Journal of Mathematics and Modeling in Finance (JMMF)
Vol. 4, No. 2, Summer & Autumn 2024
Research paper

Sensitivity assessing to data volume for forecasting: introducing similarity methods as suitable ones in feature selection methods

Mahdi Goldani[1], **Soraya Asadi Tirvan**[2]

[1] Faculty of Literature and Humanities, Hakim Sabzevari University, Sabzevar, Iran
m.goldani@hsu.ac.ir

[2] Department of Energy Economics, Allameh Tabatabai University, Tehran, Iran
s_asadi@atu.ac.ir

Abstract:
In predictive modeling, overfitting poses a significant risk, particularly when the feature count surpasses the number of observations, a common scenario in high-dimensional datasets. To mitigate this risk, feature selection is employed to enhance model generalizability by reducing the dimensionality of the data. This study evaluates the stability of feature selection techniques with respect to varying data volumes, focusing on time series similarity methods. Utilizing a comprehensive dataset that includes the closing, opening, high, and low prices of stocks from 100 high-income companies listed in the Fortune Global 500, this research compares several feature selection methods, including variance thresholds, edit distance, and Hausdorff distance metrics. Numerous feature selection methods were investigated in literature. Selecting the more accurate feature selection methods in order to forecast can be challenging [1]. So, this study examines the most well-known feature selection methods' performance in different data sizes. The aim is to identify methods that show minimal sensitivity to the quantity of data, ensuring robustness and reliability in predictions, which is crucial for financial forecasting. Results indicate that among the tested feature selection strategies, the variance method, edit distance, and Hausdorff methods exhibit the least sensitivity to changes in data volume. These methods, therefore, provide a dependable approach to reducing feature space without significantly compromising predictive accuracy. This study highlights the effectiveness of time series similarity methods in feature selection and underlines their potential in applications involving fluctuating datasets, such as financial markets or dynamic economic conditions.

Keywords: feature selection, sample size, overfitting, similarity methods.
JEL Classification: C52, C38, G17.

1 Introduction

In machine learning models, having a complete and comprehensive dataset can significantly enhance model accuracy. However, there are instances where the inclusion of irrelevant features may hinder rather than help the model's performance. In fact, the feature space with larger dimensions creates a larger number of parame-

Corresponding author
Received: 01/09/2024 Accepted: 18/12/2024
https://doi.org/10.22054/JMMF.2024.81735.1145

ters that need to be estimated. As a result, by increasing the number of parameters, the probability of overfitting in the demand model is strengthened. Therefore, the best performance generalization is achieved when a subset of the available features is used. Dimensionality reduction is the way to solve this challenge. The literature on dimensionality reduction refers to transforming data from a high-dimensional space into a low-dimensional space. One of the most well-known techniques of dimensionality reduction is Feature selection [2]. Feature selection selects a subset of relevant features for use in model construction. The filters, embedded, and wrapper methods are the three main categories of Feature selection methods. Filter methods are characterized by their independence from specific machine learning algorithms. They prioritize data relationships, making them computationally efficient and straightforward to implement. In contrast, wrappers and embedded methods rely on learning algorithms. While filters are computationally efficient and easy to implement, wrappers often achieve better performance by considering feature interactions, albeit with increased computational complexity. Embedded methods strike a balance between filters and wrappers, integrating feature selection into the training process. This integration reduces computational costs compared to wrappers, as it eliminates the need for separate iterative evaluation of feature subsets [3]. Along with feature selection methods, time series similarity methods used for clustering in machine learning can be a suitable option for selecting a suitable subset of variables. Similarity methods can serve as effective feature selection techniques by identifying redundant or irrelevant features, grouping similar features together, and quantifying the relationships between features and the target variable. By leveraging similarity measures in feature selection, one can extract the most informative features from the dataset while reducing dimensionality and improving model performance. Specifically, this study evaluates the time series similarity methods such as variance thresholds, edit distance, and Hausdorff distance metrics. These methods are chosen for their ability to capture different aspects of similarity and variability in time series data. The review of articles in this field shows that similarity methods are used in combination with feature selection methods. Similarity in time series refer to time series variables are highly time dependent [4]. presented a comprehensive review of time-series measures, classifying them into four major categories: lock-step measures (e.g., Euclidean distance and Manhattan distance), elastic measures (e.g., longest common subsequence (LCS) and dynamic time warping (DTW), Edit distance), pattern-based measures (e.g., spatial assembling distance (SpADe)), and threshold-based measures (e.g., threshold query-based similarity search (TQuEST)) [5]. Özkoç category was recommended to measure the similarity. Geometric similarity measures is one of them. such as the Hausdorff distance, Fréchet distances [6]. Similarity methods like DTW, Edit Distance, Hausdorff Distance, Euclidean Distance, and Fréchet Distance are not only powerful tools for measuring the similarity between time series but also play a significant role in dimensionality reduction. Xie et al. [7] utilized similar measures to reduce the dimensionality of datasets effectively. The current research seeks to identify feature selection methods that demonstrate minimal sensitivity to data volume, thereby ensuring robustness and reliability in predictionsan essential aspect of financial forecasting. This study enhances the existing knowledge by showcasing the effectiveness of time series similarity methods in feature selection and emphasizing their potential applications in environments with fluctuating datasets, such as financial markets or dynamic economic conditions. In this study, we will evaluate several time series similarity methods, including Dynamic Time Warping (DTW), Edit Distance, Hausdorff Distance, Euclidean Distance, Fréchet Distance, and Lasso. These methods will be compared to traditional feature selection techniques to assess their stability and effectiveness in identifying important features, mitigating overfitting, and improving model performance by extracting the most informative features from the dataset and reducing dimensionality.

In literature, the primary aim of feature selection is to eliminate irrelevant variables, particularly when the number of features exceeds the number of observations. This practice helps mitigate overfitting, ensuring that the model generalizes well to unseen data. Therefore, feature selection is a method for dealing with a small number of observations. But does the performance of feature selection methods change when the number of observations is very small? In fact, this article seeks to find the answer to this question; When we are faced with a small number of observations, the results of which of the feature selection methods can be more reliable? This issue is important because most of the existing data sets that provide annual data face the problem of a small number of observations. Therefore, finding a way to reduce the dimension of a data set when the number of observations is low can help to increase the accuracy of the models. The aim of this research is to find the most optimal method to reduce the dimension of data that has the least impact on the performance of these models.

