A COMPARATIVE ANALYSIS OF NOMINAL ENCODING TECHNIQUES FOR BINARY CLASSIFICATION TASKS

Abstract

The effective encoding of categorical data is crucial in machine learning to capture their true importance and enhance model performance. This study presents a comprehensive analysis and comparison of various encoding strategies for nominal features. The chosen metric for evaluation involves training a supermodel, specifically LightGBM, which internally handles categorical variables and determines feature importance. Additionally, a baseline model, logistic regression without hyperparameter tuning, is trained using encoded data from each encoder. Coefficients are extracted from the baseline model, with higher magnitudes indicating greater feature importance. The performance of encoders is then assessed by comparing these coefficients with the feature importance of the supermodel.

This study demonstrates the efficacy of different encoders across three datasets. The findings of this study provide valuable insights and practical guidelines for practitioners in selecting appropriate encoding methods for high cardinality categorical features. Leave One Out encoder consistently emerged as a top performer, followed by James Stein, M-Estimate, and CatBoost encoders. These findings can be applied in various domains, including finance, marketing, and data analysis pipelines, to improve the accuracy and effectiveness of machine learning models that handle categorical data.

Keywords: Machine Learning, Categorical features, Nominal data, Binary Classification

Authors

A. Lalita Kumari
Department of MCA
V.E.S. Institute of Technology
Chembur, Mumbai, Maharashtra
India.
lalitakm.official@gmail.com

Dr. Dhanamma Shanar Jagli
Assistant Professor and
Deputy HoD
V.E.S. Institute of Technology
Mumbai, Maharashtra, India.
lalitakm.official@gmail.com

I. INTRODUCTION

A Machine Learning model can only understand numbers and in reality, target variables depend on multiple variables that can be numeric and categorical (nominal or ordinal) in structured classification tasks, however representing these categorical features as numeric and feeding it to the algorithm is an art that is in progress since decades. Some papers address which encoding to go for when dealing with which type of Problem statements but there is very little evidence to show which one is the best for almost all scenarios.

In this research paper, we compare and analyze different encoding strategies within the context of machine learning. Our primary objective is to investigate the effectiveness of various encoding techniques and assess their impact on the predictive power of machine learning models. We aim to contribute to the existing knowledge and understanding of encoding strategies for categorical data through meticulous experimentation and rigorous analysis.

By conducting a comprehensive analysis, this study aims to provide valuable insights and practical guidelines for practitioners when selecting appropriate encoding methods tailored to their specific datasets. We recognize that the choice of encoding method can significantly affect computational efficiency and model performance. Hence, our research endeavors to shed light on the strengths and weaknesses of different encoding techniques, empowering practitioners to make informed decisions in their data analysis pipelines.

I. RELATED WORK

Sr. No.	Author	Title	Summary
1.	Please refer: [References-1]	Encoding Methods for Categorical Data: A Comparative Analysis for Linear Models, Decision Trees, and Support Vector Machines	This paper compares five encoding methods for categorical data (one-hot, ordinal, target, CatBoost, and count encoders) in the context of machine learning. The evaluation includes linear models, decision trees, and support vector machines (SVMs).
2.	Please refer: [References-2]	A Benchmark Experiment on How to Encode Categorical Features in Predictive Modeling	This paper evaluates encoding strategies for high cardinality features across five machine learning algorithms (lasso, random forest, gradient boosting, k-nearest neighbors, support vector machines) using datasets from regression, binary, and multiclass classification scenarios.

3.	Please refer: [References-3]	A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers	This paper compares seven categorical variable encoding techniques for classification using Artificial Neural Networks on the Car Evaluation dataset from UCI. The study finds that Sum Coding and Backward Difference Coding achieve the highest accuracy compared to other encoding techniques.

These three papers collectively contribute to the existing knowledge on encoding categorical data in machine learning. They explore the performance of various encoding methods, evaluate their effectiveness across different machine learning algorithms, and highlight the impact on predictive modeling tasks. These studies serve as valuable references to establish the context for further research in categorical data encoding and can assist researchers in selecting appropriate encoding techniques based on the specific machine learning algorithms and datasets they are working with.

