FINAL PROJECT PRESENTATION

Healthcare – Persistency of a drug

Performer's Details

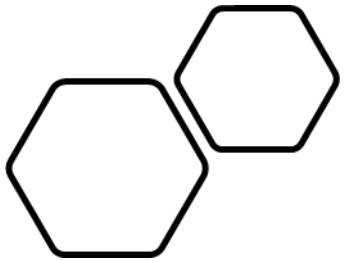
Name: Kseniya Kudzelich
Email: ksusha.konstovich@gmail.com
Country: Belarus
Specialization: Data Science
Internship Batch: LISUM34

August 29, 2024

Agenda

- Business Understanding
- Data Understanding
- Exploratory Data Analysis
- Proposed Modeling Technique
- Modeling and Evaluation

Business Understanding



Problem statement

- One of the challenge for all Pharmaceutical companies is to understand the persistency of drugs as per the physician prescription. To solve this problem ABC pharma company would like the process Automated.

Objectives

- The overall aim of the analysis part of the project is to provide insights into factors that impact the persistency of drugs, which afterwards will lay the foundation on building a suitable classification model and also propose some modelling technique to be used.

Business Understanding

- 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

```
df.head()
```
✓ 0.0s

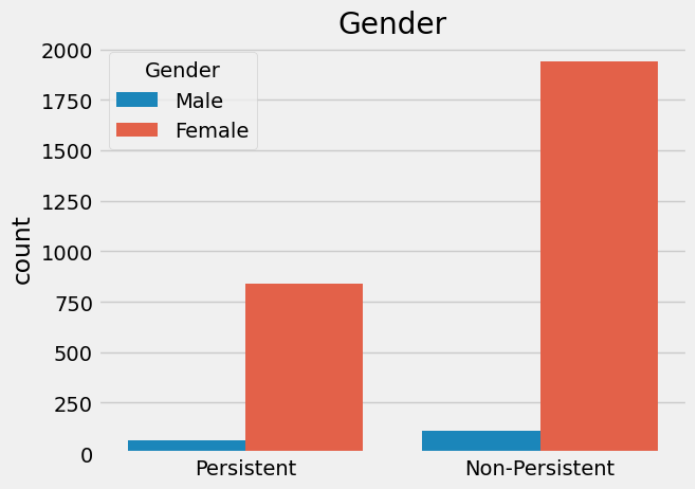
	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...
0	P1	Persistent	Male	Caucasian	Not Hispanic	West	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
1	P2	Non-Persistent	Male	Asian	Not Hispanic	West	55-65	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
2	P3	Non-Persistent	Female	Other/Unknown	Hispanic	Midwest	65-75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
3	P4	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...
4	P5	Non-Persistent	Female	Caucasian	Not Hispanic	Midwest	>75	GENERAL PRACTITIONER	Others	OB/GYN/Others/PCP/Unknown	...

5 rows × 69 columns

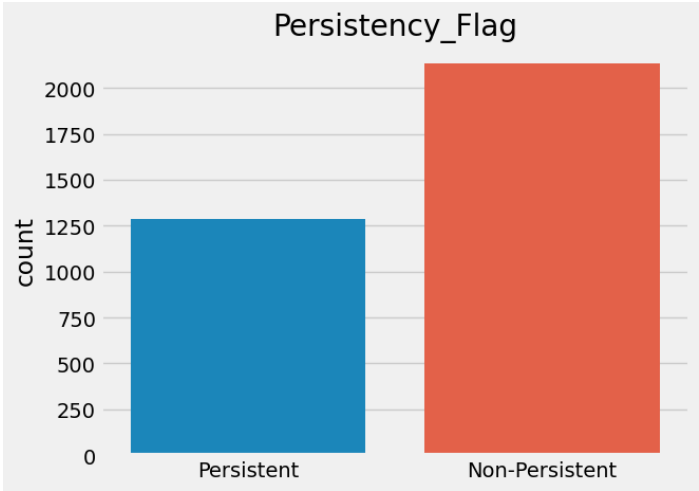
- The dataset contains 3424 rows and 69 columns.

