Multivariate Real Estate Sale Price Analysis

Abeer Mathur [#1]

Parag Patel [#2]

[#]*College of Information Sciences/Technology, The Pennsylvania State University, United States of America*

[1]aqm6537@psu.edu

[2]pzp5254@psu.edu

Abstract— This paper will cover research led by multiple sources regarding housing price analysis. This paper aims to provide insights that can guide real estate investors, lenders, and stakeholders in California. Using machine learning models, we analyze to predict property prices accurately from the three counties, Los Angeles, Orange, and Ventura, California. The data has over fifty (50) plus house features. This paper reviews three (3) research studies that contribute to this paper, with a focus on multivariate real estate price analysis.

--- Furthermore, we will be focusing on all three (3) research papers on their findings and methodologies. We will be focusing on the first research paper, which offers insight into real estate price analysis. This paper focuses on the features in and around the properties. The methods used include Ridge and Lasso regression, Neural Networks, Support Vector Machines, which provide insight into predicting real estate prices.

Keywords— XGBoost, Linear Regression, Hyperparameter Tuning, Random Forest, Support Vector Regression, Neural Networks, Support Vector Machines

I. INTRODUCTION

In this paper, the goal is to explore data from 2016 and identify key features that can be used to predict real estate sale prices for future years in 3 California counties. Our research draws from 3 main research papers. The first paper is known as the "parent paper." To provide context for the upcoming literature review, we will start with a brief overview of the three (3) research papers, which we will go more in depth in the "Literature Review" section.

The primary research paper [1] we are focusing on is called "PATE: Property, Amenities, Traffic, and Emotions Coming Together for Real Estate Price Prediction." This parent paper is our primary point of reference and provides the overall structure of the study that we conducted. In the next section, we will examine the content of the paper. We will find the methods, outcomes, and the potential significance to help for our exploration of multivariate real estate sale price analysis.

The first research paper [1] that we will be discussing will be our "parent paper." The parent paper is titled "PATE: Property, Amenities, Traffic, and Emotions Coming Together for Real Estate Price Predictions." This paper finds the impact of various characteristics on real estate prices and predicts real estate prices in Beijing, China. The unique features emphasized in this study comprise of amenities, traffic conditions, and social emotions. It uses unique features to find pairings of amenities, traffic conditions and other socioeconomic conditions to find the strongest correlation between the features and real estate prices. The results show that traffic conditions, proximity to certain environmental features like nearby tourist attractions, and social emotions have an influence on the prices. This paper used XG Boost and Linear Regression to accurately predict real estate prices based off the data. The models analyze the unique features to make predictions.

The second research paper [2] that we will be discussing is called "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis." This paper focuses on using machine learning models for rental properties. The goal of this paper is to help property owners and Airbnb customers use data to evaluate rental prices. The dataset used in this paper is a public Airbnb dataset for New York City. The authors also use sentiment analysis from reviews to improve the accuracy of the model. They also use feature selection to identify the most predictive features. This study, unlike the previous ones, concluded that only a few features, property specifications, owner information, and customer reviews on the listings, had any impact on the actual sale price unlike the 50+ features incorporated in the model above.

The third research paper [3] that we will be discussing is called "Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study." The authors focus on using machine learning models for the real estate market. They use a dataset ranging from 2010-2019 and using socio-economic factors. The authors use multiple models to foresee if the closing sale price is going to be greater than the listing price. This research can help real estate investors, bankers, and etc. to make better decisions.

Our interest in the research topic of using machine learning to gain more knowledge about real estate price prediction prices came from both of us wanting to do a topic related to real estate. After some research was done, we chose this topic. We did this as we are in a data science course and doing this paper for this course. We are reviewing and learning from research papers related to our topic. We have seen other students do prediction topics, medical related topics, and sports related topics. We found these research papers because we believe they will help us best. Their methods helped us on how to approach our method and with the knowledge captured from these papers helped with choosing our dataset.

Our purpose for focusing this paper on real estate price prediction is to get accurate results on the prices within the dataset. This can help people learn more about the real prices of real estate. The real estate price prediction can help real estate investors, the owners, and customers. The different features help with this prediction. This can inform people involved in real estate that the house features and other features help with the sale price.