The nominal encoding techniques under investigation (WoE, CatBoost Encoding, JamesStein Encoding, Leave-One-Out Encoding, and Mestimate Encoding) have unique characteristics that make them suitable for binary classification tasks. By proving their significance through empirical evaluation, this research contributes to the existing body of knowledge in the following ways:

1. **Addressing a Research Gap:** The previous research papers reviewed primarily focused on encoding methods for categorical data in general, without specific emphasis on nominal encoding techniques. This research fills this gap by specifically investigating and evaluating the performance of the selected nominal encoding methods for binary classification tasks.

2. **Comparative Analysis:** By comparing the performance of the nominal encoding techniques against each other, this research provides insights into the relative effectiveness of these methods. It offers a comprehensive analysis of their impact on binary classification accuracy.

3. **Extending the Applicability:** The previous research primarily focused on the performance of encoding techniques across specific machine learning algorithms. In contrast, this research investigates the significance of nominal encoding techniques across different algorithms, extending their applicability beyond a specific set of models.

4. **Practical Relevance:** Binary classification tasks are prevalent in various domains, such as fraud detection, disease diagnosis, and sentiment analysis. By evaluating the performance of the selected nominal encoding techniques specifically for binary classification, this research directly addresses the practical relevance of these encoding methods in real-world scenarios.

II. PROBLEM DEFINITION

The problem addressed in this paper is the lack of sufficient investigation of encoding techniques specifically tailored for classification tasks. While most research has extensively explored encoding techniques for regression tasks and neural networks, the literature on encoding techniques for classification tasks remains relatively sparse.

In classification tasks, the goal is to predict discrete class labels for input data, which presents distinct challenges compared to regression tasks. Unlike regression, where the objective is to predict continuous values, classification necessitates the identification and differentiation of distinct categories or classes. This distinction requires encoding techniques that can effectively capture and represent the discriminative information present in the input data.

Unfortunately, the existing encoding techniques, predominantly developed and optimized for regression tasks, may not be directly applicable to classification tasks. These techniques often rely on capturing and quantifying the magnitude of relationships between input features and the target variable, which is not sufficient for accurately predicting discrete class labels.

Thus, there is a pressing need to explore novel encoding techniques that are specifically tailored to classification tasks. By focusing on the unique requirements and challenges of classification, these techniques can capture and represent the intricate patterns, relationships, and discriminating factors within the input data, ultimately leading to improved classification accuracy and performance.

III. METHODOLOGY

1. This study makes use of 3 datasets. An Income dataset that has features as follows:

- **Age:** Represents the age of the individual in years.
- **Workclass:** Describes the individual's type of employment, such as Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, or Never-worked.
- **Education:** Indicates the highest level of education completed by the individual, ranging from Preschool to Doctorate.
- **Education-Num:** This represents the numerical mapping of the education level, where higher values indicate higher levels of education.
- **Marital Status:** Provides information about the individual's marital status, such as Married-civ-spouse, Divorced, Never-married, Separated, Widowed, or Married-spouse-absent.
- **Occupation:** Specifies the type of occupation the individual is engaged in, such as Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, etc.
- **Relationship:** Indicates the individual's role in the family, including Husband, Wife, Own-child, Not-in-family, Other-relative, or Unmarried.
- **Race:** Represents the individual's race or ethnicity, such as White, Asian-Pac-Islander, Amer-Indian-Eskimo, Black, or Other.

- **Sex:** Specifies the individual's gender as Male or Female.
- **Capital Gain:** Reflects the amount of capital gains reported by the individual.
- **Capital Loss:** Reflects the amount of capital losses reported by the individual.
- **Hours per week:** This represents the number of hours the individual works per week.
- **Native Country:** Indicates the individual's country of origin or citizenship.
- **Income:** The target variable, which indicates whether the individual's income exceeds $50,000 per year (<=50K or >50K).

