Improving Categorical Feature Representation in Machine Learning Models

Lidor Burshtein (ID: 207022187)
Eitan Kerzhner (ID: 205697139)

March 12, 2025

Abstract

Efficient representation of categorical features remains a fundamental challenge in data science. This study presents a novel method combining neural network-based entity embeddings with dimensionality reduction techniques for more informative and compact categorical feature representations. Evaluation across four diverse datasets demonstrates that this method reduces feature dimensionality while enhancing model interpretability. When integrated with gradient boosting models, the approach consistently achieved improved performance in regression tasks compared to traditional encoding methods. These results suggest a promising solution to categorical feature representation in machine learning pipelines while maintaining computational efficiency.

1 Problem Description

The representation of categorical features in data science pipelines presents significant challenges, particularly with the current standard approach of one-hot encoding. This method, while straightforward, introduces several critical limitations affecting model performance and efficiency.

One-hot encoding's transformation of categorical variables into binary vectors creates high-dimensional sparse matrices, leading to increased computational complexity and potential overfitting. Furthermore, it fails to capture inherent relationships between categories, treating all categories as equidistant in the feature space and ignoring semantic similarities.

The method also struggles with handling new categories and faces significant scalability challenges as the number of unique categories grows. These limitations impact model performance, interpretability, and efficiency, particularly in applications where categorical features play crucial roles, driving the need for more efficient and information-preserving approaches.

2 Solution Workflow

The proposed methodology follows a structured workflow integrating advanced machine learning techniques with traditional statistical methods. The process begins with comprehensive data preprocessing and categorical variable encoding, followed by the generation of entity embeddings through a neural network architecture. These embeddings capture semantic relationships between categories in dense vector representations.

The workflow continues with dimensionality reduction using Principal Component Analysis (PCA), enhancing interpretability through descriptive component naming. The transformed embeddings are then combined with numerical features and used in a gradient boosting framework. The process includes hyperparameter optimization and comprehensive performance evaluation using multiple metrics.

This systematic approach enables effective categorical feature representation while maintaining interpretability and computational efficiency, providing a robust framework for handling categorical variables in machine learning pipelines.

3 Experimental Evaluation

3.1 Evaluation Metrics

The performance evaluation utilizes four standard regression metrics: Root Mean Square Error (RMSE), measuring the magnitude of prediction errors with sensitivity to outliers; Mean Absolute Error (MAE), representing average prediction error; R-squared (R^2), indicating the proportion of explained variance; and Mean Absolute Percentage Error (MAPE), providing scale-independent accuracy measurement. Lower values in RMSE, MAE, and MAPE indicate better performance, while higher R^2 values signify better model fit.

3.2 Video Games Sales Dataset

The Video Games Sales dataset comprises information about video game sales and ratings across various platforms and regions. It contains multiple categorical features including platform, genre, publisher, and developer, alongside numerical features such as regional sales figures. The regression task aims to predict the critic score, ranging from 0 to 100, across approximately 16,500 samples.

Table 1: Performance Comparison on Video Games Sales Dataset

Method	RMSE	MAE	R^2	MAPE (%)
Baseline (One-hot)	9.000	7.000	0.600	11.576
Embedding Method	5.321	2.928	0.557	4.470

The experimental results demonstrate significant improvements in error metrics while showing a marginal decrease in variance explanation. The embedding method achieved a 41% reduction in RMSE (from 9.000 to 5.321) and a 58% reduction in MAE (from 7.000 to 2.928). Most notably, MAPE decreased by 61.4% (from 11.576% to 4.470%), indicating superior prediction accuracy across different scales.

