Real Estate Investment Recommendations Using Machine Learning Techniques

Janet Hu, Michail Mersinias, Luis Smith

Department of Computer Science, University of Texas at Austin

Abstract

Historically, real estate has been a major area of investment. In this paper, we provide real estate investment recommendations by using machine learning techniques combined with financial models. More specifically, we first introduce and propose Zestimate, which is an estimation of home sale prices, and is used to calculate the ZHVI home value index. By utilizing various machine learning models such as decision trees and neural networks on the Los Angeles CA housing market, we show that it is indeed an accurate estimation of home sale prices. Furthermore, using the Austin, TX, housing market as a case study, we conduct a systematic evaluation of forecasting models such as ARIMA, Prophet and a long short-term memory (LSTM) neural network for the task of predicting future home prices on the zip code area level. We add to these predictions by using the CAPM financial model in order to calculate alpha and beta metrics, which show the risk-adjusted expected returns for each zip code area. Then, we provide recommendations for real estate investment, based on the forecasted growth and the CAPM metrics of each zip code area. We experimentally show that both our forecasting predictions and the CAPM metrics manage to recommend zip code areas which achieved high actual growth rates, while they avoided the ones with low actual growth rates.

1 Introduction

As of November 2022, it is estimated that the home ownership rate amongst Americans is 66.0 percent [18]. Prospective homeowners choose a home in a particular city and zip code for a variety of reasons: cost of living, proximity to friends and

family, proximity to work, zoning for school districts, levels of crime, and other considerations.

One important consideration for choosing a home is the projected future market value of the home. A homeowner's financial interest in their home is called home equity. It is defined as the property's current market value less any liens that are attached to that property [8]. Therefore, if the current market value of the home increases over time, the home equity will increase as well. According to the U.S. Census Bureau, home equity represents 28.9 percent of the overall wealth of Americans [19]. Thus, the decision to buy a home represents one of the most important financial decisions a person will ever make. The growth of home equity will play a large role in a homeowner's future net worth as they age.

For this reason, we use a proprietary home valuation model called Zestimate® to estimate homes' market value throughout the United States [20]. We use machine learning in order to estimate the current value of a home. In this paper, we will present a case study of the Los Angeles market to show the ability of the Zestimate to estimate the sale price, with a low margin of error.

Furthermore, in this paper we present a set of tools that will be added to the Zillow app to help prospective homeowners choose a home that will be the best investment for their financial future. The Zillow Home Value Index (ZHVI), is an index that is used to measure monthly changes in property-level Zestimates across a wide variety of geographies and housing types [21]. Our first tool is the ability to forecast the expected growth of housing prices across zip codes in a given city. Our next tool helps prospective homeowners manage risk using an investment model known as the Capital Asset Pricing Model (CAPM). To demonstrate the capabilities of these tools, we will use Austin, Texas as a case study.

2 Individual Home Value Prediction (Zestimate)

2.1 Research Background and Literature Overview

The Zestimate is Zillow's best estimate of the market value of a home. Estimating the value of a home can be framed as a regression problem. Common machine learning techniques for regression include linear regression, decision tree-based methods (e.g. random forest, XGBoost, LightGBM, CatBoost) and neural network-based methods.

In linear regression, modeling of the relationship between two variables is carried out by fitting a linear equation in accordance with the observed data, considering one variable to be an independent variable, and the other to be a dependent variable. However, one main disadvantage of linear regression is that most of the naturally occurring phenomena are non-linear, therefore linear regression technique fails to fit complex data sets properly because it assumes that there exists a linear relationship among the input and output variables [9].

A decision tree is a classifier expressed as a recursive partition of the instance space. In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values [13]. Random forests are a combination of decision tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Injecting the right kind of randomness makes them accurate classifiers and regressors [3]. Gradient Boosting Decision Trees (GBDT) is an ensemble model of decision trees, which are trained in sequence. In each iteration, GBDT learns the decision trees by fitting the negative gradients (also known as residual errors). GBDT is a widely-used machine learning algorithm, due to its efficiency, accuracy, and interpretability. GBDT achieves state-of-the-art performances in many machine learning tasks, such as multi-class classification, click prediction, and learning to rank [11].

In a neural network, neurons (processing units) are connected with each other through "synaptic weights", or "weights" in short. Each neuron in a network receives

"weighted" information via these synaptic connections from the neurons that it is connected to and produces an output by passing the weighted sum of those input signals through an "activation function". Feed-forward neural networks (FFNN) is a class of neural networks where there is no "feedback" from the outputs of the neurons towards the inputs throughout the network [15].

A research study attempted to estimate the price of a home using machine learning techniques such as support vector regression, ensembles of regression trees, k-nearest neighbors, and neural networks. In the study, it was determined that outperforming models were those consisting of ensemble models of regression trees [1].

