Financial Engineering and Risk Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2024.070224
ISSN 2523-2576 Vol. 7 Num. 2

Feature Engineering-Based Random Forest Model for Predicting Chinese Gold Futures Prices

Mingming Qu[1], Yurong Shi[2]

[1]Yantai Aizhi Intelligent Technology Co., Yantai, Shandong, China
[2]Dalian Maritime University, Dalian, Liaoning, China

Keywords: Feature Engineering, Random Forest, Futures, Gold, Forecasting

Abstract: To enhance the prediction accuracy of gold futures price trends and address the challenge of low prediction accuracy amidst numerous features and noise, a novel research method combining random forest with feature engineering is introduced. Manual feature engineering methods, including Pearson coefficients, Mean Decrease in Impurity, and Mean Decrease Accuracy, are employed for feature selection. Subsequently, automatic feature engineering techniques are utilized to generate new features, which are then integrated with the Pearson coefficients. Finally, the selected features are used for modeling and regression prediction through Random Forest, from which the final conclusions are drawn. Experimental results indicate that the Random Forest based on automatic feature engineering surpasses the original Random Forest and other Random Forest models in predictive evaluation metrics.

1. Introduction

In the present fast-cash market, the forecasting of commodity future prices gets the highest value attention from the investors, banks and the economists. Gold, which symbolizes the precious metals, is periodically affected by a variety of elements such as inflation and economic crisis. Examples of this are economic policies like monetary maneuvers and inflation rates, international politics, as well as market expectations and speculative actions. The power to precisely and effectively predict the future prices of gold makes a huge difference in the strategy of investment decisions and guarding against various financial risks.

The appearance of machine learning algorithms together with the development of a large variety of financial forecasting domains has led to the emergence of very successful algorithms whose strength lies in the fact that they are capable to learn from the data. For example, the discover of random level, a bouncing reinforcing methodology, has become a famous tool for its high performing and robustness. Mist absorbance is enhanced by the information integration having the broad data range, without which the overseeing risk could escalate and make the navigation nonlinear dynamics and large number of variables difficult.

However, the performance level of the proposed model is not a direct result of the quality of the parameters. Researchers should consider the potential for messy and details-heavy data to mislead raw data at the point of use, thus compromisimg the forecasting power of data that has not been cleaned up. In this sense, the function of feature engineering becomes important, herein the selection of key characteristics of the gold prices is refined from the raw data and the rest of unneeded and redundant information is discarded. This procedure elevates the quality of model input as a sign of improvement to the accuracy and trustworthiness of the model prediction.

The purpose of this study is two-fold: firstly, to build and back-test a random forest model for gold futures prices forecasting using feature engineering as methodology; and secondly, to gain valuable insights into the gold market dynamics and inform out future investment decision making. This study tries to innovate by involving domain expertise along with existing data, incorporating feature generation by feature selection and thus considering the competitive feature identification for prioritization of inputs. Thus, the random forest model shall be employed for promising gold futures prices' prediction, and the efficiency and efficacy of the method will be assessed by comparing it with other forecasting tools.

Conversely, the purpose of this investigation not only aims at increasing the precision and continuity of gold futures price forecasts based on machines but also establishes a supportive tool for investors to make more trusted choices. In addition, in the light of this study is created potential to address the requisites for incorporating machine learning algorithms (MLAs) in financial forecasting.

2. Feature-driven Random Forest (RRF) Model for the Gold Futures Highlight

Hyperparameters influence the success or not on the machine random forest, made most of the feature variables. This study combines the task of hyperparameter tuning with data feature engineering to therefore boost the forecasting ability of random forests as regards the issue of future gold. The methodology involves data preparation, feature engineering, model training, and evaluation sequentially utilizing manual feature engineering techniques to discard features directly with the Pearson, MDI, and MDA methods [1], [2]. Subsequently, automated feature engineering generates new multi-features, which are integrated with Pearson coefficients; the optimized random forest[3], [4] is then utilized to derive prediction results. This approach facilitates identifying an improved configuration of the random forest model, enhancing its generalization capability on novel data.

