Jurnal Tekno Insentif | Vol. 19 | No. 1 | Mei 2025
DOI : https://doi.org/10.36787/jti.v19i1.1858 Halaman 30 - 44

Prediction of House Price using the Multivariate Adaptive Regression Spline Method

AYSHA ALIA ISKANDAR, DWI FITRIA AL HUSAENI, M. ZAENAL ISKANDAR SAHIDIN, LALA SEPTEM RIZA*, WAHYUDIN

Magister Pendidikan Ilmu Komputer, Universitas Pendidikan Indonesia, Indonesia
Email: lala.s.riza@upi.edu

ABSTRAK

Rumah merupakan salah satu kebutuhan pokok manusia. Saat ini, harga rumah terus meningkat, sehingga prediksi harga rumah menjadi penting bagi investor, penjual, dan pembeli dalam pengambilan keputusan. Penelitian ini menggunakan metode Multivariate Adaptive Regression Spline (MARS) untuk memprediksi harga rumah, dengan metode Generalized Linear Model via Elastic Net (GLMNET) sebagai pembanding. Studi dilakukan dalam lima tahap: persiapan data, pemilihan fungsi dasar, pembuatan model, evaluasi model, dan prediksi. Hasil penelitian menunjukkan bahwa model MARS menghasilkan nilai R-squared sebesar 0.5154666 dan Root Mean Squared Error (RMSE) sebesar 0.6263589. Variabel paling berpengaruh terhadap harga rumah adalah luas bangunan (sqft_living). MARS terbukti memberikan hasil prediksi yang lebih baik dibandingkan GLMNET. Penelitian ini diharapkan dapat membantu pengambil keputusan di bidang real estat, khususnya dalam menentukan harga properti secara lebih akurat.

***Kata kunci**: Machine Learning, Multivariate Adaptive Regression Spline (MARS), Generalized Linear Model via Elastic Net (GLMNET), Prediksi, Prediksi Harga Rumah, Regresi Linear.*

ABSTRACT

Housing is one of the basic human needs. Currently, house prices continue to rise, making price prediction important for investors, sellers, and buyers in decision-making. This study uses the Multivariate Adaptive Regression Spline (MARS) method to predict house prices, with the Generalized Linear Model via Elastic Net (GLMNET) as a comparison. The study was carried out in five stages: data preparation, basis function selection, model building, model evaluation, and prediction. The results show that the MARS model achieved an R-squared value of 0.5154666 and a Root Mean Squared Error (RMSE) of 0.6263589. The most influential variable on house prices was the building area (sqft_living). MARS produced better prediction results compared to GLMNET. This study is expected to assist decision-makers in the real estate sector, particularly in making more accurate property price predictions.

***Keywords**: Machine Learning, Multivariate Adaptive Regression Spline (MARS), Generalized Linear Model via Elastic Net (GLMNET), Prediction, Prediction of House Price, Linear Regression.*

1. PENDAHULUAN

One of the basic human needs is a house and housing. According to Sanusi et al. (2020), a house is a basic human need to improve dignity, quality of life, and livelihood. A house or shelter is one of the physiological or physical needs of humans **(Utamaningsih et al., 2019).** In determining whether to own a house, accurate price predictions are needed in determining purchase, sale, or investment. According to Singh et al. (2020) said that predicting house prices is very important in real estate market research and investment decision-making. When buying a property, one will consider the price and specifications. According to Kaushal et al. (2021), If someone wants to buy a house, they will look for an affordable house with all the features they want from a house.

In current conditions, house prices continue to increase. Many factors support an increase in house prices. According to Zhang (2021), academics usually believe that predicting the specific price for a particular estate is impractical because many factors influence the final cost. In traditional property selection, surveys can be carried out with several comparisons, but this will not provide precise accuracy. Trawinski et al. (2017) stated that there is the possibility of incorrect or subjective real estate appraisals which could harm investors or buyers.

Therefore, property price predictions are becoming increasingly important in today's digital era, especially in the real estate and financial sectors. Accurate home price predictions can help investors, sellers, and buyers make better choices. Wu (2020) said that real estate price predictions could provide a reference for investment and consumption decisions and a reference for government administrative decision-making-related departments.

