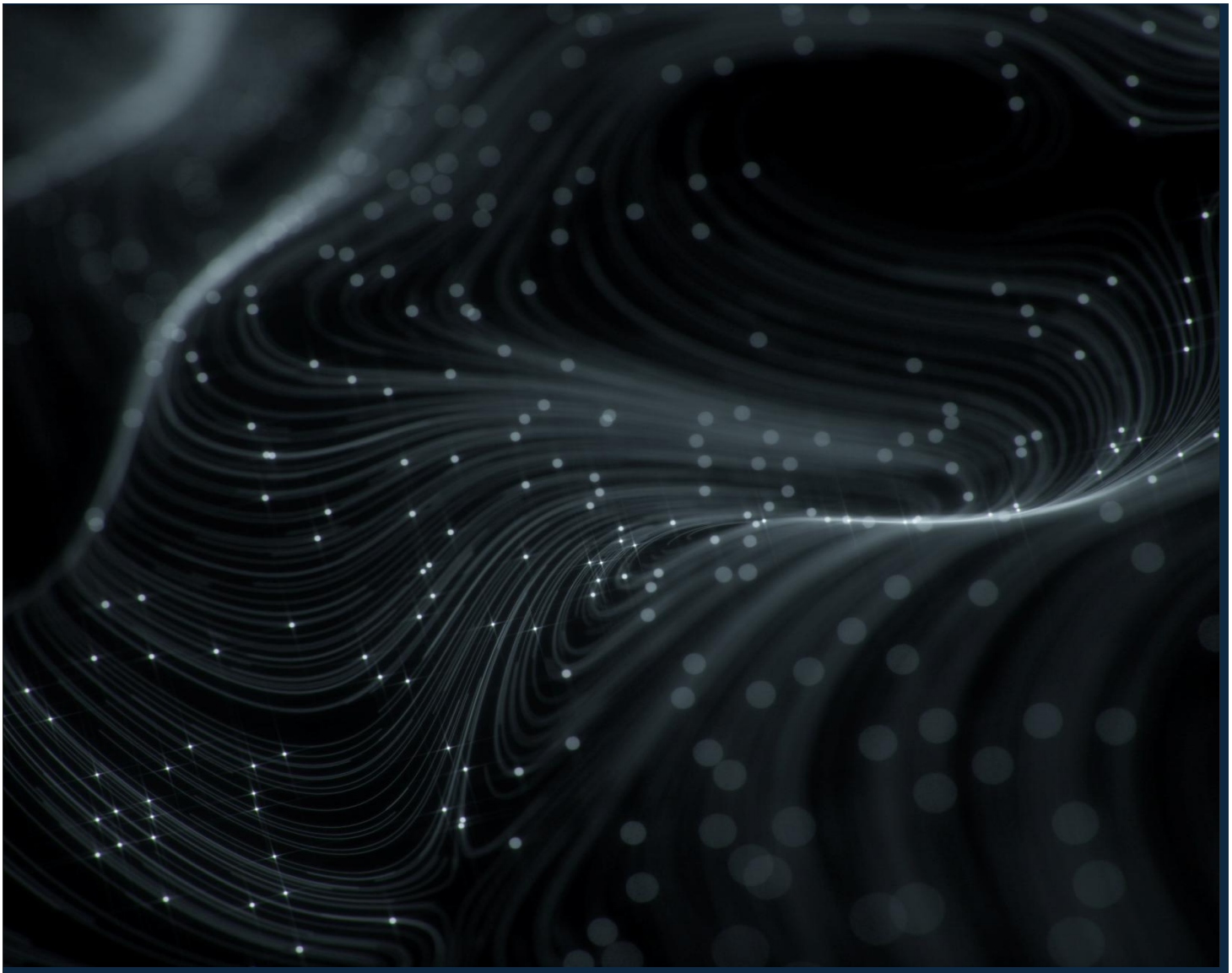

GeoVest: A Scenario-Aware Machine Learning Approach to Predicting Long-Term Land Values in America

Minai Fernando[1], Dr. Saminda Premaratne[2]

[1] Undergraduate Researcher, University of Central Lancashire (via UCL Sri Lanka)

[2] Senior Lecturer, Department of Information Technology, Faculty of Information Technology, University of Moratuwa, Sri Lanka

Abstract

Accurately predicting long term real estate values is a critical yet underexplored domain, Existing models focus on short term real estate value prediction that too only for developed regions. This study introduces GeoVest, a scenario aware machine learning approach to predicting long term land values in America. GeoVest (RMSE:0.053) focuses on both developed and underdeveloped regions taking macroeconomic indicators such as GDP, population, and employment rate for model training. This system integrates Random Forest with trend based forecasts to predict land values for up to 20 years in the future under three scenarios, best-case, middle range case and worst-case scenario. The framework is deployed as a user-facing mobile application, where users can select the location (per state), land size and duration of investment. This effectively presents the user with a range of values for a specific location of their choosing. While it does not consider localized factors such as infrastructure projects, zoning changes, etc, its strong baseline performance highlights the visibility of macroeconomic drivers for scalable land value prediction. Future work will incorporate localized development data to improve resilience and accuracy further. Overall, Geovest demonstrates the potential to make long term forecasts by combining ensemble learning and trend based analysis.

Introduction

Problem Statement

The art of investing has always been fruitful for people who understand how to invest. According to Gitman et al. (2015), 'An investment is an asset into which funds can be placed with the expectations that it will generate an income'. However, investing would only generate a high profit return (ROI) if the investor has thoroughly studied and applied investment strategies into a sector with a profitable return as hinted in Allie et al (2016). Thus, many capable individuals have deterred from investing due to insufficient knowledge for ideal investments with a high ROI.