Feature selection is a widely used technique in various data mining and machine learning applications. In the literature on feature selection, there is no study that uses similarity methods directly as feature selection methods but there are some researches that explore this concept or incorporate similarity measures into feature selection processes. For example, Zhu et al [8] In the proposed Feature Selection-based Feature Clustering (FSFC) algorithm, similarity-based feature clustering utilized a means of unsupervised feature selection. Mitra [9] proposes an unsupervised feature selection algorithm designed for large datasets with high dimensionality. The algorithm is focused on measuring the similarity between features to identify and remove redundancy, resulting in a more efficient and effective feature selection process. In the domain of software defect prediction, Yu et al. [10] emphasize the central role of similarity in gauging the likeness or proximity among distinct software modules (referred to as samples) based on their respective features. Shi et al. [11] proposed a novel approach called Adaptive-Similarity-based Multi-modality Feature Selection (ASMFS) for multimodal classification in Alzheimer's disease

(AD) and its prodromal stage, mild cognitive impairment (MCI). They addressed the limitations of traditional methods, which often rely on pre-defined similarity matrices to depict data structure, making it challenging to accurately capture the intrinsic relationships across different modalities in high-dimensional space. In the FUs [12] article Following the evaluation of feature relevance, redundant features are identified and removed using feature similarity. Features that exhibit high similarity to one another are considered redundant and are consequently eliminated from the dataset. Feature similarity measures are utilized to quantify the similarity between pairs of features. These measures help identify redundant features by assessing their degree of resemblance or closeness.

In terms of data size, theres been a bunch of studies that have addressed this issue. Vabalas [13] highlights the crucial role of sample size in machine learning studies, particularly in predicting autism spectrum disorder from high-dimensional datasets. It discusses how small sample sizes can lead to biased performance estimates and investigates whether this bias is due to validation methods not adequately controlling overfitting. Simulations show that certain validation methods produce biased estimates, while others remain robust regardless of sample size. Perry et al. [14] underscore the significance of sample size in machine learning for predicting geomorphic disturbances, showing that small samples can yield effective models, especially for identifying key predictors. It emphasizes the importance of thoughtful sampling strategies, suggesting that careful consideration can enhance predictive performance even with limited data. Cui and Goan [15] tested Six common ML regression algorithms on resting-state functional MRI (rs-fMRI) data from the Human Connectome Project (HCP), using various sample sizes ranging from 20 to 700. Across algorithms and feature types, prediction accuracy and stability increase exponentially with larger sample sizes. Kuncheva et al. [16] conducted experiments on 20 real datasets. In an exaggerated scenario, where only a small portion of the data (10 instances per class) was used for feature selection while the rest was reserved for testing, the results underscore the caution needed when performing feature selection on wide datasets. The findings suggest that in such cases, it may be preferable to avoid feature selection altogether rather than risk providing misleading results to users. Kuncheva [17] challenges the traditional feature selection protocol for high-dimensional datasets with few instances, finding it leads to biased accuracy estimates. It proposes an alternative protocol integrating feature selection and classifier testing within a single cross-validation loop, which yields significantly closer agreement with true accuracy estimates. This highlights the importance of re-evaluating standard protocols for accurate performance evaluation in such datasets.

While existing literature has introduced abundant feature selection methods. Choosing one of them is challenging for researchers. So, this study in addition to distinguishes itself by integrating time series similarity methods as approach to dimensionality reduction, as an innovation, it compares the methods in different sizes of data sets. This innovation helps to identify the performance of methods in different data sets and choose more accurate method. This focus on data volume sensitivity ensures that our findings contribute to the development of more reliable and efficient feature selection frameworks, addressing a key limitation in existing literature. A review of studies clarifies two basic issues. One, among the mentioned studies, there is no study that directly uses similarity methods as a feature selection method. Therefore, as a new proposal, this study directly uses similarity methods as a feature selection method and compares their prediction performance with feature selection methods. Second, in this study, a real data set (Financial data of the 100 largest companies by revenue), to evaluate the sensitivity of each method to the sample size and compare it with another. The rest of the paper is organized as follows: methodology is discussed in Section 2, Section 3 presents the results of the study, and Section 4 reports a discussion of findings and conclusions.

2 Methodology

This section elaborates on the methodology adopted for this research work. The complete methodology is depicted in Fig. 1 and consists of the following steps.

- Historical finance datasets of the 100 biggest companies are collected.
- In this step, appropriate features are selected using feature selection methods and similarity methods.
- Feature selection methods were used in 80 steps. Each step reduced the dataset size by 1% until just 20% of the primary dataset.
- Linear regression is trained on selected features and forecasts 10 days ahead of APPL close price.
- In the last step, Linear regression performance is evaluated through cross-validation techniques and results are documented.

2.1 Dataset

Based on the aim of this paper, to examine the Density and performance of the feature selection methods and similarity methods during high and low sample sizes, the finance dataset was chosen. A large amount of financial data is a suitable feature to examine the performance of methods in large to small amounts of data. According to the Fortune Global 500 2023 rankings, the data set of this research was secondary data including open, low, high, and close prices and the volume of the 100 biggest companies by consolidated revenue. The target value of this dataset was Apple's close price. The prediction of the closing price of this variable is done in different data sizes and the best model was selected from among the datasets.

