Trustly Case Study

Dealing with unbalanced classification

Outline

Exploratory Data Analisis

First we briefly present the description of the data:

- Summary
- Data Types
- Missing Values

Methods

Here we describe what technique we used to handle missing values, cross validation, class unbalance, and key algorithm performance metrics

Results

Finally we present and discuss the results for 3 simple models

Exploratory Data Analysis

Summary

Geolocation variable -> convert to lat-long

	SAFRA	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	CEP	TARGET
0	201901	NaN	8.1	9.99	1968	0.0	0	15.15	0.0	0.0	0	SP	São Paulo	8412006	0
1	201910	0.0	4.4	35.00	1369	0.0	0	63.98	1.0	0.0	0	RJ	Rio de Janeiro	23580304	0
2	201906	0.0	0.7	52.99	1228	0.0	0	98.84	0.0	0.0	0	MG	Belo Horizonte	30421310	0
3	201910	0.0	63.3	810.00	0	0.0	1	9237.21	0.0	0.0	0	SP	São Paulo	8253410	0
4	201902	0.0	4.1	17.50	0	0.0	1	27.70	1.0	0.0	0	ES	Vitória	29017186	0

Binary Target

Temporal variable

Missing Data

Exploratory Data Analysis

Summary

	SAFRA	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	CEP	TARGET
count	11169.000000	10437.000000	10942.000000	11169.000000	11169.000000	10263.000000	11169.000000	11008.000000	10821.000000	11057.000000	11169.000000	1.116900e+04	11169.000000
mean	201906.522339	0.106448	19.750658	531.046901	1396.048438	0.186982	0.177903	4345.434375	0.397468	0.008592	0.030531	2.006559e+07	0.010744
std	3.447787	0.308425	25.442371	906.626021	1736.590512	0.640979	0.382448	11527.310213	0.489397	0.092297	0.172051	1.019638e+07	0.103100
min	201901.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	7.050301e+06	0.000000
25%	201904.000000	0.000000	2.800000	37.520000	30.000000	0.000000	0.000000	77.865000	0.000000	0.000000	0.000000	8.412006e+06	0.000000
50%	201907.000000	0.000000	10.000000	135.000000	1321.000000	0.000000	0.000000	415.185000	0.000000	0.000000	0.000000	2.132109e+07	0.000000
75%	201910.000000	0.000000	25.300000	520.000000	1988.000000	0.000000	0.000000	2804.085000	1.000000	0.000000	0.000000	2.918217e+07	0.000000
max	201912.000000	1.000000	100.000000	8540.000000	15616.000000	11.000000	1.000000	143268.550000	1.000000	1.000000	1.000000	3.808006e+07	1.000000

1 year sample

```
df.TARGET.value_counts()
```

```
0    11049
1      120
Name: TARGET, dtype: int64
```

High class unbalance
1% not zero

Exploratory Data Analysis

Summary → Missing data Fraction

```
df.isna().mean()
```

SAFRA	0.000000
V1	0.065539
V2	0.020324
V3	0.000000
V4	0.000000
V5	0.081117
V6	0.000000
V7	0.014415
V8	0.031158
V9	0.010028
V10	0.000000
V11	0.000000
V12	0.000000
CEP	0.000000
TARGET	0.000000

6 variables with missing data

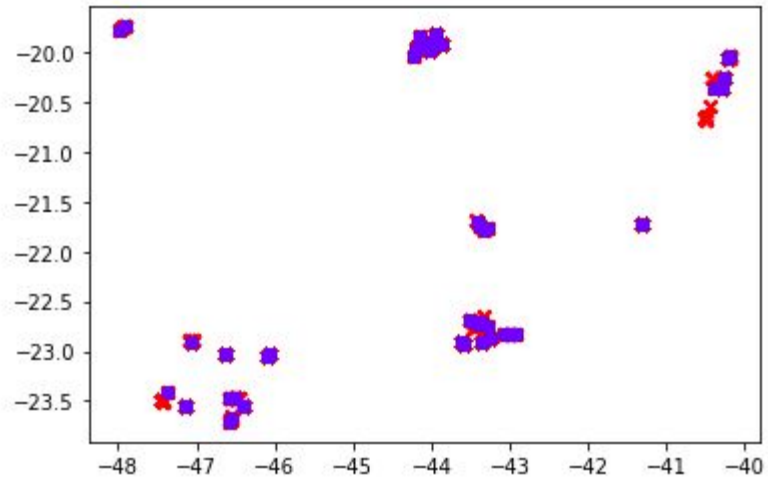
Imputing strategy:

- Similarity imputing
- By Most frequent value

Exploratory Data Analysis

Add Geolocation → Plot Data Distribution

Latitude

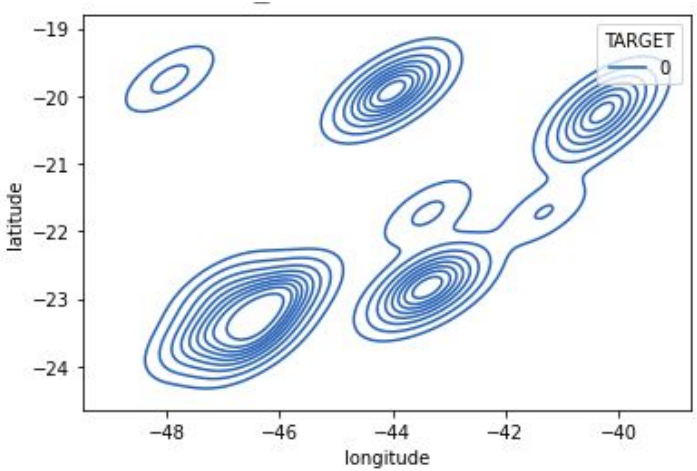


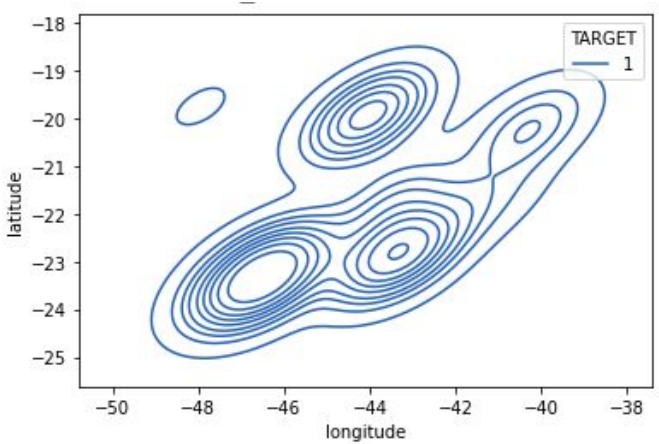
longitude

We can observe that the 11k data points are concentrated to only 100 locations (CEPS).
TARGET =1 are the red crosses. Few locations where theses stand out.

Exploratory Data Analysis

Add Geolocation → Plot Data Distribution





There are 5 main centers which the data is distributed and we can see qualitative differences for the Target Variable

Methods

Data Treatment

We use **smote** to generate new samples based on the minority class group to balance class from 1:100 to 1:1.

We imput the using the most frequent strategy

Since all selected columns are numeric we scale and center the data.

Cross Validation

We use 80% of the data to train the model and let 20%for validation.

We perform a 5 fold cross validation in the remaining 80%.

The validation data preserves the original class ratio 1:100

Metrics & Models

Metrics

- Precision
- Recall
- AVG Precision

Models

- LDA
- QDA
- XGBOOST

Results

Metrics

QDA seems to have the best performance without overfitting the data. We should conduct a grid search for XBoostedTrees to avoid the overfitting. This overfitting comes from the SMOTE process and the higher complexity of XGB model.

	name	avg_precision	avg_precision_std	avg_precision_val	precision_validation	recall_validation
0	LDA	0.797409	0.006623	0.038239	0.019481	0.500000
1	QDA	0.816682	0.006435	0.041235	0.041667	0.500000
2	xgb	0.910095	0.004440	0.014324	0.027933	0.208333

Thank you

Repository:
git@github.com:marceloosg/testcase.git