2.1. Manual Feature Engineering

(1) Pearson Correlation Analysis

Taking Au2312.SHFE as a case study, the Pearson coefficient was applied to categorize all explanatory variables, distinguishing between variables of high, moderate, weak, and negligible correlation. A Pearson coefficient greater than 0.8 indicates high associations; 0.5 - 0.8 indicate moderate associations; 0.3 - 0.5 indicates weak associations, and less than 0.3 indicate negligible correlations. These outcomes of this analysis are depicted in Figure 1.

As per the Pearson correlation coefficient, with the exception for the parameters directly associated with the commodity itself, close, low, high, open, and AU8888, and those related to AU9999, including close, low, high, open, and EMA, other variables exhibit negligible correlation. This indicates a lack of a direct or necessary linkage with the closing prices of the explanatory variables. To prevent the augmentation of computational time for the Random Forest and to mitigate the impact of irrelevant variables on experimental accuracy, variables not correlated with the closing price are excluded. The retained variables serve as the definitive explanatory variables for the model.

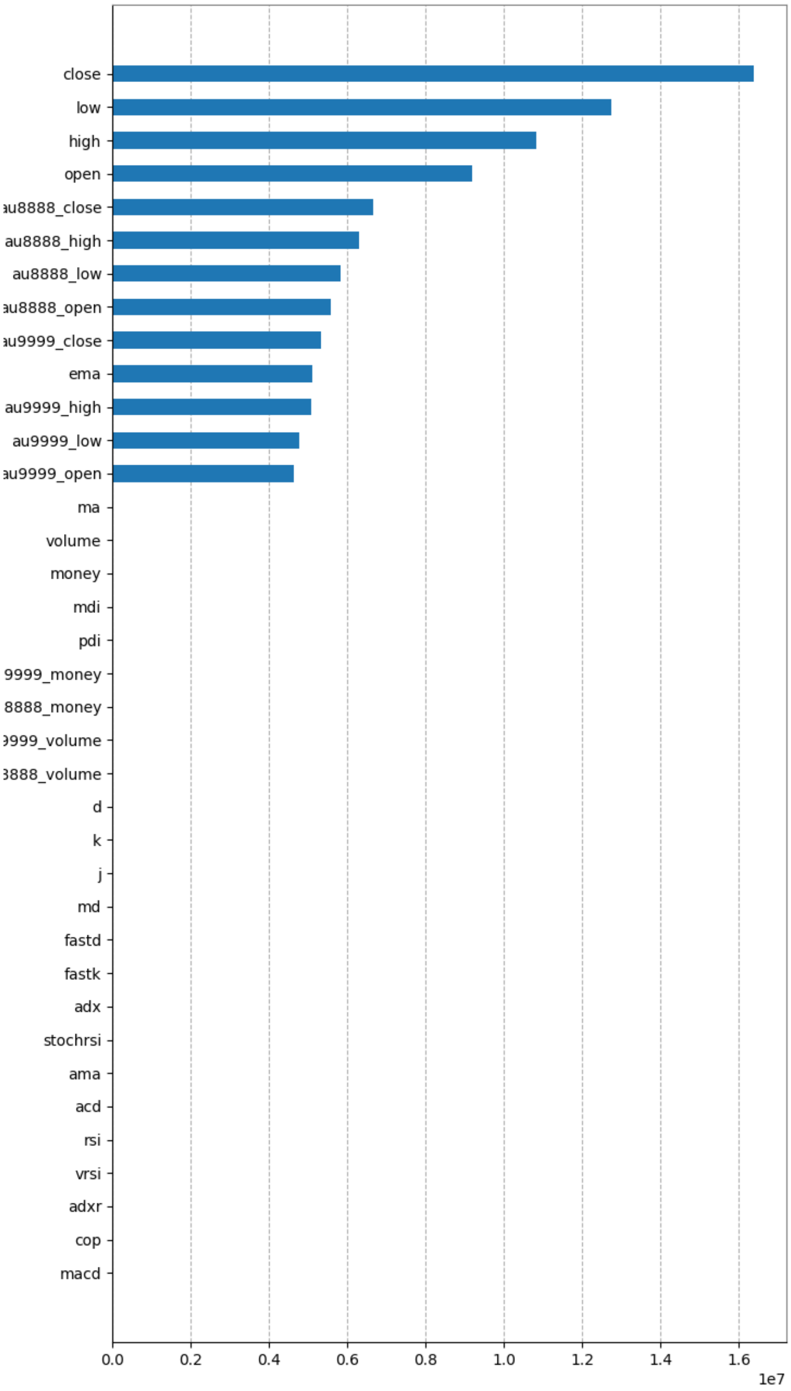


Figure 1: Pearson Coefficient.