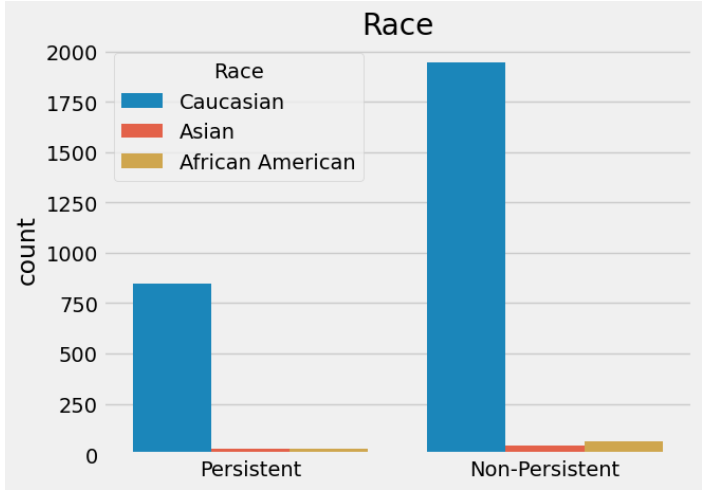
Bucket	Variable	Variable Description
Unique Row Id	Patient ID	Unique ID of each patient
Target Variable	Persistency_Flag	Flag indicating if a patient was persistent or not
Demographics	Age	Age of the patient during their therapy
	Race	Race of the patient from the patient table
	Region	Region of the patient from the patient table
	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
Clinical Factors	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)
	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 Usage During Therapy	Flag indicating if the patient had a Glucocorticoid usage during the first continuous therapy
Disease/Treatment Factor	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
	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

Exploratory Data Analysis



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





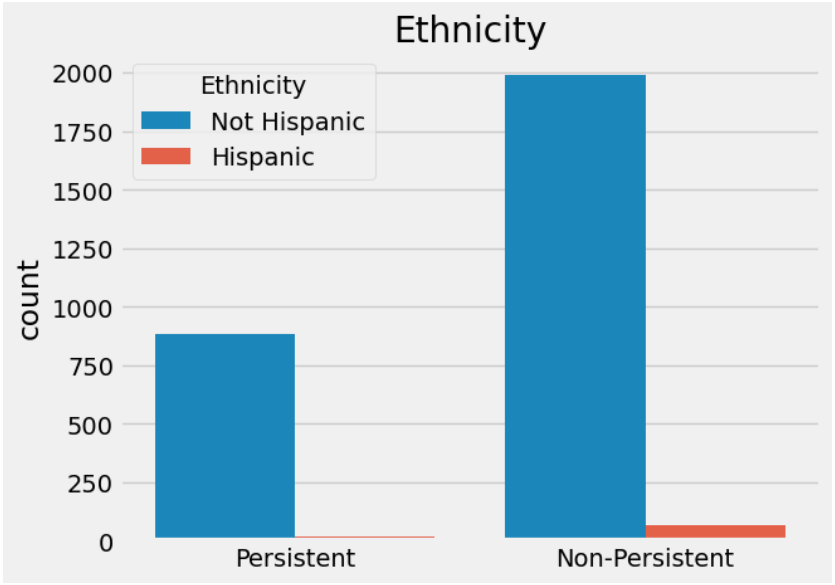
- The dataset reveal that females significantly outnumber males in both groups.

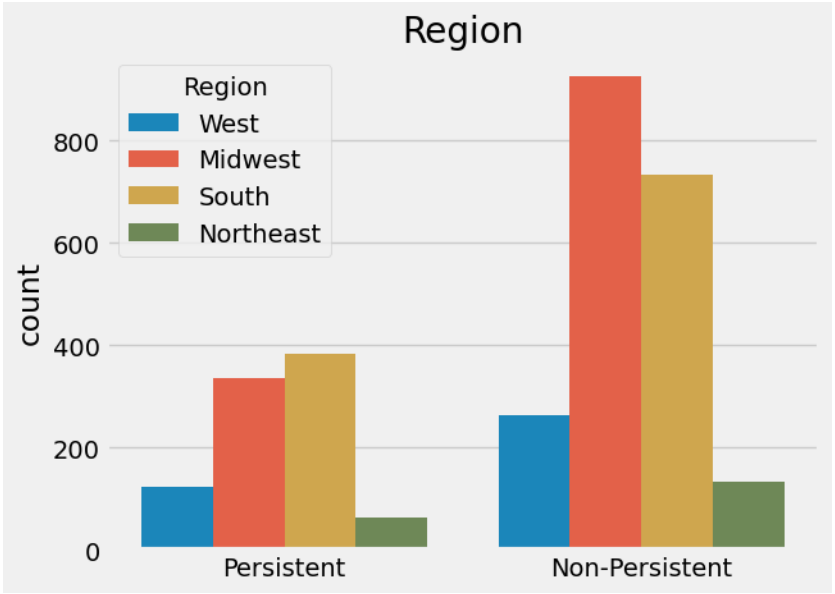
- Significantly more patients are into the Non-Persistent group compared to the Persistent group.