The development of technology can be used to make house price predictions. Previous studies that tried to make predictions with the help of technology, such as the research of Rahayuningtyas et al. (2020) said that predictions using GRNN were able to produce quite good predictions with 20% of the total data. Additionally, machine learning technology is used to predict car prices **(Pudaruth, 2014),** stock prices **(Zhang et al., 2021; Patel et al., 2015; Gegic et al., 2019; Vijh et al., 2020)**, prices seasonal goods **(Mahoto et al., 2021; Mohamed et al., 2022)**, bitcoin prices **(Chen et al., 2020)**, and property **(Ho et al., 2021)** or house prices **(Park & Bae, 2015).**

However, based on previous research, there has been no research regarding house price prediction using a sample dataset of house prices in several cities in the USA provided by Kaggle using the Multivariate Adaptive Regression Spline (MARS) method. Therefore, this research was conducted to predict house prices using the MARS method. We also compare the results using the GLMNET method as a comparative method. Furthermore, we analyze the features or variables that influence the price of the house. With the completion of this research, it is hoped that it will provide assistance and knowledge for real estate entrepreneurs and willing buyers of real estate, especially houses, in predicting house prices.

2. METHODS

This research uses the Multivariate Adaptive Regression Spline (MARS) method to create a house price prediction model. The MARS method is a nonparametric regression technique. The MARS method can model nonlinear patterns without knowing the exact shape of the nonlinearity before training the model. The core of the MARS method is the spline, a continuous piecewise polynomial function. This research also compares the results of the MARS method with the Generalized Linear Model via Elastic Net (GLMNET) method. GLMNET is an R

programming language package for performing regular regression, especially with linear regression and logistic regression models. GLMNET utilizes elastic net regulation techniques that combine the L1 penalty (Lasso) and the L2 penalty (Ridge) to overcome multicollinearity problems and to perform automatic feature selection in the model. We carry out five stages in completing case studies: data preparation, basic function selection, model building, model evaluation, and prediction. Figure 1 shows the research design for the house price prediction model in this study.

1) Data Collection

At this stage we carry out three stages, namely collecting relevant data, carrying out data pre-processing and normalization, and separating training and test datasets. We use the house price dataset provided by Kaggle (see https://www.kaggle.com/datasets/shree1992/housedata). There are 10 features provided in the dataset, namely date, price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft_above, sqft_basement, yr_built, yr_renovated, street, city, statezip, and country. Table 1 shows an example of the dataset we used in this study.

We also perform data pre-processing and normalization. We cleaned the data to address missing values and outliers. We use the code below to perform the pre-processing process. We check for null values. We also identified factor variables with less than two levels, this was done to ensure data quality and reliability in statistical analysis and machine learning. Factor variables are categorical variables that group data into different categories (levels). After that, we delete variables that only have one level. We use the preProcess() function to handle factors with one level. Next, we determine the dataset split for training and testing the model. We divide the dataset in the proportion of 80% training data and 20% test data

```
#Checking Null values
sum(is.na(maindf))

# Identifikasi variabel faktor dengan kurang dari dua level
single_level_factors    <-    sapply(trainData,    function(x)    is.factor(x)    &&
length(levels(x))<2)print(names(trainData)[single_level_factors])

# Menghapus variabel yang hanya memiliki satu level
trainData <- trainData[, !single_level_factors]

# Menggunakan preProcess untuk menangani faktor dengan satu level
preProc <- preProcess(trainData, method = c("nzv", "center", "scale"))
```

