Ntm	Risk Segment of the patient prior of the NTM Rx	Type: object Missing values: 0% Unique values: 2 Values: ['VLR_LR', 'HR_VHR']
	Risk_Segment_ During_Rx	Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate)	Type: object Missing values: 43% as 'Unknown' Unique values: 3 Values: ['VLR_LR, 'Unknown', 'HR_VHR']

	Change_Risk_Segment	Change in Risk Segment before starting with any therapy and after receiving therapy	Type: object Missing values: 65% as ‘Unknown’ Unique values: 4 Values: ['No change', 'Unknown', 'Worsened', 'Improved']
	NTM - Multiple Risk Factors	Flag indicating if patient falls under multiple risk category (having more than 1 risk) at the time of the NTM Rx (within 365 days prior from rxdate)	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N']
	Dexa_Freq_During_Rx	Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate)	Type: integer Missing values: 0% Unique values: 58 Values info: mean – 3.01, std – 8.14, min – 0, 50% – 0, max – 146 Mode: 0 (72.66%)
	Dexa_During_Rx	Flag indicating if the patient had a Dexa Scan during their first continuous therapy	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N'] Mode: 'N' (72.66%)
	Frag_Frac_Prior_Ntm	Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate)	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N'] Mode: 'N' (83.88%)
	Frag_Frac_During_Rx	Flag indicating if the patient had fragility fracture during their first continuous therapy	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N'] Mode: 'N' (87.82%)
	Gluco_Record_Prior_Ntm	Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the one year look-back from the first NTM Rx	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N'] Mode: 'N' (87.82%)
	Gluco_Record_During_Rx	Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N'] Mode: 'N' (76.49%)
Disease/Treatment Factor	Injectable_Experience_During_Rx	Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N'] Mode: 'Y' (73.66%)

	NTM - Risk Factors	Risk Factors that the patient is falling into. For chronic Risk Factors complete lookback to be applied and for non-chronic Risk Factors, one year lookback from the date of first OP Rx	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N']
	NTM - Comorbidity	Comorbidities are divided into two main categories - Acute and chronic, based on the ICD codes. For chronic disease we are taking complete look back from the first Rx date of NTM therapy and for acute diseases, time period before the NTM OP Rx with one year lookback has been applied	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N']
	NTM - Concomitancy	Concomitant drugs recorded prior to starting with a therapy(within 365 days prior from first rxdate)	Type: object Missing values: 0% Unique values: 2 Values: ['Y', 'N']
	Adherent_Flag	Adherence for the therapies	Type: object Missing values: 0% Unique values: 2 Values: ['Adherent', 'Non-Adherent'] Mode: 'Adherent' (94.94%)
	Count_Of_Risks	Total number of risks	Type: integer Missing values: 0% Unique values: 8 Values: [0, 1, 2, 3, 4, 5, 6, 7] Mode: 1 (36.27%)

Data Cleaning and Transformation

Missing values:

- *Race* – using of mode as substitution, only 2.83% are 'Other/Unknown', so it’s quite safe to use the mode to fill in the missing values.

- *Region* – using of Region mode for 'Not Hispanic', because only 1.75% missing values and 100% of them are of Ethnicity 'Not Hispanic'.

- *Ethnicity* – using of mode as substitution, only 2.66% missing values, so it's quite safe to use the mode to fill in the missing values.

- *Ntm_Speciality* – we will try 2 approaches:

1) keeping unknowns as a category since it accounts for less than 9.05% of data and see how it relates to other variables.
2) the categories that accounts for less than 0.01 of the number of observations will be replaced with 'OTHER' category.

- *Risk_Segment_During_Rx, Change_T_Score, Change_Risk_Segment* – these variables have more than 40% of missing values, consequently they will be removed.

- *Tscore_Bucket_During_Rx* - we will try 2 approaches:

1) remove this feature as it has more than 40% of missing values
2) replace unknown values with values using 'Tscore_Bucket_Prior_Ntm' column.

Outliers:

- We will try 2 approaches:

1) detect and remove outliers using IQR method.
2) keep outliers.

Categorical data:

- *Injectable Experience, Risk Factors, Comorbidity and Concomitancy* (group of variables) – 'Y' will be replaced with 1 and 'N' with 0.

- *Tscore_Bucket_Prior_Ntm* – '>-2.5' will be replaced with 1 and '<=-2.5' with 0.

- *Risk_Segment_Prior_Ntm* – 'VLR_LR' will be replaced with 1 and 'HR_VHR' with 0.

- *Ptid* – will be removed, as it has the number of unique values equal to the number of observations and this won't help modeling.

- *Age_Bucket* – will be later encoded using Ordinal Encoder with the following categories: ['<55', '55-65', '65-75', '>75']

- All other categorical features will be encoded using Label Encoder or One-hot Encoder later.

Exploratory Data Analysis

1. **Number of Risks and Treatment Persistence:**

Question: How does the number of risks affect treatment persistence?
Answer: Patients with fewer risks (0-2) are more likely to be Persistent, while those with higher risks (3-5) tend to be Non-Persistent, suggesting that an increasing number of risks may negatively impact treatment persistence.