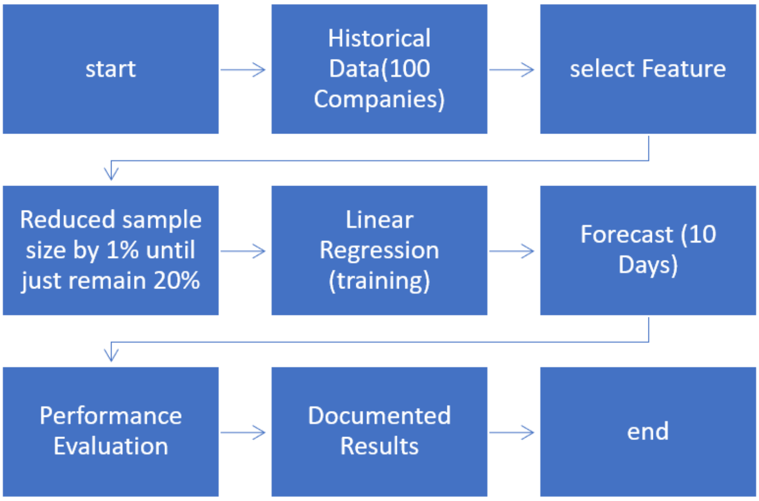
start
Historical Data(100 Companies)
select Feature
Reduced sample size by 1% until just remain 20%
Linear Regression (training)
Forecast (10 Days)
Performance Evaluation
Documented Results
end

Figure 1: The complete methodology.

The data were collected from the Yahoo Finesse site spanning from January 1, 2016, to January 28, 2024. One of the main reasons for choosing this dataset is the high quality of financial data, the large volume of data and the least number of missing values. Considering the implementation of the step-by-step data volume reduction test, it was necessary to choose such a dataset with a large data volume so that the behavior of the data is not disturbed by the data volume reduction. This research's main approach is measuring feature selection algorithms' sensitivity to sample size. For this purpose, the feature selection methods were used in 80 steps. Each step reduced the dataset size by 1% to just 20% of the primary dataset. This test is done in order to find out which method can have better results in real conditions when faced with a dataset with a small amount of data. Therefore, each step of this experiment is independent from the previous step and the results of each step are not dependent on the previous step.

2.2 feature selection methods

Once the database without missing value is obtained, the next step is to apply FS and similarity methods to choose the most relevant variables. Feature selection involves the study of algorithms aimed at reducing the dimensionality of data to enhance machine learning performance. In a dataset with N data samples and M features, feature selection aims to decrease M to M', where $M' \leq M$. Subset selection entails evaluating a group of features together for their suitability. The general procedure for feature selection comprises four key steps: Subset Generation, Evaluation of Subset, Stopping Criteria, and Result Validation. Subset generation involves a heuristic search, where each state specifies a candidate subset for evaluation within the search space. Two fundamental issues determine the nature of the subset generation process. Firstly, the successor generation determines the search's starting point, which influences its direction. Various methods, such as forward, backward, compound, weighting, and random methods, may be considered to decide the search starting points at each state [18]. Secondly, the search organization is responsible for the feature selection process with a specific strategy, such as sequential, exponential, or random search. Any newly generated subset must be evaluated based on specific criteria. Consequently, numerous evaluation criteria have been proposed in the literature to assess the suitability of candidate feature subsets. These criteria can be categorized into two groups based on their dependency on mining algorithms: independent and dependent criteria [19]. Independent criteria exploit the training data's essential characteristics without employing mining algorithms to evaluate the goodness of a feature set or feature. Based on the selection strategies and/or criteria, there are three main types of feature selection techniques. wrappers, filters, and embedded methods [20]. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of overfitting the model. Filters are similar to wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in, and specifically, a model. The table below illustrates the most well-known methods in each category.

2.3 Similarity methods

As is clear in Table 1 each method of FS has some Limitations and weaknesses. Therefore, the time series similarity methods can be a good choice as feature selection methods. Measuring similarity in time series forms the basis for the clustering and classification of these data, and its task is to measure the distance between two-time series. The similarity in time series plays a vital role in analysing temporal patterns. Firstly, the similarity between time series has been used as an absolute measure for statistical inference about the relationship between time series from different data sets [21]. In recent years, the increase in data collection has made it possible to create time series data. In the past few years, tasks such as regression, classification, clustering, and segmentation were employed for working with time series. In many cases, these tasks require defining a distance measurement that indicates the level of similarity between time series. Therefore, studying various methods for measuring the distance between time series appears essential and necessary. Among the different types of similarity measurement criteria for time series, they can be divided into three categories: step-by-step measures, distribution-based measures, and geometric methods. Table 2 describes both advantages and disadvantages of similarity methods.

The choice of feature selection method can significantly impact computational efficiency and scalability, especially when dealing with large datasets. Variance

Table 1: Feature Selection Methods

Group	**Method Name**	**Definition**	**Disadvantage**
Filters	Correlation-based	Identifies the strength and direction of the linear relationship between features and the target variable.	Assumes only linear relationships and may miss complex, nonlinear correlations.
	Variance Threshold	Eliminates features with low variance, considering them uninformative.	It cannot capture nonlinear relationships and ignores the target variable.
	Information Gain	Measures the effectiveness of a feature in classifying target labels using information theory.	They might struggle with continuous data, and the methods often require discretization.
Wrappers	Forward Selection	It is a greedy algorithm that starts with an empty set of features and adds the most predictive feature iteratively.	Overfitting is a concern, and it is sensitive to the choice of the evaluation metric.
	Backward Elimination	Backward Elimination starts with all features and removes the least significant ones iteratively.	One disadvantage of Backward Elimination is that it can be computationally expensive for large datasets.
	Recursive Feature Elimination	Recursive Feature Elimination is a feature selection method that recursively removes features and ranks them based on model performance.	One potential disadvantage of RFE is its computational expense for large datasets.
	Stepwise Selection	Stepwise Selection combines forward and backward selection to iteratively select features.	A potential disadvantage of Stepwise Selection is that it can be prone to overfitting.
	Genetic Algorithms	Genetic Algorithms are optimization methods inspired by natural selection to find the best subset of features.	A drawback is their computational complexity, especially with large feature sets.
Embedded	L1 Regularization (Lasso)	Lasso is a regularization method in machine learning that adds a penalty term to linear regression, shrinking coefficients of less important features to zero.	Sensitivity to the choice of the regularization parameter λ.
	Tree-based methods	Tree-based methods, like Random Forest and gradient-boosted trees, perform feature selection internally.	Prone to overfitting, especially with deep trees and small datasets.
	Recursive Feature Elimination with Cross-Validation (RFECV)	RFECV combines Recursive Feature Elimination (RFE) with cross-validation to find the optimal number of features.	The method can be computationally intensive due to repeated model training.
	XGBoost and LightGBM	XGBoost and LightGBM are powerful gradient-boosted decision tree methods for ranking and selecting features.	Both XGBoost and LightGBM can be sensitive to hyperparameters, and tuning is required for optimal performance.

