NANE: Neighborhood, Amenity, Noise, and Emotion feature impact in Real Estate Housing Price Prediction

Brianna Malone
The Pennsylvania State University
College of Information Science/Technology
bkm5519@psu.edu

Alvaro Tapia
The Pennsylvania State University
College of Information Science/Technology
abt5572@psu.edu

Abstract - Real estate price prediction is one of the most important challenges that society has due to the impact it has on people in every aspect. Its importance comes from the benefits of a prediction with good accuracy that would benefit potential investors, governors, and people in need of acquiring property. Many different features have been investigated with the implementation of different methodologies from AI and ML tools but not all of them convey the same results or same accuracy, investigations may have different purposes when performing price prediction. In this sense, the purpose of this project is to benefit potential buyers, by proposing an analytical combination of most of the property features available such as emotion features, noise parameters, and neighborhood variables, to train it and fit it into four different models to find the most accurate modeling method for price prediction and bring a valuable contribution. The experimentation brings us to find that XGBoost model is the most appropriate machine learning tool to determine house price prediction and that the features analyzed do bring a better accuracy than the ones tested in other research papers.

Keywords - Real Estate, Price, Prediction, Neighborhood, Noise, Emotion, ML Models, RMSE, MAE

I. INTRODUCTION

The investigation process of real estate price prediction has been going on for more than 10 years with different procedures and methods created all over the world that have been revolutionizing over the years due to the advances in technology and data analytic tools. The importance of being able to correctly predict housing prices comes from the impact Real Estate has in the current world. Presently, Real Estate investment comprises almost 10% of every country's Gross Domestic Product, in addition to that, the importance of this research analysis also comes from the erratic predicted house prices from many different platforms such as Zillow due to the lack of uninvestigated features in their tools for prediction. In this sense, accurate trend prediction can significantly benefit both government and private companies in decision-making by helping them avoid falling into unreal predictions. The usage of Artificial Intelligence (AI), and Machine Learning (ML) has made a significant impact on housing price prediction, many investigations already exist that use these tools to accurately solve this problem. Most of them only take into consideration specific characteristics and parameters based on what they plan to predict or what their goal is. However, it has been shown from past papers [1], that these investigation methods and approaches make price prediction less accurate as they do not look at the whole image of real estate property features and their importance. Many research papers [2], and [3], only use specific features for housing price prediction and they were able to achieve satisfactory results, but not good enough to be implemented in potential research purposes made by governments or private investors since they do not consider analyzing other parameters.

In this sense, this investigation is planned to use different data analytic algorithms such as feature importance and feature selection, correlation matrices, and linear models to secure the most key features from the NANE categories to increase the performance of every model that the split dataset will be fitted on in the modeling process. In the same way, to provide more options of models to increase the variety, thus there are more probabilities of finding the most accurate model among all the existing ones. For the purpose of this project, the studied models will be Neural Networks, Random Forest, Decision Trees, Logistic Regression, and XGBoost Regression, which will help to determine to what extent these features bring a higher accuracy in terms of RMSE and MAE parameters and also to identify which model is the best for real estate house price prediction to later compare it and conclude if these findings are better than from other research papers and parent paper.

II. LITERATURE REVIEW

In this “Literature Review” section, we will be discussing in detail each of the three (3) research papers that were introduced at the beginning of the document. We will discuss the paper's relevance with the topic of real estate price prediction and the results found in each paper. To build upon the results found, discuss how to further the information revealed about real estate price predictions and how it benefits different people.

[1] PATE: Property, Amenities, Traffic, and Emotions Coming Together for Real Estate Price Prediction

As mentioned in the report, the first research paper [1] that we are reviewing will be our parent paper, in which we will replicate the results of the paper later. This research paper was published in August 2022 and written by a group of students at the University of Hong Kong (HKU). This paper [1] uses various data sources like Badidu Maps, microblog posts, and transaction data of real estate from the internet to create the intended machine learning model(s).

The data can be split into four (4) categories, Property, Amenities, Traffic, and Emotions.
Property information was collected from a collection of real estate transaction data. They used the property features such as the number of rooms, number of bedrooms, etc. Amenities information was given with the use of geographical coordinates for each real estate property. They used Baidu Maps to help identify surrounding amenities within a kilometer radius of the property. Baidu Maps is a Chinese desktop/mobile web mapping service application provided by Baidu, offering satellite imagery, street maps/views, and indoor view perspectives. The total number and average distance of each of the surrounding amenities were features of the real estate. The traffic feature also uses Baidu Maps as well to obtain traffic information. For each real estate property, traffic speed information is used to reflect the efficiency of transportation surrounding the real estate. Emotion features are used to look at responses collected from a microblog as they relate to each real estate property. The researcher then ran an analysis algorithm to analyze the emotions from each response post. The posts are then classified by types: detest(dislike), happiness, anger, fear, and sadness. A calculation percentage of each emotion related to emotion features.

