Machine Learning-Based House Price Prediction via Whatsapp and Website Chatbot

Marinda Tirivashe.[1], Madyembwa Monica.[2], Shambira Leornard.[3]

Abstract

House property valuation in Masvingo rely on manual appraisals and comparative market analysis, which are often slow, inconsistent, and unable to adapt to rapid market changes. Recent advances in machine learning and digital valuation platforms have demonstrated improved estimation accuracy, though adoption in Masvingo remains limited. The aim of this study was to develop a machine learning-based model to accurately predict house prices in Masvingo Urban via WhatsApp and Website Chatbot. Primary and secondary data was collected from local real estate agencies using questionnaires, interviews, and document analysis. Data preprocessing and synthetic data generation was employed to address data sparsity. After augmenting the dataset, recalibration was performed to ensure the synthetic data matched the statistical characteristics of the original dataset and to reduce overfitting. Random Forest Regression, which outperformed other algorithms, had a Mean Absolute Error of 464.67 and R^2 of 0.8298. The algorithm predicted a high-density Masvingo property (500 m² stand, 200 m² building, seven rooms, standard amenities and standard materials) at 30,784.55. This study is essential for real estate professionals, buyers, and policymakers, providing standardised, fast, and data-driven property valuation. The study recommends that machine learning models offer a practical ICT solution for sustainable real estate valuation in Masvingo Urban.

Keywords: Property valuation; Data preprocessing; Synthetic data; Random Forest; Sustainable real estate

1. Introduction

The rapid urbanisation and economic development in Masvingo has significantly impacted the real estate market. This has increased the need for accurate and timely property valuations in the real estate industry. Currently, property valuation in Masvingo relies heavily on manual processes performed by real estate professionals. These traditional methods, although grounded in expertise are often slow and prone to human error. There are also inadequate to address the fast-paced changes of the market (Chigbu, 2021). Furthermore, these current methods struggle to adapt to economic volatility and urbanisation demands. This has resulted to inconsistencies in property valuations that can affect investment decisions and market confidence (Paradza et al., 2021). The scarcity of comprehensive and timely data on property transactions exacerbates these challenges. This highlights the limitations of conventional valuation methods and prompting the exploration of alternative solutions.

[1] Student (Bsc), Great Zimbabwe University, Zimbabwe, P. Bag 1235 Masvingo, School of Natural Sciences, Department of Maths and Computer Science, trumanmarinda@gmail.com

[2] Lecturer (Msc), Great Zimbabwe University, Zimbabwe, P. Bag 1235 Masvingo, School of Natural Sciences, Department of Maths and Computer Science, mmadyembwa@gzu.ac.zw

Globally, the real estate industry has embraced technological advancements in the field of Artificial Intelligence. This has enhanced the property valuation processes with in the real estate sector. The artificial intelligence-based models' have offered dynamic, accurate approaches that can quickly adapt to market changes. Machine learning algorithms have demonstrated a high degree of accuracy in analysing vast, complex datasets to predict future property values (Winky et al., 2021). Furthermore, the adoption of user-friendly and interactive chatbots and websites has significantly expanded the customer base (Lee, 2019). The emergence of machine learning techniques has transformed data-driven decision-making across various sectors. This has enabled the analysis of large, complex datasets with exceptional accuracy and efficiency (Madyembwa et al., 2020).

In developing countries such as Zimbabwe, the adoption of machine learning models in real estate valuation remains limited. This is primarily due to infrastructural, economic, and human resource constraints. While a few privately owned real estate, agencies have adopted automated property valuation platforms (Fang,.2023). The majority of the operators continue to rely on conventional methods to determine house prices. This study sought to address these gaps by developing a machine learning-based house price prediction model for Masvingo Urban. This promised to transform residential house prices valuation, advance economic sustainability, and support evidence-based decision-making in Zimbabwe's evolving real estate market.

2. Literature Review

This literature review examined the evolution of valuation methodologies across global, regional, and local areas. This review mainly focused specifically on Sub-Saharan Africa and Zimbabwe, where data scarcity and infrastructural challenges persist. Although challenges persist, there is increasing scope for incorporating technological solutions to enhance property valuation practices. The review also explored the role of digital platforms, such as websites and chatbots. These aspects proved to enhance property valuation through data driven interaction application systems. A comparative analysis of models was done and the research gap was identified.

2.1 House Price Prediction Models used in Developed Countries

In developed countries, real estate institutions have increasingly relied on advanced machine learning models to improve the accuracy, reliability, and efficiency. These models have extensive datasets that consist of property attributes, historical transactions, neighbourhood characteristics, and socio-economic indicator (Makoni et al., 2019). These attributes capture the complex and non-linear relationships that traditional econometric approaches often fail to address. Therefore, ensemble learning techniques have proved to improve accuracy, robustness, and generalisation in house price prediction. The ensemble learning techniques are widely applied for predictions by mitigating overfitting and accommodating high-dimensional data. Furthermore, deep learning approaches, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have been employed. They have been employed to integrate visual property features and temporal trends, thus further enhancing predictive precision (Zhang, 2023).

The K-Nearest Neighbour (KNN) algorithm has been implemented in value of a property based on the prices of similar nearby properties (Huang et al., 2017). This algorithm, utilises parameters such as location, house size, and amenities, on the closest neighbouring data points (Reddy et al., 2022). This proximity-based approach performs particularly well for local optimisation of predicting property prices based on the most similar nearby houses rather than using global averages. However, its performance tends to degrade with high-dimensional data due to the curse of dimensionality. This leads to potential inaccuracies when dealing with large and diverse multidimensional datasets (Kouiroukidis et al., 2011). Despite its ease of deployment, KNN's have proved to have limited performance in markets with significant variability or sparse data. This can be a critical drawback in its application across broader or more fluctuating real estate markets. In a study by Zhu (2023), they utilised KNN to estimate deviations in linear regression outcomes for house price predictions, enhancing accuracy by adapting to deviations in the Boston House Price data set. In his research Zhu (2023) found out that deviation regression (DR) algorithm improves house price prediction accuracy by estimating deviations by incorporating k-nearest neighbour (KNN) algorithm. In addition, another application by Samruddhi and Kumar (2020) employed KNN in a supervised learning model to predict used car prices, indicating the adaptability of KNN to various types of prediction models and its potential for high accuracy in diverse datasets. The results of this showed that the proposed KNN-based model accurately predicts used car prices with around 85% accuracy using data from the Kaggle website. While KNN is approachable and easy to deploy, its reliance on local data density can lead to inaccuracies in diverse markets.

