Data Glacier

Your Deep Learning Partner

Final Report

Week 13 - LISUM 04

José Vicente Solorzano

18/12/21

Agenda

Business Problem Background

Brief EDA

Models

Models' Metrics

Chosen Model

Deployment

Final Conclusions

Data Glacier
Your Deep Learning Partner

Business Problem Background

Data Glacier
Your Deep Learning Partner

Problem Description
ABC bank is about to launch its new product, a term deposit. Before the launching, they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Why?
To shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more.

Dataset information
The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Goal
Through a Machine Learning Model, determine if the customer is accepting the term deposit or is not.

Brief EDA

Dataset:

```
Number of rows:  41188  –  Number of columns:  21
```

Assumptions:

- “Unknown” will be treated as a category, not as NaN.
- Outliers presented in age and campaign.

Features:

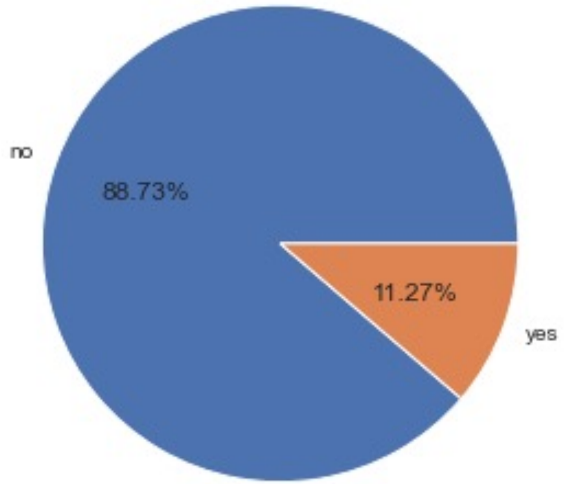
```
['age', 'job', 'marital', 'education', 'default', 'housing', 'loan',
 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays',
 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx',
 'cons.conf.idx', 'euribor3m', 'nr.employed', 'y'],
```

Target distribution:

```
Distribution of target (%):

y
no     88.734583
yes    11.265417
```

Target distribution:



Models
Models' Metrics

Data Glacier
Your Deep Learning Partner

LightGBM

CatBoost

Logistic Regression

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8960	0.8001	0.3245	0.5671	0.4126	0.3602	0.3772	0.6410
catboost	CatBoost Classifier	0.8976	0.7950	0.3104	0.5866	0.4057	0.3555	0.3772	6.8270
lr	Logistic Regression	0.8239	0.7908	0.6367	0.3468	0.4489	0.3548	0.3781	0.6580

Models
Models' Metrics

LightGBM

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8981	0.7793	0.2738	0.6054	0.3771	0.3301	0.3611
1	0.9057	0.8071	0.3426	0.6529	0.4494	0.4032	0.4284
2	0.8911	0.8076	0.2778	0.5294	0.3644	0.3111	0.3305
3	0.9022	0.8066	0.3046	0.6387	0.4125	0.3664	0.3964
4	0.8907	0.8022	0.2585	0.5316	0.3478	0.2959	0.3190
5	0.8987	0.7911	0.2831	0.6093	0.3866	0.3393	0.3691
6	0.9001	0.8164	0.3015	0.6164	0.4050	0.3574	0.3847
7	0.9008	0.7990	0.3138	0.6182	0.4163	0.3684	0.3938
8	0.8980	0.7934	0.2800	0.6026	0.3824	0.3348	0.3642
9	0.8970	0.7980	0.2954	0.5854	0.3926	0.3430	0.3670
Mean	0.8982	0.8001	0.2931	0.5990	0.3934	0.3449	0.3714
SD	0.0044	0.0099	0.0227	0.0386	0.0274	0.0291	0.0303

