Exploratory Data Analysis

<Data Science Persistency of a Drug Project>

<December 2022>

Group Details

Name	Email	Country	College	Specialization
Han-Fu Lin	hanfu.lin@mail.utoronto.ca	Canada	University of Totonto	Data science
Aişe Refia Yilmaz	aiserefia.yilmaz@gmail.com	Turkey	Beykent University	Data science
Anıl Ilgın Büyüksaraç	Anililginb@gmail.com	Turkey	Y.t.u	Data science

Agenda

Data Analysis Approach
Problem Understanding
Business Understanding
Data Preparation and Understanding
Exploratory Data Analysis
EDA Summary
Model Selection and Development
Model Building
Final Recommendation
Conclusion

Data Analysis Approach

- One of the challenge 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.
- Explore and Understand the data
- Prepare and clean the data
- Analyze the data
- Find the features of drug persistency
- Give recommendations

Problem Understanding

Since the data is very large, we know that ABC pharmaceutical company wants to run the whole process faster and to get precise results.

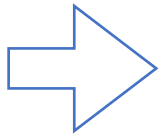
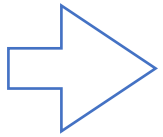
We have considered and measured many factors that do or do not affect permanence.

ML Problem:

With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

Business Understanding

- All process algorithms based on:

Problem  Model  Solution

Task

- **Problem understanding**
- **Data Understanding**
- **Data Cleaning and Feature engineering**
- **Model Development**
- **Model Selection**
- **Model Evaluation**
- **Report the accuracy, precision and recall of both the class of target variable**
- **Report ROC-AUC as well**
- **Deploy the model**
- **Explain the challenges and model selection**

DATA PREPARATION and UNDERSTANDING

- We used pandas library and Google Colab Notebook
- Data was cleaned and prepared for analyzation.

- Approaches:

  Number of NA values and missing values.

  Outliers and Skewed.

  Virtualization of data

EXPLORATORY DATA ANALYSIS

We analyzed gender, race, ethnicity, age, region, IDN Indicator values because demographic characteristics are personal and do not change.

We produced 'persistent and non-persistent of drug' histogram graphs of these properties.