In Zhang's 2023 study titled "*Housing Price Prediction Using Machine Learning Algorithm*," published in the Journal of World Economy, the author utilised a combination of machine learning models including K-Nearest Neighbours (KNN), Ridge Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) to predict house prices in Ames, Iowa, USA. This ensemble approach effectively managed the various characteristics of the dataset, allowing for a nuanced understanding of feature interactions. KNN focused on local similarity, whereas Ridge Regression addressed multicollinearity among predictors, Random Forest excelled in capturing non-linear relationships without extensive tuning, and XGBoost optimised for predictive accuracy and computational efficiency. This combined methodology significantly outperformed traditional linear models, highlighting the advantage of integrating diverse machine learning techniques to handle complex real estate datasets effectively. The primary benefit of this strategy lies in its robustness and ability to generalise across different data segments, thereby increasing the accuracy of predictions. However, the requirement for extensive data pre-processing, feature engineering, and hyper parameter tuning presents a substantial challenge, necessitating advanced expertise in data science. The study proved that there is need for further research to validate these findings using different datasets and exploring the capabilities of other machine learning algorithms in housing price prediction. Despite these complexities, the application of these models has set a benchmark for predictive accuracy in the real estate sector. Further studies, such as those by Han (2023) and Thamarai and Malarvizhi (2020), have continued to refine these techniques, exploring additional dimensions such as advanced regression and decision trees to broaden the applicability and reliability of predictive models in varying market conditions. The study by Han (2023) explored into advanced regression techniques, leveraging a sophisticated model that incorporates LASSO Regression, Elastic Net Regression, Gradient Boosting Regression, XGBoost, LightGBM, and a Stacked Model. This approach was rigorously evaluated using five-fold cross-validation within the Ames, Iowa dataset, which is known for its detailed attribute data on residential properties. Han's methodology focused on minimizing the root-mean-square logarithmic error, ultimately identifying the Gradient Boosting Regression Model as the most effective due to its lower error rates and robust performance across different subsets of data. This study underscores the effectiveness of combining multiple advanced regression techniques to enhance predictive accuracy in real estate pricing, offering substantial benefits to buyers, investors, and policymakers by providing more reliable investment insights (Han, 2023).

The research by Thamarai and Malarvizhi (2020) explored machine learning's application in predicting house prices for a small town in West Godavari district, Andhra Pradesh. Their work utilised decision tree, classification and regression, along with multiple linear regression, implemented through the Scikit-Learn Machine Learning Tool. The selection of these particular models was due to their suitability in handling the specific types of data available in the region, which included features such as the number of bedrooms, age of the house, and proximity to essential amenities like schools and shopping malls. This study has shown that the practical application of machine learning in a less typical, more localised real estate market, offering insights into how data attributes specific to a locale can significantly influence the effectiveness of predictive models. The success of their approach in a smaller market context emphasises the scalability and adaptability of machine learning models across different geographical and market conditions (Thamarai & Malarvizhi, 2020). This study demonstrated the dynamic capabilities of machine learning in real estate analytics, demonstrating that with the right model and methodology, significant advancements can be made in predicting house prices accurately across diverse environments.

The Deep Learning and ARIMA hybrid model has been implemented for house price prediction by Wang et al. (2019) to address nonlinear relationships and large-scale data challenges in the real estate market, demonstrating superior performance over traditional methods such Support Vector Regression (SVR) in complex datasets. The model combines deep learning for intricate feature extraction with ARIMA for time series forecasting of house prices, creating a robust framework for understanding and predicting real estate market trends. The main advantage of this hybrid model lies in its ability to leverage deep learning's capacity to handle and analyse large volumes of data, thus extracting deeper insights from complex patterns that might not be discernible through traditional statistical methods. Wang et al. (2019) highlighted the model's effectiveness in predicting individual house prices and general price trends in the market, noting that it aligns closely with real market dynamics, thus validating the practical application of this hybrid approach. However, the setup and maintenance of such a model are complex and require substantial computational resources, which can be a significant barrier, especially in less developed countries with limited technical infrastructure. Despite these challenges, the implementation of the Deep Learning and ARIMA model offers a promising direction for predictive analytics in real estate, particularly in highly volatile markets where traditional models might fail to capture emerging trends quickly. Further research, such as the studies conducted by Pratheeth and Vishnu (2021), could expand on the utilisation of ARIMA in conjunction with more advanced neural networks such as LSTM for real-time applications, offering a path forward for integrating these sophisticated models into everyday real estate analysis tools.

In a study by Montero-Lorenzo et al. (2018), they explored the effectiveness of parametric versus semi-parametric spatial hedonic models in house price prediction, providing a comprehensive examination of how spatial autocorrelation and heterogeneity can be integrated into real estate market analysis. These models leveraged both parametric methods, which assume a specific form for the relationship between house prices and their predictors, and semi-parametric methods that do not impose such rigid structures, allowing for a more flexible interaction between variables. The use of these advanced spatial models is particularly effective in urban settings, such as demonstrated in Madrid, Spain, where they significantly enhance predictive accuracy by accounting for spatial dependencies that traditional models might overlook. The results obtained suggest that the nonlinear models accounting for spatial heterogeneity and flexible nonlinear relationships between some of the individual or areal characteristics of the houses and their prices are the best strategies for house price prediction (Montero-Lorenzo et al., 2018). Further enhancing the understanding of spatial econometrics in house price modelling, the study by Liu (2013) on incorporating spatial and temporal dependence among housing transactions illustrates how integrating these dependencies can significantly improve the prediction power of hedonic models. Liu's approach, using spatiotemporal autoregressive models, controlled for time variation of both the attribute prices and the spatial and temporal dependence parameters, highlighting the increased accuracy in house price predictions that can be achieved when temporal dynamics are accounted for alongside spatial factors (Liu, 2013). These studies collectively emphasize the sophistication and necessity of incorporating spatial and temporal analytics into house price prediction models, providing a richer, more accurate toolkit for real estate analysis and decision-making.

In a 2023 study by Cui, the integration of Big Data-Driven Models for house price prediction was explored, utilizing extensive datasets that encompass a wide array of property characteristics such as number of bedrooms, bathrooms, amenities, and more to generate highly accurate predictive models. This research, conducted within King County, Washington, leveraged the breadth of big data to refine estimations of housing values, demonstrating the model's capacity to handle and analyse large volumes of diverse data effectively. The primary advantage of these models is their ability to process a vast amount of information, enabling the incorporation of numerous variables which significantly enhances prediction accuracy. However, these models require access to large, detailed datasets and robust computational resources, potentially limiting their application to well-documented and technologically equipped markets. Despite these challenges, Cui's implementation of Big Data-Driven Models stands out for their precision and are particularly beneficial for consumers by providing a more accurate reflection of housing values. This research represents a significant advancement in real estate analytics, offering potential solutions that address the gaps in simpler, less data-intensive models (Cui, 2023).