(2) MDI

Utilizing AU2312.SHFE as a prototype, all explanatory variables were evaluated using the MDI. The findings are depicted in Figure 2:

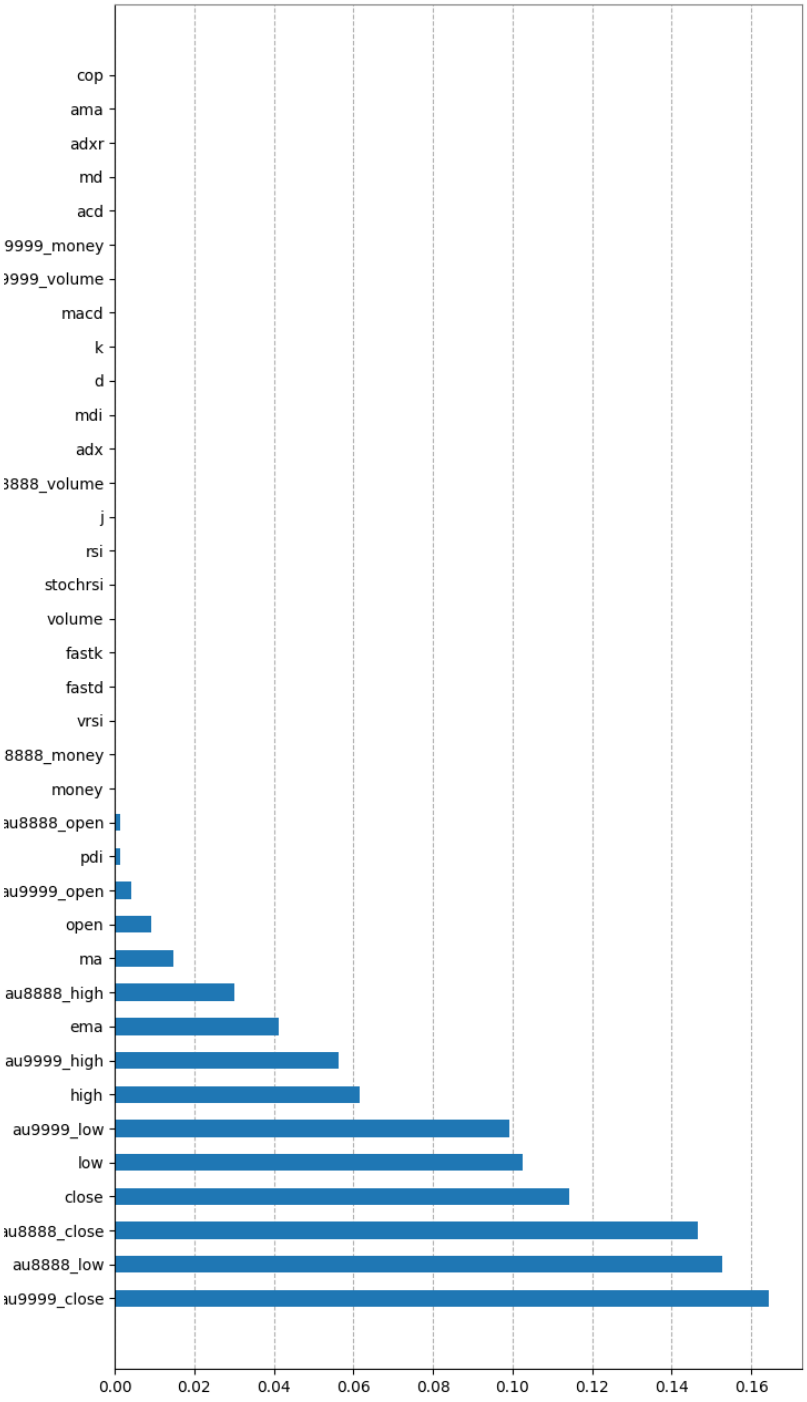


Figure 2: MDI Coefficient.

