Comprehensive hybrid regression model for financial forecasting in neutrosophic logic

Firuz Kamalov [1] *, **Said Elnaffar**[2], **Ikhlaas Gurrib**[3], **Aswani Cherukuri**[4]
[1]Department of Electrical Engineering, Canadian University Dubai, Dubai, UAE
[2]School of Engineering, Applied Science and Technology, Canadian University Dubai, Dubai, UAE
[3]Faculty of Management, Canadian University Dubai, Dubai, UAE
[4]School of Information Systems, Vellore Institute of Technology, India
Emails: firuz@cud.ac.ae; said.elnaffar@cud.ac.ae;ikhlaas@cud.ac.ae; cherukuri@acm.org

Abstract

Regression analysis is a widely used tool in several fields. In this paper, we propose a comprehensive, multi-step regression model for financial forecasting. The proposed hybrid model combines preprocessing, feature selection, and cross-validation to obtain a powerful approach to forecasting. The extension of the proposed model to neutrosophic sets is discussed. The model is applied to the case study of real estate prices. The results demonstrate the efficacy of the model.

Keywords: regression analysis; feature selection; preprocessing; financial forecasting; hybrid model; neutrosophic set

1 Introduction

Regression analysis is common task in data science. It is used in many applications including finance, economics, science, engineering, medicine and others. In regression analysis, a continuous target variable is modeled using continuous and/or categorical attributes. The objective is to infer the relationship between the attribute and target variables. Given the importance of regression analysis it has attracted a significant amount of attention from researchers. Despite the plethora of articles on the subject, there is still room for new research and exploration.

In this paper, we propose a novel hybrid regression model with demonstrated efficacy. The proposed approach comprises multiple stages. The first stage is data preprocessing which includes data cleaning, removing extreme outliers, and data transformation. The second stage is variable selection. It involves identifying the relevant attributes as well as creating new useful variables for regression model. The final stage is model testing. After building the final model, cross-validation is used to test the accuracy of the model. Although the details of each step depend on the particular data, the general procedure is the same across different applications. We illustrate the proposed hybrid regression model using real estate data. The results demonstrate the effectiveness of the model in predicting real estate prices.

The first stage of the proposed hybrid regression model is data preprocessing. It involves data cleaning, removing extreme values, and data transformation. Data cleaning involves dealing with data that is missing or erroneous. There are three common approaches to data cleaning: i) dropping the faulty observations from the dataset, ii) extrapolating the values of the missing observations (imputation), iii) leaving the the observations in the original form and employ regression models that are capable of handling data with faulty observations.

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

Dropping or leaving the faulty observations in place requires minimal preprocessing. On the other hand, data imputation is a more complex task with several available options. Data preprocessing also includes removing potential outliers. Extreme values in data can have a significant effect on the regression model. Extreme values can occur either due to erroneous measurements or a fluke event. If extreme values produce a large skew in the regression model, then it is recommended to remove them. The final stage of data preprocessing is data transformation. The raw data is often in the form that is not suitable for regression analysis. Data transformations such as scaling, log-transform, square root-transform, and others can be used to reshape the data into the form that it appropriate for regression model. Various diagnostic plots such histograms and pair-wise plots can be consulted in determining the suitable data transformation. Data preprocessing, when done correctly, can have a significant positive effect on the regression model.

The second stage of the proposed hybrid regression model is variable (feature) selection. It is the most important stage of regression analysis. The effectiveness of the regression model depends in large part on the quality of the input variables. The old saying "garbage in, garbage out" implies that without appropriate input variables it is impossible to construct a good model. There exists several approaches to feature selection: i) filter methods, ii) wrapper methods, iii) embedded methods. Filter methods employ a generic metric such as Pearson's correlation or mutual information to identify the relevant features. Wrapper methods use a specific model to evaluate the effectiveness of a feature subset. Embedded methods perform automatic feature selection as part of the model construction. Feature extraction is another part of regression analysis. Creating new variables based on the original attributes can improve the effectiveness of the model. Feature extraction can be performed either manually or via a heuristic. Manual feature extraction involves identifying new feature combinations based on plots and using cross-validation to measure their effectiveness. Algorithmic feature selection methods are based on dimensionality reduction techniques such as principal component analysis.

The final stage of the proposed hybrid regression model is parameter estimation and model testing. Parameter estimation is done using standard software such as R, Python, Excel, and others. Confidence intervals for model parameters are estimated using the standard error. The statistical significance of the model coefficients is carried out using hypothesis test. Finally, the estimated model is tested using cross-validation and other approaches. Other measures of model performance include AIC, BIC, and C_p.

Applying neutrosophic sets to linear regression involves incorporating the elements of uncertainty, indeterminacy, and partial truth inherent in neutrosophic logic into the linear regression model. This approach can be particularly useful when dealing with imprecise, inconsistent, or incomplete data, which are common in real-world scenarios.

Our paper is structured as follows. In Section 2, we provide a brief literature review related to regression analysis. Section 3 describes the dataset used in our study. Section 4 discusses the data preprocessing stage of the model. Section 5 presents attribute selection stage of the model, while Section 6 discusses attribute extraction. In Section 7, we demonstrate estimation of the model parameters and testing the resulting model. In Section 8, we discuss extending the proposed approach to neutrosophic sets. We finish with concluding remarks in Section 9.

2 Literature

Regression analysis is a well-established yet a vibrant research topic. There are exists several works that discuss the fundamentals of regression analysis[1–3] including its use in market analysis.[4] Along with classical regression analysis, modern approaches to the topic have recently gained momentum. The progress in machine learning algorithms have led to deep regression which is able to build nonlinear models based on large quantities of data. Fuzzy regression has also gained traction with resurgence of fuzzy logic across various fields of application.

The recent advances in neural networks has sparked research in deep regression models.[5–8] Deep learning based regression models have achieved state of the art performance albeit at the requirement of large amount of data. The applications extend to various fields of financial forecasting. Unlike classical regression models, deep learning is able to learn from data with limited user input. However, as a black-box model its use is restricted to limited applications.

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

More recently fuzzy logic has reemerged as an active area of research. Many studies have applied the concept of fuzzy to investigate existing phenomena from a different point of view. In particular, fuzzy regression has emerged as a vibrant research area.[9,10] The uncertainty in variables and data is encoded in fuzzy models allowing for a uniform solution. Similarly, neutrosophic regression has been proposed as an alternative method for dealing with uncertainty.[11,12]

Another interesting research avenue has been evaluation metrics used in regression analysis. Since different metrics measure different attributes of the model, the choice of the evaluation method plays an important role. Various regression evaluation metrics were analyzed in[13] who found that the coefficient of determination R^2 is more informative than other common metrics such as SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation.

