Data Glacier

Your Deep Learning Partner

Final Report

Virtual Internship

Felipe I. Crespo

Agenda

- Business Problem Background
- Brief EDA
- Models
- Models Metrics
- Chosen Model
- Deployment
- Final Conclusions

Data Glacier
Your Deep Learning Partner

Business Problem Background

Problem Description

- ABC bank is launching a new term deposit product and wants to develop a model to predict whether a customer will buy it or not.
- The model will help the bank target their marketing efforts to customers who are more likely to buy the product.

Dataset Information

- The dataset is related to direct marketing campaigns of a Portuguese banking institution.
- The campaigns were based on phone calls and aimed to sell a bank term deposit.
- Multiple contacts were often required to determine whether a client subscribed to the deposit or not.

Goal

- The goal is to develop a machine learning model that can predict whether a customer will accept the term deposit or not.

Brief EDA

Dataset

```
Rows = 41188  | Columns =  21
```

Feauters

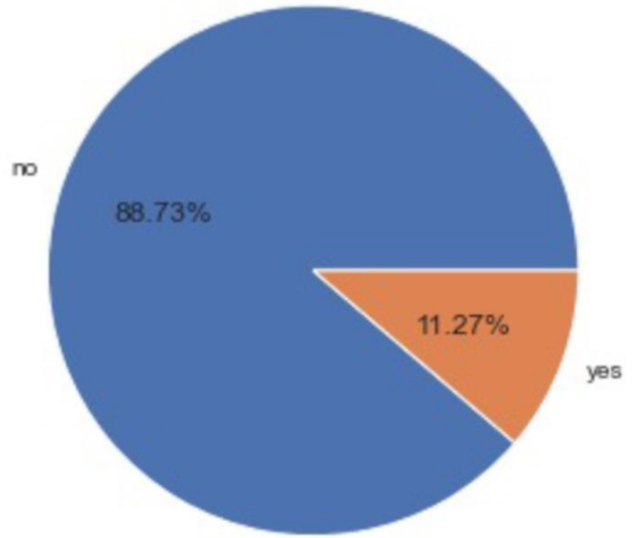
```
['age', 'job', 'marital', 'education', 'default', 'housing', 'loan',
 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays',
 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx',
 'cons.conf.idx', 'euribor3m', 'nr.employed', 'y'],
```

Assumptions

- The category "Unknown" will be treated as a distinct category and not as a missing value (NaN).
- The dataset may contain outliers in the age and campaign variables.

```
Distribution of target (%):

y
no     88.734583
yes    11.265417
```



Models

LightGBM

CatBoost

Logistic Regression

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8970	0.8007	0.3267	0.5767	0.4167	0.3649	0.3829	0.2900
catboost	CatBoost Classifier	0.8973	0.7962	0.3045	0.5855	0.4001	0.3501	0.3728	7.1710
lr	Logistic Regression	0.8231	0.7903	0.6352	0.3458	0.4476	0.3532	0.3764	0.3960
lda	Linear Discriminant Analysis	0.8210	0.7895	0.6352	0.3421	0.4446	0.3492	0.3730	0.1390

Models Metrics

LightGBM

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9001	0.8087	0.3077	0.6135	0.4098	0.3618	0.3876
1	0.8970	0.7903	0.2500	0.6000	0.3529	0.3071	0.3422
2	0.8956	0.7916	0.3179	0.5628	0.4063	0.3539	0.3713
3	0.8949	0.7834	0.2985	0.5640	0.3903	0.3387	0.3594
4	0.9008	0.8196	0.3169	0.6168	0.4187	0.3705	0.3952
5	0.8932	0.8065	0.2738	0.5528	0.3663	0.3151	0.3384
6	0.9025	0.8057	0.3169	0.6358	0.4230	0.3762	0.4036
7	0.9015	0.8020	0.3015	0.6323	0.4083	0.3619	0.3916
8	0.8956	0.8066	0.3077	0.5682	0.3992	0.3475	0.3672
9	0.9015	0.7930	0.3231	0.6213	0.4251	0.3771	0.4013
Mean	0.8983	0.8007	0.3014	0.5967	0.4000	0.3510	0.3758
SD	0.0032	0.0103	0.0217	0.0301	0.0227	0.0231	0.0225

