Final Project Report

HEALTHCARE – DRUG PERSISTENCY

EMRE KORKUSUZ

YANJUN LIN

Table of Contents

Group Description

Group Name: Data Glacier Intern Group

Name: Emre Korkusuz
Yanjun Lin

E-mail: korkusuzemre1@gmail.com
yanjun.lin.andrie@gmail.com

Country: Turkey
USA

College: Trakya University
University of California Berkeley

Specialization : Data Science

Problem Description

ABC is a pharmaceutical company that wants to understand the persistency of a drug as per the physician's prescription for a patient. This company has approached an Analytics company to automate this process of identification. This report summarizes how our team came up with a solution to automate the persistency of a drug for the client ABC.

Machine Learning Problem

With an objective to gather insights on the factors that are impacting the persistency, then build a classification model to train, test, validate, and predict based on the given dataset.

Business Understanding

The pharma company ABC wants to understand about the persistency of a drug for a patient. There are lots of Non-Tuberculous Mycobacterial (NTM) infection data. ABC company wants to know whether the drug's effects on a patient are persistent given the prescription data. Based on the persistency count from the dataset, our team from Data Glacier will analyze, model, and predict drug persistency. Then the ABC company can make strategic production decisions on such drug to maximize its revenue.

Dataset

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
	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 Factor	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

Project Lifecycle

Weeks	Deadline	Plan
Week 07	Aug 04, 2022	Problem statement and Introduction
Week 08	Aug 11, 2022	Data preprocessing
Week 09	Aug 18, 2022	Feature Extraction
Week 10	Aug 25 2022	Building the Model
Week 11	Sep 01, 2022	Model Result Evaluation
Week 12	Sep 08, 2022	Flask Development + Heroku
Week 13	Sep 15, 2022	Final Report - Code Presentation

Data Intake Report

Data Intake Report

Name: Healthcare - Persistency of a drug
Report date : 04.08.2022
Internship Batch: LISUM11
Version: 1.0
Data scientist name: Emre Korkusuz – Yanjun Lin

Healthcare_dataset.csv details:

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	csv
Size of the data	892 KB

Healthcare_dataset.xlsx details:

Total number of observations	58
Total number of files	1
Total number of features	3
Base format of the file	xlsx
Size of the data	904 KB

Data Understanding

The Healthcare Dataset includes 69 columns and 3424 rows of observations. The target variable is Persistency_Flag with Boolean type of True or False. After displaying the data, it shows that there are 2 columns data of Integer type and the rest columns are either Boolean or String data type.

PtId	object
Persistency_Flag	object
Gender	object
Race	object
Ethnicity	object
Region	object
Age_Bucket	object
Ntm_Speciality	object
Ntm_Specialist_Flag	object
Ntm_Speciality_Bucket	object
Gluco_Record_Prior_Ntm	object
Gluco_Record_During_Rx	object
Dexa_Freq_During_Rx	int64
Dexa_During_Rx	object
Frag_Frac_Prior_Ntm	object
Frag_Frac_During_Rx	object
Risk_Segment_Prior_Ntm	object
Tscore_Bucket_Prior_Ntm	object
Risk_Segment_During_Rx	object
Tscore_Bucket_During_Rx	object
Change_T_Score	object
Change_Risk_Segment	object
Adherent_Flag	object
Idn_Indicator	object
Injectable_Experience_During_Rx	object
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms	object
Comorb_Encounter_For_Immunization	object
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx	object
Comorb_Vitamin_D_Deficiency	object
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified	object
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx	object
Comorb_Long_Term_Current_Drug_Therapy	object
Comorb_Dorsalgia	object
Comorb_Personal_History_Of_Other_Diseases_And_Conditions	object
Comorb_Other_Disorders_Of_Bone_Density_And_Structure	object
Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias	object
Comorb_Osteoporosis_without_current_pathological_fracture	object
Comorb_Personal_history_of_malignant_neoplasm	object
Comorb_Gastro_esophageal_reflux_disease	object
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations	object
Concom_Narcotics	object
Concom_Systemic_Corticosteroids_Plain	object
Concom_Anti_Depressants_And_Mood_Stabilisers	object
Concom_Fluoroquinolones	object
Concom_Cephalosporins	object
Concom_Macrolides_And_Similar_Types	object
Concom_Broad_Spectrum_Penicillins	object
Concom_Anaesthetics_General	object
Concom_Viral_Vaccines	object
Risk_Type_1_Insulin_Dependent_Diabetes	object
Risk_Osteogenesis_Imperfecta	object
Risk_Rheumatoid_Arthritis	object
Risk_Untreated_Chronic_Hyperthyroidism	object
Risk_Untreated_Chronic_Hypogonadism	object
Risk_Untreated_Early_Menopause	object
Risk_Patient_Parent_Fractured_Their_Hip	object
Risk_Smoking_Tobacco	object
Risk_Chronic_Malnutrition_Or_Malabsorption	object
Risk_Chronic_Liver_Disease	object
Risk_Family_History_Of_Osteoporosis	object
Risk_Low_Calcium_Intake	object
Risk_Vitamin_D_Insufficiency	object
Risk_Poor_Health_Frailty	object
Risk_Excessive_Thinness	object
Risk_Hysterectomy_Oophorectomy	object
Risk_Estrogen_Deficiency	object
Risk_Immobilization	object
Risk_Recurring_Falls	object
Count_Of_Risks	int64

