Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

PROJECT : Healthcare - Persistency of a drug (Data Science)

BATCH CODE : LISUM22

NAME	SCHOOL	EMAIL	SPECIALIZATION AND COUNTRY
Taiwo Akingbesote	Montclair state University	akingbesotet12@montclair.edu	Data Science & USA
Kerr Tan	New York University	st4153@nyu.edu	Data Science & USA
Farzana Chowdhury	Mount Holyoke College	chowd23f@mtholyoke.edu	Data Science & USA
Aya Ibrahim	University of North Carolina	ayariyadh9@gmail.com	Data Science & USA

OUTLINE

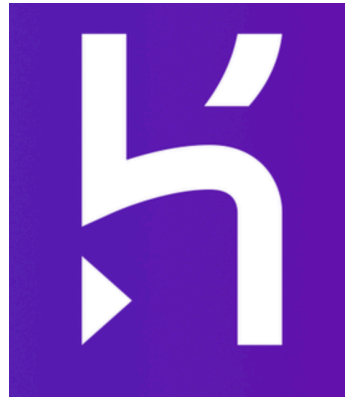
OBJECTIVE AND ANALYSIS APPROACH

DATASET PREPARATION

EXPLANATORY DATA ANALYSIS

RECOMMENDATION

MODEL BUILDING



MODEL TESTING, FLASK AND HEROKU DEPLOYMENT

OBJECTIVE AND ANALYSIS APPROACH

OBJECTIVE

- The pharmaceutical industry faces numerous challenges in understanding the persistency of drugs as per physician prescriptions. Ensuring that patients adhere to prescribed medications is crucial for their health outcomes and overall treatment effectiveness. However, non-adherence to prescribed medications can lead to suboptimal results, increased healthcare costs, and potential complications. To address this critical issue, ABC pharma company has decided to take a data-driven approach and has approached us to automate the process of identifying factors impacting drug persistency.

- Our task is to analyze a comprehensive dataset containing a diverse set of variables related to patient demographics, provider attributes, clinical factors, disease/treatment factors, comorbidities, concomitancy, and adherence information. The primary objective is to gather insights and build a robust classification model that predicts whether a patient will exhibit persistency with the prescribed drug or not. The target variable, "Persistency_Flag," will serve as the ground truth for this classification task, where it is coded as 1 if the patient is persistent and 0 if the patient is non-persistent between variables. Then build a model that classifies the dataset.

ANALYSIS APPROACH

- Analysis based on Numerical Values
- Analysis based on Patient Demographic
- Analysis based on Patients Physician/Provider
- Analysis based on Risk Factors and Change, Adherence to Therapy, & T-score Change

DATASET PREPARATION

DATASET PREPARATION

- **Data Manipulation** - After collecting the data, we noticed that there is 'N' and 'Y' for most of the one-hot encoding categorical value, it needs to change to 0 and 1 in the following step to get an accurate overview of the data.

- **Data Verification -** After manipulating the data to getting an accurate view of our data, we verified the data types and check for any duplicate value(s)