II. LITERATURE REVIEW

In this "Literature Review" section, we will be discussing each of the three (3) research papers that we introduced in the prior section of this document. We will discuss why this paper is based on these research papers. We will go over the techniques, datasets, and the results of each paper.

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

The first research paper [1] we are discussing was published in 2022 and written by a bunch of authors. The summary of the paper is to take real estate prices using unique features. The different types of features they use are property, amenities, traffic, and emotions. They use different data sets to evaluate the different features. They used 28,550 houses in Beijing, China.

The goal is using socioeconomic patterns to predict real estate prices in China. The reason to undertake this work is to use those socioeconomic patterns to predict real estate prices not only in Beijing but China as a whole. Real estate contributes to a large percentage of the GDP (Gross Domestic Product) in China. It also maximizes long term return for tenants and owners. Proximity to amenities and public transport should be considered. Also, current models are most effective, but SVM or NN can be employed. Negative impacts of gentrification such as forced displacement, discriminatory behavior by people in power, and a focus on spaces that exclude minorities. The authors formulated a multimodal to predict a house price prediction using different types of features.

The researchers referenced their past work on creating a Recurrent Neural Network that used images of houses sold in certain neighborhoods. They also had research that was performed to find the effect of socioeconomic and environmental factors on real estate prices.

The methodology used in this paper was Linear regression and XG-Boost regression. The use of both of those methods helps improve the prediction accuracy for housing price prediction. XG Boost is the most effective model to use with the highest test scores. We can learn that XG Boost regression with PATE (Property, Amenities, Traffic, Emotions) had a high R2 score as well as low scores for MSE and RMSE (Root Mean Squared Error) (Root Mean Squared Error). You can tell this is the best model to use since linear regression had a lower R2 score as well as a higher MSE and RMSE score. They used all the features to improve the performance. They also tested it with individual features to see what had the best performance, which was the traffic feature.

We can conclude that XG Boost regression model is the best to use. We also like to conclude that it is important to use several types of features in your dataset like this paper. It helps the model to achieve the best performance. Proximity to amenities and public transport should be considered since this is for the owner and tenant. We should also consider the impacts of gentrification in certain areas. This can impact the result since there are forced displacements and people with power having a certain type of behavior.

The focus on this paper is the features they used. They did not just use property features, but they used amenities that are near the property, they used how the traffic is around the property, and social emotions in the area to see if it is a good place for someone to live in.

Some suggestions we have after reviewing this paper suggest using Ridge and Lasso and Neural Networks. We also think that dataset features are extracted through region characteristics analysis. We also can use F1 scores instead of an F2 score for model performance.

[2] Airbnb price prediction using machine learning and sentiment analysis

The second research paper [2] we are reviewing was published in July of 2019 and written by people associated with Stanford University. The goal of this research paper is to figure out how many customers for their place and to figure out a reliable price using machine learning. The reason for undertaking the work is to help property owners and the customers with the pricing aspect. Methods used in the research paper are feature selection and the one with the highest r squared score was lasso feature selection. We found out that Airbnb renters can now make more informed conclusions. Lasso-based feature important analysis reduced the variance but hidden factors like biases are impossible to consider. The problem is to develop a reliable price prediction using machine learning, deep learning, and natural language processing techniques. SVR had the best R square score of 69%. This helps us know which model to use to come with the Airbnb prices. This paper is important to discipline because it used the right ML methods to produce the solution. It also helps the consumers with the pricing.

The research paper used linear regression, regression tree, random forest regression, and gradient boosting regression trees to analyze warehouse rental prices in Beijing.

The methods used were data acquisition, data processing, feature selection, modeling, and validation. For feature selection they used lasso regularization, p - value, and r squared score with lasso having the highest r squared score. Using SVR was the best method for this research. SVR helps to find a hyperplane that best fits the data points while maximizing the margin between the data points. Compared with K Means + Ridge, partitioned training into distinct clusters and applied ridge regression to each cluster.