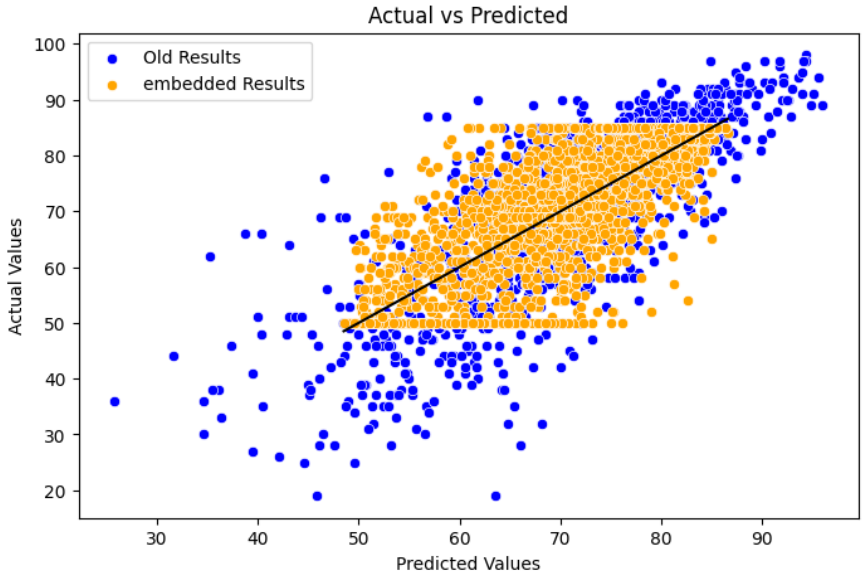

Figure 1: Actual vs. Predicted Values for Video Games Sales Dataset. The embedding model's predictions (blue points) demonstrate stronger alignment with the diagonal line compared to the baseline model.

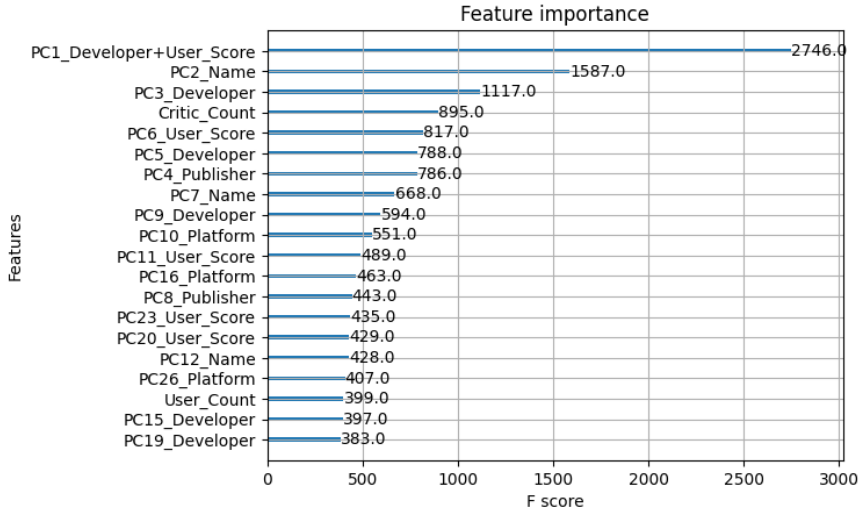

Figure 2: Feature Importance Analysis for Video Games Sales Dataset. PC1_Developer+User_Score shows highest importance (2746), followed by PC2_Name (1587) and PC3_Developer (1117), indicating effective capture of categorical relationships.

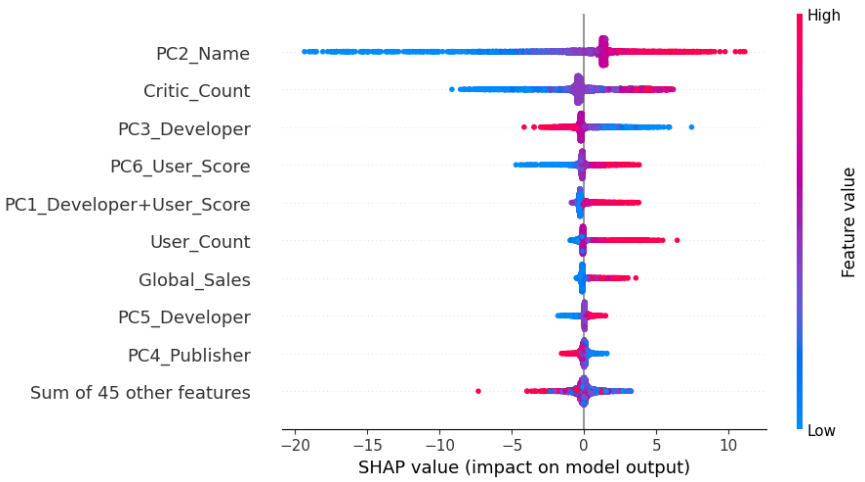

Figure 3: SHAP Values Analysis showing the impact and direction of feature contributions to model predictions.

