Sale Price Prediction

Engin Deniz Erkan
Department of Data Informatics
Middle East Technical University
Ankara, TURKEY
engin.erkan@metu.edu.tr

***Abstract*— Predicting residential property prices is a challenging problem in the field of real estate analytics that is impacted by a wide range of factors. This study focuses into two essential aspects of this prediction problem. First, the effect of several data transformation and normalization methods on machine learning algorithms' predictive capabilities is evaluated. Among the methods investigated are Z-score normalization, Min-Max normalization, and logarithmic transformation. Secondly, the impact of various feature selection strategies on these algorithms' performance in high-dimensional datasets is investigated, with a particular emphasis on Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). The principal variable in the study is the sale price of the properties, and it makes use of a dataset with 81 features that includes both numerical and categorical variables.**

Keywords— Sale Price Prediction, Machine Learning, Predictive Performance, High-dimensional Datasets, Real Estate Analytics

I. Introduction

In the real estate market, estimating the price at which residential properties are likely to sell is an essential effort that has major benefits for purchasers, sellers, and policymakers. However, there are significant problems due to the high dimensionality and complexity of real estate datasets, which can contain a wide range of numerical and categorical information. Two crucial problems in this context are addressed by this research:

1) How do different data normalization and transformation techniques influence the predictive performance of machine learning algorithms?

2) How do different feature selection algorithms influence the predictive performance of machine learning algorithms for high-dimensional datasets?

To manage the different scales and distributions of the features, efficient preprocessing methods like logarithmic transformation, Min-Max normalization, and Z-score normalization are crucial. Similarly, dimensionality reduction and enhanced model interpretability and performance depend on feature selection techniques like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE).

The Random Forest and CatBoost algorithms are used in this work to create prediction models. When training numerous decision trees, Random Forest, an ensemble learning technique, generates the mean prediction for regression tasks or the mode of the classes for classification tasks. It works well with large datasets and minimizes overfitting, therefore it's appropriate for high-dimensional data. The gradient boosting technique CatBoost performs well and requires comparatively less data preparation; it is especially skilled at handling categorical information.

A key factor in optimizing these models' performance is hyperparameter tuning. Predictive accuracy for Random Forest may be greatly increased by adjusting parameters like the number of trees, depth of each tree, and the minimum number of samples needed to divide an internal node. Achieving ideal performance with CatBoost requires altering parameters such as learning rate, depth, and number of repetitions.

This research does an extensive investigation to evaluate the effects of different feature selection strategies and preprocessing approaches on Random Forest and CatBoost model performance. The study attempts to determine the best practices for improving the accuracy of forecasts of prices for sale in high-dimensional real estate datasets by carefully tuning the hyperparameters for each model.

II. Literature Review

The paper investigates the use of machine learning algorithms in house price prediction. The study stated that there are two main research trends in the housing market analysis and house price valuation literature: hedonic-based regression approach and artificial intelligence techniques. Hedonic-based methods have been used to identify the relationship between house prices and housing characteristics. However, these methods have some limitations regarding basic model assumptions and prediction. C4.5, RIPPER, Naive Bayesian and AdaBoost algorithms are tested using three-way data splitting and 10-fold cross-validation methods to reveal the performance levels of the models. The findings show that the RIPPER algorithm has a lower error rate than other methods. Naive Bayesian has the highest error rate. This shows how different data splitting methods and cross-validation techniques affect classifier performances. This study highlights the importance of data splitting and cross-validation methods to evaluate the performance of machine learning classifiers. The RIPPER algorithm showed the best performance with a low error rate. In the future, adding more data and features may improve classifier performance [1].

The author's research focuses on forecasting home prices using an upgraded machine learning technique that takes

advantage of the House Price Index (HPI) and other key factors. It contrasts standard regression approaches with sophisticated machine learning techniques, stressing the value of innovative feature engineering and hybrid regression models. The results show that hybrid models, namely those combining Lasso and Gradient Boosting, outperform individual regression approaches, with the lowest RMSE score of 0.11260. The study emphasizes the importance of feature engineering in increasing model performance. Adding additional characteristics and selectively picking the appropriate subset improves prediction accuracy dramatically. The investigation demonstrates that, while additional features typically improve performance, the effect plateaus after a certain number of characteristics, which has been determined as about 230. Hybrid models, particularly those that combine Lasso and Gradient Boosting, outperform other algorithms by harnessing their respective strengths. This combination creates a strong prediction framework that reduces overfitting and captures complicated patterns in the data. Creative feature engineering and hybrid regression approaches are critical for making accurate house price projections. Using several regression algorithms dramatically improves prediction accuracy. Future research could investigate more external factors, such as economic indicators and demographic changes, to improve model accuracy [2].