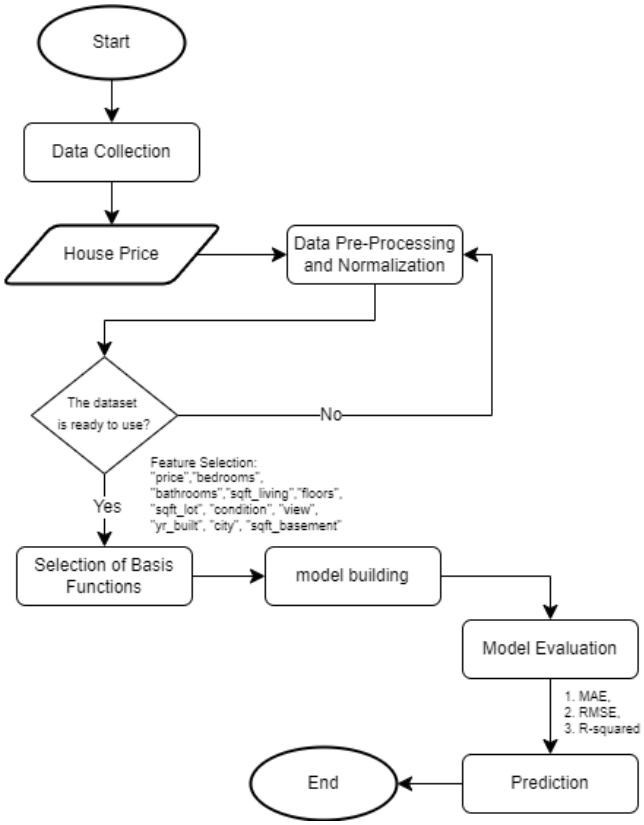

Figure 1. Research design

2) Selection of Basis Functions

The selection of Basis Functions in the MARS context is carried out to identify and construct the basic functions used to build the regression model. Basis functions are the fundamental elements that form the structure of a MARS model, allowing the model to capture non-linear relationships between predictor variables and response variables. At this stage, we determine the dataset features that will be used in the machine learning computing process. There are 11 features that we use, namely "price", "bedrooms", "bathrooms", "sqft_living", "floors", "sqft_lot", "condition", "view", "yr_built", "city", and "sqft_basement ". We also added an "oldbuilt" feature which is processed based on the year the house was built.

```
#Feature selection
maindf <- data [, c ("price","bedrooms", "bathrooms","sqft_living","floors","sqft_lot", "condition", "view", "yr_built", "city", "sqft_basement")]

#Figure out house age
maindf$oldbuilt <- as.integer(format(Sys.Date(), "%Y"))-maindf$yr_built
```

3) Model Building

At this stage, we create a MARS method by combining the basis functions that have been previously determined. MARS adds basis functions iteratively based on their contribution to

explaining the variability of the data. Each iteration adds or removes basis functions to optimize the model.

```
# Mengatur seed untuk reprodusibilitas
set.seed(123)

# Membagi data menjadi data pelatihan (80%) dan data uji (20%)
trainIndex <- createDataPartition(maindf$price, p = 0.8, list = FALSE)
trainData <- maindf[trainIndex, ]
testData <- maindf[-trainIndex, ]

# Melatih model regresi linear menggunakan caret, metode GLMNET
model <- train (price~bedrooms + sqft_living + floors + sqft_lot + condition +
oldbuilt, data = trainData, method = "glmnet")

# Melatih model regresi linear menggunakan caret, metode MARS
model <- train (price~bedrooms + sqft_living + floors + sqft_lot + condition + city
+ oldbuilt, data = trainData, method = "earth", trControl = trainControl(method =
"cv", number = 10))

# Menampilkan informasi model
print(model)
```

4) Model Evaluation

The fourth stage is model evaluation. Model evaluation is carried out using the testing dataset. We carry out analytical calculations of evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared to carry out the evaluation model.

```
# Menghitung R-squared
rsquared_value <- R2(predictions, testData$price)
print (paste ("R-squared:", rsquared_value))

# Menghitung RMSE untuk evaluasi performa
rmse <- sqrt (mean ((testData$price - predictions) ^2))
print (paste("RMSE:", rmse))
```

Table 1. Dataset for House Price Prediction (sumber https://www.kaggle.com/datasets/shree1992/housedata).

Date	Price	Bedroom	Bathroom	Sqft_living	Sqft_lot	Floors	Waterfront	View	Condition	Sqft_above	Sqft_basement	Yr_built	Yr_renovated	Street	City	Statezip	Country
2014-05-02	313000	3	1,5	1340	7912	1,5	0	0	3	1340	0	1955	2005	18810 Densmore Ave N	Shoreline	WA 98133	USA
2014-05-02	2384000	5	2,5	3650	9050	2	0	4	5	3370	280	1921	0	709 W Blaine St	Seattle	WA 98119	USA
2014-05-02	342000	3	2	1930	11947	1	0	0	4	1930	0	1966	0	26206-26214 143rd Ave SE	Kent	WA 98042	USA
2014-05-02	420000	3	2,25	2000	8030	1	0	0	4	1000	1000	1963	0	857 170th Pl NE	Bellevue	WA 98008	USA
2014-05-02	550000	4	2,5	1940	10500	1	0	0	4	1140	800	1976	1992	9105 170th Ave NE	Redmond	WA 98052	USA
2014-05-02	490000	2	1	880	6380	1	0	0	3	880	0	1938	1994	522 NE 88th St	Seattle	WA 98115	USA
2014-05-02	335000	2	2	1350	2560	1	0	0	3	1350	0	1976	0	2616 174th Ave NE	Redmond	WA 98052	USA
2014-05-02	482000	4	2,5	2710	35868	2	0	0	3	2710	0	1989	0	23762 SE 253rd Pl	Maple Valley	WA 98038	USA
2014-05-02	452500	3	2,5	2430	88426	1	0	0	4	1570	860	1985	0	46611-46625 SE 129th St	North Bend	WA 98045	USA
2014-05-02	640000	4	2	1520	6200	1,5	0	0	3	1520	0	1945	2010	6811 55th Ave NE	Seattle	WA 98115	USA
2014-05-02	463000	3	1,75	1710	7320	1	0	0	3	1710	0	1948	1994	Burke-Gilman Trail	Lake Forest Park	WA 98155	USA
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

