Churn Risk Score

Team Details.

Name	Email	Country	College/Company	Specialization
Fabian Umeh	Fabianumeh335@gmail.com	UK	Teesside University	Data Science
Rukevwe Ovuowo	rukevwe10@gmail.com	Nigeria	GBG Data science Academy	Data Science
Olutayo Oladeinbo	oladeinboolutayo@yahoo.com	UK	Teesside University	Data Science

Problem Statement

One of the challenges for all pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification. With an objective to gather insights on the factors that are impacting the persistency, it is necessary to build a classification for the given dataset, using the variable 'Persistency_Flag' as target variable and other attributes as prediction variables.

Business Understanding

ABC it is a private pharma company. Due to the problem to the persistency of drug as per the physician prescription, a data science project is applied to predict the classification of 'Persistency_Flag' variable. In other words, based on the previously patient's characteristics it is possible to predict if futures patients will use the drugs during the role treatment or if they won't.

The object of this project is providing answer of the main questions made by the company's CEO, which are:

- What is the 'Persistency_Flag' classification for future patients?

The answer for these questions is presented below:

- A dashboard with several hypotheses and insights to help the company CEO with future decisions.

The tools used for this project are: Python 3.8, Jupyter Notebook, Google Colab.

Data Understanding

Variables Description:

Here is a description of the columns in details

Unique Row Id:

- Patient ID: Unique ID of each patient;

Target Variable:

- Persistency_Flag: Flag indicating if a patient was persistent or not;
- Age: Age of the patient during their therapy;
- Race: Race of the patient from the patient table;
- Region: Region of the patient from the patient table;

Demographics:

- Ethnicity: Ethnicity of the patient from the patient table;
- Gender: Gender of the patient from the patient table;
- IDN Indicator: Flag indicating patients mapped to IDN;

Provider Attributes:

- NTM - Physician Specialty: Specialty of the HCP that prescribed the NTM Rx;
- NTM - T-Score: T Score of the patient at the time of the NTM Rx (within 2 years prior from rxdate);
- Change in T Score: Change in Tscore before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown);
- NTM - Risk Segment: Risk Segment of the patient at the time of the NTM Rx (within 2 years days prior from rxdate);
- Change in Risk Segment: Change in Risk Segment before starting with any therapy and after receiving therapy (Worsened, Remained Same, Improved, Unknown);
- 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);

Clinical Factors:

- NTM - Dexa Scan Frequency: Number of DEXA scans taken prior to the first NTM Rx date (within 365 days prior from rxdate);
- NTM - Dexa Scan Recency: Flag indicating the presence of Dexa Scan before the NTM Rx (within 2 years prior from rxdate or between their first Rx and Switched Rx; whichever is smaller and applicable);
- Dexa During Therapy: Flag indicating if the patient had a Dexa Scan during their first continuous therapy;
- NTM - Fragility Fracture Recency: Flag indicating if the patient had a recent fragility fracture (within 365 days prior from rxdate);
- Fragility Fracture During Therapy: Flag indicating if the patient had fragility fracture during their first continuous therapy;

- NTM - Glucocorticoid Recency: Flag indicating usage of Glucocorticoids (>=7.5mg strength) in the one year look-back from the first NTM Rx;
- Glucocorticoid During Therapy: Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy;
- NTM - Injectable Experience: Flag indicating any injectable drug usage in the recent 12 months before the NTM OP Rx;
- 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;

Disease/Treatment Factors:

- 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;
- NTM - Concomitancy: Concomitant drugs recorded prior to starting with a therapy (within 365 days prior from first rxdate) Adherence: Adherence for the therapies.
- Adherence : Adherence for the therapies

What type of data have you got for analysis

The data is a csv file and

What are the problems in the data (number of NA values, outliers , skewe etc)

Gather insights on the factors that are impacting the persistency of a drug during treatment.

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

- Use machine learning to classify future patients, informing if they will use the drugs during the entire treatment or if they won't.
- Create a dashboard with several hypothesis and insights to help the company CEO with future decisions.
- Visualization if they will use the drugs during treatment or not.

  The solution required in generating business insights and create a machine classification model to solve the proposed problem include

1.1 Missing values:

The dataset didn’t contain any null values. Hence we just proceeded to encoding the columns.

1.2 Approach:

In order to encode the data, the label encoder library was imported which we applied on our non-numeric columns.

Import LabelEncoder library

```
#Data Encoding
from sklearn.preprocessing import LabelEncoder
```

Encode our target column y

```
# Select our Dependent variable 'Persistency_Flag'
y = data.Persistency_Flag
```

```
y[:5]
```

```
0        Persistent
1    Non-Persistent
2    Non-Persistent
3    Non-Persistent
4    Non-Persistent
Name: Persistency_Flag, dtype: object
```

```
# Now that we have selected our dependent and independent variable,
# It is important to note that we need to encode our columns as computer only works well with numerical data
# Since we already have a dataset with 60 columns, using pandas One-Hot_Encoding wouldn't be smart.
# Hence, we use the LabelEncoder library.
```

```
#Create an object of the Label Encoder library
Le = LabelEncoder()
```

```
# Encode our target column
y = Le.fit_transform(y)
```

```
y[:5]
```

```
array([1, 0, 0, 0, 0])
```