2. A Bank Marketing dataset that has features as follows:

- **Age:** This column represents the age of the individual. It indicates the age of the person who was contacted for marketing purposes.
- **Job:** The job column refers to the occupation or profession of the individual. It provides information about the type of work they are engaged in. In this example, the person's job is described as "management."
- **Marital:** Marital column indicates the marital status of the individual. It provides information about whether the person is married, single, divorced, or in another marital status. In this case, the person is described as "married."
- **Education:** The education column represents the level of education attained by the individual. It provides information about their educational background or qualifications. In this example, the person's education is described as "tertiary," which typically refers to higher education beyond secondary school.
- **Default:** The default column indicates whether the individual has defaulted on any previous financial obligations. A "yes" value suggests that the person has defaulted, while a "no" value indicates no previous defaults.
- **Balance:** This column represents the balance of the individual's bank account. It provides information about their financial situation or the amount of money they have in their account.
- **Housing:** The housing column indicates whether the individual has a housing loan or not. A "yes" value suggests that the person has a housing loan, while a "no" value indicates the absence of a housing loan.
- **Loan:** The loan column represents whether the individual has a personal loan or not. A "yes" value suggests that the person has a personal loan, while a "no" value indicates the absence of a personal loan.
- **Contact:** This column indicates the communication contact method used to reach the individual. It provides information about how the person was contacted, such as through phone, email, or other means. In this example, the contact method is described as "unknown."
- **Day:** The day column represents the day of the month when contact with the individual took place.
- **Month:** This column indicates the month of the year when the contact occurred.
- **Duration:** The duration column represents the duration of the contact in seconds. It provides information about how long the conversation or interaction lasted between the individual and the marketing team.
- **Campaign:** This column represents the number of contacts performed during the current marketing campaign for this individual.

- **Pdays:** The pdays column indicates the number of days that passed since the last contact with the individual from a previous marketing campaign. A value of -1 suggests that the person was not previously contacted.
- **Previous:** This column represents the number of contacts performed with the individual before the current marketing campaign.
- **Poutcome:** The poutcome column indicates the outcome of the previous marketing campaign for this individual. It provides information about the result or response from previous marketing efforts. In this example, the outcome is described as "unknown."
- **Y:** The "y" column represents the target variable or the outcome variable of interest. It typically indicates whether the individual responded positively ("yes") or negatively ("no") to the marketing campaign.

3. A Diamond Dataset that has features as follows

- **Carat Weight:** This column represents the weight of the diamond, measured in carats. Carat weight is a crucial factor that determines a diamond's size and overall value.
- **Cut:** The cut refers to the quality of the diamond's cut, which directly influences its brilliance and sparkle. In this case, "Ideal" suggests the highest level of cut quality.
- **Color:** This column indicates the color grading of the diamond. The color scale typically ranges from D (colorless) to Z (light yellow or brown). In this example, the diamond has a color grade of "H."
- **Clarity:** Clarity refers to the presence of internal or external flaws, known as inclusions and blemishes, respectively. The clarity scale ranges from Flawless (FL) to Included (I). In this instance, the diamond has a clarity grade of "SI1," indicating slight inclusions.
- **Polish:** Polish refers to the diamond's surface finish quality. It determines the smoothness and reflective properties of the diamond's facets. In this case, the polish is described as "VG" (Very Good).
- **Symmetry:** Symmetry measures the precision and alignment of the diamond's facets. It assesses how well the different parts of the diamond match and interact with each other. The given example has a symmetry grade of "EX" (Excellent).
- **Report:** This column indicates the grading report issuer or certification authority for the diamond. In this case, the diamond is certified by GIA (Gemological Institute of America).
- **Price:** The price column represents the cost of the diamond in the given currency.

In this example, the diamond is priced at 5169 (without specifying the currency).

