Model Recommendation and Deployment

Project: Persistency of a Drug-Gathering insights for a Pharma company
Name: Shruthi Madgi
Email:Shruthi.madgi05@gmail.com
Country: USA
University: Visvesvaraya Technological University
Specialization: Data Science
Batch no:LISUM17
Reviewer: Data Glacier

Agenda

Problem Statement:

One of the challenge for all Pharmaceutical companies is to understand the persistency of drug as per the physician prescription whether a patient will be persistent in completing his/her dose. To solve this problem, ABC pharma company has approached an analytics company to automate the process of identification.

Objective is to build a classification model for drug persistency identification and to deploy the model on the cloud.

Data and Approach

The dataset is in .xlsx format and has 3424 points and feature vectors have 69 variables. It has 2 discrete value features/columns,1 continuous value feature and 66 categorical value columns/features including the target which is the Persistency flag.

The goal of using this dataset is to predict whether the patient will be persistent in completing his/her treatment.

Apart from individual identificators and the target variables, there are four other buckets:

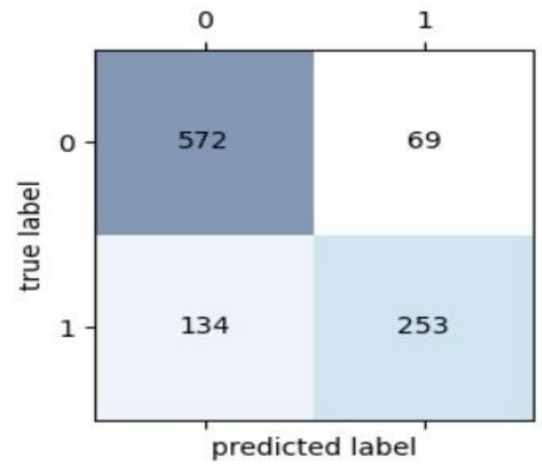
●Demographics
●Provider attributes
●Clinical factors
●Disease and treatment factors

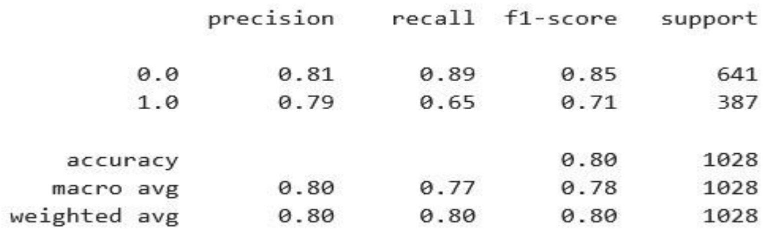
Data provided was cleaned of outliers, missing values and skewed data and was converted into a csv file.

Four models have been implemented on the basis of EDA. One among them will been chosen based on the f1 score and ROC _AUC curve.

Model recommendations

1.Decision Tree algorithm: Decision trees are a type of machine learning algorithm that work by recursively splitting the dataset based on the values of the input features, with the goal of creating subgroups that are more homogeneous in terms of the target variable (in this case, drug persistence). The max depth parameter determines the maximum number of levels or splits that the tree can have. In this case, the best results were obtained with a max depth of 1, which suggests that one of the input features had a strong predictive power in determining drug persistence. One advantage of decision trees is their interpretability, as they allow for visual inspection of the decision-making process. The Decision Tree gives an accuracy of 80%.