Investing may be made easier by introducing an investment system that tracks the current market trends, analyzes the data, and makes recommendations based on user input, that too for long-term investments.

To achieve this, artificial intelligence can be incorporated. Currently, artificial intelligence is a revolutionary concept that is actively making its way into all possible sectors in the world. For example, artificially intelligent medical robots that assist doctors in patient care (Nawrat, 2023), or artificially intelligent robots in construction (Li, 2018). This is as it can be

trained to think like a human with an instantaneous response, it can store a multiload of data, access the necessary information and present an answer in a matter of seconds. Thus, many professionals currently focus on incorporating AI into their chosen field to improve their overall efficiency. The investment industry is one such field. It stands as a ripe domain for AI prediction applications given the complex and data-intensive nature of land valuation, especially in underdeveloped regions. As stated by Tekouabou et al. (2023), Even though there is a significant growth in the number of research publications in this area of study, there has been little effort to identify its possibilities.' This provides ample opportunity to build a strong foundation for this application.

This project aims to build an ML prediction application into the investment industry that analyzes data and predicts the value of bare land in both developed and underdeveloped regions through mass appraisal for long-term investment in America.

State of the Art

The real-estate sector has had major developments in the past decade where predicting property prices with artificial intelligence have been increasingly researched upon, however, the developments are emphasized on predicting the price of houses, apartments, etc., that too for short-term.

The sole purpose of my application is to predict the price of bare land through mass appraisals for long-term investments, and with the scarcity of research found on the subject, this section will be focused on the recent developments of successful predicting models based on short-term property valuation. This is as the main concept would remain the same, as they both use ML to predict values on the factors surrounding real estate.

Commonly Used Algorithms for Real Estate Value Predictions

Firstly, there is Random Forest (RF), it is an ensemble model that works by improving on how decision trees operate. Decision Tree (DT) works by taking the input data and partitioning similar data into branches of nodes and leaves according to its predictors value before reaching the end and predicting the final value (De Ville, 2013). RF works by creating multiple trees with randomized data by bootstrapping and taking the average of all the trees as the final output. RF solves a major problem that exist in DT, this is due to multiple trees being built on randomized data, it solves the problem of bias that exists in decision trees and has a lower variance in comparison, however, since RF contains maximal tree structures, it still has a high variance, thus, making it highly unstable (Segal, 2004).

This could be controlled by using optimal settings, as highlighted by Antipov and Pokryshevskaya (2012) in their research, which used RF to predict the value of real estate in Saint Petersburg, Russia. They obtained their dataset from the Russian real-estate catalog which is constantly monitored by the publisher to ensure credibility. Their datasets consist of a little over 2000 samples for two-bedroom apartments in 2010. In addition, Antipov and Pokryshevskaya (2012) concluded that RF could stabilize outliers, and work with missing values on different levels. They also accentuated that their model was successful in predicting price efficiently because they predicted the price-per-meter than the overall price. They also clarified that RF is not prone to errors by overfitting, as previously mentioned it could only generate the best results if optimal settings are given.

Furthermore, Hong et al. (2019) used RF to predict the price of housing in south Korea, they used 50 DTs with a depth of 17, they obtained 40% of samples of apartments from the highly developed Gangnam district between 2006 to 2017. According to their experiment, the deviation of the predicted price from the actual price was only 5.5%, in addition, RF was noted to make fewer outlier predictions and the predicted price deviating more than 50% from the actual price was low as 0.5%. Hong et al. (2019) highlighted that their RF model was able to predict efficiently with a low deviation rate due to two main reasons, the first being them using a large dataset of more than 16,000 samples in a small area of 39.55km^2, the second being them using the data for only one type of real-estate, apartments, as it contains the same properties, this further enhances the model's ability to comprehend the influence of each factor on the pricing to predict an accurate value. Their model was more accurate than the model by Antipov and Pokryshevskaya (2012) due to the larger datasets used across the span of more than 10 years compared to their 1 year, however there is a possibility that their model's efficiency could have been increased further by predicting price-per-meter as emphasized by Antipov and Pokryshevskaya (2012). Hong et al (2019) also concluded that the main contributor to the value of their apartments is the development of the area in which they are located. This shows that the development of the area chosen would indeed impact on the value of the land predicted.

In continuation, Yilmazer and Kocaman (2020) used RF and multiple regression analysis (MRA) for mass appraisals in Mamak, Turkey. The study consisted of 96 independent variables and 1,200 cases for residential transactions from 2013 to 2018. Their variables were identified after extensive field work and were concluded to be the most influential on property prices by experts. They removed outliers and missing values in cases resulting 1,162 usable cases and then divided the datasets into training (80%) and testing (20%). They observed that RF handles the variable selection and optimization of the datasets automatically while bootstrapping 80% of the cases for randomized data selection and they used 100 Decisions trees after conducting Out-of-the-bad (OOB) error rate analysis.

They discovered that Random Forest is more efficient in predicting mass appraisals for property valuations with a lower RMSE of 16,486 TRY compared to MRA with a RSME of 17,975 TRY.

Moreover, Louati et al (2022) used RF, DT and Linear regression to predict the value of property in Riyadh, they collected 5496 localized data from northern Riyadh using GeoTech's DAAL website from 2018 to 2020. The data was then preprocessed and split into training (70%) and testing (30%) subsets. The accuracy of the models was analyzed and compared using Mean Absolute Error, Mean Squared Error and Median Squared Error. The lower these scores, the more accurate the model will be. They observed that RF had the highest accuracy rate followed by decision trees and then linear regression.