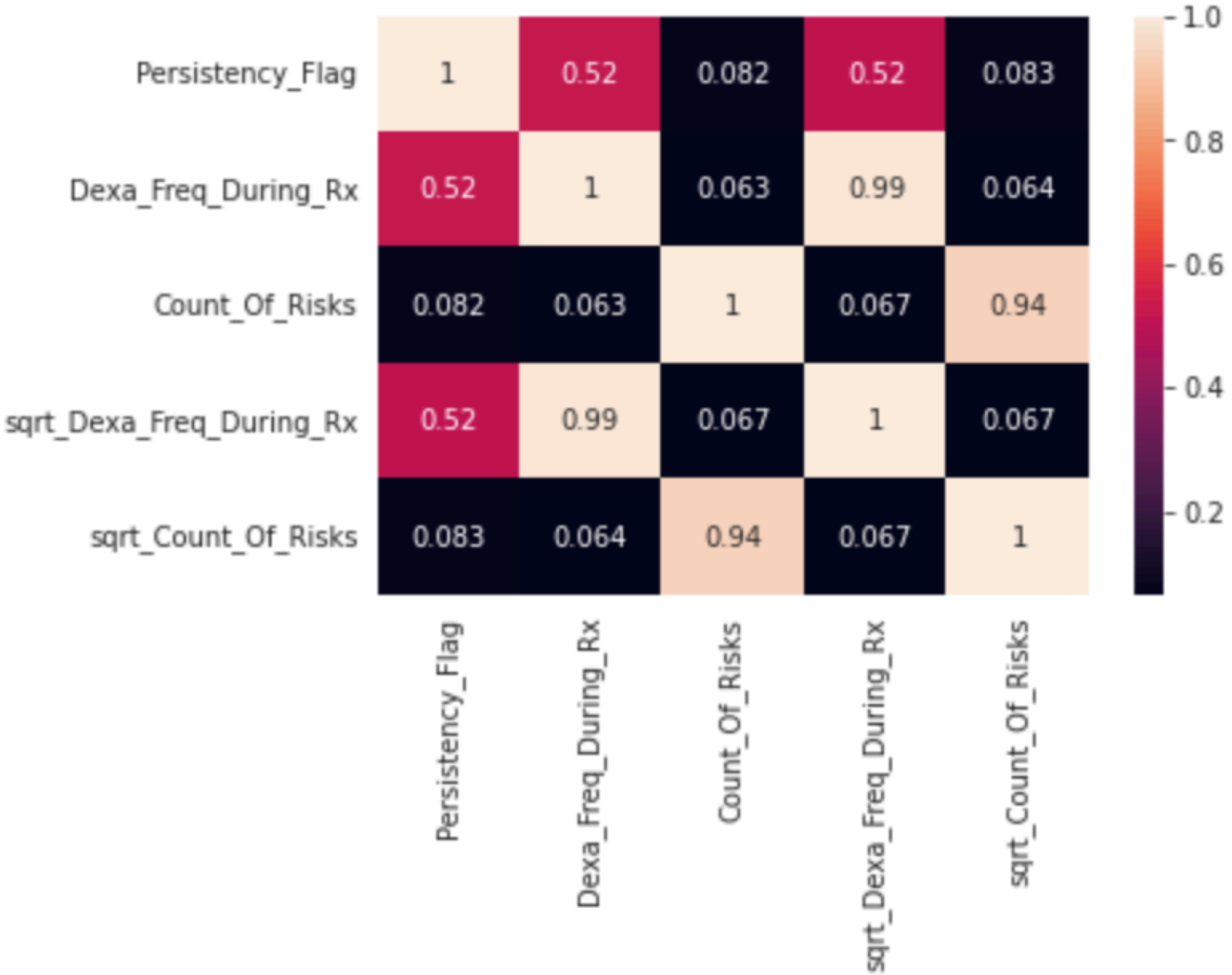
Data variables are presented in Heatmap.

According to bar graph, the highest values are seen in the midwest and south parts of the region permanent and non-permanent drug.