EDA (Exploratory Data Analysis)

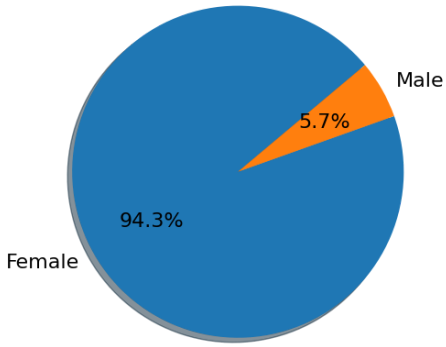
- Null Values: This dataset has no Null values
- Duplicates: This dataset has no Duplicated values
- Features: We grouped all features into 4 sub-groups as shown below

```
demographics_features = ['Gender', 'Race', 'Ethnicity', 'Region','Age_Bucket','Idn_Indicator']

provider_features = ['Ntm_Speciality', 'Ntm_Specialist_Flag','Ntm_Speciality_Bucket','Change_T_Score','Risk_Segment_Prior_Ntm',
                     'Tscore_Bucket_Prior_Ntm','Tscore_Bucket_During_Rx','Change_Risk_Segment','Risk_Segment_During_Rx']

clinical_features = ['Dexa_Freq_During_Rx', 'Dexa_During_Rx','Frag_Frac_Prior_Ntm', 'Frag_Frac_During_Rx','Gluco_Record_Prior_Ntm',
                     'Gluco_Record_During_Rx','Injectable_Experience_During_Rx','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', 'Count_Of_Risks']

treatment_features = ['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','Adherent_Flag']
```