Lastly, Khalafallah (2008) used ANNs to predict the price of short-term housing prices in Orlando, America. He used data from a span of 9 years and used several networks with a different number of hidden layers and neurons to decide on the best combination for predicting. He discovered that using a network with 1 hidden layer and 4 neurons worked best to predict the value. Their deviation from the actual price ranged from -2% to +2%, however Khalafallah (2008) noted that ANNs although very robust in approximating input and output, are unsuitable for long-term forecasts. This approach achieved a higher accuracy rate compared to Antipov and Pokryshevskaya (2012), Hong et al. (2019), Yilmazer and Kocaman (2020), and Louati et al (2022). This could be due to the structure of Artificial neural networks where all the nodes on different layers are connected to each other to output a value, however since it is a black box model, it lacks transparency on how the value is predicted which would create distrust among the users of the tool.

Both RF and ANN are highly accurate in predicting real estate prices, however, ANN cannot be interpreted, which makes it a questionable choice for an investment application, and RF has overfitting issues that could lead to a high variance in the predicted output. However, as shown in Antipov and Pokryshevskaya (2012), Hong et al (2019), Yilmazer and Kocaman (2020), and Louati et al (2022), Random Forest appears to be the most suitable model for this project due to its handling of outliers, optimization of datasets, bootstrapping random data for better accuracy and interpretability with their quantitative research methodology. However, to predict the prices for long-term, there needs to be additional features incorporated. Random Forest combined with time-series features and trend-based forecast could be used. This is because time series features will allow RF to sequentially identify patterns that exist to better predictions as it does not chronologically observe factors to output a decision. This project would also benefit by incorporating the above-mentioned methods by focusing on factors that contribute to regional progression as it would be vital in predicting land values in the future. It would be implemented with the

consideration of optimum settings for Random Forest. This could produce a robust prediction model to evaluate the price of land in the long term.

Finally, this prediction model for land would not only provide a financial advantage to investors, for example, as stressed by Yilmazer and Kocaman (2020) it is mandatory to construct data driven models for mass appraisals, this statement was backed by Louati et al (2022) as they highlighted that there is a possibility to develop a land prediction tool to enhance mass appraisal models. Thus, this project would also advance mass appraisal models in the ML prediction domain while helping investors.

Similar AI predicting tools

Currently, there are no predicting tools that predict the value of bare land for investment for long-term, however, there is a tool for short-term real-estate predictions in developed neighborhoods called Zestimate, it is an artificially intelligent prediction tool provided by Zillow, an online listing platform. It uses an extensive number of factors to determine the price of a house in a specific location. It analyzes public data to estimate a ballpark figure of the house value after being compared with historical data patterns. However, due to the volatility of the real-estate market, Zestimate only exists for short-term evaluation and Zestimate was also observed to be less accurate in poor neighborhoods (Fu et al., 2023), this is due the ease of obtaining data in developed neighborhoods. Zillow's price prediction fluctuates significantly since a major update in early 2021, this was a result of the tool analyzing the contributing factors in more detail, however this does not equate to higher efficiency. As emphasized by Malik and Fu (2023) algorithms that reacts too quickly will equate to a higher variance, which might not benefit the costumers. This is because different datasets would be given a different predicted result resulting from higher variance, the inconsistency in predictions would reduce the efficiency of the tool.

My approach would solve the gaps that exist within apps like Zestimate to achieve my goal by incorporating long-term land value predictions in both developed and underdeveloped regions.

Methodology

The Evolutionary prototyping methodology was used throughout the project. This is as it allows the development of a prototype with core features while integrating the additional features to further refine the application. This will form a minimal viable product (MVP) at the earliest convenience, allowing for supervisor feedback, aligning with the deadlines given.

The factors obtained were divided into core and secondary factors. The core factors will be incorporated initially and the secondary factors after the MVP has been developed. The factors and the details of the development process are as follows.

Table 1. Factors considered and their given category

Factor considered	Type of factor	Duration
Land value	Observed factor	2000 - 2022
The Gross domestic product (GDP) of the county/state	Core factor	2000 - 2022
Employment Rates	core factor	2000-2022
Population	core factor	2000 - 2022
Crime rates (per 100,000)	secondary factor	2022
Natural disasters	secondary factor	Accumulated since 1953
Possible disease outbreaks	secondary factor	2020-2023

The data sets for the above-mentioned factors were obtained from government websites such as the Bureau of Economic Analysis, etc. and were stored on an excel sheet after cleaning and preprocessing has been completed, this would later be imported to visual studio code for model training. Data preparation has been explained in detail below.

Implementation

Data cleaning

Firstly, after the datasets were acquired from trusted governmental sources in America, like the Bureau of statistical analysis, Bureau of economic analysis, etc. The land value, employment rate, population count, and the GDP of all U.S. states were taken for the past 22 years starting from the year 2000 till 2022/2023, they were downloaded in csv form.

The datasets were then analyzed for null values, the states with multiple null values were removed entirely, these states included Montana, New Mexico, North Dakota and District of Colombia. Apart from the above, all the remaining states were kept for training. Some of the kept states had a few missing values that were filled using interpolation and backward filling depending on the dataset. The datasets were structured to have a state column, year column, and a column for the values it holds like employment rate, etc. The state column was converted to a numeric form using one-hot encoding and the values under the value column were normalized using min-max scaler. All transformations were made using

Jupyter Notebook. The scaler, and min max values were saved in a separate folder locally for each dataset.