Other modern development in regression analysis include uncertainty models to deal with imprecise observations.[14,15] Threshold regression analysis was applied in the context of financial inclusiveness and economic growth.[16]

The applications of regression analysis have included a range of fields including financial forecasting. Performance of public sector banks in the case of India was analyzed in.[17] Sparse regression has been applied to forecast bitcoin prices.[18] Deep learning regression has been used in medical applications such as predicting Covid-19 cases.[19] In[20] the authors employed quantile regression analysis to study the effects of leverage on firm performance. A variant of quantile regression based on instrumental variables was applied in[21] to study the effect of geographical factors on Islamic banking sustainability performance. The effect of oil prices on inflation was analyzed in[22] using correlation regression analysis. Autocovariance-based regression properties were studied in.[23]

3 Data

The goal of this paper is to propose a new comprehensive, multi-stage hybrid regression model in the context of financial forecasting. To illustrate the proposed approach we consider a case study in real estate forecasting. In particular, we consider the data characterizing real estate valuation in New Taipei City, Taiwan that is available at UCI ML repository.[24] The data consists of six independent variables: "transaction date", "distance to the nearest MRT station", "house age", "number of convenience stores", "latitude", and "longitude." The dependent variable is "house price of unit area." To avoid excessive verbiage we abbreviate the variables as `date`, `mrt`, `age`, `stores`, `lat`, `long`, and `price`. The total number of observations in the data set is 414.

4 Data preprocessing

The initial stage of the proposed hybrid regression model is data preprocessing. It is one of the key parts of the model. It consists of several steps including data cleaning, removing outliers, and data transformation. Real-life data often contains missing or erroneous samples. Therefore, it is first necessary to clean the data. The simplest approach to cleaning the data is simply removing the observations that are incomplete or erroneous. However, it may substantially reduce the amount of data available for analysis. To preserve the samples with missing or incomplete information, researchers often use imputation. Imputation can be done in several different ways. The missing values can be replaced either by the sample mean or local regression outcomes. The current data set does not contain any faulty samples so there is no need to clean the data in our case.

The next step, after cleaning the data, is to remove any potential outliers that can have an inappropriate effect on the regression fit. Outliers are extreme value points that have excessive influence on the regression fit which distorts the true underlying picture. One approach to identifying outliers is by studying the data histograms. As shown in Figure 1, the `price` variable has a significant outlier. While all the `price` values lie below the threshold 80, there is one observation near 120. Since there is a huge gap between the outlier and the rest of the data. It is recommended to remove the corresponding sample from the data to avoid skewing the regression plane.

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

To confirm the effect of the outlier, we look at the diagnostic plots of a simple regression. As shown in Figure 2, the observation 271 stands out among the others. In particular, it is close to Cook's distance line. It also lies substantially apart from the rest of the data points in the Scale-Location plot. Given the extreme value of the observation 271 together with the diagnostic plots, we decide to remove the observation from the dataset. Note that by using 413 observation in place of the original 414 observation we loose very little information from the dataset. The histogram of `mrt` in Figure 1 also indicates potential outliers. However, given that the extreme values in `mrt` are not completely isolated from the rest of the data, we do not consider them as outliers.

The last stage of data preprocessing is data transformation. Real-life data is often unsuitable in raw form to carry out effective regression analysis. Therefore, data transformation is necessary to reshape the data to obtain a more accurate regression fit. There exist several approaches to data transformation including scaling and functional transformation. Scaling is a linear transformation, while functional transformation is nonlinear.

There are two common scaling techniques. First, scaling the data to the interval [0, 1] is given by the equation

$$x_{tr} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{1}$$

Second, scaling the data to have mean 0 and standard deviation 1 is given by the equation

$$x_{tr} = \frac{x - \bar{x}}{s}, \tag{2}$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. Scaling allows the optimization algorithms to run more efficiently and helps avoid potential local minima. In addition, scaling transforms all the feature variables into the same range of values which allows to compare the magnitude of model coefficients and judge their significance.

Functional transformation is another widely used technique to preprocess the data. Functional transformations aim to improve the distribution of feature values. It is often used to decompress values that are concentrated over a small interval. Popular functional transformations include log and square root functions,

$$x_{tr} = \log(x),\ x_{tr} = \sqrt{x}. \tag{3}$$

In addition, the sigmoid and inverse sigmoid functions are also used in functional transformations. As shown in Figure 1, the distribution of `mrt` is concentrated near 0. Therefore, it would be beneficial to transform the `mrt` variable via the log transform. In addition, we scale all the variables to have mean 0 and standard deviation 1. To visualize the transformed variables we draw the pairwise graphs as shown in Figure 3. The pairwise plots show no extreme outliers. We also see no significantly abnormal distribution of the points. It follows that data preprocessing was successful.

The pairwise scatter plots in Figure 3 can also be used to study relationships between the variables. The plots show that there is potential correlation between the pairs of variables (`mrt`, `price`), (`lat`, `price`), (`long`, `price`), (`stores`, `price`), (`age`, `price`), as well as (`mrt`, `age`), (`mrt`, `stores`). In particular, `mrt` and `age` appear to have a negative relationship with the outcome `price`, while `long`, `lat`, and `stores` have a positive relationship with the outcome. Although several variables are correlated with the outcome, they may not be significant in the final model if their effect is reflected by another model parameter. In fact, the correlation between `lat` and `long` may result in the exclusion of one of them from the final model.

Understanding pairwise correlations is an important step in regression analysis. To obtain a more precise measure of correlations we consider Pearson's correlation between variables as shown in Table 1.

The correlation matrix shows that `mrt` and `stores` are the most correlated variables with respect to `price`. It is surprising to see that `age` has a low correlation with the outcome.

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

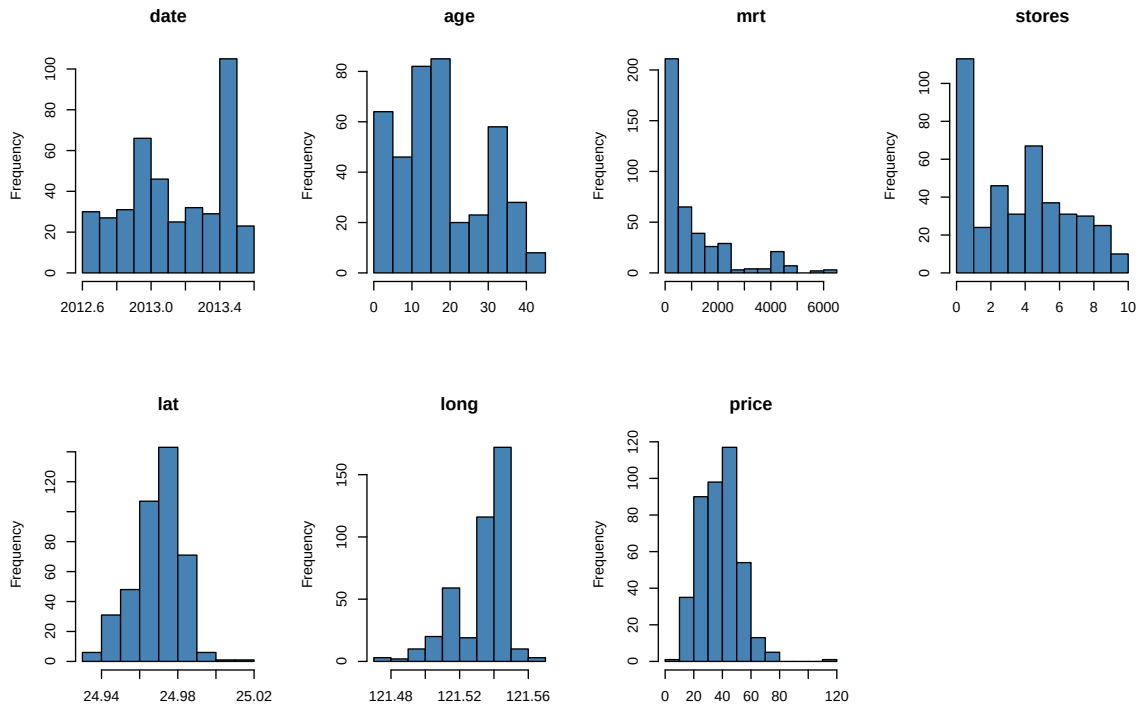

Figure 1: Outlier detection based on distribution of variable values.

