Data Glacier

Your Deep Learning Partner

Final Report

Healthcare - Persistency of a drug

Team: DG_team_project_PL-RO-KSA-EGY

Date: May 15, 2021

Agenda

1. Project GitHub repository
2. Project team and schedule
3. Problem statement
4. Business understanding and scope identification
5. Data understanding
6. EDA and feature engineering
7. Model building and model selection
8. Model evaluation
9. Model deployment

10.Model governance (food-for-thoughts)

Data Glacier

Your Deep Learning Partner

Project GitHub repository

https://github.com/Omar-Safwat/HealthCare_project

|| *(„Findings” folders – work of team members, „Week” folders – combined works)*

|-- Afshan-Findings

|-- Larisa-Findings

|-- Omar-Findings

|-- Roger-Findings

|-- Week_7

|-- Week_8

|-- Week_9

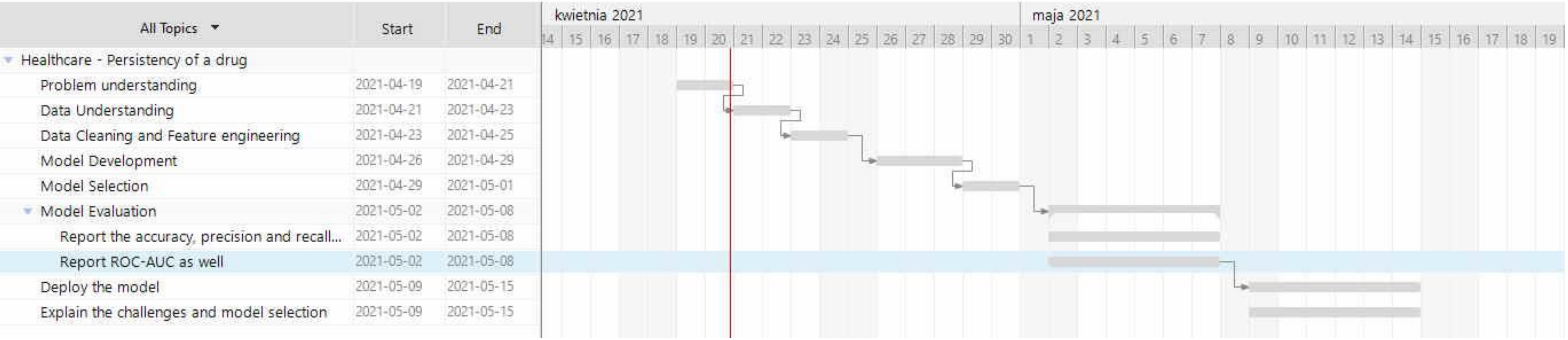
|-- Week_10-11

|-- Week_12

|-- Week_13

Project team and schedule

Name	Specialization	Country	Email
Ms. Larisa Popa	Data Science	Romania	Larisapopa4@gmail.com
Ms. Afshan Hashmi	Data Science	Kingdom of Saudi Arabia	afshanhashmi786@gmail.com
Mr. Omar Safwat	Data Science	Egypt	omarksafwat@gmail.com
Mr. Roger Burek-Bors	Data Science	Poland	roger.burek-bors@hotmail.com



All Topics	Start	End
Healthcare - Persistency of a drug		
Problem understanding	2021-04-19	2021-04-21
Data Understanding	2021-04-21	2021-04-23
Data Cleaning and Feature engineering	2021-04-23	2021-04-25
Model Development	2021-04-26	2021-04-29
Model Selection	2021-04-29	2021-05-01
Model Evaluation	2021-05-02	2021-05-08
Report the accuracy, precision and recall...	2021-05-02	2021-05-08
Report ROC-AUC as well	2021-05-02	2021-05-08
Deploy the model	2021-05-09	2021-05-15
Explain the challenges and model selection	2021-05-09	2021-05-15

Problem Statement

One of the challenge 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.

Objective:

With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset.

Business understanding and scope identification (Week_7)

1. Patients around the world are suffering from mycobacteria (NTM). The most common clinical manifestation of NTM disease is lung disease.
2. Pulmonary NTM disease diagnosis requires both identification of the mycobacterium in the patient's lungs, as well as a high-resolution CT scan of the lungs.
3. Failure to adhere to prescribed-medication regimens is one of the principal reasons patients don’t achieve the expected outcomes from their treatment. Solving this challenge has been a major goal for pharmaceutical and healthcare organizations for decades.
4. An estimated 125,000 lives are lost annually in the United States and additional healthcare expenditures of 290 billion USD are driven bynonadherence.

Data understanding (Week_8)

1. The Pharmaceutical company provided dataset called in xlsx format.
2. 69 features were discovered, among them 1 label required for classifier to rely on (Persistancy_Flag).
3. 3424 data entries were discovered.
4. 99% features were provided as categorical data and needed to be turned into numerical for ML model.
5. Missing values were present within 7 features.
6. Outliers were present within 23 features.
7. Skewed data was not present.