Exploratory Data Analysis

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

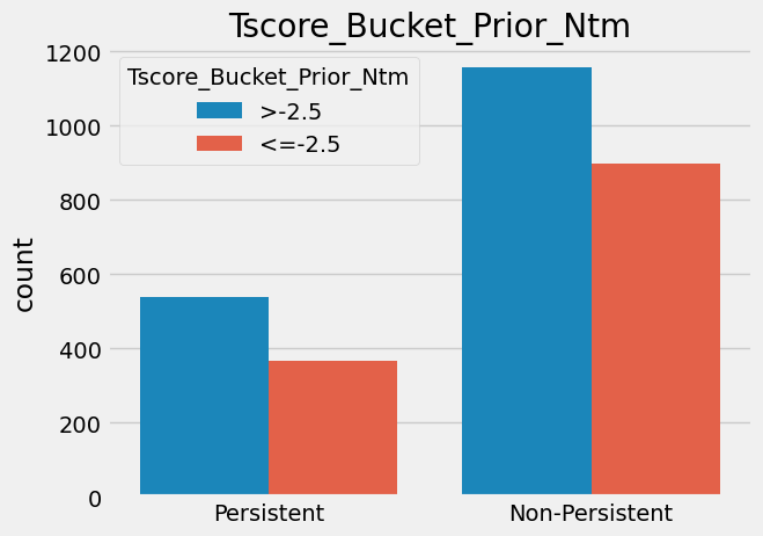
- **Patients from the Midwest and South regions dominate both groups.**