In consideration of the MDI size, it was found that the feature importance of "close," "low," "high," and "open," as well as "close," "low," "high," and "open," in addition to "ema," "ma," "pdi" for both au8888 and au9999, is significantly higher, rendering the importance of other features negligible. Consequently, less critical features were omitted to focus on those of higher importance.

(3) MDA

Au2312.SHFE served as a paradigm for evaluating all explanatory variables through MDA. The findings, depicted in Figure 3, illustrate:

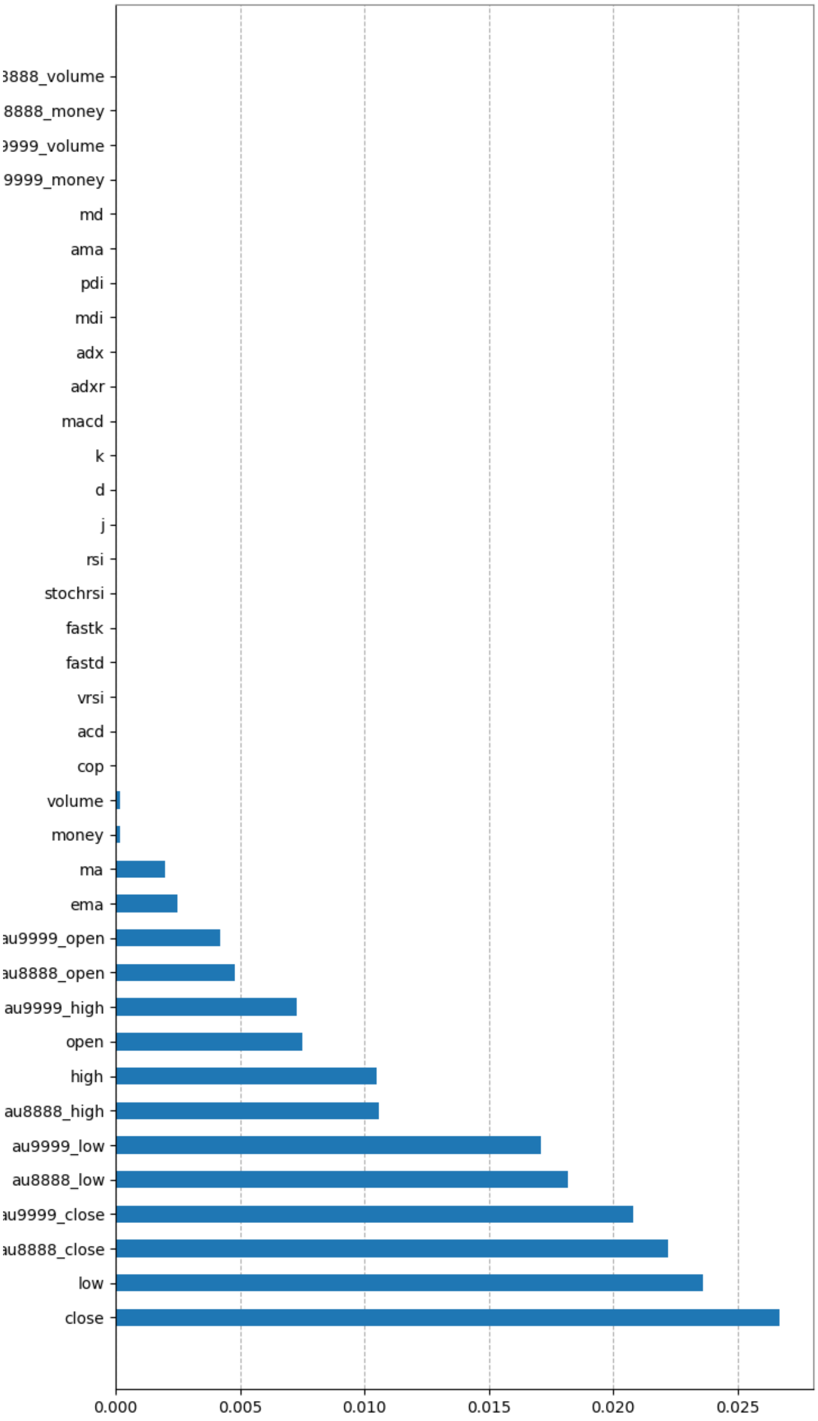


Figure 3: MDA Coefficient.

Based on the MDA evaluation, features such as "close," "low," "high," "open," alongside "ema," "ma," "volume," "money" for au8888 and au9999, emerged as substantially more crucial compared to others, leading to the exclusion of features with minimal importance in favor of preserving those deemed more significant.

2.2. Automated Feature Engineering

Contrary to manual efforts, automated feature engineering innovates by autonomously generating new candidate features from available data, subsequently selecting the optimal features for model training. This advancement is pivotal, transcending the constraints encountered in