In total, 27 features were analyzed from the collected multi-source data. To build models, they intended to perform two different methods, XGBoost regression, and linear regression. To complete this, you must train and test via both linear and XGBoost.

The researchers found when compared with only using property features, adding other features (traffic, emotions, or amenities) could enhance the model's performance. We find that utilizing all features from various characteristics (property, amenities, emotions, and traffic) attains the best results. The removal of amenities from the model shows the most important performance drop in comparison to removing emotion or traffic. This suggests that amenity features have a more significant impact on the model. Removing the traffic characteristic has a small effect on the performance. It is reasonable since the traffic feature has one dimension.

[2] Does Noise Affect Housing Prices? A Case Study in the Urban Area of Thessaloniki

The second research paper [2] we are reviewing was published in February 2023 and written by a team of researchers from the Aristotle University of Thessaloniki in Thessaloniki, Greece. This paper [2] uses aerial maps, and heat (colored-coded) maps to provide a mapping of noise ranges. The goal of this research is to see whether there is a correlation or connection between housing prices and noise level (decibels).

The researcher constructed a noise pollution dataset based on research performed and published by the Hellenic Ministry of Environment and Energy in the town of Thessaloniki, Greece. The goal is to investigate the correlation and effect of noise on housing prices and needs a feasible dataset. For this paper, researchers applied Openhouse. Openhouse is a real estate platform working in towns in Greece. It contains useful information on an extensive range of property features.

Researchers used a method referred to as Georeferencing, which tries to relate virtual graphics of maps, like the ones within the noise studies, to a ground device of geographic coordinates. Geographic Information System (GIS) software program equipment is used to carry out this task. Begin training data using machine learning models of XGBoost and light gradient boosting on property data for distinct areas of Thessaloniki to analyze the way noise affects price estimations through model techniques. To check out the correlation between housing price and noise, tree-based models specifically use choice random forests, trees, XGBoost, and light gradient boosting.

The goal is to provide a new noise pollution dataset that indicates the impact noise has on housing prices but also reveals why the effect of noise on prices significantly differs amongst different regions of the same city. A new noise dataset, as well as a new housing price dataset containing noise information for Thessaloniki, can be created after a model has been trained and reliable.

The outcomes have been displayed as a heatmap, in which discrete colors represent various levels of noise range of 5-decibel durations. In every municipality, the noise is segmented into day and night noise. The data obtained the noise factoring in both aviation disruptions and traffic. The dataset will consist of the noise (decibels) of a real estate property given its latitude and longitude coordinates to the corresponding area on a geographic map. The use of XGBoost and LGBM models verifies that noise impacts prices and it can influence a few places undoubtedly even as others negatively. Property prices grow in the town center and places close to the vicinity as noise increases. This can be attributed to the prevalent commerciality of the region. In evaluation, homes placed far from the center are impacted negatively utilizing noise.

We can improve our opinions from our parent paper [1] which uses additional features of noise concerning property location that is mentioned in the paper [2]. Through this research, we can explore further housing questions to better understand what factors into housing predictions.

[3] The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment

The third research paper [3] was published in 2018 and written by a group of researchers from the University of Trento in Trento, Italy. This paper [3] uses data from 8 Italian cities from various sources to determine if neighbor and property features have an impact on property prices. Neighborhoods are defined in the research as geographical units where activities and social interactions occur the most. The property contains a brief description and a list of characteristics about the property. Examples of some of the neighborhood features evaluated are proximity to airports, security perception, population, and proximity to industrial areas to name a few.

The analysis was formed as a multi-model problem to predict the price of houses collected from a dataset of real estate properties and added neighborhood features based on overlapping geographical boundaries. The data included in the study are home listing, geographical information (OpenStreetMap, Urban Atlas, Google Street View), and census data. In the research, they are curious to understand the economic impact of neighborhoods on housing values through a predictive model based on XGBoost. XGBoost allows for features' importance and positive/negative contributions of each feature to the predicted values. The researcher trained the model by using a K-fold cross-validation schema and gradient boosting to make sure that the model is strong in detecting unseen neighborhoods and houses. To start, they divide the original dataset into five folds, assigning three to the training set, one test set., and one to the validation.

Results indicate that when results imply that after being taken collectively, the results display the economic result of a neighborhood on housing value. Among those models used, the neighborhood features extensively improved the results, losing the percentage errors using 60%. The maximum vital organizations of predictors belong to neighborhood metrics, whilst asset predictors account for much less. We can improve and further our opinions from our parent paper [1] that uses additional features that are different from the paper [3]. Through this research, we can explore further housing issues such as affordability and gentrification.

