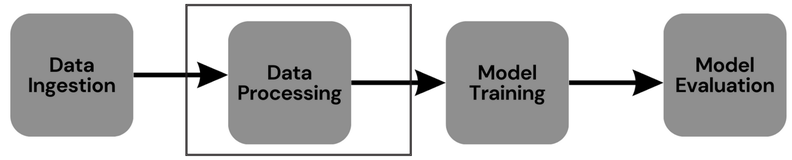
Automatic feature discovery to improve Machine Learning performance

Encoding Methods for Categorical Data: A Comparative Analysis for Linear Models, Decision Trees, and Support Vector Machines

Author: Andrei Udila
a.udila@student.tudelft.nl

Supervisors: Andra Ionescu
Asterios Katsifodimos

01 Introduction



Categorical data - describes characteristics and qualities
Not directly comparable / measurable

We need to encode it to numerical data
How?

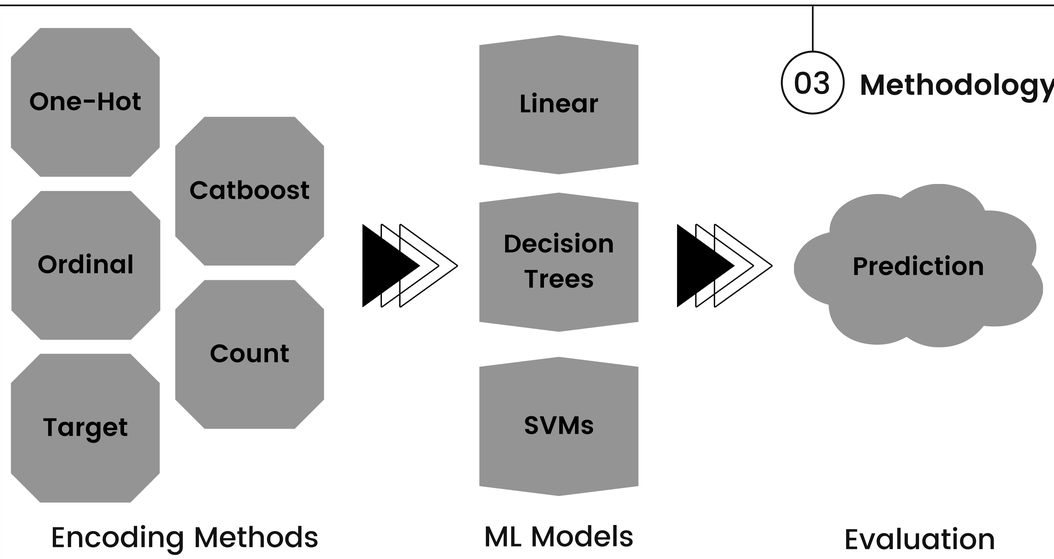
02 Objective

How can encoding categorical data improve ML performance?

Which encoding technique is most effective for which ML algorithms?

How do the results compare in terms of prediction performance and efficiency?

03 Methodology



04 Experiments

Encoder vs Encoder

Encoder + Encoder

Encoder vs AutoML

Trained ML models using one or multiple encoding methods at a time

Trained ML models with automatic data encoding by AutoGluon

05 Results

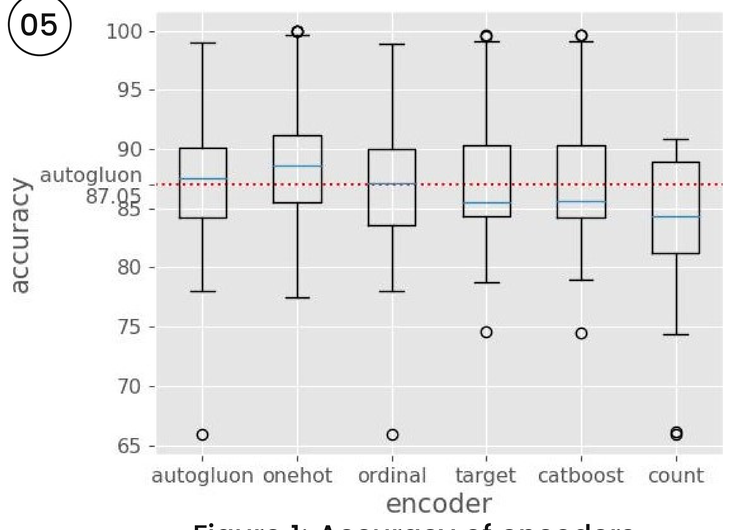


Figure 1: Accuracy of encoders

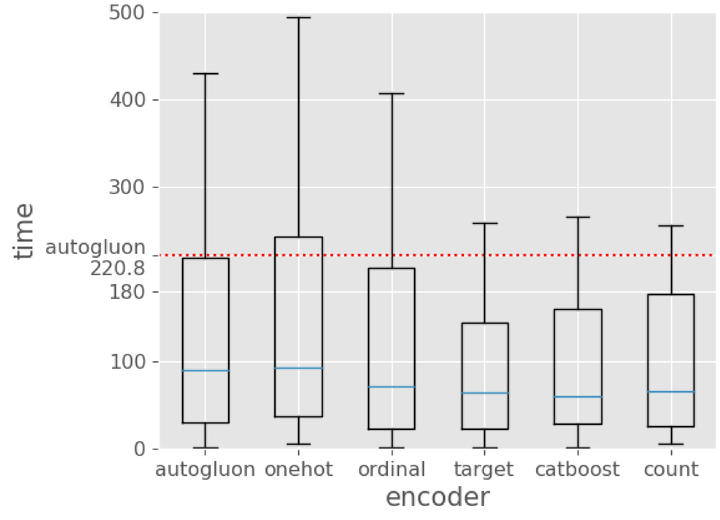


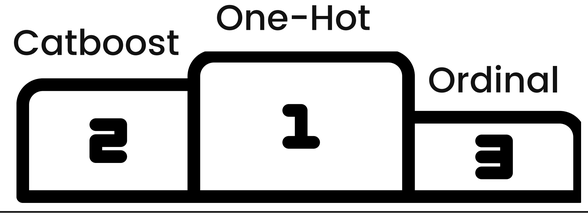
Figure 2: Runtime of encoders

06 Discussion of Results

One-Hot Encoding performs **best** in terms of **accuracy.**

Ordinal and **Catboost** perform **best** when considering both **accuracy** and **runtime.**

Combining encoders yields **worse** results than using a **single encoder.**



Related Literature
[1] Diogo Seca and Jo˜ao Mendes-Moreira. Benchmark of encoders of nominal features for regression. In Trends and Applications in Information Systems and Technologies pages 146–155, Cham, 2021. Springer International Publishing.
[2]Florian Pargent, Bernd Bischl, and Janek Thomas. A benchmark experiment on how to encode categorical features in predictive modeling. PhD thesis, Master Thesis in Statistics, Ludwig-Maximilians-Universitat Munchen, 2019
[3] Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. Bachelor's thesis, KTH, School of Electrical Engineering and Computer Science(EECS), 2018.
[4]Florian Pargent, Florian Pfisterer, Janek Thomas, and Bernd Bischl. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics, 37(5):2671–2692, mar 2022

TUDelft