2.2 Data

Our case study focuses on a full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) from 2016. The dataset used for this case study was released to the public as a kaggle competition [22]. The features used for this dataset include:

- Geographical features (e.g. "latitude", "longitude", "regionidcounty", "regionidcity", "regionidzip", etc.)
- Size features (e.g "lotsizesquarefeet", "'finishedsquarefeet15", "basementsqft", "garagetotalsqft", "yardbuildingsqft17", etc.)
- Structural features (e.g."buildingqualitytypeid", "buildingclasstypeid", "typeconstructiontypeid", "yearbuilt", "numberofstories", "unitcnt", etc.)
- Characteristics of the home (e.g."buildingqualitytypeid", "buildingclasstypeid", "typeconstructiontypeid", "yearbuilt", "numberofstories", "unitcnt", etc.)
- Tax features (e.g."buildingqualitytypeid", "buildingclasstypeid", "typeconstructiontypeid", "yearbuilt", "numberofstories", "unitcnt", etc.)
- Other miscellaneous features

The complete set of features and feature definitions can be found in the "zillow_data_dictionary.xlsx" file of the Zillow's Home Value Prediction dataset [22].

2.3 Experimental Evaluation

For our experiments, we first conducted an exploratory analysis on the dataset which consists of over 80 features related to properties of individual houses. Thus, before proceeding to modeling and prediction, we analyzed the distribution of each feature and performed necessary transformations in order to produce normal distributions and eliminate outliers. Moreover, we examined the correlation between features and eliminated redundant features which resulted in high correlation of over 75%. In the figure that follows, we present the feature importance table as it is produced by our machine learning models after the data preparation and feature engineering steps.

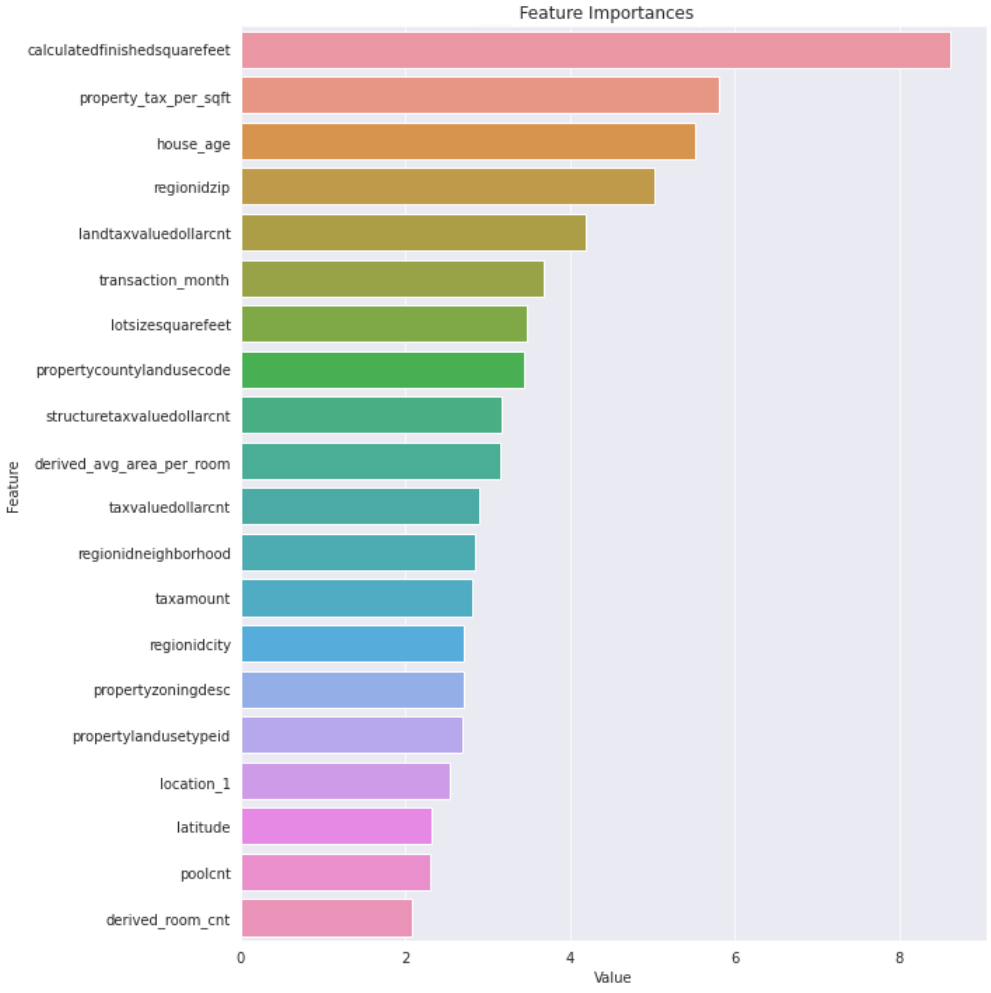

Figure 1: Top 20 features used for estimating the price value of an individual house