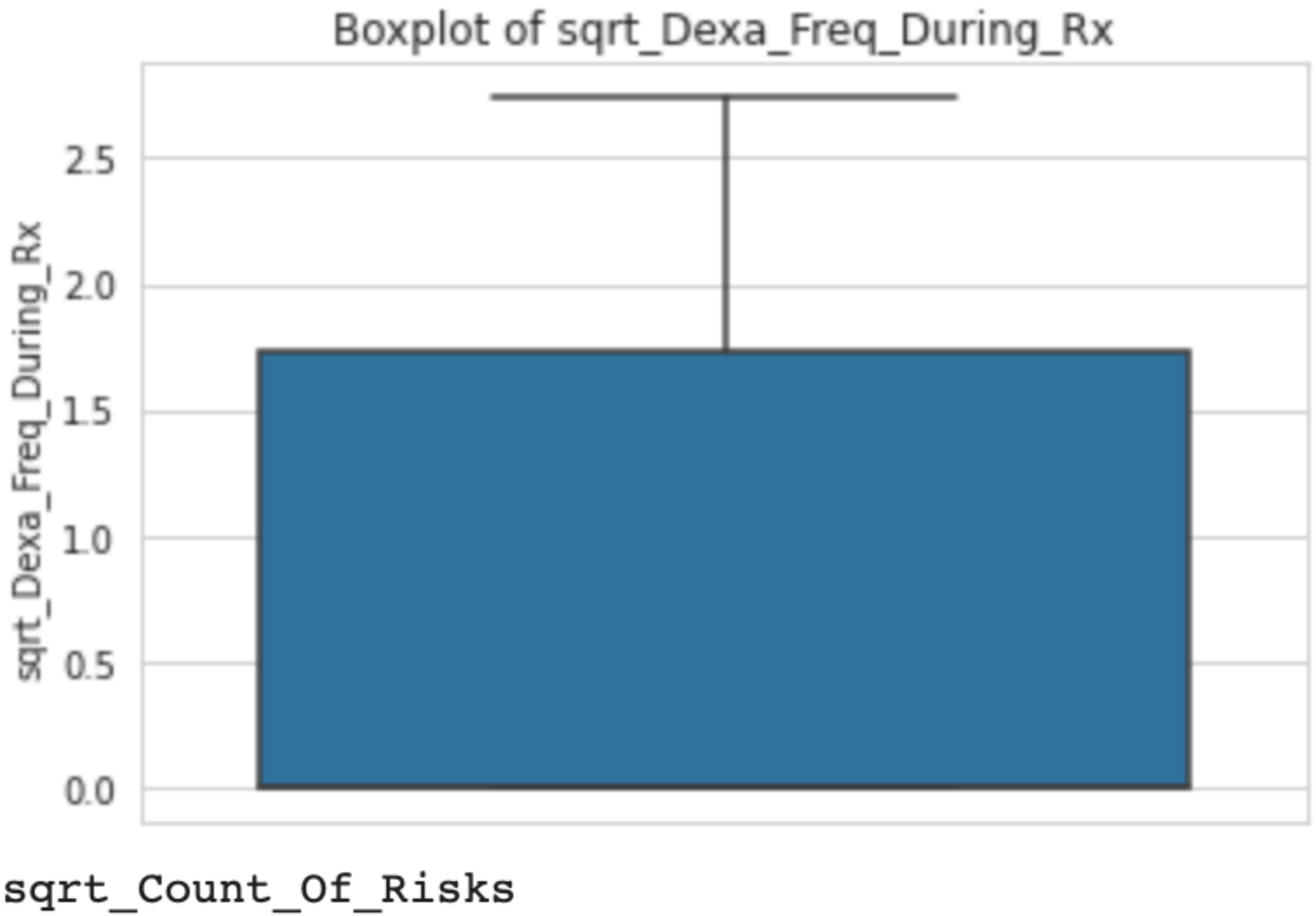
On the ethnicity persistent and non-persistent bar graph, not hispanic is clearly ahead of unknown and hispanic.