Table 2: Similarity Methods

Method	**Advantages**	**Disadvantages**	**Category**
Euclidean Distance	Most straightforward and widely used criterion for similarity measurement.	Does not support local time shifts. Inefficient for large datasets.	Step-by-step
DTW (Dynamic Time Warping)	Supports local scaling and order preservation of sequences.	Time-consuming. Sensitive to noise. High computational cost.	Elastic
LCSS (Longest Common Subsequence)	Robust against noise. Focuses on similar parts of sequences.	Strongly depends on similarity threshold. Zero matches may arise for low overlap.	Elastic
EDR (Edit Distance on Real sequence)	Robust against noise and data corruption. Handles real-value sequences.	Not metric.	Elastic
ERP (Edit Distance with Real Penalty)	Metric distance that follows the triangle inequality.	Compares locations only within a time threshold.	Elastic
Hausdorff Distance	Measures spatial similarity between two routes, considering extreme deviations.	Not suitable for trends. Complex calculations.	Geometric
Discrete Frechet Distance	Considers order and continuity of points.	Limited applicability to path comparison.	Geometric

Source: [22]

Threshold, Correlation-Based Selection, and Tree-Based Methods offer high scalability and relatively low computational complexity, making them suitable for large-scale applications. On the other hand, methods like Recursive Feature Elimination, Dynamic Time Warping, and Hausdorff Distance, while potentially more precise, are computationally intensive and may not scale well for very large datasets. Therefore, the selection of feature selection methods should balance between the computational resources available and the required prediction accuracy.

2.4 Measure the performance of methods

After selecting the most relevant variables and creating a suitable subset, the performance of each of the selected subsets was measured using A 10-fold cross-validation method. This method divided the one dataset randomly into 10 parts. The 9 parts out of 10 parts are used for training and reserved one-tenth for testing [23]. This process was repeated 10 times, reserving a different tenth for testing. During this process, the linear regression model is used for training and testing. In statistics,

Table 3: Feature Selection Methods with Complexity and Scalability

Method	**Computational Complexity**	**Scalability**	**Description**
Variance Threshold	Low	High	Simple and fast, involves computing the variance of each feature and discarding low-variance features.
Stepwise Selection	Moderate to High	Moderate	Iteratively adds/removes predictors based on criteria like AIC, BIC, or R-squared.
Correlation-Based	Low	High	Involves computing the correlation matrix, straightforward and efficient for large datasets.
Backward Elimination	High	Moderate to Low	Starts with all features and removes the least significant one iteratively.
Recursive Feature Elimination	High	Moderate	Recursively removes least important features based on model performance.
Euclidean Distance	Moderate	High	Calculates pairwise distances between feature vectors for similarity analysis.
Dynamic Time Warping	High	Moderate	Measures similarity between time series, computationally expensive for long sequences.
Simulated Annealing	High	Low	Probabilistic global optimization, computationally expensive for large-scale problems.
Tree-Based Methods	Moderate	High	Includes methods like random forests and decision trees that perform internal feature selection.
Forward Selection	Moderate to High	Moderate	Iteratively adds predictors to the model, requires repeated model evaluation.
Hausdorff Distance	High	Low to Moderate	Measures the maximum distance between two sets of points, useful for spatial similarity.
Mutual Information	Moderate	High	Measures dependency between variables, efficiently handles non-linear relationships.
Fréchet Distance	High	Low	Similar to DTW but more expensive as it considers all orders and trajectories.
Edit Distance	High	Low	Measures minimum operations to transform one sequence into another, computationally expensive.
Lasso Regression	Moderate	High	Performs feature selection and regularization, effectively reducing model complexity.

linear regression is a statistical model that estimates the linear relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. If the explanatory variables are measured with error, then errors-in-variables models are required, also known as measurement error models. A linear regression model assumes that the relationship between the dependent variable y and the vector of regressors x is linear.

$$Y = X\beta + \epsilon$$

This relationship is modeled through a disturbance term or error variable ϵ an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors [24]. Linear regression identifies the equation that produces the smallest difference between all the observed values and their fitted values. To be precise, linear regression finds the smallest sum of squared residuals that is possible for the dataset. The evaluation metrics used to appraise the performance of the regression models consisted of the coefficient of determination (R^2). R^2 evaluates the efficiency of feature selection algorithms. The coefficient of determination measures the proportion of the variance in the dependent variable that is predictable from the independent variables. The equation for R2 can be described as follows:

$$R^2 = \frac{\text{Explained variations}}{\text{Total variations}}$$

R-squared evaluates the scatter of the data points around the fitted regression line. It is also called the coefficient of multiple determination for multiple regression. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values. The approach of this article is to identify methods of feature selection and similarity that have the best performance in small data sizes. For the discussion of feature selection, in addition to the methods proposed in the research method, there are many methods in the research literature, including the combined methods that are more accurate in many cases. However, the main aim of this study is to identify methods that are computationally simple in addition to selecting the most suitable subset of data. Therefore, methods were chosen that did not have computational complexity.

3 Result

The performance of feature selection methods in different data sizes was evaluated. The average value of r-squared was measured in each step of reducing the sample size and selecting an appropriate data subset based on existing methods (Table 4).

The Var methods have the best performance based on the values of r squared. The stepwise and correlation took the next position. The lasso method has the worst performance between other methods. Among the 15 feature selection methods used in this study, the Euclidean distance method ranks 6th and the DTW method ranks 7th, which performed better than other similarity methods. The edit distance similarity method has the worst performance.