conventional feature engineering methods. Whereas traditional approaches necessitate domain experts' in-depth data analysis and manual feature design, often a laborious and expertise-dependent process, automated feature engineering streamlines the identification and incorporation of superior features, thereby enhancing both the efficiency and quality of feature engineering. Automated feature engineering is characterized as a technique for constructing processes specific to given scenarios, which operates independently of human intervention.

AutoFeat, a Python library, facilitates automated feature engineering and feature selection, along with offering models like AutoFeatRegressor and AutoFeatClassifier, which demand extensive scientific computation and robust computational resources. In this context, AutoFeatRegressor was utilized for the generation of new features.

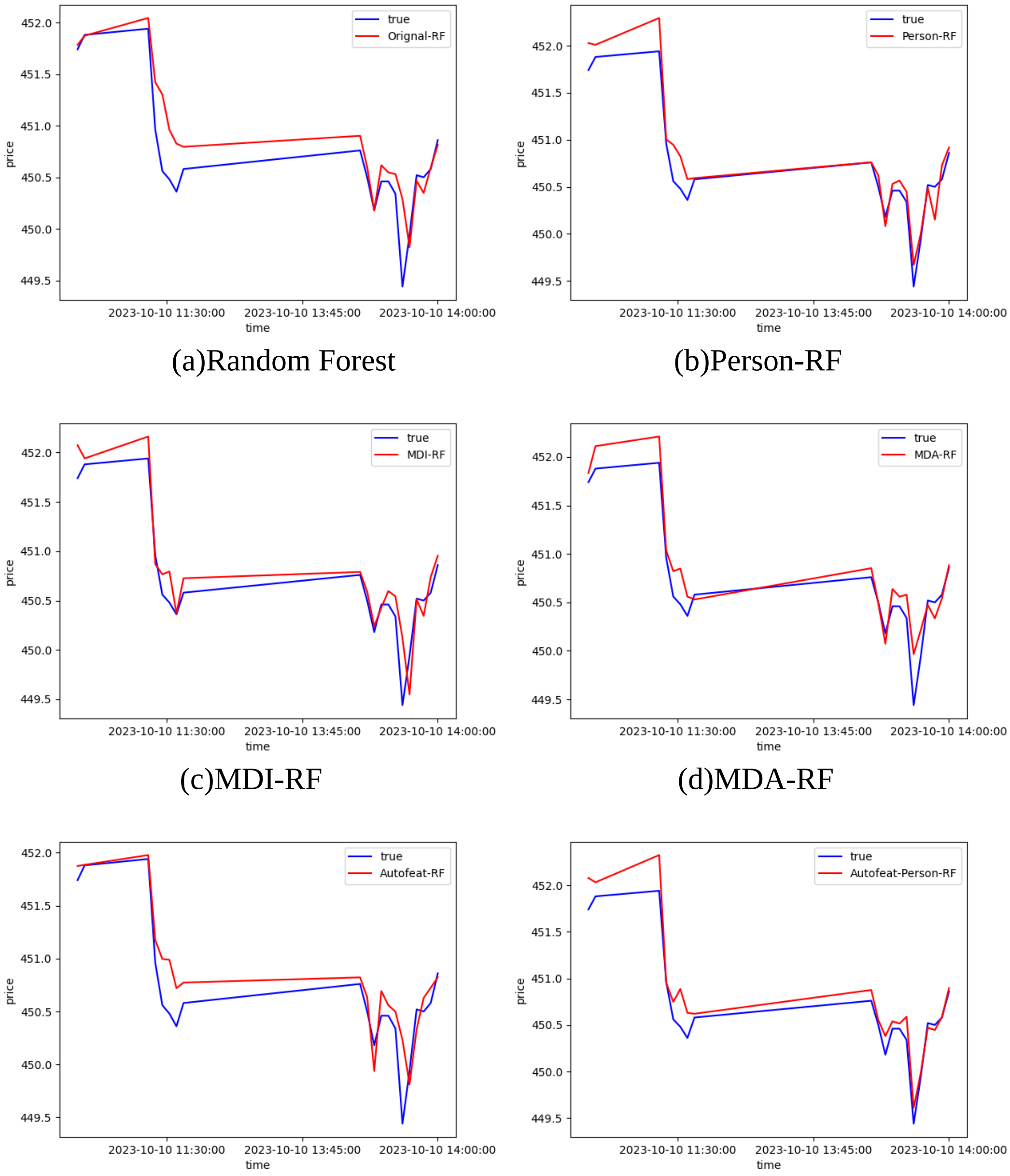
Observations from our experiments indicated that Pearson coefficients yield favorable outcomes in feature selection, and AutoFeat's performance was similarly effective. This led to the integration of both methodologies for feature generation and selection, a strategy that notably enhanced the experimental results.

