Final Project

Virtual Internship

15-May-2021

Background – Bank marketing campaign case study

- **Problem Statement:** ABC Bank wants to sell it's term deposit product to customers and before launching the product 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 ML Model:** Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more. This will save resource and their time (which is directly involved in the cost (resource billing)).
- **Data Set 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. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).
- **Objective:** Obtain a model that determines whether or not X customer will buy your product, based on past interactions with the bank and other financial institutions.

Data Exploration

Dataset shape:

- 21 Features
- 41188 instances

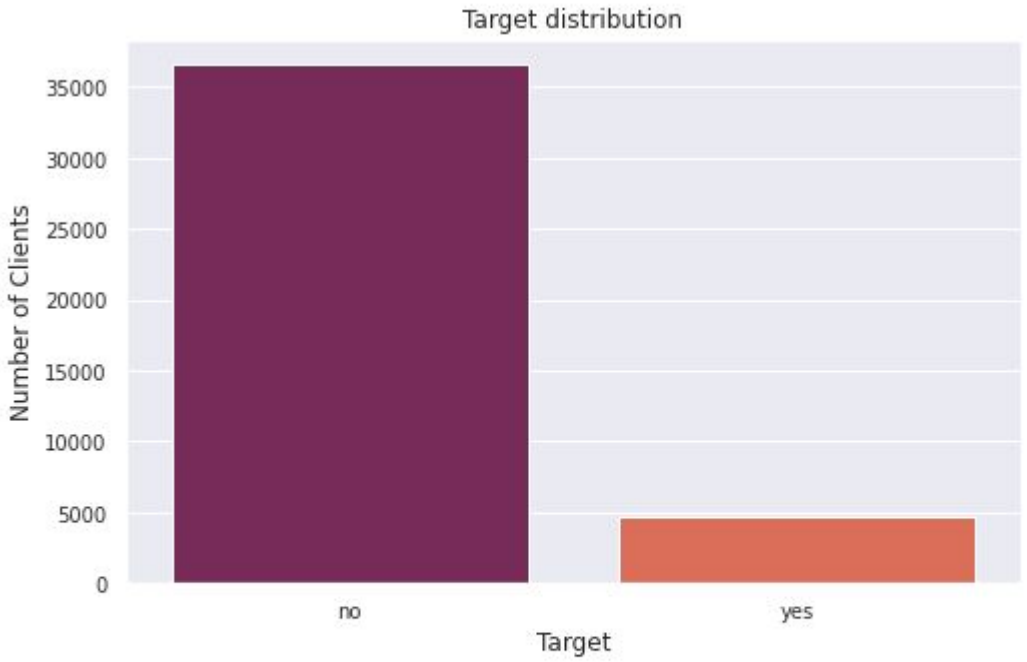
Assumptions:

- Outliers are present in age, duration, campaign, pdays.
- “Unknown” values are treated as NaN values