CatBoost

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8967	0.7753	0.2892	0.5839	0.3868	0.3374	0.3623
1	0.9022	0.8060	0.3488	0.6141	0.4449	0.3957	0.4148
2	0.8907	0.8095	0.2932	0.5249	0.3762	0.3216	0.3380
3	0.8994	0.8005	0.2831	0.6174	0.3882	0.3415	0.3726
4	0.8956	0.7961	0.2954	0.5714	0.3895	0.3386	0.3608
5	0.8998	0.7925	0.3015	0.6125	0.4041	0.3562	0.3831
6	0.8991	0.8177	0.3231	0.5966	0.4192	0.3692	0.3901
7	0.8963	0.7875	0.3169	0.5722	0.4079	0.3562	0.3749
8	0.8942	0.7837	0.2769	0.5625	0.3711	0.3206	0.3447
9	0.8911	0.7822	0.3046	0.5294	0.3867	0.3317	0.3470
Mean	0.8965	0.7951	0.3033	0.5785	0.3975	0.3469	0.3688
SD	0.0036	0.0127	0.0203	0.0315	0.0210	0.0219	0.0222

Logistic Regression

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8204	0.7690	0.5877	0.3322	0.4244	0.3276	0.3464
1	0.8189	0.8010	0.6759	0.3443	0.4562	0.3611	0.3907
2	0.8182	0.7908	0.6481	0.3387	0.4449	0.3488	0.3751
3	0.8266	0.8006	0.6154	0.3478	0.4444	0.3510	0.3710
4	0.8345	0.8018	0.6615	0.3694	0.4741	0.3851	0.4082
5	0.8339	0.7881	0.6123	0.3605	0.4538	0.3635	0.3813
6	0.8273	0.8073	0.6738	0.3584	0.4679	0.3761	0.4029
7	0.8193	0.7836	0.6369	0.3393	0.4428	0.3467	0.3712
8	0.8200	0.7758	0.6215	0.3378	0.4377	0.3415	0.3641
9	0.8210	0.7904	0.6338	0.3416	0.4440	0.3485	0.3722
Mean	0.8240	0.7908	0.6367	0.3470	0.4490	0.3550	0.3783
SD	0.0059	0.0116	0.0271	0.0113	0.0138	0.0160	0.0174