Machine learning techniques developed for predicting housing prices are examined in the paper. In particular, hybrid regression and stacked generalization methods were used. The aim of the study is to compare various advanced machine learning models and comprehensively validate the performance of these models. The results show that advanced models give optimistic results in house price prediction. The study's findings show that adding more features increases prediction accuracy. In particular, Lasso regression was highly effective in feature selection, and removing unnecessary features reduced the error of other regression methods. The study also reveals that the combination of different regression algorithms outperforms a single algorithm. This study shows that creative feature engineering and hybrid regression methods provide significant improvements in house price prediction. In the future, incorporating more variables, such as the economic cycle, population movements, and interest rates, could improve the accuracy of forecasts [3].

III. Data Preprocessing

A. *Statistics about the datasets*

There are totally 36 numerical columns, 43 categorical columns and one numerical target variable in the dataset. Training data consist of 1314 row data and test data consists of 146 row data.

The SalePrice ranges from 34900 $ to 755000 $ with a mean of approximately 180964 $.

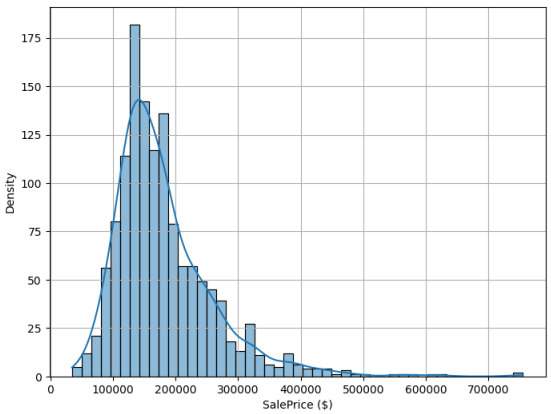

Fig. 1. Sale Price Histogram

B. *Missing Values*

Table 1 shows how much missing data is available in the train dataset and its percentage.

TABLE I. Missing Values

Columns	Missing Data Count	Missing Data Percentage (%)
PoolQC	1307	99.4673
MiscFeature	1265	96.2709
Alley	1232	93.7595
Fence	1054	80.2131
MasVnrType	787	59.8935
FireplaceQu	621	47.2603
LotFrontage	231	17.5799
GarageType	73	5.55556
GarageYrBlt	73	5.55556
GarageFinish	73	5.55556
GarageQual	73	5.55556
GarageCond	73	5.55556
BsmtFinType2	33	2.51142
BsmtFinType1	32	2.43531
BsmtExposure	32	2.43531
BsmtCond	32	2.43531
BsmtQual	32	2.43531
MasVnrArea	8	0.608828
Electrical	1	0.0761035

When the attribute definitions in the data set were examined, it was seen that the reason why these data were missing was because the house did not have certain features. Therefore, instead of treating these as missing data, an extra category was added. In this way, it is aimed to use every data in the best possible way. The missing numbers in the numeric variables were filled with zero median or mode imputation, depending on the definition of attribute.

C. Handling Categorical Variables

One popular method for managing categorical variables in machine learning problems is label encoding. Label encoding associates each distinct category with a distinct number, in contrast to other encoding techniques like one-hot encoding, which can greatly increase the dataset's complexity. When working with large data sets or when memory restrictions are an issue, this compact representation is very helpful since it conserves memory and computational resources. Furthermore, if an ordinal connection exists between categories, label encoding maintains it, which might be essential for some algorithms to correctly analyze the data. Label encoding is a suitable option for preparing categorical variables in machine learning pipelines since it is very easy to implement and can be performed effectively across several category columns in the dataset. In this study, label encoding is performed for all categorical variables.