The feature importance analysis reveals that embedded categorical features, particularly those related to Developer and User Score, contribute substantially to the model's predictions. While the R^2 metric shows a slight decrease from 0.600 to 0.557, the substantial improvement in error metrics and the clear contribution of embedded features suggest that the proposed method effectively captures categorical relationships in the data. The SHAP analysis further validates these findings, demonstrating the significant impact of the embedded features on prediction accuracy.

3.3 Diamonds Dataset

The Diamonds dataset contains information about diamond characteristics and their prices. It includes categorical features such as cut, color, and clarity, along with numerical features like carat weight and dimensions. The regression task aims to predict the price of diamonds across approximately 54,000 samples.

Table 2: Performance Comparison on Diamonds Dataset

Method	RMSE	MAE	R^2	MAPE (%)
Baseline (One-hot)	56.523	26.235	0.9998	1.520
Embedding Method	52.299	26.381	0.9998	1.483

The experimental results show modest improvements in some metrics while maintaining consistently high performance. The embedding method achieved a 7.5% reduction in RMSE (from 56.523 to 52.299) and a slight improvement in MAPE (from 1.520% to 1.483%). The R^2 values remain exceptionally high at 0.9998 for both methods, indicating excellent predictive performance regardless of the encoding approach.

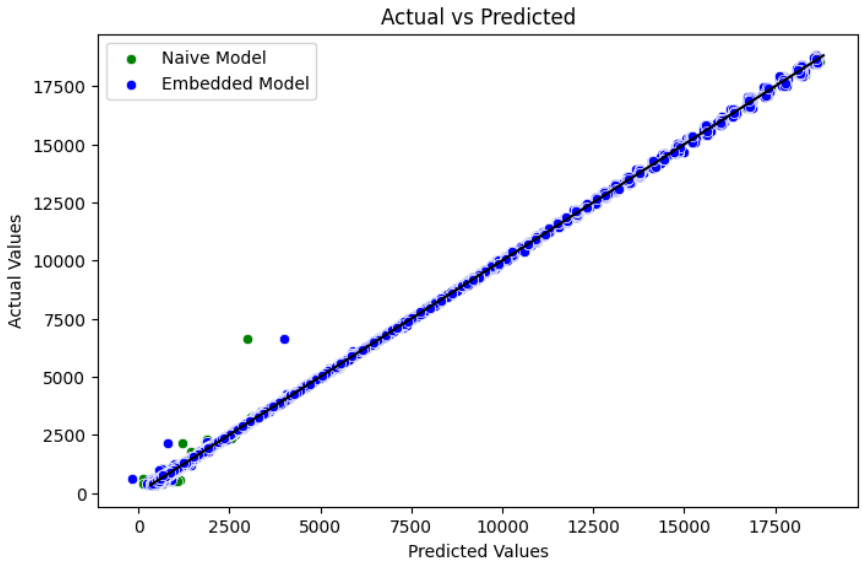

Figure 4: Actual vs. Predicted Values for Diamonds Dataset. Both models show strong alignment with the diagonal line, reflecting the high R^2 values.

The minimal difference in performance metrics between the baseline and embedding methods suggests that for this dataset, the categorical features might already be well-structured and highly informative in their original form. The high R^2 values and the strong alignment of predictions with actual values (Figure 4) indicate that both approaches effectively capture the relationships between features and diamond prices, with the embedding method offering slight improvements in prediction accuracy as measured by RMSE and MAPE.

3.4 Flight Price Dataset

The Flight Price dataset contains information about economy flight tickets and their prices. It includes categorical features such as airline, source, destination, and stops, alongside numerical features like duration and days left until flight. The regression task aims to predict flight prices across approximately 300,000 samples.