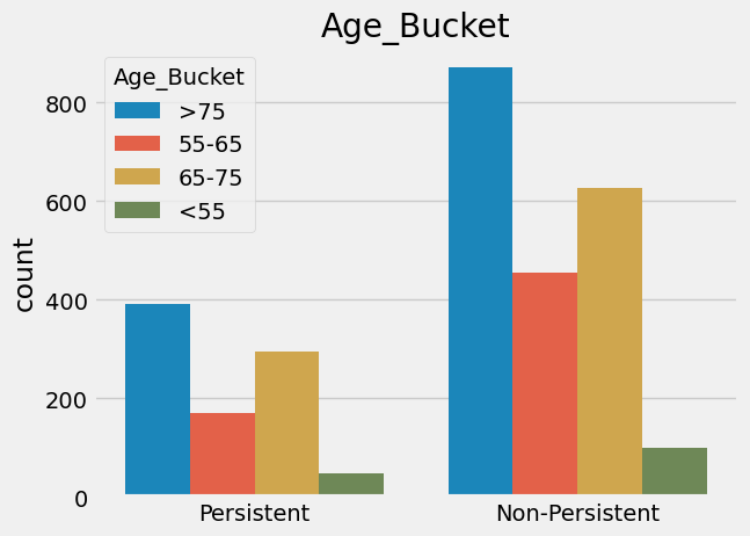


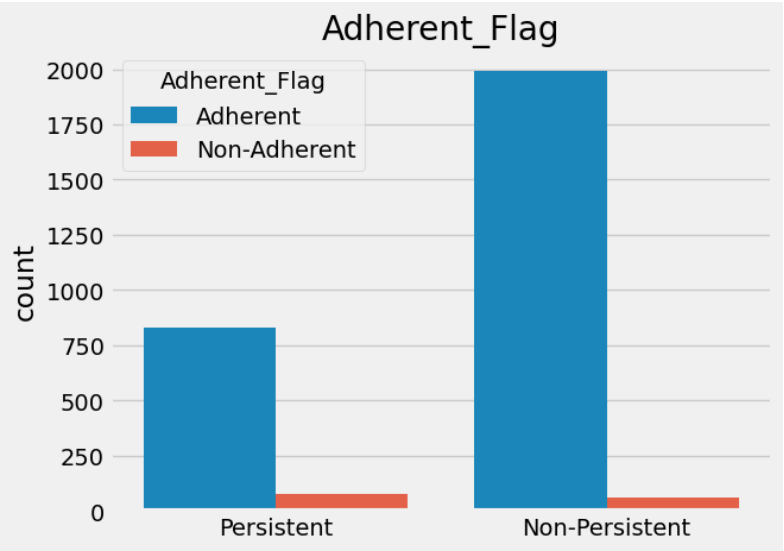


Data Glacier
Your Deep Learning Partner

Exploratory Data Analysis



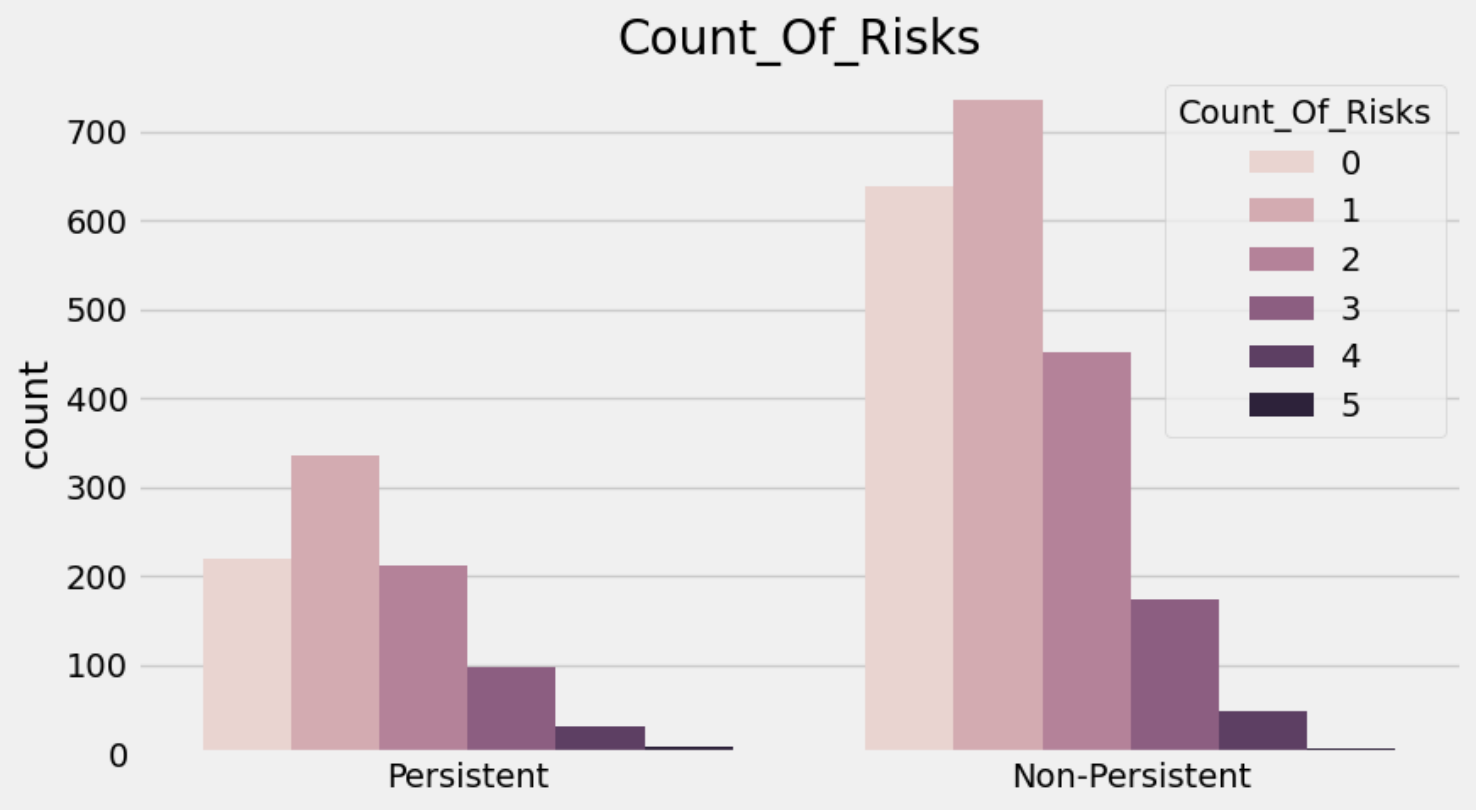




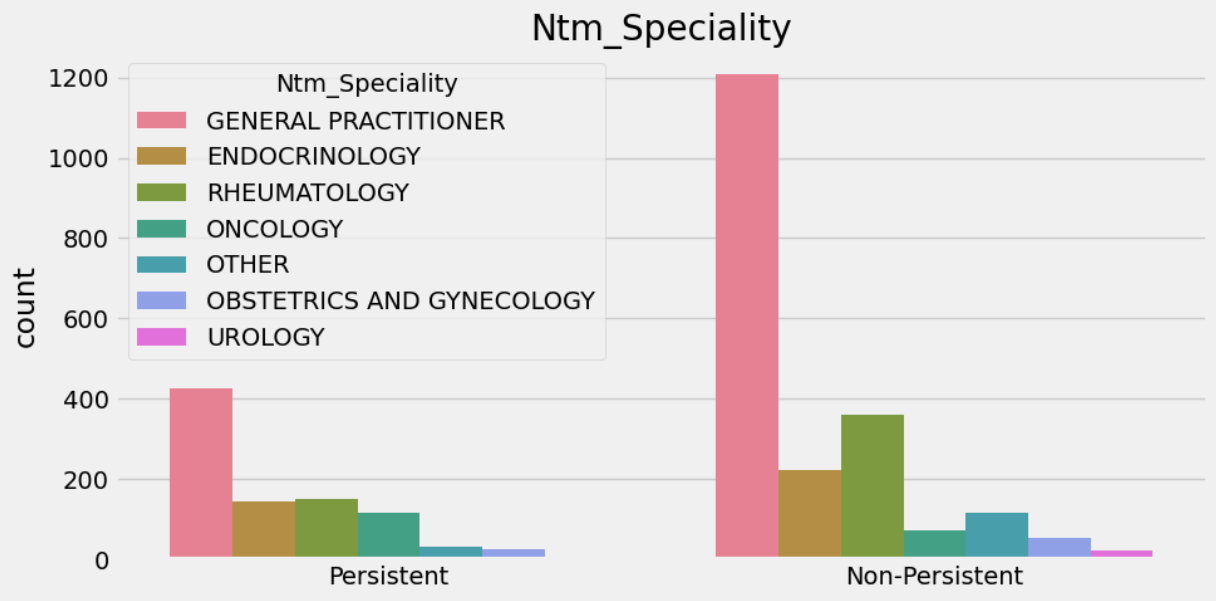
- For this study, patients older than 75 make up a significant portion of both groups.
- Patients with a T-score greater than -2.5 are more common in both groups.
- Both groups Persistent and Non-Persistent have significantly more patients who follow the prescriptions (Adherent). Most patients who are labeled as Adherent (following prescriptions) still fall into the Non-Persistent category, indicating other factors affect long-term treatment persistence.

Exploratory Data Analysis

- Patients with 0, 1 and 2 risks predominate in both Persistent and Non-Persistent groups with the highest concentration at 1 risk.

