EUROPEAN CENTRAL BANK

EUROSYSTEM

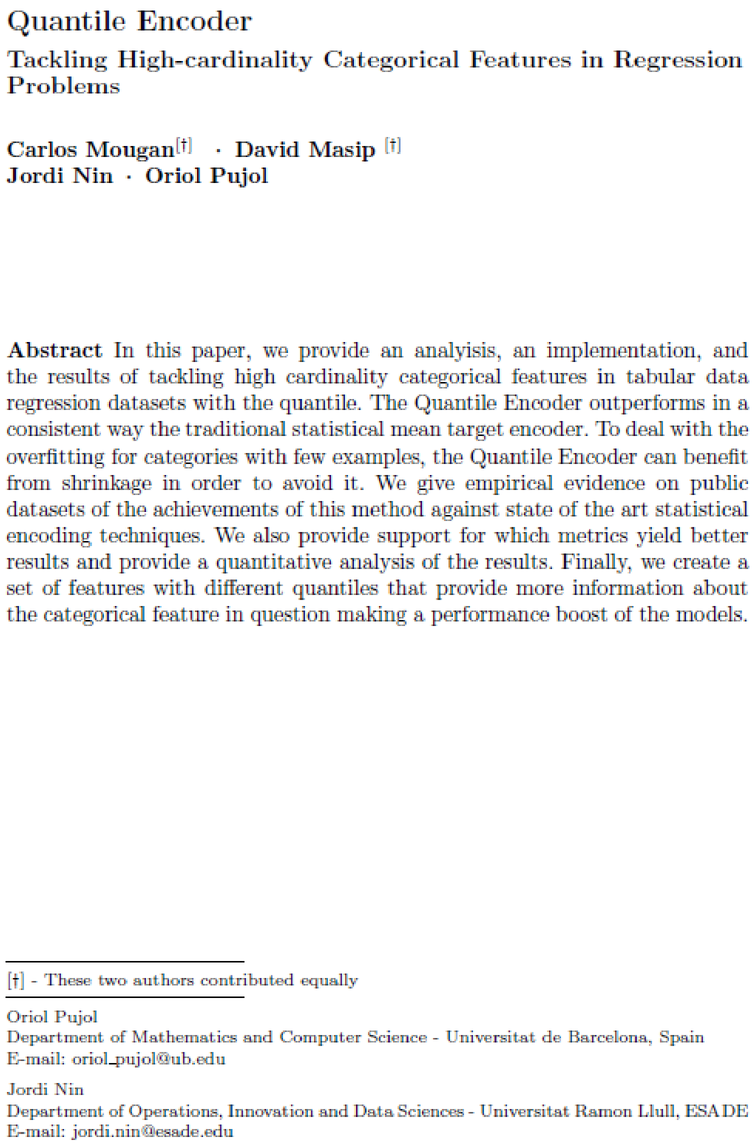
Quantile Encoder:

Tackling high cardinality categorical features in supervised learning

Researchers:
Carlos Mougan
David Masip - realxdata
Jordi Nin - ESADE
Oriol Pujol - University of Barcelona

Goals of the presentation

- Overview of a Machine Learning **pipeline**.
- Review of the most **common techniques to handle categorical data**.
- **Sktools** library
- State of the art on **features with high cardinality**.
- Introduction to **Quantile Encoder**.



Quantile Encoder

Tackling High-cardinality Categorical Features in Regression Problems

Carlos Mougan[†] · David Masip [†]
Jordi Nin · Oriol Pujol

Abstract In this paper, we provide an analyisis, an implementation, and the results of tackling high cardinality categorical features in tabular data regression datasets with the quantile. The Quantile Encoder outperforms in a consistent way the traditional statistical mean target encoder. To deal with the overfitting for categories with few examples, the Quantile Encoder can benefit from shrinkage in order to avoid it. We give empirical evidence on public datasets of the achievements of this method against state of the art statistical encoding techniques. We also provide support for which metrics yield better results and provide a quantitative analysis of the results. Finally, we create a set of features with different quantiles that provide more information about the categorical feature in question making a performance boost of the models.