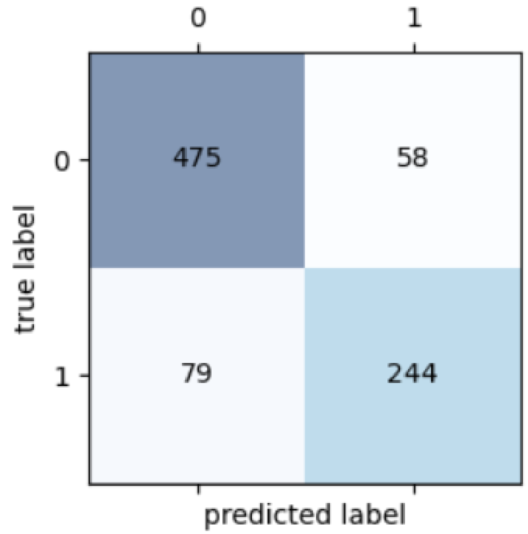


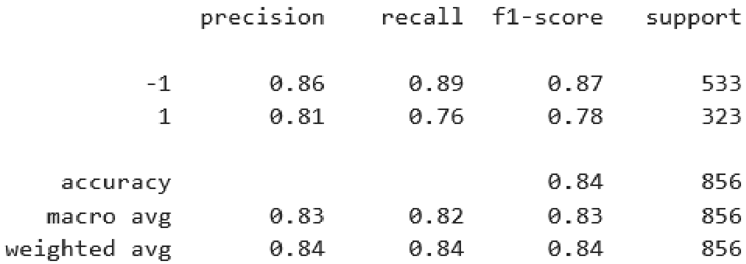


```
              precision    recall  f1-score   support

         0.0       0.81      0.89      0.85       641
         1.0       0.79      0.65      0.71       387

    accuracy                           0.80      1028
   macro avg       0.80      0.77      0.78      1028
weighted avg       0.80      0.80      0.80      1028
```

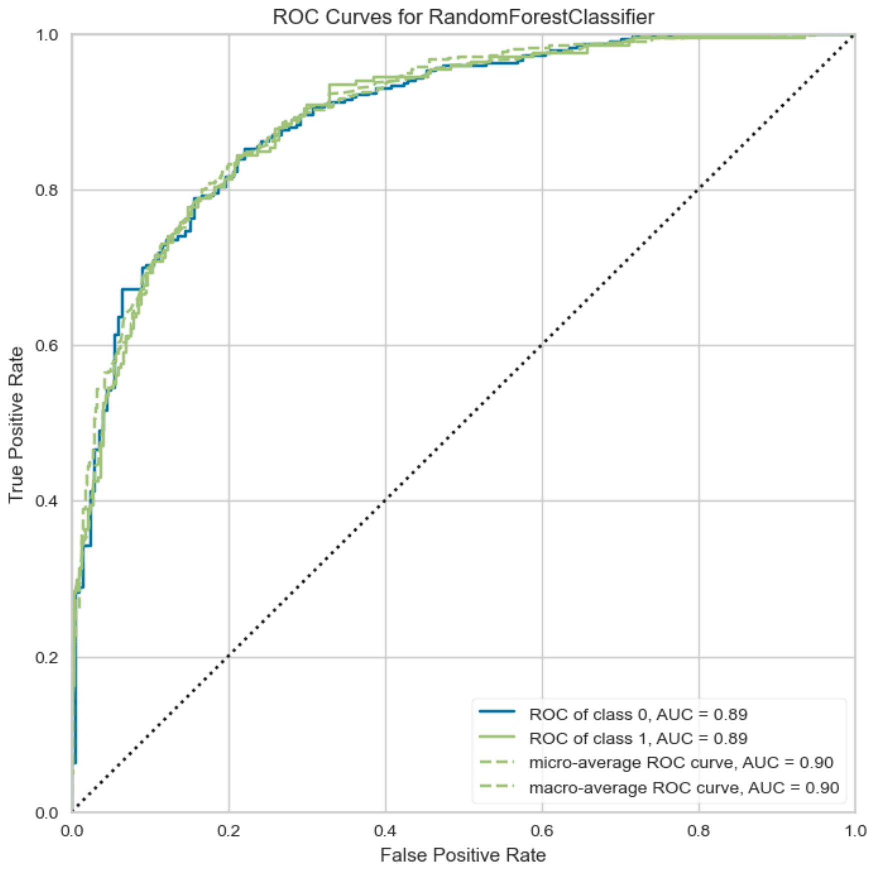
2.Support Vector Machines algorithm: Support Vector Machines (SVM) is a popular machine learning algorithm that can be used for classification and regression problems. The linear kernel is one type of SVM kernel function that works by finding the hyperplane that best separates the data into two classes (in this case, positive and negative outcomes). The accuracy of 84% on the test data suggests that the model was able to accurately classify drug outcomes. One advantage of SVMs is their ability to handle high-dimensional data, such as the 83-dimensional feature vectors used in this study. The output labels of 1 and -1 are a common convention in SVMs, with positive outcomes assigned a label of 1 and negative outcomes assigned a label of -1.