The conclusion to be drawn is that SVR is the best model to use because of its R squared, MAE, and MSE scores. Lasso feature selection was the best out the rest since it had the highest r squared score. This helps any Airbnb renters and owners because they can produce the best possible price to rent out depending on the variables used and the unique features.

Some of the suggestions we have are using Random Forrest to better compare it to see how much the regularization helps. We would also suggest adding more data from sites like VRBO, another website like Airbnb where you can rent out a place, to boost the performance of K-Means. Also, since in NYC there are different property sizes and different attributes, it is critical to categorize each and call for distinctions.

[3] Machine learning approaches to real estate market prediction problem: A case study

The third research paper [3] we are reviewing was released in August of 2020. The goal of this paper is to classify whether the property sale price is higher or lower than the appraised price using ML techniques. This paper is important because it addresses a research gap and gives new knowledge to this certain field. The value is valuable to stakeholders, including real estate investors, mortgage lenders, financial institutions, and first-time homebuyers. The data they used is from Florida's Volusia County Property Appraiser website, GDP, Consumer Price Index, Producer Price Index, House Price Index, and Effective Federal Funds Rate. The research methodologies used were data acquisition, data processing, feature selection, modeling, and validation.

The authors had referenced their past work which was using data from Fairfax County and other government websites. They also referenced forecasting the US (United States) real house price index, and housing price forecasting based on genetic algorithm and SVM.

Their methodology was using Ridge and Lasso as well as Neural Networks rather than XGBoost and several types of variables that

are used to predict the house price. The paper finds that using XGBoost classifier in predicting the classification of classes 0 and 1. When training on selected features and socio-economic factors, it demonstrated a high level of accuracy and effectiveness in classifying the two classes. The performance metric for both classes fell between 92.6% and 94.3%. The high F-1 score of 93.5% for both classes let us know the importance of the model in terms of precision and recall.

In conclusion, the model demonstrated effectiveness in distinguishing between the two classes based on the features they provided. The findings on this paper have potential applications for businesses and the government. The results from the model can help make informed decisions.

Some of the suggestions we had were about using Ridge and Lasso as well as Neural Networks. Dataset features extracted through region characteristics analysis. Properties vary by type, and it is critical to categorize each and call for distinctions.

III. CONTRIBUTIONS

Now we have talked about the research that we have conducted in the past sections, especially the LITERATURE REVIEW. Since seeing the results of the three (3) research papers, we wanted to implement more into our research. The next few sections will be about what our original research plan was, what actually happened, then the new plan for implementation.

A. ORIGINAL PLAN FOR IMPLEMENTATION

The original plan we had for implementation was to focus on the original data set that we had found. While searching for a dataset, we wanted to focus on having a lot of features. The one we found had 50 plus features. With that data set it was focused on Ames, Iowa, and the year 2006 to 2010. The issue with the years is that it is too old and not new so it would not really help with house price predictions in 2023 and upcoming years. We also had a plan for modeling where we focus on Linear Regression, XG-BOOST Regression, Random Forest Regressor, and SVR. When we had originally made our feature matrix, it was a lot, and we knew we had to reduce it. Our plan for the code part was to focus on the parent paper's code. We wanted to implement the same way they did but with our analysis and data.

B. WHAT HAPPENED

Since we knew that we did not like the dataset we had and wanted to have a dataset that was newer, we went to search for one. This was one of the hardest parts during this whole process since we wanted specific data. One of the ideas we had was trying to find a newer dataset of the Ames, Iowa data. So, we contacted the Ames, Iowa Assessor Office. We also did not have much time to waste because cleaning the data will take a while. After a few days we went back and forth trying to get an exact dataset, we came up short. We could not get what we were looking for. Then we also tried contacting the Assessor's office near Philadelphia, PA. We also came up short for the dataset near Philadelphia, PA. One of the big issues we encountered with our code was Feature Importance, we did not know which features would help our specific models the best. We also believed that we could do better with our models. Our best model had a 15% error on test data, which is not bad but again we wanted to do better.