The feature importance table gives us an overview of the features which constitute the estimated home value, which forms the basis of our research. In order to validate its accuracy, we utilize various machine learning regression techniques to predict the logerror which is the log difference between the estimated value (Zestimate) and the sale price of each individual house and constitutes a good measure of relative prediction accuracy [17]. We use linear models (linear regression), decision tree models (random forest, XGBoost, LightGBM, CatBoost)) as well as a feedforward neural network. For our metrics, we use the mean absolute error (MAE) and the root mean square error (RMSE). Finally, a 5-fold cross validation was conducted in order to ensure the statistical significance of our results. The table of results is presented below.

Model	RMSE	MAE
Linear Regression	8423902	52380
Random Forest Regressor	0.08277	0.05278
XGBoost Regressor	0.08311	0.05294
LightGBM Regressor	0.08260	0.05268
CatBoost Regressor	0.08279	0.05242
Feedforward Neural Network	0.08360	0.05262

Table 1: Performance of machine learning models on house price estimation

Linear regression performs poorly as as it is unable to capture the complexity of the task. All other models provide low RMSE and MAE scores, which is an indicator of good performance. The best performing model in terms of RMSE is the LightGBM Regressor, and the best performing model in terms of MAE is the CatBoost Regressor. However, all machine learning models display a similar degree of good performance which leads us to conclude that the estimated price value (Zestimate) is indeed an accurate representation of the sale price of an individual house and can thus be used as a reliable estimate for the housing sale prices in our research.

3 Forecasting Zip-Code Level Housing Prices

3.1 Research Background and Literature Overview

Forecasting the growth of ZHVI at a zip-code level can be modeled as a time series problem. Prediction of time series data is a challenging task due to both the difficulty of capturing existing trends and unexpected changes in future trends. In order to tackle this problem, we attempted a variety of different statistical and machine learning modeling approaches.

A traditional statistical approach to time series modeling is using an Autoregressive Integrated Moving Average (ARIMA) model. A major benefit of ARIMA is that it makes very few assumptions and is very flexible. Essentially, it uses a data-oriented approach that has the flexibility of fitting an appropriate model which is adapted from the structure of the data itself. With the assistance of autocorrelation function and partial autocorrelation function, the stochastic nature of the time series can be approximately modeled [5].

Facebook has also developed a simple, modular regression model for modeling time series called Prophet. One of the main benefits of Prophet is that it often works well with default parameters [16]. Prophet has also successfully been applied in other industries. For example, Prophet was used to accurately forecast future sales in one of the biggest retail companies in Bosnia and Herzegovina [24].

A machine learning approach that has been commonly used for time series modeling is a long short-term memory (LSTM) model. An LSTM model uses a recurrent neural network (RNN) architecture that has been designed to address the vanishing and exploding gradient problems of conventional RNNs. LSTMs have successfully been applied to sequence prediction tasks [14]. Previous research has shown that the LSTM model is a strong candidate model for predicting trends in real estate [10].

3.2 Data

The Zestimate home valuation model is Zillow's estimate of a home's market value [20]. ZHVI is a smoothed, seasonally adjusted measure of the typical home value, and is built by measuring monthly changes in property-level Zestimates. It reflects the typical value for homes in the 35th to 65th percentile range [23].

The average Zestimate within some range of home values determines the index level, meaning the index retains its interpretation as the dollar value of a typical home. For zip code level data, index appreciation can be interpreted as the zip code's total appreciation. In other words, the ZHVI appreciation can be viewed as the theoretical financial return that could be gained from buying all homes in a given zip code in one period and selling them in the next period [21]. Therefore, the growth of the index can be thought of as a good proxy of owning an individual home in that zip code.

For our case study of Austin, Texas, we used zip codes that belong to the "Austin-Round Rock-Georgetown, TX" metro area according to our produced dataset.

3.3 Experimental Evaluation

Our goal was to forecast the 1-year, 2-year, 3-year, and 4-year forecasts of the ZHVI for each zip code. This would help consumers understand the projected growth of investments in different zip codes. For our training data, we used historical data from 2008-2017. Thus, our projection periods were 2017-2018 (1-year forecast), 2017-2019 (2-year forecast) 2017-2020 (3-year forecast), 2017-2021 (4-year forecast). The different forecasting models we attempted were ARIMA, Prophet, and an LSTM model. Our expectation was that the forecasting models would perform better for shorter time intervals (e.g. 1-year, 2-year) than longer time intervals (3-year, 4-year), as it is likely easier to capture shorter localized trends.