SVM gave the accuracy of 84%.



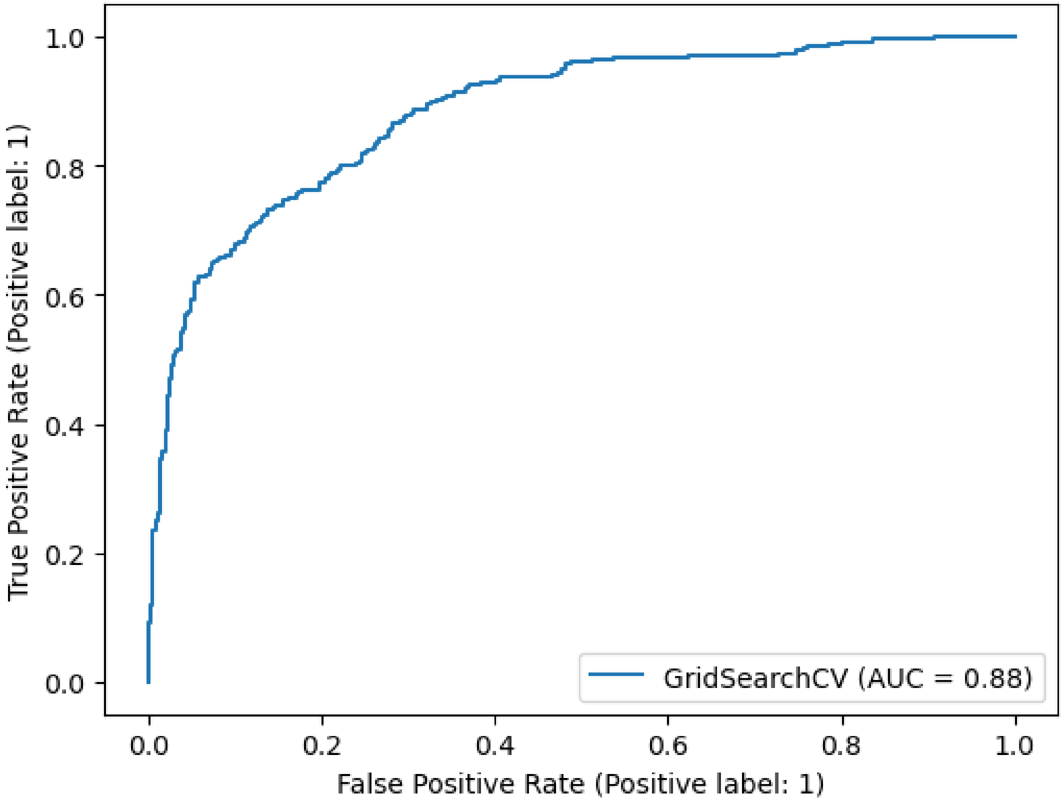

```
              precision    recall  f1-score   support

          -1       0.86      0.89      0.87       533
           1       0.81      0.76      0.78       323

    accuracy                           0.84       856
   macro avg       0.83      0.82      0.83       856
weighted avg       0.84      0.84      0.84       856
```

3.Random Forest algorithm: Random Forest is an ensemble learning method that works by combining multiple decision trees, with the goal of reducing overfitting and increasing prediction accuracy. The number of estimators parameter determines the number of decision trees that will be trained. The max depth parameter sets a limit on the depth of each individual decision tree to prevent overfitting. The accuracy of 81% on the test data suggests that the model was able to accurately classify drug outcomes. The AUC of 89% suggests that the model has good discriminative power, i.e. it is able to differentiate well between positive and negative drug outcomes.



4.Logistic Regression (LR): Model was trained on a dataset of drug persistence, using binary classification, where 0 represents Non-Persistent and 1 represents Persistent labels. Using the different combinations of set hypermeters and GridSearchCV ,an f1_score of 81% was achieved. Overall, the technique proved to be an efficient alternative for optimizing hyperparameters in the LR model, leading to better predictions of drug persistence.

ROC curve is 88%, meaning the algorithm can discriminate easily between good and negative outcomes.