Therefore, by expanding upon the work of Cui, further research by Patel et al. (2023) demonstrated the utilisation of machine learning techniques combined with big data analytics to predict housing prices in a suburban area of New Jersey. This study specifically focused on integrating socio-economic data with physical house attributes to enhance predictive accuracy, showcasing how external macroeconomic factors can be effectively incorporated into prediction models (Patel et al., 2023).

Another study by Zhang et al. (2019) employed a similar Big Data approach but focused on applying deep learning algorithms to capture complex nonlinear relationships in house pricing in the San Francisco Bay area. Their research highlights the potential of deep learning methods to provide more nuanced insights into the factors driving house prices in high-demand markets, underscoring the growing importance of advanced analytics in real estate (Zhang et al., 2019). Together, these studies illustrated the dynamic capabilities of Big Data-Driven Models in real estate analytics, demonstrating that with the right model and methodology, significant advancements can be made in predicting house prices accurately across diverse environments.

2.2 House Price Evaluation Models in Africa

Spatial Bayesian Vector Autoregressive models (SBVAR) was implemented to forecast house prices in six metropolitan areas of South Africa, utilising monthly data over a significant period from 1993 to 2005 by Gupta and Das (2008). This research demonstrated the effectiveness of incorporating spatial contiguity and Bayesian priors, which allowed for a nuanced prediction of house price trends, particularly for large middle-segment houses. The SBVAR models were shown to outperform traditional Vector Autoregressive (VAR) and Bayesian VAR models based on the Minnesota prior, especially when forecasting one to six months ahead. The primary advantage of SBVAR models lies in their ability to account for spatial dependencies, which are crucial in capturing the dynamics of regional housing markets. However, the complexity of these models and the computational intensity required for their implementation pose significant challenges. Despite these drawbacks, the use of SBVAR models in this context highlights their potential to enhance forecasting accuracy, supporting their utility in economic and real estate planning (Gupta & Das, 2008).

Adding to the rationale for employing advanced econometric models in the SADC region, Das, Gupta, and Kabundi (2010) conducted a study utilizing dynamic factor models to predict house price inflation across five metropolitan areas in South Africa. This approach utilized a large cross-section of macroeconomic time series, achieving statistically significant improvements over traditional vector autoregressive models. The dynamic factor models provided a robust framework that integrated broader economic indicators into the forecasting process, enhancing the model's predictive power and relevance for policy-making (Das, Gupta, & Kabundi, 2010). In addition to that, a study by Das, Gupta, and Kabundi in 2008 developed large-scale Bayesian Vector Autoregressive (BVAR) models to forecast real house price growth rates in South Africa. Their models, estimated under alternative hyper parameter values, not only provided accurate forecasts but also demonstrated the ability to predict downturns in the housing market, underscoring the predictive strength and strategic importance of BVAR models in economic forecasting (Das, Gupta, & Kabundi, 2008).

In 2019, Murekachiro, Mokoteli, and Vadapalli conducted a study which explored the use of Generalised Additive Models (GAMs) for predicting stock market trends in Sub-Saharan Africa, providing insights into the applicability of this methodology for real estate market prediction. While initially applied to financial markets, the principles detailed in their study offer valuable implications for real estate, particularly in their capacity to model nonlinear relationships without a predefined parametric form. This flexibility allows GAMs to effectively capture the intricate dynamics of housing prices influenced by both market and non-market factors. The researchers demonstrated that GAMs could outperform deep neural models such as LSTM and GRU in predicting daily stock market prices, suggesting similar potential for enhancing real estate price predictions due to their high accuracy and adaptability to different types of data.

In recent research by Fasanya & Awodimila (2020), titled "*Are commodity prices good predictors of inflation? The African perspective*", significant advancements were highlighted in applying machine learning techniques to predict house prices across African markets. Their research utilised a variety of algorithms such as regression trees, support vector machines, and neural networks to handle the large datasets and complex variable interactions typical in real estate data. The study demonstrated these machine learning models' capabilities to process and analyse extensive data to predict house prices with a high degree of accuracy, offering adaptability to the diverse and dynamic real estate markets found in Africa. The primary strength of these models lies in their high accuracy and flexibility in modelling complex and non-linear relationships that traditional econometric models might fail to capture. However, the successful implementation of machine learning models in house price prediction requires significant data pre-processing to clean and normalize data, as well as substantial technical expertise to select appropriate algorithms and tune their parameters effectively. This can pose a substantial barrier in emerging markets where such technical resources or data quality might not be readily available. Furthering this exploration, a study by Juneja (2023) explored the effectiveness of machine learning algorithms in predicting house prices, emphasizing the impact of data characteristics such as location, duration of house ownership, and house dimensions on prediction accuracy. This study illustrates how various machine learning models can be optimised to improve prediction reliability, crucial for developing tailored marketing strategies and enhancing consumer engagement in the real estate sector (Juneja, 2023). These insights collectively underscore the transformative potential of machine learning in enhancing house price predictions in Africa, offering a strategic advantage in a highly competitive market. The continued development and refinement of these models are essential for adapting to the evolving real estate landscape and achieving sustained economic growth in the region.

In a 2020 study conducted by Pillay, Autoregressive Integrated Moving Average (ARIMA) models were used to forecast economic variables influencing house prices, focusing on the Johannesburg Stock Exchange in South Africa. Pillay demonstrated the effectiveness of ARIMA models in providing straightforward and reliable time series forecasts, particularly valuable for predicting changes in house prices. ARIMA models, enhanced with seasonal components in this study, are well-suited for economic forecasting, where understanding trends is essential. However, their primary limitation lies in handling linear relationships only, necessitating extensions for more complex market dynamics. Despite these challenges, Pillay's research underscores the potential of ARIMA models as a reliable tool for the real estate market, provided adjustments are made for specific market characteristics (Pillay, 2020). Building upon ARIMA models, a 2019 study by Alkali assessed ARIMA against ARIMAX models, which incorporate exogenous variables, in Abuja, Nigeria. This comparative analysis highlighted ARIMAX models' enhanced ability to capture the influence of macroeconomic factors such as interest rates and GDP on house prices. Alkali's research emphasizes the improved forecasting performance of ARIMAX over standard ARIMA models, especially in capturing the complex dynamics of the residential real estate market in Abuja. This study suggests that integrating additional economic indicators into ARIMA models offers a more sophisticated approach to understanding and predicting market trends in African cities, providing a strategic advantage for stakeholders in the volatile real estate sector (Alkali, 2019).