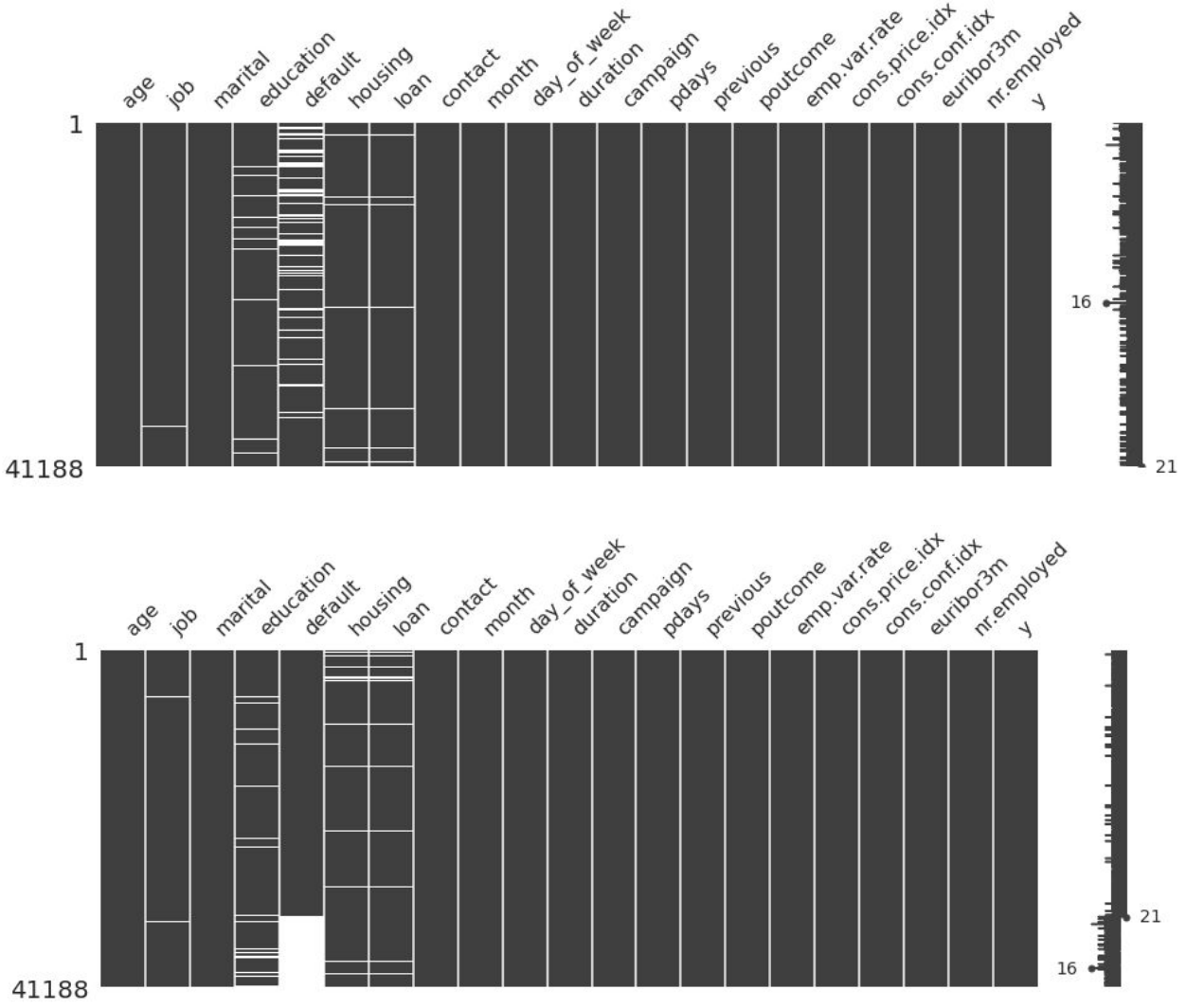
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

The dataset has 21 columns (features) and 41188 rows (instances)

Brief EDA




```
The amount of "no" target is =  36,548
The amount of "yes" target is = 4,640
The percentage of "no" target is = 88.73 %
The percentage of "yes" target is = 11.27 %
```



Recommended models

- Experiment n° 1:

This experiment was carried out by eliminating the outliers detected by the IQR method. For this case the top three models were:

  - Light Gradient Boosting Machine
  - CatBoost Classifier
  - Extreme Gradient Boosting

- Experiment n° 2:

This experiment was carried out by treating the outliers with the WOE method. For this case the top three models obtained were:

  - Light Gradient Boosting Machine
  - Gradient Boosting Classifier
  - Logistic Regresion