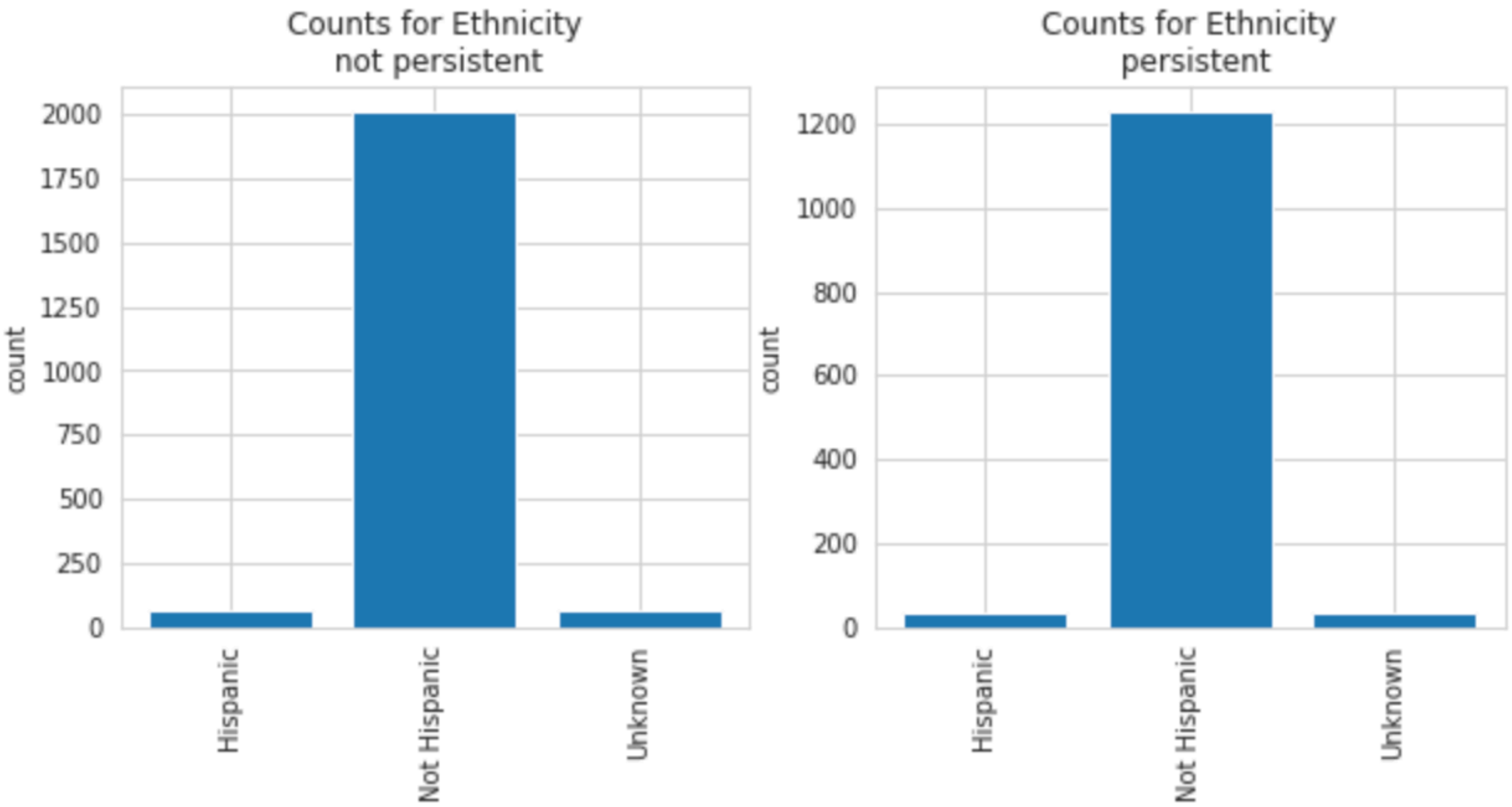
As a result of the analyzes made, we got different outputs every week and saved them.

Heatmap


Persistency_Flag
Dexa_Freq_During_Rx
Count_Of_Risks
sqrt_Dexa_Freq_During_Rx
sqrt_Count_Of_Risks
1
0.52
0.082
0.52
0.083
0.52
1
0.063
0.99
0.064
0.082
0.063
1
0.067
0.94
0.52
0.99
0.067
1
0.067
0.083
0.064
0.94
0.067
1
Persistency_Flag
Dexa_Freq_During_Rx
Count_Of_Risks
sqrt_Dexa_Freq_During_Rx
sqrt_Count_Of_Risks
1.0
0.8
0.6
0.4
0.2