Table 1: Pearson correlations

	date	age	mrt	stores	lat	long	price
date	1.00	0.02	0.10	0.01	0.03	-0.04	0.08
age	0.02	1.00	0.07	0.05	0.06	-0.05	-0.21
mrt	0.10	0.07	1.00	-0.69	-0.46	-0.65	-0.76
stores	0.01	0.05	-0.69	1.00	0.45	0.45	0.61
lat	0.03	0.06	-0.46	0.45	1.00	0.41	0.56
long	-0.04	-0.05	-0.65	0.45	0.41	1.00	0.55
price	0.08	-0.21	-0.76	0.61	0.56	0.55	1.00

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

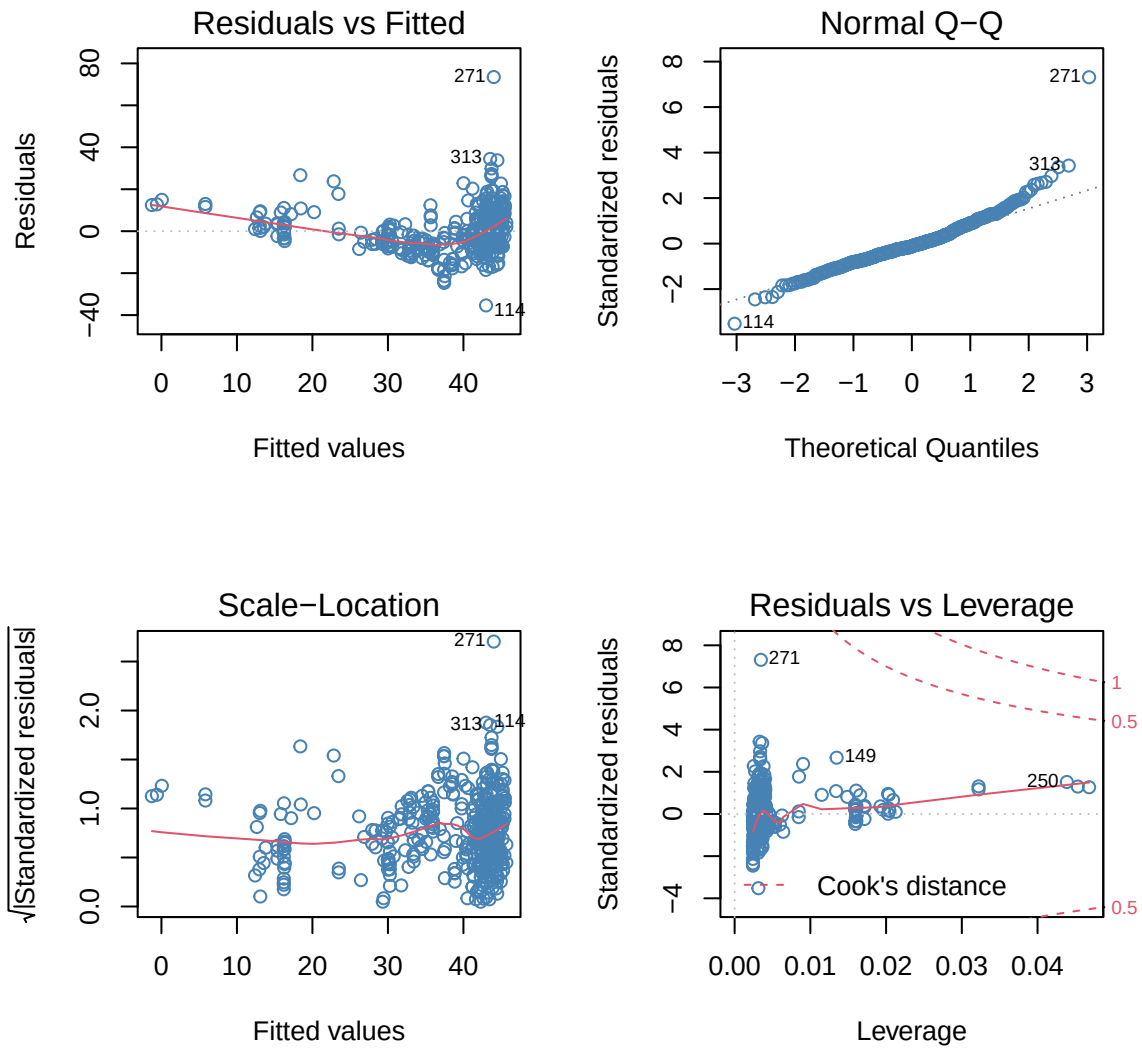

Figure 2: Diagnostic plots in simple linear regression.

https://doi.org/10.54216/IJNS.230321
Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