D. Appying Different Data Normalization and Ttansformation Techniques

To prepare datasets for feeding into machine learning models, it is essential to apply various data transformation and normalization procedures. By scaling features to have a mean of zero and a standard deviation of one, standardization makes sure that every feature adds the same amount to the model and keeps variables with greater scales from becoming dominant. On the other hand, min-max scaling constrains data points inside a restricted interval while maintaining the relative relationships between them by transforming characteristics to a specific range, usually between zero and one. Finally, to deal with skewed distributions, log transformation is used to stabilize variance and improve the symmetry of the data. By improving the data's suitability for machine learning algorithms, these preparation stages improve the interpretability and prediction capabilities of the models.

E. Feature Selection Algorithms

Two feature selection methods that are often utilized in machine learning applications are Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA). These methods are especially useful when working with high-dimensional datasets like real estate data. RFE works by fitting a model iteratively and removing the least significant features until the target feature count is achieved. This method works especially well when choosing a subset of characteristics to maximize model performance is the objective. However, PCA is a dimensionality reduction method that creates a new collection of orthogonal features known as principle components from the original data. PCA reduces the dimensionality of the data while maintaining as much of the original information as feasible by identifying the directions, or principle components, that maximize the variance in the data.

Reducing the dimensionality of the dataset and increasing model performance are two advantages of both RFE and PCA. They vary, nevertheless, in their methodology and implementation. When choosing a subset of characteristics for predictive modeling based on their significance, RFE is a better option. It gives information about the relative weight of each feature and works especially well when paired with algorithms like Random Forest that are capable of estimating feature importance automatically. PCA, on the other hand, is recommended when dimensionality reduction and identifying underlying patterns or structure in the data are the main goals. When dealing with multicollinearity among features or when displaying the data in lower dimensions is needed, it is quite helpful.

For this data set, these two feature selection algorithms were used and their effect on the prediction model was examined.

IV. Experiments

A. Naïve Model

The study's baseline approach provides a straightforward, yet useful, standard for estimating the price at which residential properties will be sold. The way this model works is that it takes the average selling price from the training dataset for each neighborhood and uses that number to predict the sale price of properties in the corresponding neighborhoods in the test dataset. This simple method takes use of the locality-based phenomenon that is typically seen in real estate markets, when homes within the same area frequently display comparable pricing trends. To evaluate the performance of the model, evaluation measures such as Mean Absolute Percentage Error (MAPE), R-squared (R2), and Root Mean Squared Error (RMSE) are calculated. These measures offer information on how well the baseline model captures the variation in selling prices between neighborhoods. Table 2 shows the results of evaluation metrics for the test dataset.

TABLE II. NAÏVE MODEL EVALUATION METRICS

Metric	Value
RMSE	46071
R2	0.61835
MAPE	17.4311

B. CatBoost Model

CatBoost is an effective gradient boosting method that is especially well-suited for real-world datasets that contain both numerical and categorical variables, such information on residential properties. It is specifically built to handle categorical features with ease. It makes use of gradient boosting, a method that constructs an ensemble of decision trees in a stepwise manner, with each new tree correcting the

errors produced by the older ones. Because CatBoost handles the encoding of categorical variables directly, it stands out for its natural support of categorical features without the need for any preprocessing. Because of this, CatBoost is a desirable option for predictive modeling jobs since it provides excellent accuracy and user-friendliness, particularly in situations where categorical variables are crucial to the final result.

The performance of the CatBoost algorithm was assessed in this study using the following metrics: Mean Absolute Percentage Error (MAPE), R-squared (R2), and Root Mean Squared Error (RMSE). The root mean square error (RMSE) provides a thorough evaluation of the accuracy of the model by calculating the average magnitude of the errors between the estimate and actual selling prices. Indicating the goodness-of-fit of the model, R2 shows how much of the variation in the target variable is explained by it. In order to determine how accurate a forecast is in relation to real values, MAPE computes the average percentage difference between predicted and actual selling prices. These metrics are crucial for evaluating the CatBoost algorithm's generalization performance and predictive ability across various datasets and hyperparameter setups.

In this work, measurements are logged during model training and hyperparameters are observed using the Weight and Biases (wandb) platform. With wandb, tracking and visualizing model performance in a variety of hyperparameter configurations is simple, making it easier to optimize hyperparameters and compare outcomes. It also offers a user-friendly interface for tracking the development of the model's training and analyzing how changes to its hyperparameters affect its performance.