III. NOVELTY/CONTRIBUTIONS

For this project, the newly taken approach concerned gathering and selecting the best parameters studied so far in Real Estate identified from the different resource papers [1], [2], and [3]. We took inspiration from every feature studied in each referenced investigation, concerning neighborhood, noise, and emotion parameters, respectively. After a vast indication into every dataset from each paper and other documents posted on different platforms, we took notice that the dataset used and implemented in the parent paper [1], already has some similar features that were studied on the other resources which can be easily segmented for the development of this project. In this sense, besides working on a different approach by analyzing how the combination of all the features investigated in past papers can return a more accurate house price prediction, our contribution can also be reflected in the different tools and machine learning models implemented in this research paper to provide a different approach and potentially find better results in price prediction.

As a first instance, the initial contribution comes from the execution of the Random Forest Classifier (RFC) tool to calculate the feature importance of every parameter in the dataset by comparing it with the price of every Real Estate property. The purpose of using this rather than implementing the F-Score tool used in the referenced parent paper is due to the effectiveness RFC has when handling multicollinearity because of the number of parameters in the dataset that are better modeled. In the same way, this instrument is also less prone to have issues with overfitting, compared to the F-Score and other models such as decision trees that are utilized for the same purpose.

Going along with the methodology process of this investigation it should be noted that differently from the parent paper, while the authors used all the parameters without taking into consideration their F-Score results, for this project it was chosen some specific parameters based on the feature importance results, consequently dropping insignificant parameters from the dataset, which showed a very low importance correlation when comparing it to the price of each property.

Subsequently, when creating the correlation matrix for this paper, it differs from the parent paper's results due to the elimination of unwanted categories that may reduce the accuracy of certain parameters when modeling the data. This can begin to be evidenced in this correlation matrix, in the parent paper code [4] it is seen that most of the features show a high correlation that is at least above average. This is a current improvement that we are contributing to this paper as in past resources (parent paper), most parameters showed a low score in their relationships versus other variables, see our final code [5]. Finally, for this specific paper, it was decided to include in the analysis 3 other different models that have not been studied before in either of the past investigated research papers [1,2,3]. These models are Neural Network, Random Forest, and Logistic Regression; however, we also implemented the other used models such as Decision trees and XGBoost Regression to potentially find better results in terms of accuracy and compare them. In this sense, our goal is to increase the machine learning model options that are available and were once created to find the best model possible by analyzing the RMSE and MAE of each of them, comparing them, and producing potentially better results in comparison to past papers.

IV. METHODOLOGY

For this project, it was developed many different data analysis tool measures as well as machine learning models to help determine the best potential features for the modeling process and explore the diversified results to find the best accurate model in terms of validation effectiveness. For this purpose, the methodology process was divided into three main blocks: A) Data analytic tools for feature importance and selection, B) Machine learning models creation, and C) Development of data visualization tools for ML model comparison.

A) The dataset selected for the research paper was already cleaned by previous authors who utilized it for other purposes, so the procedures that had to be made to begin analyzing the dataset was to perform feature importance, and feature selection.

I. **Feature importance:** Differently from the parent paper, for the purpose of this research it was implemented Random Forest Classification to determine more accurately the importance each feature from the dataset has in the price of every real estate property. The main reason this was done is due to the higher effectiveness Random Forest in comparison to other tools such as F-Score. The main difference is that this tool handles multicollinearity and is less prone to have issues with overfitting, which can be an issue due to the size of the dataset.

II. **Feature selection:** Firstly, before doing any analysis, it had to be removed the "id" parameter because it was completely unnecessary to have in the datagram since we are not planning to analyze every sole property, we are doing a whole study and all properties. After doing this, it was developed a correlation matrix using the seaborn package available in Python 3.0 which accurately creates a good-looking visualization of the relationships between every parameter so we can better determine what variables have a higher significance in

the study of the project which would also mean that when choosing these features for the modeling process, it will also return a probable higher accuracy than when using all parameters such us in the parent paper [1]. Therefore, the least significant features were discarded from each area of the NANE parameters.

B) In regard to the machine learning modeling process, it was decided to develop four different models for the investigation of this paper, these were Neural Networks, Decision Trees, XGBoost, and Logistic Regression. Every model followed certain hyperparameters to bring the best performance from each of them. It was used four models to expand the variety of models so we can have more models to test. It is important to note that before implementing any model it was necessary to split it with the following guidelines: Train set with 75%, Test set with 10%, and Validation set with 15%. In this sense, these split sets will be trained and modeled in the following machine learning models.

I. **Neural Network Model:** The first intentional model developed for this investigation was a NN model due to previous demonstrated effectiveness in past research papers. The hyperparameters used for this model were 64 units as an input layer, a 32-unit layer for the hidden layer, and a 1-unit output layer for regression. It should be noted that the package used for the coding was TensorFlow's Keras API. In addition to that, the model was trained for 50 epochs and with a batch size of 32 for the training set. The predicted validation and test values were later tested with their respective parameters to find this model's RMSE and MAE.