3. Experiments and Results

To substantiate the predictive accuracy of the random forest model for gold futures prices as discussed in this study, predictions were made for Au2312.SHFE. Data spanning from '2022-04-25' to '2023-10-11' at 15-minute intervals was selected for analysis, segregating the final 20 samples as the test set and allocating the remainder for training purposes.

The experimental outcomes for Au2312.SHFE are illustrated within Fig. 4, where the blue solids indicate real numbers, while its predicted scores from the corresponding model are represented by a red solid line. The predictive outcomes displayed in Figure 4 reveal that models based on Pearson's coefficients outperform alternative models, with the random forest model, enhanced by AutoFeat + Pearson, demonstrating particularly superior predictive capabilities.

The results depicted in Figure 4 demonstrate that the random forest method inherently possesses a commendable fitting capability, effectively capturing the fluctuations of gold futures prices to render superior predictions. The enhancement of the random forest model through feature selection methodologies notably diminishes the error by excluding the impact of irrelevant variables on gold prices. Improved performance of random forest model with different feature selection techniques has been observed. The performance surpasses that of the unmodified model which lacks feature selection. The performance is directly linked to the high-quality feature-selection mode which is being used to eliminate non-significant features, which in turn means that the features only contain useful ones with no data noise or redundancy. This approach enables a model to pinpoint main features rather than considering each of them. This in turn generalizes the model even further but also limits predictive mistakes. The evolution of the feature selection procedures through time allows the model to better describe the gold price fluctuations. By methodically selecting the set of optimum features, the model will emphasize those elements that unconditionally contribute to the price volatility. The resulting accuracy will be significantly enhanced and the margin of prediction error will be substantially reduced. The further evaluation of varied feature selection models' effectiveness and accuracy would be done with the metrics, such as explained_variance_score, MSE, RMSE, MAPE, and R^2. I have used them for comparative analysis. The definitions and formulas of these metrics are delineated as follows:



(a)Random Forest (b)Person-RF

(c)MDI-RF (d)MDA-RF

(e)Autofeat-RF (f)Autofeat-person-RF

Figure 4: Comparison of Model Predictions for Au2312.SHFE.