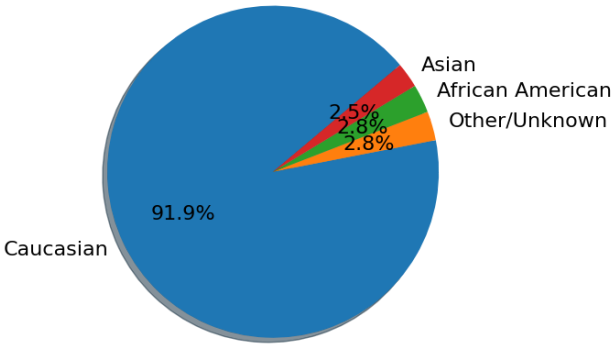
Features Analysis

Demographics Features Analysis

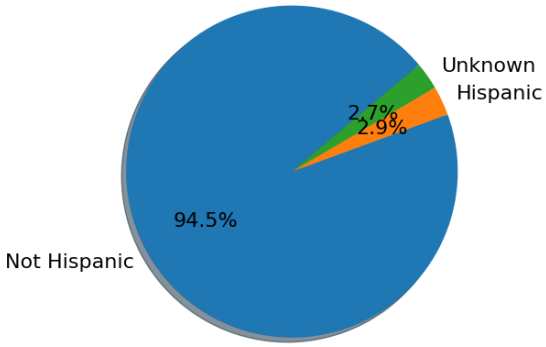
Gender Distribution



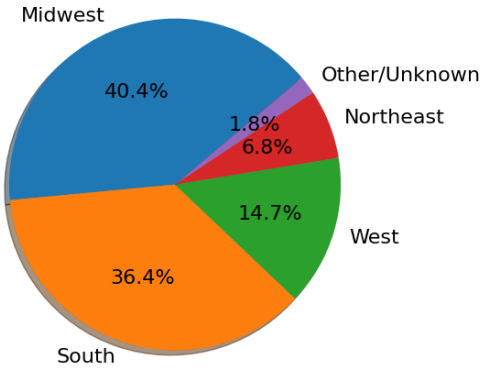
Race Distribution



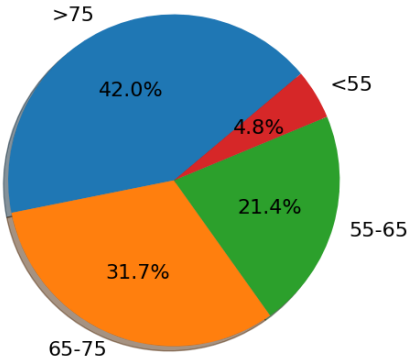
Ethnicity Distribution



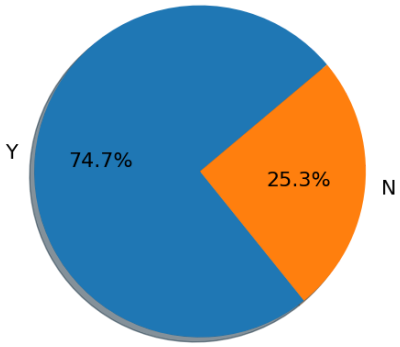
Region Distribution

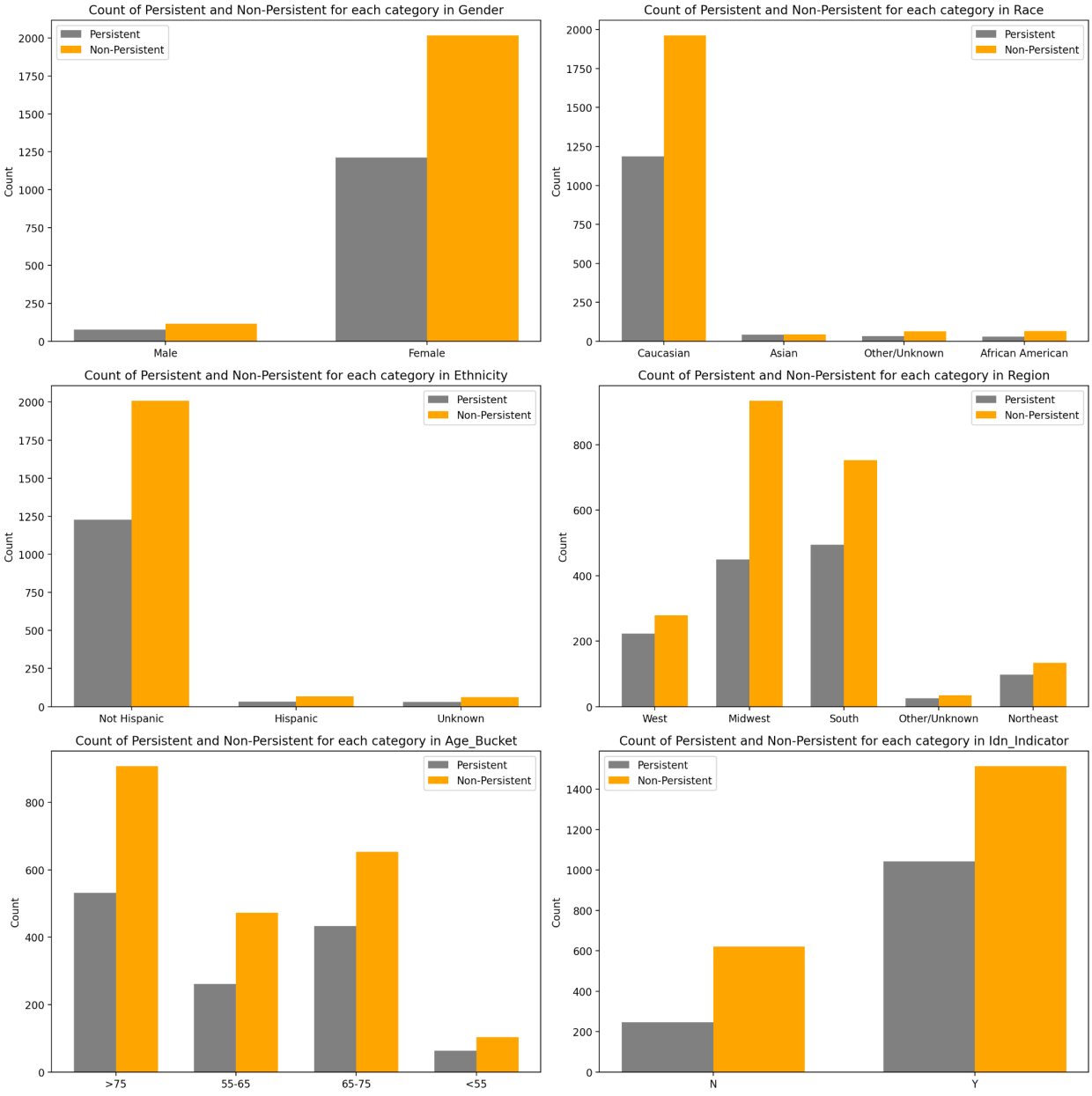


Age_Bucket Distribution



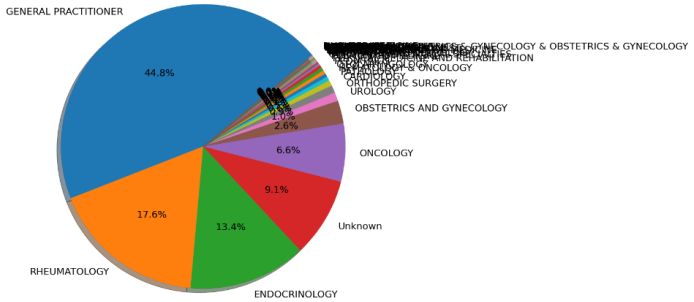
Idn_Indicator Distribution



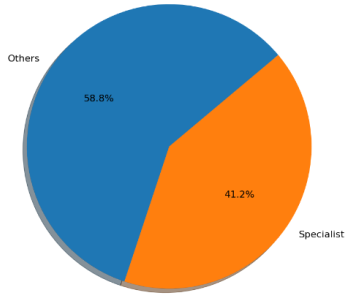

Count of Persistent and Non-Persistent for each category in Gender
Persistent
Non-Persistent
Count
Male
Female
Count of Persistent and Non-Persistent for each category in Race
Caucasian
Asian
Other/Unknown
African American
Count of Persistent and Non-Persistent for each category in Ethnicity
Not Hispanic
Hispanic
Unknown
Count of Persistent and Non-Persistent for each category in Region
West
Midwest
South
Other/Unknown
Northeast
Count of Persistent and Non-Persistent for each category in Age_Bucket
>75
55-65
65-75
<55
Count of Persistent and Non-Persistent for each category in Idn_Indicator
N
Y