C. NEW PLAN FOR IMPLEMENTATION

The new plan for implementation was to focus on getting a new dataset as soon as possible. We did not want to waste any time since we would have to clean the data and then run our code again. The first thing we did was find a new dataset that had relevant features. We had found one at first, but we believed that we could find a better one. With some luck, we had found the one that we thought would best suit us and our project. We will talk more about that new data set in the Data Set section. Since finding the new data, we had cleaned it and implemented it into our code. We wanted to focus on hyperparameter tuning, validation testing, and finding the best features for each model. We also wanted to implement time series analysis. We think that if we implement time series analysis, it will work best with our dataset since we have three cities. Next, we wanted to focus on a few months with the two years we had within our data. We want to train our model for 2016 on the key features that had the strongest contributions to the tac dollar value count. We also had our validation set for 2017. Then with that we had our time series analysis where it shows the actual price and predicted price. With the implemented time series analysis, it understands the trend and patterns in the actual and predicted prices over the time period given. Overall, time series analysis is used throughout the code for data visualization, which allows you to see the regression models. It plots the predictions against the actual prices over time. We will talk more about that in the Implementation section.

IV. THE DATA SET

The data set we had used was focused on three counties in California. The dataset was full of real estate properties in Los Angeles, Orange County, and Ventura California. The dataset is from 2016 and 2017. The reason we chose this dataset specifically is because it had relevant property and surrounding features that we had wanted. It also included the basics of the house price, longitude, latitude, square feet, bedroom count, bathroom count, etc. Shown below are a couple of plots and attributes that we used.

	parcelid	logerror	transactiondate	airconditioningtypeid	architecturalstyletypeid	basementsqft	bathroom
5	11509835	-0.2705	2016-01-02	1.000000	7.229885	713.581395	
6	12286022	0.0440	2016-01-02	1.816372	7.229885	713.581395	
7	17177301	0.1638	2016-01-02	1.816372	7.229885	713.581395	
8	14739064	-0.0030	2016-01-02	1.816372	7.229885	713.581395	
9	14677559	0.0843	2016-01-03	1.816372	7.229885	713.581395	
...	...	...	...	...	...	...	
90270	10774160	-0.0356	2016-12-30	1.000000	7.229885	713.581395	
90271	12046695	0.0070	2016-12-30	1.816372	7.229885	713.581395	
90272	12995401	-0.2679	2016-12-30	1.816372	7.229885	713.581395	
90273	11402105	0.0602	2016-12-30	1.816372	7.229885	713.581395	
90274	12566293	0.4207	2016-12-30	1.816372	7.229885	713.581395	

Figure 1: The dataset shows a glimpse of the 2016 real estate data.

	parcelid	logerror	transactiondate	airconditioningtypeid	architecturalstyletypeid	basementsqft	bathroomcnt	bedroomcnt
5	11509835	-0.2705	2016-01-02	1.000000	7.229885	713.581395	4.0	4.0
6	12286022	0.0440	2016-01-02	1.816372	7.229885	713.581395	1.0	2.0
7	17177301	0.1638	2016-01-02	1.816372	7.229885	713.581395	2.5	3.0
8	14739064	-0.0030	2016-01-02	1.816372	7.229885	713.581395	1.0	2.0
9	14677559	0.0843	2016-01-03	1.816372	7.229885	713.581395	2.0	2.0
...	...	...	...	...	...	...	...	...
90270	10774160	-0.0356	2016-12-30	1.000000	7.229885	713.581395	1.0	1.0
90271	12046695	0.0070	2016-12-30	1.816372	7.229885	713.581395	3.0	3.0
90272	12995401	-0.2679	2016-12-30	1.816372	7.229885	713.581395	2.0	4.0
90273	11402105	0.0602	2016-12-30	1.816372	7.229885	713.581395	2.0	2.0
90274	12566293	0.4207	2016-12-30	1.816372	7.229885	713.581395	1.0	3.0

Figure 2: The dataset shows a glimpse of the 2017 real estate data.

