Data Glacier

Your Deep Learning Partner

Final Project Report

Project Name: Healthcare – Persistency of a Drug

Group Name: Decent Healthcare Analysis

Name: Tolga Yaz

E-Mail: tolgayaz1991@gmail.com

Country: Turkey

Specialization: Data Science

Background – Persistency of a Drug

- **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.**

- **The planned project lifecycle:**

1. Problem Description, Business Understanding and Data Intake Report Preparation
2. Data Understanding and Exploratory Data Analysis
3. Eda Presentation and Proposal of Modeling Technique(s)
4. Model Selection and Model Building
5. Evaluation of Results and Final Project Report

Basics of the Examined Data and Exploratory Data Analysis Approach

- **1 data file in .xlsx format with 2 sheets were obtained:**

1. **Healthcare_dataset.xlsx**: Includes 2 sheets:

a. **Feature Description:** Includes descriptions of main features provided.

b. **Dataset :** Includes the data provided.

- **Characteristics of the data provided:**
- 3424 Observations,
- 69 Features.

- **Exploration Approach:**

1. The data was transformed into Python Pandas Dataframe to be processed.
2. The dataframe were observed carefully to identify and deal with the nulls and duplicates.
3. Analysis were done upon the dataframe file via usage of Python Libraries and Jupyter Notebooks.

Exploratory Data Analysis Summary (Dealing with Missing Values)

- 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.

- When the data were examined, no dominant outlier problem was seen.

- When the data were examined, it was seen that there are some values named “Unknown”.

- The “Unknown” values were changed into NaN values to easily examine them via Pandas.

- Since we have many columns and the data are sensitive, we dropped the columns with >200 missing values and got a new dataframe without these missing values. For the "Ethnicity" column, we used mode value to fill missing values.

- The shape of the data after exploratory data analysis:
(3202, 34) with a target variable (Persistency_Flag) and 33 predictors

- **The shape of the data before exploratory data analysis: (3424, 69)**
- The shape of the data before exploratory data analysis: (3202, 34)

Model Selection and Model Building (Random Forest and Logistic Regression)

- Since the data is tabular and the target value with most of the other features are categorical, the recommended ml model was "Random Forest".

- As a comparison model, Logistic Regression model was also built and tried.

- The performance of Random Forest Model was better than Logistic Regression after hyperparameter tuning.

- While deciding the success of the models, we examined the outputs below:

1. Confusion Matrix
2. Precision
3. Recall
4. F1 Score
5. Accuracy

Confusion Matrices

For Random Forest	Actual Positive	Actual Negative
Predicted Positive	528	67
Predicted Negative	120	246

For Logistic Regression	Actual Positive	Actual Negative
Predicted Positive	515	80
Predicted Negative	114	252

Precision, Recall, f1 Score and Accuracy

	Precision	Recall	F1 Score	Accuracy
Random Forest	0.80	0.78	0.79	0.81
Logistic Regression	0.79	0.78	0.78	0.80

As seen Random Forest is more successful than Logistic Regression.

Thanks