Sales Prediction Based on Lasso Regression

Mengyu Xu*

Department of Computational and Applied Mathematics, University of Chicago, Chicago, the United States

*Corresponding author: mxu09@uchicago.edu

Abstract. Sales prediction is a critical aspect for businesses across diverse fields, providing them with the means to operate efficiently and achieve success. It constitutes an integral component of the decision-making and planning processes within a business. Several forecasting models are available for sales prediction, with most machine learning models performing exceptionally well. However, the suitability of these models can vary depending on the dataset provided. While many datasets are not overly complex and contain a limited number of variables, others are more intricate, featuring numerous variables, many of which may be irrelevant and could potentially skew the results. Therefore, the goal is to eliminate these irrelevant variables and identify those that are more closely correlated with the prediction task. The Least Absolute Shrinkage and Selection Operator (LASSO) regression model emerges as a valuable tool for the removal of irrelevant data. This study employed the Lasso regression model to analyze a housing sales dataset and discovered that it provided an excellent fit for prediction without overfitting the training set. This study presents an overview of the advantages of LASSO, discusses its limitations, and offers insights into its potential for future development.

Keywords: LASSO regression; machine learning; sales prediction.

1. Introduction

The necessity for accurate sales prediction has grown to be an essential requirement in the successful management of most businesses [1]. Sales prediction is a process of predicting future sales for products in various fields, such as fashion products [2], E-Commerce markets [3], business firms [4], retail company [5] and even individual products in supermarkets [6]. Sales prediction is essential for business success, enabling businesses to plan based on forecasts proactively. However, it is undeniable that predictions inherently carry some degree of bias or, in other words, a margin of error. Nonetheless, the choice of different models for sales forecasting can yield varying levels of accuracy, ranging from low to high. It is advisable to integrate additional tasks into the current model to enhance the accuracy of prediction models [7].

Next, this study will present some models employed by earlier researchers and discuss their corresponding outcomes. Firstly, in the realm of E-Commerce market sales prediction, a newly developed model known as the Seq2Seq and Transformer architecture using deep learning has emerged to tackle the sales forecasting challenge for Corporación Favorita. This model not only delivers the highest performance but also does so at the most cost-effective theoretical computational expense. These findings indicate that, for this specific use case, the Seq2Seq trimmed model stands out as the recommended choice, owing to its remarkable efficiency and effectiveness [8].

In the realm of E-commerce, it is worth noting that deep learning models are not the exclusive choice. Machine learning, another widely adopted and popular approach, exhibits extensive versatility, enabling its application not only in the E-commerce field but also in real-world supermarket commerce [3, 6]. In the field of E-commerce, three other commonly used models, including Incentive-Auto-Regressive-Integrated-Moving-Average (I-ARIMA), Long-Short-Term Memory (LSTM), and Artificial-Neural-Network (ANN), are introduced. While the I-ARIMA model falls under deep learning, LSTM and ANN are considered machine learning techniques. These three approaches can accommodate varying accuracy requirements and different data types. Examining diverse datasets from various E-Commerce domains indicates that LSTM excels over others, including contemporary machine learning options, in terms of predictive precision. Upon comparing

these models, researchers have observed that machine learning models perform more accurately than deep learning models. The lower accuracy is because conventional time-series analysis-based methods primarily rely on business records, overlooking the spatial relationships between adjacent retail locations [3]. In real-world supermarket business, four machine learning models are introduced: XGBoost, ARIMAX, LSTM, and Facebook Prophet. Generally, XGBoost and LSTM exhibited superior performance with lower errors. However, Facebook Prophet excelled in accuracy during holidays, making it ideal for holiday forecasts. LSTM adapted swiftly to holiday dynamics, enhancing performance.

Weather data did not significantly improve models and sometimes worsened results. Thus, results are inconclusive, highlighting the model choice's dependence on time and forecasting goals [6]. Undoubtedly, as shown by the earlier examples, it is clear that machine learning models are more suitable for forecasting sales. Some data could interfere with our model's performance when using a dataset for predictions. In order to optimize our predictions, it is crucial to eliminate these disruptive factors. After conducting several research, it was found that Lasso regression can serve this purpose for optimization. Therefore, this paper applies the least absolute shrinkage and selection operator regression model, also called as the LASSO regression model. The reason is that in predictive research, the Lasso technique enhances predictions by shrinking regression coefficients and simplifies models by setting some coefficients to zero [9]. By using LASSO, it would be a good fit to deal with datasets where number of unknown parameters is relatively larger than observations [10]. The paper is structured into five sections: Section 2 will introduce the data and methodology, section 3 will discuss the results and provide analysis along with graphs, section 4 will explore limitations and future outlooks, and section 5 will conclude the paper.