2.3 House Price Prediction Models in Zimbabwe

Research on house price prediction in Zimbabwe has gained traction with the utilization of various statistical and machine learning models aimed at understanding and forecasting the dynamics of the real estate market within this region. ARIMA and ARCH/GARCH models were applied to forecast stock prices on the Zimbabwe Stock Exchange (Mutendadzamera & Mutasa, 2014). Although focused on stocks, the methodologies are relevant to house prices due to similar speculative market dynamics. These models, particularly ARCH/GARCH, are advantageous for capturing time-varying volatilities, an essential feature in Zimbabwe's fluctuating economic environment. The study confirmed that GARCH models, with their robust handling of volatility, provide superior forecasts compared to ARIMA models, suggesting their potential applicability in predicting house prices. However, their complexity in terms of setup and interpretation might limit their usability for non-specialists. Despite its widespread application, the Hedonic Pricing Model in Zimbabwe faces several challenges, primarily due to the volatile economic environment that can dramatically influence property value factors. Furthermore, the model requires comprehensive and reliable data to be effective, a frequent hurdle in Zimbabwe's sometimes irregular real estate market where official records and data transparency may lack. These limitations suggest that while the Hedonic Pricing Model provides valuable insights into property valuation, its effectiveness is contingent on the quality of the underlying data and the stability of the economic conditions (Makoni, 2012). This example underscores the need for continuous improvement in data collection and economic stabilization to enhance the reliability of traditional house pricing models in Zimbabwe. Such advancements would not only improve the accuracy of property valuations but also contribute to a more transparent and robust real estate market.

Another traditional method utilised in Zimbabwe for estimating property values is Comparative Market Analysis (CMA). This approach involves comparing the prices of properties that are similar and in close proximity to determine the value of a specific property. A practical implementation of this method can be seen in the real estate operations around major urban centers like Harare, where real estate agents frequently employ CMA to advise both sellers on setting listing prices and buyers on making competitive offers. Chikolwa (2015) explored the application of CMA in Zimbabwe's residential markets, particularly focusing on the suburbs of Harare. The study demonstrated that while CMA is highly reliant on the availability of current market data, it provides a realistic and timely valuation that reflects the current market conditions. However, the challenge in Zimbabwe is often the lack of access to up-to-date and accurate property transaction data, which can skew valuations and lead to either overpricing or under-pricing of properties. This limitation is compounded by Zimbabwe's economic fluctuations, which can quickly change property values and make previous comparisons obsolete.

The Cost Approach is another traditional method used in Zimbabwe for property valuation, particularly effective for new constructions or properties where few comparable exist. This method estimates a property's value by determining how much it would cost to construct a comparable property from scratch, considering current material, labour costs, and land value, minus any depreciation. A relevant study by Moyo (2018) analysed the application of the Cost Approach in the valuation of newly built properties in Bulawayo, demonstrating its utility in settings where detailed cost data is available and current. Moyo's research revealed that the Cost Approach provided an essential valuation tool in scenarios where other methods such as CMA or Hedonic might not be applicable due to a lack of comparable market data. However, the method's reliance on accurate and up-to-date construction cost data, along with proper depreciation models, poses significant challenges. In Zimbabwe, where economic instability can cause rapid fluctuations in material and labour costs, keeping this data current is particularly challenging. Additionally, the method does not account for market conditions or buyer preferences, which can sometimes result in valuations that do not reflect the market's actual state.

Weightless Neural Network (WNN) was utilized in a research by Mpofu (2006) to forecast stock prices for Kingdom Financial Holdings in Zimbabwe. The study demonstrated the neural network’s capability to adapt to the non-linear patterns typical of financial markets, which are comparable to those found in real estate pricing. Mpofu's work offers a promising direction for adopting advanced machine learning techniques in Zimbabwe's housing market, potentially providing more dynamic and responsive modelling solutions. The application of neural networks could significantly improve predictive accuracy by learning from complex datasets and capturing intricate relationships between house features and market prices (Mpofu, 2006). This study together with study by Mutendadzamera & Mutasa (2014) highlight the diversity of modelling techniques that can be adapted to Zimbabwe’s real estate market, each offering unique insights and tools to tackle the challenges posed by its economic landscape. Therefore, both traditional econometric models and innovative machine learning approaches hold considerable promise for enhancing the accuracy and reliability of house price predictions in Zimbabwe.

2.4 Research on Chatbots and Websites

Chatbots, particularly those that operate over popular messaging platforms like WhatsApp, heavily rely on machine learning algorithms to improve interaction quality. The use of Natural Language Processing (NLP) technologies enables chatbots to deliver more accurate and human-like responses. For instance, Dhanashri Hase et al. (2023) developed a WhatsApp Chat Analyser that uses Python libraries and Natural Language Processing concepts to analyse chat data and provide insights through visualisations, demonstrating the practical applications of machine learning in processing conversational data from WhatsApp.

Csaky (2019) introduced the use of advanced neural network architectures such as transformers to improve the conversational abilities of chatbots. This approach significantly enhanced the chatbot's capability to manage complex dialogues, making interactions appear more human-like. Furthermore, Prakasam et al. (2023) introduced an AI-powered healthcare chatbot on WhatsApp designed to facilitate medical appointment scheduling and consultations. This chatbot bridges the gap between patients and healthcare providers, offering a streamlined communication process. In addition to chatbots, the use of websites in the real estate domain has also seen significant advancements. Websites provide a robust platform for users to interact with property valuation models and access real-time data.

The Z-estimate model which is an **A**utomated Valuation Model (AVM**)** developed by Zillow has been a pioneering effort in online property valuation, leveraging a combination of machine learning and statistical models to predict home values. The algorithm incorporates a wide array of data points, including historical transaction data, tax assessments, and user-submitted information, to provide real-time property value estimates (Bourassa et al., 2010). This model's strength lies in its ability to continuously refine predictions with new data inputs, ensuring up-to-date valuations. The proprietary nature of the algorithm also means that its internal workings are not fully transparent, which can lead to scepticism among users regarding the reliability of the estimates (Bourassa et al., 2010). However, its accuracy can be affected by data quality and availability, particularly in less-documented or volatile markets.

Elsewhere, Redfin (2024), employed a data-driven valuation model that integrates big data and machine learning to estimate property values accurately. The model utilizes a comprehensive dataset that includes property details, market trends, and user interactions, updated frequently to reflect current market conditions (Rossi & Biddle, 2013). This model offered real-time insights and high accuracy in property valuations. The frequent updates and large dataset enhance the model's responsiveness to market changes. The reliance on user-generated data also introduces potential biases that need to be managed carefully (Rossi & Biddle, 2013). However, similar to Zillow, the accuracy of Redfin's estimates can be impacted by data quality and regional differences in market dynamics.