Furthermore, the study is taken ahead by using X_num binary classification tasks generated by combinations of categorical nominal columns of this dataset. Only those classes that have more than 500 samples are being considered for the Income dataset and those classes having more than 100 samples are being considered for the Diamond and Bank-Marketing dataset. These samples are randomly chosen from the dataset. On the randomly chosen samples, the LightGBM algorithm is used, which will be considered the supermodel for our study.

After the LightGBM model is trained only those models are taken into account whose accuracy score is more than 0.9. The selected model's feature importance attribute (LightGBM library) is used to identify the top n features that have contributed highly to the accuracy of the model. If the feature importance attribute doesn't return any categorical columns in the top n results, then that model will be discarded as that model's accuracy solely depends on numerical columns and categorical columns have significantly less role in the prediction power of the model. If categorical columns are returned those columns will be looked into the encoders that are being tested. Encoding using five nominal encoders such as Leave one out encoder, CatBoost encoder, James Stein encoder, M-Estimate encoder, and Weight of Evidence.

A logistic regression algorithm is then modeled on the data without any hyperparameter tuning for creating baseline models for comparative analysis. After the model is trained the coefficient attribute of the logistic regression model encoded with each encoder separately is used and it is sorted in a descending order following a simple heuristic that the greater the magnitude of the co-efficient greater its say in the model's prediction power, which gives the rank and it is used to identify the importance of those features. If the LightGBM model has a certain number of categorical columns in top n and after encoding using various encoders and using logistic regression if there aren't any categorical columns with high coefficients then it can be said that the logistic regression model has failed to capture the essence of the categorical columns. If the encoded models have categorical columns with high coefficients in the top_n then it can be said that the essence is captured. In this study exact categorical columns which are present in the LightGBM model having an accuracy of more than 90% in the validation set that consists of 20% of all the data are looked up in the encoded models' top_n important columns and their rankings are not considered for keeping the experiment sophisticated. Only Nominal and Numeric Features are considered and Ordinal features are discarded as the study is to test the nominal encoders and applying them on ordinal data or ordinal encoders on Nominal data is of no use.

In this research paper, the spotlight has been placed on five lesser-known techniques that have demonstrated remarkable potential in handling binary classification tasks with remarkable accuracy. By shedding light on these methods, the aim is to contribute to the advancement of the field and inspire further exploration and adoption of these powerful tools by researchers and practitioners alike.

IV. EXPERIMENTAL SETUP

1. **Encoding Techniques:** Five nominal encoding techniques are evaluated: Weight of Evidence (WoE), CatBoost Encoding, JamesStein Encoding, Leave-One-Out Encoding, and Mestimate Encoding. These techniques are applied to the nominal features of the dataset.

2. **Supermodel and Baseline Model Selection:** The LightGBM (LGBM) model will be chosen as the supermodel due to its ability to handle categorical features effectively and its potential for high performance in classification tasks.

- The logistic regression model will be selected as the baseline model, representing a simpler and more interpretable approach.

3. Experimental Procedure:

- **Data Preparation:**
  - The dataset will undergo preprocessing, which includes dropping rows with missing values.
  - Only nominal features and continuous features will be considered for the experiment, excluding ordinal features.

- **Supermodel Training and Evaluation:**
  - For each combination of nominal features, the LGBM model will be trained as a binary classifier.
  - The accuracy of the LGBM model will be evaluated for each binary classification task.
  - If the accuracy of the LGBM model exceeds 90%, indicating a high-performing supermodel, the corresponding combination of nominal features will be selected for baseline modeling.