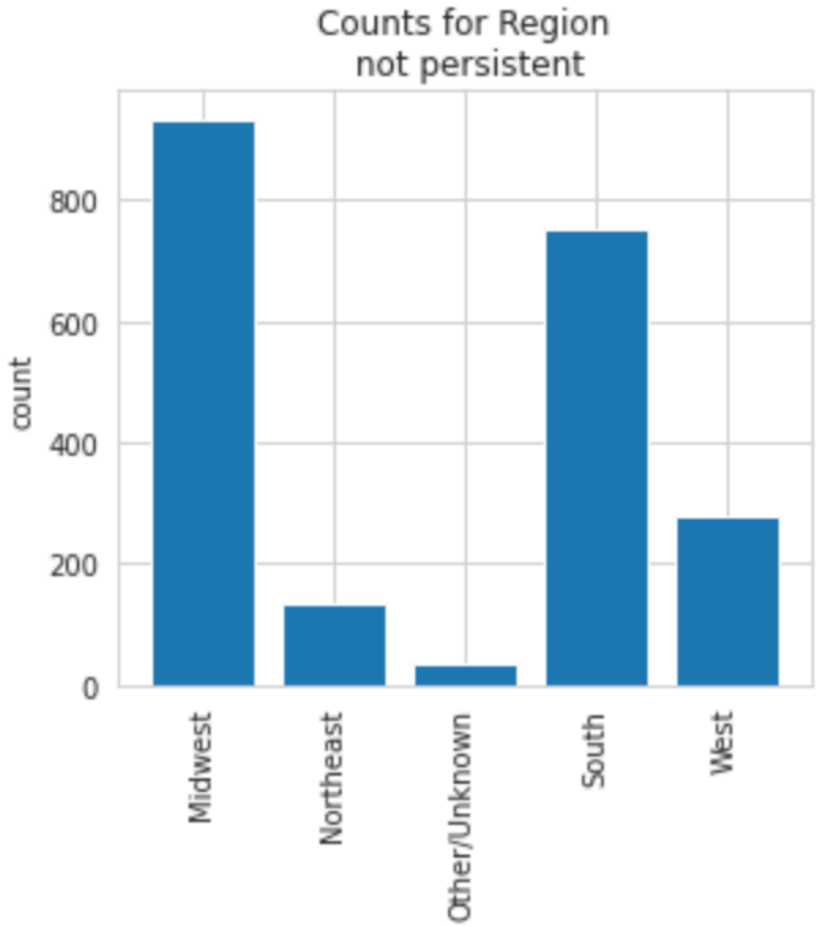
Outlier Test

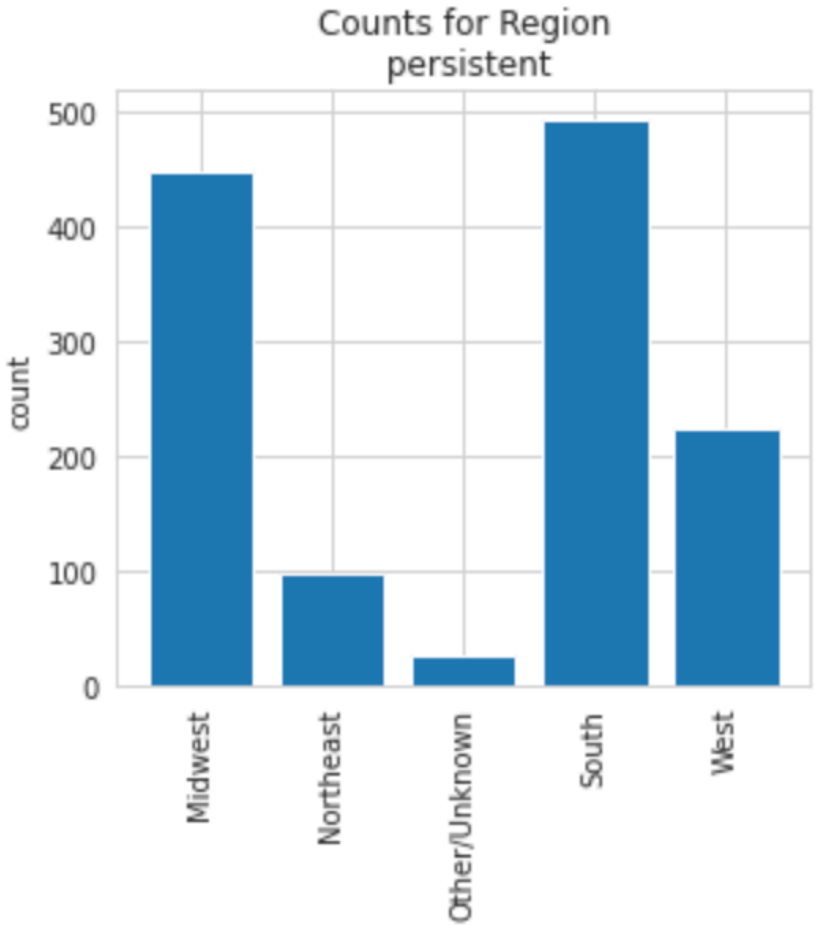