The ARIMA model was created using the pmdarima python library. The library contains automated hyperparmeter tuning that finds the optimal parameters for that ARIMA model to fit the training data. The Prophet model was created using the

prophet python library. Manual hyperparameter tuning was attempted, however, the default parameters resulted in the best fitting model.

The LSTM models were created using PyTorch. Preprocessing was done to normalize the data to remove the underlying increasing trend and to scale the values to a range between [-1, 1]. Since the data has an overall increasing trend over time, we make our data stationary by calculating the difference between each data point in a given time series. We then scale those values to a range between -1 and 1 since the default non-linear activation function for LSTM layers in PyTorch is tanh which has a range of [-1, 1]. Each model takes an input of 48 months to predict the projection period specified (1, 2, 3, or 4 years). For example, in the 1-year forecast LSTM model the training data consists of x, y pairs where x is an array of length 48 (for 48 months) and the y is an array of length 12 (for 12 months). To calculate the predicted values at the original scale from the output of the model, we apply the inverse of the normalization function to rescale our outputs. We then take the most recently observed ZHVI value from the input period and use the predicted differences to calculate the full predicted results at the original scale.

Experimentation with the number of months to include in the input array and manual hyperparameter tuning was performed to find the best overall models. The final model architecture consisted of a single LSTM layer with a hidden state of 128 features followed by a final linear layer that converted the output of the last LSTM layer into the desired output size. The models were each trained for a total of 200 epochs using MSE loss using an Adam optimizer.

To understand the importance of choosing a zip code with strong growth potential, we can illustrate using a hypothetical $200,000 investment into a home in Austin in 2017 as seen in Figure 2. We can see the expected value of the investment in 2021 ranges widely depending on if the best performing zip code, the median performing zip code, or the worst performing zip code were to be chosen.

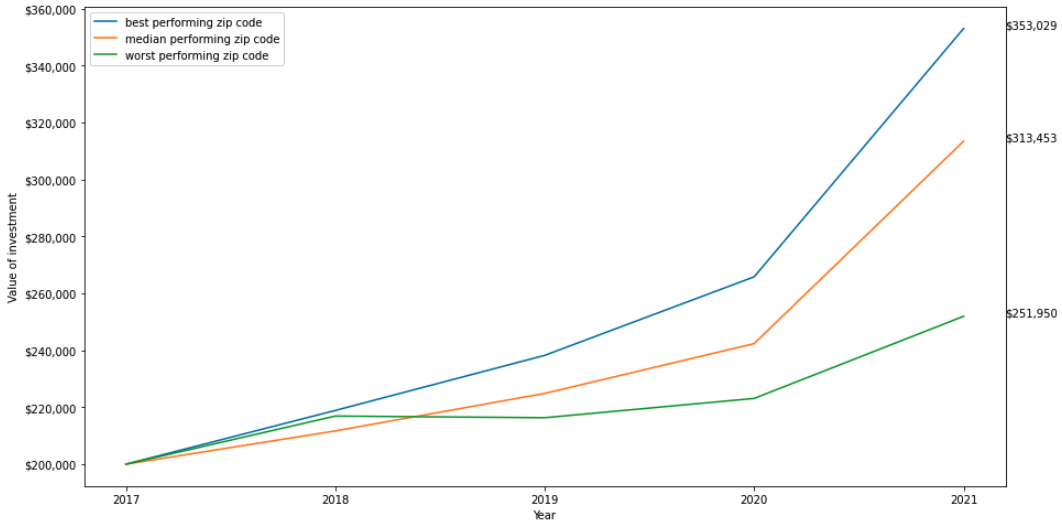

Figure 2: Value of a hypothetical $200,000 investment from 2017 until 2021

Thus, having the ability to forecast which zip codes are likely to grow in the near future could give prospective homeowners useful information that could aid them in their decision making process.

3.3.1 Evaluation Metrics

Our chosen evaluation metrics were Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). In independent surveys that were performed over a timeline of 25 years, these three evaluation metrics were concluded to be the most popular for forecasting. There is still no consensus on a single best metric.

RMSE and MAE are scale-dependent metrics, as errors have physical dimensions and are expressed in the units of the data under analysis. Therefore, since we are estimating the ZHVI in a zip code, which reflects the typical value for homes in that zip code, we can understand RMSE and MAE in those contexts. For example, MAE would express the dollar value of the mean (absolute) error of a forecasted ZHVI from the actual ZHVI. RMSE would express the dollar value of the root mean squared error of a forecasted ZHVI from the actual ZHVI.

MAPE on the other hand is a scale-independent metric. It is a measure that describes something conceptually similar to a mean absolute percentage error, although it is not symmetrical, as it puts a heavier penalty on negative errors [2].

