Week 10 Project Deliverables

Group Name: Decent Healthcare Analytics

Name: Tolga Yaz

E-Mail: tolgayaz1991@gmail.com

Country: Turkey

Specialization: Data Science

Problem Description:

- One of the challenges for pharmaceutical companies is to understand the persistency of drugs as per the physician prescription.
- To solve this problem a pharma company wants to automate this process of identification via using analytics.
- **Main Goal:** Predicting drug persistency of patients to provide better healthcare.
- **Main Objective:** Building a classification model to predict drug persistency of patients via using provided data and data science methods.

Github Repo Link: https://github.com/tolgayaz1991/DgInternshipProjects

Exploratory Data Analysis

About the Analysis

- The data provided are mostly composed of categorical features (67 of 69 features are categorical with data type of object). And only 2 features have data type of int64.
- The missing data were dealt last week.
- After some simple exploration of the data as seen in the ipynb file, the analysis was concluded with correlation analysis.
- For correlation analysis, categorical to numerical transformation was applied when needed.
- Then, the features which do not affect the result of the target variable were found and dropped.
- The 28 features that were dropped because their lack of efficiency to the target result are:

['Gender', 'Ethnicity', 'Ntm_Specialist_Flag', 'Gluco_Record_Prior_Ntm', 'Frag_Frac_Prior_Ntm', 'Risk_Segment_Prior_Ntm', 'Tscore_Bucket_Prior_Ntm', 'Idn_Indicator', 'Injectable_Experience_During_Rx', 'Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias', 'Comorb_Osteoporosis_without_current_pathological_fracture', 'Concom_Cholesterol_And_Triglyceride_Regulating_Preparations', 'Concom_Anti_Depressants_And_Mood_Stabilisers',

'Risk_Type_1_Insulin_Dependent_Diabetes','Risk_Untreated_Early_Menopause', 'Risk_Patient_Parent_Fractured_Their_Hip', 'Risk_Smoking_Tobacco', 'Risk_Chronic_Malnutrition_Or_Malabsorption', '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_Recurring_Falls', 'Risk_Untreated_Chronic_Hyperthyroidism', 'Risk_Immobilization']

- **Below is a screenshot from correlation analysis (the values are pearson correlations):**

	Persistency_Flag
Persistency_Flag	1.000000
Dexa_Freq_During_Rx	0.330503
Count_Of_Risks	0.081699
Gluco_Record_During_Rx_code	0.213078
Dexa_During_Rx_code	0.495773
Frag_Frac_During_Rx_code	0.113836
Comorb_Encounter_For_Screening_For_Malignant_Neoplasms_code	0.317499
Comorb_Encounter_For_Immunization_code	0.318606
Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx_code	0.287595
Comorb_Vitamin_D_Deficiency_code	0.178660
Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified_code	0.233583
Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx_code	0.210992
Comorb_Long_Term_Current_Drug_Therapy_code	0.358023
Comorb_Dorsalgia_code	0.221552
Comorb_Personal_History_Of_Other_Diseases_And_Conditions_code	0.224808
Comorb_Other_Disorders_Of_Bone_Density_And_Structure_code	0.252569
Comorb_Personal_history_of_malignant_neoplasm_code	0.173350
Comorb_Gastro_esophageal_reflux_disease_code	0.226601
Concom_Narcotics_code	0.200429
Concom_Systemic_Corticosteroids_Plain_code	0.244364
Concom_Fluoroquinolones_code	0.189717
Concom_Cephalosporins_code	0.221705
Concom_Macrolides_And_Similar_Types_code	0.217520
Concom_Broad_Spectrum_Penicillins_code	0.194313

- **The shape of the data after eda:**

(3202, 34) with a target variable (Persistency_Flag) and 33 predictors