CatBoost

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8984	0.8007	0.3108	0.5941	0.4081	0.3584	0.3810
1	0.8994	0.7942	0.2809	0.6149	0.3856	0.3390	0.3701
2	0.8949	0.7858	0.3241	0.5556	0.4094	0.3560	0.3717
3	0.8921	0.7848	0.2769	0.5422	0.3666	0.3143	0.3356
4	0.9025	0.8187	0.3415	0.6236	0.4414	0.3929	0.4144
5	0.8949	0.7977	0.3138	0.5604	0.4024	0.3497	0.3675
6	0.9001	0.7949	0.2985	0.6178	0.4025	0.3551	0.3833
7	0.8928	0.7910	0.2615	0.5519	0.3549	0.3045	0.3299
8	0.8977	0.8090	0.3231	0.5833	0.4158	0.3648	0.3840
9	0.9001	0.7856	0.3138	0.6108	0.4146	0.3661	0.3905
Mean	0.8973	0.7962	0.3045	0.5855	0.4001	0.3501	0.3728
SD	0.0033	0.0104	0.0235	0.0293	0.0239	0.0243	0.0237

Logistic Regression

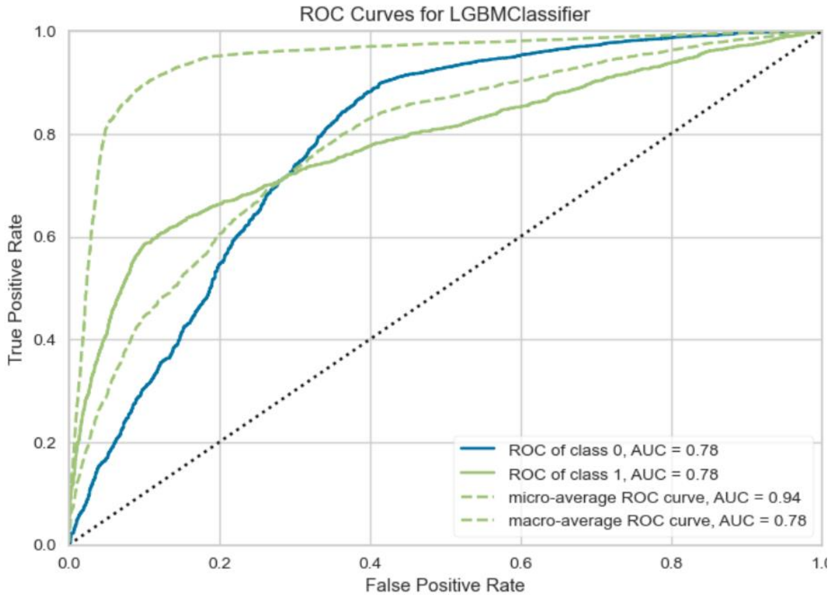
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8315	0.7995	0.6462	0.3614	0.4636	0.3729	0.3951
1	0.8123	0.7691	0.6173	0.3241	0.4251	0.3257	0.3499
2	0.8130	0.7886	0.6451	0.3302	0.4368	0.3384	0.3658
3	0.8030	0.7712	0.5908	0.3062	0.4034	0.2993	0.3225
4	0.8456	0.8227	0.6800	0.3932	0.4983	0.4147	0.4364
5	0.8280	0.7933	0.6431	0.3548	0.4573	0.3651	0.3879
6	0.8248	0.7841	0.6400	0.3490	0.4517	0.3580	0.3814
7	0.8283	0.7993	0.6246	0.3524	0.4506	0.3581	0.3787
8	0.8266	0.8056	0.6677	0.3563	0.4647	0.3724	0.3986
9	0.8179	0.7697	0.5969	0.3299	0.4250	0.3273	0.3476
Mean	0.8231	0.7903	0.6352	0.3458	0.4476	0.3532	0.3764
SD	0.0114	0.0165	0.0269	0.0230	0.0252	0.0305	0.0303