5) Prediction Model

The final step is to predict the dataset model. We use the predict() function to perform the prediction step.

```
# Membuat prediksi menggunakan data uji
predictions <- predict (model, newdata = testData)

# Menampilkan beberapa prediksi
print(predictions)

# Plot prediksi vs. nilai sebenarnya
ggplot() + geom_point(data = testData, aes(x = price, y = predictions), color =
"blue") + geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed")
+ labs(x = "Actual Score", y = "Predicted Score", title = "Actual vs. Predicted
Score")

x = lm (formula = price~bedrooms + sqft_living + floors + sqft_lot + condition +
city + oldbuilt, data = maindf)
summary(x)
```

3. RESULTS AND DISCUSSION

3.1. Accuracy Test and Prediction Results

We conducted experiments by testing the accuracy of several feature combinations using 2 methods, namely the GLMNET and MARS methods. Accuracy testing is carried out using simple functions found in the R programming language, namely R2() for R-Squared values and sqrt(mean()) for RMSE values. Table 2 shows an experimental table of R-Squares and RMSE value results from several feature combinations along with their average values in the training process.

Table 2. Experimental table - R-Squared and RMSE values.

No	Combination of Features	Result			
		GLMNET		MARS	
		R-Squared	RMSE	R-Squared	RMSE
1	bedrooms + bathrooms + sqft_living + floors + sqft_lot + condition + city + oldbuilt	0.3514294	0.7778154	0.4739143	0.675889
2	bedrooms + sqft_living + floors + sqft_lot + condition	0.2909282	0.8493189	0.4059363	0.7041254
3	bedrooms + sqft_living + floors + sqft_lot + oldbuilt	0.3223301	0.7762791	0.4184008	0.6898235
4	bedrooms + sqft_living + floors + sqft_lot + condition + city + oldbuilt	0.4110614	0.6806642	0.5154666	0.6263589
5	bedrooms + bathrooms + sqft_living + floors + sqft_lot + condition + oldbuilt	0.2985517	0.8154845	0.4149394	0.6989579
6	bedrooms + sqft_living + floors + sqft_lot + condition + oldbuilt	0.3358965	0.7545535	0.4454765	0.685958
7	bedrooms + bathrooms + sqft_living + floors + sqft_lot + condition + oldbuilt	0.2660824	0.9320492	0.3891602	0.7121714
	Rata-Rata	0.3251828	0.798024	0.441471	0.681879

In Table 2, it can be seen that the largest R-Squared value and the smallest RMSE value are in the feature combination model number 4, where the R-Squared is 0.5154666 and the RMSE is 0.6263589. The R program code for the best value accuracy results is

```
model <- train (price~bedrooms + sqft_living + floors + sqft_lot + condition + city
+ oldbuilt, data = trainData, method = "earth", trControl = trainControl(method =
"cv", number = 10))
```

The average R-Squared value using the MARS method is greater than the GLMNET method, the average RMSE value is also smaller. Therefore, the MARS method is superior in this study. Apart from that, the results of the house price prediction accuracy test using the MARS method can be seen in Figure 2.

```
Multivariate Adaptive Regression Spline

3680 samples
   7 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3312, 3312, 3312, 3312, 3312, 3312, ...
Resampling results across tuning parameters:

  nprune  RMSE       Rsquared   MAE
   2      0.6661027  0.4100584  0.3183232
   9      0.6366081  0.4959517  0.2653022
  17      0.6263589  0.5154666  0.2580837

Tuning parameter 'degree' was held constant at a value of 1
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were nprune = 17 and degree = 1.
```

Figure 2. Accuracy test results using the MARS method

In Figure 2, nprune 2 is defined as low nprune, where the model is very simple, the risk of overfitting is low, but may fail to capture complex patterns (underfitting). In nprune 9, it is defined as mid-level nprune, where models with a balance between complexity and simplicity provide more accurate predictions with a moderate risk of overfitting. And at nprune 17, it is defined as high nprune, where the model is very complex and able to capture complex patterns in the data, but with a higher risk of overfitting.

Apart from that, Figure 2 shows the accuracy results using the MARS method on feature combination number 4. In nprune 17, the resulting R-Squared value is 0.5154666, indicating that around 51.55% of the variance or variation in the dependent variable can be explained by the regression model. which has been made. In other words, more than half of the variability in the data can be explained by the independent variables included in the model. This value indicates that the model has moderate explanatory power. It's not a low value, but it's not a very high one either. This means that the model is quite good at capturing the relationship between the independent and dependent variables, but there is still almost 48.45% of the variance that cannot be explained by the model.