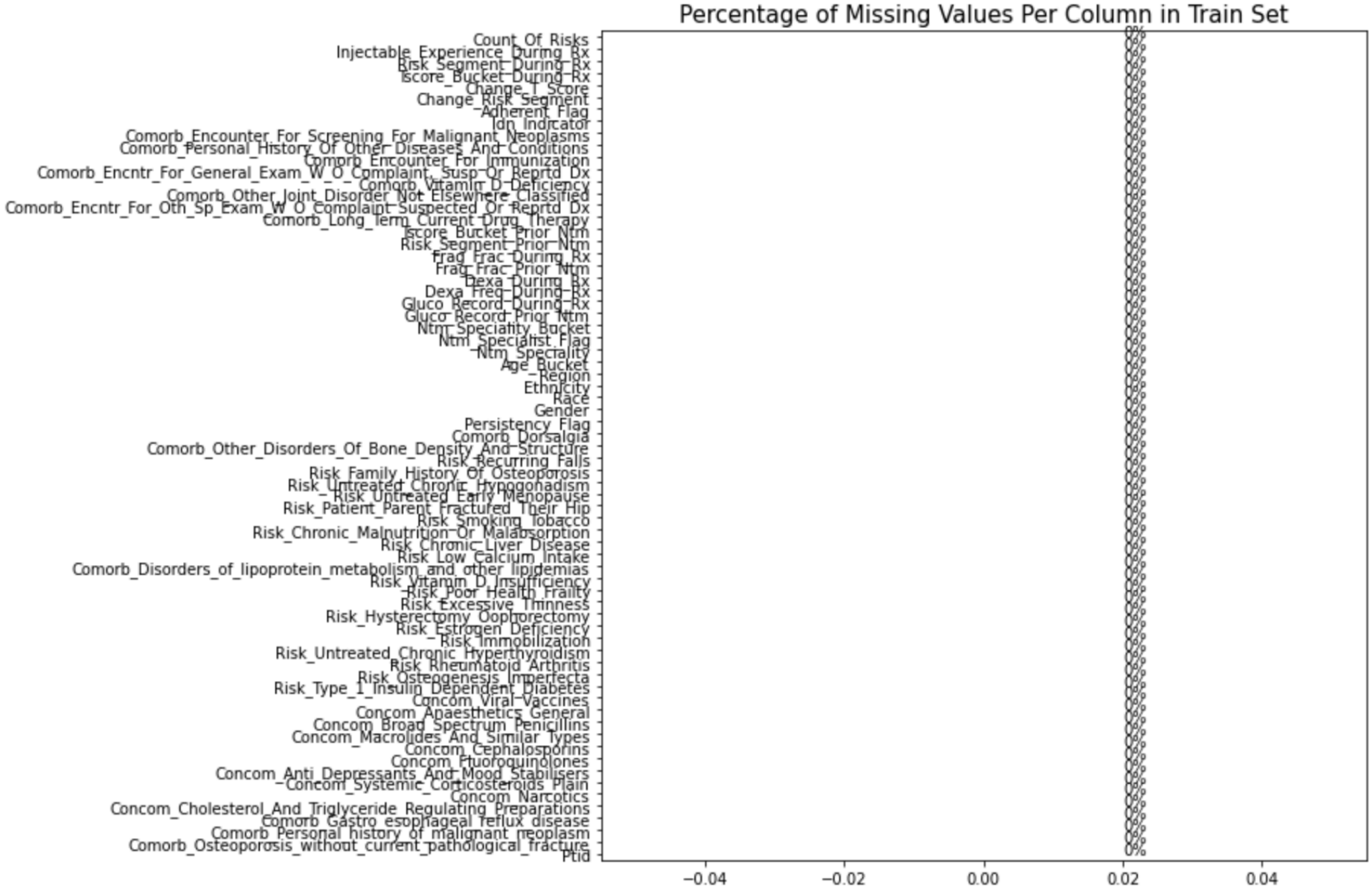
Although the values of the two graphs are different, they are quite similar in shape.