3.3.2 Evaluation Results

The different forecasting models were evaluated using the average of the evaluation metrics across zip codes across different yearly prediction time intervals. The analysis was broken up into short-term forecasts (1-year, 2-year) and long-term forecasts (3-year, 4-year).

The results of the short-term forecasts across the different models can be seen in Table 2. Prophet performed best across all evaluation metrics, followed by LSTM and then ARIMA. We can interpret Prophet's MAE results as being off by an average of $3,917 for 1-year forecasts and an average of $6,623 for 2-year forecasts. In absolute percentage terms, the MAPE suggests that Prophet was off on average by 1.0% for 1-year forecast and 1.6% for 2-year forecasts.

Prediction Interval	Model	RMSE	MAE	MAPE
1-year	ARIMA	7770	6611	0.0167
	Prophet	4601	3917	0.0100
	LSTM	5176	4423	0.0121
2-year	ARIMA	14311	12282	0.0305
	Prophet	7974	6623	0.0163
	LSTM	8714	7536	0.0211

Table 2: Short-Term Forecast Model Comparison

A visual comparison of the actual 1-year growth rates and the projected 1-year growth rates can be seen in Figure 3 and Figure 4. In general, we can see that Prophet did an adequate job of predicting the general regions that were expected to grow the fastest, although it tends to have overestimated the actual growth rates. Prophet's predictions were within an absolute error of 2.0% for 68.5% of the zip codes.

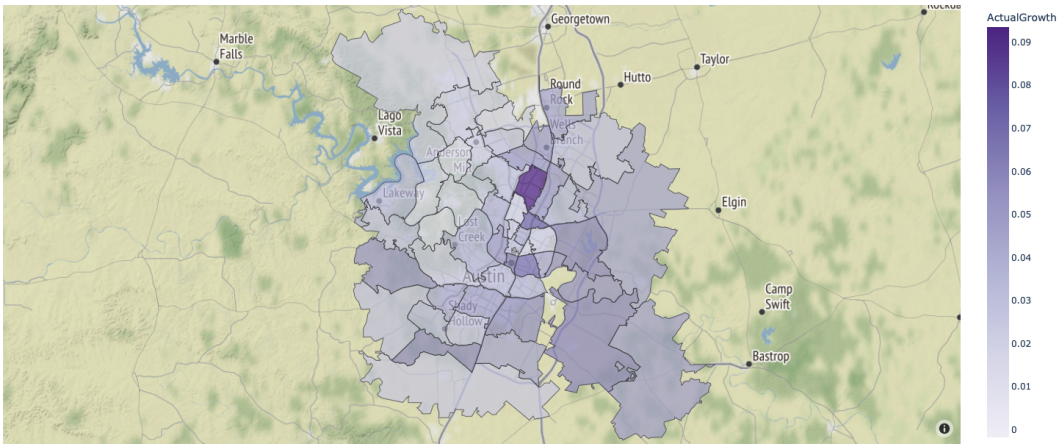

Figure 3: Actual 1-year ZHVI Growth

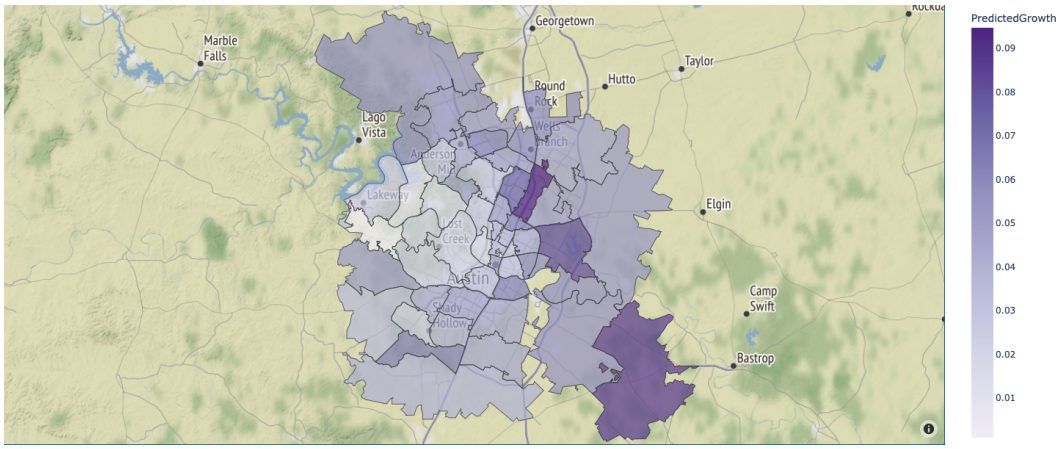

Figure 4: Predicted 1-year ZHVI Growth by Prophet

The results of the long-term forecasts across the different models can be seen in Table 3. Prophet performed best across MAE and MAPE evaluation metrics for the 3-yr forecasts, while the LSTM performed best on RMSE for the 3-yr forecasts. The LSTM performed best across all evaluation metrics for the 4-yr forecasts. We can interpret Prophet's 3-year MAE results as the forecasts being off from the actual zip code ZHVI by an average of $11,757 and its MAPE results as the forecasts being off by 2.6% on average. We can interpret LSTM's 4-year MAE results as the forecasts being off from the actual zip code ZHVI by an average of $39,754 and its MAPE results as the forecasts being off by 7.2% on average.

