feedzai

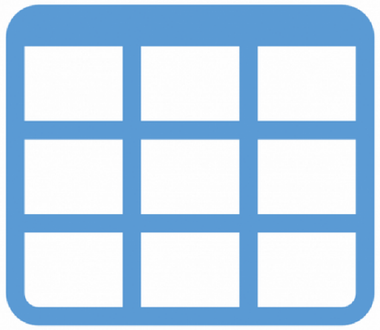
TÉCNICO LISBOA

Tabular Data Normalization for DL

Rita Leite, nº 92646

TABULAR DATA

- Data organized in rows and columns
- Most common type of data
- Heterogeneous by nature



CHALLENGES

- Data Quality
- What to do with Categorical Features?
    1. One-hot Encoding
    2. Label Encoding
- What about Numerical features?

ROADMAP

NORMALIZATION FOR NUMERICAL FEATURES

Z-score Normalization
Robust Normalization
MinMax Normalization

EMBEDDINGS

Embeddings for Categorical Features
Embeddings for Numerical Features

ENCODINGS FOR CATEGORICAL FEATURES

Label Encoding
Target Encoding
Count Enconding
CatBoost Encoding

SPECIALIZED ARCHITECTURES

TabTransformer

DATA DESCRIPTION

BAF

Bank Account Fraud Detection

- Feedzai's publicly available dataset
- Each column corresponds to an online Bank Account opening application
- 5 out of the 31 variables are categorical
- Prevalence of 1.025% on training set

CTR

Click Through Rate Prediction

- All 21 features are categorical
- Some of them contain over thousands of possible categories
- Prevalence on training set of 16.50%

Fraud

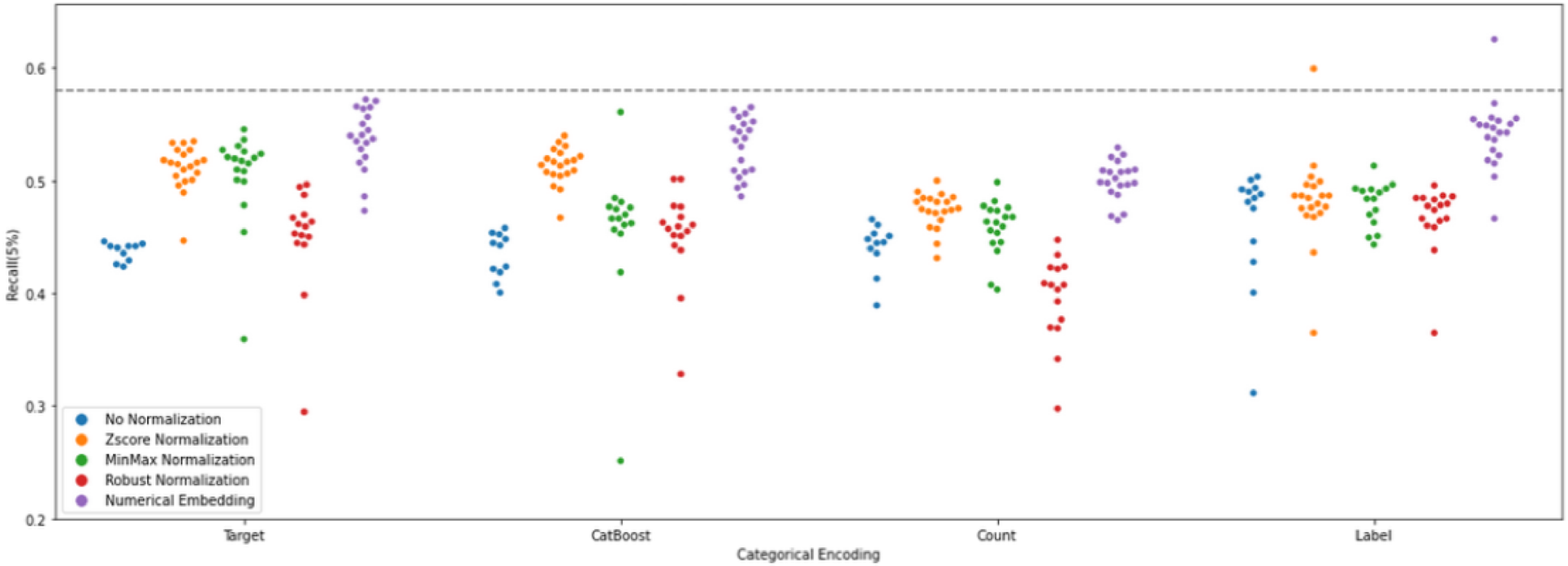
Credit Card Fraud Detection

- Features are result of PCA on original data
- 29 continuous features
- Prevalence of 0.18% on training set

EXPERIMENT SET UP

- LightGBM was chosen as the baseline method
- Random Search for Hyperparameters
- An MLP is used as the base for the normalization techniques
- The selection of the best models was done according to the AUC

RESULT ANALYSIS



Swarmplot showing the Recall on a 5% FPR per type of categorical encoding. Results shown are for the test set of the BAF dataset. Dashed line represents recall for best LightGBM model on test set

RESULT ANALYSIS

BAF

Model	AUC	Recall(5%)
LightGBM	0.89902	0.57983
Best MLP	0.896530	0.57073
TabTransformer	0.87952	0.507
MLP w/ Cat. Embeddings	0.86556	0.47549

CTR

Model	AUC	Recall(5%)
TabTransformer	0.74296	0.20425
LightGBM	0.74151	0.213
Best MLP	0.74125	0.20589
MLP w/ Cat. Embeddings	0.73988	0.21022

ROCAUC for test set and Recall with FPR at 5%

1. Categorical Embeddings

Embeddings are more effective when features are more informative

RESULT ANALYSIS

BAF

Model	AUC	Recall(5%)
LightGBM	0.89902	0.57983
Best MLP	0.896530	0.57073
TabTransformer	0.87952	0.507
MLP w/ Cat. Embeddings	0.86556	0.47549

CTR

Model	AUC	Recall(5%)
TabTransformer	0.74296	0.20425
LightGBM	0.74151	0.213
Best MLP	0.74125	0.20589
MLP w/ Cat. Embeddings	0.73988	0.21022

ROCAUC for test set and Recall with FPR at 5%

1. Categorical Embeddings

   Embeddings are more effective when features are more informative

2. TabTransformer

   The TabTransformer is not effective when dealing with low dimensional categorical features, most likely due to its complexity

RESULT ANALYSIS

BAF

LightGBM w/ Num. Embeddings	Categorical Encoding	AUC	Recall(5%)
	Label Encoding	0.8994	0.5843
	CatBoost Encoding	0.89852	0.57283
	Target Encoding	0.89839	0.57563
	Count Encoding	0.89668	0.57353
	LightGBM	0.89902	0.57983

ROCAUC for test set and Recall with FPR at 5%

1. Categorical Embeddings

   Embeddings are more effective when features have higher cardinality

2. TabTransformer

   The TabTransformer is not effective when dealing with low dimensional categorical features, most likely due to its complexity

3. LightGBM

   The LightGBM does not benefit from performing data transformation beforehand

CONCLUSION

Supervised methods are a great way to create meaningful representations of both numerical and categorical features.
With the right data preprocessing and normalization techniques, a simple MLP is enough to match the performance of LightGBM
No solution fits all!

feedzai

TÉCNICO LISBOA

Questions