Meanwhile, the RMSE value of 0.6263589 indicates that the average model prediction error is around 0.63 units (depending on the units of the dependent variable used). The lower the RMSE value, the better the model in terms of prediction accuracy. This RMSE value can be considered good or bad depending on the context and scale of the data used. For example, in the context of temperature prediction (in degrees), an RMSE of 0.63 might be considered good enough, while in the context of house price prediction (in dollars), the same RMSE might be considered inadequate. The following is the code and results of the linear model (lm) for feature combination number 4.

```
lm(formula = price~bedrooms + sqft_living + floors + sqft_lot + condition + city +
oldbuilt, data = maindf)
```

Figure 3 shows a linear model (lm) where the dependent variable is price, and the independent variables (predictors) are bedrooms, sqft_living, floors, sqft_lot, condition, view, and oldbuilt. The data comes from a dataframe called maindf. Apart from that, it can be concluded as follows:

i. Residual standard error: 489,500 at 4550 degrees of freedom. It shows the distance of the average observed value from the regression line.

ii. Multiple R-squared: 0.2543. About 25.32% of the variability in prices is explained by the model.

iii. Adjusted R-squared: 0.2463. shows that the regression model used can explain around 24.63% of the variability in the dependent variable after taking into account the number of predictor variables. This provides a more accurate assessment of the model's explanatory power, particularly when considering model complexity.

iv. F-statistic: 31.67 at 49 and 4550 degrees of freedom, p-value: < 2.2e-16. This shows that the overall model is statistically significant ($p < 0.05$). This means that there is a significant relationship between the independent variables in the model and the dependent variable, providing confidence that the model is relevant for prediction or further analysis.

```
Call:
lm(formula = price ~ bedrooms + sqft_living + floors + sqft_lot +
    condition + city + oldbuilt, data = maindf)

Residuals:
     Min       1Q   Median       3Q      Max
-2193471  -105052   -11437    67155 26352566

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.395e+05  2.250e+05  -1.065  0.28708
bedrooms       -4.770e+04  1.006e+04  -4.743 2.17e-06 ***
sqft_living     2.679e+02  1.069e+01  25.060  < 2e-16 ***
floors          3.264e+04  1.698e+04   1.922  0.05467 .
sqft_lot       -1.959e-01  2.242e-01  -0.874  0.38216
condition       2.735e+04  1.202e+04   2.275  0.02296 *
oldbuilt        9.223e+02  3.509e+02   2.628  0.00861 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 489500 on 4550 degrees of freedom
Multiple R-squared:  0.2543,    Adjusted R-squared:  0.2463
F-statistic: 31.67 on 49 and 4550 DF,  p-value: < 2.2e-16
```

Figure 3. Accuracy results with linear models

Table 3 shows a sample of prediction results using the best model, namely with a combination of feature number 4 using the MARS method. Table 3 presents the results of the initial house prices listed in the dataset, as well as the predicted house prices based on the best prediction model that we have previously created using the MARS method. The predicted price is influenced by 10 variables or features, namely bedrooms, bathrooms, Sqft_living, floors, sqft_lot, condition, view, city, sqft_basement, and old_built.

The prediction results have an R-squared of 0.427036490654884, which indicates that the model explains approximately 42.7% of the variance in the dependent variable. This is a fairly good level of explanation but shows that there is still room for improvement and further research. Meanwhile, the resulting RMSE value is 648344.50456136, which shows that the

average model prediction error is around 648344.5 in dependent variable units. This value can be said to be very high and requires a mode improvement process such as adding more relevant variables, changing the shape or transformation of variables, or using more complex modeling techniques.

Table 2. Prediction results using the MARS method.

price	Bedrooms	Bathrooms	Sqft_living	floors	Sqft_lot	condition	view	city	sqft_basement	old built	price_prediction
313 000	3	1.50	1340	1.5	7912	3	0	Shoreline	0	69	1.497.656
420 000	3	2.25	2000	1.0	8030	4	0	Bellevue	1000	61	2.234.217
335 000	2	2.00	1350	1.0	2560	3	0	Redmond	0	48	1.507.706
482 000	4	2.50	2710	2.0	35868	3	0	Maple Valley	0	35	3.024.486
257 950	3	1.75	1370	1.0	5858	3	0	Des Moines	0	37	1.528.963
435 000	4	1.00	1450	1.0	8800	4	0	Bellevue	0	70	1.620.826
495 000	4	1.75	1600	1.0	6380	3	0	Seattle	470	65	1.787.810
675 000	5	2.50	2820	2.0	67518	3	0	Issaquah	0	45	3.147.871
604 000	3	2.50	3240	2.0	33151	3	2	Federal Way	0	29	3.615.537
750 000	3	2.50	2390	1.0	6550	4	2	Seattle	950	69	2.669.980
…	…	…	…	…	…	…	…	…	…	…	…