The explained_variance_score, namely the explained variance score, measures the extent to which the model accounts for the fluctuations observed in the dataset. A value of 1 indicates perfect model performance; lower scores signify diminishing effectiveness.

$$explainedvariance\left(y,\hat{y}\right)=1-\frac{Var\left\{y-\hat{y}\right\}}{Var\left\{y\right\}} \tag{1}$$

Where, the true value is denoted as y, the predicted value as $\hat{y}$, and variance as var.

The MAE reflects the error average magnitude in a collection of these types of prediction, without considering their direction. It represents the mean absolute difference between the model and the predicted score. The MAE being the lowest represents the higher precision of the models' predictions;

the MAE is given by n that is the total amount of specimens available, y_i is the true score, while $\hat{y}_i$ is the predicted score.

$$MAE = \frac{1}{n}\sum_{i=0}^{n-1}\left|y_i - \hat{y}_i\right| \tag{2}$$

The mean square error (MSE) refers to the statistical mean of the squared difference between the estimated scores and the actual scores. It results from the variance and bias of estimator and their value relation to the true scores. By decreasing in the measured MSE, a model of superior quality with more accurate predictions is obtained. The sample size is represented by n, whereas the true value and the predicted value are represented by y_i and $\hat{y}_i$, respectively.

$$MSE = \frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \hat{y}_i\right)^2 \tag{3}$$

The root of square error (RMSE) measures the root of the square of the difference between the predicted score and the actual score.

Hence, interpretation of the measure suggests, the quantity of prediction errors' standard deviation which indicates how much are they distributed around zero. A lower value for the model's RMSE both means that it is more accurate and that the unit is the same as that of the original dataset.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}\left(y_i - \hat{y}_i\right)^2} \tag{4}$$

The basic descriptor of the mean absolute percentage error is between actual and predicted scores. A MAPE value below 10% is generally indicative of high prediction accuracy.

$$MAPE = \frac{1}{n}\sum_{i=0}^{n-1}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{5}$$

The coefficient of determination, R^2 (R-Square), denotes a statistic representing the combined difference between the absolute score and the predicted score, similar to the MSE; The denominator refers to the squared difference between the true scores and their mean, which corresponds to the variance (Var). The R-Squared value, ranging between [0,1], serves as an indicator of model fit: a value of 0 signifies poor fit, whereas 1 indicates an error-free model. Generally, a higher R^2 value denotes better model fitting.

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=0}^{n-1}\left(y_i - \bar{y}_i\right)^2} \tag{6}$$

Examination for predictors in Table 1 reveals that models with explained_variance_score exceeding 0.9, such as the Pearson feature selection model, the MDA feature selection model, and the AutoFeat-Pearson feature selection model, demonstrate superior explanatory power for price fluctuations. Notably, the AutoFeat-Pearson feature selection model, with a score of 0.948214253373187, exhibits the highest degree of explanatory ability. Further analysis shows the AutoFeat-Pearson feature selection model yielding the lowest MAE (0.1444000000000017), indicating minimal deviation between actual and forecasted scores. Similarly, this model presents the lowest MSE (0.036095399999998314) and RMSE (0.18998789435118837), suggesting the closest approximation and lowest average inverse of predicted and actual scores, respectively. Its minimum MAPE score of 0.0003203169201916095, again attributed to the AutoFeat-Pearson feature selection model, denotes minimized error between actual and predicted scores. Lastly, the R^2 value closest to 1, standing at 0.8978555419722952, belongs to the AutoFeat-Pearson feature selection model, indicating the best model fit. Comparative analysis indicates that while other feature selection models underperform relative to the AutoFeat-Pearson feature selection model, they nevertheless outperform the baseline random forest model lacking feature selection. This suggests the utility and effectiveness of feature selection for enhancing the random forest model's performance, with AutoFeat-Pearson emerging as the optimal feature selection methodology.

