Modelling

Persistency of a drug

Team : OpenML

Members :
- Juan Carlos
- Laith Adi
- Gerson Orihuela
- Walquer Valles

Date : 15-May-2021

Agenda

- Business Problem
- Decision Tree Model
- Random Forest Model
- Logistic Regression Model
- SVM Model
- Conclusions

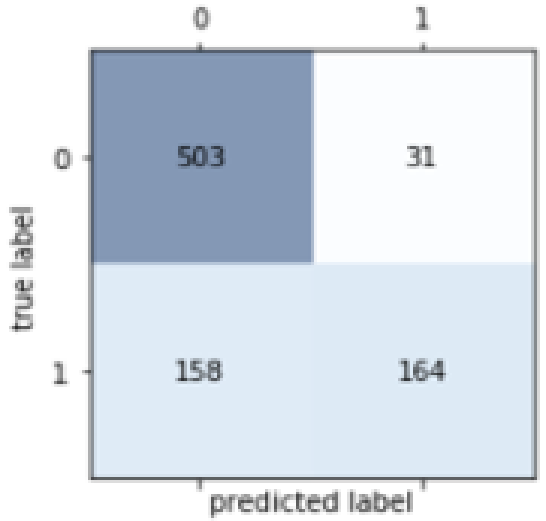
Business problem

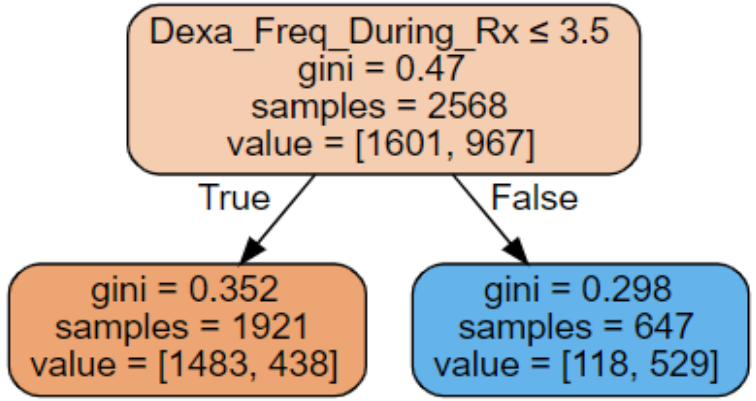
ABC Pharma contacted us to carry out an analysis in order to have a deeper understanding on the factors impacting the **persistence** of their drug. The aim is to know if a patient, based on his/her information, will follow the prescription of the physician and continue taking the drug for all the treatment time.

Decision tree model

Decision tree model was used, the best scores corresponded with a tree of depth 1. the results are the following:
- 78% accuracy was obtained for training data.
- 77% accuracy was obtained for testing data.

The results for the confusion matrix are the following:





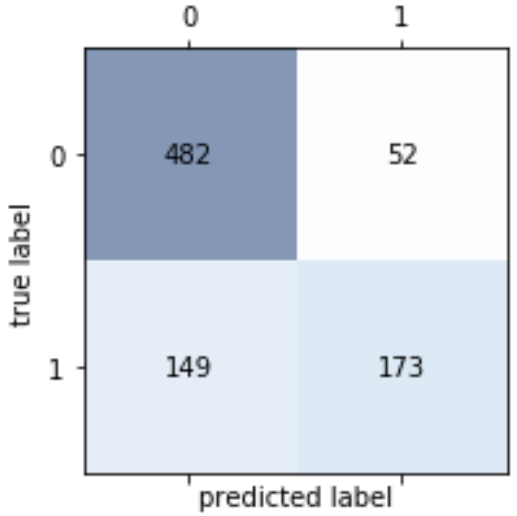
Dexa_Freq_During_Rx is clearly the variable that has the most value in terms of predictive power. It could be interesting to see if we can get similar results without it. Take into account the Dexa_Freq_During_Rx is kind of another type of treatment. So it would be interesting to be able to predict persistence without it.