II. **Decision Trees Model:** It was also implemented decision tress model due to the constant application it mostly has in other papers due to its easy usage in real estate price prediction. The hyperparameters used for this model consisted of setting the random state of the model with 42 which ensures reproducibility. This is important because without a fixed seed the model can produce inaccurate results due to randomness. Like the NN model, this process was also tested in the test set and validation test to measure its accuracy.

III. **XGBoost Model:** Similarly, to Decision Trees, the XGBoost model also utilized a random state parameter set to 42. The main characteristic about this model is that it helps to easily handle overfitting when handling missing values of the dataset, as well as its efficiency when working with large datasets. The split dataset was later fitted into this model to find the RMSE and MAE accuracy parameters.

IV. **Logistic Regression:** The last implemented model was logistic regression; the main reason of its application was to vary from the linear regression found in the parent paper to find if it brings better accuracy than other methods**.** Like the last two models, the logistic regression set a 42 random state, with a same sequence during each run. To code this model it was also needed to make use of the scikit-learn package available in Python 3.0 which gathers programming information of most of the machine learning models. Like previous models it was also found the RMSE, and MAE parameters.

C) Finally, it was developed different data visualization tools to compare the results of each model and help determine the best machine learning model that has the best accuracy in terms of RMSE and MAE that will later help to generate effective conclusions.

I. **Linear Graph Models:** For every developed model, a scatterplot with its respective trend line was also created to visualize the differences between the predicted prices and the actual prices so we can identify in a visual way how effective each model was. Where the closer the dots were to the trend line, the more accurate the model is to determine house price prediction. These plots show all discrepancies between each value dot and underline the patterns of the data. This step in the methodology process is crucial because it will be used to compare it with the results from the parent paper [1].

II. **Comparison Table:** For better interpretation, a table was also created that contained all the information of the accuracy measure tools such as RMSE and MAE from every single model. To facilitate the recognition of the best model that has the best accuracy among all the models, it was highlighted with a blue color.

V. IMPLEMENTATION & RESULTS

For the implementation section, it will compare the results gathered from the parent paper [1], and the results achieved during the investigation process to evidence our contribution and demonstrate the difference in achievements. The results achieved from the parent paper's code [4] give a thorough explanation and analysis of various aspects that play a role in house price prediction. This section will primarily analyze the achieved code for this research paper [5] to expand this study about the impact NANE features have in the prediction of house prices.

As a first instance, the result of the original parent code [4] provided multiple plots that detailed the results for which variables had a greater contribution to house price prediction which focused primarily on Beijing, China. One of the initial codes from that original parent paper code produces is that of the feature selection process which is detailed in a bar chart created after a coding process using F-Score. The main implementation achieved for this paper was to take a different approach by using a Random Forest Classifier to achieve a better result. As it can be evidenced in both graphs (view Fig. 1 and Fig. 2), the results from the feature importance are slightly different from one another where but this is efficient enough to determine what features to drop from every NANE section, which were the lowest ones from each area. We also tried to interrelate both graphs to see if some low-value parameters were repeated to be classified as low-value variables to also discard them from the dataset.

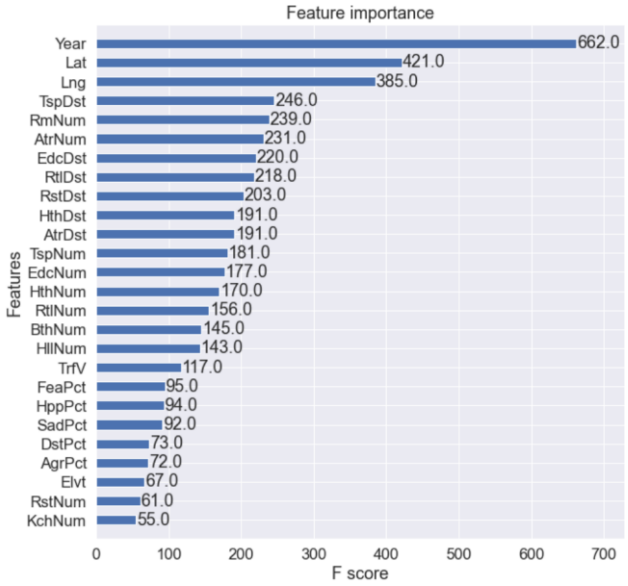

Fig. 1 F-Score Feature Importance Bar Plot (Parent Paper [1])

