Supplementary Material:
Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features

1 Implementations

Feature encoding techniques are implemented in several general preprocessing packages. We provide an overview of existing implementations for R in Table 1. For python, an extension to scikit-learn (Pedregosa et al., 2011), *category encoders* (McGinnis et al., 2018) is available. To enable a fair and reliable comparison in our study, we implemented all methods outlined above on top of the `mlrCPO` package. All code can be found in our **online repository** (https://github.com/compstat-lmu/paper_2021_categorical_feature_encodings).

encoding method	regularization	package

indicator	–	`stats`
		`embed`
		`mlrCPO`
		`mlr3pipelines`

hash	–	`FeatureHashing` `embed`

impact	smoothing	`embed`
		`mlrCPO`
		`mlr3pipelines`

impact	cross-validation	`vtreat`
		`mlr3pipelines (via vtreat)`

regularized impact (glmm)	–	`embed` `mlr3pipelines`

Table 1: Overview over existing encoding implementations in R. Implementations can deviate from the algorithms described in our manuscript by minor implementation details. Indicator encoding encloses a variety of methods (one-hot, dummy, helmert encoding etc.)

2 Pseudocode for all encoding strategies

Algorithm 1 to 10 contain the pseudocode for all encoding strategies studied in our manuscript in order to improve reproducibility and to provide further hints towards subtle differences in encoders and implementations.

Algorithm 1 Integer Encoding

```
Training:
compute random permutation $\boldsymbol{int} = (int_1, \ldots, int_k, \ldots, int_L)^T$ of $(1, \ldots, L)^T$
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do $\hat{x}_i^{train} = int_k$ with $x_i^{train} = l_k$, $k = 1, \ldots, L$
Prediction:
for $x^{new}$ do
    if $x^{new} \in \mathcal{L}^{train}$ then $\hat{x}^{new} = int_k$ with $x^{new} = l_k$, $k = 1, \ldots, L$ else $\hat{x}^{new} = NA$
```

Algorithm 2 Frequency Encoding

```
Training:
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do $\hat{x}_i^{train} = \frac{N_l}{N}$ with $x_i^{train} = l$
Prediction:
for $x^{new}$ do
    if $x^{new} \in \mathcal{L}^{train}$ then $\hat{x}^{new} = \frac{N_l}{N}$ with $x^{new} = l$ else $\hat{x}^{new} = 1$
```

Algorithm 3 One-Hot Encoding

```
Training:
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
    for all $l \in \mathcal{L}^{train}$ do $\hat{x}_{il}^{train} = I(x_i^{train} = l)$
Prediction:
for $x^{new}$ do
    for all $l \in \mathcal{L}^{train}$ do
        if $x^{new} \in \mathcal{L}^{train}$ then $\hat{x}_l^{new} = I(x^{new} = l)$ else $\hat{x}_l^{new} = 0$
```

Algorithm 4 Dummy Encoding

```
Training:
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
    for all $l \in \mathcal{L}^{train} \setminus l_{ref}$ do $\hat{x}_{il}^{train} = I(x_i^{train} = l)$
Prediction:
for $x^{new}$ do
    for all $l \in \mathcal{L}^{train} \setminus l_{ref}$ do
        if $x^{new} \in \mathcal{L}^{train}$ then $\hat{x}_l^{new} = I(x^{new} = l)$ else $\hat{x}_l^{new} = NA$
```

Algorithm 5 Hash Encoding

```
Training: require $hash.size \in \mathbb{N}$
for all $l \in \mathcal{L}^{train}$ do $ind_l = (hash(l) \mod hash.size) + 1$, $ind_l \in \mathbb{N}$, $hash(l) \in \mathbb{N}$
1. define matrix $\boldsymbol{D}^{N \times hash.size}$ with $d_{ih} = 1$ if $ind_l = h$ and $d_{ih} = 0$ if $ind_l \neq h$, $x_i^{train} = l$
2. $\boldsymbol{D} \rightarrow \tilde{\boldsymbol{D}}^{N \times V}$ with $V \leq hash.size$, by dropping constant columns in $\boldsymbol{D}$
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do $\hat{x}_{iv}^{train} = \tilde{d}_{iv}$
Prediction:
for $x^{new}$ do
    $ind^{new} = (hash(x^{new}) \mod hash.size) + 1$
    $\boldsymbol{d}^{new}$ of length $hash.size$ with $d_h^{new} = 1$ if $ind^{new} = h$ and $d_h^{new} = 1$ if $ind^{new} \neq h$
    $\boldsymbol{d}^{new} \rightarrow \tilde{\boldsymbol{d}}^{new}$ of length $V$, by dropping columns which were constant in $\boldsymbol{D}$
    for all $V$ columns in $\tilde{\boldsymbol{D}}$ do $\hat{x}_v^{new} = \tilde{d}_v^{new}$
```

