House Price Prediction in Paris using Machine Learning

Abstract:

The house price prediction in the dynamic and complex real estate market of Paris has several major concerns including the impact of various factors on property values, such as location, property characteristics, economic indicators, and market trends. Conventional prediction methods often fall short in addressing these complexities, lacking innovative approaches for accurate and comprehensive forecasting. Machine learning is an emerging approach that may apply algorithms and statistical models in enabling systems to learn from data and make predictions to overcome the limitations of conventional methods and achieve more accurate and reliable house price predictions in smart cities. In this research, an efficient machine learning model is proposed for accurate and reliable prediction of house prices while providing a useful and facilitating mechanism to stakeholders in the real estate market. The proposed model also provide assistance to buyers, sellers, and investors by making decision making, and achieving a higher Mean Squared Error (MSE) and Mean Absolute Error (MAE) in Linear Regression, and the lower MSE and MAE values with the Random Forest and Decision Tree algorithm, representing their efficiency in predicting house prices at real time.

Introduction:

The demand for houses has grown rapidly over the years as people's living standards improved, making it one of the essential needs alongside food, water, and more. While some vision houses as investments, most people around the world buy houses primarily as accommodations and for their livelihoods.

The House Price Index (HPI) is a widely used metric for tracking residential housing price fluctuations in various countries, like the US Federal Housing Finance Agency HPI, S&P/Case-Shiller price index, UK National Statistics HPI, UK Land Registry's HPI, UK Halifax HPI, UK Rightmove HPI, and Singapore URA HPI. The HPI employs a weighted, repeat-sales methodology, which computes average price changes by concentrating on repeated sales or refinancing of the same properties. This data is derived from the analysis of repeat mortgage transactions on single-family properties that have been purchased by Fannie Mae or Freddie Mac since January 1975. With the use of analytical tools, housing economists can apply this information to estimate differences in mortgage default rates, prepayments, plus housing affordability inside specific geographic regions [1]. The HPI is an imprecise indicator for forecasting the price of a specific house since it includes all transactions, and factors like district, age, as well as number of floors must be considered instead of relying exclusively on past repeat sales.

According to [2], housing markets are a significant factor in a nation's currency and an important predictor of the overall health of the economy. Although homeowners contribute to the country's finances by purchasing items like furniture and domestic appliances, builders and employees motivate economic growth by acquiring raw materials to satisfy the demand for housing. Furthermore, a flourishing construction sector and an adequate supply of homes demonstrate customers' ability to make significant investments and the overall economic health of a country.

With the rise of Big Data, machine learning has emerged as a crucial approach for precisely predicting house prices based on attributes, regardless of historical data. While most of the research works have demonstrated the efficiency of machine learning in this regard, few have explored the grouping of different models. S. Lu et al. directed an experiment applying a hybrid regression method, although requiring extensive parameter modification for optimal results [3-5].

In [6], the authors utilized genetic algorithms to forecast rental prices in real estate, considering geographic locations and four property features. They found that genetic algorithms outperformed multivariate regression in terms of prediction accuracy. Similarly, according to the authors in [7], a machine learning-based house price prediction system was established to assist house sellers and real estate agents. Their analysis, based on data from the Multiple Listing Service of the Metropolitan Regional Information Systems, confirmed that the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm yielded additional accurate predicting results associated to other models.

In a research, Gupta et al. [8] applied time series models to predict the U.S. real HPI, as the inclusion or exclusion of 10 or 120 additional quarterly macroeconomic series. Their findings specified that including fundamental economic variables meaningfully improved forecasting accuracy, mainly for the 10 variables inside the dynamic stochastic general equilibrium model.

Methodology:

House price prediction in smart cities faces significant challenges due to challenges regarding accuracy, the complex nature of the real estate market, and the need for data-driven solutions. Existing methods frequently struggle to deliver precise forecasts, hampering decision-making for stakeholders. Paris' real estate market is renowned for its complexity and rapid changes, making it challenging to just depend on conventional methods. In this research work, an efficient model is proposed that aims to overcome these challenges in smart cities by leveraging the potential of machine learning algorithms. The proposed model is represented in figure 1.