V. IMPLEMENTATION

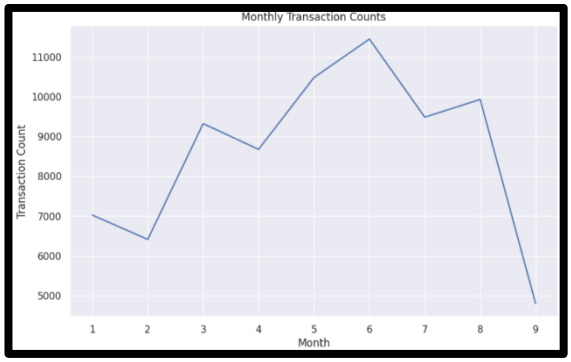

Figure 3: This figure showcases the months that had the highest transactions over the two years and why we chose the period of (March-August) summer to perform our study.

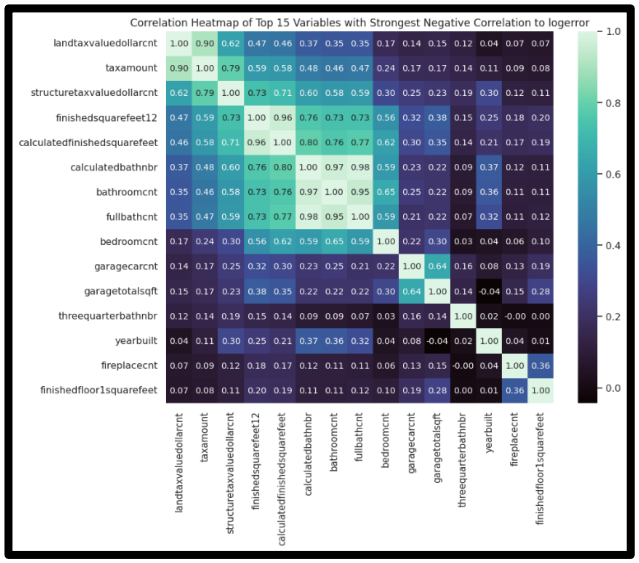

Figure 4: The correlation matrix showcasing the strongest correlated features for predicting our target vector "taxvaluedollatcnt".

In our implementation of the code, we narrowed down on what we chose to train our models on. We decided to not split the data up by counties, but instead only focus on the busiest season in real estate sales. As *Figure 3* showcases, we identified the period of March through August to be the busiest and thus the most plentiful in data. This allowed us to address any "seasonal factors" that may come into play on the edges of the fiscal year.

We also much like our parent paper, based on data over a set period of months, identified the 15 strongest correlating features to our target vector "taxvaluedollatcnt". In a Pearson Correlation Matrix, the closer the correlation value *r* is to 1 or -1 the stronger the positive or negative correlation that feature has with the target, respectively. We selected these 15 "key features" and trained our data on them. We would like to note that we split our datasets into simply training and validation instead of the train, test, and validation. This choice was made due to the fact that the data being used to train the models only came from a singular year.

A. TIME SERIES ANALYSIS

In the original code we found from our "parent paper," it did not have time series analysis. As we talked about in the CONTRIBUTIONS and DATA SET section, there are two sets of data ranging from 2016 to 2017. Since there are two sets with different years, we wanted to implement time series analysis in our code.

Time series analysis uses transaction dates as the x-axis. In this case, transactions dates represent time in the time series data. It plots both actual prices and predicted prices on the y-axis. By plotting both actual prices and predicted prices over time, you can see the trends in the data. Time series analysis often examines the residuals, in our case that would be actual prices and predicted prices.

In Figure 5, the time series graph shows the KNN Regression model comparing actual property prices with the predicted prices over time, from April to August 2017. This model follows the general trend of the actual prices but fails to capture the sharp spikes and seasonal trends in the data. KNN Regression model struggles with fine details and sudden changes in the market.