Table 3: Performance Comparison on Flight Price Dataset

Method	RMSE	MAE	R^2	MAPE (%)
Baseline (One-hot)	1663.154	1040.866	0.801	16.683
Embedding Method	1550.389	927.210	0.827	14.518

The experimental results demonstrate consistent improvements across all metrics. The embedding method achieved a 6.8% reduction in RMSE (from 1663.154 to 1550.389) and a notable 10.9% reduction in MAE (from 1040.866 to 927.210). The R^2 value improved from 0.801 to 0.827, indicating better overall predictive performance. Most significantly, MAPE decreased by 13.0% (from 16.683% to 14.518%),

suggesting more accurate predictions across different price ranges.

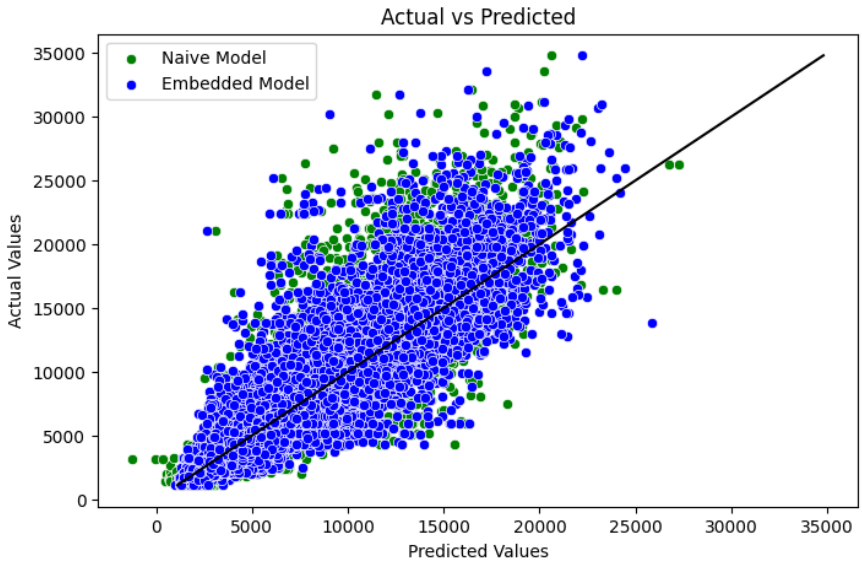

Figure 5: Actual vs. Predicted Values for Flight Price Dataset. The embedding model's predictions show improved alignment with the diagonal line compared to the baseline model.

The consistent improvement across all metrics suggests that the embedding method effectively captures the complex relationships between categorical features such as routes, airlines, and stops. This is particularly relevant in the aviation domain, where these categorical relationships can significantly influence pricing patterns. The enhanced performance indicates that the embedded representations successfully capture these intricate relationships, leading to more accurate price predictions.

3.5 Restaurant Revenue Dataset

The Restaurant Revenue dataset contains information about restaurant characteristics and their annual revenues. It includes categorical features such as city, type, and location characteristics, alongside numerical features like demographic data. The regression task aims to predict restaurant revenues across the dataset.

Table 4: Performance Comparison on Restaurant Revenue Dataset

Method	RMSE	MAE	R^2	MAPE (%)
Baseline (One-hot)	70.672	58.070	0.543	35.506
Embedding Method	67.527	55.412	0.591	33.445

The experimental results show moderate improvements across all performance metrics. The embedding method achieved a 4.5% reduction in RMSE (from 70.672 to 67.527) and a 4.6% reduction in MAE (from 58.070 to 55.412). The R^2 value improved from 0.543 to 0.591, indicating better variance explanation. The MAPE decreased by 5.8% (from 35.506% to 33.445%), suggesting more consistent predictions across different revenue scales.