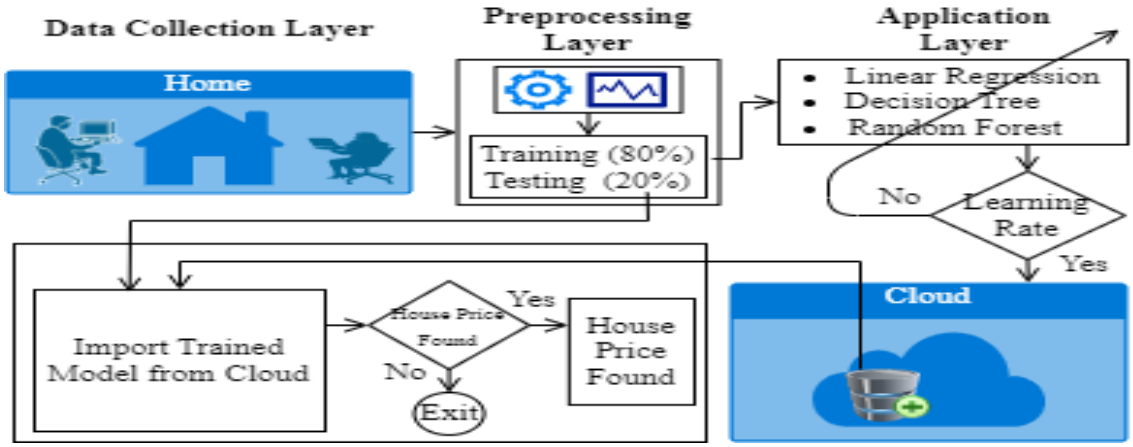

Figure 1: Proposed model for house price prediction

Figure 1 illustrates a two-phase process: training and validation. The training phase consists of three layers; data collection layer, preprocessing layer, and application layer respectively. The data collection layer is responsible for collecting house price data as input. The subsequent layer is the preprocessing layer, where the collected data is prepared via multiple preprocessing techniques (handling missing values, moving average, and normalization...) and divided into 80% for training and 20% for testing dataset. The training dataset is then passed to the application layer, which employs multiple machine learning algorithms to predict whether the learning rate has been met. If not, the application layer is retrained, if yes, the trained outcome is stored in the cloud. In the validation phase, testing data is collected from the preprocessing layer, and the trained model is imported from the cloud for prediction purposes. It checks if the house price has been successfully predicted or not. If not, the process exists, if yes, a message indicating the successful prediction is shown.

Simulation:

In this research work, the compelling need for accurate house price prediction is addressed, which plays a crucial role in the real estate industry. An efficient model is proposed which is applied on a dataset having 80% training, and 20% testing samples to simulate, explore, and evaluate effective capabilities and its potential to revolutionize house price forecasting, and to informed decision-making in the real estate market. The simulations of the proposed model are being represented. The basic libraries of the model applying machine learning algorithms, including random forest, decision tree, and linear regression. These libraries enable accurate predictions, interpretability, and insights, leveraging ensemble methods, hierarchical structures, and linear relationships.

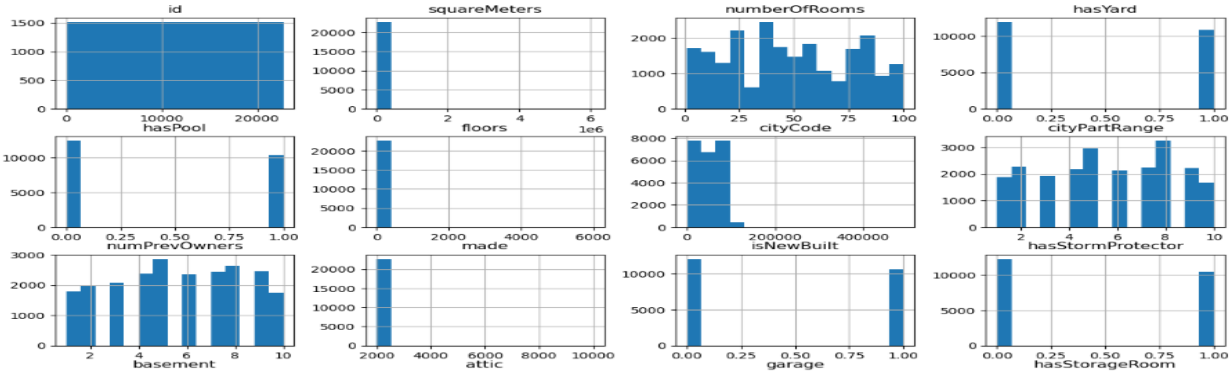

Figure 2: Histogram of the dataset

```
def run_regression_analysis(X, y):
    # Reshape y to a 1-dimensional array
    y = y.values.ravel()
    # Split the data into training and testing sets
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    # Initialize the regression models
    linear_regression = LinearRegression()
    decision_tree_regression = DecisionTreeRegressor()
    random_forest_regression = RandomForestRegressor()
    support_vector_regression = SVR()

    # Fit the regression models
    linear_regression.fit(X_train, y_train)
    decision_tree_regression.fit(X_train, y_train)
    random_forest_regression.fit(X_train, y_train)
    support_vector_regression.fit(X_train, y_train)

    # Predict on the test set
    linear_regression_predictions = linear_regression.predict(X_test)
    decision_tree_predictions = decision_tree_regression.predict(X_test)
    random_forest_predictions = random_forest_regression.predict(X_test)
    support_vector_predictions = support_vector_regression.predict(X_test)

    # Calculate evaluation metrics
    linear_regression_mse = mean_squared_error(y_test, linear_regression_predictions)
    decision_tree_mse = mean_squared_error(y_test, decision_tree_predictions)
    random_forest_mse = mean_squared_error(y_test, random_forest_predictions)
    support_vector_mse = mean_squared_error(y_test, support_vector_predictions)

    linear_regression_mae = mean_absolute_error(y_test, linear_regression_predictions)
    decision_tree_mae = mean_absolute_error(y_test, decision_tree_predictions)
    random_forest_mae = mean_absolute_error(y_test, random_forest_predictions)
```

Figure 3: Code of the proposed model using machine learning algorithms.