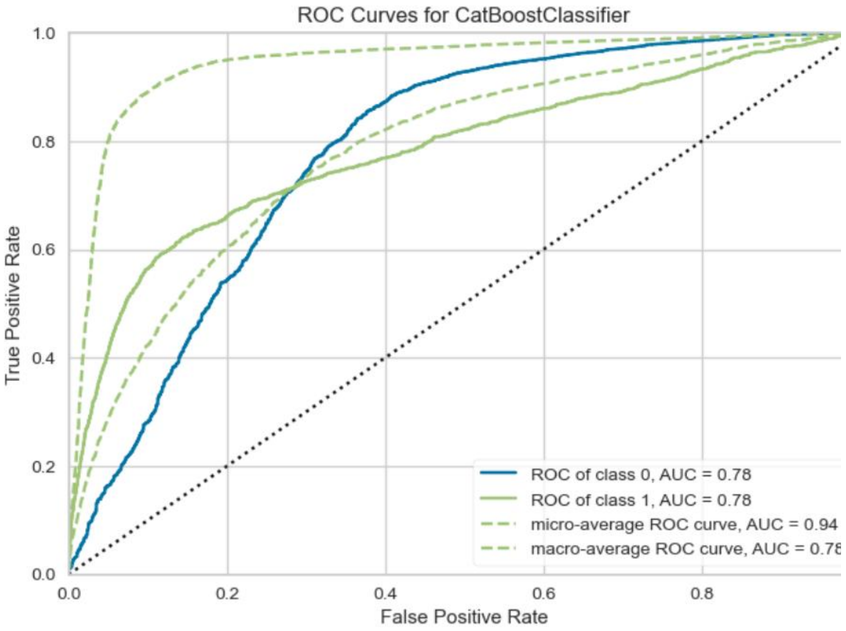
Models Metrics

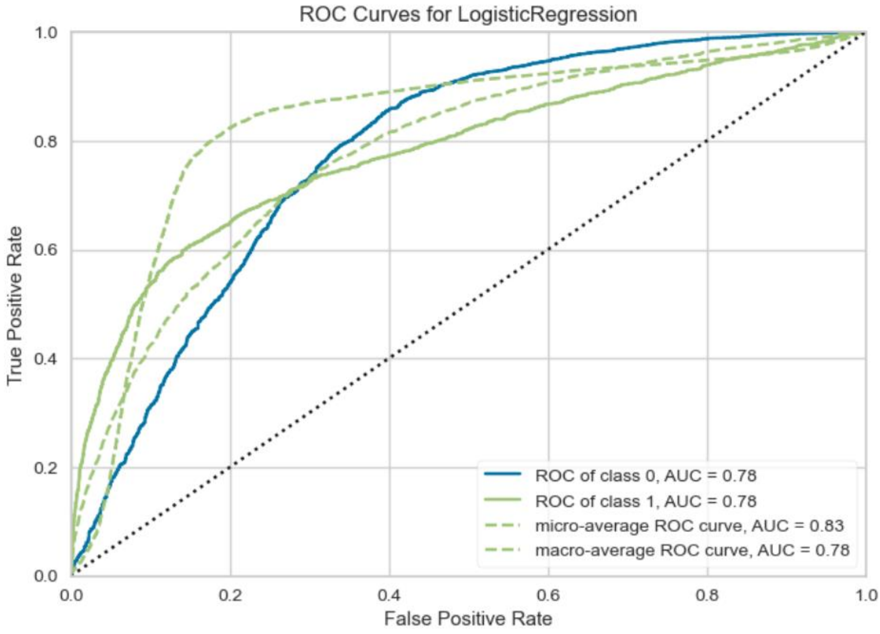
LightGBM

CatBoost

Logistic Regression

