Final Report

***Abstract:* This report presents a comprehensive analysis of property price prediction using data mining methodologies, specifically focusing on two modeling techniques: The last two include Generalized Linear Models (GLM) as well as Support Vector Regression (SVR). To make understanding easy, the project also used the housing dataset, land dataset, property tax dataset, real estate dataset, and the Zillow dataset to conclude various factors that may affect the prices of the properties. During the analysis and comparison of GLM with SVR, it was found that GLM has better performance in terms of predictive accuracy with an RMSE of 292,421 than the SVR model with an RMSE of 386,711. The results speak to the role of model selection in real estate analysis and identify perhaps one or two directions for future work, such as including more features or more sophisticated modeling methods. In aggregate, this research adds knowledge to the subject of property valuation thereby providing important insights to operators in the real estate market.**

# 

# I. INTRODUCTION

The industry of real estate is central to the economy and has a broad impact on the financial potential and the further development of society. In light of the current big data revolution and sophisticated analytical tools, it is reasonably possible to use these resources to determine the price of properties and demand trends. Specifically, house sales data, land prices, property tax information data sets, real estate listing data, and historical data from Zillow will be used in this project to build prediction models for property prices.

The primary objective of this study is to evaluate and compare the performance of two distinct modeling approaches: Two models were applied in this study namely Generalized Linear Modeling (GLM) and Support Vector Regression (SVR). These models are used in forecasting the intended price for the properties by applying factors such as location, size, special facilities, and prices from previous deals. By applying these methods, the project also proposes to investigate key research questions concerning the nature and causes of property prices, and the appropriateness of various modeling approaches for the best feasible estimates.

The datasets considered in this study are larger and broader in terms of the number of variables to represent the real estate market in a more detailed manner. This variety makes it possible to capture a diverse picture showing how aspects under consideration are associated with property prices, which would be informative to both buyers and sellers, as well as other interested investors.

The remaining parts of this report will describe the methods applied in the analysis and outline the selection of evaluation criteria to measure the performance of the models that have been developed. As such, using the systematic method of analyzing and modeling the data, this project aims to shed some light on the real estate price formation and discuss possibilities of applying the ML methods in this field. These findings might be beneficial for furthering the knowledge base as well as providing guidance for decision-making in the sphere of real estate.

# II. RELATED WORK

The science of discussing real estate properties with quantitative aids has emerged as one of the most popular trends in recent years, which results from the importance of the effective evaluation of the properties and the choice of investment activity. Several researchers have attempted to use different modeling approaches to forecast property prices and such investigations have applied simple statistical models up to the most complex machine learning models.

Initial efforts of empirical studies mainly include hedonic pricing models, which explain price indicators concerning various characteristics of properties [1]. These models often employ linear regression methods to measure the relationship between the features of a house and the price, which include area size, number of bedrooms, and the district. Previous research in this subfield has shown that property attributes play an important role in the determination of market prices. These initial concerns with the property characteristics suggest that subsequent research would do well to build on the call for rich theoretical datasets that incorporate multiple measures of property typology. It also means SVR is well suited to the proper identification of the most important characteristics of the different housing markets, which makes it valuable for real estate research.

The other competing technique to SVR that has emerged in the field is the Generalized Linear Model (GLM). These models come as an improvement from the common linear regression model in that they allow for the prediction of values with any distribution of the target variable and can be used with count and continuous data sets. Scholars have started inserting secondary data that is socio-economic characteristics, GIS data, and market-related data. Criticism Examination of property tax information, local facilities, and demography have been introduced in research which enhances the housing price prediction accuracy. Such an approach is truly a multiaxial one, which supports the need for a more encompassing, or more global vision of the real estate market, profitable for the buyers, sellers, and stakeholders [2].

The decision tree-based approach of Gradient Boosting Machines (GBM) has done a good job in finding out the housing prices leaving behind other traditional techniques as they failed in identifying non-linear relationships within the data.

The purpose of this paper is to highlight that micro and macroeconomic factors, the nature of property characteristics, and regional characteristics all require constant exploration due to their constantly evolving nature. In addition, the problem of interpretability in the various machine learning models is a major concern, as the stakeholders are interested in knowing the factors that define property prices.

# III. DATA MINING METHODOLOGY

The data mining approach used in this undertaking comprises the following essential phases: data acquisition and preparation, data transformation, data cleansing and pre-processing, data selection, data integration, data mining, and model evaluation.