Providers Features Analysis

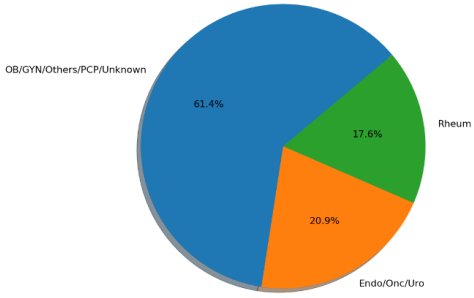
Ntm_Speciality Distribution



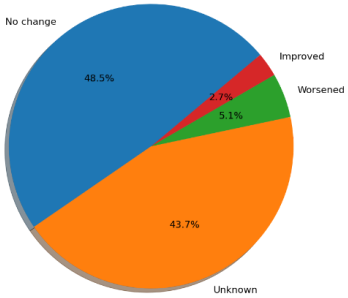
Ntm_Specialist_Flag Distribution



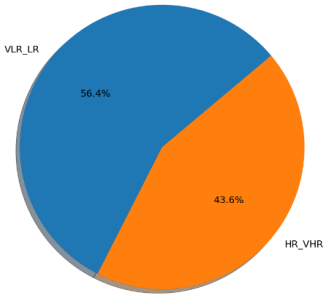
Ntm_Speciality_Bucket Distribution



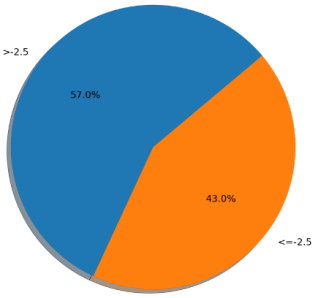
Change_T_Score Distribution



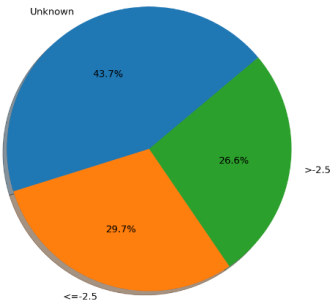
Risk_Segment_Prior_Ntm Distribution



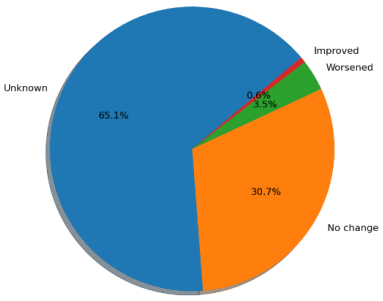
Tscore_Bucket_Prior_Ntm Distribution



Tscore_Bucket_During_Rx Distribution

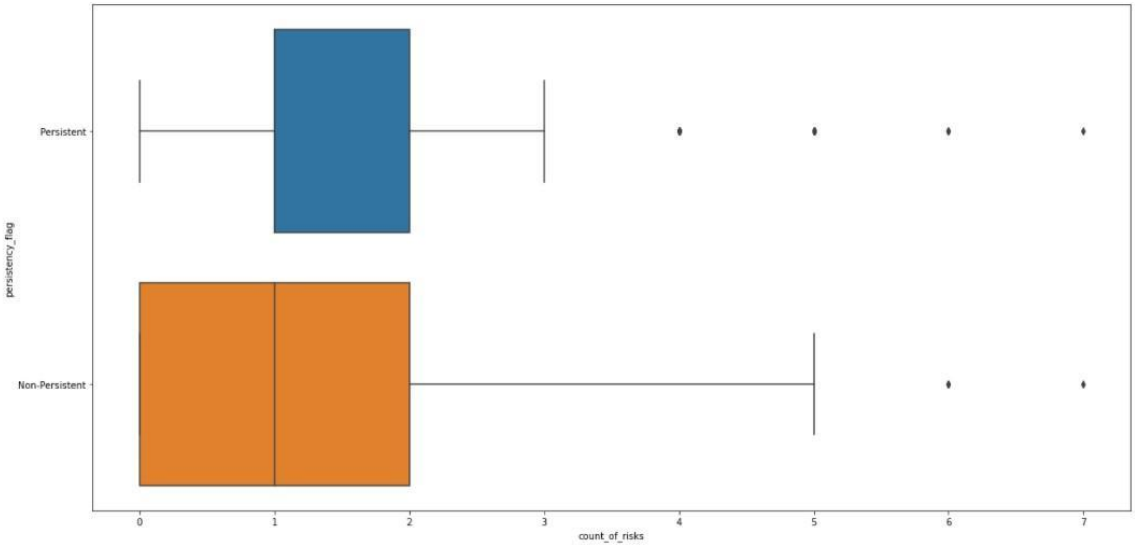


Change_Risk_Segment Distribution



Problems

- Outliers: We have only two numerical columns and both of them have some outliers.
    - **count_of_risks:**



    - **dexa_freq_during_rx:**

Dataset describe:

In general, the relations between our data and the output of mathematical calculations are attached.