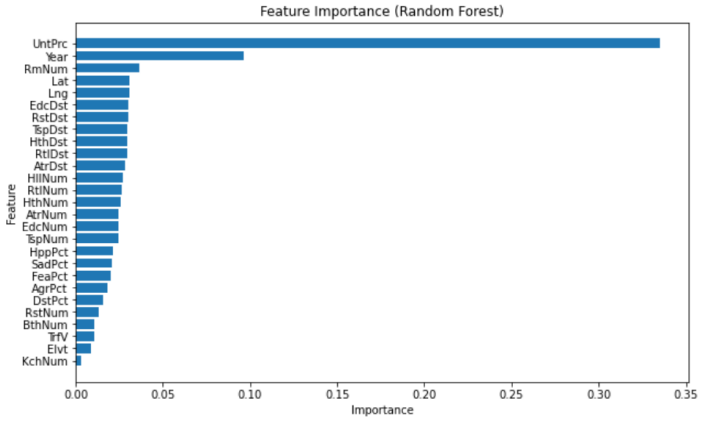

Fig. 2 Random Forest Classifier Importance Bar Plot

The bar plot produced using XGBoost provides an F-Score ranging from 0 to 1. Higher values indicate better performance; crucial to ranking each of the variables. From these results, we narrowed down the list of variables to focus on that had a greater impact on price valuations. We decided to reduce the number of amenities because it is ambiguous to say how many amenities are in a city, distance is more important. Only including the number of tourist attractions because that is the only amenity that has an impact. Removing some emotion parameters (the least important) based on feature importance. When done filtering, we are left with 19 variables to train, test, and validate.

The correlation matrix uses a heatmap to see relationships between each of the final features selected. The relationship between features is measured by R-value correlation. The R-value is a number range between −1 and 1. The closer to 1, the greater the positive correlation, closer to −1, the more negative correlation. In this sense, the variables dropped were: 'Elvt', 'KchNum', 'HllNum', 'EdcNum', 'HthNum', 'RtlNum', 'RstNum', 'DstPct', 'AgrPct'.

To validate the right selection of features the correlation matrix for both the parent paper and this project was performed to find potential relationships between parameters and corroborate its importance in the investigation. The following graphs were developed.

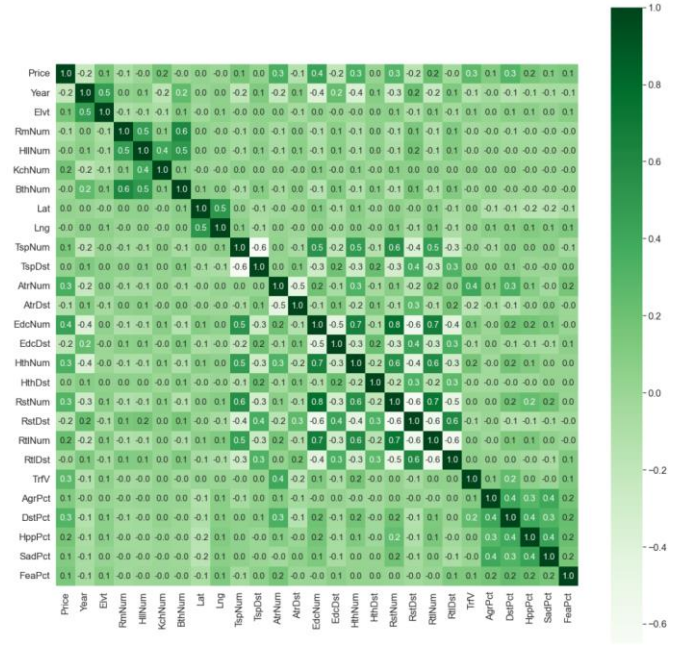

Fig. 3 Correlation Matrix (Parent Paper)