2. Data and Method

2.1. Data

This research paper utilizes open data generously shared by Surprise Housing to delve into the world of sales trends. Surprise Housing, a renowned American company known for its data-driven approach to buying and selling houses at profitable margins, is currently focusing on the Australian real estate market. In pursuit of this objective, Surprise Housing's data experts have compiled a dataset of Australian property sales. This dataset encompasses 81 variables, including ID, MSSubClass, MSZoning, LotFrontage, LotArea, Street, Alley, and more. To meet the research requirements and enhance the robustness of the analysis, all the variables provided by the organization have been utilized, even though the dataset contains twelve variables with missing values: LotFrontage, Alley, MasVnrArea, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, and MiscFeature. The rationale behind this choice is that incorporating more variables may introduce more significant variability in the accuracy of the data. Consequently, employing Lasso regression allows for a better distinction between interfering variables, facilitating their exclusion and helping achieve the desired research objectives.

2.2. Model

Lasso regression, as previously discussed, enhances predictive models by applying a regularization technique that shrinks regression coefficients and simplifies the model by setting specific coefficients to zero. To achieve this, this study has employed Lasso regression using scikit-learn's Lasso model. Lasso regression is a variant of linear regression that incorporates a regularization term into the linear regression equation. This regularization term, controlled by the alpha parameter (α), plays a crucial role in mitigating overfitting by penalizing the magnitudes of coefficients associated with each feature. When α is set to zero ($\alpha=0$), Lasso regression behaves identically to standard linear regression, which poses a risk of overfitting as all features contribute fully. However, as α increases, regularization becomes more pronounced, driving certain feature coefficients closer to zero. This step effectively performs feature selection by emphasizing the most informative variables while maintaining control

over the model's complexity. For our model, one has set the alpha parameter to 0.001. In addition to the alpha parameter, Lasso regression inherits various other parameters from the base 'linear_model.Lasso' class in scikit-learn. In our current configuration, these parameters retain their default settings, including 'fit_intercept' and 'normalize'. To assess the model's performance and determine if Lasso regression is an appropriate choice, one employs three evaluation methods: R-squared (R^2), Root Mean Squared Error (RMSE), and Cross-Validation (CV). These metrics collectively provide insights into the model's goodness of fit, predictive accuracy, and generalization capability. In basic terms, as R^2 approaches 1, it signals a stronger model fit (typically, R^2>0.5 is regarded as acceptable). As for RMSE, it quantifies the average prediction error, with lower RMSE values signifying a more precise fit. To implement cross-validation effectively, such as using k-fold cross-validation, assess the model's performance across diverse data subsets. This approach promotes both consistency and acts as a safeguard against overfitting.

3. Results and Discussion

The dataset contains a combination of categorical and numerical values. To begin with, this study extracted the numerical values from the dataset for a preliminary examination. Out of the 81 variables, 38 are numerical (comprising 3 float and 35 integer variables), while the remaining 43 are categorical, often referred to as "object" variables. It's important to note that even though some variables may be in numerical format, they are essentially categorical because these numbers do not represent a meaningful numeric relationship but rather categorize data into distinct groups. If these variables were treated as continuous numerical data for predictive modeling, they could potentially introduce noise and adversely affect the accuracy of the model.