Realtor.com which is an online real estate marketplace, has integrated predictive analytics into its platform to enhance property valuation accuracy and user experience. Therefore, by analysing user behaviour, search patterns, and extensive market data, the platform's machine learning algorithms can predict property values and trends with notable precision (Ling & Archer, 2018). This integration not only improves the accuracy of valuations but also provides personalized recommendations to users, leveraging large datasets to tailor insights to individual preferences. However, the reliance on user behaviour data introduces privacy concerns, and the effectiveness of predictive analytics can vary based on the completeness and accuracy of the input data.

Trulia's use of interactive mapping and valuation tools exemplifies the integration of spatial analysis with machine learning for real estate valuations. The platform's algorithms analyse spatial data, property characteristics, and market conditions to estimate property values and display valuation trends through interactive maps (Goodman & Thibodeau, 2008). This approach enhances user engagement and decision-making by providing a visual and detailed understanding of market dynamics. However, the effectiveness of this model is contingent on the quality and granularity of the spatial data available. In areas with less detailed geographic data, the accuracy of valuations may be compromised.

The reviewed studies highlight significant advancements in using websites for property valuation through machine learning and predictive analytics. The success of platforms such as Zillow, Redfin, Realtor.com, and Trulia underscores the potential benefits of integrating a machine learning model with a website interface for house price prediction system. Therefore, by adopting similar methodologies, this promises to enhance the accuracy, accessibility, and user engagement to the system. The combination of a WhatsApp chatbot and a website interface further provides a comprehensive solution that caters to different user preferences, ensuring a seamless and efficient user experience.

2.5 Research Gap

Despite significant global advancements in machine learning–based house price prediction, a critical gap remains in the applicability and practical deployment of these models in data-scarce and economically volatile environments. The existing studies are predominantly situated in data-rich and economically stable markets. Furthermore, while ensemble learning models such as Random Forest have demonstrated superior extrapolative performance globally, there is limited empirical evidence on their reliability when trained using synthetically augmented datasets calibrated to reflect real-world market conditions. In Zimbabwe, prior research and practice continue to rely heavily on manual valuation methods. These are Comparative Market Analysis and the Cost Approach, with minimal integration of automated, machine learning-driven valuation systems. Furthermore, although digital platforms and chatbots are widely used in other sectors, their integration with machine learning-based property valuation remains largely unexplored. Consequently, there is a lack of end-to-end, deployable valuation frameworks that combine data preprocessing, predictive modelling, and accessible digital interfaces tailored to the Zimbabwean real estate industry. This limits the practical adoption of advanced valuation emerging advanced technological models. Therefore, by integrating these studies, the literature review provides a comprehensive understanding of how websites can be utilized in the real estate domain. This enhances property valuation through advanced machine learning models. This literature review supports the development and implementation of the proposed house price prediction model in Masvingo, utilizing both a WhatsApp chatbot and a website interface.

3. Methodology

The House Price Prediction System was developed to address the heavy reliance on manual appraisals and human expert judgment in Masvingo Urban’s real estate market. Primary and secondary data sources were implemented to ensure comprehensive model development. The primary data was collected through interviews, focus group discussions, and structured questionnaires, administered to local real estate agents and property valuers. Secondary data was obtained from online property listings, government land and zoning records, and socioeconomic indicators relevant to housing demand and valuation. The study employed a sample size of 150, determined to provide sufficient representation of property characteristics within Masvingo Urban. Purposive sampling approach was used to select participants and data sources, ensuring diversity across different property types and locations. Data was collected for attributes such as location, size, number of bedrooms and bathrooms, property condition, and sale prices. The initial dataset comprised approximately 1,200 entries. Although the dataset was valuable, it was of limited size and contained anomalies such as missing values and out-of-range entries.

Therefore, to enhance data quality and efficiency of the machine learning model performance data preprocessing techniques were employed. Missing values were imputed using statistical methods, outliers were detected and removed, and the dataset was normalised. Furthermore, error metrics such as Mean Squared Error (MSE) were employed to identify and correct data inconsistencies. Due to the limited size of the original dataset, a synthetic dataset was generated, expanding the dataset to 40,000 entries to provide sufficient variability for model training. However, reliance on synthetic data led to overfitting of the machine learning model. This was evidenced by the high R^2 score of 0.8298 in preliminary evaluations, indicating potential limitations when applied to entirely new, real-world data. This challenge was addressed through careful calibration of synthetic samples and regularization strategies to improve model generalization. . Therefore, to address this. the synthetic dataset was recalibrated to match the statistical distribution of the real data. This reduced the artificial variance introduced during augmentation. The high R^2 score was reduced to 0.8298.

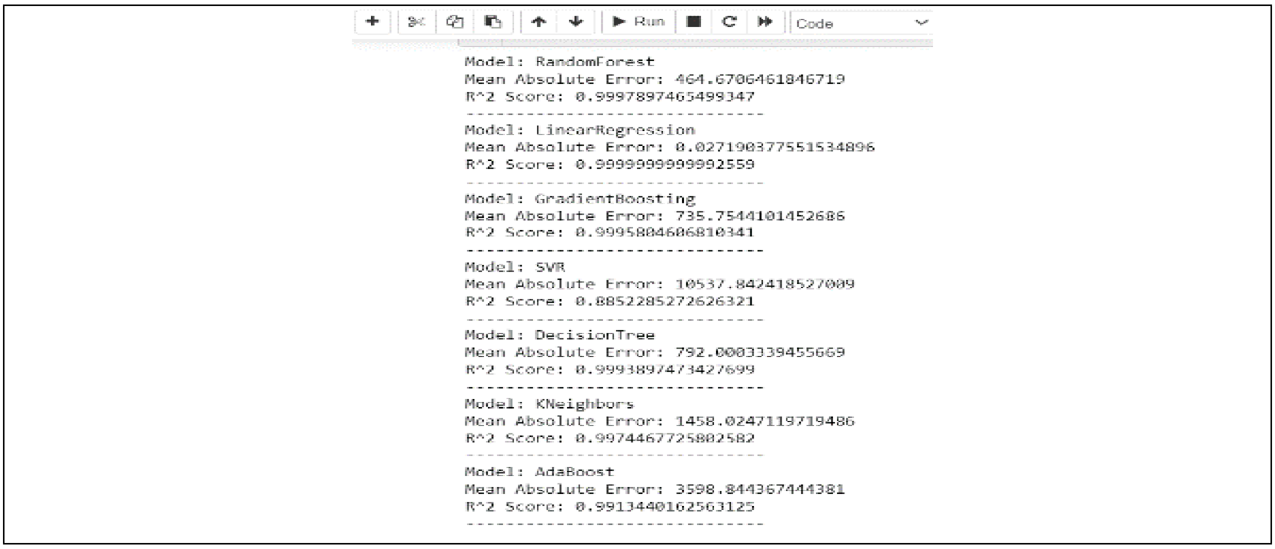
A comparative analysis of machine learning algorithms was conducted, using linear regression, gradient boosting, and Random Forest Regression. The Random Forest algorithm The model achieved a Mean Absolute Error (MAE) of 464.67 and consistently outperformed other tested models in preliminary tests. The trained model was integrated into an interactive WhatsApp Chatbot system, enabling real-time house price predictions within Masvingo Urban. While the model demonstrates promising predictive performance, further training and validation using larger, real-world datasets are necessary to assess its commercial viability and ensure reliability for broader application.