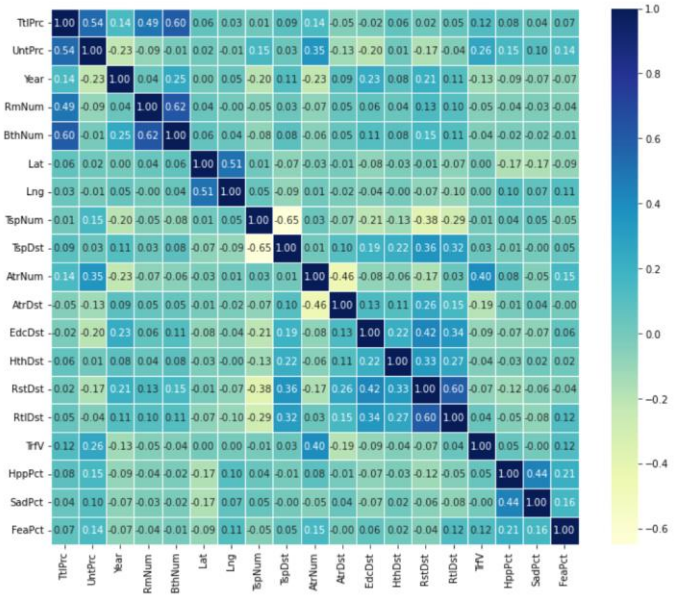

Fig. 4 Correlation Matrix

The correlation plot is presented as a heat map that displays the correlation between multiple variables as a color-coded matrix. It is like a color chart showing how closely related different variables are. As it is possible to evidence, the coded correlation matrix in Fig. 4, done for this project has fewer parameters to analyze due to the feature selection process. We can see that this slight modification of reducing unwanted features brings a higher correlation between each parameter, while if we look on Fig.3, most of the squares with values have a colored box that shows a low correlation. It should be noted that this is only a visual corroboration of the importance of the features, therefore the actual process that demonstrates the process are the models where the split dataset is fitted and trained.

For the machine learning implementation, we want to look deeper into the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Both metrics are used to measure error sizes between predicted to actual values. The higher the value for both metrics, the larger the error. The goal of machine learning is to test the data on different machine learning devices to see which produces lesser error. The smaller the RMSE and MAE the better performance for the model.

Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$

Fig. 5 Accuracy Measurements Formula

As it was explained previously, it was decided to run a machine learning analysis on four different types of models using the same filtered data set (19 variables) in our new code [5]. The four machine learning methods tested were Neural Network, Decision Tree, XGBoost, and Logistic Regression. The results comparison was developed in two blocks, the first one is a made data visualization by creating linear graph models for each respective model that shows the effectiveness of the model based on the trend line and point values. In this sense, every model had a respective linear graph that showed the actual relationship between real prices and predicted prices based on the validation method process of the model. In total there were four different graphs, one for each machine learning model.

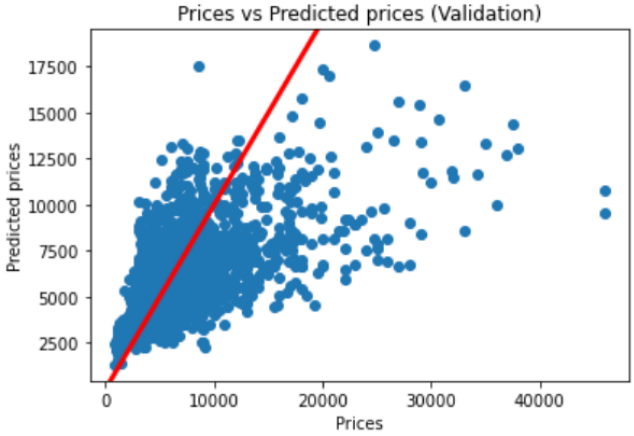

Fig. 6 Neural Network Linear Graph Model

As it is possible to visualize in the Neural Network graph (Fig. 6), most of the data points are dispersed all over the graph area, where in most cases the created trend line for this model is far away. This demonstrates itself to be a very inefficient model besides all the attempts made to choose the best hyperparameters for this specific model to try to increase its accuracy. This model's interpretation could be that either the features chosen should not be worked on this type of machine learning model, or that using Neural Network is inefficient when studying real estate price prediction.

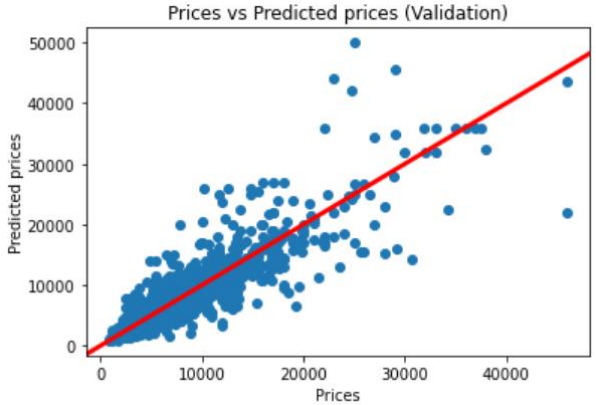

Fig. 7 Decision Trees Linear Graph Model

Here we can prove that when modeling the split data using Decision Trees, there is some improvement visualized, where most data points are closer to the trend line. However, there are still some random errors or unexpected results that can be represented in the cumulative amount of data points that are away of the trend line. With this we can interpret that even though there is an improvement in the modeling process, there is still a low accuracy rate.