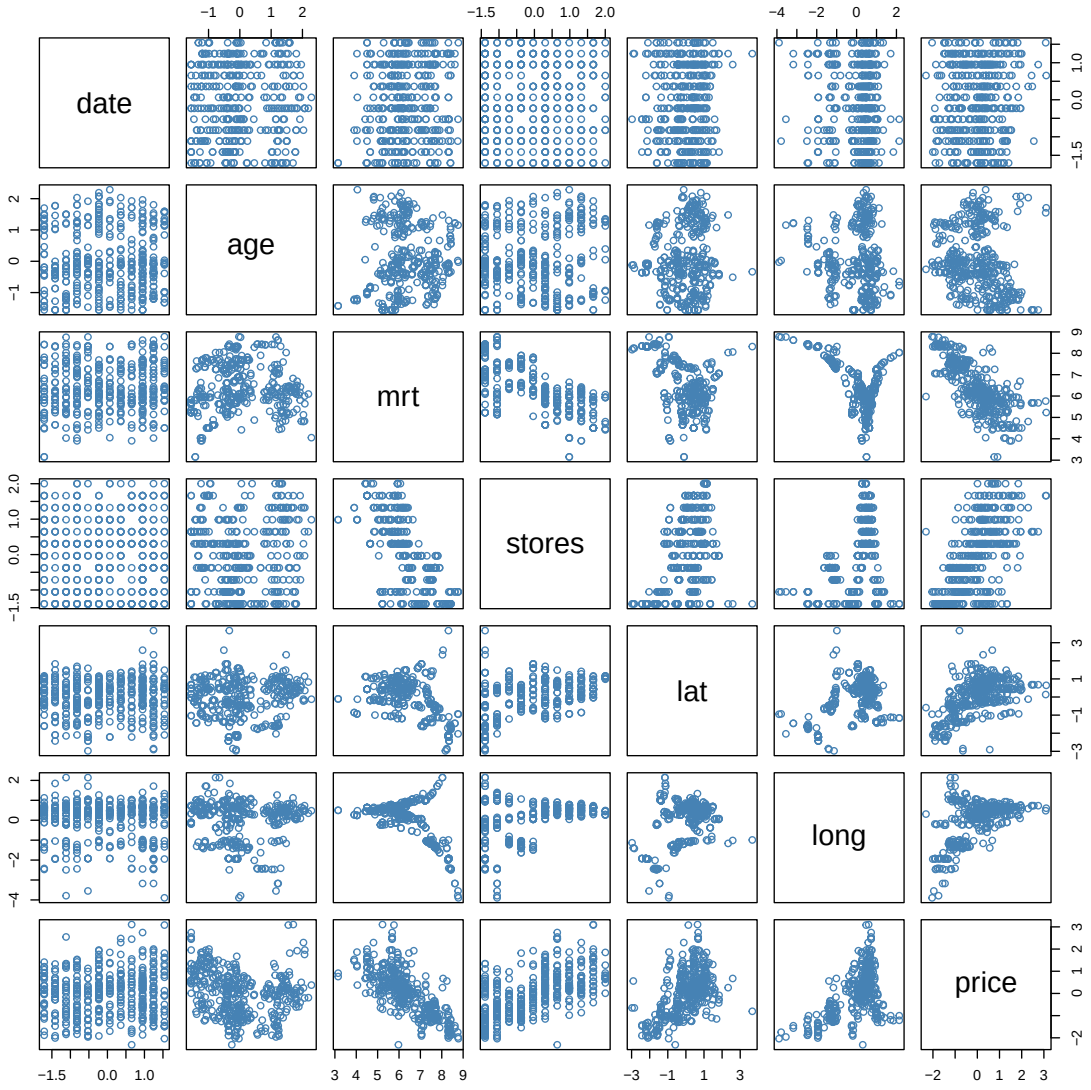

Figure 3: Pairwise distributions of transformed variables.

https://doi.org/10.54216/IJNS.230321
Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

5 Variable selection

Variable (feature) selection and extraction is arguably the most critical step in the proposed hybrid regression model. Variable selection refers to identifying the optimal subset of input variables to model the outcome variable.[25] Variable extraction refers to creating new variables based on the existing features. Effective variable selection and extraction can lead to significant positive results in regression analysis.

The simplest approach to variable selection is to use all the variables in the dataset to fit a linear regression and select the variables with low p-values. As shown in Table 2, all the variables with exception of `long` have low p-values and are statistically significant in the model. Thus, based on the above approach, we select the variables `date`, `mrt`, `age`, `stores`, and `lat` in the optimal subset.

One of the common problems in feature selection is feature interaction. While a variable can have a significant association with the outcome, its interaction with other variables may change the extent of its effect. One way variable interaction is manifested is via colinearity. To identify colinearity among the variables we utilize variance inflation factor (VIF). It is commonly used to measure multicollinearity among the predictor variables. VIF value over 5 or 10 is often considered high and indicates the presence of multicollinearity. In our case, a shown in Table 3, the VIF values are well below 5 which means that there is no significant amount of multicollinearity in the feature subset.

Although selecting variables based on the p-values is simple and efficient, there exist a plethora of more sophisticated heuristics for selecting the optimal subset of features. We utilize four of the most commonly used approaches to identify the optimal subset in our dataset. In particular, we employ forward selection, backward selection, exhaustive search, and seqrep methods to identify the optimal subset. In forward selection the optimal subset is constructed iteratively by adding one feature at a time starting with a subset of size 1, while in backward selection the optimal subset is obtained iteratively by removing one feature at a time starting with the set of all features. Exhaustive search checks every possible combination of features to identify the optimal subset. Figure 4 shows R^2, adjusted R^2, C_p, BIC, and RSS values over a range of model complexity. Note that the forward, backward, and exhaustive search methods produce the same results. The plots of the metrics for evaluating model quality show that the optimal subset size is either $m = 5$ or $m = 6$. In particular, the maximum of R^2 and the adjusted R^2 as well as the minimum of RSS are reached at $m = 5$ and $m = 6$ for all four selection methods. On the other hand, the minimum value of BIC and C_p is achieved at $m = 5$ for all four selection methods. We conclude that $m = 5$ is the optimal size of subset.

Figure 5 shows the variables selected in each optimal subset according to the size of the subset. As shown in Figure 5, the variables selected at each subset size are consistent for backward, exhaustive, and forward selection methods. In particular, the variables are selected in the following order: 1) `mrt`, 2) `lat`, 3) `age`, 4) `date`, 5) `stores`, 6) `long`. The variables selected by seqrep are the same as the other methods, except for the subset of size 4. The results of the feature selection algorithms are in line with the simple approach above using p-values. We conclude that the optimal feature subset based on the above approach consists of the attributes `mrt`, `lat`, `age`, `date`, and `stores`.

To validate the optimal subset selected by the above approach we perform a bootstrap sampling experiment. In this experiment, the data is repeatedly sampled into training and testing subsets. Then a regression model is fitted using the optimal subset of features. Finally, the test set is used to measure the accuracy of the model. The bootstrap sampling is conducted 100 times. The results of the bootstrap sampling experiments are presented in Figure 6. As shown in Figure 6, the minimum average bootstrap test error is achieved at $m = 5$ and $m = 6$. To be precise, the average MSE at $m = 5$ is 0.3236, while the average MSE at $m = 6$ is 0.3232 for all four methods. This is generally consistent with the metric plots obtained using above, where R^2, adjusted R^2 and RSS where also equal at $m = 5$ and $m = 6$. However, the BIC plot above indicated that $m = 5$ is the optimal subset size. Since $m = 5$ achieves the same test error as $m = 6$, but with fewer variables, it is the preferred choice for the optimal subset size.

Figure 7 shows the membership of the optimal subsets over all the 100 bootstrap experiments. We observe that the order and consistency of the selected variables based on the bootstrap resampling is the same for backward, exhaustive, and forward algorithms. The first three variables selected are in order: `mrt`, `lat`, and `age`. The fourth variable selected is most often `date` followed by `stores`. These observations are consistent with the results obtained using previous approaches.