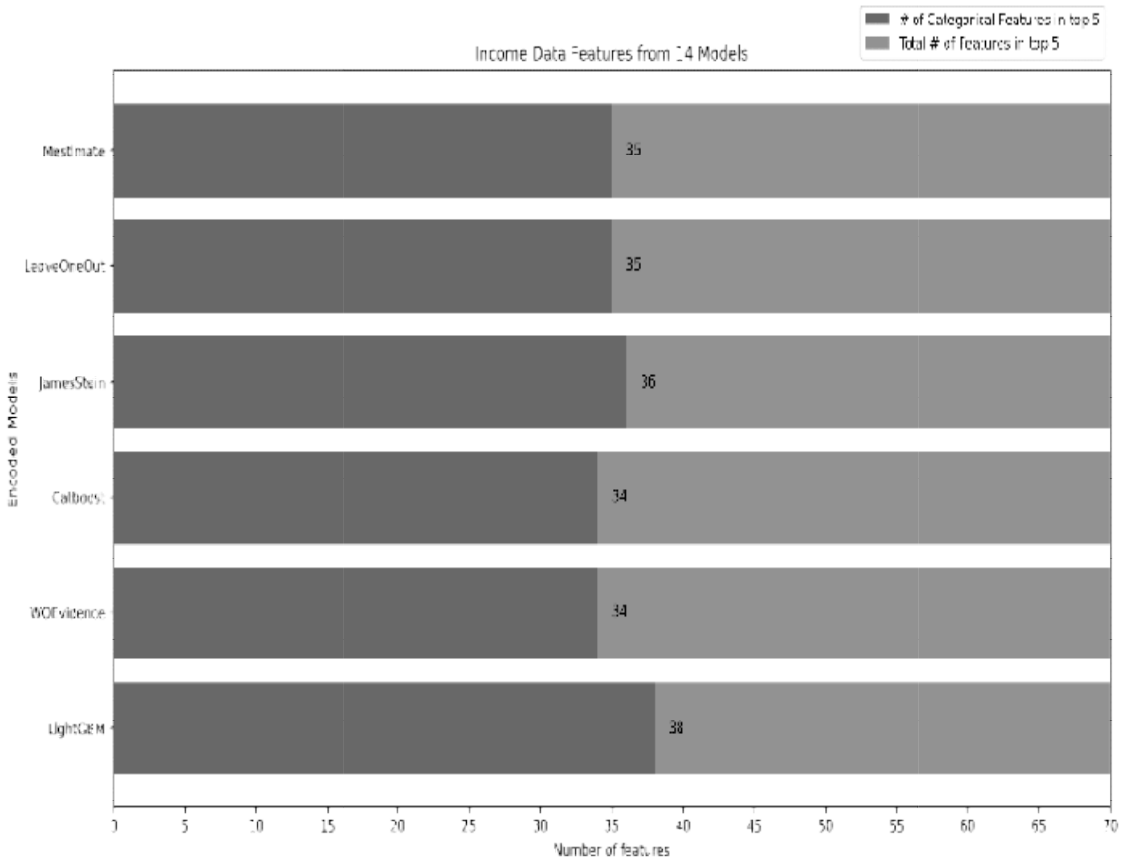
- **Baseline Model Training and Evaluation:**
  - For the selected combinations of nominal features, the logistic regression model will be trained and evaluated as the baseline model for binary classification tasks.
  - The accuracy of the Encoded Logistic regression models and LGBM model will be computed to assess the performance of the baseline model.

- **Hyperparameter tuning:**
  - Both the baseline model (logistic regression) and the LGBM supermodel will be trained using their default hyperparameters.
  - For logistic regression, the default settings will be used, such as LogisticRegression(C=0.1, solver="lbfgs", max_iter=10000).
  - For the LGBM model, the default hyperparameters feature_name='auto' and categorical_feature='auto' will be employed, allowing the model to automatically handle categorical features.

V. RESULTS AND ANALYSIS

In the Income data set 14 binary classification models had accuracy greater than 89 and it had 38 categorical features (70 total features) that hold importance in those models. James Stein Encoding has encoded the most categorical features 36 out of 38 from the LightGBM models and then Leave One Out and M-Estimate share the second rank in this scoring evaluation with 35 features each.

CatBoost and Weight of Evidence both have encoded 34 features out of 38 encoded by the LightGBM model which has an accuracy greater than 89.