Based on the results, Support Vector Machine is chosen because of the accuracy and its ability to handle high dimensional data.

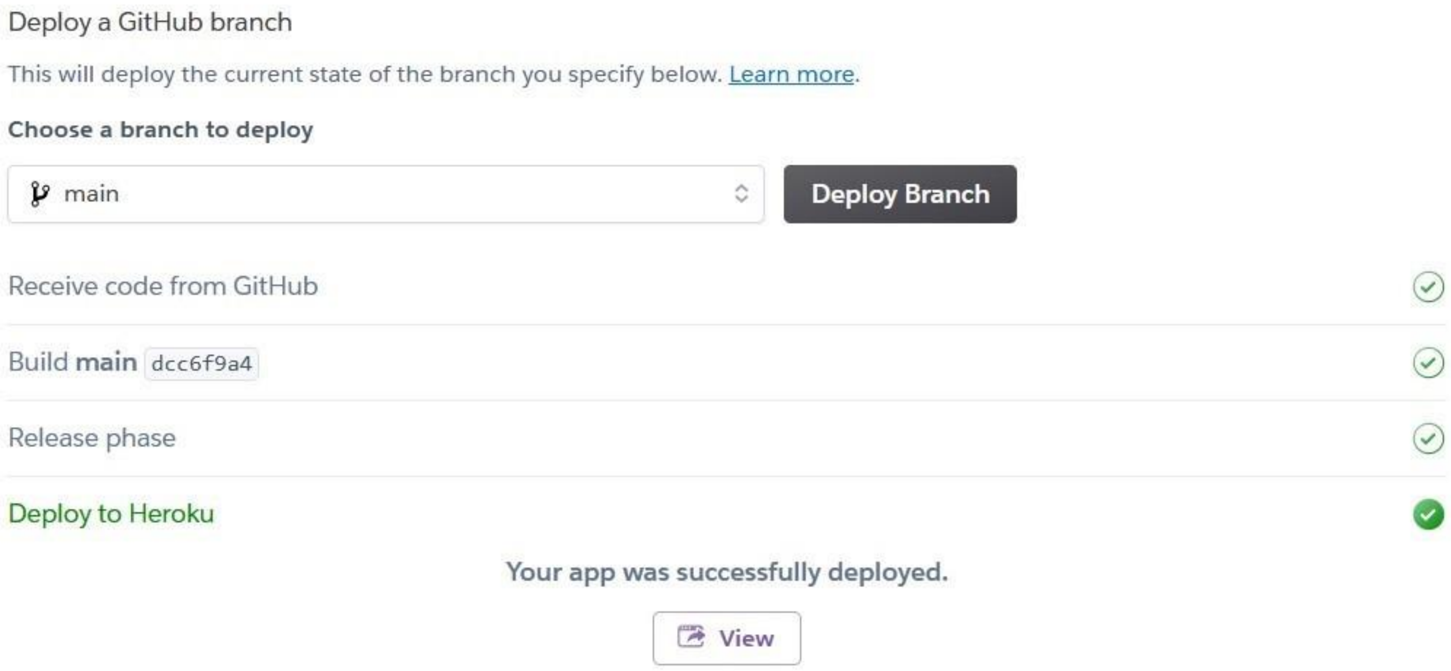
DEPLOYING ON HEROKU:

1.To deploy the model on Heroku, first test and run the model on the local machine using Flask.

2. To deploy a model on Heroku, there are two important files that need to be in the root directory of Github:
a. Requirements.txt ,which lists all the requirements of your Flask for Heroku to read. Requirements is generated by using pipreqs or pip freeze command.
b.Procfile, which is a file without any extension. Heroku reads Procfile and web command in it to know what server to run for. The Procfile must contain web or worker command. Procfile is case sensitive.

3.Once the files are uploaded on to the GitHub repository, link the repository to the Github account.(A different GitHub repo was used because of conflicting files in the original Data Glacier Repo).

If everything is done right, the app should work on the platform.
The app is active on: https://patient-persistence.herokuapp.com/



Deploy a GitHub branch

This will deploy the current state of the branch you specify below. Learn more.