The learning rate and depth are two crucial hyperparameters that are adjusted in the CatBoost algorithm in addition to the evaluation metrics. During gradient descent, the learning rate determines the step size at each iteration, which affects the rate of convergence and the amount of updates to the model parameters. While a greater learning rate helps speed training, it can also cause instability or divergence. A lower learning rate can lead to delayed convergence but can help prevent overshooting the ideal solution. The maximum depth of every decision tree in the ensemble is set by the depth parameter. Although deeper trees might better capture complex interactions in the data, they may also make overfitting more likely, particularly in smaller datasets. To achieve optimal model performance, it is important to balance the learning rate and depth since these hyperparameters interact to determine the trade-off between generalization ability and model complexity. The work aims to determine the optimal mix of learning rate and depth for optimizing prediction accuracy and resilience in the CatBoost models through hyperparameter tuning. Table 3 shows the results of the catboost algorithm on the same data that has undergone different preprocessing processes.

TABLE III. CATBOOST MODEL EVALUATION METRICS

Dataset	Learning Rate	Depth	RMSE	R2	MAPE
Standard_RFE	0.01	4	27109.8	0.867851	10.7309
Standard_RFE	0.05	4	26192.7	0.876641	10.3899
Standard_RFE	0.01	6	25740.5	0.880864	10.0986
Standard_RFE	0.05	6	25537.7	0.882734	9.95418
Standard_PCA	0.01	4	25503.9	0.883044	11.0552
Standard_PCA	0.05	4	26054.2	0.877942	11.7007
Standard_PCA	0.01	6	27431.8	0.864694	11.3805
Standard_PCA	0.05	6	27000.5	0.868915	11.1594
MinMax_RFE	0.01	4	71671.9	0.0763495	39.5909
MinMax_RFE	0.05	4	73449.9	0.029952	39.2245
MinMax_RFE	0.01	6	69343.1	0.135397	38.8437
MinMax_RFE	0.05	6	67302.7	0.18553	36.4377
MinMax_PCA	0.01	4	46926.8	0.60404	20.6881
MinMax_PCA	0.05	4	48588.8	0.575496	21.7378
MinMax_PCA	0.01	6	47926.4	0.586991	20.9379
MinMax_PCA	0.05	6	48924	0.569618	21.078
Log_RFE	0.01	4	27022.2	0.868704	10.4241
Log_RFE	0.05	4	24830.3	0.88914	9.97545
Log_RFE	0.01	6	24930.5	0.888244	9.83567
Log_RFE	0.05	6	24065.6	0.895863	9.65022
Log_PCA	0.01	4	55115.9	0.453784	22.733
Log_PCA	0.05	4	55756.9	0.441005	22.6248
Log_PCA	0.01	6	57240	0.410872	23.6369
Log_PCA	0.05	6	58575.8	0.383056	24.4044

C. Random Forest Model

For regression and classification problems, the Random Forest algorithm is a flexible and effective ensemble learning technique that is often utilized, especially when dealing with high-dimensional and diverse datasets like real estate data. During training, it creates a large number of decision trees, from which it produces the mean prediction for regression tasks or the mode of the classes for classification tasks. Because Random Forest can handle both numerical and categorical characteristics well without requiring a lot of preprocessing, it is especially well-suited for these kinds of datasets. It manages feature interactions and non-linear connections automatically. It also has an integrated feature significance estimate function, which makes it possible to determine which features have the biggest impact on prediction performance. For real estate price prediction applications, Random Forest stands out as a dependable and understandable algorithm that provides excellent accuracy and resistance to complicated and diverse information.

The depth of the trees and the total number of trees in the forest are two important hyperparameters that were tuned in this study. In order to prevent overfitting and restrict the complexity of individual decision trees, the depth parameter sets the maximum depth of each decision tree in the ensemble. While a deeper tree can capture more complicated relationships in the data but may raise the danger of

overfitting, a shallow tree with limited depth may result in underfitting. The number of trees in the forest is specified by the number of trees parameter, which also affects the overall resilience and complexity of the model. While adding more trees increases computational cost, it often improves model stability and performance. The work aims to determine the best combination that improves prediction accuracy and generalization performance across various datasets and feature selection methodologies by fine-tuning these hyperparameters using a grid search strategy. During model training, the Weight and Biases (wandb) platform is used for hyperparameter tuning and recording metrics. It offers a useful interface for monitoring and displaying the impact of various hyperparameter configurations on model performance. Table 4 shows the results of the catboost algorithm on the same data that has undergone different preprocessing processes.