Prediction Interval	Model	RMSE	MAE	MAPE
3-year	ARIMA	20956	17260	0.041064
	Prophet	16355	11757	0.025873
	LSTM	15351	12628	0.031855
4-year	ARIMA	71665	43767	0.078400
	Prophet	80465	45051	0.074373
	LSTM	67161	39754	0.07166

Table 3: Long-Term Forecast Model Comparison

As seen in the results, Prophet was the overall best performing model, with the best overall evaluation metric results for the 1-year, 2-year, and 3-year predictions. Prophet has already been successfully adopted for forecasting time series in industries such as retail, and the success of the forecasts show that it is a suitable model for predicting real estate trend data. LSTM performed better for the 4-year predictions, perhaps signaling it may have been able to better capture longer term trends than traditional time series models and Prophet.

The goal is for our zip code forecasts to be reliable for prospective homeowners. For this reason, we will use Prophet as our primary model for forecasting and plan to only show the 1-year and 2-year forecasts. Given that the forecasts are off on average by 1.0% for 1-year forecast and 1.6% for 2-year forecasts, this gives prospective homeowners fairly reliable information that can aid them in their decision making.

Finally, as far as the recommendations are concerned, our forecast indicates high growth in East Austin as well as in particular zip code areas of West Austin and City Center. This predicted growth was indeed validated, as these particular geographical areas achieved a high actual growth. Furthermore, while the growth of North Austin was overestimated, our forecast correctly predicted that South Austin will not achieve considerable growth and that turned out to be the actual case.

4 Capital Asset Pricing Model

4.1 Research Background and Literature Overview

Another way to evaluate an investment is by utilizing the Capital Asset Pricing Model (CAPM) [7] which describes the relationship between systematic risk, or the general perils of investing, and expected return for assets. It is a finance model that establishes a linear relationship between the required return on an investment and risk. The model is based on the relationship between an asset's beta and alpha. Beta represents the risk-free rate. Alpha represents the risk premium, or the expected return on the market minus the risk-free rate. CAPM, along with several variations of it, has been successfully applied to real estate investing [4] as a way to evaluate investments based on not only expected return, but also on risk. In our work, we apply CAPM to real estate investing through regression [12] in order to calculate the alpha and beta metrics [6] for each zip code area. The equation has the following form:

$$Y = Beta \cdot X + Alpha \tag{1}$$

In the equation listed above, Y is the performance of the asset, which in our case is the corresponding zip code area, and X is the performance of the benchmark index, which in our case is the entire Austin market. Beta is a measure of volatility relative to a benchmark index, while alpha is the excess return on an investment after adjusting for market-related volatility.

Therefore, a high value of alpha is always desirable as it indicates a higher return of investment. On the other hand, the ideal value of beta is dependent on the investor's risk profile. An investor who values high growth, and is willing to accept risk, prefers a high value of beta as it will provide the highest return of investment. Furthermore, as beta is a multiplier rather than an addition element, a high value of beta will often affect Y to a greater degree than alpha and will be prioritized over alpha; although alpha will still be required to have a positive value. Thus, an investor

who values high growth and the risk that comes with it will opt for a high beta value and a positive alpha value. However, a risk-averse investor will prefer a low beta, which will ensure that the asset is not heavily influenced by the fluctuations of the market, and a high alpha which will provide a good return of investment. This is summarized in the table below.

Investor Type	Returns	Risk	Beta	Alpha
A	High	Accepted	High	Positive
B	Medium	Reluctant	Low	High

Table 4: Investor types and their preference profiles

In our work, we apply CAPM in order to provide zip code areas as recommendations for both investor types which wish to invest in Austin-Round Rock-Georgetown, TX metro area.

4.2 Data

For our experiments, we use the same data as in section 3.2 and once again we perform our case study of Austin, Texas by using zip codes that belong to the ”Austin-Round Rock-Georgetown, TX” metro area according to our dataset.

4.3 Experimental Evaluation

In order to calculate the alpha and the beta for each zip code area, we performed data preparation in the dataset in order to ensure the resulting data frame is in the right format. We also filled in unreported values using linear interpolation; a method of curve fitting using linear polynomials to construct new data points within the range of a discrete set of known data points.

