Julia Donato

Drug Persistency Project- Week 8: Data Problems

Individual Project- Julia Donato, julia.donato21@gmail.com, USA, Indiana University-Bloomington, Data Science

Problem Description:
Drug persistency is a challenge for pharmaceutical companies and understanding the factors that impact it can be difficult. My goal is to develop a machine learning model that can predict drug persistency based on physician prescription data.

Data Understanding:
Dataset: Healthcare_dataset.xlsx
Shape: (3424, 69)

Bucket	Column Name	Type	Information	Solutions
Unique Row Id	Patient ID	Object	Unique to each patient	
Target Variable	Persistency_Flag	Object	Variables: 'Persistent', 'Non-Persistent' NA: None	
Demographics	Gender	Object	Variables: 'Male', 'Female' NA: None	
	Race	Object	Variables: 'Caucasian', 'Asian', 'Other/Unknown', 'African American' NA: 2.83% 'Other/Unkown'	Mode imputation-common solution, ok to do since NaN such a small percent of data
	Ethnicity	Object	Variables: 'Not Hispanic', 'Hispanic', 'Unknown' NA: 2.66% 'Unkown'	Mode imputation-common solution, ok to do since NaN such a small percent of data
	Region	Object	Variables: 'West', 'Midwest', 'South', 'Other/Unknown', 'Northeast' NA: 1.75% 'Other/Unkown'	Mode imputation-common solution, ok to do since NaN such a small percent of data
	Age_Bucket	Object	Variables: '>75', '55-65', '65-75', '<55'	
	Idn_Indicator	Object	Values: 'N', 'Y' NA: None	

Provider Attributes	Ntm_Speciality	Object	Variables: 'GENERAL PRACTITIONER', 'Unknown', 'ENDOCRINOLOGY', 'RHEUMATOLOGY', 'ONCOLOGY', 'PATHOLOGY', […] 'TRANSPLANT SURGERY', 'PLASTIC SURGERY', 'CLINICAL NURSE SPECIALIST', 'OTOLARYNGOLOGY', 'HOSPITAL MEDICINE', 'ORTHOPEDICS', 'NEPHROLOGY', 'GERIATRIC MEDICINE', 'HOSPICE AND PALLIATIVE MEDICINE', 'OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY', 'VASCULAR SURGERY', 'CARDIOLOGY', 'NUCLEAR MEDICINE' NA: 9.05% Unkown	Mode imputation-common solution, ok to do since NaN such a small percent of data
	Ntm_Specialist_Flag	Object	Others', 'Specialist' NA: None	
	Ntm_Speciality_Bucket	Object	Values: 'OB/GYN/Others/PCP/Unknown', 'Endo/Onc/Uro', 'Rheum' NA: None	
Clinical Factors	Gluco_Record_Prior_Ntm	Object	Values: 'N', 'Y' NA: None	
	Gluco_Record_During_Rx	Object	Values: 'N', 'Y' NA: None	
	Dexa_Freq_During_Rx	int64	Values: [0, 2, 7, 3, 5, 20, 13, 1, 6, 12, 4, 10, 25, 11, 18, 21, 15, 28, 22, 37, 14, 8, 9, 17, 81, 42, 16, 30, 19, 45, 27, 24, 58, 26, 23, 33, 110, 36, 34, 88, 66, 32, 118, 48, 69, 38, 40,	

			68, 52, 50, 146, 44, 35, 39, 108, 54, 72, 29], NA: None	
	Dexa_During_Rx	Object	Values: 'N', 'Y' NA: None	
	Frag_Frac_Prior_Ntm	Object	Values: 'N', 'Y' NA: None	
	Frag_Frac_During_Rx	Object	Values: 'N', 'Y' NA: None	
	Risk_Segment_Prior_Ntm	Object	Values: 'VLR_LR', 'HR_VHR' NA: None	
	Tscore_Bucket_Prior_Ntm	Object	Values: '>-2.5', '<=-2.5' NA: None	
	Risk_Segment_During_Rx	Object	Values: 'VLR_LR', 'Unknown', 'HR_VHR' NA: 43.72% 'Unkown'	When Risk_Segment_During_Rx = 'Unkown', replace 'Unkown' with the value of Risk_Segment_Prior_Ntm
	Tscore_Bucket_During_Rx	Object	Values: '<=-2.5', 'Unknown', '>-2.5' NA: 43.72% 'Unkown'	When Tscore_Bucket_During_Rx = 'Unkown', replace 'Unkown' with the value of Tscore_Bucket_Prior_Ntm
	Change_T_Score	Object	Values: 'No change', 'Unknown', 'Worsened', 'Improved' NA: 43.72% 'Unkown'	The above actions will lead to the 'Unkown' values changing to 'No Change'
	Change_Risk_Segment	Object	Values: 'Unknown', 'No change', 'Worsened', 'Improved' NA: 65.1% 'Unkown'	The above actions will lead to the 'Unkown' values changing to 'No Change'
Disease/Treatment Factor	Injectable_Experience_During_Rx	Object	Values: 'Y', 'N' NA: None	
	NTM - Risk Factors	Object	Values: 'Y', 'N' NA: None	
	NTM - Comorbidity	Object	Values: 'Y', 'N' NA: None	
	NTM - Concomitancy	Object	Values: 'Y', 'N' NA: None	

	Adherent_Flag	Object	Values: 'Adherent', 'Non-Adherent' NA: None	

Alternatively method to try for Risk_Segment_During_Rx through Change_Risk_Segment is to delete this columns and see how this affects the data set. This is because the NaN or 'Unkown' variables make up such a large proportion of the data (>40%)

Github Repo Link:
https://github.com/julia-donato/DG-Final-Project