Best Models Metrics

VotingClassifier (Blending) for experiment n°1

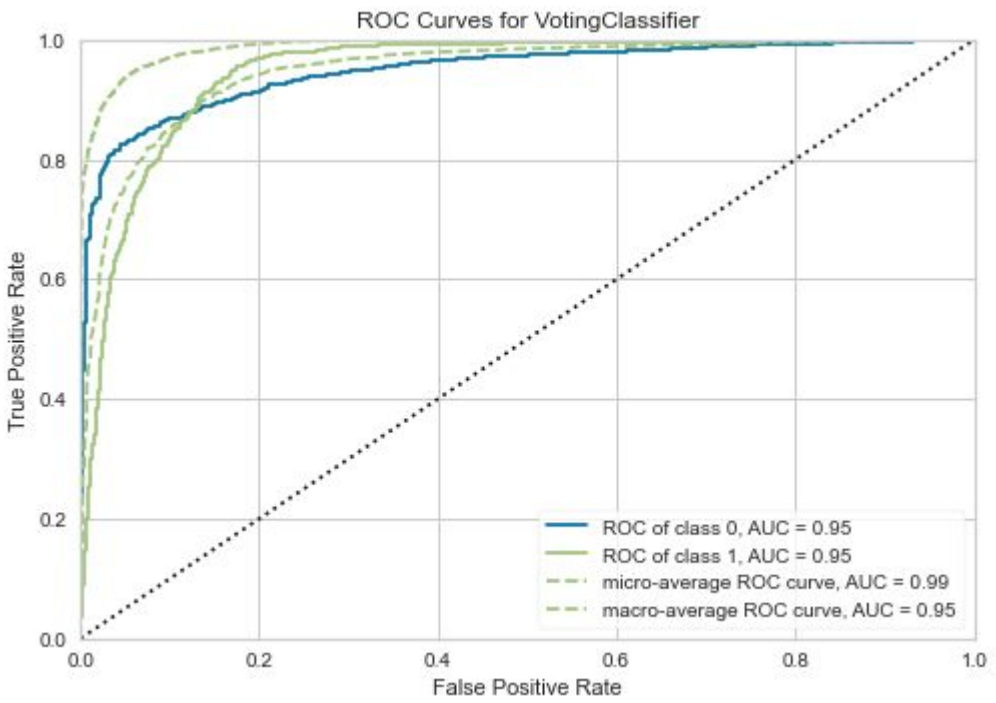
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9424	0.9562	0.5545	0.5600	0.5572	0.5264	0.5264
1	0.9449	0.9481	0.5000	0.5882	0.5405	0.5115	0.5134
2	0.9372	0.9549	0.5900	0.5130	0.5488	0.5153	0.5167
3	0.9365	0.9340	0.4900	0.5104	0.5000	0.4661	0.4662
4	0.9398	0.9385	0.4800	0.5393	0.5079	0.4760	0.4769
5	0.9359	0.9442	0.4500	0.5056	0.4762	0.4422	0.4430
6	0.9391	0.9519	0.5149	0.5361	0.5253	0.4927	0.4929
7	0.9437	0.9512	0.5248	0.5761	0.5492	0.5192	0.5199
8	0.9333	0.9378	0.4257	0.4886	0.4550	0.4197	0.4208
9	0.9398	0.9431	0.5149	0.5417	0.5279	0.4958	0.4960
Mean	0.9393	0.9460	0.5045	0.5359	0.5188	0.4865	0.4872
SD	0.0035	0.0073	0.0452	0.0305	0.0320	0.0335	0.0334

LightGBM for expermient n°2

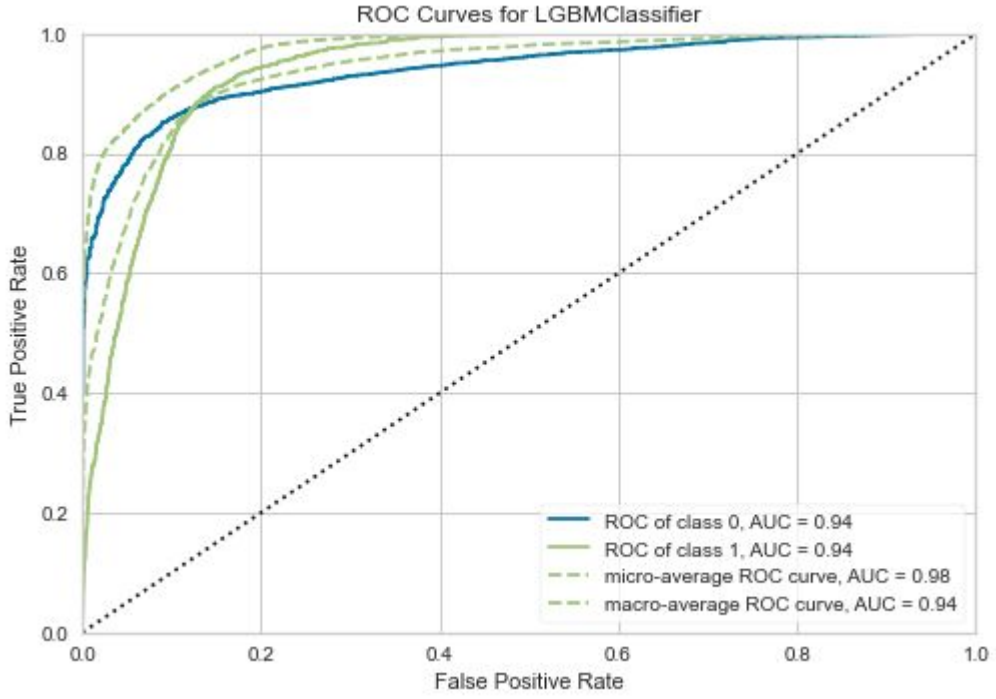
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8904	0.9321	0.5498	0.5709	0.5602	0.4976	0.4977
1	0.9025	0.9387	0.6111	0.6157	0.6134	0.5576	0.5576
2	0.8913	0.9282	0.5926	0.5674	0.5797	0.5173	0.5175
3	0.9067	0.9397	0.6407	0.6291	0.6349	0.5814	0.5814
4	0.8927	0.9307	0.6074	0.5714	0.5889	0.5272	0.5276
5	0.8922	0.9284	0.5889	0.5719	0.5803	0.5185	0.5186
6	0.8927	0.9264	0.5481	0.5804	0.5638	0.5027	0.5030
7	0.8978	0.9355	0.6000	0.5956	0.5978	0.5393	0.5393
8	0.8978	0.9306	0.5926	0.5970	0.5948	0.5364	0.5364
9	0.9110	0.9436	0.5852	0.6695	0.6245	0.5743	0.5759
Mean	0.8975	0.9334	0.5916	0.5969	0.5938	0.5352	0.5355
SD	0.0067	0.0054	0.0261	0.0312	0.0234	0.0271	0.0272