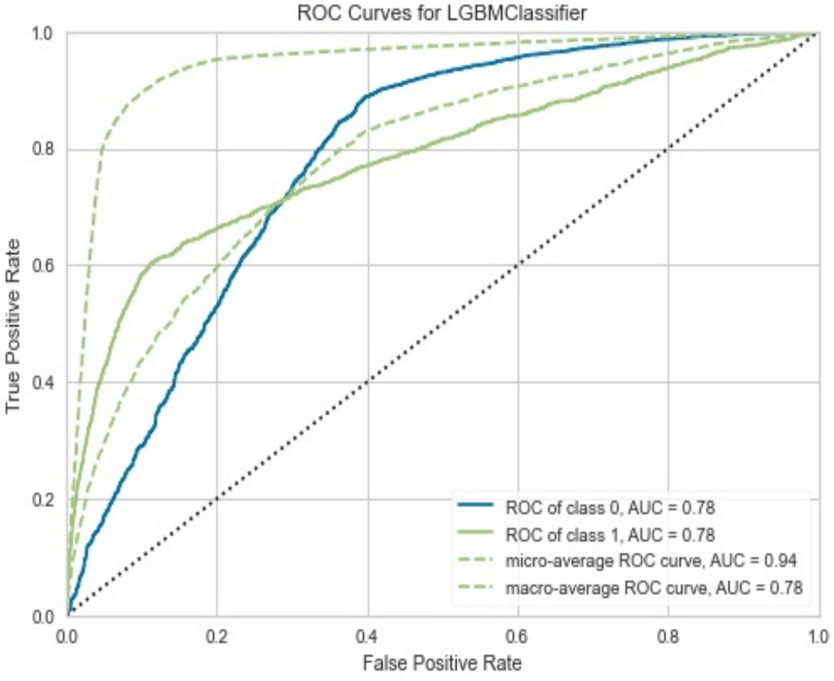
Models
Models' Metrics

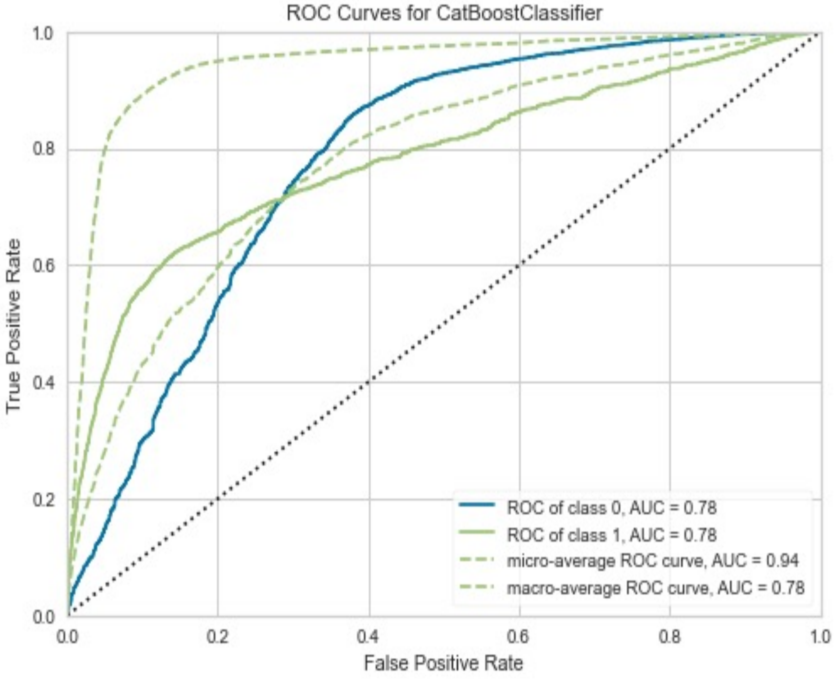
Data Glacier
Your Deep Learning Partner

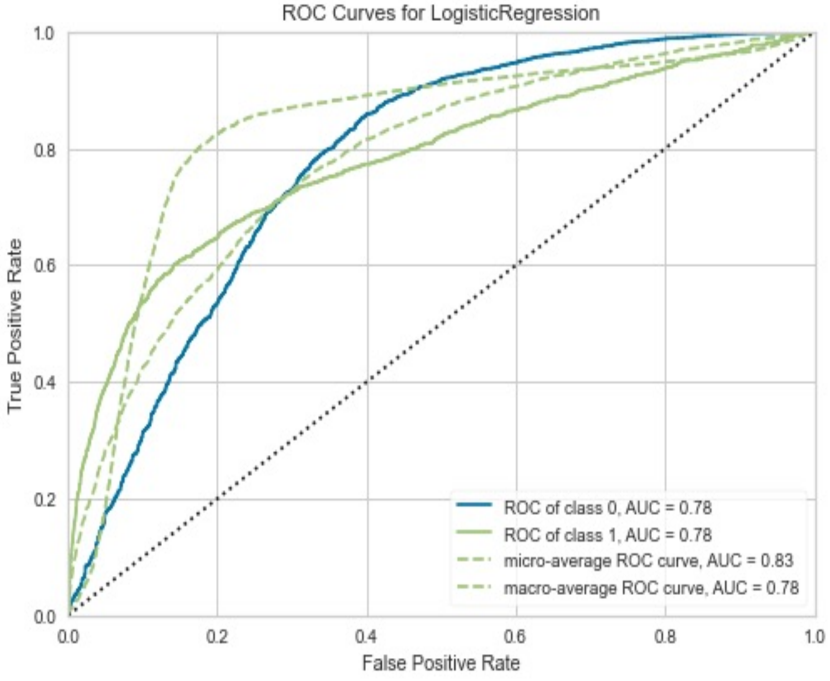
LightGBM

CatBoost

Logistic Regression





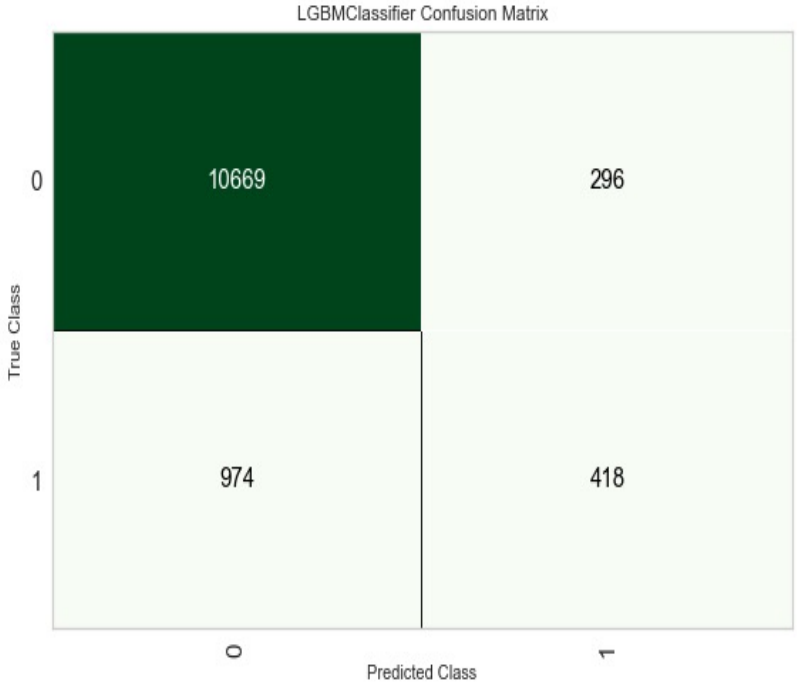


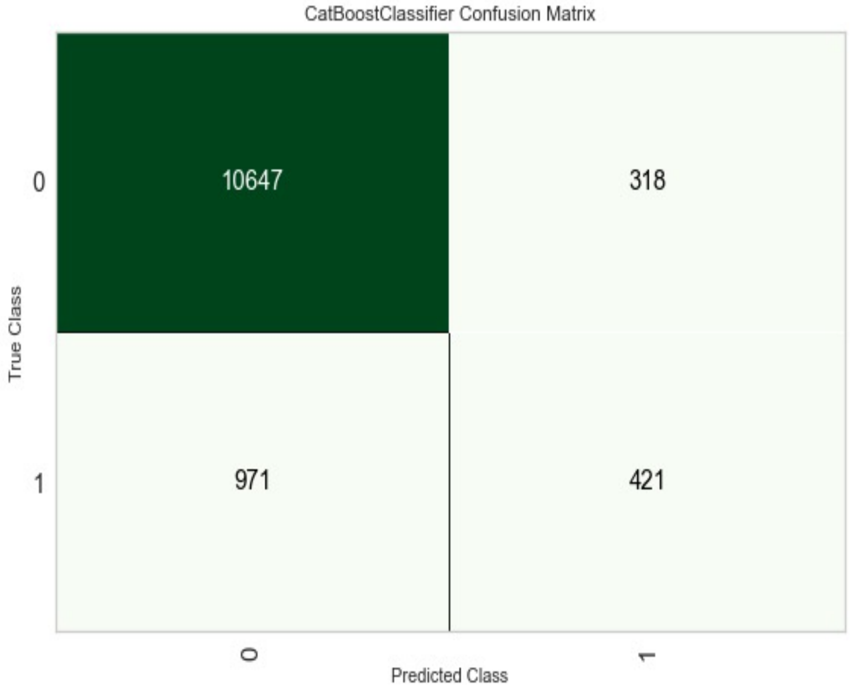
Models
Models' Metrics

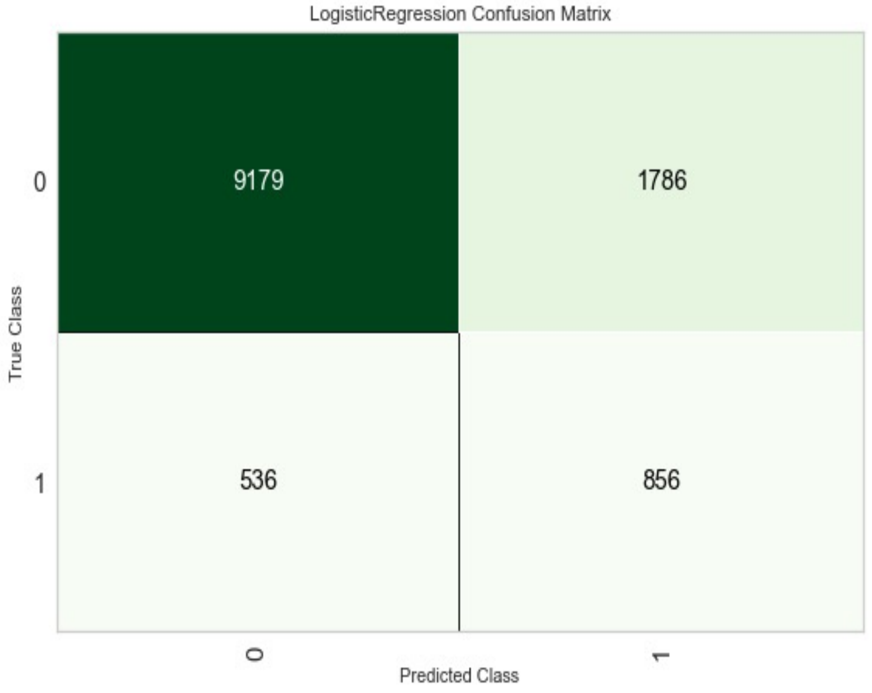
LightGBM

CatBoost

Logistic Regression





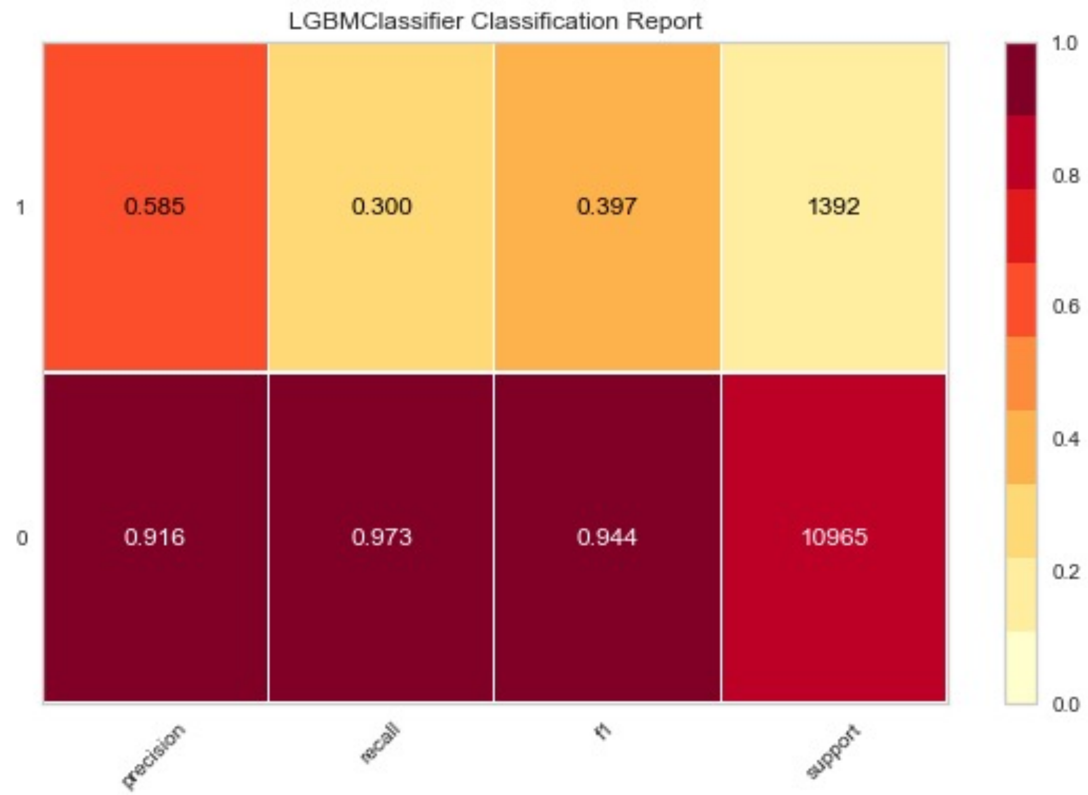


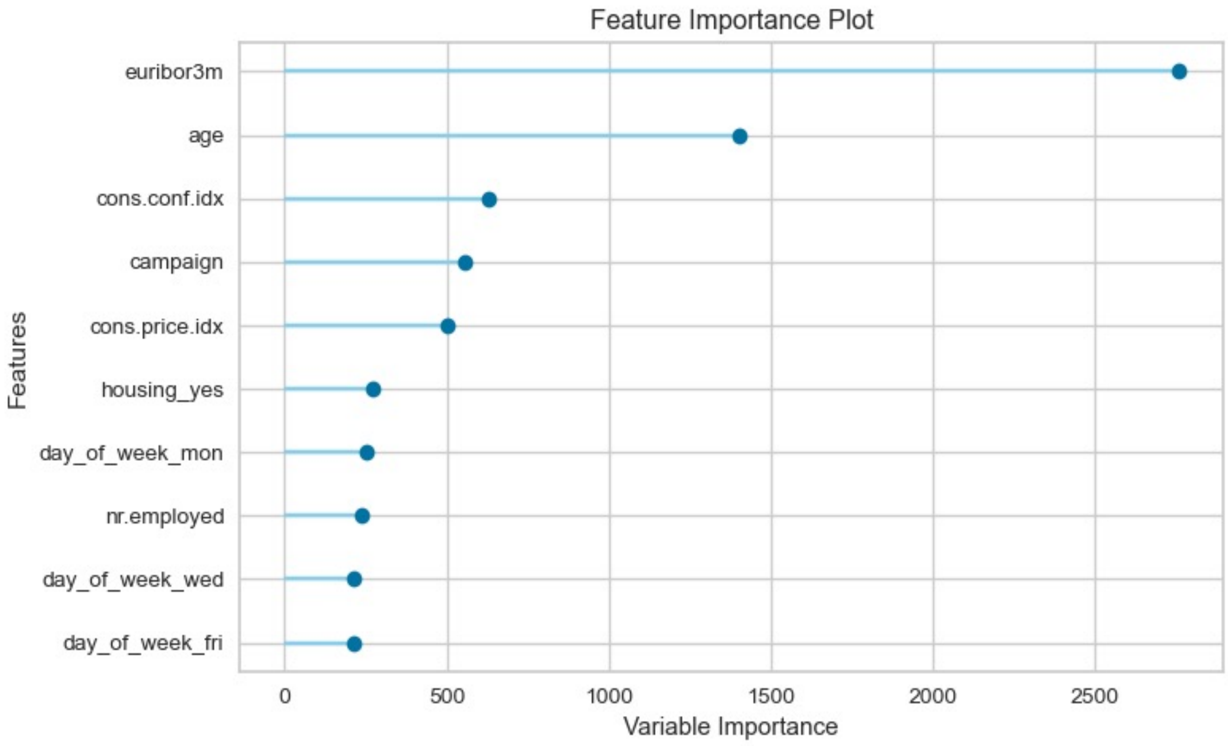
Chosen Model

LightGBM