Table 4: The average value of R-squared

Method	**R-squared**	**Sensitivity to Data Size**	**Computational Complexity**
Variance Threshold	0.996092	Low	Low
Stepwise	0.995537	Moderate	Moderate
Correlation	0.994809	Moderate	Low
Backward Elimination	0.994060	Moderate	High
Recursive Feature Elimination	0.993215	Moderate	High
Euclidean Distance	0.991367	Low	Moderate
Dynamic Time Warping	0.991285	Moderate	High
Simulated Annealing	0.990358	Low	High
Tree-Based Methods	0.987900	High	Moderate
Forward Selection	0.980137	Moderate	Moderate
Hausdorff Distance	0.977381	Low	High
Mutual Information	0.977135	Low	Moderate
Fréchet Distance	0.976615	Moderate	High
Edit Distance	0.951547	Low	High
Lasso	0.704437	High	Moderate

Figure 2 shows the r-squared value of different feature selection methods in different datasets. In fact, by reducing the sample size in each step, different data sets were selected according to different methods. The horizontal axis of the graph represents the remaining percentage of the number of observations. In each step, by reducing the number of samples and choosing one, considering that the number of observations has decreased, the performance of the regression model decreases and the r-squared value in the number of low samples is lower than the number of high hub samples. In general, the value of r-squared was low in all methods at low percentages, and as the percentage increased, its value increased, and the performance of the 15 existing methods was similar and close to each other. However, the r-squared value of the lasso method was dramatically lower than other methods.

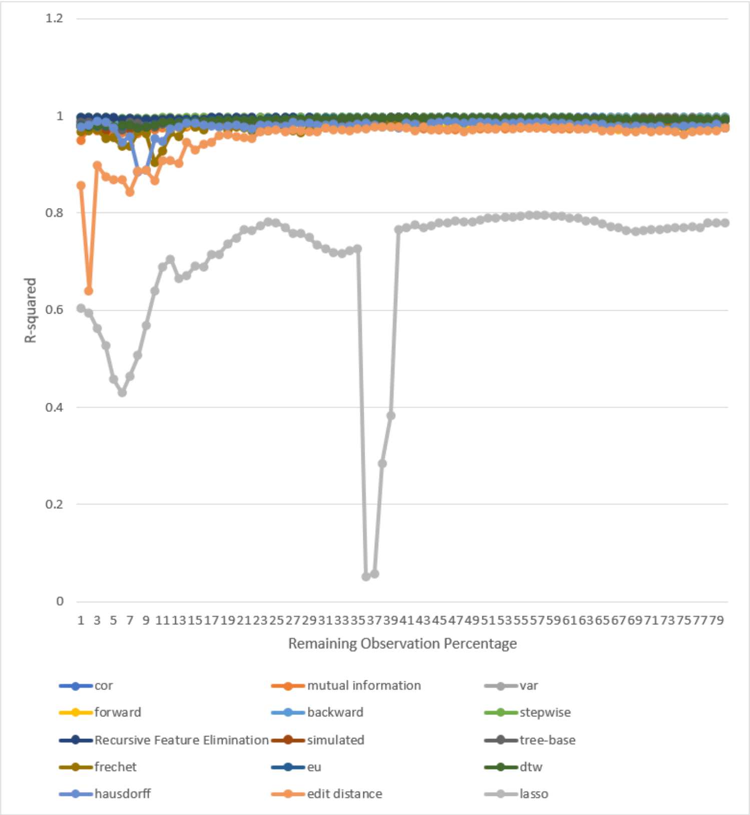

Figure 2: The value of r-squared of feature selection methods in the number of different observations

The figure shows the r-squared value of filtered methods. The trend line of each of these graphs was drawn. Regarding the slope of the trend line, among the three existing methods, the mutual information method had the lowest slope and sensitivity to the number of observations. However, the fluctuations of the r-squared value were high, which made this method less reliable. On the other hand, although the slope of the trend line of the var method was slightly higher than the mutual information method, the r-squared changes during the change in the number of observations were less than the other methods of this group. Variance Threshold (var) exhibited the highest R-squared values, indicating its effectiveness in retaining the most informative features. Its simplicity and ability to eliminate low-variance features, which are often less informative, contribute to its high performance. Stepwise and Recursive Feature Elimination showed strong performance due to their iterative approach, which systematically refines feature subsets. Although computationally intensive, their consideration of feature interactions enhances their predictive accuracy. Correlation-Based Selection performed well due to its efficiency in identifying and retaining features with strong linear relationships with the target variable. Its low computational complexity and high scalability make it suitable for large datasets. Dynamic Time Warping (DTW) and Euclidean Distance (EU) similarity methods performed better than other similarity-based methods. DTW's ability to handle temporal misalignments and Euclidean Distance's straightforward computation contributed to their relatively high performance. The Lasso method had the lowest R-squared values, which could be due to its sensitivity to the choice of the regularization parameter. Additionally, its performance may be affected by the high dimensionality and complexity of the financial data used in this study. Edit Distance and Hausdorff Distance methods showed lower performance compared to other similarity methods. The computational intensity and sensitivity to noise in Edit Distance and the limitations in capturing the trend of time series in Hausdorff Distance may explain their lower R-squared values. Tree-Based and Simulated Methods showed moderate performance. Tree-based methods, while efficient with parallel processing capabilities, may not have captured the temporal dependencies effectively. Simulated methods, such as Simulated Annealing, are computationally expensive, which may limit their scalability for very large datasets.

3.1 Sensitivity to Data Volume

The study also evaluated the sensitivity of each method to changes in data volume. The Variance Threshold, Correlation-Based Selection, and Simulated methods demonstrated the least sensitivity, maintaining relatively stable R-squared values across different sample sizes. This stability is crucial for applications in dynamic environments such as financial markets, where data volume can fluctuate significantly.

The performance results of the Wrappers methods are shown in Figure 4. Five known methods from this group were reviewed. Among these, the value of r-squared fluctuated greatly during the change of the number of samples in forward, recursive feature elimination, and stepwise methods. Among the two backward and simulated methods, the simulated method had less fluctuation with a lower slope.