4. Results

This section presented the experimental outcomes of the proposed house price prediction system and evaluated the effectiveness of the implemented machine learning models and user interfaces. The results focused on assessing Random Forest Regression prediction accuracy. This section demonstrated the practical application of the system model through both WhatsApp and website chatbot platforms.

4.1 Regression Model Performance Analysis

Figure 1 illustrated the performance of the seven regression algorithms that were quantitatively evaluated using Mean Absolute Error (MAE) and the coefficient of determination (R^2). MAE measured the average magnitude of prediction errors, which implied that smaller values highlights better algorithm performance. R^2 revealed how well the model's predictions match the actual data. Therefore, high R^2 implied that the model fits the data better. In this comparative analysis, Linear Regression and Random Forest initially showed the highest model performance, with MAE values of 0.027 and 464.67; and R^2 values of 0.9999 and 0.9998, respectively. The extremely high R^2 suggested potential overfitting. Random Forest was considered the best despite a slightly higher MAE. The Random Forest was considered the best as it captures complex, non-linear patterns and is less sensitive to outliers. The original dataset collected from four local real estate agencies contained approximately 500 property records, which was insufficient for model training. Therefore, to improve generalizability, data augmentation techniques were applied to generate synthetic records, which expanded the dataset to 10,000 entries. The synthetic dataset was subsequently recalibrated to match the statistical distribution of the original data, by mitigating biases and reducing artificial variance introduced during augmentation. After these adjustments, the Random Forest model achieved MAE of 464.67 and an R^2 of 0.8298, demonstrating improved stability and predictive reliability.




```
Model: RandomForest
Mean Absolute Error: 464.6706461846719
R^2 Score: 0.9997897465499347
------------------------------
Model: LinearRegression
Mean Absolute Error: 0.027190377551534896
R^2 Score: 0.9999999999992559
------------------------------
Model: GradientBoosting
Mean Absolute Error: 735.7544101452686
R^2 Score: 0.9995804606810341
------------------------------
Model: SVR
Mean Absolute Error: 10537.842418527009
R^2 Score: 0.8852285272626321
------------------------------
Model: DecisionTree
Mean Absolute Error: 792.0003339455669
R^2 Score: 0.9993897473427699
------------------------------
Model: KNeighbors
Mean Absolute Error: 1458.0247119719486
R^2 Score: 0.9974467725802582
------------------------------
Model: AdaBoost
Mean Absolute Error: 3598.844367444381
R^2 Score: 0.9913440162563125
------------------------------
```

Figure 1: Performance of Metrics

4.2 Web-Based Property Feature Input Interface

Figure 2 illustrated the web-based property feature Input Interface used to capture key housing parameters applied in residential property valuation. The parameters included property location, stand size, a building size, number of rooms, amenities, and construction materials. These attributes reflected typical considerations used by real estate agents when assessing housing value. The structured and standardised input format ensured consistency in data entry, improved data quality, and enabled the regression models to generate reliable and market-aligned property value estimates. This demonstrated the importance of well-defined property features in accurate housing price prediction.

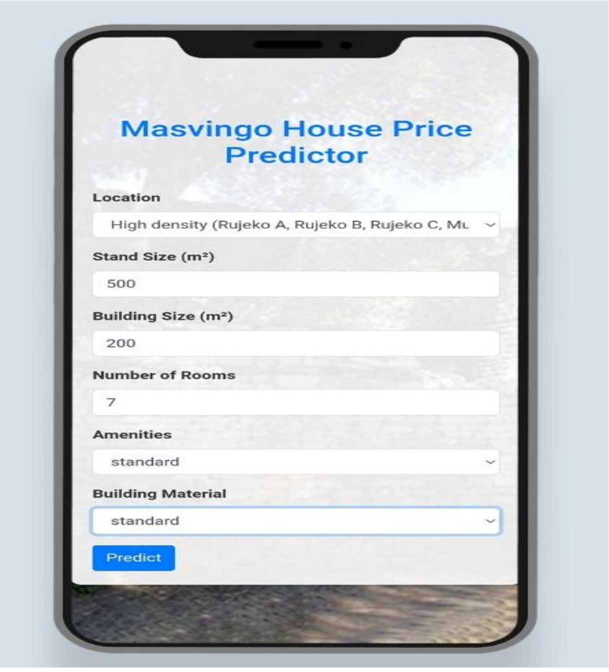


Figure 2: Web-based Property Feature Input

4.3 Predicted Results using the Website Interface

The developed website successfully implemented the random forest regression algorithm in predicting the provided key parameters (Figure 3). The Random Forest Regression model, demonstrated the best performance after data augmentation and recalibration (MAE = 464.67; $R^2 = 0.8298$), was integrated into the web platform to generate consistent and reliable valuation outputs. The key parameters that were used are high-density residential property. 500 m² stand size; 200 m² building size; seven roomed houses; standard amenities and standard construction materials. The algorithm using the website was able to display the predicted a property value of $30,784.55. This result demonstrated the model's ability to translate structured property features into realistic, market-aligned price estimates.

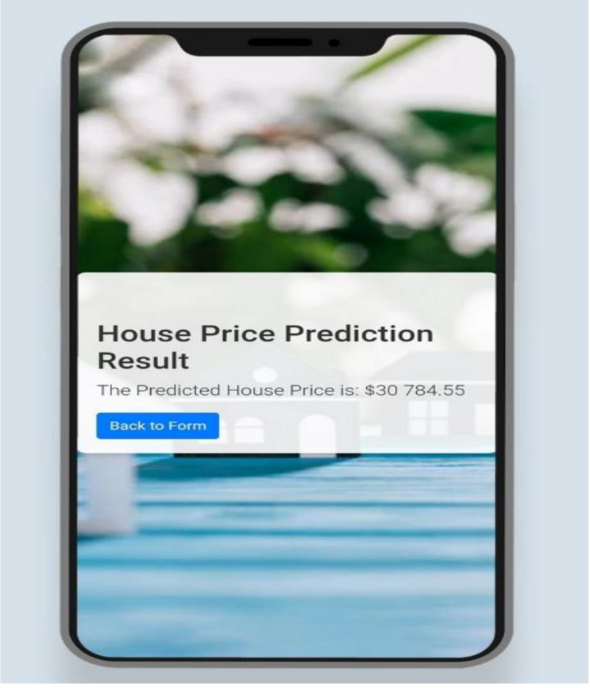


Figure 3: Website Chatbot Predicted Value

The algorithm successfully generated property value predictions through the WhatsApp chatbot system, as illustrated in Figure 4. The same input parameters used in Figure 2 were applied within the chatbot interface.