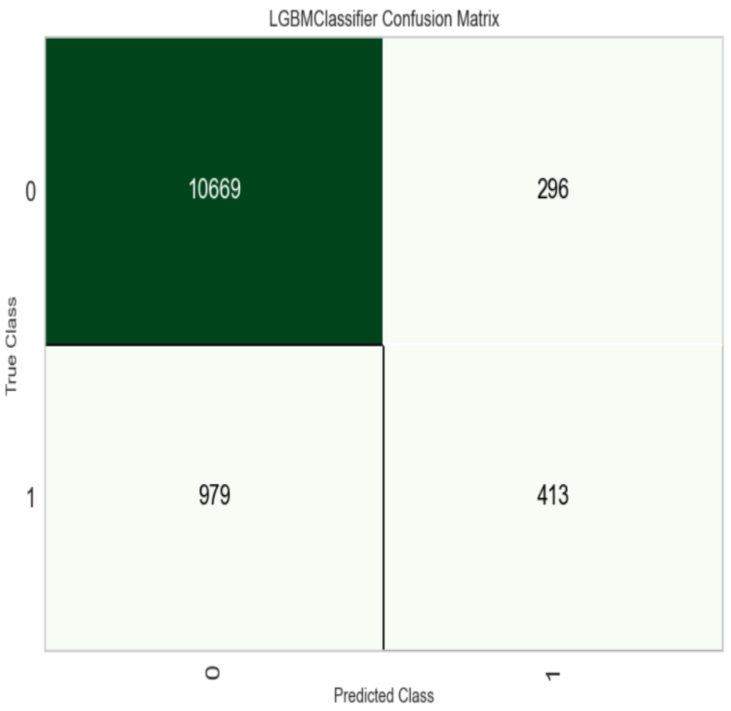




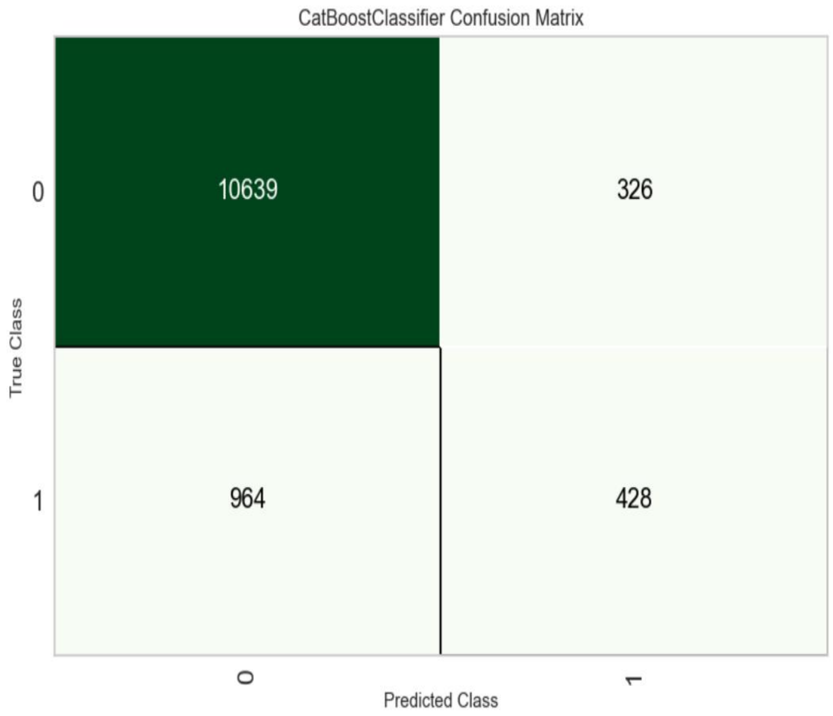


Models Metrics

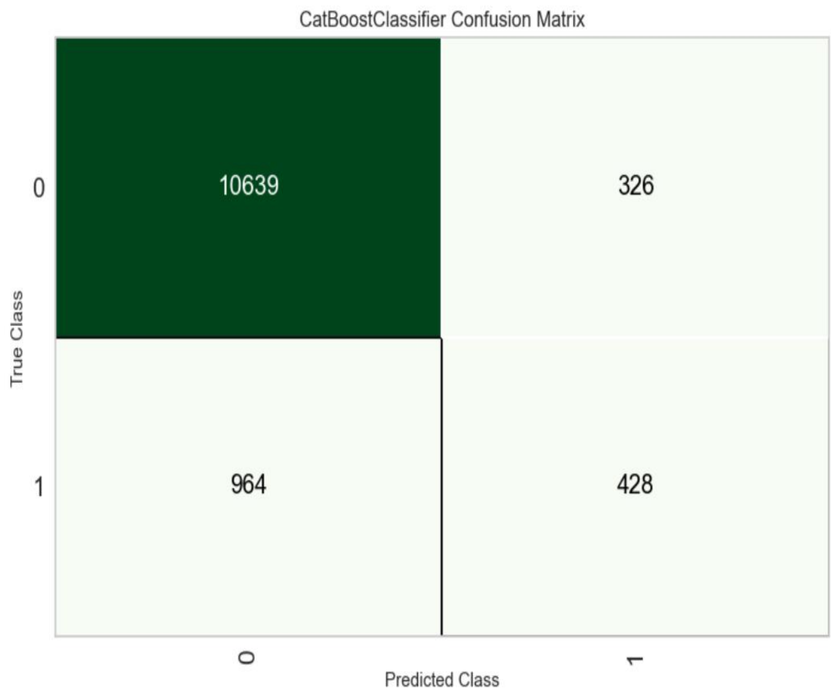
LightGBM