Model Training

To begin with, after the datasets were cleaned, transformed and normalized, a few suitable models were chosen. Although RF was deemed to be the most suitable model for this project, a couple of other models were trained as well to ensure a proper comparison between them and Random Forest.

Moving on, the choice of models must address ethical issues such as AI transparency. Since the chosen model would be used in an informative investment application, it is essential for the user to understand how the model works to predict the land values in order for the user to trust the application and its predictions. To achieve this, it is imperative that the chosen model is interpretable, but, due to the highly accurate nature of deep learning and black box models, some were considered, but with at least some instances of interpretability when it comes to the predictions, as stated by Buocz (2018), even human interpretations cannot be fully justified. Thus, ML/AI predictions, if accurate, could be considered.

The following are the models that were chosen:

Table 2. Chosen Models

Model	Sequence of training
Long-Short term memory (LSTM)	First model
An ensemble of LSTM and XGBoost	Second model
An ensemble of Random Forest and trend-based forecasting	Third model

The mentioned models were trained on 20 years of data that contains the employment rate, population, Gross domestic product (GDP), and land values per year. The land value was kept as the target value.

The first choice for a model was Long-Short term model, it is a type of Recurrent Neural Network (RNN) designed to handle sequential data and long-term dependencies. It was selected due to its ability to remember past data, as the chosen methodology was evolutionary prototyping, each identified factor, both core and secondary features for land

value prediction could be added sequentially to create and enhance the prediction accuracy.

LSTM itself was first used; LSTM used an 80/20 split; with 2 LSTM layers and a dense output layer with a learning rate of 0.001. It used early stopping if validation loss does not improve.

The second model was an ensemble of LSTM + XGBoost. This model introduced an additional stage after the LSTM model was trained that used the output of the LSTM model as the input to train the XGBoost model. XGBoost built multiple decision trees sequentially that were learnt from the errors of the previous tree to enhance its accuracy.

The third model was an ensemble model of trend-based forecast and Random Forest. The historical data was fed into trend-based forecast to predict future feature values for 20 years, these values included employment rate, population, and GDP. These values were capped so extreme values on either end would not be predicted. Random Forest was then trained on the historical dataset to understand feature importance, and then the predicted future values by trend-based forecasts were fed into the Random Forest model to predict the future land values. Throughout the process of training Random Forest, land value was the target variable and not used for training, this had a high accuracy. GridSearchCV was used to find the optimum settings for RF to produce the lowest RMSE.

Throughout the development of the application, the frontend and backend were simultaneously developed using the evolutionary prototyping methodology to ensure a working prototype exists at all times. The frontend was completed using the flutter framework consisting of a minimalistic yet professional user interface with the required explanations on how the model works to ensure user transparency, the predicted values of the states were divided into two categories: the most lucrative and least lucrative states. These were displayed to properly inform the user of the expected ROI for the states under each category. Flask server was used to connect the frontend and backend.

Discussion

The following are the in-detail descriptions and results of all the models used,

Long-Short Term Model (First Model):

It had an RMSE of 0.1053, an MSE of 0.0111, and MAE of 0.0809, and an R^2 of 0.0015. This shows that this model proved to have a moderate absolute prediction error but is unable to generalize well as shown by the small R^2 value, this suggests that the model requires additional changes to increase the overall efficiency.

Ensemble Model of Long-Short Term and Gradient Boosting (Second Model):

The second model resulted in a higher RMSE of 0.1119, MSE of 0.0125, MAE of 0.0826, and a lower R^2 of -0.1281 in comparison to the first model. This suggests that incorporating Xgboost with LSTM resulted in a worse model than using LSTM alone as the R^2 was a negative value, the failure is likely to be caused due to LSTM's weak predictions being used as inputs for the Xgboost so the residuals lacked meaningful patterns for Xgboost to learn from. Thus, Xgboost may have overfitted to noise than relying on reliable patterns as LSTM had a low variance explainability to begin with.

Ensemble Model of Trend-based Forecasts and Random Forest (Third Model):

This model had an RMSE of 0.0533, MSE of 0.0028, MAE of 0.0371, and R^2 of 0.8249. This shows that the model has a high accuracy rate with a lower RMSE, MSE and MAE score and a high R^2 score that means the model can explain 82.49% of the variance in the data. This ensemble model achieved the highest accuracy among all the models trained.

A snap analysis was also conducted to understand the main feature used by the RF model to make its predictions. It showed that the main feature used by the model for predictions was population, followed by GDP and then by employment rate.

However, to further assure this feature importance rank as credible, an analysis was conducted on the weight of feature importance throughout the historical dataset. This was done to determine if keeping the feature importance at this stable rank would support reliable predictions. The results are the following:

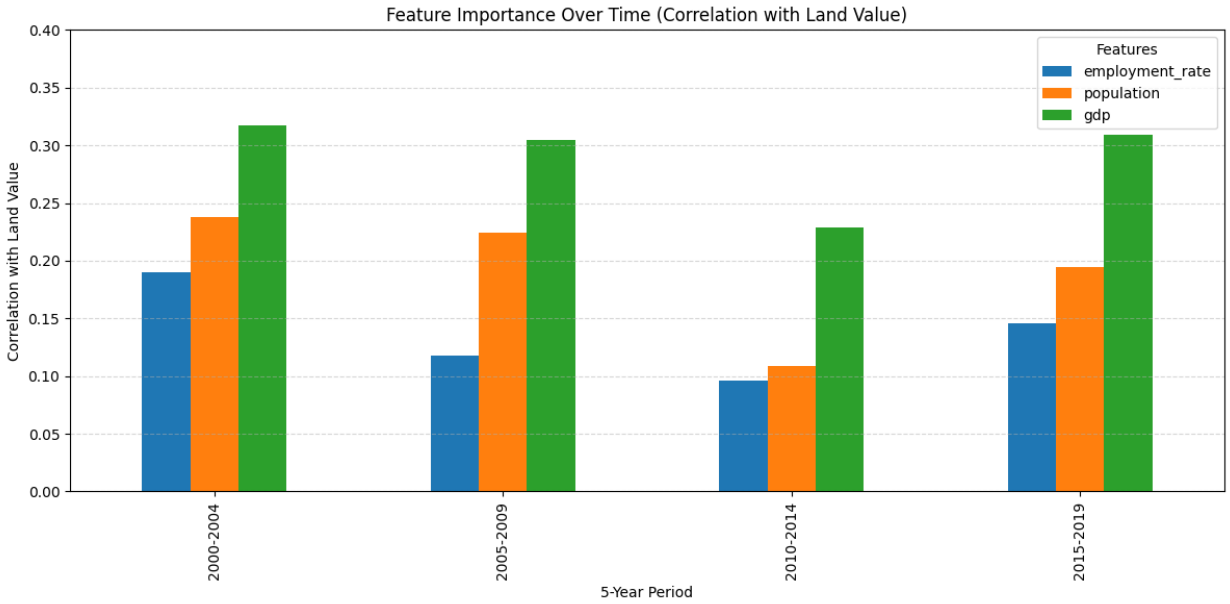

Figure 1. Feature importance throughout 20 years

As seen in figure 1, when it comes to ranking feature importance, GDP dominated for the past 2 decades, followed by population and then employment rate. Interestingly, according to the snap analysis of the model, the population was the most prominent when predicting land prices, followed by GDP and lastly by employment rate. This difference arises because the model prioritizes features that consistently explain variation in land prices across the states, rather than those that merely dominated historical trends. Population directly affects land demand, making it a more immediate and reliable predictor, whereas GDP influences prices more indirectly. This aligns with real-world dynamics, as higher population levels typically increase demand for land, thereby driving up land values. This concludes that maintaining the feature ranking the model used would enable the model to predict reliably.

The RMSE, MSE, MAE AND R^2 of the models are shown below:

Table 3. Evaluation metrics used for each model

	Trend-based + Random Forest	LSTM	LSTM + Xgboost
RMSE	0.053	0.1053	0.1119
MSE	0.0028	0.0111	0.0125
MAE	0.0371	0.0809	0.0826
R^2	0.8249	0.0015	-0.1281

This shows that the most accurate model was the ensemble model that used Trend-based forecasting and Random Forest (Third Model). It has the lowest RMSE of 0.053 which indicates that the model average prediction error is 5.3% of the data's range. The MAE of 0.0371 explains that the model is 3.71% off per prediction and the R^2 of 0.8249 entails that the model can explain 82.49% of the variability in the data. And the worst model was the LSTM + Xgboost with the highest RSME, MSE and MAE of 0.1119, 0.0125 and 0.0826 respectively and the lowest R^2 value of -0.1281 indicating the model does not predict within a good range of the actual value and its explanation of the variability is very poor. With the above considerations, the third model made of random forest and trend-based forecast was finalized as the best model to use for the investment application.

After finalizing the model, an extra precautionary measure was incorporated to ensure a more accurate prediction for a given year. The selected model was used to predict under three scenarios instead of just one value, it predicted a best case, middle-range case, and worst case. In the best-case scenario, the model predicted the best possible value for land based on the factors, and in the worst-case scenario the model will predict the worst possible value, for both cases, it depends on the year specified by the user and the size of

land input. This was made possible due to the trend-based forecast predicting best, middle, and worst case for the features during the initial predictions.

Below is the final land value prediction visualized with both the historical and the predicted output for all the states used:

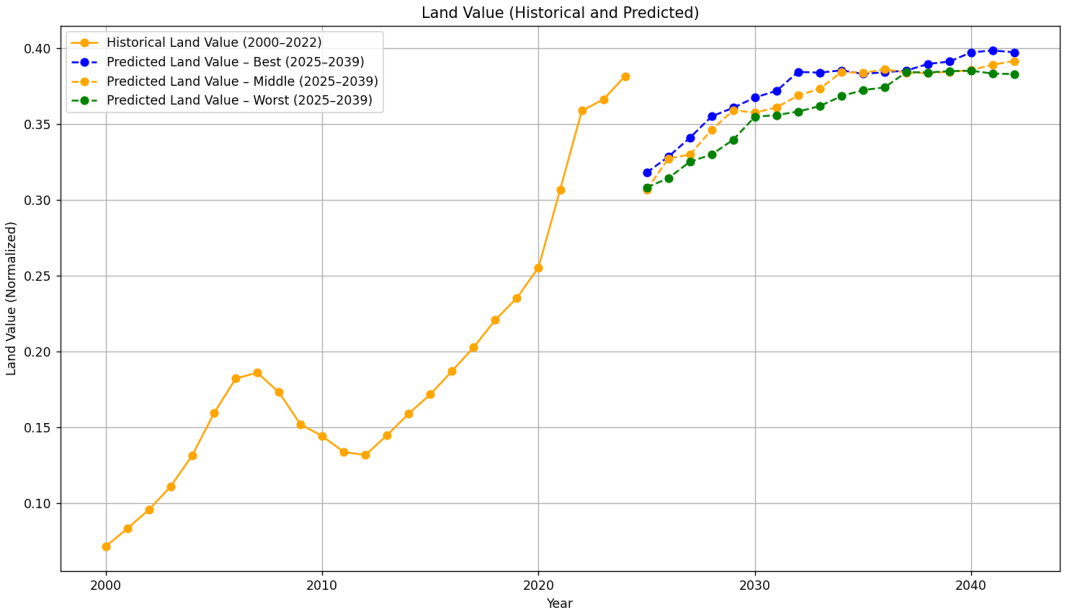

Figure 2. visualization of the historical and predicted output

The lines in blue represent the best-case scenario, the lines in orange represent the middle-range scenario, and the lines in green represent the worst-case scenario for the predicted values. The yellow lines represent historical land values, and the gap in between is the years used for testing the model. The difference between the end plot of the historical land values and the start of the predicted land values is explained by the 5.3% average deviation from the data as shown by the RMSE of 0.053.

Below are the actual land values vs predicted land values in comparison

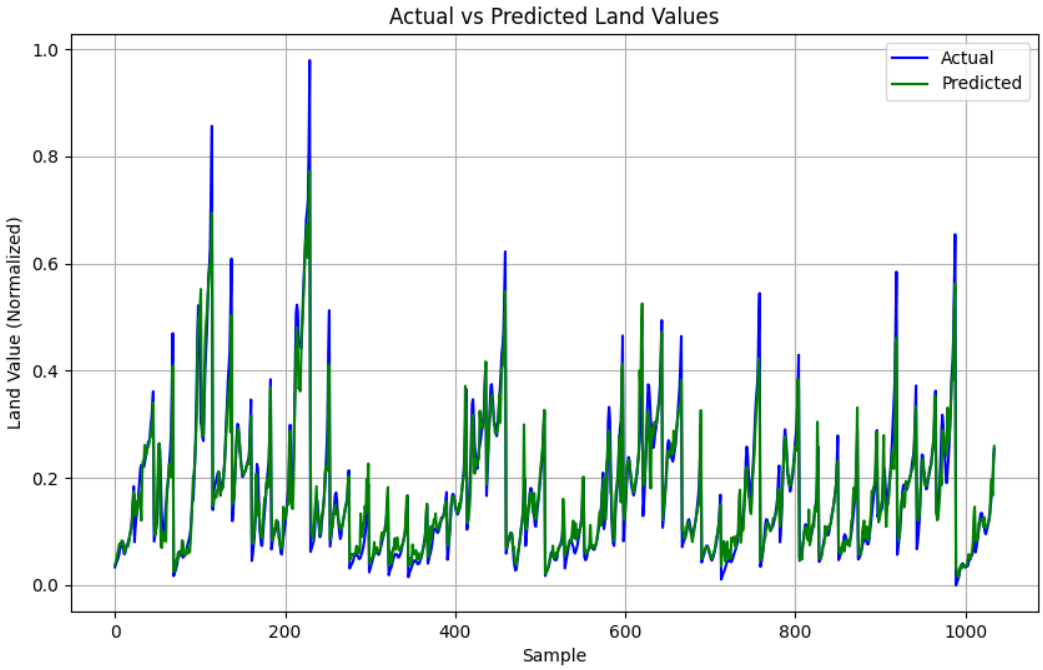

Figure 3. Actual vs Predicted land values

Alternative Testing

The testing process happened in a few stages. Firstly, there was testing for the algorithm itself. When training the model, there was already an 80/20 split for training and testing per state to test the model on unseen data. The RMSE, MSE, MAE AND R^2 were calculated, and its efficiency was noted and compared above.

Scenario-based testing was also conducted where artificial feature values were presented to the model to predict land value accordingly. This was done to determine how the model performs under different economic conditions, and it was evaluated using its own RMSE. It predicts the values under the best-case scenario, middle-range scenario and the worst-case scenario. The following are the results:

Table 4. results of the scenario-based testing

Scenario	RMSE	Key Observations
Best-case	0.0254	Performs best, accurately captures growth trends
Mixed case	0.0709	Struggles with inconsistency, predictions show higher variance. Underestimates downturn impact, needs better training on negative trends
Worst-case	0.0441	Performs second best, accurately captures growth trends

This shows that the model is best suited to predict land values in economic booms, followed by economic recessions and lastly mixed economic conditions. This can be attributed to the fact that the historical training data mainly consisted of positive or negative trends rather than a mix of both.

Conclusion

The original problem discovered was the lack of knowledge by the public to financially invest and build a profit in the long run. To combat this issue, a mobile application was developed to predict the price of land for long-term investment while informing the users about how the prediction is achieved. Different models were trained and tested to obtain the model with the highest accuracy. A professional front-end was also developed to ensure a reliable way to display predictions to the user. The front and backend were integrated to complete the application.

The model with the highest accuracy was selected and used for predicting land values for the next 20 years. The application successfully displays locations across the U.S and accepts user input, sends the input to the backend through flask server, retrieves the prediction for that year and sends it back to the frontend and displays it. The prediction accuracy was high and the front-end satisfies the prediction display.

The approach used can be generalized for similar problems such as long-term regression predictions, but the final accuracy will depend on the quality of the datasets used to train the model.