Choose a branch to deploy

main

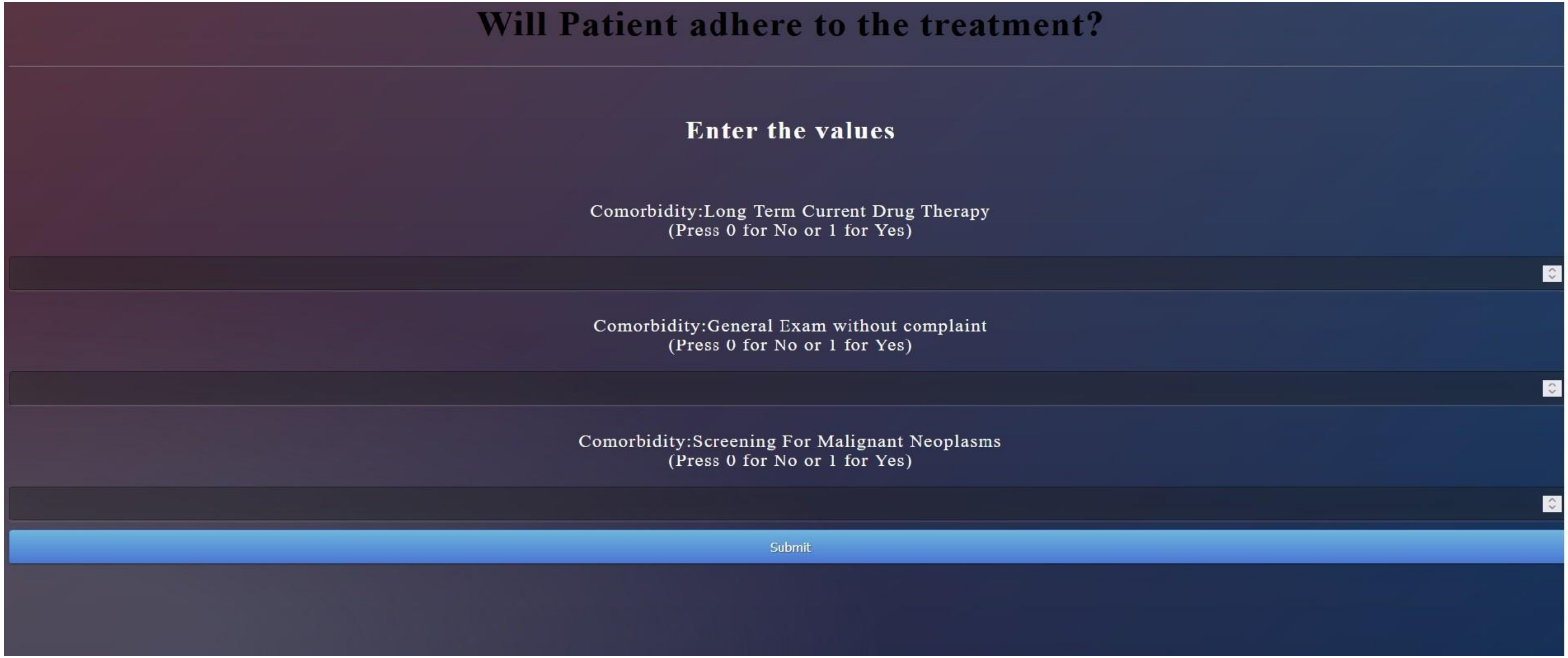
Receive code from GitHub ✓

Build **main** `dcc6f9a4` ✓

Release phase ✓

Deploy to Heroku ✓

Your app was successfully deployed.



Will Patient adhere to the treatment?

Enter the values

Comorbidity:Long Term Current Drug Therapy
(Press 0 for No or 1 for Yes)

Comorbidity:General Exam without complaint
(Press 0 for No or 1 for Yes)

Comorbidity:Screening For Malignant Neoplasms
(Press 0 for No or 1 for Yes)

Entering the values related to the comorbidity should give an answer to whether a patient will be persistent or not. If the values entered are negative or in float , the prompt returns to enter the valid values. Persistency is calculated by SVM; 1 as a result indicates Persistent and-1 as Non persistent.

GitHub: https://github.com/shru0405/DataGlacierInternship/tree/main/Week_13