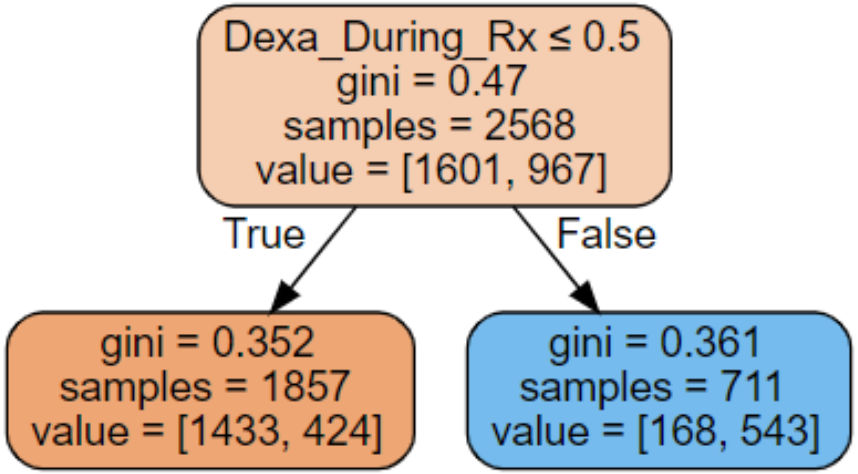
Decision tree model

What if we drop Dexa_Freq_During_Rx?

A tree of depth 1 was obtained. the results are the following:

- 76% accuracy was obtained for training data.
- 76% accuracy was obtained for testing data.





Dexa_During_Rx is a pretty similar to Dexa_Freq_During_Rx. We suppose they refer to the same treatment.

Decision tree model

Without Dexa_Freq_During_Rx and Dexa_During_Rx, comorbidity factors appear as important predictors in a tree of depth 5.

The following chart shows the variables aside from Dexa_Freq_During_Rx and Dexa_During_RX in order of predictive power of persistence in taking the drug.

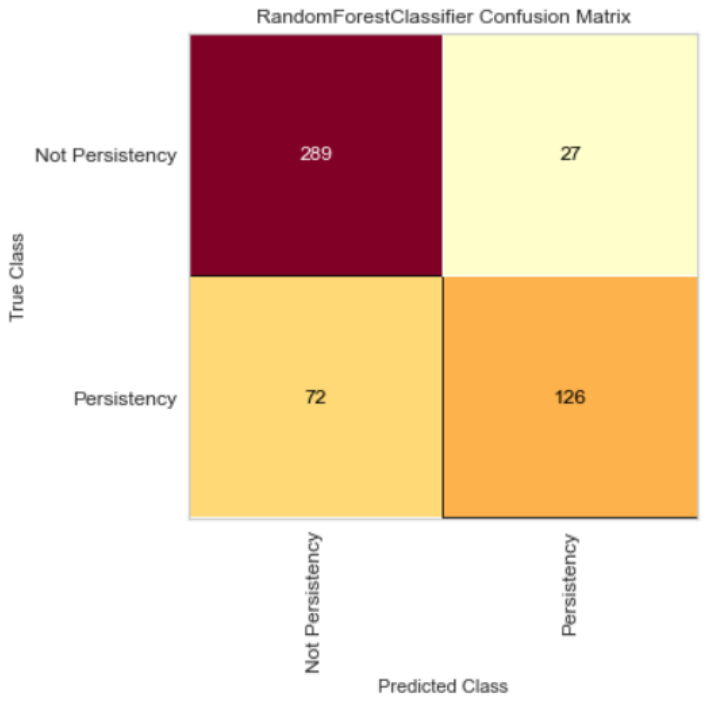
Comorb_Long_Term_Current_Drug_Therapy	0.362902
Comorb_Encntr_For_General_Exam_W_O_Complaint,_...	0.197766
Comorb_Encounter_For_Screening_For_Malignant_N...	0.144858
Comorb_Encounter_For_Immunization	0.049737
Concom_Systemic_Corticosteroids_Plain	0.047695
Concom_Broad_Spectrum_Penicillins	0.026718
Concom_Viral_Vaccines	0.026413
Adherent_Flag	0.023916
Comorb_Vitamin_D_Deficiency	0.021226