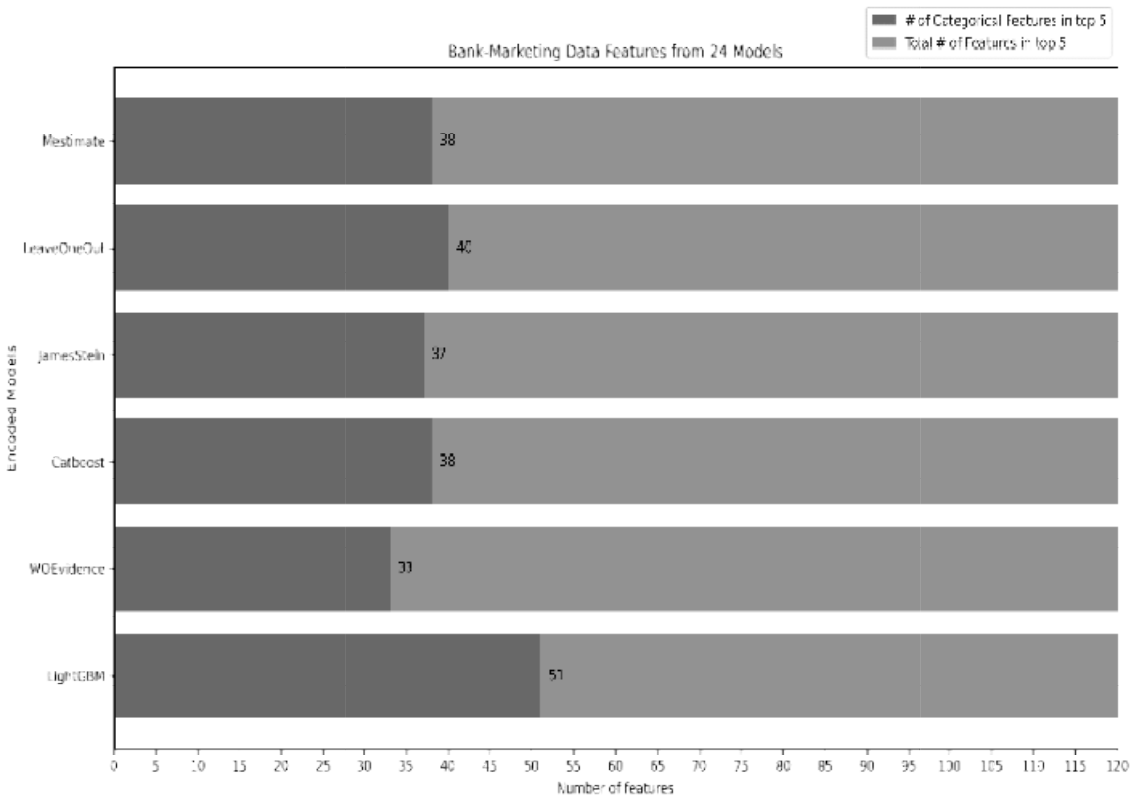


In the Diamond dataset, 10 models were having accuracy greater than 89, and 20 Categorical features were observed as important features (Top 5) from the total 50 Features of all the 10 models.

Leave One Out Encoder Tops the encoding score by successfully encoding 17 out of 20 Categorical features. James Stein, CatBoost & M-Estimate encoders have encoded 16 of the categorical features each, and Weight of Evidence has captured 16 of the Categorical Features.

In the Bank-Marketing dataset, 24 Models were trained which had accuracy greater than .89 (LightGBM) and captured 51 categorical features as Important (Top 5) out of 120 total Features.

Leave One Out Encoder again topped the rankings by successfully encoding 40 out of the 51 Encodings of LightGBM. M-Estimate & CatBoost shares the second spot encoding 38 of the Categorical features. James Stein has captured 37 Categorical Features and Weight of Evidence has again performed below average with 33 Categorical Features.