Table 1: Predictive Indicators for Au2312.SHFE by Various Models

Methods	**Explained_variance_score**	**MAE**	**MSE**	**RMSE**	**MAPE**	**R^2**
Random Forest	0.793254097618405	0.22150000000000034	0.10669459999999778	0.32664139358017347	0.0004918097359323012	0.6980706103413999
Person-RF	0.9211720377161989	0.15800000000000977	0.04002200000000078	0.20005499243958091	0.0003505375142318674	0.8867438648917848
MDI-RF	0.8771058942882369	0.17084999999999298	0.05406774999999823	0.23252472986759543	0.00037926300288268017	0.846996541927018
MDA-RF	0.9187872195621742	0.16764999999999758	0.04458724999999665	0.21115693216183232	0.0003721210100887179	0.8738249060490908
Autofeat-RF	0.8432975357692711	0.21189999999999998	0.07844849999999924	0.2800865937527165	0.00047053418447880115	0.7780027506112498
Autofeat-person-RF	0.948214253373187	0.1444000000000017	0.036095399999998314	0.18998789435118837	0.0003203169201916095	0.8978555419722952

For a more direct comparison of the fit among different models, this study presents Figure 5, which facilitates an intuitive assessment of each model's fitting capability. It becomes evident that the predicted scores from the original random forest model diverge significantly from the true scores. Conversely, models that underwent feature selection exhibit predictions much closer to the actual scores, with the AutoFeat-Pearson feature selection model achieving the nearest approximation to the true scores.

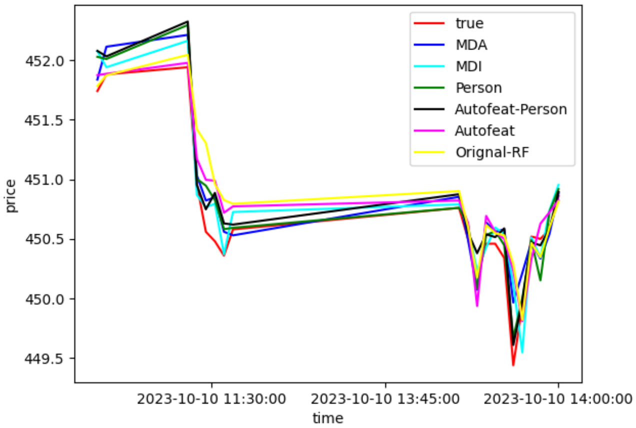


Figure 5: Model Prediction Comparison for Au2312.SHFE.

In summary, Figures 4 and 5, along with Table 1, indicate that among manual feature engineering approaches, the Pearson feature selection model is superior. Moreover, the AutoFeat-Pearson feature selection model, integrating AutoFeat feature engineering, stands out as the most effective across all considered feature selection methodologies. This finding corroborates the initial hypothesis that a combination of manual feature engineering for Pearson and automated feature engineering through AutoFeat could enhance feature selection precision within the Random Forest model, thereby optimizing predictive performance.

4. Conclusion

This study integrates manual and automated feature engineering techniques with the random forest model to investigate the trend of gold futures prices. Feature selection was initially conducted using manual engineering methods, including Pearson coefficients, MDI, and MDA. Subsequent application of automated feature engineering produced new features, which, when combined with Pearson coefficients, significantly enhanced the random forest model's performance, as evidenced by the experimental data.

Given the dynamic nature of parameter indices and the expanding pool of features it possesses, further studies can focus on selecting a varied features so as to refine model design and comparative analysis. Additionally, the increasing complexity of model evaluation and comparison, attributable to the growing number of features, suggests a need for developing more comprehensive and precise metrics and methods. These advancements will facilitate a deeper understanding of the performance variances across different feature sets and model configurations.

References

[1] Zhengxu Yan, Chao Qin, Gang Song. Random forest model stock price prediction based on Pearson feature selection [J]. Computer Engineering and Applications, 2021, 57(15):286-296.

[2] Yu Ai. Research on the prediction of CSI 300 index trend based on random forest optimization [D]. Shandong University, 2020.

[3] Leo Breiman. Randomforests [J].Machine Learning, 2001, 45(1).

[4] Nana Lin, Jiangtao QIN. Research on predicting the rise and fall of A-share stocks based on random forests[J]. Journal of Shanghai University of Technology, 2018, 40(3):267-273.