Afterwards, we performed linear regression and fit the best line across the data points of each zip code area. A common issue with linear regression is that it is often not sufficient to capture the complexity of the data. However, after calculating the R squared metric, we report that the best fit line of linear regression resulted in a

median value of 0.925, with 95% of the zip code areas reporting a high, acceptable value of over 0.7. Therefore, CAPM can indeed be applied to our case through linear regression and the results are statistically significant. The coefficients of the linear regression represent the beta and the alpha for each zip code area, and we present them in the figures below.

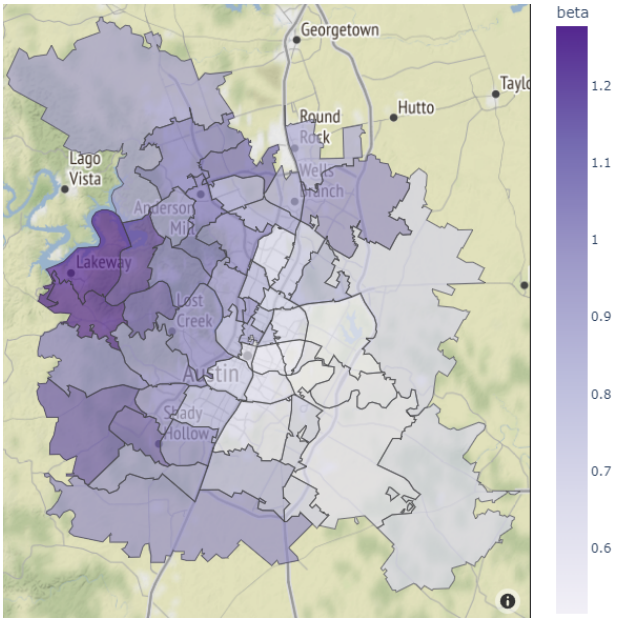

Figure 5: Beta per Zip Code Area (CAPM)

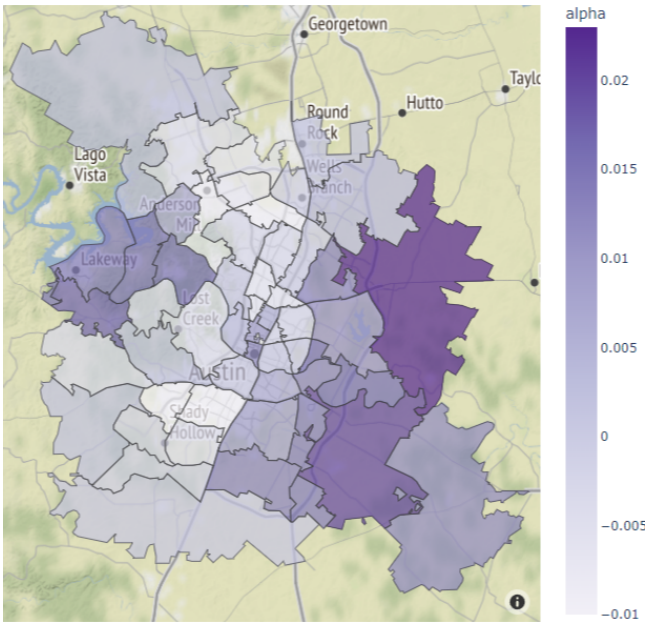

Figure 6: Alpha per Zip Code Area (CAPM)

From the figures displayed above, we can make certain observations. First, there is a clear correlation between geographical areas and alpha and beta parameters, which indicates that growth is determined by geography. Specifically, there is a clear divide between West Austin and East Austin, with North Austin, South Austin and City Center also forming distinct clusters. In the table that follows, we present the beta and alpha values for each geographical area as well as the suitability of the particular geographical area per investor type and the investment priority it is given in our recommendations.

Model	Beta	Alpha	Investor A Priority	Investor B Priority
North Austin	Medium	Low	Low	Low
South Austin	Low	Medium	Low	Low
East Austin	Low	High	Low	High
West Austin	High	High	High	Low
City Center	Medium	High	High	Medium

Table 5: Alpha and Beta for different geographical areas of Austin

Therefore, for type A investors, West Austin is the most risk-adjusted recommended area. The Lakeway area of West Austin, specifically, provides the highest beta value, along with a high alpha value, thus making it the ideal target for a type A investor. Another good investment target for type A investors is the City Center, which yields a medium beta and high alpha, meaning a high return of investment with less risk. Furthermore, North Austin is a lower priority recommendation, aimed for type A investors who cannot afford to invest in either West Austin or City Center, as it still provides a medium beta and a low but still mostly positive alpha, which indicates a relatively good investment.

As far as type B investors are concerned, East Austin is the recommended area for investment as the low beta equals low risk, while the high alpha ensures high returns of investment. In case the type B investor is willing to accept a higher degree of risk, the City Center is recommended due to the relatively low risk with a medium beta, while providing a high return of investment due to its high alpha.