2. **Number of Risks by Age and Gender:**

Question: How does the number of risks vary across different age groups and genders?
Answer: The number of risks is relatively similar across age groups and genders, with most patients having between 0 and 2 risks, indicating no significant difference in risk distribution by age and gender.

3. **Number of Risks by Region:**

Question: How does the number of risks vary across different regions?
Answer: The number of risks is consistent across all regions, with the majority of patients having between 0 and 2 risks, suggesting that regional differences do not significantly affect the distribution of risks.

4. **Number of Risks by Physician Specialty:**

Question: How does the number of risks vary by the specialty of the physician?
Answer: Patients seen by General Practitioners and Rheumatologists tend to have a wider distribution of risks, including higher numbers, whereas patients seen by specialists in other fields have a more concentrated distribution of lower risks.

5. **Frequency of DEXA During Treatment:**

Question: How does the frequency of DEXA during treatment relate to treatment persistence?
Answer: The majority of patients have a DEXA frequency of zero during treatment, and this trend is consistent in both Persistent and Non-Persistent groups, indicating that most patients do not undergo DEXA scans regardless of their treatment persistence.

6. **Gender Distribution and Treatment Persistence:**

Question: How does treatment persistence differ between males and females?
Answer: From the gender distribution graph, it is observed that females significantly outnumber males in both groups.

7. **Ethnic and Racial Distribution:**

Question: How does treatment persistence vary among different ethnic and racial groups?
Answer: The majority of Non-Persistent patients are Caucasian, while Asian and African American patients constitute a minority.

8. **Regional Differences:**

Question: In which regions are patients more likely to persist with treatment?
Answer: Patients from the Midwest and South regions dominate both groups.

9. **Age Distribution:**

Question: How does treatment persistence vary by age groups?
Answer: Patients older than 75 make up a significant portion of both groups.

10. **Physician Specialty:**

Question: How does treatment persistence differ among patients seen by physicians of different specialties?
Answer: Patients seen by General Practitioners significantly outnumber those seen by other specialists in both groups.

11. **Bone Density (T-score):**

Question: How does the T-score prior to treatment affect treatment persistence?
Answer: Patients with a T-score greater than -2.5 are more common in both groups.

12. Experience with Injectable Therapy:

Question: How does experience with injectable therapy impact treatment persistence?

Answer: The majority of patients in both groups have experience with injectable therapy, indicating its importance in treatment.

13. Risks and Treatment Persistence:

Question: How does the number of risks affect treatment persistence?

Answer: Patients with fewer risks are more likely to persist with treatment.

14. Adherence and Treatment Persistence:

Question: How is treatment persistence related to following prescriptions from physicians?

Answer: Most patients who are labeled as Adherent (following prescriptions) still fall into the Non-Persistent category, indicating other factors affect long-term treatment persistence.

Recommendations for modeling

Data Preprocessing: encode categorical variables and normalize numerical features to ensure consistency and equal contribution to the model.

Model selection: since, from the point of view of machine learning, the task is to perform a binary classifier (persistent or non-persistent), we recommend testing models for the prediction, taking into account interpretable and simple models as well.

Therefore, we will try the following models:

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. Random Forest
5. Gradient boosting (LightGBM or XGBoost)

Model evaluation: use cross-validation, performance metrics (accuracy, precision, recall, F1-score), and confusion matrix analysis to comprehensively evaluate model performance.

Hyperparameter tuning: Employ grid search for tuning.

Interpretation and Validation: Use SHAP values or another method for model interpretation, and validate the model on test data to ensure generalizability.

Data Preprocessing

The data was divided into train (70%), validation (15%) and test (15%) sets.

The following encoders were used for encoding different features: Ordinal Encoding (for ‘Age_Bucket’), Label Encoding (for target ‘Persistency_Flag’), One-Hot Encoding (for all other categorical features).

Scaling was not performed because all our columns have a small range of values.

Modeling and Evaluation

For experiments, models such as Logistic Regression, SVM, Decision Tree, Random Forest, LGBM and KNN were chosen.

Hyperparameters was tuned using GridSearchCV tool.

F1-score (weighted) was chosen as the metric.

The results on the test dataset:

Model	**F1-score (weighted)**	**Accuracy**
Logistic Regression	0.82	0.83
SVM	0.81	0.81
Decision Tree	0.77	0.78

Random Forest	0.82	0.83
LGBM	0.79	0.80
KNN	0.78	0.80

It is turned out that Logistic Regression and Random Forest models are the best performing models within our data.

As a result, **Logistic Regression** was chosen as a final model, because it is faster than Random Forest.

For Logistic Regression, the higher the coefficient value for a certain feature, the more significant it is and the more it affects the target variable. Therefore, in this way we selected the 10 most important features and analyzed them.

GitHub Repo Link

Project Link: https://github.com/kkudzelich/Data-Science-Intern

Submitted by: Kseniya Kudzelich
Date: 29 August 2024