Comparative Analysis of Dimensionality Reduction Techniques for Regression Modeling

Parisa Khanjani

Pkhanjani22@ubishops.ca

Department of Computer Science, Bishop's University

Abstract

This report presents a comparative study of various dimensionality reduction methods and their impact on regression analysis. By examining Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP), we aim to illustrate how these techniques influence the performance of linear regression models in terms of mean squared error (MSE).

1. Introduction

In the realm of data science, the curse of dimensionality is a significant barrier, often complicating models with high computational costs and overfitting risks. Dimensionality reduction techniques offer a pathway to simplify datasets while retaining critical information. This study focuses on three such methods: PCA, t-SNE, and UMAP.
Each offers unique advantages and challenges in data transformation and visualization. By comparing these techniques in the context of their impact on linear regression models, this report seeks to illuminate their respective efficacies and guide to enhance predictive modeling and data visualization strategies for complex datasets.

Objectives: The primary goal is to identify how dimensionality reduction methods optimize linear regression model performance by minimizing MSE. We're looking for a method that makes our predictions as close as possible to the real data.

2. Background

2.1 Data Visualization

Data Visualization, by using visual elements like charts, graphs, and maps, provides an accessible way to see and understand trends, outliers, and patterns in data. In complex datasets, where multidimensional data points cloud meaningful insights, effective visualization can help stakeholders make informed decisions based on those insights.

Visualization techniques range from basic charts, like bar, line, and scatter plots to complex data visualizations such as heat maps and advanced multivariate analysis.

2.2 Regression Analysis

Regression Analysis examines the relationship between a dependent variable and one or more independent variables by fitting a mathematical model to observed data. Once a model is established, it can be used to predict values of the dependent variable based on new data for the independent variables. Linear regression assumes a linear relationship between input and output variables. However, when dealing with datasets that contain numerous features, which may lead to overfitting, techniques like Lasso Regression become essential. Lasso Regression extends ordinary linear regression by adding a penalty equal to the absolute value of the magnitude of coefficients. This regularization can reduce the number of features by driving some coefficients to zero, effectively selecting more meaningful features and improving model performance. In contrast, non-linear regression is used when data is more complex and a curvilinear relationship exists between variables.

2.3 Dimensionality Reduction

Dimensionality reduction techniques are critical in handling high-dimensional data, often encountered in fields such as bioinformatics, finance, and social science. These methods reduce the number of random variables under consideration by obtaining a set of principal variables and not only they help in visualizing complex datasets but also significantly improve the efficiency of predictive models by alleviating issues of multicollinearity and the curse of dimensionality in large datasets.

PCA is the most widely used linear dimension reduction technique. It transforms a large set of variables into a smaller one that still contains most of the information in the large set. By doing so, PCA captures the greatest variance in the data with the fewest number of principal components. Each component in PCA is a linear combination of every feature; the first principal component has the highest possible variance, and each succeeding component has the highest variance possible under the constraint that it is orthogonal to the preceding components.

t-SNE is a sophisticated non-linear technique mainly used for exploring high-dimensional data. It reduces dimensions by converting affinities of data points to probabilities. The similarities in the high-dimensional space are represented by Gaussian joint probabilities, and the low-dimensional counterparts are represented by Student's t-distributions. This approach allows t-SNE to capture much more complex patterns in the data, though it is computationally intensive and typically not recommended for extrapolation or interpretation of component axes.

UMAP is a relatively new technique that is similar to t-SNE but surpasses it in terms of scalability and preserving more of the global structure of the data. UMAP works by constructing a high-dimensional graph representation of the data then optimizes a low-dimensional graph to be structurally similar. This technique is particularly powerful for data visualization and clustering as it helps reveal structures in data, which might include grouping or hierarchy.

In summary, data visualization provides the tools to effectively summarize and present data insights. Regression analysis offers methodologies to predict and understand the relationships within the data. Dimensionality reduction facilitates these processes by simplifying complex, high-dimensional datasets into manageable, insightful representations. Together, these techniques form the backbone of data-driven decision-making in contemporary analytics practices.