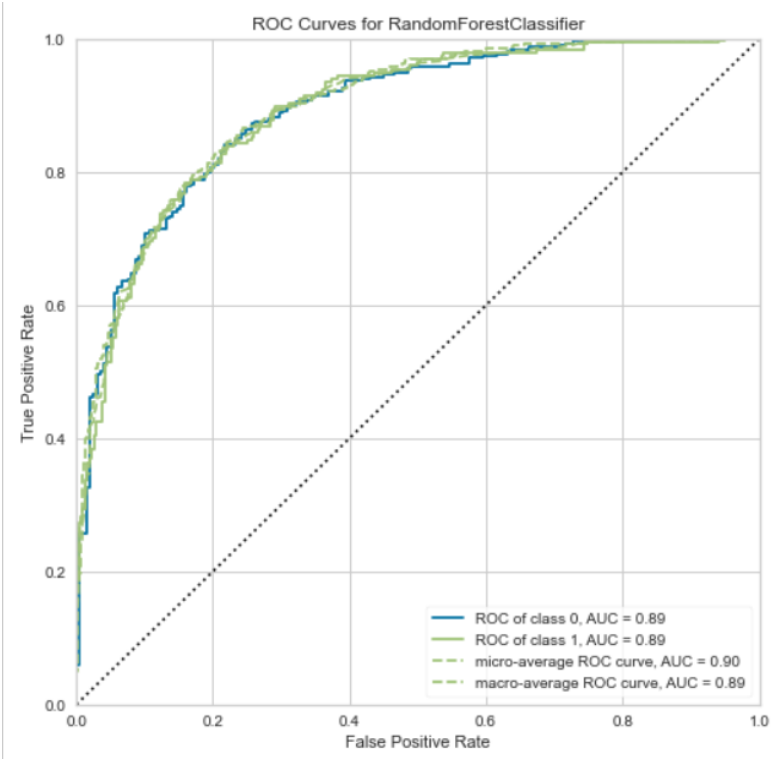
Random Forest Model

Using the chi-square statistic, some variables were eliminated. For the modelling, Random Forest model was used and the results are the following:

- 88% accuracy was obtained for training data.
- 80% accuracy was obtained for testing data.

The results for the confusion matrix and the curve ROC in the testing data are the following:

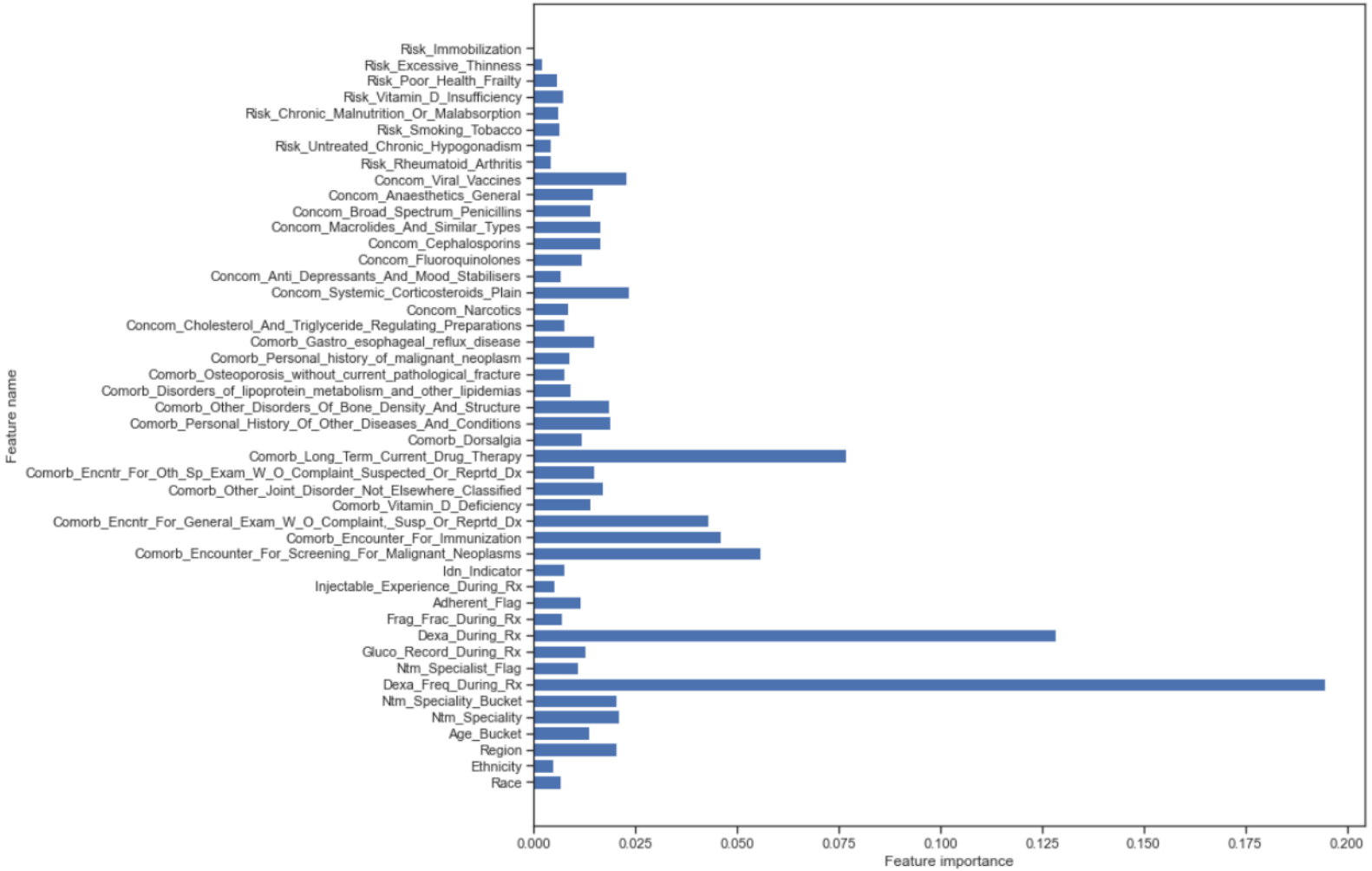




Data Glacier
Your Deep Learning Partner

Random Forest Model

The influence of the variables on the target variable is shown in the following image:



The Dexa_freq_during_rx, Dexa_During_RX and Comorb_Long_Term_Current_Drug_Therapy are the most influential.

Data Glacier
Your Deep Learning Partner

Logistic Regression Model

The following screenshot (top) displays the different parameter and their values when using exhaustive search (GridSearchCV) for tuning the hyperparameters.

```
params = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100, 1000],
    "solver": ["newton-cg", "lbfgs", "liblinear", "sag", "saga"],
    "penalty":["l1","l2"]
}
```

	C	penalty	solver	Accuracy
0	0.001	l1	newton-cg	NaN
1	0.001	l1	lbfgs	NaN
2	0.001	l1	liblinear	0.619549
3	0.001	l1	sag	NaN
4	0.001	l1	saga	0.619549
...	...	...	...	...
65	1000.000	l2	newton-cg	0.809187
66	1000.000	l2	lbfgs	0.812693
67	1000.000	l2	liblinear	0.809187
68	1000.000	l2	sag	0.809577
69	1000.000	l2	saga	0.811524

The bottom screenshot displays the different combinations of set hyperparameters and the resulting accuracy when using GridSearchCV. Next slide the best parameters and score will be shown.

Logistic Regression Model

The following are the best parameters and score, based on accuracy, from the GridSearchCV API from Sk-Learn.

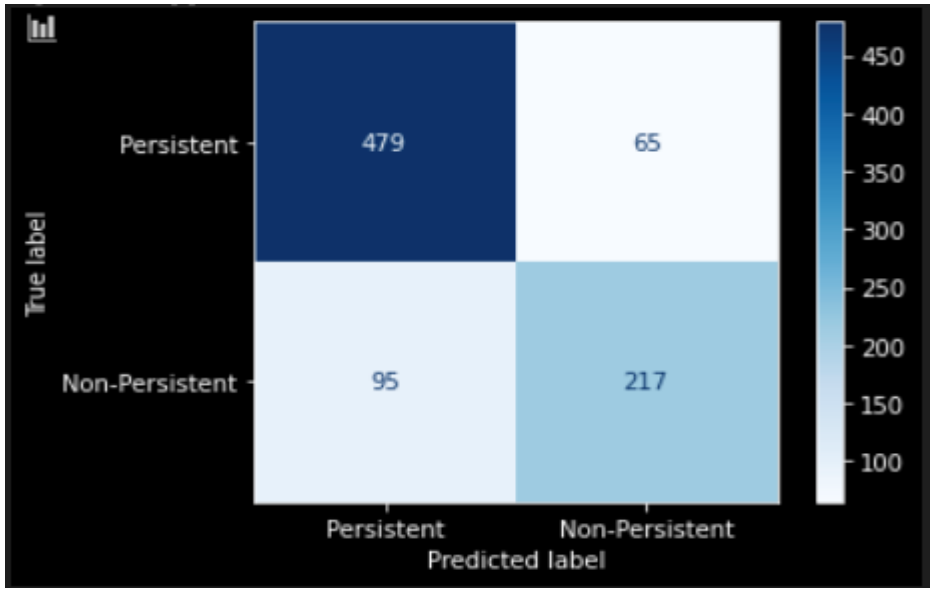
```
{'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
0.8158099528978088
```

These results are based on the fitting of the training sets.