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

Table 2: Feature selection based on regression p-values

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	3.00	0.26	11.44	0.00
date	0.13	0.03	4.63	0.00
age	-0.20	0.03	-7.07	0.00
mrt	-0.47	0.04	-11.50	0.00
stores	0.13	0.04	3.21	0.00
lat	0.26	0.03	8.02	0.00
long	0.04	0.04	1.09	0.27

Table 3: Variance inflation factors.

date	age	mrt	stores	lat	long
1.0277	1.0274	2.7641	2.0320	1.3731	1.7896

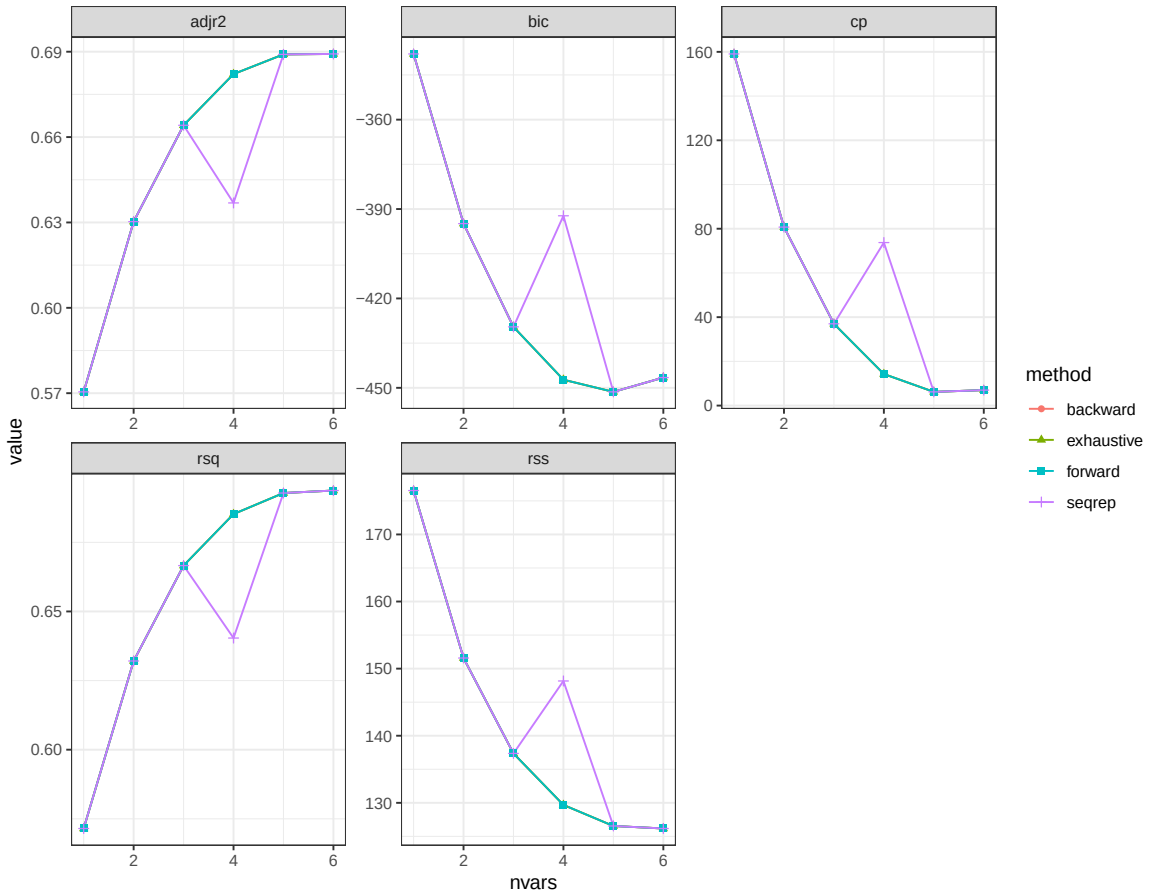

Figure 4: Feature subset evaluation for different feature selection method.

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

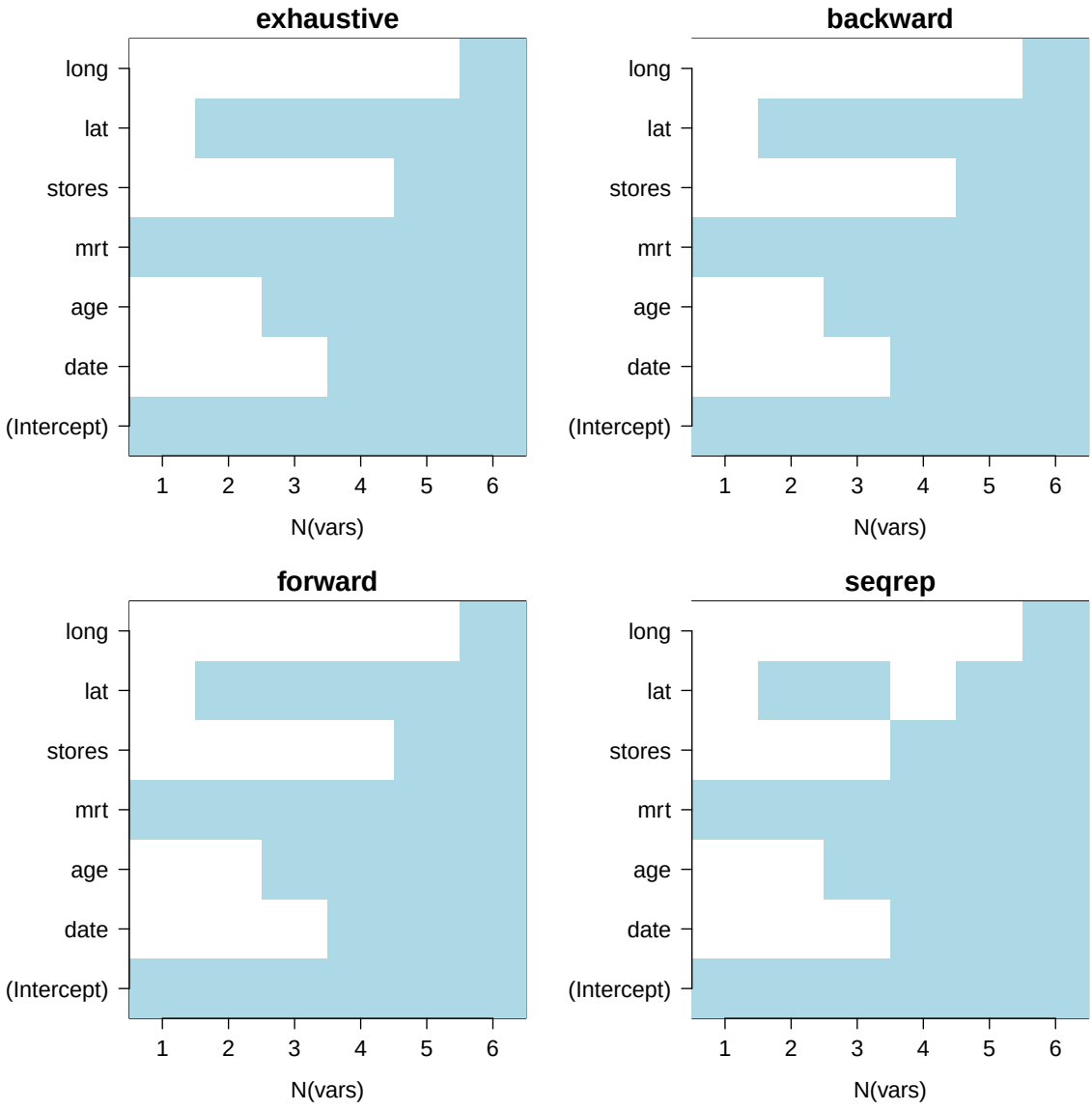

Figure 5: Selected features for different subset sizes.

https://doi.org/10.54216/IJNS.230321
Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

6 Feature extraction

The next stage in the proposed hybrid regression model is feature extraction. In this step, the existing features are used to construct new features that can help infer the outcome. The most commonly extracted features are polynomial combinations of the original variables. In our paper, we investigate first order polynomial feature interactions. In particular, we consider the set of features Q={`age.age`, `age.date`, `age.lat`, ..., `lat.lat`}.

