Week 9: Problem Understanding

Team Name:				
Name	Email	Country	College/Company	Specialization
Bao Khanh Nguyen	Nguyenkhanhbao8695@gmail.com	USA	American Energy Project	Data Science
Seyedeh Marzieh Hosseini	shosseini@uni-potsdam.de	Germany	University of Potsdam	Data Science
Guillermo Leija	leija.guillermo@gmail.com			Data Science
Zain Ul Haq		Germany		Data Science

Project Life Cycle

Tasks	**08/11/2021 Week 0**	**15/11/2021 Week 1**	**22/11/2021 Week 2**	**29/11/2021 Week 3**	**6/12/2021 Week 4**
Week 7					
Week 8					
Week 9					
Week 10					
Week 11					
Week 12					

Problem Description

ABC is a pharmaceutical business that wants to know the persistency of a drug after a physician has prescribed it for a patient. This company has approached an analytics firm to automate the identifying procedure. This analytics firm has entrusted our team with the task of developing a solution to automate the persistence of a medicine for the client ABC.

Business Understanding

One of the long-lasting business issues in the world of pharmaceutical companies is the persistency of drugs which can significantly affect the outcome of medical treatments. One of the important factors that is related to persistency is the adherence of the patient to the prescribed regimens, meaning if the patient is committed to the prescribed regimens or not. There is a lot of information about Non-Tuberculous Mycobacterial (NTM) infections. In fact, related studies show that around 50%-60% of the patients with different illnesses in US miss doses, take the wrong doses, or drop off treatment in the first year. Additionally, the illness, either chronic or acute can be related to the adherence and persistency of drugs.

ABC company also one of pharmaceutical companies, wants to know how long a medicine will last in a patient's system (persistency of a drug). Based on prescription data, the ABC corporation needs to determine whether a patient is persistent or not. ABC pharma would manufacture medicines in that number based on the persistency count so that they could operate their firm effectively and avoid the risks of NTM infections.

Data Intake Report

Name: Health care- Data Science Specialization
Report date: 5 November 2021
Internship Batch: LISUM04
Version:1.0
Data intake by: Seyedeh Marzieh Hosseini
Data intake reviewer: Bao Khanh Nguyen
Data storage location:

Tabular data details: https://github.com/Khanhbao8695/HealthCar_DS2021

Total number of observations	3424
Total number of files	1
Total number of features	69
Base format of the file	xlsx
Size of the data	898KB

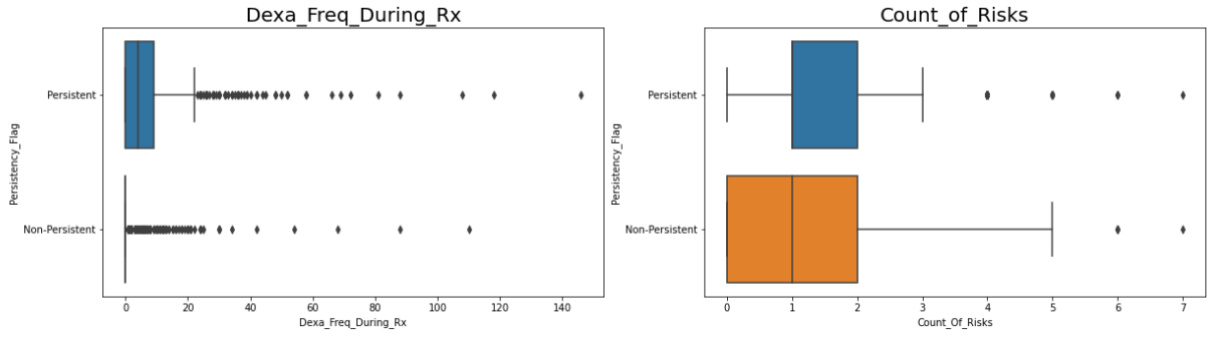
GitHub Repository:

Project Link: https://github.com/Khanhbao8695/HealthCar_DS2021

Data Types

There are 69 features in this dataset from more than 3424 inputs. The majority of the data types in this dataset are "object" types with more than 67 features and only 2 features are “inte64” data type.

Data Problems



For "Dexa Freq During Rx," a graph illustrates that this variable has a lot of skewness and Kurtosis (Platykurtic), which considers a lot of outliers. Furthermore, the data for "Count of

Risk" has a moderate skewness and is moderately kurtosis (Platykurtic), indicating that there are few outliers.

Data Transformation to resolve outliers

My first approach to deal with the skewness and outliers for these variables is using IQR Score. To remove outliers, this approach uses the IQR values calculated before. Anything outside of the range of (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR) is considered an outlier and should be eliminated.

For the Old Shape (3424,69) for both of Count of Risk" and “Dexa Freq During Rx" variable but after removed the outliers with this method, the new shape only (2964,69), which remove 460 data outside of the range of (Q1 - 1.5 IQR) and (Q3 + 1.5 IQR).

Data Transformation with non-number data

For Persistent Flag column, there are two types of data: Persistent and Non-Persistent, but if we just leave with object types will be difficult for later training models so I will assign value 0 and 1 to Non-Persistent and Persistent in Persistent Flag column. In addition, most of the value in this dataset have Y and N value, which also difficult to train model, so I will change Y and N value to 1 and 0 as well.

To do this we will use OneHotEncoder() provided by sklearn. Basically, it will transform a categorical column from this (example to describe this approach):

marital	housing
single	yes
divorced	no
married	no

...into something like this...

marital_single	marital_divorced	marital_married	housing_yes	housing_no
1	0	0	1	0
0	1	0	0	1
0	0	1	0	1

Here is the code:

```
encoder = OneHotEncoder(sparse=False)
cat_cols = ['Gender', 'Race', 'Ethnicity', 'Region','Ntm_Speciality' ,'Ntm_Speciality_Bucket', 'Age_Bucket',
            'Adherent_Flag', 'Change_Risk_Segment',
            'Change_T_Score','Tscore_Bucket_During_Rx', 'Risk_Segment_During_Rx','Tscore_Bucket_Prior_Ntm',
           'Risk_Segment_Prior_Ntm']

# Encode Categorical Data
df_encoded = pd.DataFrame(encoder.fit_transform(df[cat_cols]))
df_encoded.columns = encoder.get_feature_names(cat_cols)

# Replace Categotical Data with Encoded Data
df_health_ready = df.drop(cat_cols,axis=1)
df_health_ready = pd.concat([df_encoded, df_health_ready], axis=1)

# Encode target value
df_health_ready['Persistency_Flag'] = df_health_ready['Persistency_Flag'].apply(lambda x: 1 if x == 'Persistent' else 0)

print('Shape of dataframe:', df_health_ready.shape)
df_health_ready.head()
```

```
Shape of dataframe: (2942, 131)
```

And the Result will look like this

Out[44]:

	Gender_Female	Gender_Male	Race_African American	Race_Asian	Race_Caucasian	Race_Other/Unknown	Ethnicity_Hispanic	Ethnicity_Not Hispanic	Ethnicity_Unknown	Regic
0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	
1	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	
2	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	
3	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	
4	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	

Next Step is Applying Classification algorithms on the dataset