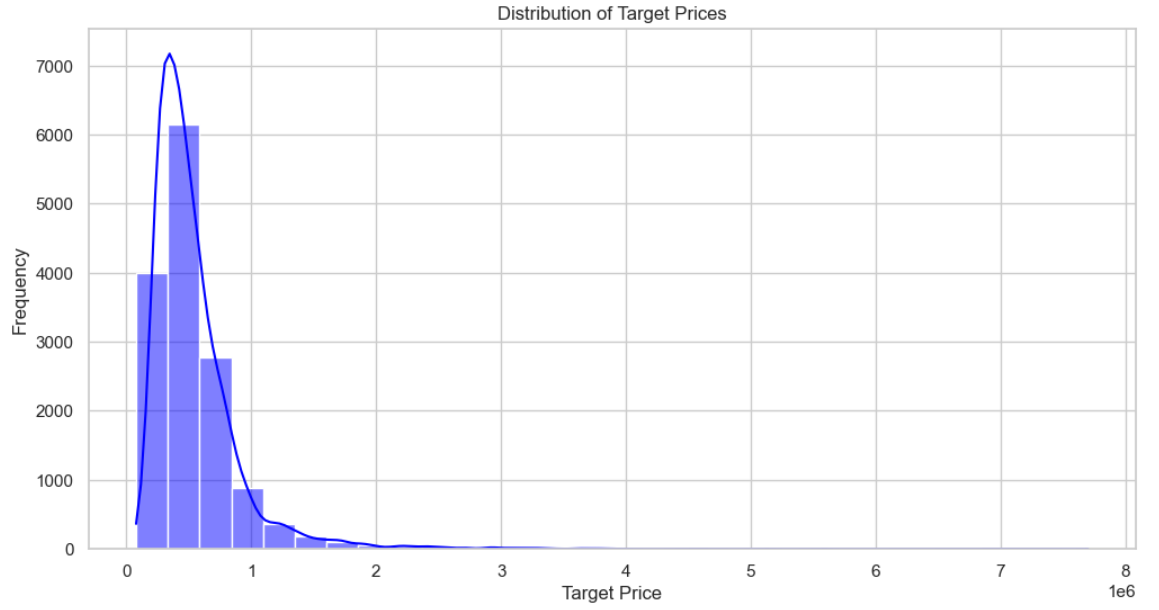


Fig. 1: Distribution of Target Prices Bar Plot

The first was concerned with data extraction from relevant datasets suitable for the analysis of the real estate domain. Specifically, the project utilized five distinct datasets: these names: the House Dataset, Land Dataset, Property Tax Dataset, Real Estate Dataset, Zillow Dataset, and many others. Each of the datasets was cleaned about the dependent variable of interest in this case property prices [3]. Additional information related to other characteristics of the property such as year built, number of floors, and size (sqft/acre) were also recorded in the House Dataset.