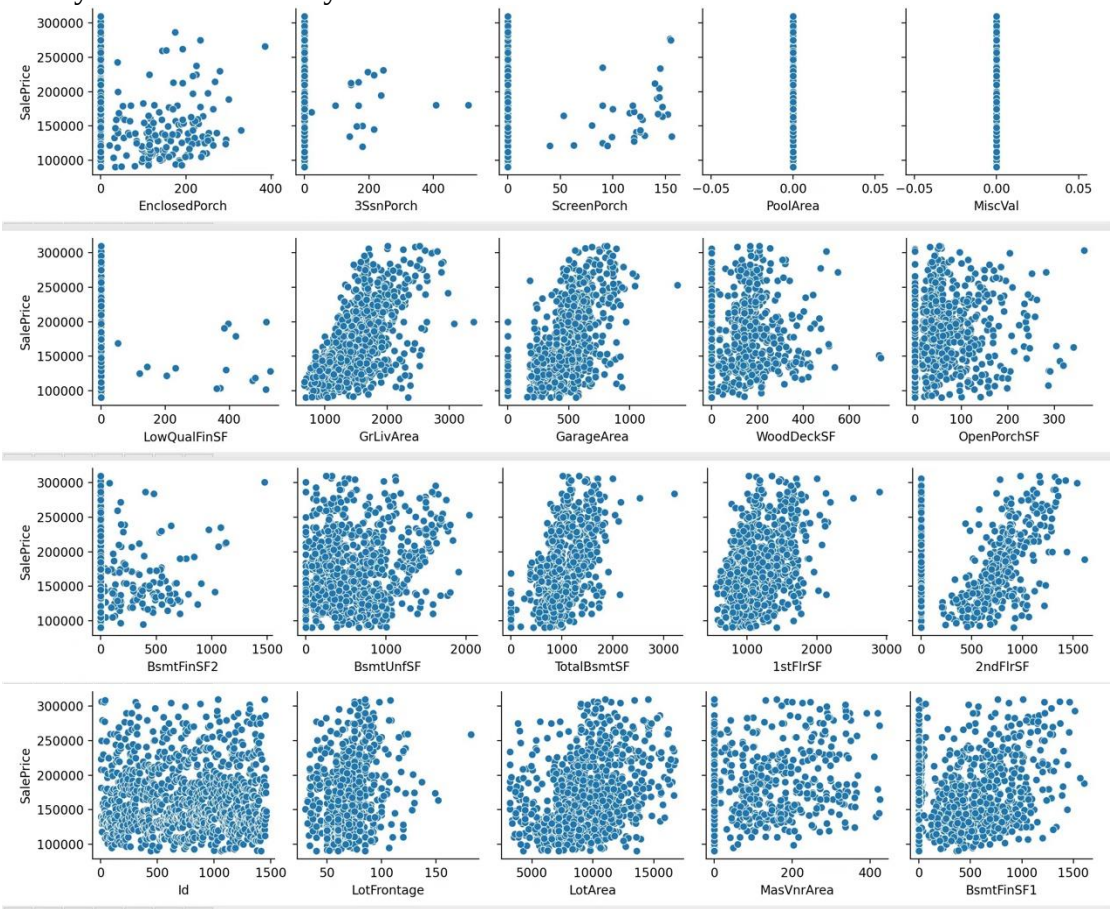

Fig. 1 Scatter of variables.

After removing the interfering variables, one can proceed to analyze the remaining variables that have a higher correlation with our target variable. First and foremost, it's essential to identify any

outliers within this subset of data. Outliers can have a significant impact on data accuracy and should be addressed as part of the analysis process. After identifying variables with outliers, our aim is to optimize and minimize the presence of these outliers. The approach employed involves calculating the first quartile (Q1) corresponding to the 5th percentile and the third quartile (Q3) corresponding to the 95th percentile. Subsequently, one computes the Interquartile Range (IQR) using the formula IQR = Q3 - Q1. The dataset is then filtered to retain only the rows where the respective data points fall within the range [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]. This step effectively eliminates outliers from the target feature, enhancing data integrity and analysis accuracy. Following the optimization of outliers, One can proceed to examine the correlations between variables. The visualizations presented below provide a more intuitive way to understand the relationships between these variables and the target variable. These Fig. 1 representations offer valuable insights into the associations between the variables and the target variable.

A heatmap serves as an effective tool for visually illustrating the strength of correlations between pairs of variables as well. When the correlation coefficient approaches 1, it signifies a robust and positive relationship between the variables. Conversely, when the correlation coefficient nears 0, it signifies a lack of discernible influence or a weak association between the variables. Seen from Fig. 2, one can easily identify the features that exhibit lower correlations with the target variable (item_outlet_sales).