The chosen model was LightGBM. Our criteria for choosing was based on AUC, because is the best indicator for binary problems.

Some additional metrics of our chosen model:





Deployment

Similar to week 5, the deployment was on Heroku as a simple API that predicts the acceptance of bank customers. As we already know, the Machine Learning Model is Light GBM.

HEROKU

Final Conclusions

- We tried different treatment for outliers and missing values, but finally the best results came from WOE treatment, and to keep ”unknown” as its own category. We experimented considering them as nulls, but the metrics got worse.

- Through Pycaret, we tried several models, but we went deeper in the first three: LightGBM, Catboost and Logistic Regression. After analysis, the model we chose for our model was LightGBM. It arrives to better predictions.

- It would be interesting to analyse Threshold optimization, because some metrics, such as precision and recall are low.

Recommended Model

We implemented pycaret library to know the model would perform better in our business problem.

The criterion was based mainly in AUC (area under the curve), the results were:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8960	0.8001	0.3245	0.5671	0.4126	0.3602	0.3772	0.6410
catboost	CatBoost Classifier	0.8976	0.7950	0.3104	0.5866	0.4057	0.3555	0.3772	6.8270
lr	Logistic Regression	0.8239	0.7908	0.6367	0.3468	0.4489	0.3548	0.3781	0.6580

We are going to try with the 2 best models:

1) Light Gradient Boosting Machine
2) CatBoost Classifier

Next week we are going to implement some Grid Search and hyperparameters optimization to get the ideal model.

Thank You

Data Glacier

Your Deep Learning Partner