The observed performance differences among the encoding techniques can be attributed to their underlying principles and algorithms. Leave One Out Encoder's success may be attributed to its ability to prevent overfitting and capture informative patterns in the data. The M-Estimate encoder's robustness could stem from its ability to handle missing values and its flexibility in adjusting the smoothing parameter. The specific reasons behind the performance variations of the James Stein, CatBoost, and Weight of Evidence encoders would require further investigation.

The findings of this research hold significant implications for real-world applications. The following points outline the practical significance of the study:

1. **Marketing and Customer Segmentation:** By applying optimal encoding techniques, businesses can improve customer segmentation, leading to more targeted marketing campaigns and increased customer engagement by segregating customers based on their behaviors and current situation as features and target being the desired segmentation for the business case

2. **Credit Risk Assessment and Fraud Detection:** Utilizing recommended encoding methods enables financial institutions to accurately assess credit risk and detect fraud, reducing financial losses and maintaining the integrity of financial systems.

3. **Healthcare and Medical Diagnosis:** Effective encoding of healthcare data enhances medical diagnosis and treatment, leading to more accurate disease identification and personalized patient care.

VI. DISCUSSION

1. **Dependency on Top K Importance:** The evaluation criterion based on selecting the top 10 or top 20 classes to determine the effectiveness of encoding techniques may overlook instances where the encoded representation fails to capture the essence of lower-ranked classes. This limitation might result in an incomplete assessment of the encoding techniques' performance in representing categorical features.

2. **Comparison with LightGBM Model:** While comparing the encoding techniques with a LightGBM model based on feature importance can provide insights, it is important to acknowledge that feature importance rankings are model-specific and may not fully capture the quality of encoding for all models. The comparison may be limited to the specific characteristics and behavior of LightGBM, potentially limiting the generalizability of the findings.

3. **Potential for Encoding Complementing Unimportant Features:** In some cases, a well-performing model may not necessarily indicate that the encoding technique has successfully captured the essence of a feature. It is possible that the encoded representation inadvertently complements an unimportant feature, resulting in a false sense of effectiveness. This limitation highlights the importance of interpreting the results carefully and considering the overall contribution of features and their encodings.

4. **Influence of Average-Performing Models:** The presence of only good categorical associations in average-performing models may skew the assessment of encoding techniques. This bias can result in overestimating the effectiveness of the encoding techniques due to their alignment with a limited set of important categorical features. The evaluation might not accurately reflect the encoding techniques' performance across a broader spectrum of models and datasets.

5. **Inadequate Representation in Worst-Performing Models**: While assessing the performance of encoding techniques, it is crucial to acknowledge that the worst-performing models may or may not have good categorical associations. The poor performance might be attributed to the inability of the model to capture the essence or represent the features properly, rather than solely relying on the encoding techniques. The limitations of the worst-performing models might not be solely attributed to the encoding techniques themselves.

6. **Encoding Technique Selection:** This research focused on evaluating a specific set of nominal encoding techniques. There are numerous other encoding methods available that were not considered in this study. Future research should explore additional encoding techniques and compare their performance against the ones investigated in this paper.

7. **Hyperparameter Tuning**: The experiments were conducted using the default hyperparameters for both the logistic regression baseline model and the LGBM supermodel. While this allows for a fair comparison, further investigation into hyperparameter tuning for these models may yield improved performance.

8. **Generalizability of Results:** The findings of this research may be specific to the chosen datasets, models, and encoding techniques. Therefore, the generalizability of the results to different datasets or models should be validated through further experimentation.

VII. FUTURE WORK

1. **Evaluation of Additional Encoding Techniques:** The exploration of other encoding techniques, such as entity embedding, target encoding variations, or advanced neural network-based encoders, would provide a comprehensive understanding of their effectiveness and applicability across various machine learning tasks.

2. **Comparative Analysis with Different Datasets**: Conducting similar experiments on diverse datasets from various domains can shed light on the generalizability and robustness of the observed performance of encoding techniques. This would provide insights into the impact of dataset characteristics on the choice of encoding techniques.