Algorithm 6 Leaf Encoding

```
Training: require number of cross-validation folds $K \in \mathbb{N}$
fit CART tree on $\mathcal{D}^{train}$ with complexity pruning based on $K$-fold cross-validation
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do $\tilde{x}_i = t$ with $x_i^{train}$ in terminal node $t$
Prediction:
for $x^{new}$ do
    if $x^{new} \in \mathcal{L}^{train}$ then $\tilde{x}^{new} = t$ with $x^{new}$ in terminal node $t$
    else $\tilde{x}^{new} = b$ where $b$ indicates the biggest terminal node
```

Algorithm 7 Impact Encoding Regression

```
Training: require smoothing parameter $\epsilon \in \mathbb{R}$
for all $l \in \mathcal{L}^{train}$ do $\delta_l = \frac{\sum_{i:x_i^{train}=l} y_i^{train} + \epsilon \cdot \bar{y}^{train}}{N_l + \epsilon} - \bar{y}^{train}$ with $\bar{y}^{train} = \frac{\sum_{i=1}^{N} y_i^{train}}{N}$
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do $\hat{x}_i^{train} = \delta_l$ with $x_i^{train} = l$
Prediction:
for $x^{new}$ do
    if $x^{new} \in \mathcal{L}^{train}$ then $\hat{x}^{new} = \delta_l$ with $x^{new} = l$ else $\hat{x}^{new} = 0$
```

Algorithm 8 Impact Encoding Classification

```
Training: require smoothing parameter $\epsilon \in \mathbb{R}$
for all $c \in \mathcal{C}$ do
    $p_c = \frac{N_c}{N}$, $p_c^{new} = \frac{N_c + \epsilon}{N + 2\epsilon}$, $logit_c = \log(\frac{p_c}{1 - p_c})$, $logit_c^{new} = \log(\frac{p_c^{new}}{1 - p_c^{new}})$
    $\delta_c^{new} = logit_c^{new} - logit_c$
    for all $l \in \mathcal{L}^{train}$ do
        $p_{lc} = \frac{\sum_{i:x_i^{train}=l} I(y_i^{train} = c) + \epsilon}{N_l + 2\epsilon}$, $logit_{lc} = \log(\frac{p_{lc}}{1 - p_{lc}})$
        $\delta_{lc} = logit_{lc} - logit_c$
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do $\hat{x}_{ic}^{train} = \delta_{lc}$ with $x_i^{train} = l$
Prediction:
for $x^{new}$ do
    for all $c$ in $\mathcal{C}$ do
        if $x^{new} \in \mathcal{L}^{train}$ then $\hat{x}_c^{new} = \delta_{lc}$ with $x^{new} = l$ else $\hat{x}_c^{new} = \delta_c^{new}$
```

Algorithm 9 GLMM Encoding Regression

```
Training: require $n.folds \in \mathbb{N}$
fit simple random intercept model: $y_i^{train} = \beta_{0l} + \epsilon_i = \gamma_{oo} + u_l + \epsilon_i$ on $\mathcal{D}^{train}$
with $u_l \overset{iid}{\sim} N(0, \tau^2)$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $x_i^{train} = l$, $l \in \mathcal{L}^{train}$
if $n.folds = 1$ then
    for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
        $\hat{x}_i^{train} = \hat{\beta}_{0l}^{\mathcal{D}^{train}}$ with $x_i^{train} = l$
else use $n.folds$ cross-validation scheme to make training sets $\mathcal{D}_1^{train}, \ldots, \mathcal{D}_{n.folds}^{train}$
    and fit simple random intercept model on each $\mathcal{D}_m^{train}$
    for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
        $\hat{x}_i^{train} = \hat{\beta}_{0l}^{\mathcal{D}^{train}}$ with $x_i^{train} = l$ based on the model m
        with $(x_i^{train}, y_i^{train}) \notin \mathcal{D}_m^{train}$
Prediction:
for $x^{new}$ do
    if $x^{new} \in \mathcal{L}^{train}$ then
        $\hat{x}^{new} = \hat{\beta}_{0l}^{\mathcal{D}^{train}}$ with $x^{new} = l$ based on full model fitted on $\mathcal{D}^{train}$
    else $\hat{x}^{new} = \hat{\gamma}_{00}$ based on full model fitted on $\mathcal{D}^{train}$
```