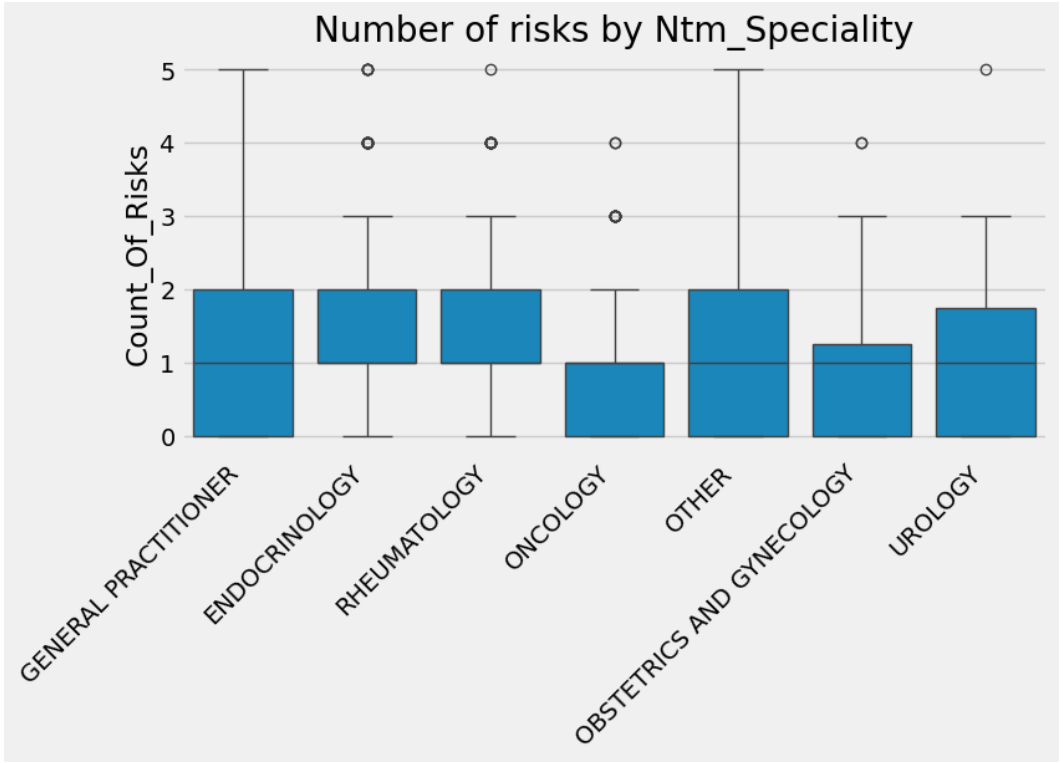


Exploratory Data Analysis



- Patients seen by General Practitioners and Rheumatologists tend to have a wider distribution of risks, including higher numbers, compared to seen by specialists in other fields.

- General Practitioner speciality prevail in both groups.



Data Glacier

EDA Summary

Data Glacier
Your Deep Learning Partner

EDA SUMMARY

- The dataset contains 3424 rows and 69 columns.
- Significantly more patients are into the Non-Persistent group compared to the Persistent group.
- The dataset reveal that females significantly outnumber males in both groups.
- The majority of patients in both groups are Caucasian.
- The non-Hispanic ethnic group is the most common in the study.
- Patients from the Midwest and South regions dominate both groups.
- Patients older than 75 make up a significant portion of both groups.
- Both groups Persistent and Non-Persistent have significantly more patients who follow the prescriptions (Adherent).
- Most patients who are labeled as Adherent (following prescriptions) still fall into the Non-Persistent category, indicating other factors affect long-term treatment persistence.

Proposed Modeling Technique

Data Glacier
Your Deep Learning Partner

PROPOSED MODELING TECHNIQUE

- **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)
6. KNN

Proposed Modeling Technique

Data Glacier
Your Deep Learning Partner

PROPOSED MODELING TECHNIQUE

- **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.

Modeling and Evaluation

Data Glacier
Your Deep Learning Partner

Modeling and Evaluation

- The results on the test dataset:

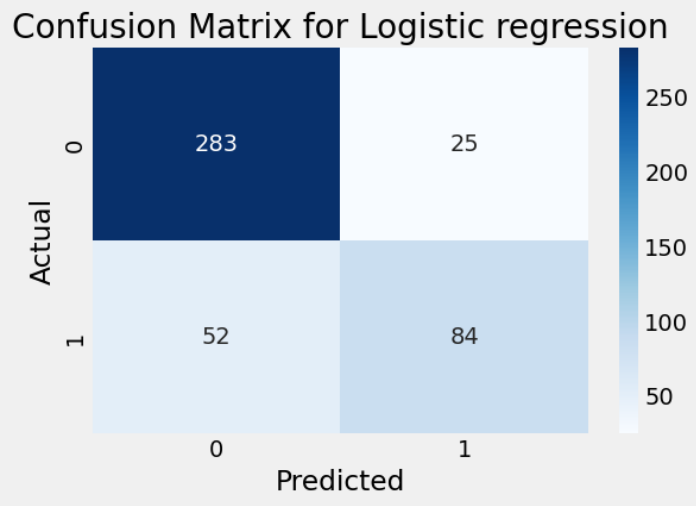
Model	F1-score (weighted)	Accuracy
Logistic Regression	0.82	0.83
SVM	0.81	0.81
Decision Tree	0.77	0.78
Random Forest	0.82	0.83
LGBM	0.79	0.80
KNN	0.78	0.80

- 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.