3. **Hyperparameter Tuning:** Investigating the effects of hyperparameter tuning for both the baseline model and the supermodel can lead to improved performance. Techniques like grid search or Bayesian optimization can be employed to find the optimal hyperparameter configurations for each model.

4. **Evaluation of Interactions between Encoding Techniques and Models:** Exploring the interaction effects between specific encoding techniques and different machine learning models can provide deeper insights into their complementary strengths and weaknesses. This analysis can guide the selection of encoding techniques for specific models and improve overall model performance

5. **Real-World Application Studies:** Conducting experiments on real-world datasets in specific application domains, such as healthcare, finance, or social sciences, can validate the effectiveness of the identified encoding techniques and further explore their practical significance and impact.

By addressing these limitations and exploring future research directions, the field of categorical data encoding can advance, leading to improved encoding techniques and enhanced performance of machine learning models in real-world applications.

VIII. CONCLUSION

We investigated and compared different encoding strategies for categorical features in the context of machine learning. Our objective was to explore the effectiveness of various encoding techniques and analyze their impact on the predictive power of machine learning models.

Based on our analysis and findings, Leave One Out Encoder consistently stood out as the top-performing encoding technique. It performed above average in all three experiments conducted on the Diamond, Bank-Marketing, and Income datasets. Leave One Out Encoder proved to be effective in encoding a high number of categorical features, demonstrating its reliability and robustness.

The M-Estimate encoder also showcased above-average performance across all three experiments. It consistently performed well in encoding the categorical features, further establishing its effectiveness in handling high cardinality data.

James Stein encoder emerged as the second-best candidate, showing strong performance in the Diamond and Bank-Marketing datasets. However, its performance in the Bank-Marketing dataset was influenced by the extremely poor performance of the Weight of Evidence (WOE) encoder, which affected the average. Nonetheless, the James Stein encoder demonstrated its capabilities in effectively encoding categorical features.

CatBoost encoder displayed average performance in the Diamond dataset, above-average performance in the Bank-Marketing dataset, and below-average performance in the Income dataset. While its performance was not extraordinary, it consistently performed adequately throughout the experiments.

Unfortunately, the weight of Evidence (WOE) encoder consistently exhibited poor performance across all three experiments. It consistently performed below average and did not showcase the effectiveness required for encoding high cardinality categorical features.
In conclusion, Leave One Out Encoder proved to be the top choice for encoding categorical features, consistently delivering above-average performance. The M-Estimate encoder also demonstrated reliability and effectiveness. James Stein encoder showed promise but was affected by the performance of the WOE encoder in one experiment. CatBoost encoder performed adequately, while the Weight of Evidence encoder consistently fell short in terms of performance. These findings provide practitioners with valuable insights and practical guidelines for selecting appropriate encoding methods tailored to their datasets, ultimately improving the accuracy and effectiveness of machine learning models in real-world applications.

REFERENCES

[1] Udilă, Andrei. "Encoding methods for categorical data: A comparative analysis for linear models, decision trees, and support vector machines." (2023).
[2] Pargent, Florian, Bernd Bischl, and Janek Thomas. "A benchmark experiment on how to encode categorical features in predictive modeling." München:Ludwig-Maximilians-Universität München (2019).
[3] Potdar, Kedar, Taher S. Pardawala, and Chinmay D. Pai. "A comparative study of categorical variable encoding techniques for neural network classifiers." International Journal of computer applications 175, no. 4 (2017): 7-9.
[4] Dahouda, Mwamba Kasongo, and Inwhee Joe. "A deep-learned embedding technique for categorical features encoding." IEEE Access 9 (2021): 114381-114391.
[5] Hancock, John T., and Taghi M. Khoshgoftaar. "Survey on categorical data for neural networks." Journal of Big Data 7, no. 1 (2020): 1-41.
[6] Cerda, Patricio, Gaël Varoquaux, and Balázs Kégl. "Similarity encoding for learning with dirty categorical variables." Machine Learning 107, no. 8-10 (2018): 1477-1494.