CatBoost



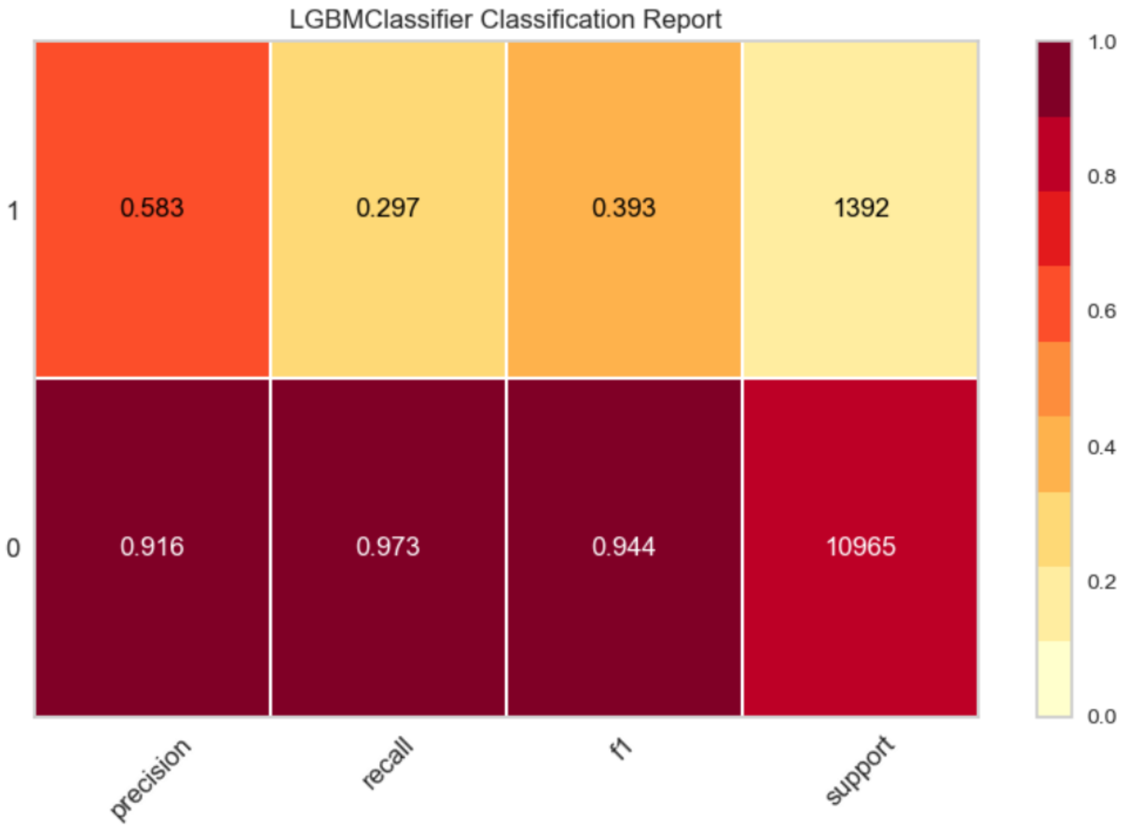
Logistic Regression

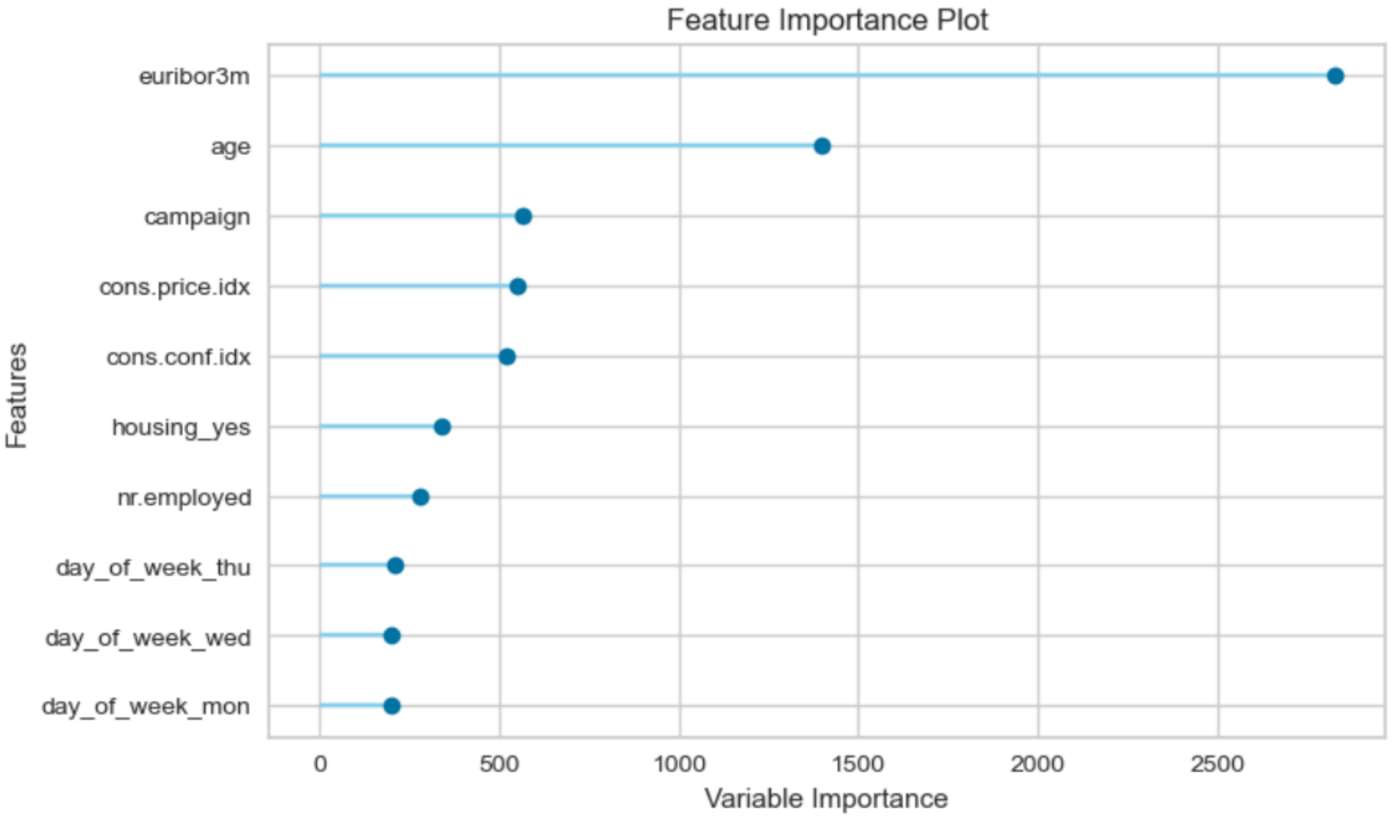


Chosen Model

LightGBM

The LightGBM model was selected based on its performance in AUC, which is considered the most reliable indicator for binary classification problems.