Out[11]:

	Dexa_Freq_During_Rx	Count_Of_Risks
count	3424.000000	3424.000000
mean	3.016063	1.239486
std	8.136545	1.094914
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	1.000000
75%	3.000000	2.000000
max	146.000000	7.000000

Dataset isnull().sum():

The isnull command is attached, which allows us to check whether there is empty data in the branches of our data, if any, and gives us the total.

Out[13]:

```
Ptid                                0
Persistency_Flag                    0
Gender                              0
Race                                0
Ethnicity                           0
                                   ..
Risk_Hysterectomy_Oophorectomy      0
Risk_Estrogen_Deficiency            0
Risk_Immobilization                 0
Risk_Recurring_Falls                0
Count_Of_Risks                      0
Length: 69, dtype: int64
```

Dataset value_counts:i

The output of the code that lists us in detail how many columns written in the contents of our data is attached.

In [14]:

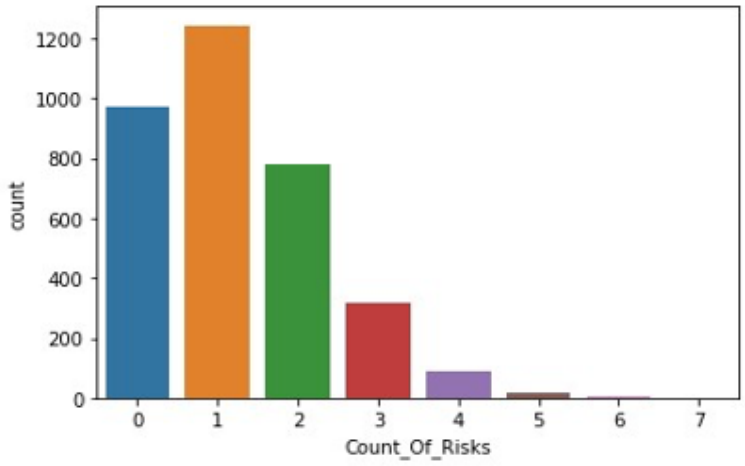
```
for f in kendi_ozeligi:
    tab = veri[f].value_counts()
    print('%s:\t%s' % (f, ', '.join([ ("%s(%d)" %(tab.index[i], tab.values[i])) for i in range(len(tab))]) ))
```

```
Ntm_Speciality_Bucket:  OB/GYN/Others/PCP/Unknown(2104), Endo/Onc/Uro(716), Rheum(604)
Frag_Frac_Prior_Ntm:    N(2872), Y(552)
Concom_Anti_Depressants_And_Mood_Stabilisers:   N(2465), Y(959)
Comorb_Other_Disorders_Of_Bone_Density_And_Structure:   N(2906), Y(518)
Risk_Excessive_Thinness:        N(3357), Y(67)
Ethnicity:      Not Hispanic(3235), Hispanic(98), Unknown(91)
Comorb_Personal_history_of_malignant_neoplasm:  N(2775), Y(649)
Adherent_Flag:  Adherent(3251), Non-Adherent(173)
Concom_Viral_Vaccines:  N(3071), Y(353)
Risk_Immobilization:    N(3410), Y(14)
Ntm_Speciality: GENERAL PRACTITIONER(1535), RHEUMATOLOGY(604), ENDOCRINOLOGY(458), Unknown(310), ONCOLOGY(225), OBSTETRICS AND GYNECOLOGY(90), UROLO
GY(33), ORTHOPEDIC SURGERY(30), CARDIOLOGY(22), PATHOLOGY(16), HEMATOLOGY & ONCOLOGY(14), OTOLARYNGOLOGY(14), PEDIATRICS(13), PHYSICAL MEDICINE AND
REHABILITATION(11), PULMONARY MEDICINE(8), SURGERY AND SURGICAL SPECIALTIES(8), PSYCHIATRY AND NEUROLOGY(4), NEPHROLOGY(3), ORTHOPEDICS(3), PLASTIC
SURGERY(2), VASCULAR SURGERY(2), HOSPICE AND PALLIATIVE MEDICINE(2), GERIATRIC MEDICINE(2), GASTROENTEROLOGY(2), TRANSPLANT SURGERY(2), CLINICAL NUR
SE SPECIALIST(1), OCCUPATIONAL MEDICINE(1), HOSPITAL MEDICINE(1), OPHTHALMOLOGY(1), PODIATRY(1), EMERGENCY MEDICINE(1), RADIOLOGY(1), OBSTETRICS & O
BSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY(1), NEUROLOGY(1), PAIN MEDICINE(1), NUCLEAR MEDICINE(1)
Risk_Untreated_Chronic_Hypogonadism:    N(3297), Y(127)
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms: N(1891), Y(1533)
Injectable_Experience_During_Rx:        Y(3056), N(368)
Gender: Female(3230), Male(194)
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx:     N(2633), Y(791)
Change_Risk_Segment:    Unknown(2229), No change(1052), Worsened(121), Improved(22)
Dexa_During_Rx: N(2488), Y(936)
Tscore_Bucket_Prior_Ntm:        >-2.5(1951), <=-2.5(1473)
Risk_Segment_During_Rx: Unknown(1497), HR_VHR(965), VLR_LR(962)
Risk_Family_History_Of_Osteoporosis:    N(3066), Y(358)
Risk_Rheumatoid_Arthritis:      N(3294), Y(130)
Persistency_Flag:       Non-Persistent(2135), Persistent(1289)
Comorb_Dorsalgia:       N(2645), Y(779)
Concom_Cephalosporins:  N(2821), Y(603)
Risk_Vitamin_D_Insufficiency:   N(1788), Y(1636)
Comorb_Vitamin_D_Deficiency:    N(2331), Y(1093)
Gluco_Record_Prior_Ntm: N(2619), Y(805)
Risk_Chronic_Malnutrition_Or_Malabsorption:     N(2954), Y(470)
Risk_Osteogenesis_Imperfecta:   N(3421), Y(3)
Risk_Untreated_Chronic_Hyperthyroidism: N(3422), Y(2)
Frag_Frac_During_Rx:    N(3007), Y(417)
Tscore_Bucket_During_Rx:        Unknown(1497), <=-2.5(1017), >-2.5(910)
Risk_Hysterectomy_Oophorectomy: N(3370), Y(54)
Region: Midwest(1383), South(1247), West(502), Northeast(232), Other/Unknown(60)
Risk_Segment_Prior_Ntm: VLR_LR(1931), HR_VHR(1493)
Idn_Indicator:  Y(2557), N(867)
Comorb_Personal_History_Of_Other_Diseases_And_Conditions:       N(2747), Y(677)
Comorb_Osteoporosis_without_current_pathological_fracture:      N(2507), Y(917)
Concom_Systemic_Corticosteroids_Plain:  N(2451), Y(973)
Concom_Narcotics:       N(2191), Y(1233)
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx:        N(2072), Y(1352)
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified:   N(2425), Y(999)
Comorb_Long_Term_Current_Drug_Therapy:  N(2607), Y(817)
Risk_Recurring_Falls:   N(3355), Y(69)
Concom_Cholesterol_And_Triglyceride_Regulating_Preparations:    N(2242), Y(1182)
Comorb_Encounter_For_Immunization:      N(1911), Y(1513)
```

