WEEK 10 Deliverables
Data Glacier Team: The Data Dynamos
Batch Code: LISUM04
Remote Internship Cohort: September to December 2021
This report is submitted as a text link to Canvas

1. Team member's details:
   Jennifer Turley
   Jennifer.turley@ucdconnect.ie
   Ph +353 83 065 7252
   University College Dublin, Ireland
   Data Science

   Queen Echerenwa
   Quin_codes@outlooks.com
   +234 806 352 8847
   Data Science

   Ateeb Aqil Aziz
   Atteeb.aqil@gmail.com
   Ph +92 3044693262
   Data Science

Github repo link:

The code and corresponding Jupyter notebook printout PDF can be found at
https://github.com/DataDynamosTeam/DrugPersistencyProject/tree/JTs-branch

Note the code is identified with leading characters W10

Problem Description:

Drug persistency with a medical treatment is the continuation by the patient of a treatment prescribed by a medical provider. When a patient fails to be *persistent* (by quitting the treatment prematurely) this can be expected to have a negative impact on the medical outcome for that patient. This project examines drug persistency for a treatment for improving bone density, with the goal to understanding and clarifying variables that have an impact on persistency.

Utilizing a dataset of 3434 observations of primarily categorical data, the data is first cleaned and transformed to numerical formats. Then classification models from Scikit-learn, Imbalanced-learn, and Statsmodels are applied in an effort to predict drug persistency from the other variables provided.

Exploratory Data Analysis (EDA) Performed on the Data:

The data cleansing and transformation accomplished in Weeks 8 and 9 included correcting some miscoded names (cleansing) and conversion of categorical values to numerical formats (transformation). Some new variables were created from old variables. Bar charts of individual variables against the Persistency_Flag variable were also created. These bar charts visually demonstrated which independent variables might contribute to the prediction of the persistency. These charts are included in the Week 8 .ipynb Jupyter notebook.

Preliminary models were run which in combination with the bar charts permitted the creation of a reduced list of features of interest for predicting persistency.

This reduced list is as follows:

```
features = ['Gluco_Record_During_Rx', 'Dexa_During_Rx', 'Frag_Frac_During_Rx', 'Risk_Segment_During_Rx', 'Adherent_Flag', 'Idn_Indicator',
           'Injectable_Experience_During_Rx', 'Comorb_Encounter_For_Screening_For_Malignant_Neoplasms', 'Comorb_Encounter_For_Immunization',
           'Comorb_Encntr_For_General_Exam_W_O_Complaint_Susp_Or_Reprtd_Dx', 'Comorb_Vitamin_D_Deficiency',
           'Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified', 'Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx',
           'Comorb_Long_Term_Current_Drug_Therapy', 'Comorb_Dorsalgia', 'Comorb_Personal_History_Of_Other_Diseases_And_Conditions',
           'Comorb_Other_Disorders_Of_Bone_Density_And_Structure', 'Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias',
           'Comorb_Osteoporosis_without_current_pathological_fracture', 'Comorb_Personal_history_of_malignant_neoplasm',
           'Comorb_Gastro_esophageal_reflux_disease', 'Concom_Cholesterol_And_Triglyceride_Regulating_Preparations', 'Concom_Narcotics',
           'Concom_Systemic_Corticosteroids_Plain', 'Concom_Anti_Depressants_And_Mood_Stabilisers', 'Concom_Fluoroquinolones',
           'Concom_Cephalosporins', 'Concom_Macrolides_And_Similar_Types', 'Concom_Broad_Spectrum_Penicillins', 'Concom_Anaesthetics_General',
           'Concom_Viral_Vaccines', 'Risk_Rheumatoid_Arthritis', 'Risk_Untreated_Chronic_Hyperthyroidism', 'Risk_Untreated_Chronic_Hypogonadism',
           'Risk_Smoking_Tobacco', 'Risk_Chronic_Malnutrition_Or_Malabsorption', 'Risk_Chronic_Liver_Disease', 'Risk_Low_Calcium_Intake',
           'Risk_Vitamin_D_Insufficiency', 'Risk_Poor_Health_Frailty', 'Risk_Excessive_Thinness', 'Risk_Estrogen_Deficiency', 'Risk_Immobilization',
           'Dexa_Freq_During_Rx_Bucket_Flag','Change_RiskSeg_Worsened','Change_RiskSeg_Improved', 'Change_RiskSeg_Unk' , 'ChangedTScore_Worsened',
           'ChangedTScore_Improved', 'ChangedTScore_Unk', 'AsianRace_Flag',  'Midwest_Flag', 'Ntm_Speciality_MyBuckets1', 'Ntm_Speciality_MyBuckets2' ]
```

Since using predictor variables that are themselves correlated (with each other) is a potential problem when applying models that assume these predictors are linearly independent (such as in logistic regression), it was important to examine potential correlations between the proposed predictor variables. One correlation table was produced for all the independent variables: this allowed for checking the level of correlation between all of the independent variables. Since many variables were already eliminated in Week 9, including those with different codings of the same information (such as for speciality of the prescribing doctor – given individually and also in buckets), there was no excessive correlation between the remaining list of features to be used as predictors.

Next, correlation heatmaps were created, with the features list split into three smaller sets for ease of viewing the resulting charts (to prevent excessively small squares). The Persistency_Flag variable was included in the heatmaps since it is of value to visualize its correlation which the predictor variables, although the correlation has already been assessed numerically in Week 9.

The heatmaps are shown below:

correlation heatmap, subset 1 of features

correlation heatmap, subset 2 of features
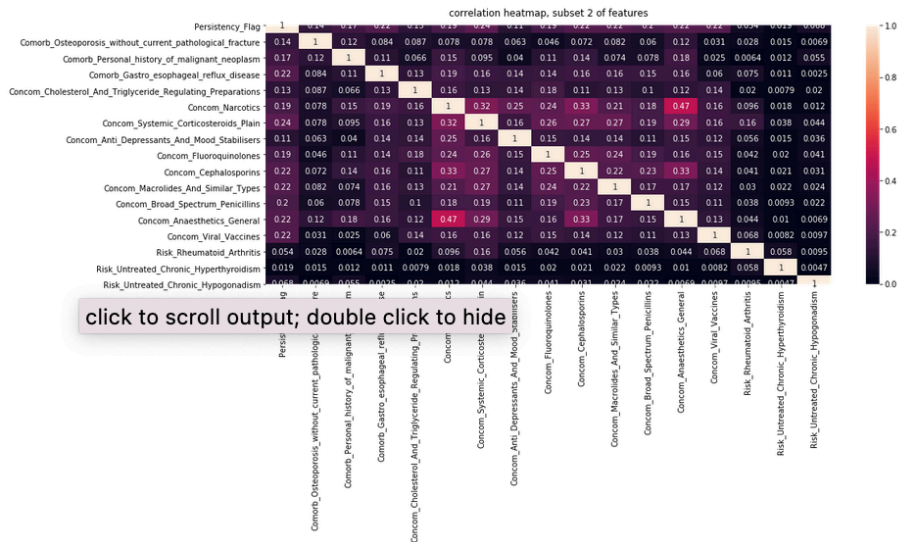
click to scroll output; double click to hide

correlation heatmap, subset 3 of features

Final Recommendation:

Several models performed comparably with classification accuracy around 80%. Some had lower accuracy but higher ROC AUC. However, toward the main goal of gaining insight into features influencing persistency, the logistic regression and random forest models were the most helpful. Further information on these results will be provided in the following weeks' deliverables.