In Figure 8, we present the regression error for a sequence of models starting with the original set of attributes ($m = 6$) and iteratively adding 1 attribute from the set Q. As expected, the addition of new variables reduces the train MSE of the model. However, the decreased MSE must be considered with caution as the increase in the number of parameters can lead to overfitting. Upon examining Figure 10, we observe that the only feature that significantly reduces the train MSE is the variable `age.age`. The remaining variables do not display a sufficiently steep reduction in error to be considered further.

To verify if the addition of the variable `age.age` is statistically significant, we compare the performance of the model with the original 6 inputs to the performance of the second model consisting of the original inputs plus the the variable `age.age`. As shown in Table 4, the addition of the variable `age.age` increases the model R^2 and adjusted R^2. It also decreases the model error (RSE), BIC, and AIC. Thus, the variable `age.age` improves our model across all criteria. Finally, the ANOVA test comparing the two models - with and without the variable `age.age` - produces the F-statistic 1.218e-08 which is statistically significant. It indicates that the addition of the variable `age.age` improves the model beyond random chance.

7 Model testing

The last stage of the hybrid regression model is parameter estimation and model evaluation. The optimal set of features selected based on the above discussion are `date`, `mrt`, `age`, `stores`, `lat`, and `age.age`. We fit a linear regression model to the data using the given set of attributes. As shown in Table 6, the p-values of the model parameters show that all the features are highly significant. The adjusted R^2 of the model is 0.712 and the model F-statistic is 2.2e-16.

Table 7 shows the 95% confidence interval for the model parameters. We observe that `age` and `mrt` variables have a negative effect on the real estate price. As the age or the distance to MRT increase, the price of the house decreases (on average). The other variables have a positive relation to the house price. It is worth mentioning the coefficient of `age.age` is positive which moderates the negative effect of the linear term `age`.

To estimate the out-of-sample error of the model we employ bootstrap sampling. The data is repeatedly divided (100 times) into bootstrap-sampled training and test subsets. The linear regression model is fitted on the training set using the selected subset of features from above. Then the MSE is calculated on the corresponding test set. The mean test MSE over 100 trials is 0.3016 and the standard deviation is 0.0345.

8 Neutrosophic regression

A neutrosophic set is a concept in the field of mathematics and logic, particularly in the area of fuzzy logic. It was introduced by Smarandache[26] to handle uncertainty, indeterminacy, and inconsistency in information. The concept extends the idea of fuzzy sets and intuitionistic fuzzy sets, which were designed to deal with vagueness and imprecision.

In traditional set theory, an element either belongs to a set or does not. Fuzzy set theory extended this by allowing degrees of membership, ranging from 0 (completely outside the set) to 1 (completely inside the set). Intuitionistic fuzzy sets further extended this concept by introducing a degree of non-membership in addition to the degree of membership. Neutrosophic sets take this a step further by introducing a third component: the degree of indeterminacy. It means that for each element in the universe of discourse, a neutrosophic set assigns three parameters:

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

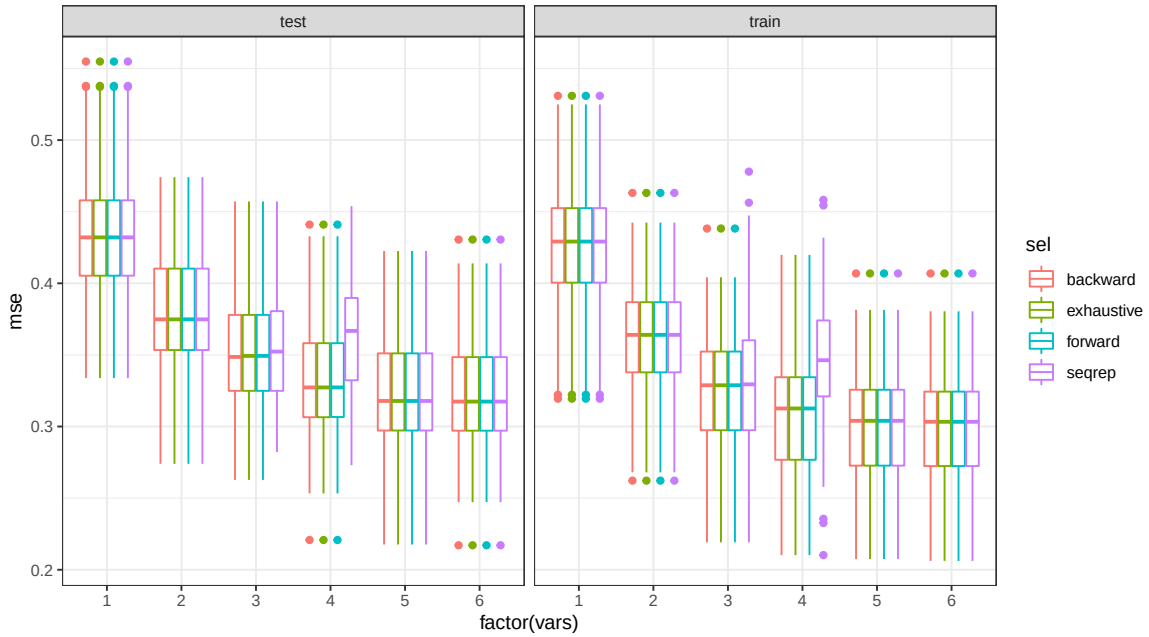

Figure 6: Bootstrap regression errors for different feature subset sizes.

Table 4: Comparison of models with and without age.age variable.

	Model 1	Model 2
r.squared	0.69	0.72
adj.r.squared	0.69	0.71
sigma	0.56	0.54
aic	698	667
bic	730	703

Table 5: Analysis of variance (ANOVA) for models with and without age.age variable.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	406	126.18				
2	405	116.45	1	9.73	33.84	0.0000

Table 6: Variable relevance based on regression p-values.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5598	0.2421	10.57	0.0000
date	0.1331	0.0268	4.97	0.0000
age	-0.2666	0.0293	-9.09	0.0000
mrt	-0.4267	0.0360	-11.85	0.0000
stores	0.1319	0.0377	3.50	0.0005
lat	0.2668	0.0305	8.74	0.0000
age.age	0.1690	0.0292	5.79	0.0000

Table 7: The 95% confidence interval for the model parameters.