EXPLANATORY DATA ANALYSIS

Data Glacier
Your Deep Learning Partner

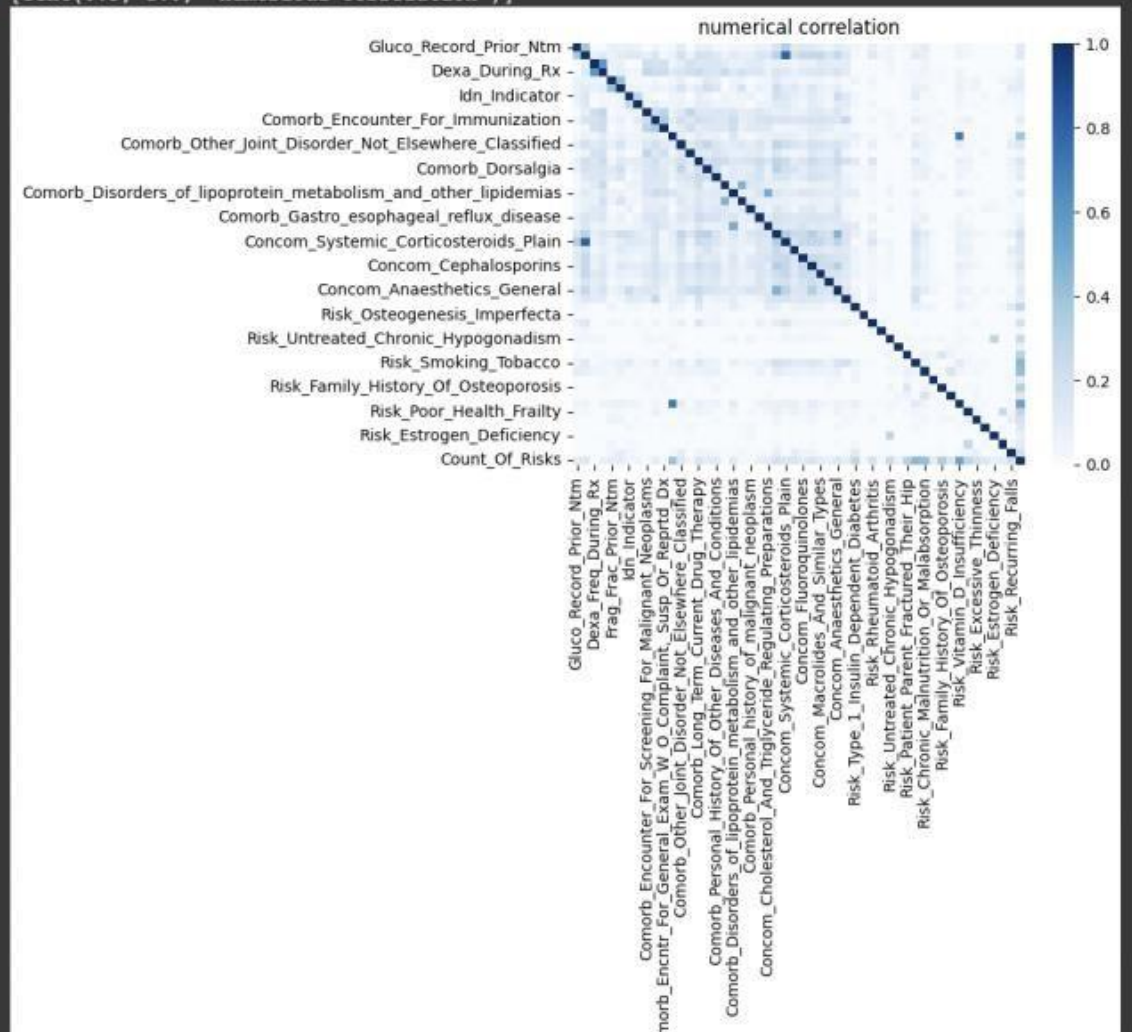
Analysis based on Numerical Values

Numerical Values Analysis

```
[ ] numerical = [col for col in data.columns if data[col].dtype == 'int64']
    data_numerical = data[numerical]
```

```
sns.heatmap(data_numerical.corr(),cmap='Blues',vmin=0,vmax=1).set(title='numerical correlat
```

```
[Text(0.5, 1.0, 'numerical correlation')]
```

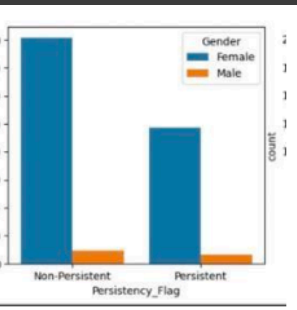


FINDINGS

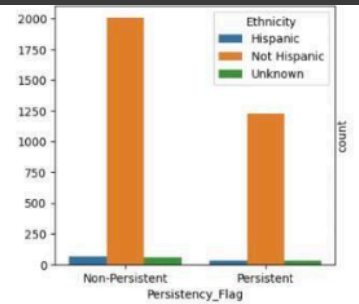
No findings were made with this analysis

Data Glacier
Your Deep Learning Partner

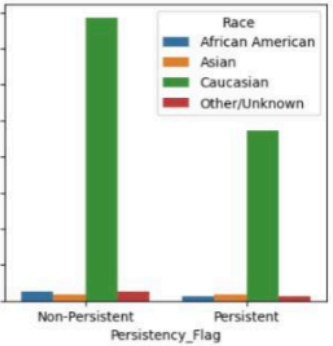
Analysis based on Patient Demographic



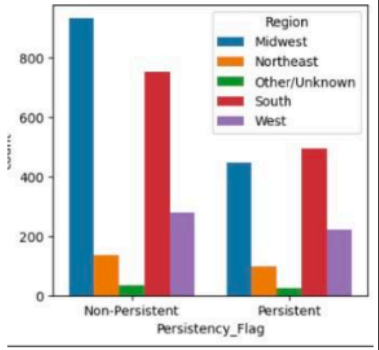
Gender



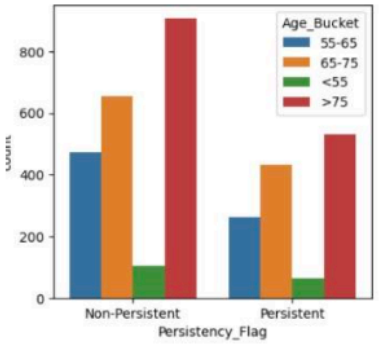
Ethnicity



Race



Region



Age

Findings

1) There are higher persistency and non-persisten counts in Females than in Males with non-persistency being higher.
2) There are higher persistency and non-persisten counts in Caucasians among all other races with n persistency being higher.
3) There are higher persistency and non-persisten counts in non-Hispanic people among all other ethnicities with non-persistency being higher.
4) The highest persistency counts in order among regions is in the South, Midwest, and West regions And the highest non-persistency counts in order is in the Midwest, South, and West regions.
5) The highest persistency and non-persistency count in order are among patients of the following age groups: >75, 65-75, and 55-65.