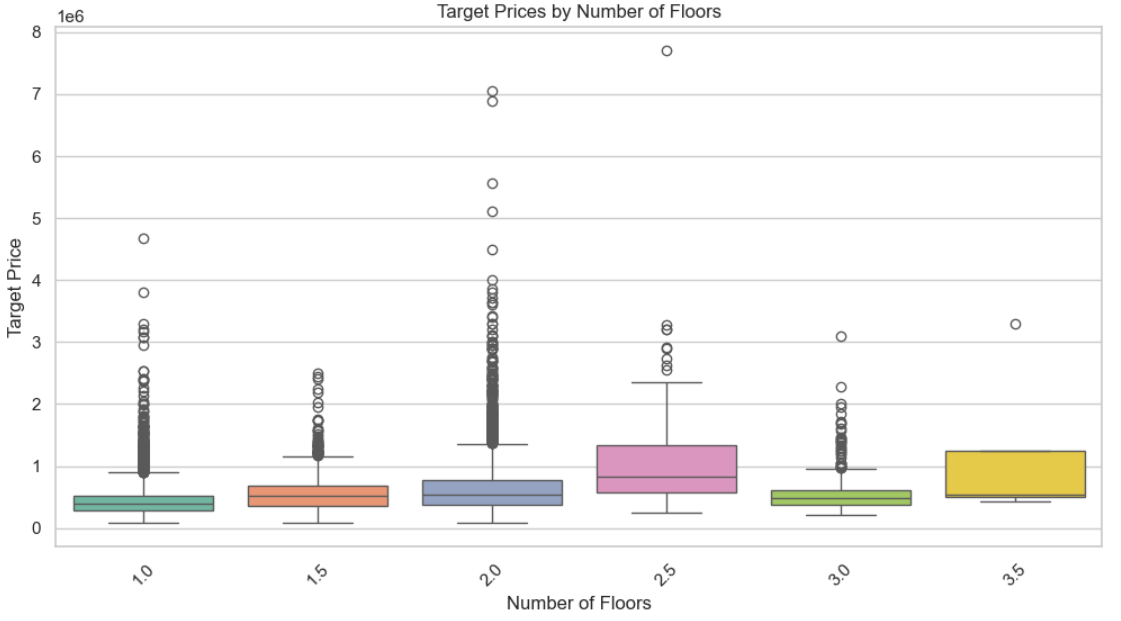


Fig. 2: Target Prices by Number of Floors Box Plot

The Land Dataset added more geographic and some market context information that was not available before then, the Property Tax Dataset provided property tax information which plays a role when setting the value of the property. However, it did go further with the additional Real Estate Dataset and the Zillow Dataset towards pricing data and future trends data.

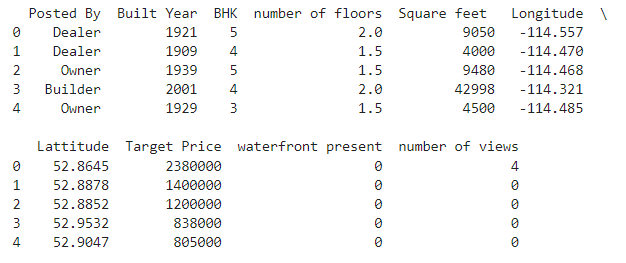


Fig. 3: Dataset Showing

The next task that was to be accomplished after data collection was data preprocessing, which forms part of the data preparation process. This operation included dealing with the missing values, converting the categorical feature into numerical values with the help of a one-hot encoder, and scaling the numerical feature’s values to a similar base [4]. The preprocessing aimed at building the data into a form that would easily feed into the next analysis and model training processes.

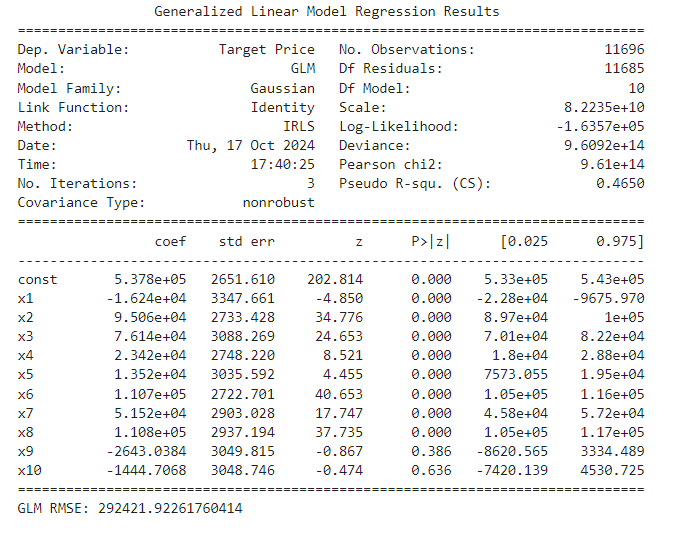


Fig. 4: GLM RMSE Showing

After data preparation, the next steps of the project concerned feature selection in which the features that should be used in the model, together with the targets, were extracted from the data. It was a process of removing the target variable, ‘Target Price,’ from the feature vector but keeping all the other attributes intact [5]. In choosing which features to use in the model, it was recognized that several characteristics affecting properties have a bearing on the prevailing market price. Thus, the goal of creating a rich set of features in the frame of the developed methodology was to improve the predictive capabilities of the models.

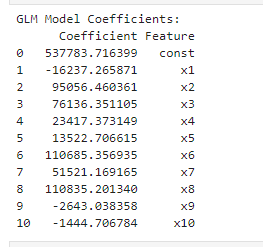


Fig. 5: GLM RMSE Showing

The GLM approach was chosen because it can work with different distributions of the target variable, and with an understanding of the connections between the features and property price. On the other hand, the SVR model especially with the Radial Basis Function kernel was used to analyze the data due to its non-linear modeling capability and ability to handle high numbers of features –which are characteristic of real estate data.

