Title

Created by Khabibullin Salikh
Maybe seen by Dji
s.khabibullin@innopolis.university
g.dlamini@innopolis.university
gcinzoe04@gmail.com

1 Motivation

In classification I must built model, which will predict class with minimum of Loss function (mistakes). In regression I must built model, which will predict exactly target feature. Also I must do this ML Task, that Dji will think, that Salikh is not a bad student.

2 Data

In both tasks there are too many outlies. In classification there is strong imbalance in data. So main goal is to take this moment into account. In regression given features help a little to predict a target. So I think PyTorch is more useful here.

3 Exploratory data analysis

I calculate correlation matrix and drop some dependent features. I also could built some categorical features instead of dropping them. It's also important to choose right metrics, to validate and val(or test) data. It's also important to choose right metrics, to validate and val(or test) data.

4 Task

Classification:

I used metrics: "Roc-Auc", "Accuracy", "Precision", "Recall". As I understood there are imbalance in data. Class "0" is 13.6 larger than Class "1". I chose undersampling. Of course, I could use libraries like TSNE, but they help a little with real data. I used LogisticRegression with different data.

Regression:

I didn't use features "bitrate-mean", "bitrate-std" because we predict bitrate)) I used metrics: "MAE", "R2", "RMSE". I used LinearRegression, PolynomialFeatures, Lasso, Ridge, RandomForestRegressor, CatBoostRegressor.

5 Results

In Classification task best model was LogisticRegression with class-weight = "balanced" and penalty = 'l2'.

In Regression task best model was CatBoostRegressor (other models gave same results)

I have 2 pictures which showw my shy results. I will put them on github.

Table 1. For Classification

Model	Roc-Auc	Acc.	Precision	Recall
Model 1	0.569	0.941	0.714	0.142
Model 2	0.563	0.940	0.707	0.130
Model 3	0.563	0.940	0.707	0.130
Model 4	0.709	0.858	0.236	0.539
Model 5	0.709	0.857	0.236	0.539
Model 6	0.709	0.857	0.236	0.539
Model 7	0.710	0.874	0.262	0.521
Model 8	0.710	0.874	0.262	0.521
Model 9	0.710	0.874	0.262	0.521

6 Data Imbalance

It's important to win Imbalance moment because in opposite situation it will have strong affect on model and this model will be bad.

7 Conclusion

I built different models and chose the best of them. If I know more about data I could built some else features which could help my model.

You can listen song .mp3, which I downlowd on github.