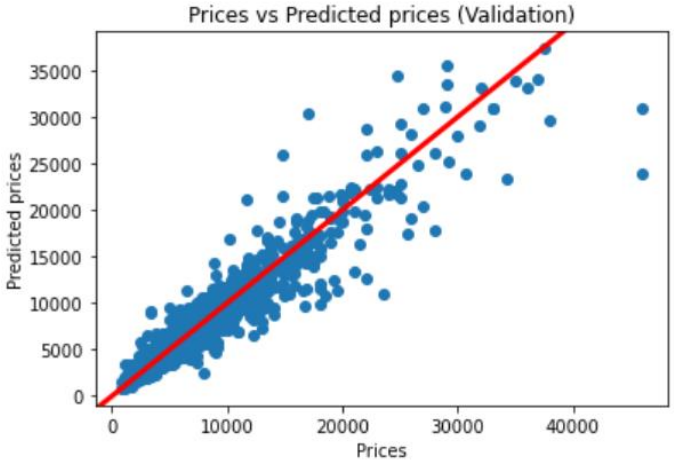

Fig. 8 XGBoost Regression Linear Graph Model

In this case, XGBoost Regression demonstrated to outperform all the previous fitted models, as it is evidenced that most data points are closer to the trend line, and differently from the Decision Tree model, here there is a less cumulative number of data points away from his red line visualized in the graph. It can be assured that it follows a strict and positive correlation between real estate predicted prices and actual prices.

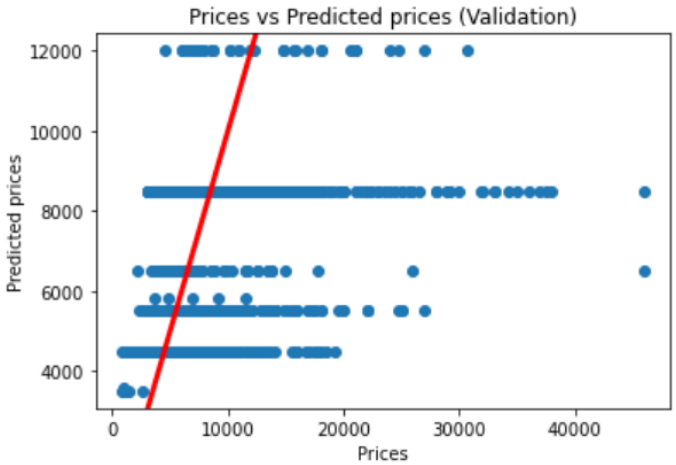

Fig. 9 Logistic Regression Linear Graph Model

Finally, the Logistic Regression modeling gives an inaccurate result that cannot be interpreted. The main reason for this is because this type of model is not suitable for regression tasks, and using these metrics will most of the time result in misinterpretation of the fitted data. In this sense, it can be argued that the usage of logistic regression modeling for real estate price prediction is unnecessary due to its inefficiency.

Fig. 10 XGBoost Regression Linear Graph Model (Parent Paper)

When taking the parent paper into consideration [1], it can be visually evidenced when comparing Fig. 8 with Fig. 10 that for both cases the best machine learning modeling process is XGBoost. However, from this visual perspective it can be identified that the created graph for this research paper has closer data points to the trend line than the resulting graph from the parent paper [1]. In this sense, even though the model type is the same, the main difference from the dataset was the elimination of non-important features at the beginning of the project, which is now the cause of this difference in the accuracy of price prediction showing a better result from our end.

In order to corroborate the results given by the previous data visualization tool, we should analyze the second block of data processing, where it was created a table (Table 1) that groups all the information from the results of every model, consisting in the output of the validation accuracy metrics known as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

	Model	RMSE	MAE
0	Neural Networks	2111.851460	3560.549775
1	Decision Trees	2036.076236	986.976190
2	XGBoost Reg	1399.420120	766.125927
3	Logistic Reg	3944.891032	2324.963585

Table 1. Comparison Table of Results

From the table above, we can see that XGBoost regression gives the lowest MSE and RMSE values compared to other models used in our implementation, by having the lowest values of Root-Mean-Square Deviation (RMSE) with a value of 1399.42, and a Mean Absolute Error (MAE) of 766.13. In this sense, it can be argued that our implementation yielded the same results as the parent paper where it is showed that XGBoost models are the best to use with training/testing our housing price prediction.

However, when comparing the actual quantitative results to the RMSE and MAE of the parent paper [1] that can be seen in Table 2, we have that our results from this project are better since they show to have a lower RMSE and MAE. When technically comparing both tables, we have that our RMSE and MAE is 1399.42, and 766.13 respectively, while the ones from the parent paper [1], are 8829, and 5721, respectively. In this sense, since the accuracy matrices values achieved from this paper are lower than the accuracy parameters from the parent paper [1], which means that the average magnitude error is less, and therefore the model accuracy with the features selected outperform other models developed in the past and from other research papers.