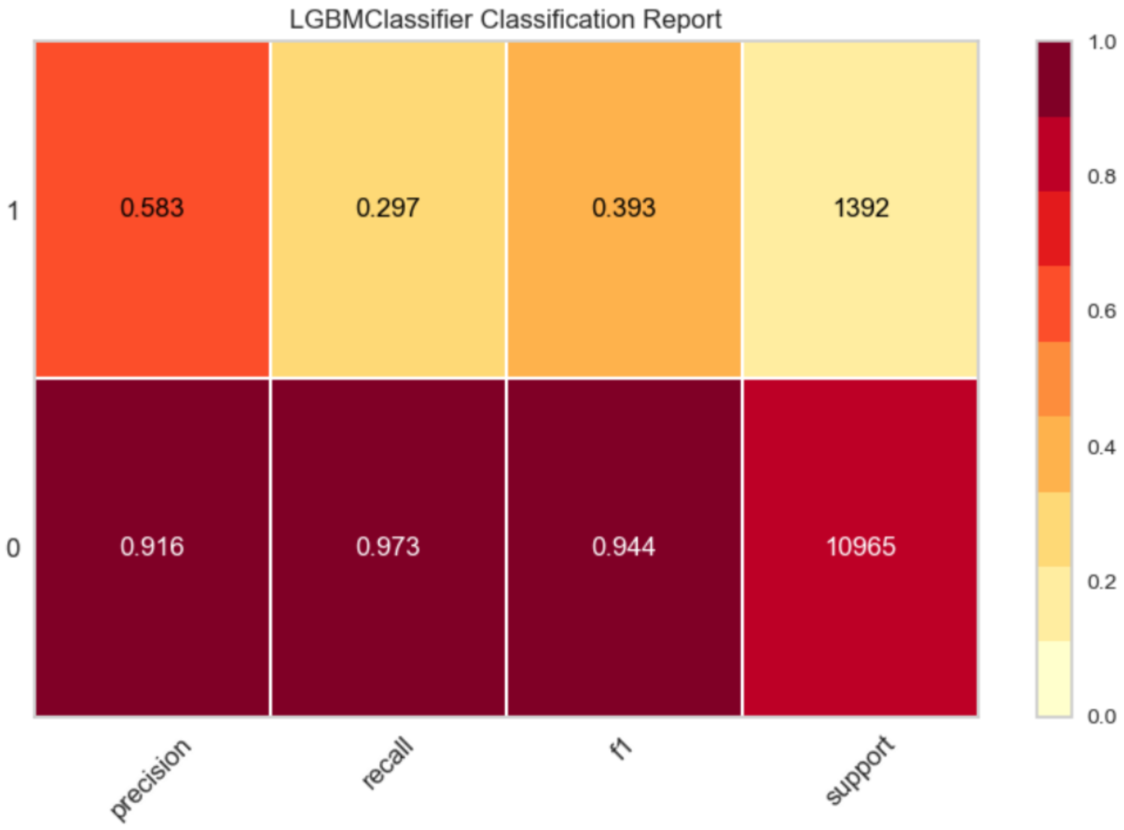


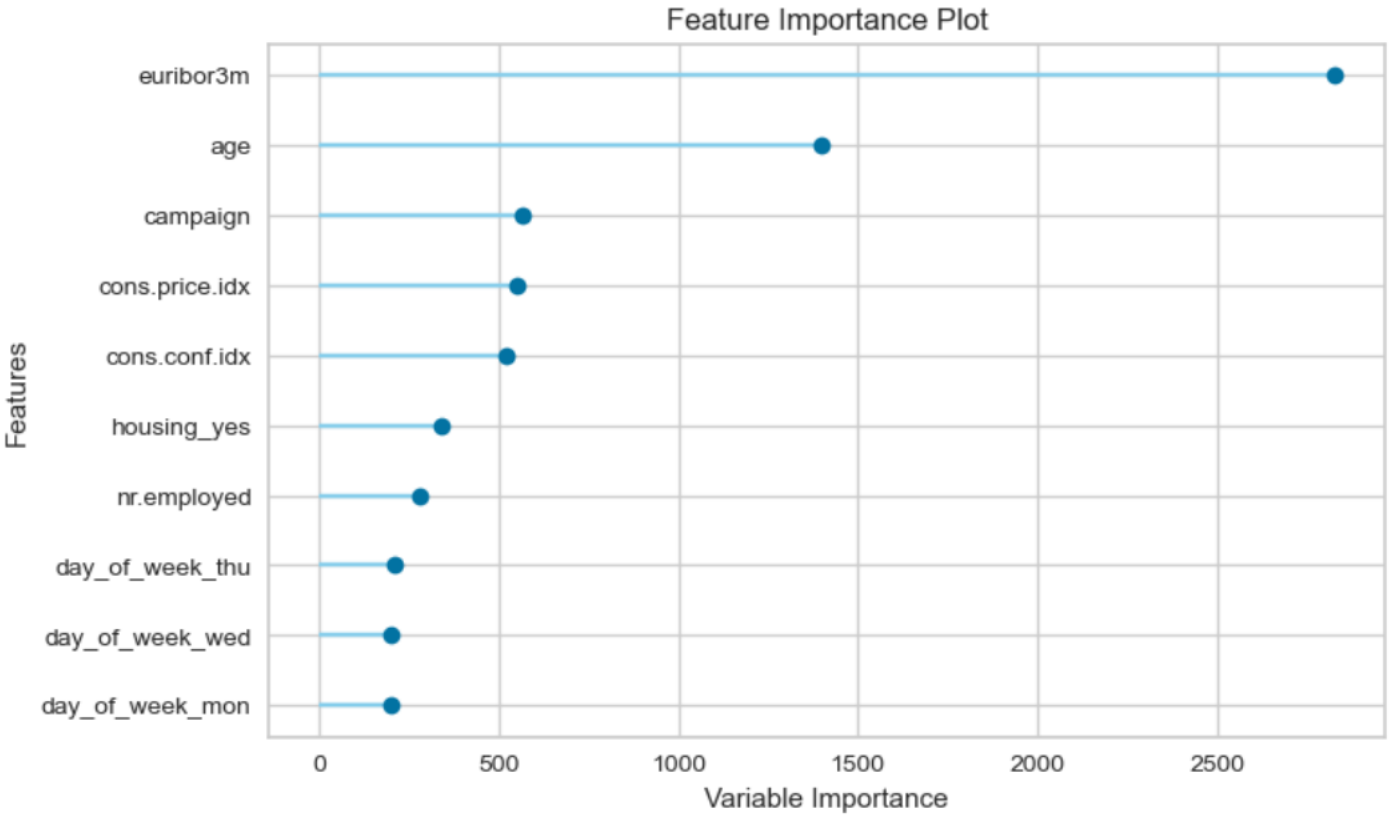


Chosen Model

LightGBM

The LightGBM model was selected based on its performance in AUC, which is considered the most reliable indicator for binary classification problems.





Deployment

HEROKU

We deployed the Light GBM Machine Learning Model as a simple API on Heroku, similar to our Week 5 deployment. This API allows for the prediction of bank customer acceptance.

Final Conclusions

- We tested different techniques for handling outliers and missing values, including treating "unknown" as a separate category and applying Weight of Evidence (WOE) treatment. The best results were obtained using these methods, while treating "unknown" as nulls resulted in worse metrics.

- From a range of models tested using Pycaret, our analysis focused on the three top-performing models: LightGBM, Catboost, and Logistic Regression. After thorough consideration, we ultimately chose LightGBM as our final model due to its superior predictive performance.

- While our model showed good performance overall, there may be room for improvement through threshold optimization, particularly for metrics such as precision and recall, which had lower scores.

Thank You

Data Glacier

Your Deep Learning Partner