	2.5 %	97.5 %
(Intercept)	2.08	3.04
date	0.08	0.19
age	-0.32	-0.21
mrt	-0.50	-0.36
stores	0.06	0.21
lat	0.21	0.33
age.age	0.11	0.23

https://doi.org/10.54216/IJNS.230321
Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

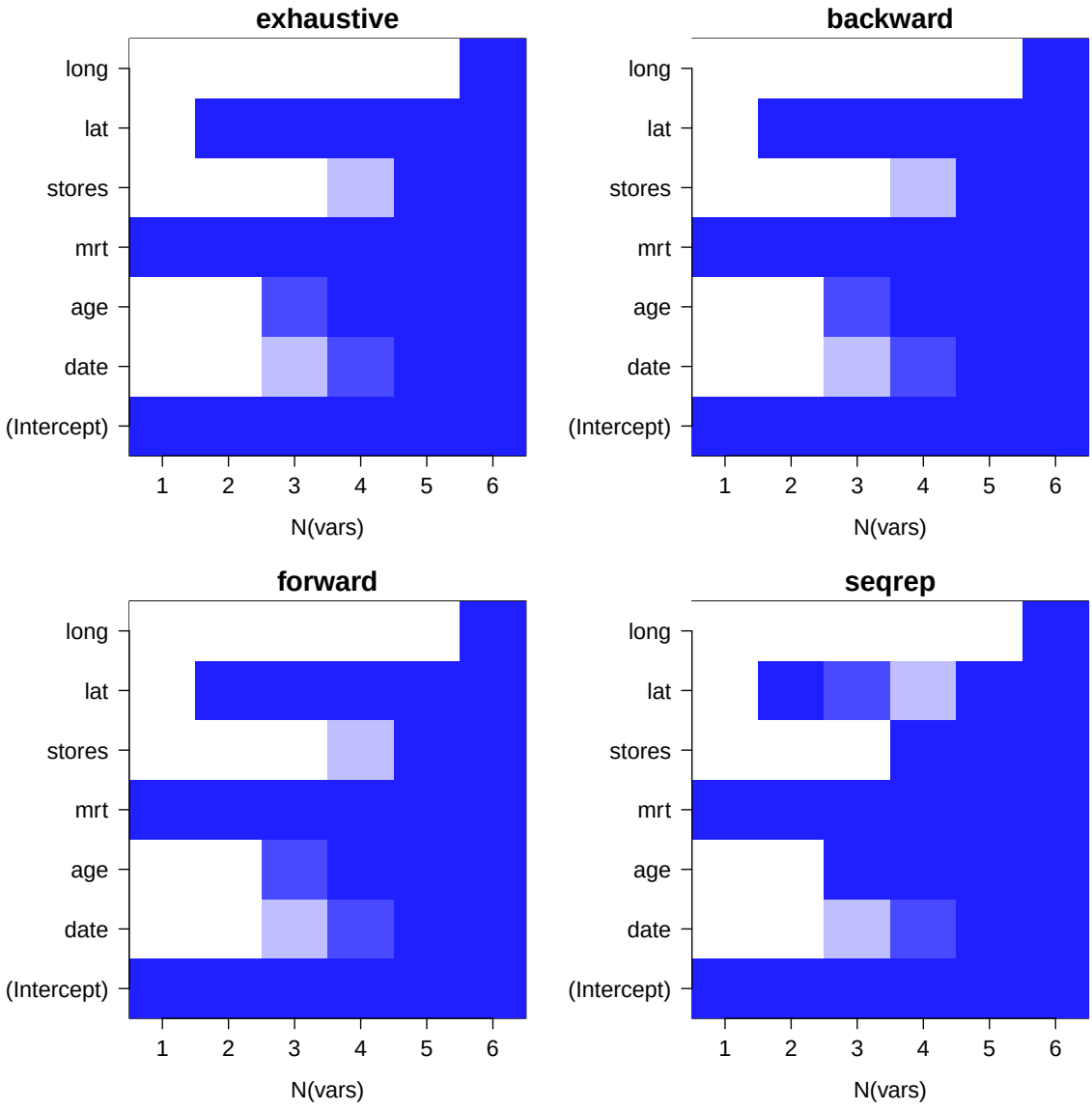

Figure 7: Selected features for different subset size based on bootstrap simulation.

https://doi.org/10.54216/IJNS.230321
Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

1. Truth Membership (T) represents the degree to which the element belongs to the set. It's similar to the degree of membership in fuzzy sets.
2. Indeterminacy Membership (I) is a unique feature of neutrosophic sets. It represents the degree of indeterminacy, or the extent to which it is unknown or indeterminate whether the element belongs to the set.
3. Falsity Membership (F) represents the degree to which the element does not belong to the set, akin to the degree of non-membership in intuitionistic fuzzy sets.

Each of these parameters is a value in the interval [0, 1], but unlike in fuzzy and intuitionistic fuzzy sets, the sum of these values can be greater than 1 in neutrosophic sets which allows for a much richer and more flexible representation of uncertainty and partial information.

Integrating neutrosophic sets into linear regression means embedding the aspects of uncertainty, indeterminacy, and partial truth, which are key to neutrosophic logic, into the framework of linear regression. It is especially beneficial in situations involving data that is imprecise, inconsistent, or incomplete, circumstances frequently encountered in real-world settings.[27,28] An outline of how neutrosophic sets can be integrated into a linear regression framework is given below:

1. In a neutrosophic context, each data point in the regression model would have associated degrees of truth, indeterminacy, and falsity which means that for each observed value y_i corresponding to x_i, there are neutrosophic values $T(y_i)$, $I(y_i)$, and $F(y_i)$ representing the truth, indeterminacy, and falsity of the data point, respectively. The linear regression model is typically of the form $y = \beta_0 + \beta_1 x + \epsilon$, where ϵ is the error term. In a neutrosophic framework, this model needs to be adapted to handle the neutrosophic values. The parameters (like β_0 and β_1) and error terms may also be represented as neutrosophic numbers to encapsulate uncertainty and indeterminacy.
2. Traditional parameter estimation like ordinary least squares might be modified to account for the neutrosophic nature of the data which could involve minimizing the sum of squared residuals in a way that accounts for the truth, indeterminacy, and falsity values of each observation. The estimation process would need to handle the uncertainty and indeterminacy in the data. It might require specialized optimization techniques that can work with the neutrosophic values.
3. Standard error metrics (like R-squared, mean squared error) might need to be redefined or adapted to account for the neutrosophic values. Interpreting the results of a neutrosophic linear regression model requires understanding the role of indeterminacy and uncertainty in the model's predictions and coefficients.
4. Integrating neutrosophic logic into linear regression adds a layer of complexity to the model formulation, estimation, and interpretation. The computation for parameter estimation and model evaluation might be more demanding than traditional regression analysis.

9 Conclusion

In this paper, we proposed a multi-step hybrid regression model. The proposed model provides a comprehensive approach to regression for financial forecasting. We illustrated the procedure for the proposed regression model in the context of real estate data. The final model is constructed using the set of features `date`, `mrt`, `age`, `stores`, `lat`, and `age.age`. The model mean test bootstrap MSE is 0.3016 with standard deviation 0.0345.

The comprehensive approach to regression analysis ensures stable and accurate forecasts following the best practices in the literature. The proposed model encompasses three critical stages: data preprocessing, feature selection, and model cross validation. Each stage is further subdivided into various tasks leading to a full regression analysis.

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

References

[1] Arkes, J. (2023). Regression analysis: a practical introduction. Taylor & Francis.

[2] Gunst, R. F., & Mason, R. L. (2018). Regression analysis and its application: a data-oriented approach. CRC Press.

[3] Lawrence, K. D. (2019). Robust regression: analysis and applications. Routledge.

[4] Sarstedt, M., Mooi, E., Sarstedt, M., & Mooi, E. (2019). Regression analysis. A concise guide to market research: The process, data, and methods using IBM SPSS Statistics, 209-256.