From the group of embedded methods, two methods were investigated (Fig5). The Lasso method had a relatively lower slope than Tree-based. However, the r-squared value of this method is lower than the tree-based method.

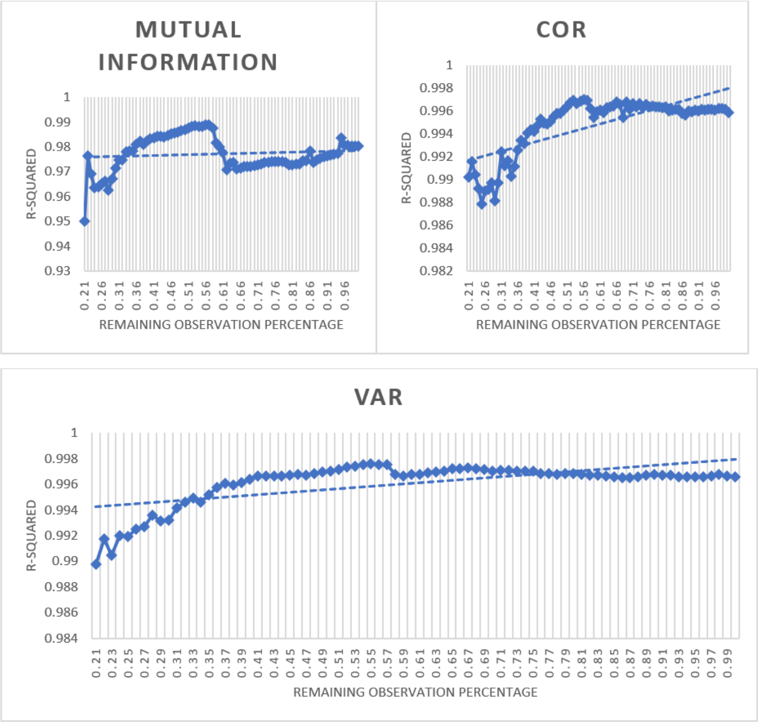

Figure 3: The value of r-squared of filtered feature selection methods

Figure 6 shows the performance of 5 similarity methods. among these five methods, the edit distance method had the lowest slope. Similarity methods had minor fluctuations during data size reduction.

3.2 Discussion

size. These factors are crucial in evaluating the robustness and reliability of feature selection methods in dynamic and high-dimensional datasets.

- **Performance Evaluation**
  Among the evaluated methods, the Variance Threshold, Stepwise, and Correlation methods consistently demonstrated the highest R-squared values. This indicates their effectiveness in retaining the most informative features and enhancing the predictive accuracy of the models. These methods are computationally efficient and scalable, making them suitable for large datasets.

  The sensitivity analysis revealed that the Mutual Information, Variance Threshold, Simulated Annealing, Edit Distance, and Hausdorff methods exhibited less sensitivity to changes in data size. This stability is essential for applications in fluctuating environments, such as financial markets, where data avail-

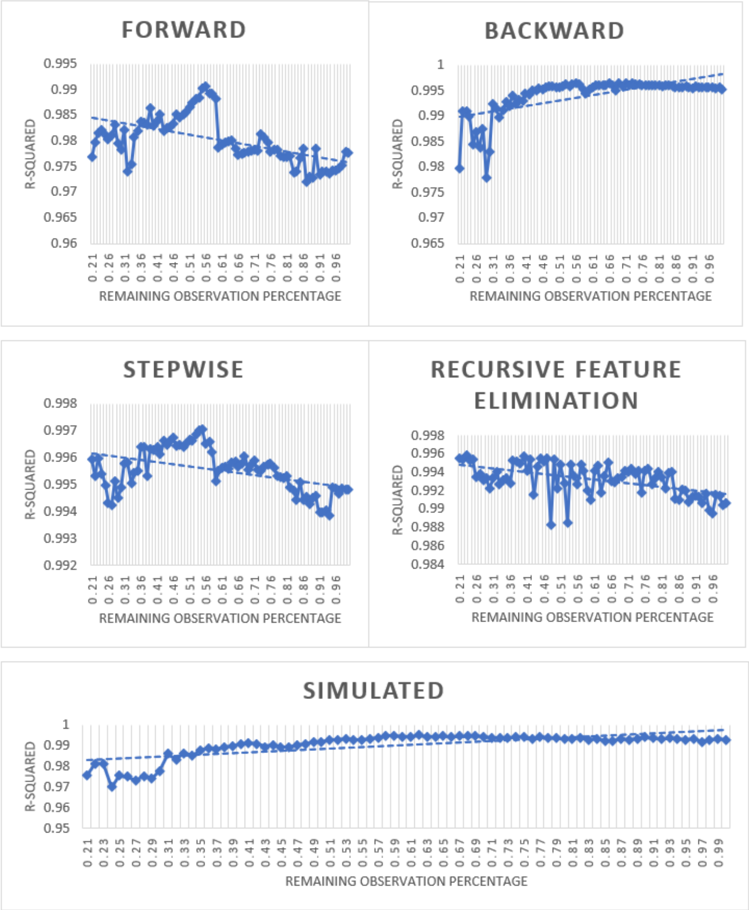

Figure 4: The value of r-squared of Wrappers methods

ability can vary significantly. The less sensitive methods ensure consistent model performance despite changes in sample size. In terms of fluctuation, the Variance Threshold, Simulated Annealing, and Edit Distance methods showed minimal performance variation across different sample sizes. This consistency further underscores their robustness and reliability in real-world scenarios where data volume is not controlled.