It is also imperative to note that while GeoVest achieves an overall RMSE of 0.053 on the test set using GDP, population, and employment rate, it may not capture localized, discrete shocks (e.g., major infrastructure projects, rezoning, or natural disasters) due to having limited training factors. To assess robustness, we performed Performance testing (RMSE, MAE, MSE, R^2), temporal cross-validation, Snap analysis, and scenario-based testing. However, we recommend incorporating project- and hazard-level proxies (building permits, types of projects completed e.g. highways, roads, etc.) and disaster indices) and to better quantify uncertainty for investors.

Further Improvements

For further improvement, more factors could be used for training the model, for example, projects completed, zoning laws, etc. The factors used could also be made dynamic, for example, for the current year, the factors could be updated automatically instead of using

the outdated values if there ever was a change. The trend-based and Random Forest model could also be integrated with gradient tree boosting to further enhance the model's accuracy. The final outputs of the ensemble model could be used as an input for gradient tree boosting, it can then calculate the residuals from RF to predict land values that are more accurate in value.

References

Allie, J., West, D. and Willows, G. (2016) 'The value of financial advice: An analysis of the investment performance of advised and non-advised individual investors', Investment Analysts Journal, 45(sup1), pp. S63–S74. doi: 10.1080/10293523.2016.1201292.

Almaslukh, B. (2020). 'A gradient boosting method for effective prediction of housing prices in complex real estate systems', in 2020 International Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, pp. 217–222.

Antipov, E.A. and Pokryshevskaya, E.B. (2012) 'Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics', Expert Systems with Applications, 39(2), pp. 1772–1778.

Buocz, T.J., 2018. Artificial Intelligence in Court. *Legitimacy Problems of AI Assistance in the Judiciary". Retskraft–Copenhagen Journal of Legal Studies*, *2*(1), pp.41-59.

Casali Y, Aydin NY, Comes T (2022) Machine learning for spatial analyses in urban areas: a scoping review. Sustain Cities Soci 85:104050

Choy Lennon HT, Ho Winky KO (2023) The use of machine learning in real estate research. Land 12(4):740

Couper, M.P., Baker, R. and Mechling, J., 2011. Placement and design of navigation buttons in web surveys. *Survey Practice*, *4*(1).

Damodaran, A. (2003) Investment philosophies: Successful strategies and the investors who made them work. Vol. 185. New York: John Wiley & Sons.

De Ville, B. (2013) 'Decision trees', Wiley Interdisciplinary Reviews: Computational Statistics, 5(6), pp. 448–455. doi: 10.1002/wics.1278.

Dondo, J. (2022) 'Most novice investors counting on advisors for education', Wealth Professional. Available at: https://www.wealthprofessional.ca/investments/socially-responsible-investing/most-novice-investors-counting-on-advisors-for-education/367761 (Accessed: 20 November 2024).

Fu, R., Huang, Y., Mehta, N., Singh, P.V. and Srinivasan, K. (2023) 'Unequal Impact of Zestimate on the Housing Market', SSRN. [Sl]. Available at: SSRN.

Gitman, L.J. *et al.* (2015) *Fundamentals of investing*. Pearson Higher Education AU. Available at: https://books.google.lk/books?hl=en&lr=&id=DB3iBAAAQBAJ&oi=fnd&pg=PP1&dq=investing+&ots=KkXcgFrwdK&sig=foo7wIqEjGqSPu1j51ezdM_e28c&redir_esc=y#v=onepage&q=investing&f=false

Guo JQ, Chiang SH, Liu M, Yang CC, Guo KY (2020) Can machine learning algorithms associated with text mining from internet data improve housing price prediction performance? Int J Strateg Prop Manag 24(5):300–312

Hitz, S. (2022) Redfin Strategic Analysis. Honors thesis. University of Nebraska-Lincoln. Available at: https://digitalcommons.unl.edu/honorstheses/ (Accessed: 2 November 2024).

Hong, J., Choi, H. and Woo, S.K. (2019) 'A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea', International Journal of Strategic Property Management, 24(3), pp. 140–152.

Hooda, I. and Chhillar, R.S., 2015. Software test process, testing types and techniques. *International Journal of Computer Applications*, *111*(13).

Khalafallah, A. (2008) 'Neural network-based model for predicting housing market performance', Tsinghua Science and Technology, 13(S1), pp. 325–328.

Lorenz F, Willwersch J, Cajias M, Fuerst F (2022) Interpretable machine learning for real estate market analysis. Real Estate Economics, Early View

Li, R.Y.M. (2018). 'Robots for the construction industry', in An Economic Analysis on Automated Construction Safety. Singapore: Springer. doi: 10.1007/978-981-10-5771-7_2.

Louati, A., Lahyani, R., Aldaej, A., Aldumaykhi, A. and Otai, S. (2022) 'Price forecasting for real estate using machine learning: A case study on Riyadh city', Concurrency and Computation: Practice and Experience, 34(6), p. e6748.

Malik, N. and Fu, R. (2023) 'Why does my Zestimate fluctuate? Platform design of Algorithmic Pricing Models', Platform Design of Algorithmic Pricing Models (28 March 2023).

Natural hazard disclosures (nhds) (no date) NATURAL HAZARD DISCLOSURE, LLC. Available at: https://www.nhdreport.com/nhd (Accessed: 31 October 2024).

Nawrat, Z. (2023) 'Introduction to AI-driven surgical robots', Artificial Intelligence Surgery, 3(2), pp. 90–97.

Pan, S.J. and Yang, Q. (2010) 'A survey on transfer learning', IEEE Transactions on Knowledge and Data Engineering, 22(10), pp. 1345–1359. doi: 10.1109/TKDE.2009.191.