Data cleansing and transformation (Week_9)

1. Missing data in total: 7278 values which means 3%.

Approaches used:

- prediction using regression technique
- imputation using KNN technique

2. Outliers in total: in 23 of 69 features which means 33%.

Approaches used:

- deletion of feature
- deletion of row containing outlier

3. Outcome:

- EDA dataset: cleaned_data.csv with categorical values
- Training dataset: data_training.csv with numerical values

EDA and feature engineering (Week_10-11)

8 hypothesis were constructed to gain knowledge on the subject and formulate final recommendation on features and model selection. Main topics included:

- Importance of CT scan
- Risks impact
- Demographics
- Effect of provider attributes
- Effect of glucose
- Side effects
- Most frequent risks
- Comorbidities and persistence

Model building and model selection (Week_12)

1. Following models were developed by team members:
    - Decision Trees
    - K-Nearest Neighbors (KNN)
    - Logistic Regression
    - Support Vector Machines (SVM)

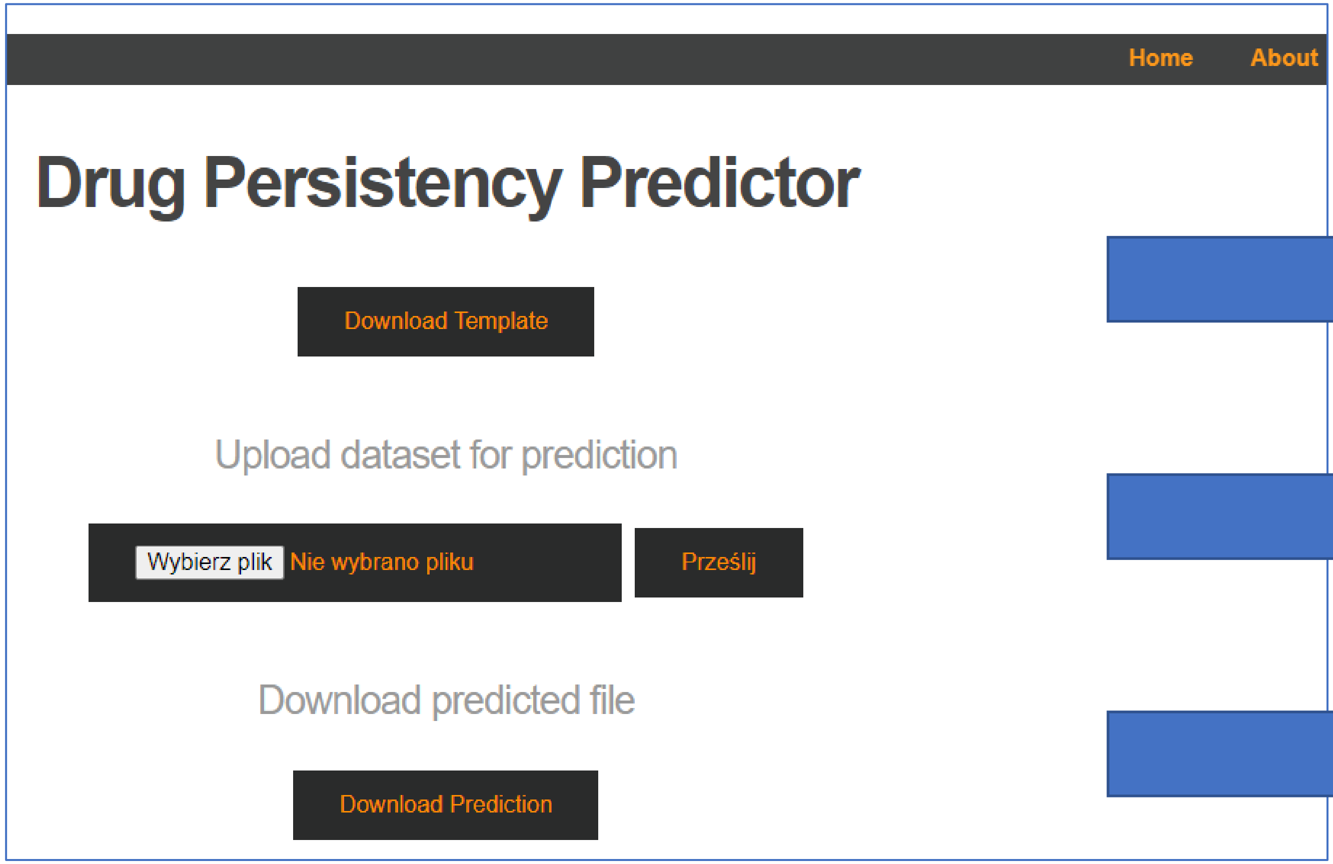
2. Following evaluation metrics were used to assess models:
    - Accuracy
    - Precision and Recall
    - F1/F2 Score
    - Lift and Gain
    - KS statistics
    - AUC-ROC

Model evaluation (Week_13)

		Decision Trees	KNN	Logistic Regression	SVM
Accuracy (highest)		82%	79%	84%	84%
Precision	O (False)	0.76	0.53	0.86	0.85
	1 (True)	0.84		0.80	0.77
F1-score	O (False)	0.74	0.61	0.87	0.90
	1 (True)	0.86		0.78	0.53
ROC AUC score		-	-	0.90	0.83

Model deployment (Week_13)

Following cloud solution was implemented on Heroku



Drug Persistency Predictor

Upload dataset for prediction

Download predicted file

As data needs to be in specific format user can download template.

User provides filled template to predictor.

After ML computation user can download file with prediction.

https://healthcare-project-plroksaegy.herokuapp.com/

Model governance (food-for-thoughts)

1. Training should be provided to the organization on how to use the ML model.
2. Each submittal of data on webapp should contribute to expansion of data set.
3. Quarterly new model will be trained based on expanded dataset.

Thank You