Figure 3 is the representation of the regression analysis code that utilizes multiple machines learning algorithms, containing linear regression, random forest, and decision tree, to predict relationships between variables. These algorithms empower accurate predictions, apprehending both linear and non-linear patterns, and offer insights into the meaning of diverse features in the regression procedure.

```
Linear Regression MSE: 4189364877822.2305
Decision Tree MSE: 62796147699.056435
Random Forest MSE: 48095198865.86323
Support Vector MSE: 8665970796451.026
Linear Regression MAE: 1756174.0459301684
Decision Tree MAE: 11822.394874615082
Random Forest MAE: 14040.63274854176
Support Vector MAE: 2542292.247895283
Linear Regression R-squared: 0.5152572858408329
Decision Tree R-squared: 0.9927339880955406
Random Forest R-squared: 0.9944350043703086
Support Vector R-squared: -0.0027214929245902297
```

Figure 4: Proposed model performance in terms of RMSE

Figure 4 represents the performance evaluation of the proposed model using several algorithms such as linear regression, random forest, and decision tree. It also compares the relevant metrics to assess the efficiency of the model across these algorithms.

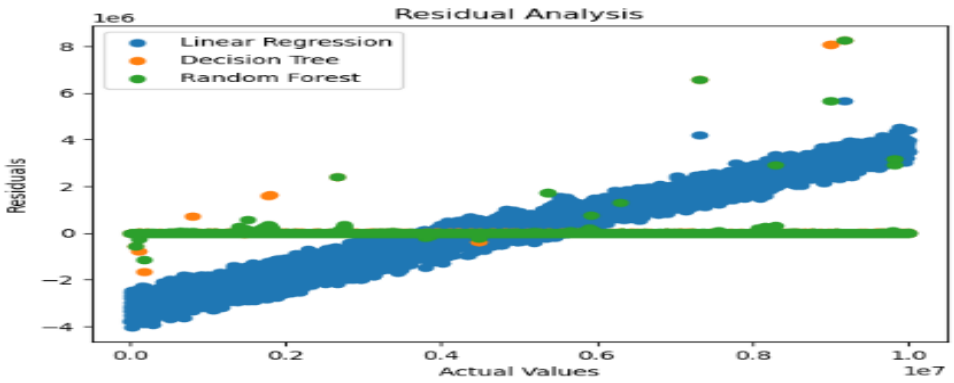

Figure 5: Proposed model performance comparison of machine learning algorithms

Figure 5 is the representation of the regression analysis graph's varying performance of different algorithms, with linear regression yielding the highest MSE and MAE, while decision tree and random forest algorithms exhibit lower MSE and MAE values. It clearly describes based on the given metrics that the decision tree algorithm shows the better performance in terms of lower MSE and MAE values as compared to linear regression and random forest algorithms.

Conclusion:

The major challenge in this research work is the dynamic nature of real estate markets in house price prediction, that may be challenging for conventional methods. Machine learning is an emerging approach that excels in analyzing large data, capturing real-time patterns, and enabling more accurate predictions of house prices. An efficient model is proposed for conducting the simulations using various machine learning algorithms, containing linear regression, random forest, and decision tree, which can gain valuable insights in performance of house price prediction. The simulation results reveal that Linear Regression shows higher MSE and MAE, highlighting its limitations in apprehending the complexities of housing datasets. Whereas the Random Forest and Decision Tree algorithms demonstrate superior performance, with lower MSE and MAE values, representing their efficiency in predicting house prices in real time.

References:

1. Index, H.P., 2015. Federal Housing Finance Agency. Retrieved from.
2. Temur, A.S., Akgün, M. and Temur, G., 2019. Predicting housing sales in Turkey using ARIMA, LSTM and hybrid models.
3. Fan, C., Cui, Z. and Zhong, X., 2018, February. House prices prediction with machine learning algorithms. In Proceedings of the 2018 10th International Conference on Machine Learning and Computing (pp. 6-10).
4. Phan, T.D., 2018, December. Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In 2018 International conference on machine learning and data engineering (iCMLDE) (pp. 35-42). IEEE.
5. Mu, J., Wu, F. and Zhang, A., 2014, August. Housing value forecasting based on machine learning methods. In Abstract and Applied Analysis (Vol. 2014). Hindawi.
6. Del Giudice, V., De Paola, P. and Forte, F., 2017. Using genetic algorithms for real estate appraisals. Buildings, 7(2), p.31.

7. Park, B. and 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), pp.2928-2934.
8. Gupta, R., Kabundi, A. and Miller, S.M., 2011. Forecasting the US real house price index: Structural and non-structural models with and without fundamentals. Economic Modelling, 28(4), pp.2013-2021.