OpenML team

Name	Email	Country	College/Company	Specialization
Juan Carlos	juanca.gutierrez@outlook.com	Spain	everis	Data Science
Laith Adi	Laith_adi@hotmail.com	Canada	Laurier University	Data Science
Gerson Orihuela	yovanni.orihuela@gmail.com	Peru	Inspira IT	Data Science
Walquer Valles	wx.vr@outlook.com	Peru	KeepCoding	Data Science

PROBLEM DESCRIPTION:

We have been provided with a dataset containing 69 medical variables and we are going to look for null values, weird values, format errors, outliers, incorrect data types and similar problems that may have leaked into the dataset.

DATA UNDERSTANDING:

We have got 3424 data points and feature vectors have 69 variables. Our intention is to build a model that predicts if a given patient will persist on his/her treatment or not. Having this, our target is the “Persistency Flag” variable.

Besides individual identificators and the target variable, there are other four buckets:

- Demographics
- Provider attributes
- Clinical factors
- Disease and treatment factors

Bucket	Columns	Information
Demographics	Gender	Type: Object No missing values # of unique values: 2 Values: “Male”, “Female” Mode: Female (3230/3424 or 94.33%)
	Age_Bucket	Type: Object No missing values # of unique values: 4 Values: >75, 55-65, 65-75, <55 Mode: >75 (1439/3424 or 42.03%)
	Race	Type: Object Missing values: “Other/Unknown” (97/3424 or

		2.83%) # of unique values: 4 Values: [Caucasian, Asian, Other/Unknown, African American] Mode: “Caucasian” (3148/3424 or 91.94%)
	Region	Type: Object Missing values: “Other/Unknown” (60/3424 or 1.75%) # of unique values: 5 Values: West, Midwest, South, Other/Unknown, Northeast Mode: “Midwest” (1383/3424 or 40.39%)
	Ethnicity	Type: Object Missing values: “Unknown” (91/3424 or 2.66%) # of unique values: 3 Values: “Not Hispanic”, “Hispanic”, “Unknown” Mode: “Not Hispanic” (3235/3424 or 94.48%)
	Idn_Indicator	Type: Object No missing values # of unique values: 2 Values: “Y”, “N” Mode: “Y” (2557/3424 or 74.68%)
Provider Attributes	Ntm_Specialty	Type: Object Missing values: “Unknown” (310/3424 or 9.05%) # of unique values: 36 Values: 'GENERAL PRACTITIONER', 'Unknown', 'ENDOCRINOLOGY', 'RHEUMATOLOGY', 'ONCOLOGY', 'PATHOLOGY', 'OBSTETRICS AND GYNECOLOGY', 'PSYCHIATRY AND NEUROLOGY', 'ORTHOPEDIC SURGERY', 'PHYSICAL MEDICINE AND REHABILITATION', 'SURGERY AND SURGICAL SPECIALTIES', 'PEDIATRICS', 'PULMONARY MEDICINE', 'HEMATOLOGY & ONCOLOGY', 'UROLOGY', 'PAIN MEDICINE', 'NEUROLOGY', 'RADIOLOGY', 'GASTROENTEROLOGY', 'EMERGENCY MEDICINE', 'PODIATRY', 'OPHTHALMOLOGY', 'OCCUPATIONAL MEDICINE', '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' Mode: “General Practitioner” (1535/3424 or 44.83%)
Clinical factors	Ntm_Speciality	Type: Object % missing values: 9.05% as Unknown # of unique values: 36 Values: 'GENERAL PRACTITIONER', 'Unknown', 'ENDOCRINOLOGY', 'RHEUMATOLOGY', […],'VASCULAR SURGERY', 'CARDIOLOGY', 'NUCLEAR MEDICINE' Mode: GENERAL PRACTITIONER (1535/3424 or 44.83%)
	Ntm_Specialist_Flag	Type: Object % missing values: 0% # of unique values: 2 Values: 'Others', 'Specialist' Mode: Others (2013/3424 or 58.79%)
	Ntm_Speciality_Bucket	Type: Object % missing values: 0% # of unique values: 3 Values: 'OB/GYN/Others/PCP/Unknown', 'Endo/Onc/Uro', 'Rheum' Mode: OB/GYN/Others/PCP/Unknown (2104/3424 or 61.45%)
	Gluco_Record_Prior_Ntm	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y' Mode: N (2619/3424 or 76.49%)
	Gluco_Record_During_Rx	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y' Mode: N (2522/3424 or 73.66%)
	Dexa_Freq_During_Rx	Type: Int64 % missing values: 0% # of unique values: 58 Values info: Mean 3.01, std 8.13 min 0.00 25% 0.00 50% 0.00 75% 3.00 max 146.0 Mode: 0 (2488/3424 or 72.66%)