The Random Forest Regression model, predicted the market value of the property as $30,784.55 using the WhatsApp Chatbot.

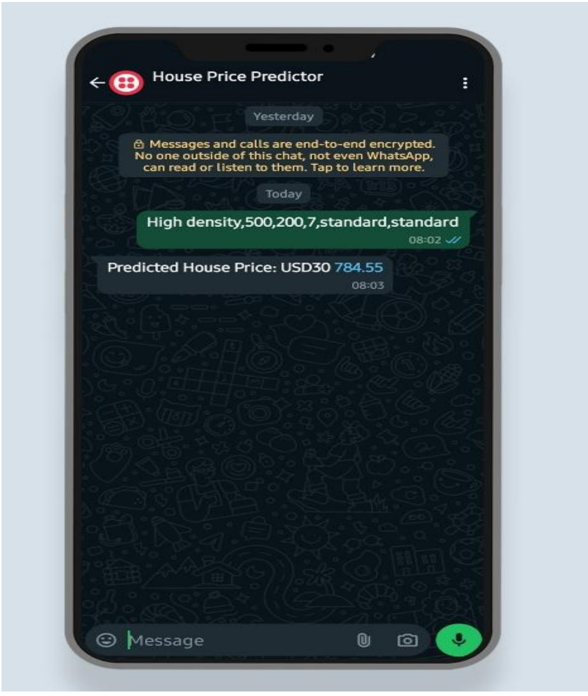


Figure 4: WhatsApp Chatbot Prediction Interface

5. Discussion

The study demonstrates that the Random Forest Regression model effectively captured the multidimensional factors influencing house prices in Masvingo Urban. Using key attributes such as stand size (500 m²), building size (200 m²), number of rooms (seven), and standard amenities, the model predicted a property value of USD 30,784.55. This is consistent with earlier studies which established that ensemble learning techniques outperform traditional valuation approaches by capturing complex, non-linear relationships among housing attributes (Wang et al., 2019; Hoang et al., 2021).

The strong performance of the Random Forest model (MAE = 464.67; R^2 = 0.8298, after recalibration) aligns with findings by Zhang (2023) and Han (2023), who also reported that tree-based ensemble models provide robust and reliable predictions even in heterogeneous housing markets. Similar to their studies conducted in developed markets, this research confirms that Random Forest models remain effective in emerging markets when appropriate data preprocessing and calibration techniques are applied.

However, unlike studies conducted in data-rich environments such as Iowa (Zhang, 2023) or Washington (Cui, 2023), this study operated under conditions of data scarcity. To address this limitation, synthetic data generation and recalibration were employed. While high R^2 values initially suggested overfitting, recalibration significantly improved model generalization. This finding supports earlier observations that machine learning models in African contexts require careful preprocessing to avoid misleading performance metrics (Fasanya and Awodimila, 2020).

The study further supports the argument that local similarity-based models such as K-Nearest Neighbours are effective in modelling neighbourhood-level price patterns (Huang et al., 2017; Zhu, 2023). However, the superior performance of Random Forest in this study suggests that ensemble approaches are more suitable for Masvingo's heterogeneous housing market, where property characteristics and pricing patterns vary significantly across locations. This extends prior work by demonstrating that ensemble learning can compensate for sparse and noisy data common in developing economies.

From a systems perspective, the integration of the prediction model into both a web-based interface and a WhatsApp chatbot significantly enhanced accessibility and usability. Consistent with this finding, Oluwafemi et al. (2021) and Prakasam et al. (2023), also reported that chatbot-based systems improve access to digital services, particularly in regions with limited physical infrastructure. The WhatsApp-based deployment is especially relevant in Zimbabwe, where mobile penetration is high and desktop-based platforms are less accessible.

Furthermore, the results directly address challenges identified in earlier Zimbabwean studies. Makoni (2012) and Moyo (2018) highlighted that traditional valuation methods such as Comparative Market Analysis and the Cost Approach are slow, subjective, and highly sensitive to economic volatility. The current study demonstrates that a machine learning-driven approach can provide faster, more consistent, and data-driven valuations, thereby extending existing valuation practices rather than replacing them entirely.

Overall, the findings of this study align with global trends in real estate technology, as observed in platforms such as Zillow and Redfin, which leverage machine learning for automated valuation (Bourassa et al., 2010; Rossi & Biddle, 2013). However, this research uniquely contributes to the Zimbabwean context by demonstrating that accurate, real-time property valuation is achievable even in data-constrained environments through careful model selection, data augmentation, and multi-platform deployment.

6. Future Work

The house price forecasting model using machine learning has shown a lot of promise as a valuable business tool for the Masvingo real estate market. Random Forest algorithm, proved to accurately predict property value, and to assist real estate professionals to perform their jobs in a better way. The model has promised the potential to unlock even more value for Zimbabwe's real estate industry. Future work should incorporate data from multiple geographical locations in Zimbabwe to evaluate the performance of the model. There is need for quantitative assessments of construction material quality, beyond visual inspection. Furthermore, future work should incorporate the usage of non-destructive testing (NDT) methods. These are Ultrasonic Pulse Velocity (UPV), Rebound Hammer tests, and thermal imaging. These could be employed on building materials to evaluate the integrity, hardness, and durability of building material quality. Therefore, by integrating these material quality measurements into the Random Forest model this would allow the system predictions to be more precise and standardized.

7. Acknowledgement

The authors thankfully acknowledge the peer reviewers and editorial team of POTRAZ for their critical feedback and guidance, which substantially improved the quality of this work. We also extend our sincere appreciation to local real estate agencies in Masvingo for providing essential data and insights that informed the study. We are indebted to colleagues and co-researchers who contributed to data collection, analysis, and scholarly discussions. Finally, the authors acknowledge Great Zimbabwe University for the academic environment and resources that enabled the successful completion of this research.