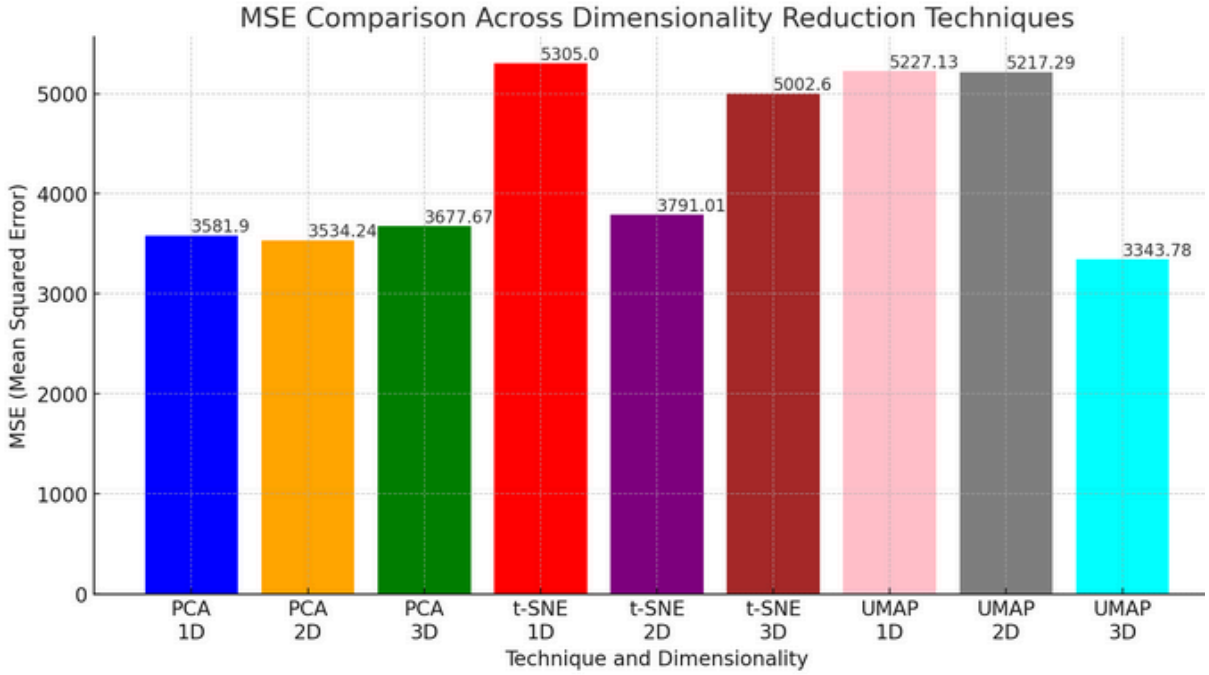
3. Experiments and Evaluation

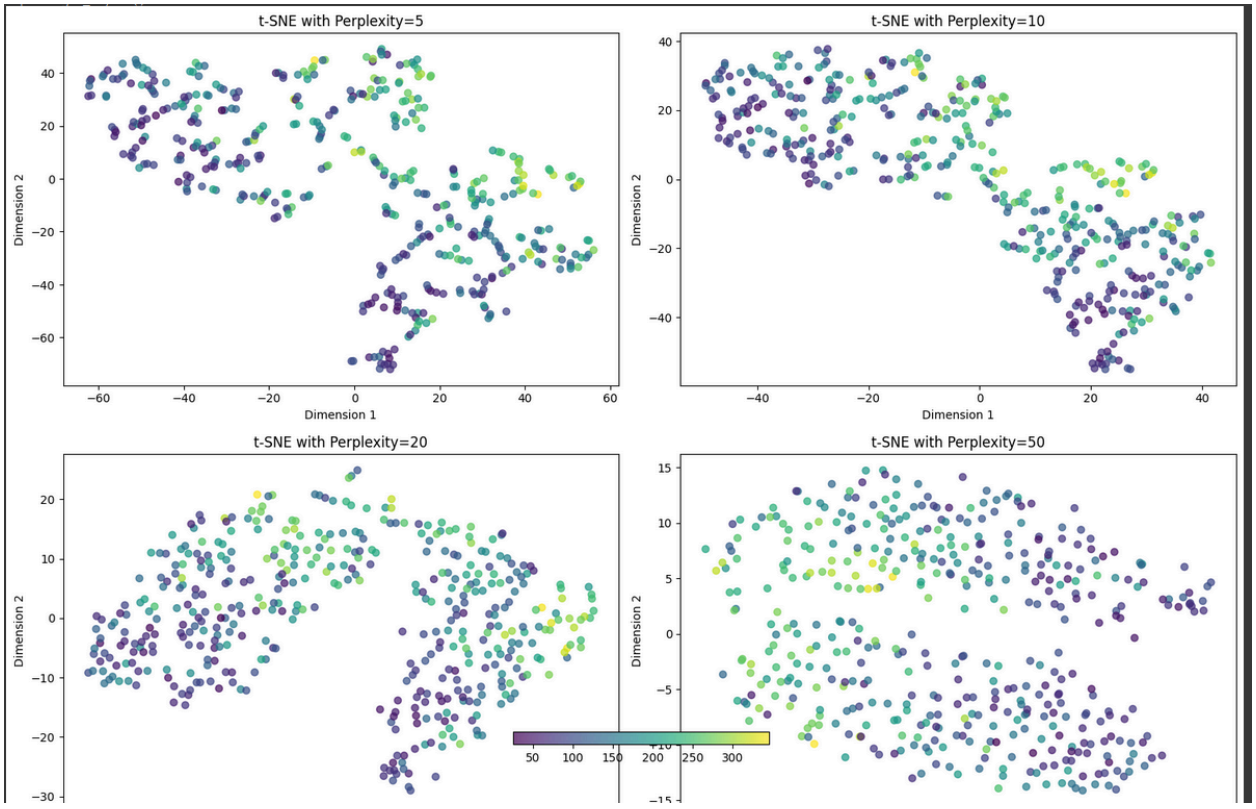
3.1. Dataset Description

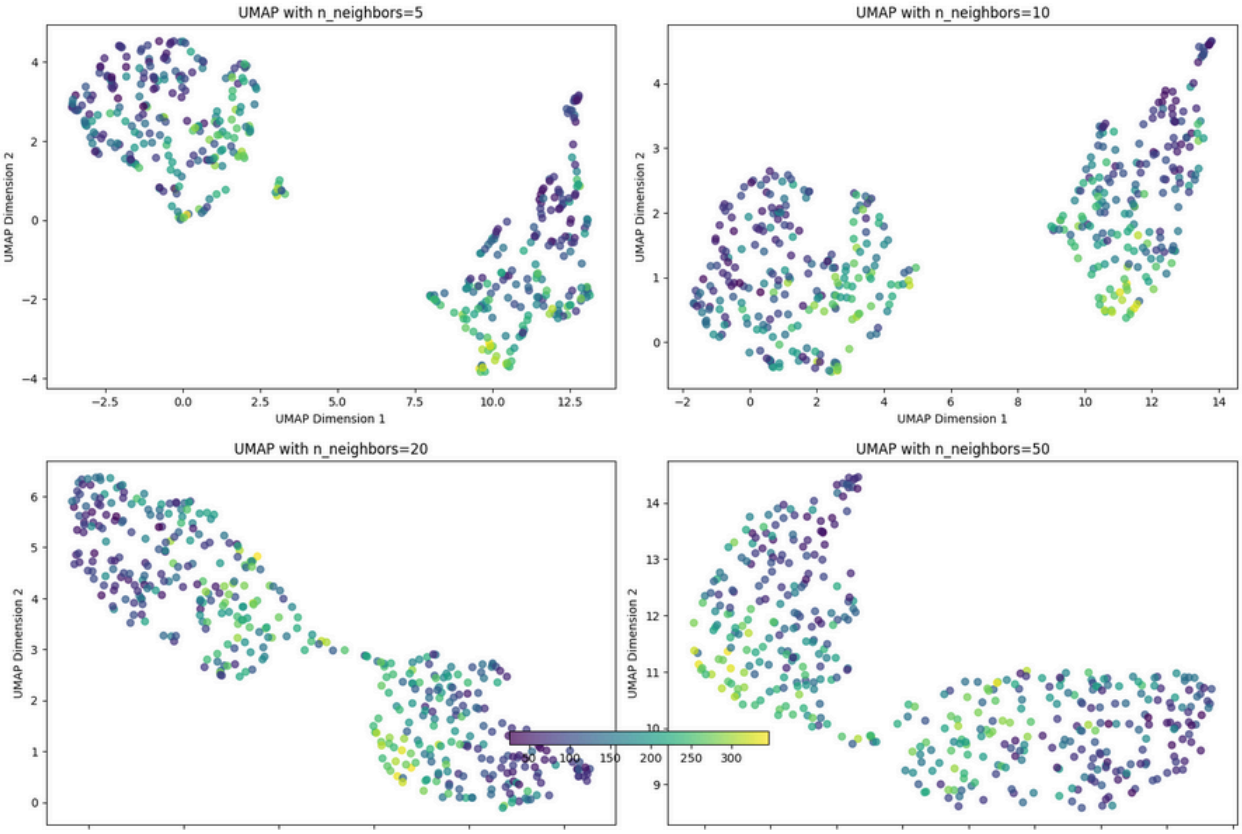
Each row in the dataset corresponds to a single patient's medical record. The dataset includes both numerical and categorical data. Numerical data includes age, BMI, blood pressure, and various blood serum measurements. The categorical data encoded as numerical values, particularly in the 'SEX' column. The final column, 'Y', represents a target variable, which is a measure of diabetes progression after one year. The other columns are used as predictor variables in a regression analysis to determine their influence on the target variable.