3.2. Data Visualization Analysis

Figure 4 shows the correlogram of housing dataset. We use the ggcorrplot() function to create the correlogram. Correlogram with ggcorrplot() serves to create a clear and informative visualization of the correlation matrix (Londe et al., 2022). This helps in identifying relationships between variables in the dataset visually. Meanwhile, Figure 5 shows a scatterplot matrix of the variable's bedrooms, sqft_living, floors, and condition. The scatter plot in the matrix shows the relationship between two variables, this helps in identifying patterns or linear relationships between these variables (Friendly, 2002).

The different colors in Figure 4 have meaning where the color "tomato2" is used to show a strong negative correlation, the color "white" shows a neutral correlation, and the color "springgreen3" shows a strong positive correlation. The correlation value shows the relationship between two variables at x and y coordinates. Variables with high correlation have a significant relationship and can be considered for further analysis. Meanwhile, variables with low correlation or close to zero do not have a significant linear relationship and may be less relevant to the purpose of the analysis being carried out.

Based on the correlation diagram analysis shown in Figure 4, it is known that the number of beds has the highest correlation with the building area variable with a correlation value of 0.6.

The condition variable has the largest correlation value with the age of the building, the house price variable has the largest correlation value with the building area variable, and the building area has the highest correlation with the number of floors. In this research, it is known that house prices have a high correlation with the number of beds, building area, number of floors, land area and views of the house. Other correlation values can be seen in Figure 4.

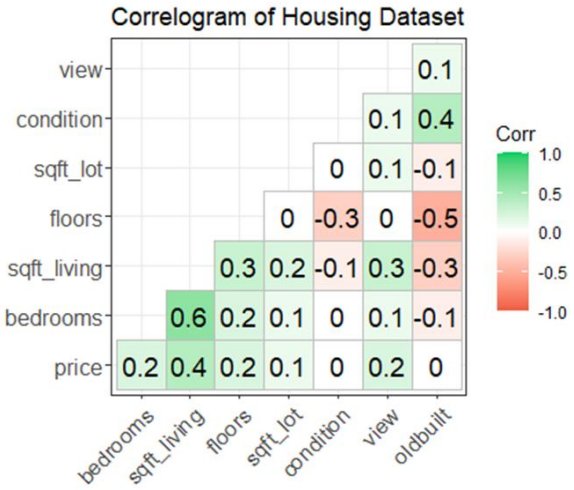

Figure 4. Correlogram of housing dataset.

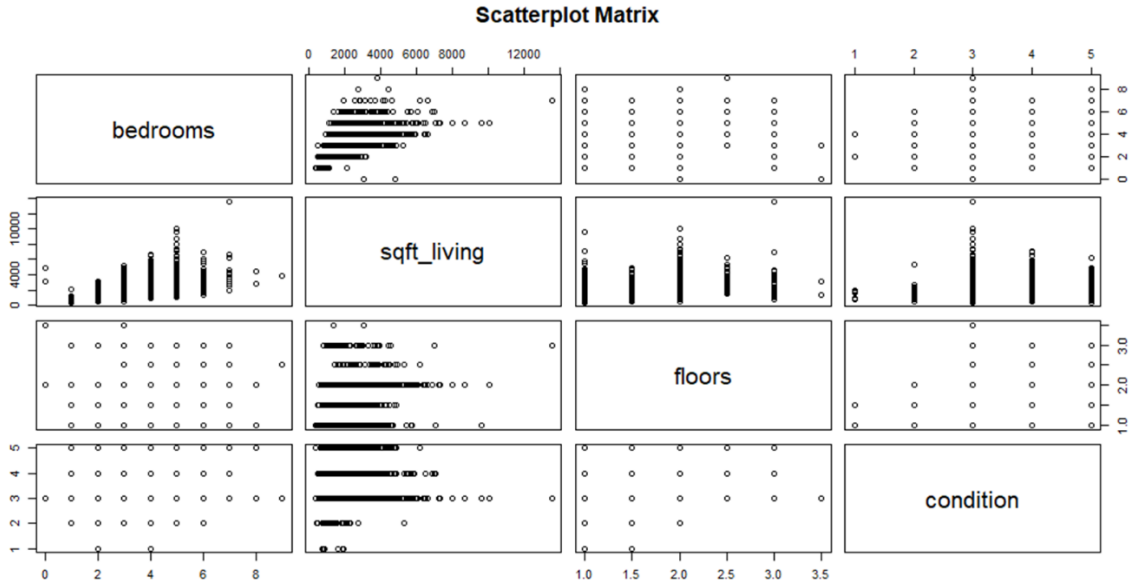

Figure 5. Scatterplot matrix variable bedrooms, sqft_living, floors, dan condition.