[5] Kamalov, F. (2020). Forecasting significant stock price changes using neural networks. Neural Computing and Applications, 32, 17655-17667.

[6] Kamalov, F., Smail, L., & Gurrib, I. (2020, December). Forecasting with deep learning: S&P 500 index. In 2020 13th International Symposium on Computational Intelligence and Design (ISCID) (pp. 422-425). IEEE.

[7] Lathuilière, S., Mesejo, P., Alameda-Pineda, X., & Horaud, R. (2019). A comprehensive analysis of deep regression. IEEE transactions on pattern analysis and machine intelligence, 42(9), 2065-2081.

[8] Rajabi, S., Roozkhosh, P., & Farimani, N. M. (2022). MLP-based Learnable Window Size for Bitcoin price prediction. Applied Soft Computing, 129, 109584.

[9] Chachi, J., Taheri, S. M., & D'Urso, P. (2022). Fuzzy regression analysis based on M-estimates. Expert Systems with Applications, 187, 115891.

[10] Chukhrova, N., & Johannssen, A. (2019). Fuzzy regression analysis: systematic review and bibliography. Applied Soft Computing, 84, 105708.

[11] Nagarajan, D., Broumi, S., Smarandache, F., & Kavikumar, J. (2021). Analysis of neutrosophic multiple regression. Neutrosophic Sets and Systems, 43, 44-53.

[12] Omar, R. H., Kayali, M. N., & Zeina, M. B. (2023). Neutrosophic Multinominal Logistic Regression Technique for Optimizing Adaptive Reuse of Historical Castles. International Journal of Neutrosophic Science, 21(3), 56-6.

[13] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623.

[14] Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. Econometrica, 88(1), 265-296.

[15] Yao, K., & Liu, B. (2018). Uncertain regression analysis: An approach for imprecise observations. Soft Computing, 22, 5579-5582.

[16] Nizam, R., Karim, Z. A., Rahman, A. A., & Sarmidi, T. (2020). Financial inclusiveness and economic growth: New evidence using a threshold regression analysis. Economic research-Ekonomska istraživanja, 33(1), 1465-1484. Chicago

[17] Pervez, A., & Ali, I. (2022). Robust regression analysis in analyzing financial performance of public sector banks: a case study of India. Annals of Data Science, 1-15.

[18] Gurrib, I., Kamalov, F., Starkova, O., Elshareif, E. E., & Contu, D. (2023). Drivers of the next-minute Bitcoin price using sparse regressions. Studies in Economics and Finance.

[19] Zgheib, R., Chahbandarian, G., Kamalov, F., El Messiry, H., & Al-Gindy, A. (2023). Towards an ML-based semantic IoT for pandemic management: A survey of enabling technologies for COVID-19. Neurocomputing, 528, 160-177.

[20] Ghardallou, W. (2023). The heterogeneous effect of leverage on firm performance: a quantile regression analysis. International Journal of Islamic and Middle Eastern Finance and Management, 16(1), 210-225. Chicago

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

[21] Budiman, T., Febrian, E., & Azis, Y. (2022). The effect of geographical factors on Islamic banking sustainability performance: an instrumental variable quantile regression analysis. Asian Economic and Financial Review, 12(2), 70-88. lied Soft Computing, 84, 105708.

[22] Abuselidze, G. (2022, July). The Influence of Changes in Oil Prices at the Inflation Levels: A Correlation-Regression Analysis. In International Conference on Computational Science and Its Applications (pp. 45-57). Cham: Springer International Publishing.

[23] Kamalov, F. (2020). A note on the autocovariance of p-series linear process. Gulf Journal of Mathematics, 9(2), 40-45.

[24] Real estate valuation data set. (2018). UCI Machine Learning Repository. https://doi.org/10.24432/C5J30W.

[25] Thabtah, F., Kamalov, F., Hammoud, S., & Shahamiri, S. R. (2020). Least Loss: A simplified filter method for feature selection. Information Sciences, 534, 1-15.

[26] Smarandache, F. (1999). A unifying field in Logics: Neutrosophic Logic. In Philosophy (pp. 1-141). American Research Press.

[27] Ouallane, A. A., Bakali, A., Bahnasse, A., Broumi, S., & Talea, M. (2022). Fusion of engineering insights and emerging trends: Intelligent urban traffic management system. Information Fusion.

[28] Govindan, K., Ramalingam, S., & Broumi, S. (2021). Traffic volume prediction using intuitionistic fuzzy Grey-Markov model. Neural Computing and Applications, 33(19), 12905-12920.

https://doi.org/10.54216/IJNS.230321

Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024

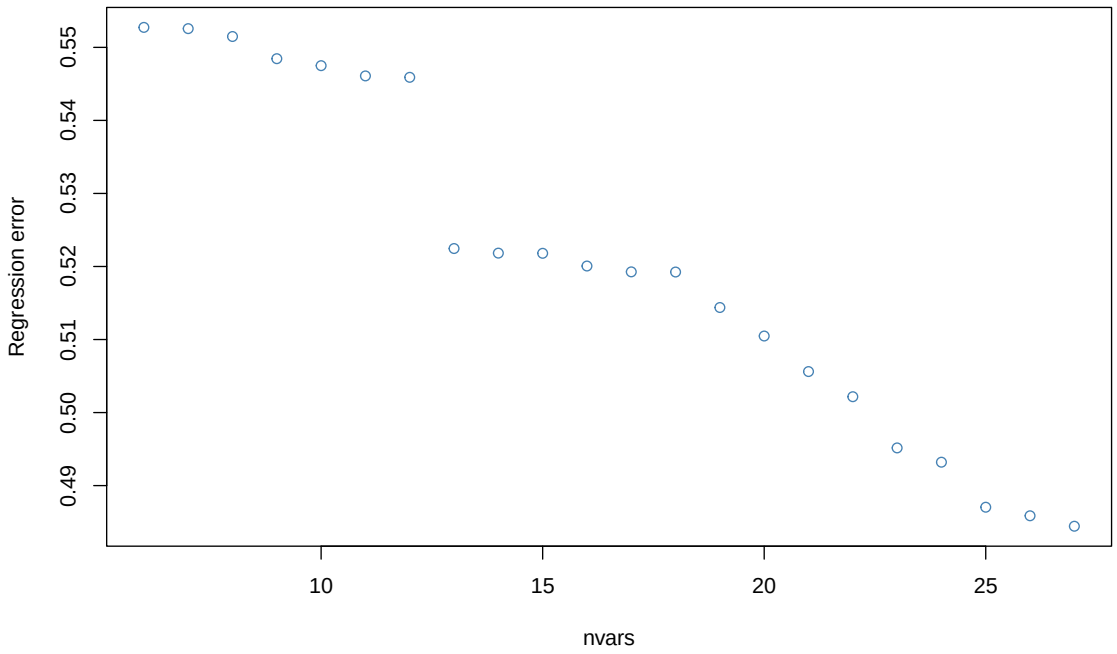

Figure 8: Regression error relative to the number of model variables.

https://doi.org/10.54216/IJNS.230321
Received: July 21, 2023 Revised: November 21, 2023 Accepted: February 09, 2024