3.2. Dimensionality Reduction Method Results

- PCA's parameters (like the number of components) directly affect the amount of variance captured. More components can capture more variance but can also lead to overfitting if irrelevant variance is included. In this dataset, PCA generally improved from 1D to 2D but slightly worsened in 3D.
- Parameters like perplexity and the number of components influence how t-SNE balances attention between local and global similarities. Higher perplexity can lead to focusing on more global aspects, but may not always align with the underlying predictive structure needed for regression. Comparatively to the other techniques, t-SNE's performance was poor in 1D, and improved in 2D, but was not optimal in 3D.
- Parameters such as n_neighbors and min_dist control UMAP's emphasis on local versus global structure. n_neighbors controls how UMAP balances local versus global structure in the data, with lower values emphasizing local structure, which may not always provide the best features for regression tasks. The UMAP approach was less effective in 1D and 2D but was most effective in 3D. This suggests that UMAP is particularly good at capturing complex, non-linear structures when considering more dimensions.

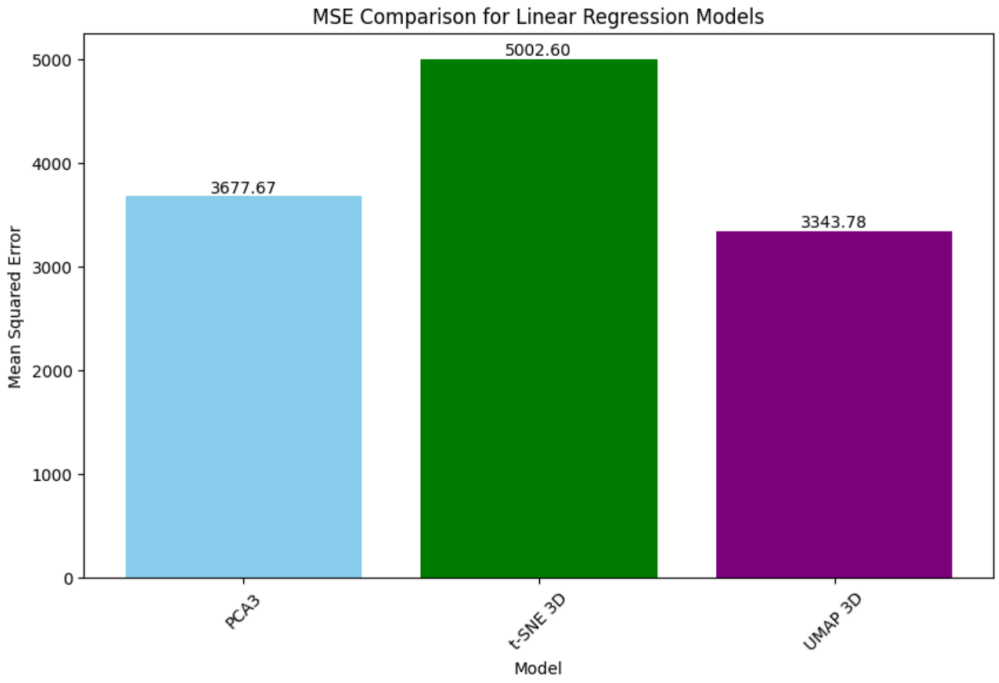

MSE Comparison Across Dimensionality Reduction Techniques
MSE (Mean Squared Error)
Technique and Dimensionality
PCA 1D
PCA 2D
PCA 3D
t-SNE 1D
t-SNE 2D
t-SNE 3D
UMAP 1D
UMAP 2D
UMAP 3D
3581.9
3534.24
3677.67
5305.0
3791.01
5002.6
5227.13
5217.29
3343.78


t-SNE with Perplexity=5
t-SNE with Perplexity=10
t-SNE with Perplexity=20
t-SNE with Perplexity=50
Dimension 1
Dimension 2



3.3. Comparative Analysis

The bar chart comparison of MSE across PCA, t-SNE, and UMAP models offers clear insights into their predictive accuracies after dimensionality reduction on a diabetes dataset. PCA shows decent performance with an MSE of 3677.67, while t-SNE lags with the highest error at 5002.60, suggesting it may not preserve critical variance for regression tasks in this context. UMAP stands out, achieving the lowest MSE of 3343.78, indicating its superior capability in retaining useful data structures for predictions. The success of methods used to reduce data dimensions is greatly influenced by the specific characteristics of the dataset, and for this particular set of data, UMAP's detailed approach to manifold learning appears to be exceptionally fitting and suggesting that the diabetes dataset contains non-linear interdependencies that are better captured in a multi-dimensional space.