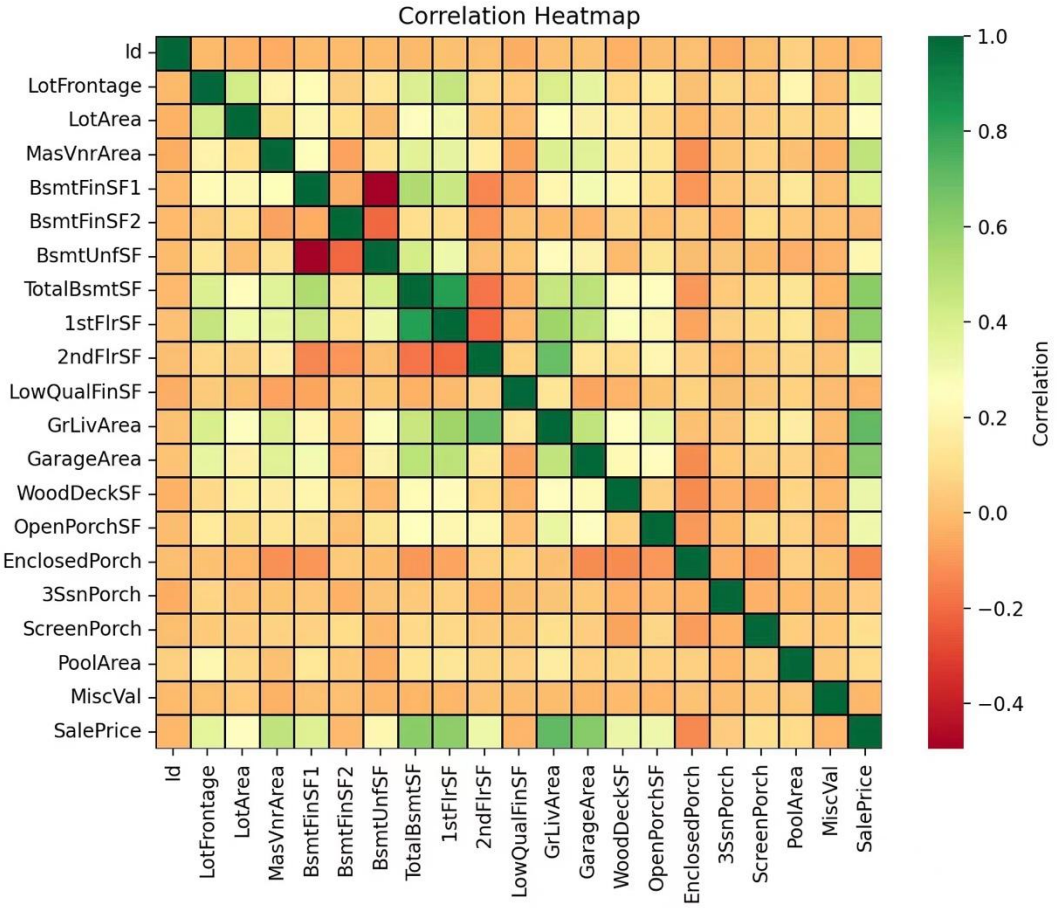

Fig. 2 Correlation heatmap.

Now, this study addresses the missing values denoted as "NA" in the dataset. Our approach will vary based on the type of variable: For categorical variables, this study will add a "No" prefix to indicate the absence of the categorical feature. For numerical variables, this study will replace missing values with "0". However, there's a special scenario where a column predominantly contains "NA" values. In such cases, this study will opt to remove that column entirely to prevent significant bias in our final predictions. Moving forward, this study is addressing the numerical values that were initially

separated from the categorical values. When considering the year of establishment, one calculates the number of years the outlet has been in operation by subtracting the provided year from the current year (2023). This transformation provides a more continuous representation of the data, enhancing our comprehension of it. Additionally, one converts all other variables into an 'object group' by applying the '.astype('object')' method to the respective strings. Regarding variables that have only two distinct categorical responses, one can efficiently map these variables by assigning '0' or '1' to the respective values. This mapping process allows us to represent these values as numerical values. In the case of the remaining categorical variables, one employs a technique known as the creation of dummy variables to convert them into numerical values. Dummy variables, also referred to as indicators or binary variables, play a crucial role in statistical modeling and data analysis, particularly in regression analysis and machine learning. These variables are utilized to represent categorical data, facilitating the transformation of such variables into a numeric format that is compatible with various algorithms and models.