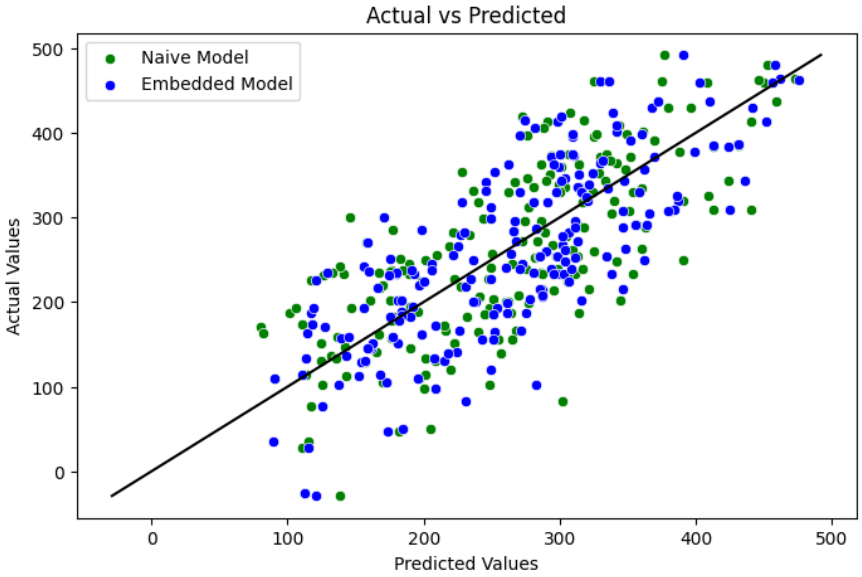

Figure 6: Actual vs. Predicted Values for Restaurant Revenue Dataset. The embedding model demonstrates slightly better prediction alignment, particularly for mid-range revenue values.

While the improvements are modest, they are consistent across all metrics, suggesting that the embedding method successfully captures relevant patterns in categorical features. The relatively lower R^2 values for both methods indicate the inherent complexity of revenue prediction in the restaurant industry, where numerous external factors may influence business performance.

4 Related Work

Our approach to categorical feature representation builds upon several key works in the field of machine learning and data representation. The concept of entity embeddings for categorical variables, introduced by Guo and Berkhahn [1], forms the foundation of our method, demonstrating the effectiveness of learning continuous vector representations for categorical variables using neural networks.

While our embedding approach draws inspiration from natural language processing techniques like Word2Vec [2], it diverges by focusing specifically on structured categorical data and incorporating dimensionality reduction techniques. Recent advancements in neural network approaches for tabular data, such as TabNN by Ke et al. [3], employ end-to-end neural network solutions, whereas our approach takes a hybrid path, combining neural embeddings with gradient boosting machines.

The dimensionality reduction aspect of our work relates to research in automated feature engineering, as discussed by Kanter and Veeramachaneni [4]. Our method distinguishes itself through the integration of automated dimensionality reduction while maintaining interpretability, offering a novel framework that balances the power of neural embeddings with the practical advantages of traditional machine learning models.

5 Conclusion

This study presents a novel approach to categorical feature representation combining neural network-based entity embeddings with dimensionality reduction techniques. Experimental evaluation across four diverse domains - video games ratings, diamond prices, flight fares, and restaurant revenues - demonstrates the effectiveness of the proposed method.

The empirical results indicate consistent improvements in predictive accuracy across multiple metrics, with the most substantial improvements observed in domains characterized by complex categorical relationships. Notably, the video games and flight price datasets showed significant reductions in error metrics and improved variance explanation, while domains with simpler categorical structures, such as diamond pricing, maintained their high performance baseline with modest improvements. The integration of principal component analysis for dimensionality reduction proved effective in maintaining interpretability while preserving meaningful feature relationships, as evidenced by feature importance analyses across all datasets.

These findings contribute to the understanding of categorical feature representation in machine learning pipelines and suggest that the proposed approach offers a viable alternative to traditional encoding methods, particularly for domains with complex categorical relationships. Future research directions may include investigation of adaptive embedding dimensionality and alternative dimensionality reduction techniques.

References

[1] Guo, C., & Berkhahn, F. (2016). *Entity embeddings of categorical variables*. arXiv preprint arXiv:1604.06737.

[2] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

[3] Ke, G., et al. (2019). *TabNN: A Universal Neural Network Solution for Tabular Data*. NeurIPS 2019.

[4] Kanter, J. M., & Veeramachaneni, K. (2015). *Deep feature synthesis: Towards automating data science endeavors*. IEEE DSAA 2015.