- **Best Overall Method**
  According to the three criteria: R-squared value, sensitivity to data size, and fluctuation, the Variance Threshold method emerged as the best overall method. It consistently performed well across all metrics, making it a reliable choice for feature selection in predictive modeling. Among the similarity-based methods, the Hausdorff and Edit Distance methods showed good perfor-

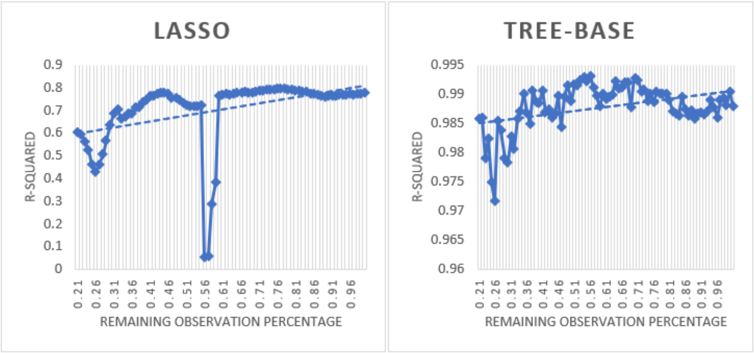

Figure 5: The value of r-squared of embedded methods

mance, indicating their potential as viable alternatives to traditional feature selection techniques.

- **Implications and Applications**
  This study focuses on improving the performance of data-driven models by selecting the most appropriate features, which is crucial for accurate and reliable predictions. The findings have significant implications for various applications, particularly in financial forecasting, where robust feature selection can lead to better investment decisions and risk management.

- **Limitations and Future Research**
  Despite the promising results, it is important to acknowledge certain limitations of this study. First, while we evaluated several feature selection methods, many existing hybrid methods were not investigated. Hybrid methods, which combine the strengths of multiple techniques, may offer improved performance and should be explored in future research. Second, the results presented in this study are based on a specific dataset of financial data from Fortune Global 500 companies. The performance of feature selection methods may vary with different datasets, particularly those from other domains or with different characteristics. Future studies should validate the findings using diverse datasets to ensure the generalizability of the results. Additionally, the computational complexity of some methods, such as Dynamic Time Warping and Fréchet Distance, may limit their scalability for very large datasets. Further research could explore optimization techniques to enhance the efficiency of these methods.

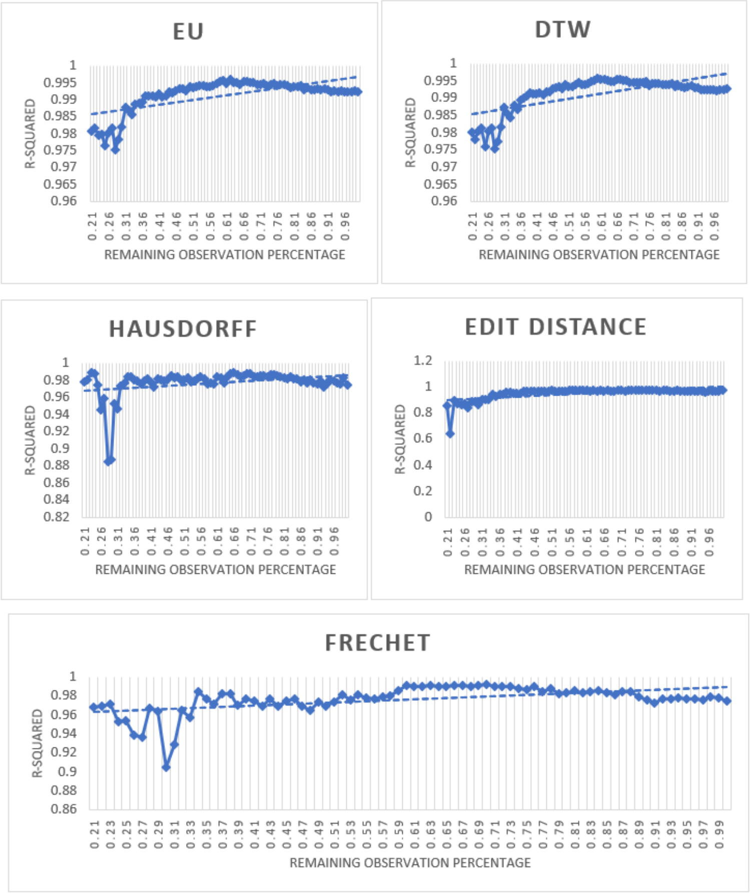

Figure 6: The value of r-squared of similarity methods

4 Conclusion

The aim of this study is to select feature selection techniques that have little sensitivity to low data size and select a subset of data that has high predictive performance in low data size. As mentioned in the results section, based on the dataset used in this study, the performance of standard feature selection methods fluctuates in different data volumes, which can reduce the level of confidence in these techniques. Among the ten standard feature selection methods, two variance and simulated methods are more stable than others. The graphs show that the similarity methods introduced as an alternative to the feature selection methods are less volatile than these methods, which increases the confidence in the results of these methods when the data size is different. Therefore, according to the first approach of this study, the similarity methods are more reliable than the usual feature selection methods.

Of course, it is essential to note that these results are only related to one data set, and the results may change in other data sets. This study will present an overall assessment of feature selection methods by pointing out their sensitivity to different sizes of data and introducing the similarity-based approaches as robust alternatives. Unlike previous studies, which often incorporate similar measures into clustering or hybrid methods, this one uniquely applies to them as stand-alone feature selection techniques. Results point to the fact that approaches like Hausdorff Distance and Edit Distance have low sensitivity with an increase in the volume of data, showing consistent performances. This therefore supports Zhu et al. [8] and Mitra et al. [9], who considered similarity for reducing redundancy and clustering features but without considering its direct predictive value. Moreover, the resistance of Variance Threshold and similarity-based approaches extends prior work such as Vabalas et al. [13] and Kuncheva et al. [17], examining small-sample feature selection with other biases that were methodological protocol in nature rather than specific algorithmic. The traditional methods, like Lasso Regression, performed very poorly for data with changing conditions, further strengthening the observation of Perry et al. [14] that the techniques should be resilient enough to capture fluctuating sample sizes. This study provides new insights, extending not only the applicability of similarity methods to feature selection but also providing a comprehensive sensitivity analysis. These findings hold significant implications for dynamic environments like financial forecasting and highlight a need for future validation across diverse domains. The second and more critical approach that the article sought to test is to examine the sensitivity of the methods to the change in data size. Indeed, any method with the least minor sensitivity to data size change will be chosen. According to trendline results, the variance, correlation methods, simulated methods, edit distance, and Hausdorff are less sensitive to observation size. Considering that time series similarity methods had the most minor fluctuation among other feature selection methods, these methods can be used as reliable methods for feature selection. Similarity methods, such as the Hausdorff and edit distance approaches, emerged as the most stable among the various feature selection techniques evaluated. This robustness across different data sizes underscores their reliability and suitability for this research context. Their consistent performance indicates that they can effectively handle fluctuations in observation numbers without significant loss of predictive accuracy. This resilience is crucial in ensuring the robustness and generalizability of predictive models, particularly in dynamic environments such as financial markets. Consequently, these methods stand out as promising tools for feature selection, offering researchers a dependable approach to identifying relevant variables for predictive modeling tasks, such as forecasting Apple's closing price.