After converting all values to numerical values, one now could apply the lasso regression model to the new dataset. This study first splits the original dataset into train and test dataset. The training dataset teaches and refines the machine learning model, while the test dataset checks how well the model predicts new, unseen data. This split helps avoid the model from getting too specialized, where it's great with the training data but falters with new information, ensuring it can provide useful predictions in real-life situations. This study randomly assigned 80% of the data by the train dataset and the rest 20% be the test dataset. Then applying GridSearchCV to modify model with optimal value of alpha. After applying the lasso progression, this study plots and gets the comparison graph for the train and test data. Fig. 3 illustrates a good fit for both the train and test data. The solid black line represents the train data, while the blue dashed line represents the test data.

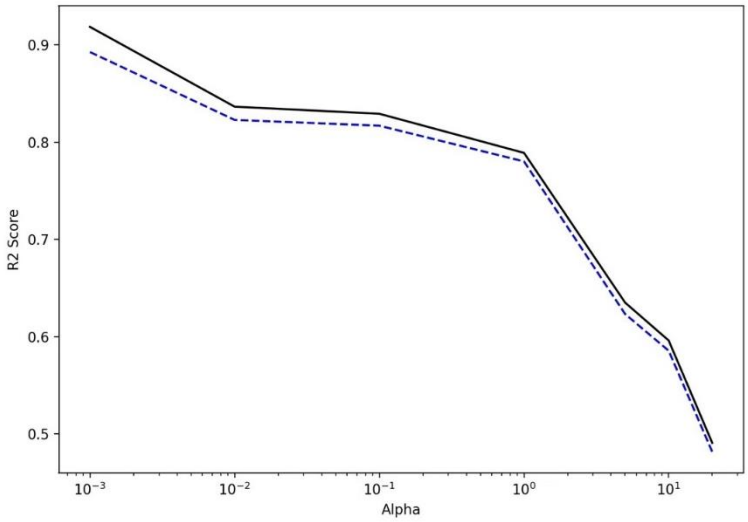

Fig. 3 R-square as a function of alpha.

As previously mentioned, this study has employed three evaluation methods: R-squared, Root Mean Squared Error (RMSE), and Cross-Validation (CV) to test the performance of LASSO to see whether it is a good fit for the model. Upon analyzing the results, a train and test R-squared value around 0.9 evidently show that the LASSO model is a great fit and exhibits a high level of predictive accuracy. The similarity between the R-squared values for both the training and test datasets further confirms its appropriateness as a model. The root mean squared error (RMSE) is a commonly used metric for assessing the performance of regression models, measuring the root mean square of the errors between model predictions and actual observations. The smaller the value of RMSE, the better, as it measures the magnitude of prediction errors made by the model. In general, when RMSE is closer to 0, it indicates that the model's predictions are more accurate. Smaller RMSE values are indicative of superior performance, signifying minimal prediction errors. In the case, this study has achieved a remarkably low RMSE value of 0.0085, underscoring the model's precision. In addition

to R-squared and RMSE, this study has employed another essential evaluation technique to test the performance of the model called Cross-Validation (CV). Cross-Validation is a method used to evaluate the performance of the predictive model. The method partitions the dataset into different sections, typically a training set and a validation set, and then repeatedly training and evaluating the model. The goal is to attain a more precise assessment of the model's performance, particularly its ability to make accurate predictions on new, unseen data. Our results reveal consistently robust R-squared values across all folds, ranging from 0.8649 to 0.9286, with a mean R-squared value of 0.9001. These findings affirm the model's high accuracy and suitability for the task at hand.

4. Limitations and Prospects

LASSO serves a fundamental purpose by introducing L1 regularization into linear regression. This regularization technique plays a crucial role in driving irrelevant coefficients towards zero, thereby facilitating feature selection. However, it's crucial to emphasize the pivotal role of parameters in linear regression. When these parameters become excessively large or when we're dealing with an excessive number of features, the result can easily tip into the realm of overfitting, a scenario one certainly aim to avoid. One of the primary strengths of LASSO regression is its ability to address the limitations of both least squares estimation and stepwise regression. By doing so, it not only enhances feature selection but also effectively resolves the intricate problem of multicollinearity among features. However, when dealing with a group of highly correlated features, the Lasso regression method tends to favor the selection of a single feature from the correlated group, introducing inherent biases in feature selection and potentially causing instability in the outcomes. Furthermore, as previously mentioned that Lasso operates as an optimization technique built upon the foundation of linear regression, its performance may be less optimal when dealing with nonlinear relationships between predictor variables and the target variable.

