Drug Persistency Project

Group Name: Cool Data Scientists Team

Group Members: Jamila Hamdi, H. Melis Tekin Akcin, Yousef Elbayoumi, Mukhammadjon Kholmirzev

Date: 15-Aug-2021

Agenda

- Business problem
- EDA recommendation
- Model building
- Model selection
- Performance metrics
- Final recommendation

Business problem

Data Glacier
Your Deep Learning Partner

- 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.

➔To have a better business understanding, our aim is to:

- ✓ Understand our dataset deeply
- ✓ Examine the features to gather some more knowledge
- ✓ Search for the relationships between the features and their effects on our target variable.
- ✓ Look at the demographics, clinical factors, provider attributes and disease/treatment factors.

EDA recommendation

Based on the Exploratory Data Analysis done on the dataset, following recommendations are given to the ABC pharma company :

- ❑ According to Cleaning: The data is considered clean pretty much.
- ❑ According to Region: The data mostly as “Not Persistent”.
- ❑ According to Correlations: The data doesn’t have good correlations due to encoding the data, we replaced Y and N with 0 and 1.
- ❑ According to Statistics: Similar to the correlations, we can’t comment much here due to the same reason.
- ❑ Obviously, this is a classification problem, the team is considering several ML models to build and test, such as KNN, MLP, Decision Tree, Random Forest, etc.

Model building

Data Glacier
Your Deep Learning Partner

❑ In the model building stage, various **regression techniques** were used to classify Drug Persistency of subjects based on the predictor variables.

❑ After preparing the features of the machine learning model to predict data by transforming categorical data into numbers, now it is the turn of the model building step.

❑ To do this, we split the dataset into train and test set using **Scikit-learn library** into 80% for training data and the rest (20%) for the testing, with training samples shape is (2739, 64) and the testing shape is (685, 64).

❑ We used seven different regression techniques:

- **Support Vector Machines (SVM)**
- **K-Nearest Neighbor (KNN)**
- **Logistic Regression**
- **Multi Layer Perceptron (MLP) Classifier**
- **Decision Tree**
- **Gradient Boosting**
- **Random Forest**

Model selection

Data Glacier
Your Deep Learning Partner

✓ In order to select best model from seven models, we looked at the performance metrics used which are: **Accuracy, Precision, Recall, f1-Score, Support and AUC**. The results for all ML classifiers are shown below :

- **Multi Layer Perceptron (MLP)**

	precision	recall	f1-score	support
0	0.83	0.80	0.81	431
1	0.68	0.72	0.70	254
accuracy			0.77	685
macro avg	0.75	0.76	0.76	685
weighted avg	0.77	0.77	0.77	685

- **Random Forest**

	precision	recall	f1-score	support
0	0.83	0.87	0.85	431
1	0.76	0.69	0.72	254
accuracy			0.80	685
macro avg	0.79	0.78	0.79	685
weighted avg	0.80	0.80	0.80	685

Model selection

Data Glacier
Your Deep Learning Partner

- **Support Vector Machines (SVM)**

	precision	recall	f1-score	support
0	0.63	1.00	0.77	431
1	0.00	0.00	0.00	254
accuracy			0.63	685
macro avg	0.31	0.50	0.39	685
weighted avg	0.40	0.63	0.49	685

- **K-Nearest Neighbor (KNN)**

	precision	recall	f1-score	support
0	0.71	0.76	0.73	431
1	0.53	0.47	0.50	254
accuracy			0.65	685
macro avg	0.62	0.61	0.62	685
weighted avg	0.64	0.65	0.65	685

- **Logistic Regression**

	precision	recall	f1-score	support
0	0.82	0.86	0.84	431
1	0.74	0.69	0.71	254
accuracy			0.79	685
macro avg	0.78	0.77	0.78	685
weighted avg	0.79	0.79	0.79	685

Model selection

Data Glacier
Your Deep Learning Partner

- **Decision Tree**

	precision	recall	f1-score	support
0	0.76	0.73	0.74	431
1	0.57	0.62	0.59	254
accuracy			0.69	685
macro avg	0.67	0.67	0.67	685
weighted avg	0.69	0.69	0.69	685

- **Gradient Boosting**

	precision	recall	f1-score	support
0	0.83	0.86	0.85	431
1	0.75	0.71	0.73	254
accuracy			0.80	685
macro avg	0.79	0.78	0.79	685
weighted avg	0.80	0.80	0.80	685

Model selection

Data Glacier
Your Deep Learning Partner

Based on the previous analysis we recommend using two top model which are almost close in term of Accuracy to solve this problem:

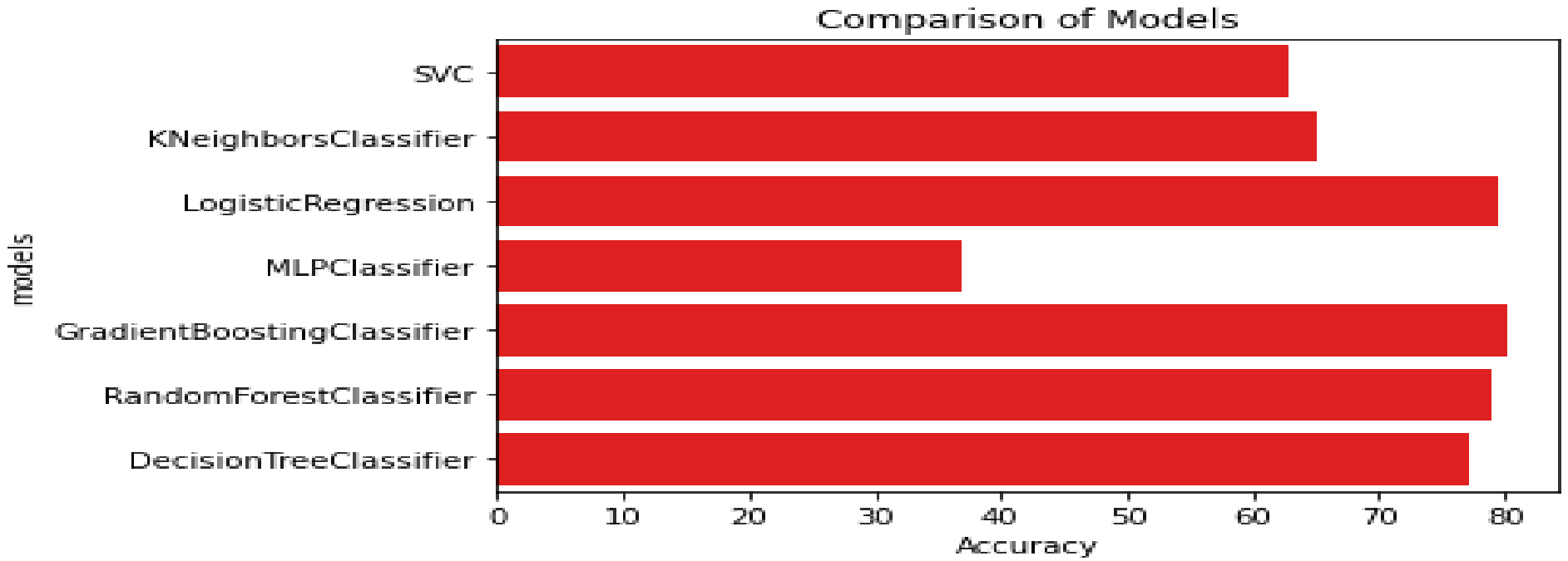
- **Gradient Boosting**
- **Logistic Regression**

The best model that detect Patient Persistency, is **Gradient Boosting** with an Accuracy of **80%**.

Model selection

Data Glacier
Your Deep Learning Partner

The picture below show the comparison of seven models used with their Accuracy result. It shows that Logistic Regression and Gradient Boosting classifiers are almost close in term of Accuracy.



Performance metrics

Data Glacier
Your Deep Learning Partner

❑ The evaluation of the Machine Learning techniques models was performed by creating the Confusion Matrix, and , then, measuring the Accuracy, Precision, Recall, and f1-Score from the Confusion Matrix.

❑ The parameters of the Confusion Matrix are shown below:

- **TP** = True Positive
- **TN** = True Negative
- **FP** = False Positive
- **FN** = False Negative

Performance metrics

Data Glacier
Your Deep Learning Partner

❑The formula of metrics used are explained below:

- **Accuracy** = TP+TN/TP+FP+FN+TN
- **Precision** = TP/TP+FP
- **Recall** = TP/TP+FN
- **F1 Score** = 2*(Recall * Precision) / (Recall + Precision)

❑ AUC (**Area Under The Curve**) ROC (**Receiver Operating Characteristics**) curve is one of the most important evaluation metrics for checking any classification model's performance at various threshold settings.

Performance metrics

Data Glacier
Your Deep Learning Partner

❑ ROC tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

❑ This curve plots two parameters:

- **True Positive Rate** (**TPR**) (also called recall) and is as follows:

TPR=TPTP+FN

- **False Positive Rate** (**FPR**) is defined as follows: FPR=FPFP+TN

Final recommendation

Data Glacier

Your Deep Learning Partner

- We have prepare the dataset for the classification problem. We set X by dropping the target variable "Persistency_Flag" and we set y as the target variable. Then we have used "Ordinal" and "Label" Encoder to encode the variables.
- As usual, we split the dataset as train set and test set.
- We have applied Logistic Regression, KNN, Random Forest Model, Neural Network, Gradient Boosting Model, Support Vector Machines and Classification Trees Models.

Final recommendation

Data Glacier

Your Deep Learning Partner

❑As seen in the above, we have compared their accuracy scores and we obtained the following:

✓ Gradient Boosting Model is the best fit model to our dataset with accuracy score 0.80.

✓We can also apply Random Forest Model with accuracy score 0.79

✓We can use Logistic Regression model with accuracy score 0.79 and cross validated score with 10 splits 0.78.

Thank You