Comparing the recommended results to the actual ZHVI growth, as it is depicted in Figure 3, we notice that our recommendations of East Austin to type B investors and West Austin to type A investors, as well as the City Center, yielded good returns. This is because zip code areas located in East Austin achieved the highest actual growth, while zip code areas located in West Austin and the City Center also achieved considerable actual growth. On the other hand, most zip code areas of North Austin and South Austin achieved low actual growth and were correctly recommended with the lowest priority.

5 Conclusion and Future Work

In conclusion, we have experimentally shown that Zestimate, and in extension ZHVI, is a reliable estimation of home sale prices. Furthermore, our proposed real estate investment recommendation engine is based on forecasting algorithms which manage to predict zip code areas of significant future growth with high accuracy, while largely avoiding to recommend the ones with low future growth. The CAPM financial model further validates these recommendations and adjusts them to the investment profile of each investor, based on their tolerance of risk. Finally, the recommendations produced by both the forecasted algorithms and the CAPM financial model manage to yield high investment returns and thus show that machine learning models, combined with financial ones, can indeed capture the trend of something as complex as the housing market on the zip code area level.

As part of our future work, we plan to expand our current recommendation list to include recommendations based on different geographical levels, such as country level, state level, city level and neighbourhood level. We also plan to implement ensemble machine learning or deep learning models, which may provide better results. Finally, we plan to investigate the incorporation of external data sources such as population growth, socioeconomic statistics, crime rate and restaurant health inspection data in order to further optimize the performance of our models for real estate investment recommendations.

References

[1] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso. Identifying real estate opportunities using machine learning. *Applied sciences*, 8(11):2321, 2018.

[2] A. Botchkarev. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*, 2018.

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] D. W. Draper and M. C. Findlay. Capital asset pricing and real estate valuation. *Real Estate Economics*, 10(2):152–183, 1982.

[5] S. L. Ho and M. Xie. The use of arima models for reliability forecasting and analysis. *Computers & industrial engineering*, 35(1-2):213–216, 1998.

[6] Investopedia. Alpha vs. Beta: What's the Difference?, . URL `https://www.investopedia.com/ask/answers/102714/whats-difference-between-alpha-and-beta.asp/`.

[7] Investopedia. Capital Asset Pricing Model (CAPM) and Assumptions Explained, . URL `https://www.investopedia.com/terms/c/capm.asp/`.

[8] Investopedia. Home Equity: What It Is, How It Works, and How You Can Use It, . URL `https://www.investopedia.com/terms/h/home_equity.asp`.

[9] M. Iqbal. Application of regression techniques with their advantages and disadvantages. *Elektron. Mag*, 4:11–17, 2021.

[10] J. Jiao, S. J. Choi, and W. Xu. Tracking property ownership variance and forecasting housing price with machine learning and deep learning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5175–5184. IEEE, 2021.

[11] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[12] Property Investment. Applying the CAPM to Real Estate: Alpha and Beta Estimates by Property Type. URL https://property-investment.net/2020/08/11/applying-capm-to-real-estate-alpha-and-beta-estimates-by-property-type/.

[13] L. Rokach and O. Maimon. Decision trees. In *Data mining and knowledge discovery handbook*, pages 165–192. Springer, 2005.

[14] H. Sak, A. Senior, and F. Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*, 2014.

[15] M. H. Sazli. A brief review of feed-forward neural networks. *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, 50(01), 2006.

[16] S. J. Taylor and B. Letham. Forecasting at scale. *The American Statistician*, 72 (1):37–45, 2018.

[17] C. Tofallis. A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8): 1352–1362, 2015.

[18] US Census Bureau. QUARTERLY RESIDENTIAL VACANCIES AND HOMEOWNERSHIP, THIRD QUARTER 2022, . URL https://www.census.gov/housing/hvs/files/currenthvspress.pdf.

[19] US Census Bureau. The Wealth of Households: 2017, . URL https://www.census.gov/content/dam/Census/library/publications/2020/demo/p70br-170.pdf.

[20] Zillow. What is a Zestimate?, . URL https://www.zillow.com/z/zestimate/.

[21] Zillow. Zillow Home Value Index Methodology, 2019 Revision: What's Changed & Why, . URL https://www.zillow.com/research/zhvi-methodology-2019-highlights-26221/.

[22] Zillow. Zillow Prize: Zillow's Home Value Prediction (Zestimate), . URL https://www.kaggle.com/competitions/zillow-prize-1/data/.

[23] Zillow. Zillow Research Data, . URL https://www.zillow.com/research/data/.

[24] E. Zunic, K. Korjenic, K. Hodzic, and D. Donko. Application of facebook's prophet algorithm for successful sales forecasting based on real-world data. *arXiv preprint arXiv:2005.07575*, 2020.