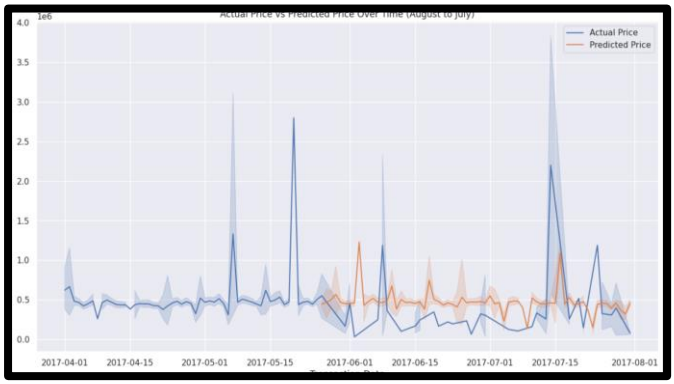

Figure 5: This picture showcases the time series graph of predictions made by the XG Boost model, which was our *worst* performing model on the 2017 dataset.

In the following diagram, figure 6, the time series graph shows a comparison of actual versus predicted property prices from the Random Forest model between the period of April 2017 to August 2017. This model shows a closer alignment between actual and predicted prices compared to Figure 5. The predicted line is very close to the actual line, which suggests that this model is better with the complexity of the data. Overall, this graph gives us a more accurate modeling of property prices through Random Forest.

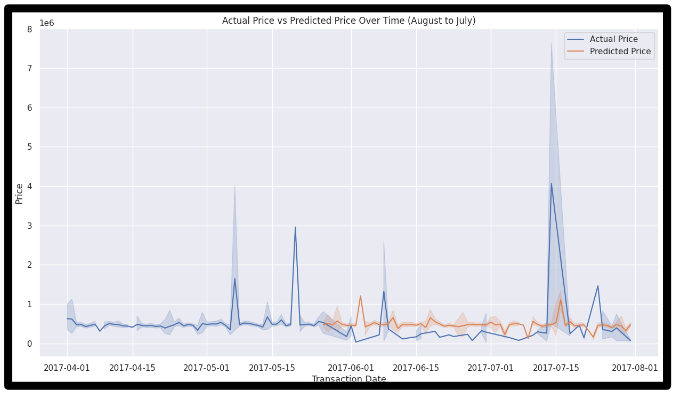

Figure 6: This picture showcases the time series graph of predictions made by the random forest model, which was our best performing model on the 2017 dataset.

Time series analysis has been a key part of our research, it allowed us to understand how property prices change over time. Using this method helped us see trends, like when prices go up or down during the year. By comparing the actual property prices with predicted property prices over the months, we have gotten a clearer picture of the market's natural flow. Time series analysis is valuable for anyone to understand when and why property prices fluctuate.

B. MACHINE LEARNING MODELS

Linear Regression: A technique used to model a relationship between a dependent variable "y" and independent variables using a straight line.

Lasso: An iteration of linear regression that adds a penalty to the complexity of the model and makes some features zero to avoid overfitting.

Ridge Regression: A technique that adds a regularization parameter *alpha* that prevents the model from overfitting. In the figure below we showcase how we also optimized for the best *alpha* value.

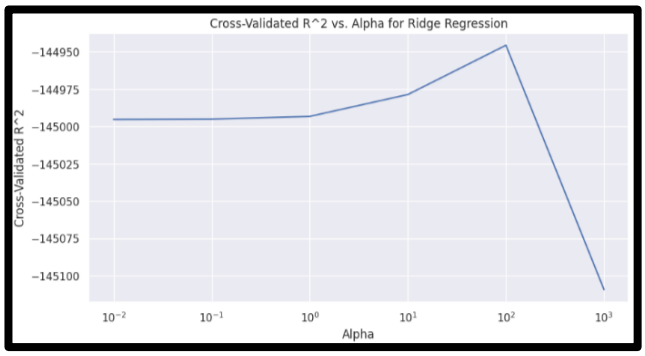

Figure 8: Showcases how the optimal alpha value was picked for Ridge Regression.

KNN Regression: A technique that uses the average number of nearest neighbors (k) of a data point to predict its value.

XG Boost: A technique that combines multiple decision trees to make a prediction and adjusts the gradient with each iteration to improve the result.

Random Forrest: Much like XG Boost is an ensemble learning method. It utilizes bootstrapping and random variable selection to reduce variance and improve accuracy.