	Dexa_During_Rx	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y' Mode: N (2488/3424 or 72.66%)
	Frag_Frac_Prior_Ntm	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y' Mode: N (2872/3424 or 83.88%)
	Frag_Frac_During_Rx	Type: Object % missing values: 0% # of unique values: 2 Values: 'N', 'Y'
	Risk_Segment_Prior_Ntm	Type: Object % missing values: 0% # of unique values: 2 Values: 'VLR_LR', 'HR_VHR'
	Tscore_Bucket_Prior_Ntm	Type: Object % missing values: 0% # of unique values: 2 Values: '>-2.5', '<=-2.5'
	Risk_Segment_During_Rx	Type: Object % missing values: 43% as Unknown # of unique values: 3 Values: 'VLR_LR', 'Unknown', 'HR_VHR'
	Tscore_Bucket_During_Rx	Type: Object % missing values: 43% as Unknown # of unique values: 3 Values: <=-2.5', 'Unknown', '>-2.5'
	Change_T_Score	Type: Object % missing values: 43% as Unknown # of unique values: 4 Values:'No change', 'Unknown', 'Worsened', 'Improved'
	Change_Risk_Segment	Type: Object % missing values: 65% as Unknown # of unique values: 4 Values: 'Unknown', 'No change', 'Worsened', 'Improved'
Disease/treatment factors	NTM - Injectable Experience	Type: Object No null values # of unique values: 2 Values: “Y”, “N”
	NTM - Risk Factors	Type: Object

		No null values # of unique values: 2 Values: “Y”, “N”
	NTM - Comorbidity	Type: Object No null values # of unique values: 2 Values: “Y”, “N”
	NTM - Concomitancy	ype: Object No null values # of unique values: 2 Values: “Y”, “N”
	Adherence	Type: Integer No null values # of unique values: 8 Values: 0, 1, 2, 3, 4, 5, 6, 7

APPROACHES TO OVERCOME DATA ERRORS

Race variable - missing values

- use the mode as an imputer. Two reasons why:
    1. only 2.83% (97 instances out of 3424) are “Other/Unknown”. So, it feels safe to use the mode to fill in the values for now.
    2. The mode accounts for 91.94% of the data. And if we were to group the data by ethnicity, the mode accounts for 93.45% (3023 instances out of 3235) for “Not Hispanic” and 61.22% (60 instances out of 98) for “Hispanic”.
- For those reasons, it's safe to assume that it is likely that the “Other/Unknown” values can be treated as the mode.

Region variable - missing values

- use the Region mode for “Not Hispanic” Ethnicity group. Reasons:
    - 100% of “Other/Unknown” values in the Region variable, the instances Ethnicity falls under “Not Hispanic”

Ethnicity variable - missing values

- use the mode as an imputer. Reason:
    - the mode accounts for 94.48% (3235 instances out of 3424) of the values for Ethnicity.
    - There are only 2.66% of missing values so the number is not alarmingly large to reconsider what we use for the missing values. The mode should be safe/good enough.

Ntm_Speciality variable - missing values

- We will try two approaches: To keep unknowns as a category since it accounts for less than 10% of the data and see how it relates to other variables.

- An alternative approach will be to use the mode instead.

NTM - Injectable Experience, Risk Factors, Comorbidity and Concomitancy (group of variables) - handling categorical data

- “Y” will be replaced with 1 and “N” with 0

Risk_Segment_During_Rx, Tscore_Bucket_During_Rx, Change_T_Score and Change_Risk_Segment missing values

- These variables have more that 40% missing values, consequently they’ll be eliminated

GITHUB REPO: https://github.com/jaycee-ds/Drug_Persistency_ABC_Pharma