[†] - These two authors contributed equally

Oriol Pujol
Department of Mathematics and Computer Science - Universitat de Barcelona, Spain
E-mail: oriol_pujol@ub.edu

Jordi Nin
Department of Operations, Innovation and Data Sciences - Universitat Ramon Llull, ESADE
E-mail: jordi.nin@esade.edu

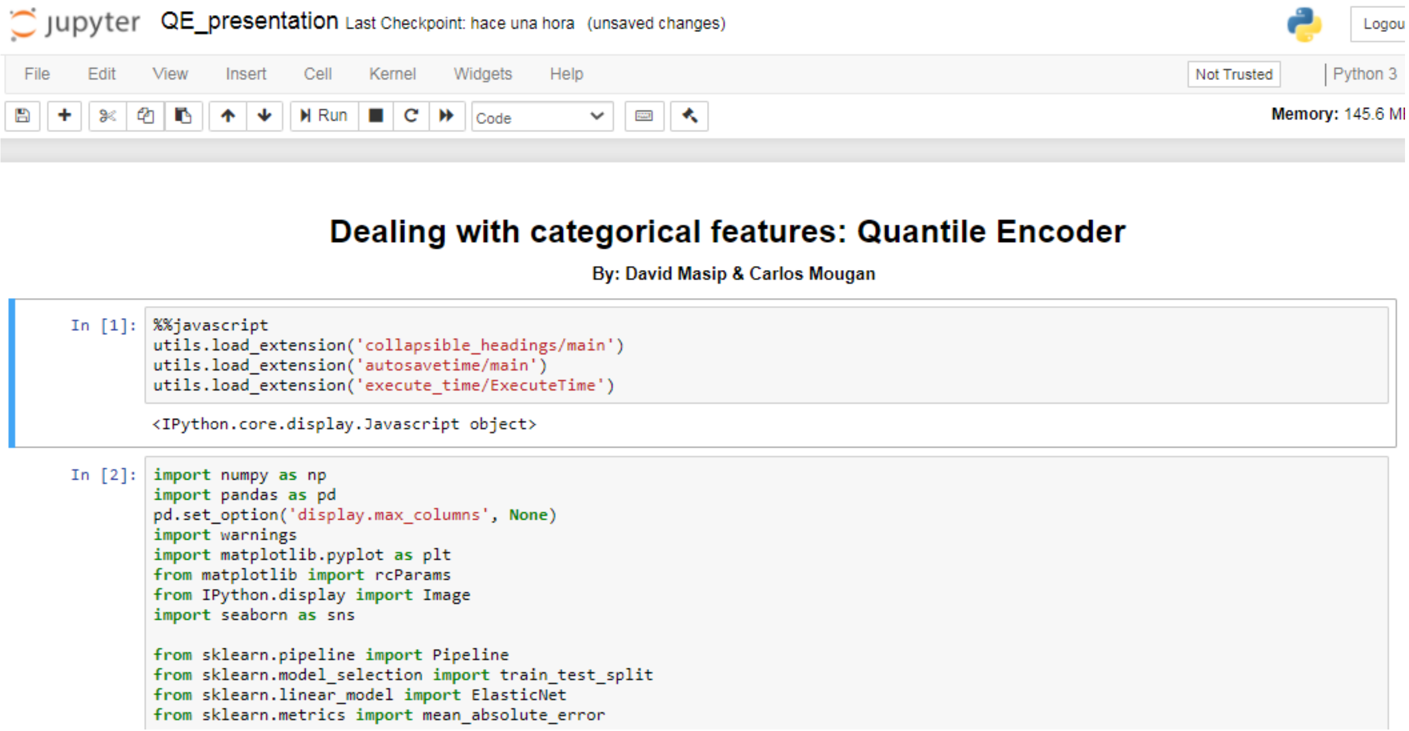
1. Quantile Encoder

- In review for **The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases –** ECML PKDD 2021

- **Paper** at https://darwin.escb.eu/livelink/livelink/app/nodes/306435319

- **Sktools** library**:** https://sktools.readthedocs.io/

Presentation on Github



Dealing with categorical features: Quantile Encoder

By: David Masip & Carlos Mougan

In [1]:

```
%%javascript
utils.load_extension('collapsible_headings/main')
utils.load_extension('autosavetime/main')
utils.load_extension('execute_time/ExecuteTime')
```

```
<IPython.core.display.Javascript object>
```

In [2]:

```
import numpy as np
import pandas as pd
pd.set_option('display.max_columns', None)
import warnings
import matplotlib.pyplot as plt
from matplotlib import rcParams
from IPython.display import Image
import seaborn as sns

from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.linear_model import ElasticNet
from sklearn.metrics import mean_absolute_error
```

Conclusions

Presentation takeaways

- How to **deal with categorical features** in supervised learning
- Machine Learning **pipeline** example
- **Sktools & category encoders** python libraries

Paper contributions

- **Encoding**: Quantile Encoder
- **Optimization**: Not all encodings are optimal for all metrics and loss functions
- **Feature engineering**: Set of features quantile features (Summary Encoder)

Quantile Encoder

Finish

Feel free to drop any questions