VI. RESULTS

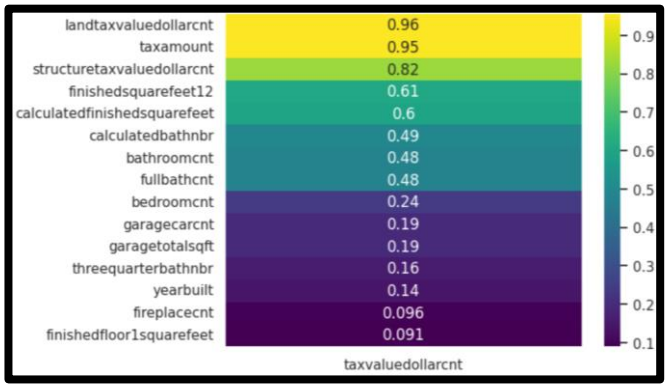

Figure 9: This figure shows the "key features" that contribute to the sale price of the home ("taxvaluedollatcnt").

In this study our core purpose was to identify key features that would most accurately predict sale prices and pair them with the most effective machine learning model. As can be seen in figure _, the strongest predictor of the value of the land upon which it was constructed. With regards to the actual features of the house, the amount of square footage that is actually finished plays a large role in determining how high the final sale price of the home will be. These features are fairly intuitive as the buyers who are able to see a more finished product are able to better envision living in that space. Unsurprisingly, the number of bedrooms, bathrooms, and the size of the garage are all factors that have a strong positive correlation with house prices. What was interesting to see was that the customers don't particularly care about a furnished basement as much as they care about the fireplace. This indicates that the *appearance* of a more "homely" or "finished" or "lived in" home are stronger determinants of home price than shear metrics on the size of certain rooms. It was also surprising to see that the basement, also referred to as "finishedfloor1squarefeet", has the least positive impact on sale price. This was surprising to see as often times finished basements, according to conventional transactions, tend to add significant value to the home's sale price.

Model	Train RMSE	Validation RMSE Normalized	Train R^2	Validation R^2
Linear Regression	0.3127806852451063	0.24	0.9329729127122606	0.9666352214082314
Lasso	0.3127806921580613	0.24	0.9329729097494481	0.9666352389860839
KNN Regression	0.3943913932933794	0.47	0.893432359209648	0.8783145171368244
Ridge Regression	0.3128014955024155	0.24	0.932963993380596	0.9666373265424204
XGBoost	0.1540570062424864 54	0.53	0.9837393496165592	0.8200577172175549
XGBoost Tuned	0.3128014955024155	0.45	0.8646129074744804	0.8200577172175549
Random Forest	0.3128014955024155	0.23	0.9902109764550691	0.9689701322668891

Figure 10: This figure showcases the train, test, and validation RMSE. The figure also shows the R^2 value which is an indicator of how well the statistical model was able to predict the outcome.

As it can be seen in the figure above Ridge Regression performed the best out of all the models. It had a R^2 of 0.969 on its validation set meaning it was a great predictor of outcome and a root mean squared error (RMSE) of only 0.23. On the other hand, XG Boost performed the worst out of all the models with an R^2 value of 0.98 and an RMSE of 0.53. This is surprising as XG Boost is an ensemble model much like random forest. What was surprising to note was that tuning the XG Boost model did not improve the results by any significant margin. This may be due to Grid Search being utilized or maybe the optimal range of hyperparameters was not provided. Coming back to the best model, Random Forrest, it is important to note the potential reasons of the model's high performance. We believe it was due to multicollinearity. Multicollinearity refers to certain variables being related, "garagecnt" and "garagesqft" for example may get bundled together. Multicollinearity does not occur in the case of random forest as each variable is selected individually. This makes the model much more robust.

VII. NOVELTY

The principal novelty introduced in this paper is time series analysis into the domain of real estate price prediction. Prior research in real estate analysis has been focused on variables and cross-sectional data, this research focuses on the application of time series analysis techniques to understand and forecast real estate prices accurately in three Californian counties: Los Angeles, Orange, and Ventura.