Modeling and Evaluation

Logistic Regression

- Regularization has been applied
- **F1-score**: 82%, **Accuracy**: 83%



Classification Report for Logistic regression:

	precision	recall	f1-score	support
0	0.84	0.92	0.88	308
1	0.77	0.62	0.69	136
accuracy			0.83	444
macro avg	0.81	0.77	0.78	444
weighted avg	0.82	0.83	0.82	444

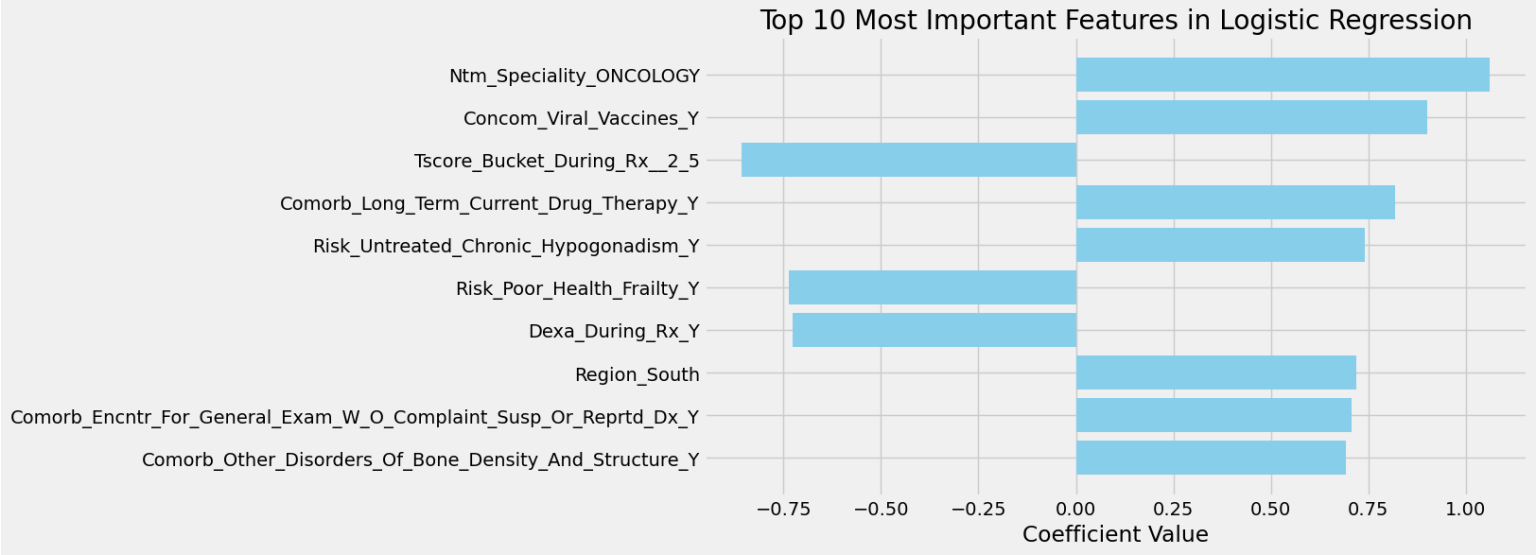
Data Glacier
Your Deep Learning Partner

Modeling and Evaluation

Data Glacier
Your Deep Learning Partner

Feature Importance

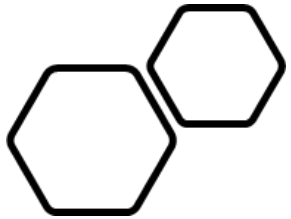
- The most important features of the model include the patient's belonging to the specialty "Oncology“, receiving viral vaccines, undergoing long-term current drug therapy, and having untreated chronic hypogonadism, which significantly increase the probability of a positive outcome.
- Negative coefficients indicate the negative impact of a health condition, such as T-score results, poor health frailty, and Dexa scan during treatment on the outcome.
- The region of residence and specific medical examinations also influence the predictions of the model, though to a lesser extent.
- These results highlight the importance of taking into account medical and regional factors when assessing the likelihood of certain outcomes and developing a treatment strategy.



Repository details

- Repo link: GitHub repo

Data Glacier
Your Deep Learning Partner



Thank You!