Algorithm 10 GLMM Encoding Binary Classification

```
Training: require $n.folds \in \mathbb{N}$
fit simple glmm: $E(y_i^{train}) = h(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$, $\eta_i = \beta_{0l} = \gamma_{00} + u_l$ on $\mathcal{D}^{train}$
with $u_l \overset{iid}{\sim} N(0, \sigma^2)$, $y_i^{train} \overset{ind}{\sim} Be(h(\eta_i))$ and $x_i^{train} = l$, $l \in \mathcal{L}^{train}$
if $n.folds = 1$ then
    for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
        $\hat{x}_i^{train} = \hat{\beta}_{0l}^{\mathcal{D}^{train}}$ with $x_i^{train} = l$
else use $n.folds$ cross-validation scheme to make training sets $\mathcal{D}_1^{train}, \dots, \mathcal{D}_{n.folds}^{train}$
    and fit simple glmm on each $\mathcal{D}_m^{train}$
    for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
        $\hat{x}_i^{train} = \hat{\beta}_{0l}^{\mathcal{D}^{train}}$ with $x_i^{train} = l$ based on the model $m$
        with $(x_i^{train}, y_i^{train}) \notin \mathcal{D}_m^{train}$
Prediction:
for $x^{new}$ do
    if $x^{new} \in \mathcal{L}^{train}$ then
        $\hat{x}^{new} = \hat{\beta}_{0l}^{\mathcal{D}^{train}}$ with $x^{new} = l$ based on full model fitted on $\mathcal{D}^{train}$
    else $\hat{x}^{new} = \hat{\gamma}_{00}$ based on full model fitted on $\mathcal{D}^{train}$
```

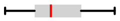
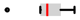
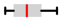
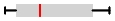
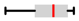
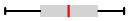
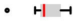
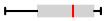
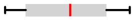
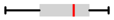
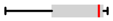
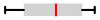
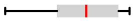
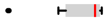
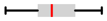
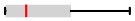
Algorithm 11 GLMM Encoding Multiclass Classification

```
Training: require $n.folds \in \mathbb{N}$
define response matrix $\boldsymbol{Y}^{N \times C}$ with $y_{ic} = 1$ if $y_i^{train} = c$ and $y_{ic} = 0$ if $y_i^{train} \neq c$:
$\tilde{\mathcal{D}}^{train} = \{(x_i^{train}, y_{i1}^{train}, \dots, y_{iC}^{train}), \dots, (x_N^{train}, y_{N1}^{train}, \dots, y_{NC}^{train})\}$
for all $C$ classes do
    fit simple glmm: $E(y_{ic}^{train}) = h(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$, $\eta_i = \beta_{0l} = \gamma_{00} + u_l$ on $\boldsymbol{y}_c^{train}$ from $\tilde{\mathcal{D}}^{train}$
    with $u_l \overset{iid}{\sim} N(0, \sigma^2)$, $y_{ic}^{train} \overset{ind}{\sim} Be(h(\eta_i))$ and $x_i^{train} = l$, $l \in \mathcal{L}^{train}$
if $n.folds = 1$ then
    for all $C$ models do
        for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
            $\hat{x}_{ic}^{train} = \hat{\beta}_{0l}^{\tilde{\mathcal{D}}^{train}}$ with $x_i^{train} = l$
else use $n.folds$ cross-validation scheme to make training sets $\tilde{\mathcal{D}}_1^{train}, \dots, \tilde{\mathcal{D}}_{n.folds}^{train}$
    for all $C$ classes do
        fit simple glmm on $\boldsymbol{y}_c^{train}$ from each $\tilde{\mathcal{D}}_m^{train}$
        for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
            $\hat{x}_{ic}^{train} = \hat{\beta}_{0l}^{\tilde{\mathcal{D}}^{train}}$ with $x_i^{train} = l$ based on the model $m$ for class $c$
            with $(x_i^{train}, y_i^{train}) \notin \tilde{\mathcal{D}}_m^{train}$
Prediction:
for $x^{new}$ do
    for all $C$ class models do
        if $x^{new} \in \mathcal{L}^{train}$ then
            $\hat{x}_c^{new} = \hat{\beta}_{0l}^{\tilde{\mathcal{D}}^{train}}$ with $x^{new} = l$ based on full model for class $c$
        else $\hat{x}_c^{new} = \hat{\gamma}_{00}$ based on full model for class $c$
```

Note on GLMM Encoders: To speed up the computation of the GLMM encoders, we followed the performance tips from the lme4 vignette (https://cran.r-project.org/web/packages/lme4/vignettes/lmerperf.html): we did not compute derivations (`calc.derivs = FALSE`) and used the `NLOPT_LN_BOBYQA` optimizer from the nloptr package with liberal stopping criteria (`maxeval = 1000`, `xtol_abs = 1e-6`, `ftol_abs = 1e-6`).