Figure 6 shows important variables that influence prediction results using the MARS method. Determining variable importance in the dataset is carried out to improve model performance (Ahrazem Dfuf, 2022). Reducing the number of features or variables by retaining only the most relevant ones can improve model performance. This can also reduce overfitting, especially on datasets with a very large number of features. By removing irrelevant features, the model becomes simpler and better at generalizing to new data. Irrelevant features can interfere with model performance, identifying and removing these features helps in improving prediction accuracy (Syaputra et al., 2023). Apart from that, knowing the most influential

features helps in understanding the conditions in which the model makes predictions. With fewer features, the time required to train the model is shorter, which is especially beneficial for complex models or very large datasets.

When it comes to user needs, understanding the most important features can provide valuable insight into the problem domain at hand. In the case taken in this research, it is known that the sqft_living feature has the highest importance value, namely 100, this shows that in predicting house prices it is known that sqft_living is more important than other aspects such as city location, and house condition., bedrooms, and the age of the building. The results of this analysis can help decision-makers in the real estate sector. Figure 7 shows the univariate linear regression plot between sqft_living and price.

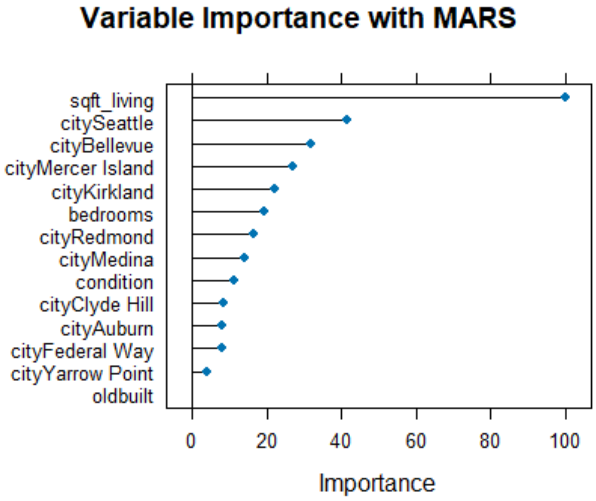

Figure 6. Variable importance with MARS Method.

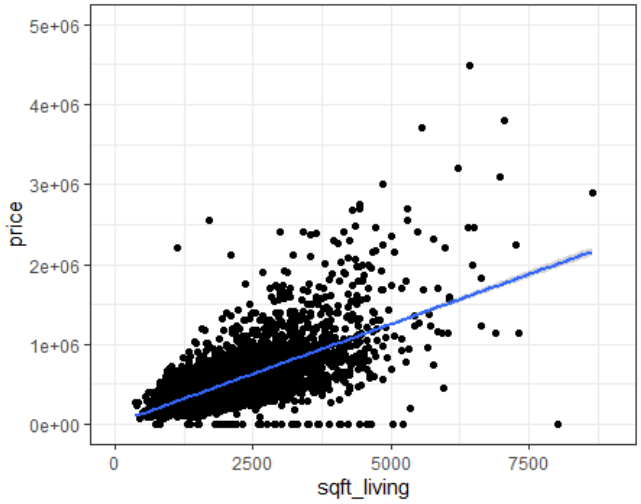

Figure 7. Plot univariate linear regression between sqft_living and price.

3.3. Comparison with Previous Research

In the research of Naser et al. (2022), the MARS method is used to predict the compressive strength of environmentally friendly concrete. In the training dataset process, the MARS model

produced an RMSE of 4,192, an R-squared of 0.889, and a MAE of 3,334. The performance metrics in this research prove the superiority of the MARS model when compared to Random Forest and Support Vector Machine. The R-squared and RMSE results are better than the results we got. However, we try to analyze and test a variety of important features and variables during the testing process. Meanwhile, in research (Oduro et al., 2015), researchers used the Boosting-Multivariate Adaptive Regression Splines (B-MARS) algorithm to increase the accuracy of MARS modeling. The best R-squared result produced was 93% or 0.93 with an RMSE of 0.00011. In research (Cheng & Cao, 2014), researchers used evolutionary multivariate adaptive regression splines (EMARS) to efficiently predict building energy performance. EMARS is a combination of multivariate adaptive regression splines (MARS) and artificial bee colonies (ABC). In EMARS, MARS performs learning and curve fitting and ABC performs optimization to determine the most suitable parameter settings with minimal prediction error. The resulting R-squared results are almost close to 1, namely 0.989 for the cooling load (CL) and 0.998 for the heating load (HL). Meanwhile, the largest average RMSE value produced was 2.49.