Persistency and Non-Persistent

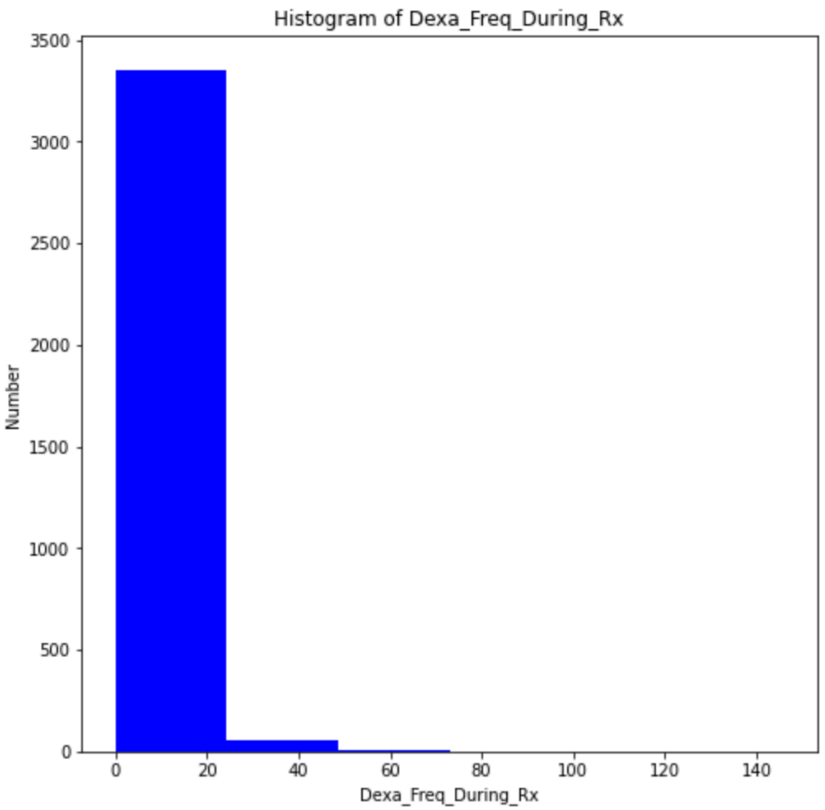


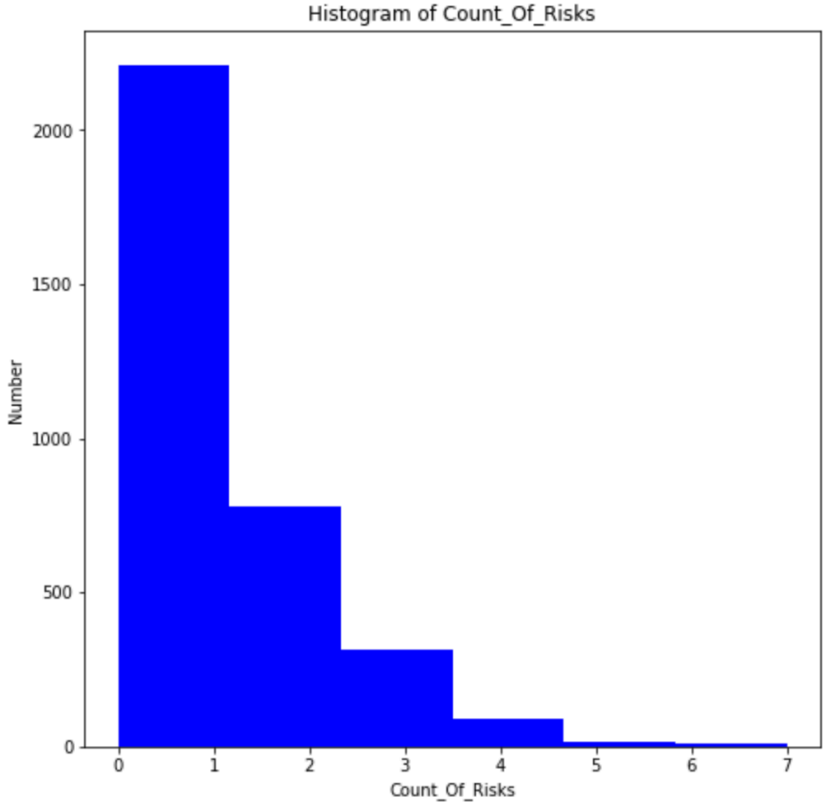


There are not very big difference between each graph.