Encoding other non-numeric column:

```
#Columns to Encode
CTE = ['Gender', 'Race', 'Ethnicity', 'Region', 'Age_Bucket', 'Ntm_Speciality', 'Ntm_Specialist_Flag',
       'Ntm_Speciality_Bucket', 'Gluco_Record_Prior_Ntm', 'Gluco_Record_During_Rx', 'Dexa_During_Rx',
       'Frag_Frac_Prior_Ntm', 'Frag_Frac_During_Rx', 'Risk_Segment_Prior_Ntm', 'Tscore_Bucket_Prior_Ntm',
       'Risk_Segment_During_Rx', 'Tscore_Bucket_During_Rx', 'Change_T_Score', 'Change_Risk_Segment',
       'Adherent_Flag', 'Idn_Indicator', 'Injectable_Experience_During_Rx',
       'Comorb_Encounter_For_Screening_For_Malignant_Neoplasms', 'Comorb_Encounter_For_Immunization',
       'Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx', 'Comorb_Vitamin_D_Deficiency',
       'Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified',
       'Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx', 'Comorb_Long_Term_Current_Drug_Therapy',
       'Comorb_Dorsalgia', 'Comorb_Personal_History_Of_Other_Diseases_And_Conditions',
       'Comorb_Other_Disorders_Of_Bone_Density_And_Structure',
       'Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias',
       'Comorb_Osteoporosis_without_current_pathological_fracture', 'Comorb_Personal_history_of_malignant_neoplasm',
       'Comorb_Gastro_esophageal_reflux_disease', 'Concom_Cholesterol_And_Triglyceride_Regulating_Preparations',
       'Concom_Narcotics', 'Concom_Systemic_Corticosteroids_Plain', 'Concom_Anti_Depressants_And_Mood_Stabilisers',
       'Concom_Fluoroquinolones', 'Concom_Cephalosporins', 'Concom_Macrolides_And_Similar_Types',
       'Concom_Broad_Spectrum_Penicillins', 'Concom_Anaesthetics_General', 'Concom_Viral_Vaccines',
       'Risk_Type_1_Insulin_Dependent_Diabetes', 'Risk_Osteogenesis_Imperfecta', 'Risk_Rheumatoid_Arthritis',
       'Risk_Untreated_Chronic_Hyperthyroidism', 'Risk_Untreated_Chronic_Hypogonadism',
       'Risk_Untreated_Early_Menopause', 'Risk_Patient_Parent_Fractured_Their_Hip', 'Risk_Smoking_Tobacco',
       'Risk_Chronic_Malnutrition_Or_Malabsorption', 'Risk_Chronic_Liver_Disease',
       'Risk_Family_History_Of_Osteoporosis', 'Risk_Low_Calcium_Intake', 'Risk_Vitamin_D_Insufficiency',
       'Risk_Poor_Health_Frailty', 'Risk_Excessive_Thinness', 'Risk_Hysterectomy_Oophorectomy',
       'Risk_Estrogen_Deficiency', 'Risk_Immobilization', 'Risk_Recurring_Falls']
```

```
# Encode the other cartegorical variables
for col in CTE:
    X[col] = Le.fit_transform(X[col])
```

```
X.head()
```

	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	Gluco_Record_Prior_Ntm	Gluco_Record_Duri
0	1	2	1	4	3	5	0	1	0	
1	1	1	1	4	0	5	0	1	0	

It was later noticed that the number of persistent cases are fewer compared to the number of people with Non-persistent cases. To solve this imbalance in the dataset, we imported the SMOTE library. The SMOTE library is used create random samples of the minority class and add to the dataset, so there is an even number between the minority and majority class.

```
#There is an imbalance in the dataset.
#We use SMOTE to settle the imbalance in the dataset
```

```
# Create an object of our SMOTE library
sm = SMOTE(random_state=2)
```

```
# Performing oversampling on our train set
X_train,y_train = sm.fit_resample(X_train,y_train)
```

Result

The target variable is persistency_flag and the evaluation metric used as per highlighted by the task is stated below.

score = 100 * f1_score(test, predictions, average="macro")

AUC

ROC

The results of the machine learning models are summarized below

Model	Model parameters	SCORE	PRECISION	RECALL	F1-Score	AUC
Logistic Regression	Max_iter=1500, solver='newton-cg'	77.82%	0(Class) – 83% 1(Class) – 70%	0(Class) 81% 1(Class) – 72%	0(Class) – 82% 1(Class) - 71%	76.7%
SVM	C=10, probability=True	78.89%	0(Class) – 82% 1(Class) – 73%	0(Class) – 85% 1(Class) – 69%	0(Class) – 83% 1(Class) – 71%	77%
KNN	n_jobs=5, n_neighbours=100	77.72%	0(Class) – 77% 1(Class) – 78%	0(Class) – 91% 1(Class) – 56%	0(Class) – 84% 1(Class) – 66%	73.5%
Gradient Boost	learning_rate=0.01, loss='exponential', max_depth=70, max_features=1 ,n_estimators=200	80%	0(Class) – 83% 1(Class) – 74%	0(Class) – 84% 1(Class) – 72%	0(Class) – 84% 1(Class) – 73%	78.2%

Summary and Recommendation

The dataset contains 3424 rows and 69 columns.

The number of cases where the drugs proved to be non-persistent were higher compared to number od persistency cases.

The dataset reveal that more females partook in this analysis than male.

People of Caucasian race when compered to other races were the most common in the study.

The non-Hispanic ethnic group were the most common in the study.

There were more people from the Midwest and South region compared to other regions.

For this study, most people selected are greater than 75 years of age.

People with a Tscore of >-2.5 have a higher chance of drug being non-persistent.