Bibliography

[1] Y. Hmamouche, P. Przymus, A. Casali, and L. Lakhal, *GFSM: a feature selection method for improving time series forecasting*, Int. J. Adv. Syst. Meas., (2017).

[2] E. W. Newell and Y. Cheng, *Mass cytometry: blessed with the curse of dimensionality*, Nat. Immunol., 17 (2016), pp. 890–895. doi:10.1038/ni.3485.

[3] B. Remeseiro and V. Bolon-Canedo, *A review of feature selection methods in medical applications*, Comput. Biol. Med., 112 (2019). doi:10.1016/j.compbiomed.2019.103375.

[4] E. Ergüner Özkoç, *Clustering of Time-Series Data*, IntechOpen, (2021). doi:10.5772/intechopen.84490.

[5] A. Alqahtani, M. Ali, X. Xie, and M. W. Jones, *Deep Time-Series Clustering: A Review*, Electronics, 10 (23) (2021), 3001. doi:10.3390/electronics10233001.

[6] J. L. Vermeulen, *Geometric similarity measures and their applications [dissertation]*, Utrecht University, (2023).

[7] H. Xie, J. Li, and H. Xue, *A survey of dimensionality reduction techniques based on random projection*, arXiv, (2017). Available from: https://arxiv.org/abs/1706.04371.

[8] X. Zhu, Y. Wang, Y. Li, Y. Tan, G. Wang, and Q. Song, *A new unsupervised feature selection algorithm using similarity-based feature clustering*, Comput. Intell., 35 (1) (2019), pp. 2–22. doi:10.1111/coin.12192.

[9] P. Mitra, C. A. Murthy, and S. K. Pal, *Unsupervised feature selection using feature similarity*, IEEE Trans. Pattern Anal. Mach. Intell., 24 (3) (2002), pp. 301–312. doi:10.1109/34.990133.

[10] Q. Yu, S. Jiang, R. Wang, and H. Wang, *A feature selection approach based on a similarity measure for software defect prediction*, Front. Inf. Technol. Electron. Eng., 18 (11) (2017), pp. 1744–1753. doi:10.1631/FITEE.1601322.

[11] Y. Shi, C. Zu, M. Hong, L. Zhou, L. Wang, X. Wu, et al., *ASMFS: Adaptive-similarity-based multi-modality feature selection for classification of Alzheimer's disease*, Pattern Recognit., 126 (2022), 108566. doi:10.1016/j.patcog.2022.108566.

[12] X. Fu, F. Tan, H. Wang, Y. Zhang, and R. W. Harrison, *Feature similarity based redundancy reduction for gene selection*, In: Proceedings of the International Conference on Data Mining (Dmin), (2006), pp. 357–360.

[13] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, *Machine learning algorithm validation with a limited sample size*, PLoS One, 14 (11) (2019), e0224365.

[14] G. L. Perry and M. E. Dickson, *Using machine learning to predict geomorphic disturbance: The effects of sample size, sample prevalence, and sampling strategy*, J. Geophys. Res. Earth Surf., 123 (11) (2018), pp. 2954–2970. doi:10.1029/2018JF004640.

[15] Z. Cui and G. Gong, *The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features*, Neuroimage, 178 (2018), pp. 622–637. doi:10.1016/j.neuroimage.2018.06.001.

[16] L. I. Kuncheva, C. E. Matthews, A. Arnaiz-González, and J. J. Rodríguez, *Feature selection from high-dimensional data with very low sample size: A cautionary tale*, arXiv, (2020). Available from: https://arxiv.org/abs/2008.12025.

[17] L. I. Kuncheva and J. J. Rodríguez, *On feature selection protocols for very low-sample-size data*, Pattern Recognit., 81 (2018), pp. 660–673. doi:10.1016/j.patcog.2018.03.012.

[18] J. Doak, *An evaluation of feature selection methods and their application to computer security [Technical Report]*, CSE-92-18, (1992).

[19] H. Liu and L. Yu, *Toward integrating feature selection algorithms for classification and clustering*, IEEE Trans. Knowl. Data Eng., 17 (4) (2005), pp. 491–502. doi:10.1109/TKDE.2005.66.

[20] C. F. Tsai and Y. T. Sung, *Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches*, Knowl. Based Syst., 203 (2020), 106097. doi:10.1016/j.knosys.2020.106097.

[21] U. Mori, A. Mendiburu, and J. A. Lozano, *Similarity measure selection for clustering time series databases*, IEEE Trans. Knowl. Data Eng., 28 (1) (2015), pp. 181–195. doi:10.1109/TKDE.2015.2462369.

[22] M. Goldani, *A review of time series similarity methods*, In: Proceedings of the 3rd International Conference on Innovation in Business Management and Economics, (2022).

[23] S. Palkhiwala, M. Shah, and M. Shah, *Analysis of machine learning algorithms for predicting a student's grade*, J. Data Inf. Manag., 4 (2022), pp. 329–341. doi:10.1007/s42488-022-00078-2.

[24] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*, 3rd ed., Hoboken: John Wiley & Sons, 2012.

How to Cite: Mahdi Goldani[1], Soraya Asadi Tirvan[2], *Sensitivity assessing to data volume for forecasting: introducing similarity methods as suitable ones in feature selection methods*, Journal of Mathematics and Modeling in Finance (JMMF), Vol. 4, No. 2, Pages:115–134, (2024).