3.4. Discussion

In our analysis of dimensionality reduction techniques, we observed that the effectiveness of these methods greatly depends on the specific structure of the dataset. While PCA, with its linear approach, provided some level of success, it struggled to capture the complex, non-linear relationships present in the diabetes dataset. Similarly, t-SNE, which prioritizes local structures, did not necessarily lead to better predictive accuracy, indicating a potential mismatch with the dataset's requirements. UMAP stood out as the most appropriate technique, offering a balanced consideration of both local and global structures and showing that a thoughtful approach to dimensionality reduction can significantly improve predictive modeling.

The integration of LASSO into this project brings additional refinement by implementing variable selection and regularization. LASSO's ability to drive the coefficients of less important variables towards zero as the regularization parameter, alpha, increases, is crucial. This not only helps prevent overfitting but also pinpoints the most influential variables. By choosing the optimal alpha, LASSO ensures that essential variables are retained while superfluous ones are excluded, thus enhancing both the accuracy and interpretability of our model. This makes LASSO a key component in not just refining our model but in maximizing the contributions of essential variables, complementing the dimensionality reduction insights provided by techniques like UMAP.

4. Conclusion

In summary, these findings emphasize that the choice of dimensionality reduction technique can significantly impact predictive modeling outcomes. They also highlight the importance of considering the intrinsic geometry of the data, which may be better captured by some methods in higher dimensions. For this diabetes dataset, UMAP's 3D reduction aligns best with the underlying data structure, leading to the most accurate predictions in the context of a linear regression framework. This aligns with the broader understanding that while PCA is often the first choice for linear data reduction, UMAP (and to some extent t-SNE) can be powerful alternatives for datasets with complex, non-linear patterns.