With the continuous development of machine learning and data science, the ability of Lasso regression in feature selection and handling multicollinearity among variables has made it an important tool in predictive modeling. While the Lasso model's performance may occasionally fall short of expectations in specific scenarios, there's a strong belief that ongoing optimization endeavors will pave the way for significant enhancements. Moreover, the future holds intriguing prospects for enhancing predictive accuracy by seamlessly integrating Lasso with other machine learning techniques, such as Long-Short-Term Memory (LSTM) and random forests. Hybrid models can leverage the strengths of different algorithms to handle complex relationships and feature interactions, ultimately leading to more robust and resilient predictions. Furthermore, an intriguing avenue for future development lies in expanding LASSO's capabilities beyond its primarily linear orientation. If it can evolve to effectively capture nonlinear relationships, it has the potential to outperform current expectations and become an even more indispensable tool in the data scientist's toolkit. Moreover, consider the potential applications in fields like healthcare, finance, and climate science, where the inherent complexity of data often defies linear modeling. In these domains, an evolved Lasso with nonlinear capabilities could revolutionize decision-making and predictive accuracy, bringing about substantial advancements. As the synergy of data science and machine learning continues to unfold, Lasso's role is poised to expand, ensuring it remains a driving force in predictive modeling and data-driven decision-making. Keeping pushing the boundaries of innovation, Lasso's future is looking brighter than ever, bringing exciting possibilities for data scientists and analysts around the world.

5. Conclusion

To sum up, LASSO steps up as a solid choice for predicting sales. Our analysis revealed that it adeptly sifted through the multitude of variables, 81 in total, to pinpoint the most influential features. This ability to identify key predictors empowers us to allocate resources and efforts more efficiently. In real estate data, it's normal to find features that are highly correlated such number of bedroom and

kitchen above ground, number of fireplaces in our dataset. Lasso handles this tricky stuff with finesse, making sure researchers don't rely too much on linked things and making the model tougher and easier to understand. Predicting property prices is crucial for both the real estate industry and homebuyers. Lasso-based models provide data-driven guidance for property deals, helping buyers and sellers make smarter pricing and negotiation decisions. It's worth noting that while Lasso excels in many scenarios, it may not be the optimal choice when dealing with datasets characterized by nonlinear relationships among variables. However, this serves as an opportunity to compare different modeling approaches to select the one that best fits the dataset's unique characteristics. In summary, leveraging Lasso for housing price prediction not only enhances forecasting accuracy but also provides valuable insights into the pivotal factors influencing property prices. This approach signifies the application of data science and machine learning to the realm of real estate, making them more data-driven and intelligent.

References

[1] Rothe J T. Effectiveness of sales forecasting methods. Industrial Marketing Management, 1978, 7(2): 114-118.

[2] Chen D, Liang W, Zhou K, et al. Sales Forecasting for Fashion Products Considering Lost Sales. Applied Sciences, 2022, 12(14): 7081.

[3] Ajaykrishna S, Suganya T S, Rao B, et al. Online Sales Prediction in E-Commerce Market Using Machine Learning. 2023 4th International Conference on Signal Processing and Communication (ICSPC). IEEE, 2023: 47-51.

[4] Wang C H, Gu Y W. Sales Forecasting, Market Analysis, and Performance Assessment for US Retail Firms: A Business Analytics Perspective. Applied Sciences, 2022, 12(17): 8480.

[5] Saha P, Gudheniya N, Mitra R, et al. Demand forecasting of a multinational retail company using deep learning frameworks. IFAC-PapersOnLine, 2022, 55(10): 395-399.

[6] Fredén D, Larsson H. Forecasting daily supermarkets sales with machine learning. CCSI, 2020.

[7] Dalrymple D J. Sales forecasting methods and accuracy. Business Horizons, 1975, 18(6): 69-73.

[8] Vallés-Pérez I, Soria-Olivas E, Martínez-Sober M, et al. Approaching sales forecasting using recurrent neural networks and transformers. Expert Systems with Applications, 2022, 201: 116993.

[9] Musoro J Z, Zwinderman A H, Puhan M A, et al. Validation of prediction models based on lasso regression with multiply imputed data. BMC medical research methodology, 2014, 14(1): 1-13.

[10] Gauraha N. Introduction to the lasso: A convex optimization approach for high-dimensional problems. Resonance, 2018, 23(4): 439-464.