Percentage of Missing Values Per Column in Train Set
Count_Of_Risks
Injectable_Experience_During_Rx
Risk_Segment_During_Rx
Tscore_Bucket_During_Rx
Change_T_Score
Change_Risk_Segment
Adherent_Flag
Idn_Indicator
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms
Comorb_Personal_History_Of_Other_Diseases_And_Conditions
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
Tscore_Bucket_Prior_Ntm
Risk_Segment_Prior_Ntm
Frag_Frac_During_Rx
Frag_Frac_Prior_Ntm
Dexa_During_Rx
Dexa_Freq_During_Rx
Gluco_Record_During_Rx
Gluco_Record_Prior_Ntm
Ntm_Speciality_Bucket
Ntm_Specialist_Flag
Ntm_Specialty
Age_Bucket
Region
Ethnicity
Race
Gender
Persistency_Flag
Comorb_Dorsalgia
Comorb_Other_Disorders_Of_Bone_Density_And_Structure
Risk_Recurring_Falls
Risk_Family_History_Of_Osteoporosis
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_Low_Calcium_Intake
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias
Risk_Vitamin_D_Insufficiency
Risk_Poor_Health_Frailty
Risk_Excessive_Thinness
Risk_Hysterectomy_Oophorectomy
Risk_Estrogen_Deficiency
Risk_Immobilization
Risk_Untreated_Chronic_Hyperthyroidism
Risk_Rheumatoid_Arthritis
Risk_Osteogenesis_Imperfecta
Risk_Type_1_Insulin_Dependent_Diabetes
Concom_Viral_Vaccines
Concom_Anaesthetics_General
Concom_Broad_Spectrum_Penicillins
Concom_Macrolides_And_Similar_Types
Concom_Cephalosporins
Concom_Fluoroquinolones
Concom_Anti_Depressants_And_Mood_Stabilisers
Concom_Systemic_Corticosteroids_Plain
Concom_Narcotics
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations
Comorb_Gastro_esophageal_reflux_disease
Comorb_Personal_history_of_malignant_neoplasm
Comorb_Osteoporosis_without_current_pathological_fracture
Ptid
0%
−0.04
−0.02
0.00
0.02
0.04

Persistency and Non-Persistent





Histogram of Dexa_Freq_During_Rx and Count_of_Risks are positively skewed distribution.

Model Selection and Development

- Logistic Regression result

```
print(f'AUC is {roc_auc_score(y_val,y_pred)}')
print(f'F1 score is {f1_score(y_val,y_pred>0.5)}')
```

```
AUC is 0.8860176017601761
F1 score is 0.8209806157354619
```

- Decision Tree result

```
print(f'AUC is {roc_auc_score(y_val,y_pred)}')
print(f'F1 score is {f1_score(y_val,y_pred>0.5)}')
```

```
AUC is 0.8202805280528053
F1 score is 0.7541371158392436
```

- LightGBM result

```
print(f'AUC is {roc_auc_score(y_val,y_pred)}')
print(f'F1 score is {f1_score(y_val,y_pred>0.5)}')
```

```
AUC is 0.9268206820682068
F1 score is 0.8711111111111111
```

Model Building

- Categorical data were converted into numerical data.

- Tree Classifier, Random Forest Classifier and LightGBM Classifier regression techniques were used.

EDA Summary

- Although there are many different factors apart from the analyzes we have made with the data we have, the result we have obtained regarding drug permanence is very high.

- There are two situations here. Firstly, this rate may decrease or increase with the data we do not have. Secondly, we can consider the 92 accuracy rate and values close to it.

Final Recommendation

- The LightGBM model is the most suitable model because it has the highest AUC and F1 score.

- It will give the best results in persistent and non-persistent drug effects compared to other models.

Thank You