Dataset Count_of_Risks countplot:

Attached is the chart of the risks that emerge from the results of our data in seaborn.

```
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa24cbcaed0>
```



Manipulations on the dataset

When our data is multiplied by 0.01 percent, when we create another data and assign our original data to this data, when we start this new data from the number that comes out, the changes and mathematical arrangements in our data are visible.

```
In [28]: len(veri) * 0.01
Out[28]: 34.24
In [29]: yuzdeDoksanDokuzDf = veri.sort_values("Count_Of_Risks",ascending = False).iloc[34:]
In [30]: yuzdeDoksanDokuzDf.describe()
```

Out[30]:

	Dexa_Freq_During_Rx	Count_Of_Risks
count	3390.000000	3390.000000
mean	3.020944	1.202065
std	8.165475	1.030744
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	1.000000
75%	3.000000	2.000000
max	146.000000	4.000000

Relationship between two columns in data:

For our data, the average of the mathematical columns relative to each other and the study of the relationships between them is attached.

```
In [40]: veri.groupby("Dexa_Freq_During_Rx").mean()["Count_Of_Risks"]

Out[40]: Dexa_Freq_During_Rx
0     1.192524
1     1.500000
2     1.250000
3     1.260870
4     1.294118
5     1.552632
6     1.448598
7     1.462366
8     1.239437
9     1.343750
10    1.636364
11    1.200000
12    1.596154
13    0.736842
14    1.157895
15    0.666667
16    1.285714
17    0.714286
18    1.142857
19    0.666667
20    1.600000
21    1.142857
22    1.538462
23    2.000000
24    1.200000
25    2.000000
26    1.100000
27    0.000000
28    1.428571
29    1.000000
30    1.428571
32    1.666667
33    2.000000
34    1.000000
35    3.000000
36    0.800000
37    1.000000
38    0.000000
39    1.500000
40    2.000000
42    3.000000
44    0.000000
45    2.000000
48    1.000000
50    0.000000
52    1.000000
54    2.000000
58    1.500000
66    0.000000
68    1.000000
69    1.000000
```

Conversion of data to mathematical monuts:

The output, in which the objects written in the columns in our data are transformed into mathematical expressions, is attached.

In [42]:
```
level_substitution = {}

def levels2index(levels):
    dct = {}
    for i in range(len(levels)):
        dct[levels[i]] = i
    return dct

df_num = veri.copy()

for c in kendi_ozeligi:
    level_substitution[c] = levels2index(veri[c].unique())
    df_num[c].replace(level_substitution[c], inplace=True)

df_num
```

Out[42]:

	Ptid	Persistency_Flag	Gender	Race	Ethnicity	Region	Age_Bucket	Ntm_Speciality	Ntm_Specialist_Flag	Ntm_Speciality_Bucket	...	Risk_Family_History_Of_Osteoporosis
0	0	0	0	0	0	0	0	0	0	0	...	0
1	1	1	0	1	0	0	1	0	0	0	...	0
2	2	1	1	2	1	1	2	0	0	0	...	0
3	3	1	1	0	0	1	0	0	0	0	...	0
4	4	1	1	0	0	1	0	0	0	0	...	0
...	...	...	...	...	...	...	...	...	...	...	...	...
3419	3419	0	1	0	0	2	0	0	0	0	...	0
3420	3420	0	1	0	0	2	0	1	0	0	...	0
3421	3421	0	1	0	0	2	0	2	1	1	...	0
3422	3422	1	1	0	0	2	1	1	0	0	...	0
3423	3423	1	1	0	0	2	2	1	0	0	...	0

3424 rows × 69 columns

The help text about our codes that turn into math commands is attached.