Data	Method	R^2	Adjusted R^2	MAE	MSE	RMSE
Training set	Linear regression w/ only P	0.1674	0.1671	18284	551713399	23489
	Linear regression w/o A	0.2520	0.2515	16947	495668829	22264
	Linear regression w/o T	0.3730	0.3722	15499	415469247	20383
	Linear regression w/o E	0.3636	0.3629	15582	421702012	20535
	Linear regression w/ PATE	**0.3797**	**0.3789**	**15391**	**411032381**	**20274**
	XGBoost regression w/ only P	0.9095	0.9095	5206	59965069	7744
	XGBoost regression w/o A	0.9184	0.9183	4941	54069367	7353
	XGBoost regression w/o T	0.9331	0.9330	4416	44350499	6660
	XGBoost regression w/o E	0.9319	0.9318	4477	45145802	6719
	XGBoost regression w/ PATE	**0.9343**	**0.9342**	**4387**	**43549356**	**6599**
Testing set	Linear regression w/ only P	0.1626	0.1618	17905	530648157	23036
	Linear regression w/o A	0.2437	0.2424	16696	479250332	21892
	Linear regression w/o T	0.3591	0.3572	15324	406118449	20152
	Linear regression w/o E	0.3510	0.3494	15358	411213070	20278
	Linear regression w/ PATE	**0.3651**	**0.3632**	**15235**	**402302484**	**20057**
	XGBoost regression w/ only P	0.8560	0.8558	6244	91267181	9553
	XGBoost regression w/o A	0.8646	0.8644	6080	85802314	9263
	XGBoost regression w/o T	0.8751	0.8747	5773	79153827	8897
	XGBoost regression w/o E	0.8740	0.8737	5814	79830419	8935
	XGBoost regression w/ PATE	**0.8770**	**0.8766**	**5721**	**77956264**	**8829**

Table 2. Comparison Table of Results (Parent Paper)

VI. CONCLUSION & FUTURE WORK

Real Estate now more than ever has been is being one of the main targets for families, investors, and the government, due to the impact it has, as it is not only 10% part of the GDP from every country, but also every person needs a place to live and therefore investments can be made to generate income by either selling properties with the right price, or buying them knowing the potential price it could have. Many different platforms have tried to establish a price or value to every property at least in the United States with the option of Zillow or Trulia. However, in many cases there are concerns about the accuracy of those prices, that have led to many circumstances where prices were completely inaccurate to the actual house price. Many researchers have also tried to develop different ways of using machine learning models for real estate price prediction, but many different variations exist so that it has been difficult to define what is the best model for this field. For this purpose, this project was developed to combine the best information from previous research papers to use machine learning to develop the best model possible with the best features that need to be studied when trying to predict house prices.

In this paper, it was found that after proceeding with different experimentations of machine learning models, it was possible to corroborate along with the parent paper [1], that XGBoost is the best model when performing house price prediction, this was evidenced based on the data visualization from of the linear graph and the quantitative results that show a low RMSE and MAE. In the same way, it is possible to conclude that the selected features from the NANE parameters were the correct and adequate ones and should always be considered when predicting house prices due to the impact it has when fitting them into the model. NANE features are more impactful than regular amenity features which have improved machine learning models performance.

We hope that this work was essential for many researchers and investigators that way want to implement the same processes but with a different dataset or try to improve the actual findings to bring better results to benefit others and use this as an inspiration. We also found potential investigations that could be done in the future such as combine these textual features with image features to validate better the results, also potentially investigate how this process can be utilized also when analyzing the time variable such us performing price prediction but for the future in time, and finally it could be implemented the variable of users reviews on properties to also use it as an additional feature that may bring better results than the achieved.

VII. REFERENCES

[1] Zhao, Ying-Zheng, et al. “PATE: Property, Amenities, Traffic, and Emotions Coming Together for Real Estate Price Prediction.” arXiv (Cornell University), Cornell University, Aug. 2022, https://arxiv.org/pdf/2209.05471v2.pdf

[2] Kamtziridis, Georgios, et al. “Does Noise Affect Housing Prices? A Case Study in the Urban Area of Thessaloniki.” arXiv (Cornell University), Cornell University, Feb. 2023, https://arxiv.org/pdf/2302.13034v1.pdf

[3] De Nadai, Marco, and Bruno Lepri. “The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment.” 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, Oct. 2018, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8631423

[4] abt5572. “DS340W-Group10/DS 340W Parent Paper - Code.ipynb at Main · Abt5572/DS340W-Group10.” GitHub, 2023, github.com/abt5572/DS340W-Group10/blob/main/DS%20340W%20Parent%20Paper%20-%20Code.ipynb. Accessed 23 Nov. 2023.

[5] abt5572. “DS 340W Final Code.ipynb at Main · Abt5572/DS340W-Group10.” GitHub, 2023, github.com/abt5572/DS340W-Group10/blob/main/DS%20340w%20Group%2010%20Final%20Code.ipynb. Accessed 23 Nov. 2023.