3 Benchmark Datasets

Table 2: Benchmark Datasets and Dataset Characteristics

OmlId	Name	Cl	N	NA%	Num	Bin	Cat	HighCardLevels	Entropy
41211	ames-housing	0	2930	0.00	34	2	44	10,10,16,16,17,28	
41445	employee_salaries	0	9228	0.03	1	2	3	37,385,694	
41210	avocado-sales	0	18249	0.00	8	1	2	54	
41437	wine-reviews	0	129971	6.22	1	0	6	19,43,425,707,1229,16757	
41444	medical_charges	0	163065	0.00	1	0	5	51,100,306,1977,3201	
41267	particulate-matter-ukair-2017	0	394299	0.00	4	0	5	12,30,53	
41251	flight-delay-usa-dec-2017	0	457892	0.00	2	0	7	12,31,52,52,294,294	
41255	nyc-taxi-green-dec-2016	0	1224158	0.00	4	3	7	241,260	
41283	churn	2	5000	0.00	14	2	3	10,51	
981	kdd_internet_usage	2	10108	0.39	0	48	20	10,11,46,77,119,129	
4135	Amazon_employee_access	2	32769	0.00	0	0	9	67,128,177,343,343,449,2358,4243,7518	
41434	Click_prediction_small	2	39948	0.00	3	0	6	6064,19228,19803,22381,25321,30114	
1590	adult	2	48842	0.95	6	1	7	14,16,41	
1114	KDDCup09_upselling	2	50000	68.50	174	4	30	14,...,5073,5713,13990,15415,15415	
41162	kick	2	72983	6.39	14	3	15	12,16,33,37,74,134,153,863,1063	
41442	open_payments	2	73354	22.61	0	1	4	513,1460,2255,4365	
41447	road-safety-drivers-sex	2	233964	10.78	2	1	3	380,20397	
41224	porto-seguro	2	595212	2.45	27	23	8	10,12,18,104	
188	eucalyptus	5	736	3.20	14	0	5	12,14,16,27	
41446	Midwest_survey	10	2778	1.66	0	21	5	1008	
41212	hpc-job-scheduling	4	4331	0.00	5	0	2	14	
41216	video-game-sales	12	16598	0.25	6	0	2	31,578	
41440	okcupid-stem	3	50789	15.97	2	1	16	12,12,15,15,18,32,45,48,184,208,7019	
4541	Diabetes130US	3	101766	0.00	13	9	25	10,10,18,73,717,749,790	

Note:
OmlId = Id on OpenML, Name = name on OpenML, Cl = classes (0: regression), N = observations, NA% = percentage of missing values, Num = numeric features, Bin = binary features, Cat = categorical features, HighCardLevels = number of levels for each categorical feature with at least 10 levels (some levels are not displayed for KDDCup09_upselling), Entropy = box-plot of normalized Shannon-entropy across levels (smaller = larger imbalance).

Table 3: Win Percentages of One-hot over Dummy Encoding

Learner	Binary Class	Multi Class	Regression
LASSO	80	94	75
RF	83	71	71
GB	63	71	54
KNN	67	59	50
SVM	68	82	52

Note. Percentage of encoding conditions in which one-hot performs better than dummy per task setting.

4 Comparison of One-Hot and Dummy Encoding

A frequently asked question is, whether one-hot encoding or dummy encoding is to be preferred. While dummy encoding clearly is commonly preferred for non-regularized linear models, the answer is less clear for general machine learning algorithms studied in our setting.

From the results shown in Table 3 we can observe, that for most algorithms and datasets, one-hot encoding is to be preferred over dummy-encoding.

5 Additional encoders not mentioned in the manuscript

In addition to the encoders mentioned in the manuscript, the benchmark results available in our online repository (https://github.com/compstat-lmu/paper_2021_categorical_feature_encodings) contain two additional experimental encoders that were newly developed. We do not mention them in our manuscript because they did not perform well, and in hindsight, we are not satisfied with their design. We briefly mention them here for transparency.