In [43]:
```
level_substitution
```

Out[43]:
```
{'Ntm_Speciality_Bucket': {'OB/GYN/Others/PCP/Unknown': 0,
  'Endo/Onc/Uro': 1,
  'Rheum': 2},
 'Frag_Frac_Prior_Ntm': {'N': 0, 'Y': 1},
 'Concom_Anti_Depressants_And_Mood_Stabilisers': {'N': 0, 'Y': 1},
 'Comorb_Other_Disorders_Of_Bone_Density_And_Structure': {'N': 0, 'Y': 1},
 'Risk_Excessive_Thinness': {'N': 0, 'Y': 1},
 'Ethnicity': {'Not Hispanic': 0, 'Hispanic': 1, 'Unknown': 2},
 'Comorb_Personal_history_of_malignant_neoplasm': {'N': 0, 'Y': 1},
 'Adherent_Flag': {'Adherent': 0, 'Non-Adherent': 1},
 'Concom_Viral_Vaccines': {'N': 0, 'Y': 1},
 'Risk_Immobilization': {'N': 0, 'Y': 1},
 'Ntm_Speciality': {'GENERAL PRACTITIONER': 0,
  'Unknown': 1,
  'ENDOCRINOLOGY': 2,
  'RHEUMATOLOGY': 3,
  'ONCOLOGY': 4,
  'PATHOLOGY': 5,
  'OBSTETRICS AND GYNECOLOGY': 6,
  'PSYCHIATRY AND NEUROLOGY': 7,
  'ORTHOPEDIC SURGERY': 8,
  'PHYSICAL MEDICINE AND REHABILITATION': 9,
  'SURGERY AND SURGICAL SPECIALTIES': 10,
  'PEDIATRICS': 11,
  'PULMONARY MEDICINE': 12,
  'HEMATOLOGY & ONCOLOGY': 13,
  'UROLOGY': 14,
  'PAIN MEDICINE': 15,
  'NEUROLOGY': 16,
  'RADIOLOGY': 17,
  'GASTROENTEROLOGY': 18,
  'EMERGENCY MEDICINE': 19,
  'PODIATRY': 20,
  'OPHTHALMOLOGY': 21,
  'OCCUPATIONAL MEDICINE': 22,
  'TRANSPLANT SURGERY': 23,
  'PLASTIC SURGERY': 24,
  'CLINICAL NURSE SPECIALIST': 25,
  'OTOLARYNGOLOGY': 26,
  'HOSPITAL MEDICINE': 27,
  'ORTHOPEDICS': 28,
  'NEPHROLOGY': 29,
  'GERIATRIC MEDICINE': 30,
  'HOSPICE AND PALLIATIVE MEDICINE': 31,
  'OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY': 32,
  'VASCULAR SURGERY': 33,
  'CARDIOLOGY': 34,
  'NUCLEAR MEDICINE': 35},
 'Risk_Untreated_Chronic_Hypogonadism': {'N': 0, 'Y': 1},
 'Comorb_Encounter_For_Screening_For_Malignant_Neoplasms': {'N': 0, 'Y': 1},
 'Injectable_Experience_During_Rx': {'Y': 0, 'N': 1},
 'Gender': {'Male': 0, 'Female': 1},
 'Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx': {'Y': 0,
```

Correlation Analysis

Model Training & Testing

Classifiers Used

Classifiers used include models from Linear classifier, Ensemble & Boosting Models, and Neural Network model.

Linear Classifiers:

- Ridge Classifier
- SGD Classifier
- Logistic Regression Classifier

Ensemble & Boosting Models:

- Bagging Classifier
- Gradient Boosting Classifier
- Random forest
- ExtraTrees Classifier
- AdaBoost
- XGBoost Classifier
- Stacking Classifier

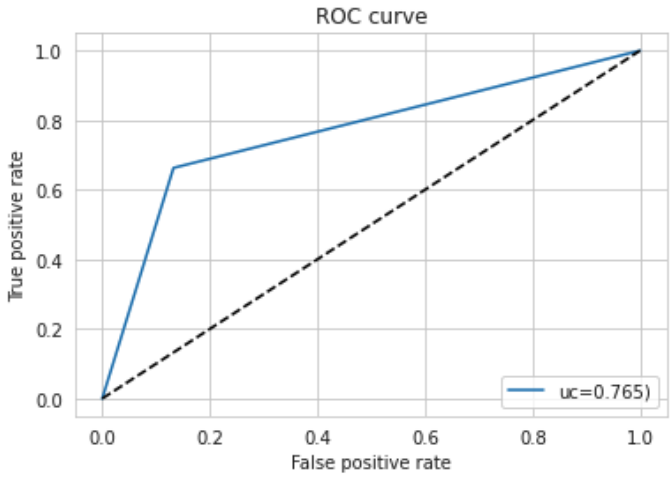
Neural Network:

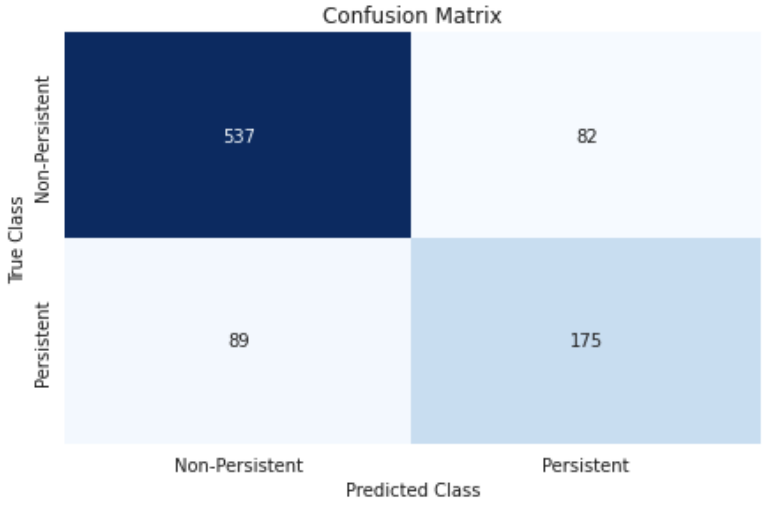
- Multi-layer Neural Network
- Multi-layer Perceptron

Best performing models are listed as follow:

Ridge Classifier

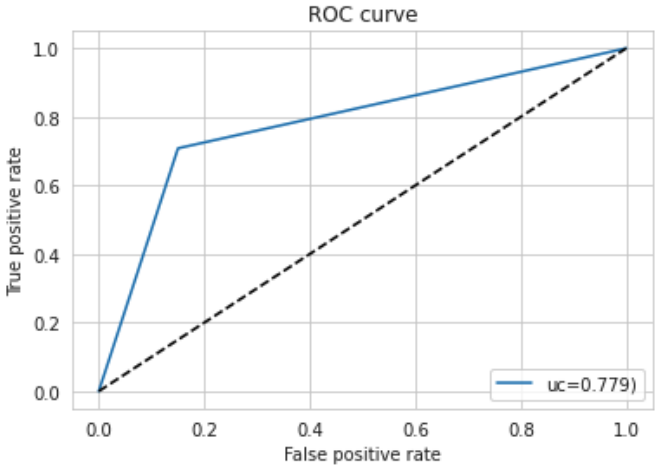
```
Accuracy : 0.8063420158550396
Precision : 0.6809338521400778
Recall : 0.6628787878787878
F1 Score : 0.6717850287907869
                precision    recall  f1-score   support

Non-Persistent       0.86      0.87      0.86       619
    Persistent       0.68      0.66      0.67       264

      accuracy                           0.81       883
     macro avg       0.77      0.77      0.77       883
  weighted avg       0.80      0.81      0.81       883

AUC : 0.7652035296421402
```

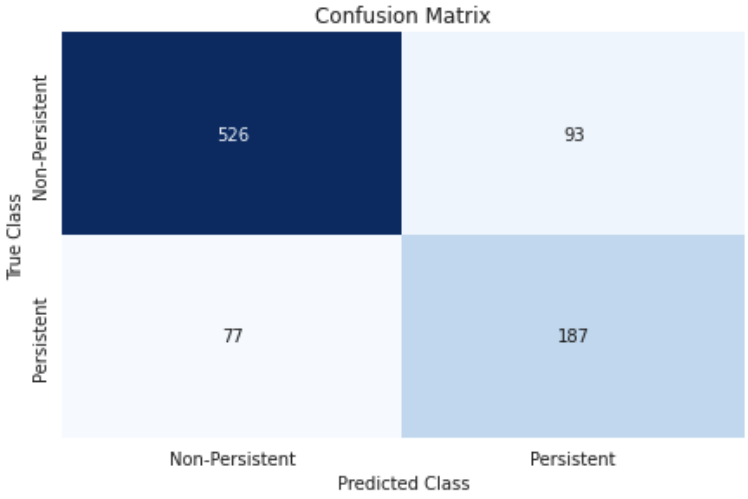




AdaBoost

```
Accuracy : 0.8074745186862967
Precision : 0.6678571428571428
Recall : 0.7083333333333334
F1 Score : 0.6875
                precision    recall  f1-score   support

Non-Persistent       0.87      0.85      0.86       619
    Persistent       0.67      0.71      0.69       264

      accuracy                           0.81       883
     macro avg       0.77      0.78      0.77       883
  weighted avg       0.81      0.81      0.81       883

AUC : 0.7790455035002694
```





XGBoost

```
Accuracy : 0.8063420158550396
Precision : 0.6795366795366795
Recall : 0.6666666666666666
F1 Score : 0.6730401529636711
                precision    recall  f1-score   support

Non-Persistent       0.86      0.87      0.86       619
    Persistent       0.68      0.67      0.67       264

      accuracy                           0.81       883
     macro avg       0.77      0.77      0.77       883
  weighted avg       0.81      0.81      0.81       883

AUC : 0.7662897145934302
```