Best Models Evaluation

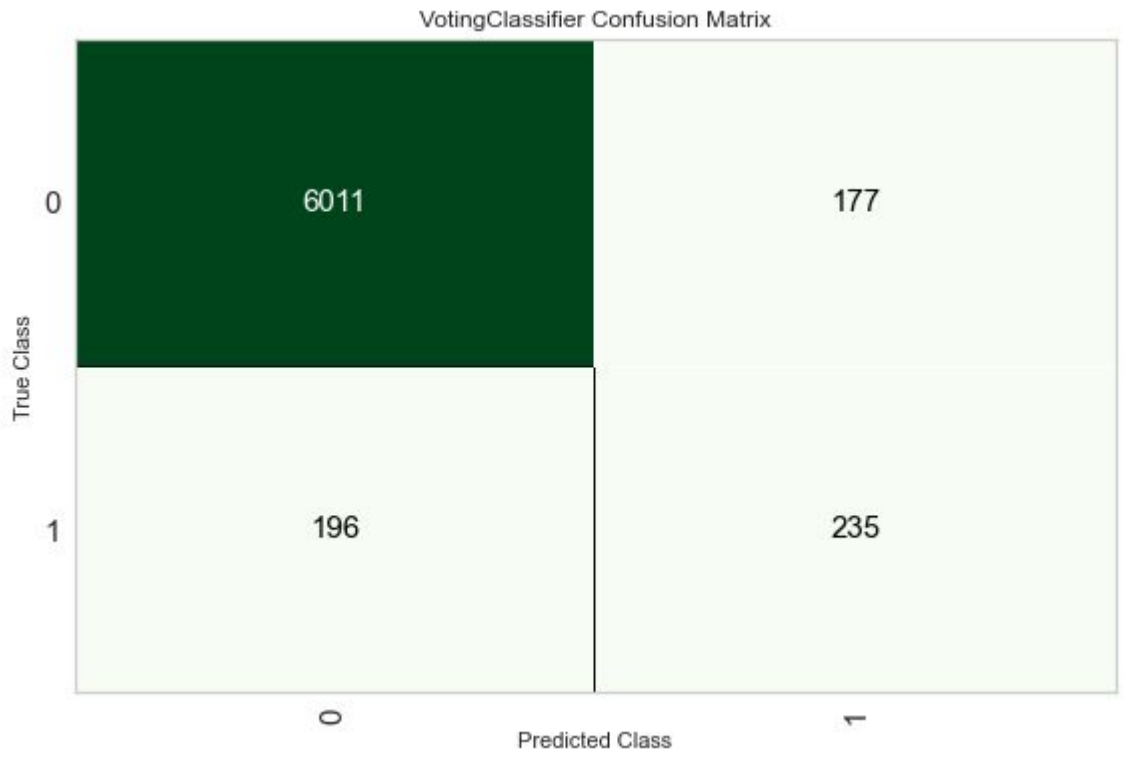
Experiment n°1

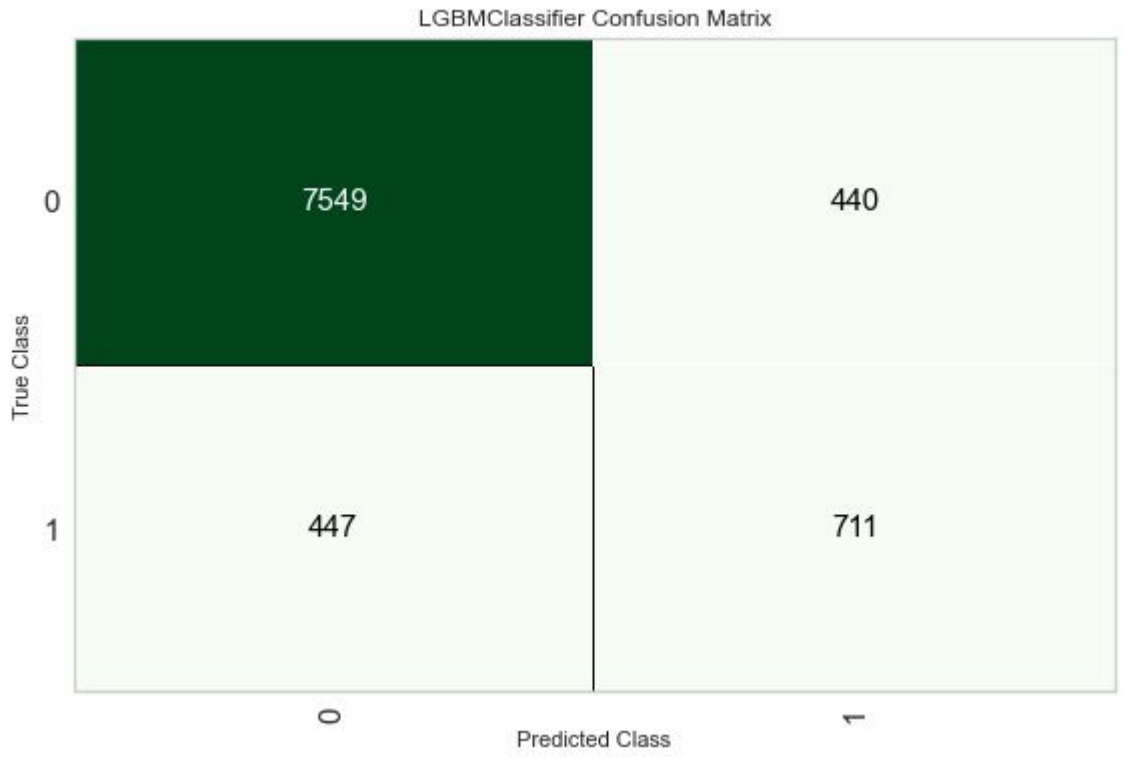


Experiment n°2



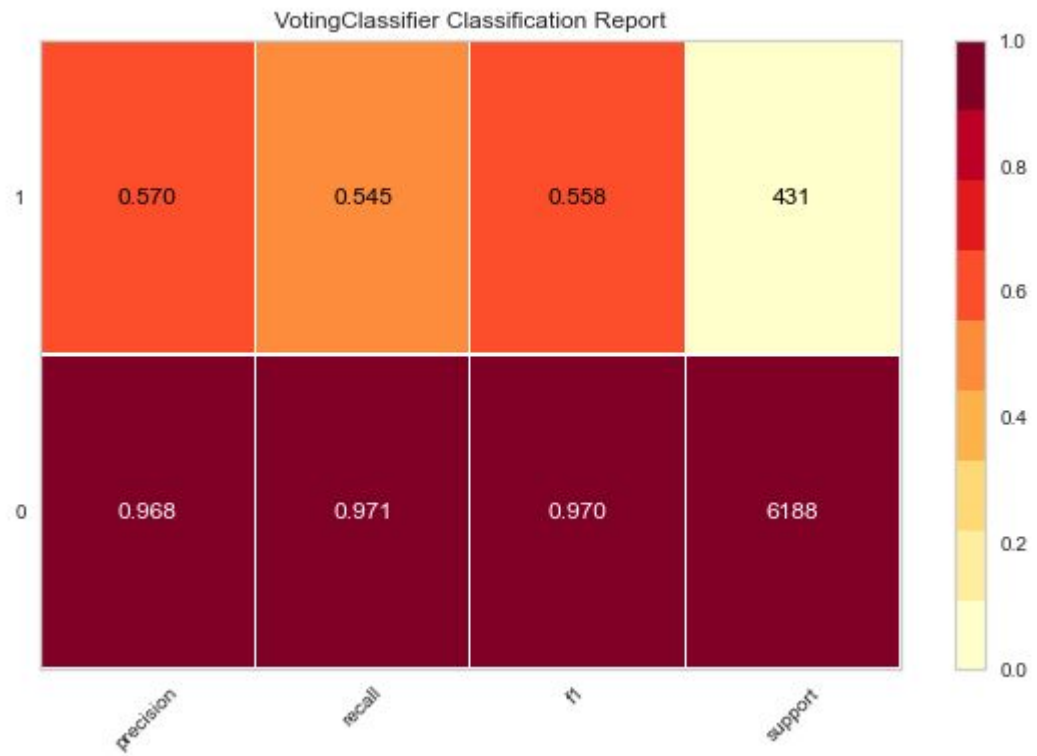
Best Models Evaluation

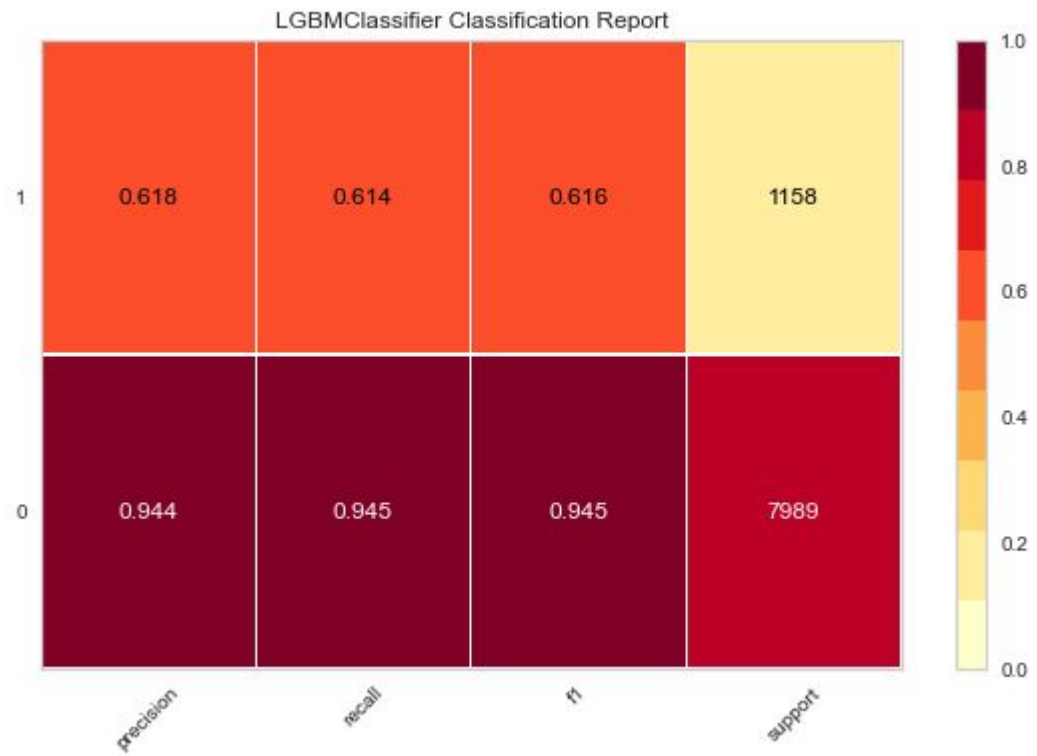




no: 0
yes: 1

Best Models Evaluation





Best Model

Blending: VotingClassifier

- The best model is found treating outliers with IQR method and deleting all NaNs values from the dataset.
- The best model is found blending four models: lightGBM Classifier, CatBoost Classifier, XGBoost Classifier, Random Forest Classifier. The resulting Voting Classifier is better than individual models.

Model Deployment

https://bankmkt.herokuapp.com/

Conclusions

- Different machine learning experiments were developed to find the best model to predict whether a given bank customer will buy a fixed term deposit.
- Blending of LightGBM, CatBoost, Random Forest and XGBoost yield the best model for predicting fixed term deposit.
- The Voting Classifier obtained has good Accuracy and AUC values but poor Recall and Precision values.
- Threshold optimization will be required for further improvement of the model performance.

Thank You