References

Alkali, M A. (2019). "Assessing the forecasting performance of ARIMA and ARIMAX models of residential prices in Abuja Nigeria." Asia Proceedings of Social Sciences 4.1 (2019): 4-6.

Csaky, R. (2019). "Deep learning based chatbot models." arXiv preprint arXiv:1908.08835.

Cui, T. F. (2023). Essays in Housing Markets. University of Pennsylvania.

Das, S., and Alain K. (2008). "Is a DFM well-suited for forecasting regional house price inflation?."

Das, S., Rangan, G., and Alain, K. (2010). "The blessing of dimensionality in forecasting real house price growth in the nine census divisions of the US." Journal of Housing Research 19.1 : 89-109.

Fasanya, I. O., and Crystal, P. A. (2020). "Are commodity prices good predictors of inflation? The African perspective." Resources Policy 69 : 101802.

Gao, H., Danlu, C., Tianhong, L., Felix, W., Laurens, M., Kilian Q. Weinberger Multi-Scale Dense Networks for Resource Efficient Image Classification. https://arxiv.org/abs/1703.09844

Goodman, A. C., and Thomas G. T. (2008). "Where are the speculative bubbles in US housing markets?." Journal of Housing Economics 17.2 : 117-137.

Gupta, R., and Sonali, D. (2008). "Spatial Bayesian methods of forecasting house prices in six metropolitan areas of South Africa." South African Journal of Economics 76.2: 298-313.

Han, Sumin, et al. (2023) "Improving Real Estate Appraisal with POI Integration and Areal Embedding." arXiv preprint arXiv:2311.11812.

Hase, Dhanashri, et al. (2023). "WhatsApp Chat Analysis Based on NLP Using Machine Learning." International Journal of Innovative Research in Engineering.

Hoesli, M., Steven C. B., Donato, S., and Sumei, Z. (2010). "Land Leverage and House Prices." In 46th Annual AREUEA Conference Paper.

Ismail O. fasanya, Crystal P. Awodimila Are commodity prices good predictors of inflation? The African perspective https://www.sciencedirect.com/science/article/abs/pii/S0301420720301367

Juneja, Sonia, et al. (2023). "House price prediction using machine learning algorithms." International Journal for Research in Applied Science & Engineering Technology (IJRASET).

Kouiroukidis, N., and Georgios, E. (2011). "The effects of dimensionality curse in high dimensional knn search." 2011 15th panhellenic conference on informatics. IEEE.

Liu, X. (2013). "Spatial and temporal dependence in house price prediction." The Journal of Real Estate Finance and Economics 47.2: 341-369.

Makoni, S. C. (2019). Neighborhood Socioeconomic Stratification: An Analysis of Factors to Classify Neighborhoods through Principal Components Analysis, Clustering, and Classification Techniques. MS thesis. Utica College.

Madyembwa, M., Kernan M., and Sibonile, M. (2020). An Automated Data Pre-processing Technique for Machine Learning in Critical Systems. International Journal of Electronic Engineering and Computer Science Vol. 5, No. 1, pp. 1-9 http://www.aiscience.org/journal/ijeecs

Montero, L., José, M., Beatriz L. I., and Antonio P. (2009). "Estimating commercial property prices: an application of cokriging with housing prices as ancillary information." Journal of Geographical Systems 11.4: 407-425.

Moyo, V. (2018). "A Practical Approach To Estimating The Cost Of Equity For South African Firms Undertaking Cross-Border Investments." The International Business & Economics Research Journal (Online) 17.2: 35.

Murekachiro, D., Thabang, M., and Hima, V. (2019). "Predicting emerging and frontier stock markets using deep neural networks." Proceedings of SAI Intelligent Systems Conference. Cham: Springer International Publishing.

Mpofu, N. (2006). "Forecasting stock prices using a weightless neural network." Journal of Sustainable Development in Africa 8.1: 115-119.

Mutendadzamera, S., and Farikayi K. Mutasa. "Forecasting stock prices on the Zimbabwe Stock Exchange (ZSE) using Arima and Arch/Garch models." International Journal of Management Sciences 3.6 (2014): 419-432.

Partson Paradza, Joseph Awoamim Yacim, Benita Zulch. CONSISTENCY AND FAIRNESS OF PROPERTY VALUATION FOR COMPENSATION FOR LAND AND IMPROVEMENTS IN ZIMBABWE https://repository.up.ac.za/server/api/core/bitstreams/fd451bbb-c7e2-4c69-8e75-00f14b165768/content

Patel, Tanu, and Jyoti Sahu. "Comparative analysis of regression techniques for accurate house price prediction using machine learning: a statistical analysis method." 2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS). IEEE, 2023.

Pillay, Surendran. "Determining the optimal Arima model for forecasting the share price index of the Johannesburg Stock Exchange." Journal of Management Information and Decision Sciences 23.5 (2020): 527-538.

Prakasam, S., et al. "Design and development of ai-powered healthcare whatsapp chatbot." 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN). IEEE, 2023.

Pratheeth, S., and Vishnu Prasad. "Stock price prediction using machine learning and deep learning." 2021 IEEE Mysore Sub Section International Conference (MysuruCon). IEEE, 2021.

R. S. Latha et al., "Stock Movement Prediction using KNN Machine Learning Algorithm," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-5, doi: 10.1109/ICCCI54379.2022.9740781

Samruddhi, K., and R. Ashok Kumar. "Used car price prediction using K-nearest neighbor based model." Int. J. Innov. Res. Appl. Sci. Eng.(IJIRASE) 4.3 (2020): 2020-686.

Thamarai, M., and S. P. Malarvizhi. "House Price Prediction Modeling Using Machine Learning." International Journal of Information Engineering & Electronic Business 12.2 (2020).

Uchendu Eugene CHIGBU Urban-rural land linkages: Concept explained using 'continuum' theory and metaphor https://www.fig.net/resources/proceedings/fig_proceedings/fig2022/papers/ts01b/TS01B_chigbu_11762.pdf

Wang, Feng, et al. "House price prediction approach based on deep learning and ARIMA model." 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). IEEE, 2019.

Winky K.O. Ho, Bo-Sin Tang and Siu Wai Wong Predicting property prices with machine learning algorithms. JOURNAL OF PROPERTY RESEARCH 2021, VOL. 38, NO. 1, 48–70 https://doi.org/10.1080/09599916.2020.1832558

Zhang, Ling. "Housing price prediction using machine learning algorithm." Journal of World Economy 2.3 (2023): 18-26.

Zhu, J., Gong, Y., Liu, C., Du, J., Song, C., Chen, J., & Pei, T. (2023). Assessing the effects of subjective and objective measures on housing prices with street view imagery: A case study of Suzhou. Land, 12(12), 2095.