4. CONCLUSION

The research results show that prediction results using the MARS model have better results compared to the GLMNET model. The average R-squared value produced by the MARS method is greater than the GLMNET method, namely 0.5154666 with an average RMSE value of 0.6263589. This research also shows that house prices have the largest correlation with sqft_living. We found that the variable that has the most influence on house prices is the sqft_living variable or house area. The variable sqft_living has the highest import value, namely 100. This research shows that knowing sqft_living is more important than other aspects such as city location, condition of the house, bedrooms, and age of the building in determining house prices. Despite the results we got, this research still has many shortcomings and requires many improvements to improve the performance and accuracy of prediction results.

DAFTAR RUJUKAN

Ahrazem Dfuf, I. (2022). Application of machine learning to variable importance analysis.

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395.

Cheng, M. Y., & Cao, M. T. (2014). Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. *Applied Soft Computing Journal*, *22*, 178–188.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, *56*(4), 316-324.

Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. *TEM Journal*, *8*(1), 113.

Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, *38*(1), 48-70.

Kaushal, A., & Shankar, A. (2021, April). House price prediction using multiple linear regression. *Proceedings of the International Conference on Innovative Computing & Communication (ICICC)*. http://dx.doi.org/10.2139/ssrn.3833734

Londe, V., Reid, J. L., Farah, F. T., Rodrigues, R. R., & Martins, F. R. (2022). Estimating optimal sampling area for monitoring tropical forest restoration. *Biological Conservation*, *269*, 109532.

Mahoto, N. A., Iftikhar, R., Shaikh, A., Asiri, Y., Alghamdi, A., & Rajab, K. (2021). An intelligent business model for product price prediction using machine learning approach. *Intelligent Automation & Soft Computing*, *30*(1), 148-159.

Mohamed, M. A., El-Henawy, I. M., & Salah, A. (2022). Price prediction of seasonal items using machine learning and statistical methods. *Computers, Materials & Continua*, *70*(2), 3473-3489.

Naser, A. H., Badr, A. H., Henedy, S. N., Ostrowski, K. A., & Imran, H. (2022). Application of multivariate adaptive regression splines (MARS) approach in prediction of compressive strength of eco-friendly concrete. *Case Studies in Construction Materials*, *17*(July), e01262.

Oduro, S. D., Metia, S., Duc, H., Hong, G., & Ha, Q. P. (2015). Multivariate adaptive regression splines models for vehicular emission prediction. *Visualization in Engineering*, *3*(1), 3-13.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928-2934.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, *42*(1), 259-268.

Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, *4*(7), 753-764.

Rahayuningtyas, E. F., Rahayu, F. N., & Azhar, Y. (2021). Prediksi harga rumah menggunakan general regression neural network. *Jurnal Informatika*, *8*(1), 59-66.

Sanusi, R. M., Ansori, A. S. R., & Wijaya, R. (2020). Prediksi harga rumah di kota bandung bagian timur dengan menggunakan metode regresi. *eProceedings of Engineering*, *7*(3), 9381.

Singh, A., Sharma, A., & Dubey, G. (2020). Big data analytics predicting real estate prices. *International Journal of System Assurance Engineering and Management*, *11*, 208-219.

Syaputra, A., Muslim, B., & Prawira, N. S. (2023). Implementasi metode support vector machine dengan algoritma genetika pada prediksi konsumsi energi untuk gedung beton bertulang. *Faktor Exacta*, *16*(2), 142-153.

Utamaningsih, A., Monika, G. & Yenika (2019). Motivasi kerja karyawan dalam kajian teori kebutuhan maslow. *J. Ilmiah Poli Bisnis*, *11*(2), 133–142.

Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia computer science*, *167*, 599-606.

Wu, Z. (2020). Prediction of California house price based on multiple linear regression. *Academic Journal of Engineering and Technology Science*, *3*(7.0).

Zhang, J., Li, L., & Chen, W. (2021). Predicting stock price using two-stage machine learning techniques. *Computational Economics*, *57*, 1237-1261.

Zhang, Q. (2021). Housing price prediction based on multiple linear regression. *Scientific Programming*, *2021*, 1-9.