Most real estate analysis often focuses on property features, economic factors, and demographics. This research focuses on time series data, it captures temporal patterns and trends that impact real estate prices. The novelty is recognizing how time has influenced the real estate market. It is affected by economic cycles, seasonality, and external factors, which can be better understood through time series

analysis. For this particular study we decided to look at the summer of 2016 and 2017. This is because during our exploratory data analysis, we concluded that the summer is when the greatest number of home sales occur for both years. Our graphs showcase the 2016 data during the months of March – August as well as our predictions for 2017 over the same periods based on various models.

Another aspect of our research is the use of advanced machine learning techniques that are not typically applied to the real estate market. For example, we have explored the use of multiple machine learning methods and advanced regression techniques. These techniques include Random Forest and XGBoost, which have been fine tuned to optimize their performance for real estate data, presenting a novel approach in predicting property prices with higher accuracy.

This paper broadens its novelty by thinking about time series analysis on investment decisions. By forecasting prices accurately, this paper helps real estate investors, lenders, and stakeholders in California with insight into the market, which can help make informed decisions. This research we have conducted goes beyond conventional ways to conduct property price prediction. Our integration allows for a more holistic view of the real estate market.

VIII. CONCLUSION & FUTURE WORK

In conclusion, this research demonstrates the application of machine learning techniques to predict real estate prices in three California countries: Los Angeles, Ventura, and Orange. We incorporated methodologies from research studies, accomplished an understanding of property valuation. The novel introduction of time series analysis has allowed for a deeper exploration, enhancing the model's accuracy. These findings have implications for stakeholders, offering a data-driven approach to decision-making.

Throughout our real estate analysis research, we have done research, code study, data acquisition, implementation, and documentation. While we have implemented our ideas, we have had other concepts during this research journey. These ideas enhance the project's depth, which are for future integration for this research.

The first idea for future work is sentiment analysis. The models could include more details like reviews, rating and other more "emotional" features that play into a purchase. The pro for doing sentiment analysis is having a more holistic understanding of customer purchase. We can also see how much the perception of value plays into a home's value. The con is there is limited data for each more and that may lead to skewed outcomes. The reviews could also be biased and can be heavily skewed by external forces. The sentiments add other layers to the model and improve the model's performance. IT can further capture complex relationships we may not initially consider.

Our second idea for future work is feature combination. Feature combination is to combine information from existing features to create one new feature. It captures the relationships and patterns between the two individual features. The pro of doing feature combination is it can improve the model performance since it captures complex relationships. The con is the data can be sparse if there are two features that do not relate to each other. There is also a risk of overfitting. The possible outcome of incorporating feature combination in our project can lead to improved model performance by creating those new features that capture intricate data relationships. This will help predictive accuracy and provide deeper insights into the data.

Our third idea for future work is splitting by county for our research. This would be beneficial since it lets us know the geographical variations in the real estate markets. Each county could have different factors influencing the house price. It would help improve the overall accuracy and effectiveness of the predictive model. The pro of doing this is getting local market insights and it would tailor the model. The con is it would reduce the sample size of each county. There would be overfitting and complexity. A possible outcome would be splitting data by county in house price prediction can give insights into county market dynamics, targeting marketing efforts. This can lead to improved prediction accuracy with each county market.

IX. REFERENCES

[1] Zhao, Y., Ravi, R., Shi, S., Wang, Z., Zhao, J., & Lam, E. (2022, August 29). PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction. https://arxiv.org/pdf/2209.05471.pdf

[2] Kalehbasti, P. R., Nikolenko, L., & Rezaei, H. (2019, July 29). Airbnb price prediction using machine learning and sentiment analysis. arXiv.org. https://arxiv.org/abs/1907.12665

[3] Jha, S. B., Pandey, V., Jha, R. K., & Babiceanu, R. F. (2020, August 22). Machine learning approaches to real estate market prediction problem: A case study. arXiv.org. https://arxiv.org/abs/2008.09922

[4] IndigoPurple. "IndigoPurple/Pate: Pate: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction." *GitHub*, github.com/IndigoPurple/PATE. Accessed 28 Nov. 2023.

[5] Mathur, Abeer, and Parag Patel. "Abeermathur7/Realestatemodeling." *GitHub*, 11 Oct. 2023, github.com/Abeermathur7/RealEstateModeling.