TABLE IV. RANDOM FOREST MODEL EVALUATION METRICS

Dataset	Max Depth	Num Estimators	RMSE	R2	MAPE
Standard_RFE	4	50	31671.4	0.819638	13
Standard_RFE	4	100	31038.9	0.826771	12.8742
Standard_RFE	6	50	29797	0.840355	11.6864
Standard_RFE	6	100	28955	0.84925	11.5711
Standard_PCA	4	50	29636.6	0.842069	11.941
Standard_PCA	4	100	29274.9	0.845901	11.9751
Standard_PCA	6	50	29176.4	0.846936	11.9414
Standard_PCA	6	100	28556.9	0.853366	11.8738
MinMax_RFE	4	50	54411.5	0.467657	25.5641
MinMax_RFE	4	100	55587.5	0.444398	25.7164
MinMax_RFE	6	50	60633	0.338959	27.3622
MinMax_RFE	6	100	62099.3	0.306599	27.5155
MinMax_PCA	4	50	48801.8	0.571765	18.7634
MinMax_PCA	4	100	48411	0.578596	18.8918
MinMax_PCA	6	50	52992.4	0.495063	20.0922
MinMax_PCA	6	100	51977.8	0.514213	20.0724
Log_RFE	4	50	31878.7	0.81727	12.6911
Log_RFE	4	100	31722.6	0.819054	12.5644
Log_RFE	6	50	29914.4	0.839094	11.2205
Log_RFE	6	100	29051.3	0.848246	11.0945
Log_PCA	4	50	55895.9	0.438215	22.429
Log_PCA	4	100	55277.8	0.450571	22.4761
Log_PCA	6	50	55875.2	0.438632	22.7194
Log_PCA	6	100	55058.5	0.454922	22.8431

V. DISCUSSIONS

Two main research questions were the focus of the analysis: how various feature selection algorithms impact the predictive performance of machine learning models for high-dimensional datasets, and how various data normalization and transformation techniques impact the predictive performance of machine learning algorithms. Both the CatBoost and Random Forest models were used in the evaluation, along with different combinations of the feature selection strategies Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) and normalization techniques Standard, MinMax, and Logarithmic Transformation.

The outcomes showed that the model's performance varied significantly depending on the data preparation techniques selected. The CatBoost model outperformed the Random Forest model with a MAPE of 11.0945 when looking at the major performance indicator, Mean Absolute Percentage Error, or MAPE. CatBoost's MAPE was 9.65022. With a MAPE of 17.4311, the naïve baseline model was significantly surpassed by both models, demonstrating the efficacy of the machine learning techniques applied.

The combination of Logarithmic Transformation and Recursive Feature Elimination (RFE) proved to be the most effective for both CatBoost and Random Forest models, demonstrating its effectiveness in managing the high-dimensional real estate dataset. On the other side, the least successful coupling was that of RFE and MinMax scaling, indicating that this specific combination might not be appropriate for the dataset in question.

It is interesting to note that, for all normalizing methods, PCA and RFE did not clearly outperform each other. For example, RFE perform better than PCA when paired with Logarithmic Transformation for both models, even though PCA performed better in some circumstances when scaled normally. This suggests that the kind of normalization used on the data might affect how successful feature selection techniques are, depending on the context.

The effects of various transformation and normalizing methods also differed. Logarithmic Transformation paired with RFE consistently outperformed other methods, even though Standard Scaling with PCA regularly produced good results. Nevertheless, out of all the combinations, Logarithmic Transformation plus PCA tended to perform worst. This variability emphasizes how feature selection and normalization techniques should be chosen with the unique properties of the dataset and machine learning model in mind.

In summary, the study showed that some combinations, like Logarithmic Transformation with RFE, can greatly improve model performance even if it did not specify a single normalization or feature selection method that is always better. Practitioners working with high-dimensional datasets may find this insight useful as it implies that maximizing prediction accuracy requires thorough testing of various preprocessing techniques.

The most significant factors influencing the model's predictions, according to the SHAP values of the best-performing model, are YearBuilt (original building date), GrLivArea (above grade living space square feet), and Overall Quality. These qualities are in accordance with widely accepted real estate valuation theories, which hold that the age, size, and quality of construction all play a major role in determining a property's market worth. The prevalence of these variables in the SHAP study indicates that the model is capable of capturing the key elements that influence

residential property values, confirming their crucial significance in precisely forecasting property prices.

