Group Name: **Attack on Data**

Team Members:

Name	Email	Country	College/Company	Specialization
Armin Khayati	Northatlas@gmail.com	UAE	-	Data Science
Ezzuldin Zaky	Ezzulding.zaky@gmail.com	UAE	American University of Sharjah	Data Science
Orcun Sami Tandogan	tandogan.orcun@metu.edu.tr	Turkey	Middle East Technical University	Data Science

Problem description

One of the challenges 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.

Data cleansing and transformation on data

- Handling **Unknown** values for Race, Region, and Ethnicity Variables
    - Using mode as an imputer as an imputer on ***Race*** and ***Ethnicity*** variables.
    - For ***Region*** variable, because most of the people with **Unknown** Region have **Not Hispanic** Ethnicity, and Most of people with Not Hispanic Ethnicity, have Midwest Region, we will replace Unknown Regions with **Midwest**.

- Handling **Rare Labels**: Finding categories less than 5 percent in each variable, then merging those categories into one or drop them if the variable only has 2 categories (e.g., Y/N) and cardinality of one them is less than 5 percent.
- Grouping integer values of Count_Of_Risks variable into two **bins**.

Group	Variable	Categories to Merge
Variables Chosen to Merge Categories	Race	African American, Asian
	Age_Bucket	<55, 55-65
	Ntm_Speciality	OBSTETRICS AND GYNECOLOGY , UROLOGY , ORTHOPEDIC SURGERY , CARDIOLOGY , PATHOLOGY , HEMATOLOGY & ONCOLOGY , OTOLARYNGOLOGY , PEDIATRICS , PHYSICAL MEDICINE AND REHABILITATION , PULMONARY MEDICINE , SURGERY AND SURGICAL SPECIALTIES , PSYCHIATRY AND NEUROLOGY , NEPHROLOGY , ORTHOPEDICS , PLASTIC SURGERY , VASCULAR SURGERY , HOSPICE AND PALLIATIVE MEDICINE ,

		GERIATRIC MEDICINE , GASTROENTEROLOGY , TRANSPLANT SURGERY , CLINICAL NURSE SPECIALIST , OCCUPATIONAL MEDICINE , HOSPITAL MEDICINE , OPHTHALMOLOGY , PODIATRY , EMERGENCY MEDICINE , RADIOLOGY , OBSTETRICS & OBSTETRICS & GYNECOLOGY & OBSTETRICS & GYNECOLOGY , NEUROLOGY , PAIN MEDICINE , NUCLEAR MEDICINE
	Change_T_Score	Improved, Worsened
	Change_Risk_Segment	Improved, Worsened
	Count_Of_Risks	Bin 1 is [0,1,2,3] and bin 2 is [4,5,6,7]
Variables Chosen to Drop	Ethnicity	
	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_Chronic_Liver_Disease	
	Risk_Low_Calcium_Intake	
	Risk_Excessive_Thinness	
	Risk_Hysterectomy_Oophorectomy	

	Risk_Estrogen_Deficiency
	Risk_Immobilization
	Risk_Recurring_Falls
	Dexa_Freq_During_Rx

- **One hot encoding** all the variables after doing above tasks

Code Review

Code By	Review By
Armin Khayati	Ezzuldin Zaky
Ezzuldin Zaky	Orcun Sami Tandogan
Orcun Sami Tandogan	Armin Khayati

Github Repo:

https://github.com/Arminkhayati/dataglacier_internship