Analysis based on Patients Physician/Provider

Findings

1) The highest non-persistency and persistency counts among patients occurred with thos whose providers are general practitioners and rheumatologists with endocrinologists and unknown specialties coming next.

2) The highest non-persistency and persistency counts also occurred among patients whose providers' flag
was categorized as non-specialists.

3) The highest non-persistency and persistency counts also occurred among patients whose providers' bucket
was categorized as OB/GYN/Others/PCP/Unknown.

- Based on this information, it's hard to detect what specialty led to the most non-persistency.
  However, Generally, those who were general practitioners or non-specialists had higher non-persistency patient counts.

Data Glacier
Your Deep Learning Partner

Analysis based on Risk Factors and Change, Adherence to Therapy, & T-score

Findings

- With this analysis, we looked for other factors that may influence target variable and plotted their graphs to visualize them. They include.
  - Risk_Segment_Prior_Ntm & Risk_Segment_During_Rx: The risk segment of patients before they started their treatment (prior to receiving the NTM medication) with VLR_LR and HR_VHR representing Very Low Risk/Low Risk and Very High Risk/High Risk respectively.
  - Change_Risk_Segment: If there was any change in Risk Segment.
  - Change_T_Score: The T- score is a measurement used to assess bone density in the context of osteoporosis. This value indicates the change in the patient's bone density relative to that of a healthy adult.
  - Adherent Flag: Adherence status of patients to their prescribed therapies and whether they followed the prescribed medication.

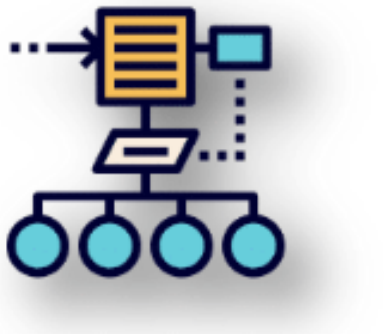
FINDINGS

1) There are higher non-persistency counts among patients who have low risk factors prior to taking their medication.

2) There are higher non-persistency counts among patients who have low risk and unknown factors during taking their medication.

3) Both the change in risk segment and the change in T-score are mostly either unknown or had no change with the count being higher for patients who were non-persistent.

4) Although many patients were adherent to their medication, there was still higher non-persistency among them.



MODEL BUILDING

MODEL BUILDING

After our analysis, we concluded that using the demographic analysis, physician anal[illegible] and Risk Factors and Change, Adherence to Therapy, & T-score Change Analysis

STEPS IN BUILDING OUR MODEL

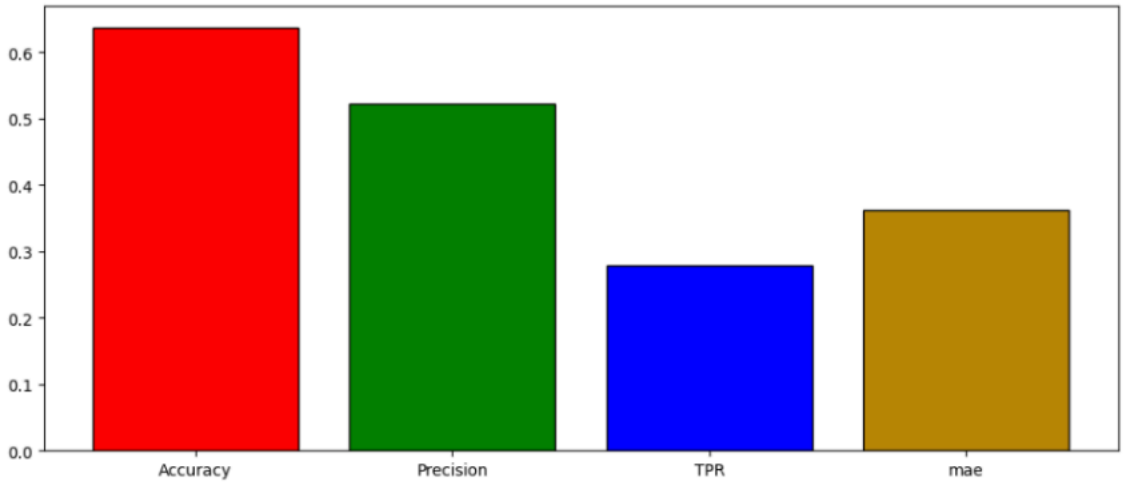
- ***Feature Engineering - This part includes formatting important data that proved to be very sensitive in determine the persistency***
- ***Encoding Categories - After feature engineering, we encode the region, ntm speciality and persistency flag and created a features list that takes users input***
- ***Splitting the dataset into test and train***
- ***Building the model and evaluating the Model using the confusion matrix***

```
# Make predictions
y_pred = model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

```
Accuracy: 0.635036496350365
```

VISUALIZING THE MODEL



Data Glacier
Your Deep Learning Partner