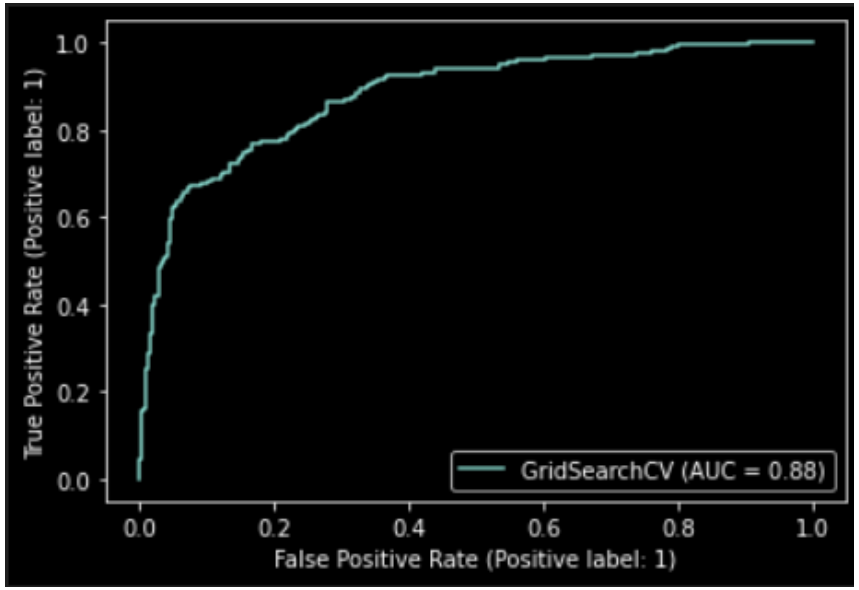
Logistic Regression Model

f1 score is used to evaluate the logistic regression model. Along with the confusion matrix and ROC curve.

- f1 score: 0.81087
- Confusion matrix



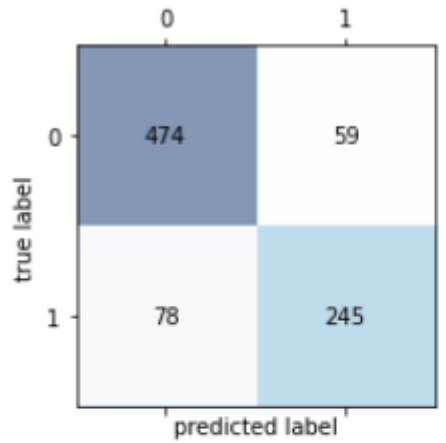
- ROC curve



Data Glacier
Your Deep Learning Partner

SVM Model

Support Vector Machines algorithm to classify the persistence of patients (1 for positives and -1 for negatives). A linear kernel has been used, obtaining an **accuracy of 83.5 %** over testing data (25 % out of the whole dataset).



	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

Conclusions

- The Dexa_freq_during_rx, is the variable with greater power to predict persistent and non-persistent results. This variable if followed in impontance by Dexa_During_RX and Comorb_Long_Term_Current_Drug_Therapy.
- After those the following chart summarize the importance of other variables to predict the target variable.

Comorb_Long_Term_Current_Drug_Therapy	0.362902
Comorb_Encntr_For_General_Exam_W_O_Complaint,_...	0.197766
Comorb_Encounter_For_Screening_For_Malignant_N...	0.144858
Comorb_Encounter_For_Immunization	0.049737
Concom_Systemic_Corticosteroids_Plain	0.047695
Concom_Broad_Spectrum_Penicillins	0.026718
Concom_Viral_Vaccines	0.026413
Adherent_Flag	0.023916
Comorb_Vitamin_D_Deficiency	0.021226

- The best model to be used to make predictions is the SVM model with **83.5 % of accuracy.**

Thank You