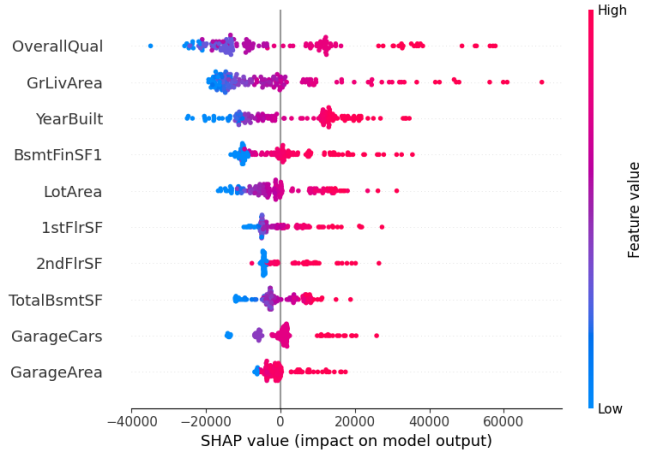

Fig. 2. SHAP Plots of Best Performing Model

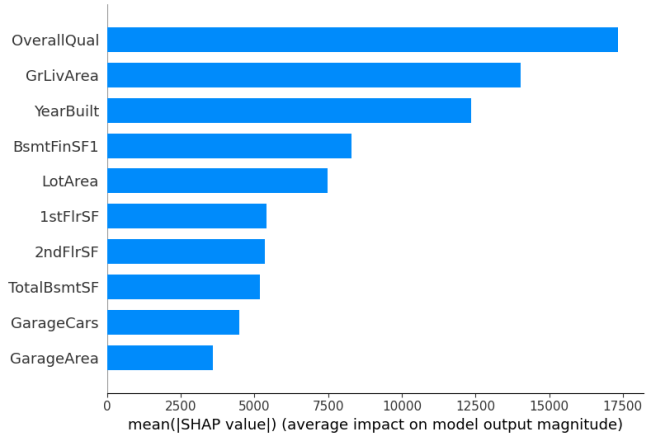

Fig. 3. SHAP Values of Best Performing Model

VI. Conclusions

This study examined how various feature selection, transformation, and data normalization methods affected the CatBoost and Random Forest models' ability to predict high-dimensional real estate datasets. The findings demonstrated how preprocessing techniques have a major impact on model performance. The most successful method was found to be the combination of Logarithmic Transformation and Recursive Feature Elimination (RFE), which produced the lowest MAPE for both models. On the other hand, the least effective result was obtained when MinMax scaling and RFE were combined. These results highlight how crucial it is to choose the right preprocessing methods based on the machine learning model and the dataset. All things considered, this study offers insightful information about how to improve real estate price forecast accuracy by using the best feature selection and data preparation techniques.

VII. Future Works

The results of this study might be expanded upon in a number of ways by future investigations. Examining other machine learning algorithms and how they interact with different data preparation methods is one possible avenue to find more reliable models for high-dimensional datasets. In order to capture more complex correlations in the data, integrating advanced feature engineering techniques, such as polynomial features and interaction terms, is another field that needs additional research. Predictive performance may also be improved by utilizing ensemble approaches, which combine the advantages of several models. Finally, the generalizability and feasibility of these feature selection and preprocessing methods may be confirmed by using them on various kinds of datasets.

References

[1] B. Park and J. K. Bae, “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,” Expert Systems with Applications, vol. 42, no. 6, pp. 2928–2934, Apr. 2015, doi: https://doi.org/10.1016/j.eswa.2014.11.040.

[2] Q. Truong, M. Nguyen, H. Dang, and B. Mei, “Housing Price Prediction via Improved Machine Learning Techniques,” Procedia Computer Science, vol. 174, pp. 433–442, 2020, doi: https://doi.org/10.1016/j.procs.2020.06.111.

[3] S. Lu, Z. Li, Z. Qin, X. Yang, and R. S. M. Goh, “A hybrid regression technique for house prices prediction,” IEEE Xplore, Dec. 01, 2017. https://ieeexplore.ieee.org/document/8289904