Renigier-Biłozor, M., Janowski, A. and d'Amato, M. (2019) 'Automated valuation model based on fuzzy and rough set theory for real estate market with insufficient source data', Land Use Policy, 87, p. 104021. doi: 10.1016/j.landusepol.2019.104021.

Robertson, S. and Robertson, J., 2012. *Mastering the requirements process: Getting requirements right*. Addison-wesley.

Schoch, L., 2021. Reflections on the Parkinson's Project: learning experiences with the Gibbs' Reflective Cycle.

Segal, M.R. (2004) Machine Learning Benchmarks and Random Forest Regression. UCSF: Center for Bioinformatics and Molecular Biostatistics.

Stufflebeam, D.L. and Coryn, C.L., 2014. *Evaluation theory, models, and applications*. John Wiley & Sons.

Talabis, M.R.M., Martin, J. and Wheeler, E. (2013) *Information security risk assessment toolkit: Practical assessments through data collection and data analysis*. Waltham, Mass: Syngress**.**

Teoh EZ, Yau WC, Ong TS, Connie T (2022) Explainable housing price prediction with determinant analysis. Int J Housing Markets Analysis

Tekouabou, S.C.K., Gherghina, Ş.C., Kameni, E.D. *et al* (2023). AI-Based on Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey. *Arch Computat Methods Eng* 31, 1079–1095. https://doi.org/10.1007/s11831-023-10010-5

Text - H.R.6580 - 117th Congress (2021-2022): Algorithmic Accountability Act of 2022 (2022) Congress.Gov. Available at: https://www.congress.gov (Accessed: 31 October 2024).

The Fair Housing Act (2023) Civil Rights Division. Available at: https://www.justice.gov/crt/fair-housing-act-1 (Accessed: 31 October 2024).

Walczak, S. (2019). 'Artificial Neural Networks', in Advanced Methodologies and Technologies in Artificial Intelligence, Computer Simulation, and Human-Computer Interaction. doi: 10.4018/978-1-5225-7368-5.ch004.

Yilmazer, S. and Kocaman, S. (2020). 'A mass appraisal assessment study using machine learning based on multiple regression and random forest', Land Use Policy, 99, p. 104889.

Bibliography

Baur K, Rosenfelder M, Lutz B (2022) Automated real estate valuation with machine learning models using property descriptions. Exp Syst Appl 213:119147

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F. and Ratti, C. (2021). 'Understanding house price appreciation using multi-source big geo-data and machine learning', Land Policy, 111, p. 104919. doi: 10.1016/j.landusepol.2021.104919.

Li X, Kao C (2022) Spatial analysis and modeling of the housing value changes in the us during the covid-19 pandemic. J Risk Financial Manag 15(3):139

Louati A, Lahyani R, Aldaej A, Aldumaykhi A, Otai S (2022) Price forecasting for real estate using machine learning: a case study on Riyadh city. Concurr Comput 34(6):e6748

Lahmiri, S., Bekiros, S. and Avdoulas, C. (2023). 'A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization', Decision Analytics Journal, 6, p. 100166. doi: 10.1016/j.dajour.2023.100166.

Li, X. and Kao, C. (2022). 'Spatial analysis and modeling of the housing value changes in the US during the COVID-19 pandemic', Journal of Risk and Financial Management, 15(3), p. 139. doi: 10.3390/jrfm15030139.

Mora-Garcia, R.T., Cespedes-Lopez, M.F. and Raul Perez-Sanchez, V. (2022). 'Housing price prediction using machine learning algorithms in COVID-19 times', Land, 11(11), p. 2100. doi: 10.3390/land11112100.

Rischard, M., Branson, Z., Miratrix, L. and Bornn, L. (2021) 'Do school districts affect NYC house prices? Identifying border differences using a Bayesian nonparametric approach to geographic regression discontinuity designs', Journal of the American Statistical Association, 116(534), pp. 619–631. doi: 10.1080/01621459.2020.1772080.

Seagraves, P., 2023. Real Estate Insights: Is the AI revolution a real estate boon or bane?. *Journal of Property Investment & Finance*, (ahead-of-print).

Salih, A.M. and Wang, Y. (2024) 'Are Linear Regression Models White Box and Interpretable?', arXiv. Available at: https://arxiv.org (Accessed: [Insert Date]).

Sangani, D., Erickson, K. and Al Hasan, M. (2017). 'Predicting Zillow estimation error using linear regression and gradient boosting', in 2017 IEEE 14th International Conference on

Mobile Ad Hoc and Sensor Systems (MASS). IEEE, pp. 530–534. doi: 10.1109/MASS.2017.90.

Sisman, S. and Aydinoglu, A.C. (2022). 'Improving performance of mass real estate valuation through application of the dataset optimization and spatially constrained multivariate clustering analysis', Land Policy, 119, p. 106167. doi: 10.1016/j.landusepol.2022.106167.

Wang, Z., Wang, Y., Sensen, W. and Zhenhong, D. (2022) 'House price valuation model based on geographically neural network weighted regression: The case study of Shenzhen, China', ISPRS International Journal of Geo-Information, 11(8), p. 450. doi: 10.3390/ijgi11080450.

Zhan, C., Wu, Z., Liu, Y., Xie, Z. and Chen, W. (2020). 'Housing prices prediction with deep learning: An application for the real estate market in Taiwan', in 2020 IEEE 18th International Conference on Industrial Informatics (INDIN). IEEE, vol. 1, pp. 719–724. doi: 10.1109/INDIN45582.2020.9442300.