Algorithm 12 Cluster Encoding

```
Training: require number of desired levels $V \in \mathbb{N}$
if task is regression then
    for all $l \in \mathcal{L}^{train}$ do $\bar{y}_l^{train} = \frac{\sum_{i:x_i^{train}=l} y_i^{train}}{N_l}$
        define $\boldsymbol{v}_l = (\bar{y}_l^{train}, \delta_l)^T$ with $\delta_l = \left| N_l - \frac{\sum_{k=1}^{L} N_k}{L} \right|$
if task is classification then
    for all $l \in \mathcal{L}^{train}$ do
        for all $c \in \mathcal{C}$ do $s_{lc} = \sum_{i:x_i^{train}=l} I(y_i^{train} = c)$
        define $\boldsymbol{v}_l = (s_{l1}, \dots, s_{lC}, \delta_l)^T$ with $\delta_l = \left| N_l - \frac{\sum_{k=1}^{L} N_k}{L} \right|$
1. compute distance matrix $\boldsymbol{D}^{L \times L}$ with $d_{jk} = ||\boldsymbol{v}_j - \boldsymbol{v}_k||$
2. fit hierarchical cluster analysis on $\boldsymbol{D}$
3. prune dendrogram to obtain $V$ combined levels

for all $x_i^{train} \in \boldsymbol{x}^{train}$ do $\tilde{x}_i^{train} = v$ with $x_i^{train}$ in leaf $v$
Prediction:
for $x^{new}$ do
    if $x^{new} \in \mathcal{L}^{train}$ then $\tilde{x}^{new} = v$ with $x^{new}$ in leaf $v$ else $\tilde{x}^{new} = NA$
```

Algorithm 13 RF Encoding Regression and Binary Classification

```
Training: require number of trees $B \in \mathbb{N}$
fit random forest with trees $T_1, ..., T_B$
for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
    if $x_i^{train}$ inbag for all $B$ trees then $\hat{x}_i^{train} = \frac{1}{B} \sum_{b=1}^{B} T_b(x_i^{train})$
    else $\hat{x}_i^{train} = \frac{1}{B_i^{OOB}} \sum_{b:\ x_i^{train}\ OOB\ T_b} T_b(x_i^{train})$
Prediction:
for $x^{new}$ do
    if $x^{new} \in \mathcal{L}^{train}$ then $\hat{x}^{new} = \frac{1}{B} \sum_{b=1}^{B} T_b(x^{new})$
    else $\hat{x}^{new} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{L} \sum_{l \in \mathcal{L}^{train}} T_b(l)$
```

Algorithm 14 RF Encoding Multiclass Classification

```
Training: require number of trees $B \in \mathbb{N}$
fit random forest with trees $T_1, ..., T_B$
for all $C$ classes do
    for all $x_i^{train} \in \boldsymbol{x}^{train}$ do
        if $x_i^{train}$ inbag for all $B$ trees then $\hat{x}_{ic}^{train} = \frac{1}{B} \sum_{b=1}^{B} T_b^c(x_i^{train})$
        else $\hat{x}_{ic}^{train} = \frac{1}{B_i^{OOB}} \sum_{b:\ x_i^{train}\ OOB\ T_b} T_b^c(x_i^{train})$
Prediction:
for $x^{new}$ do
    for all $C$ classes do
        if $x^{new} \in \mathcal{L}^{train}$ then $\hat{x}_c^{new} = \frac{1}{B} \sum_{b=1}^{B} T_b^c(x^{new})$
        else $\hat{x}_c^{new} = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{L} \sum_{l \in \mathcal{L}^{train}} T_b^c(l)$
```

5.1 Additional datasets not mentioned in the manuscript

In addition to the datasets mentioned in the manuscript, the benchmark results available in our online repository (https://github.com/compstat-lmu/paper_2021_categorical_feature_encodings) contain three additional datasets: *KDD98* (OmlId: 41435), *sf-police-incidents* (OmlId: 41436), *Traffic_violations* (OmlId: 41443). A substantive amount of conditions failed for those rather big datasets due to memory problems, which is why we remove them from the results discussed in our manuscript. We briefly mention them here to provide full transparency.

References

McGinnis, W., hbghhy, Tao, W., andrethrill, Siu, C., Davison, C., Bollweg, N., 2018. scikit-learn-contrib/categorical-encoding: Release for zenodo. Zenodo. doi:10.5281/zenodo.1157110

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others, 2011. Scikit-learn: Machine learning in python. Journal of machine learning research 12, 2825–2830.