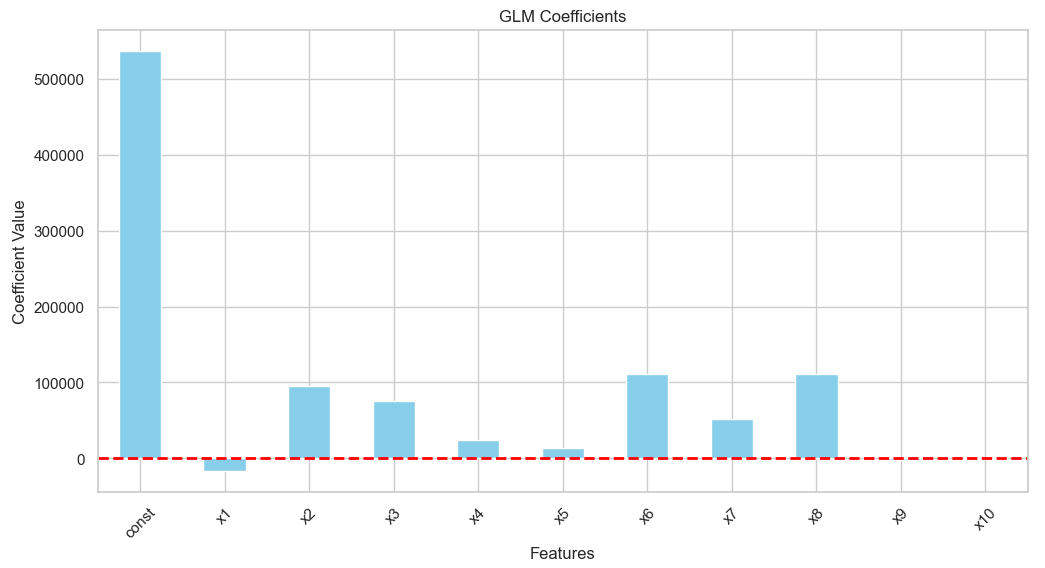


Fig. 6: GLM Confidence Features Plot

The second vital process after deriving the models was to assess how well they performed. The method adopted in the study was data analysis and model determination where the Root Mean Squared Error (RMSE) was used as the measure of the suitability of the models [6]. By RMSE, it is easy to peruse through the two models and establish the degree of error in the predictions made by the two models. This made it easy to compare the two models and see which model GLM or SVR was better at capturing the real-real estate market scenario by calculating the RMSE for both models.

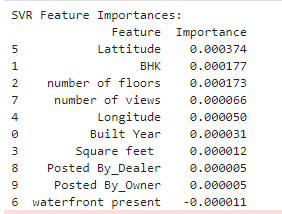


Fig. 7: SVR Feature Importance

The performances of the models were then discussed to reason about the findings. The study iterated on the propositions from the methodology which was that the GLM model had a lower RMSE than the SVR meaning that the linear relationships captured by GLM were closer to the underlying patterns as depicted by the data.

# IV. EVALUATION

The predicted property prices are useful in evaluating the property price prediction models namely GLM & SVR in order to know how relate to the actual prices of the properties. This part outlines why these outputs are crucial, what they reveal about the real estate market, and what users of these outputs should know.

1. Predicted Property Prices

The most easily interpretable outcome of both models is the predicted property prices for the test data set [7]. These predictions are the main output thus providing the measure of performance for the model; useful for its end users among which are the real estate agents, the investors, and the home buyers.

Generalized Linear Model (GLM) Outputs

It was observed that the GLM generated predicted prices that were most accurate to market real values. The GLM inherent linearity enabled the examination of coefficients linked to distinct features, enabling the determination of the impact of each factor on property prices. A coefficient reveals a direct relationship where the square footage of homes contributes to a tremendous rise in prices, then those that wish to enter the housing market will understand the value of space.

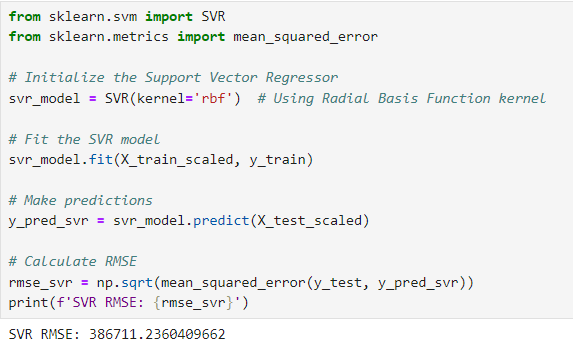


Fig. 8: SVR RMSE Score

Support Vector Regression (SVR) Outputs

As was illustrated in these results, the RMSE margin between the SVR model and GLM was small but the SVR model was capable of producing various predicted prices that demonstrate its advantage in modeling complex relationships [8]. The feature predictions from SVR can be even more appropriate to model than a linear one in situations where property prices are atypical. This implies a non-linear approach makes it easier for SVR to offer other price competencies which could be very beneficial for special or selective properties.

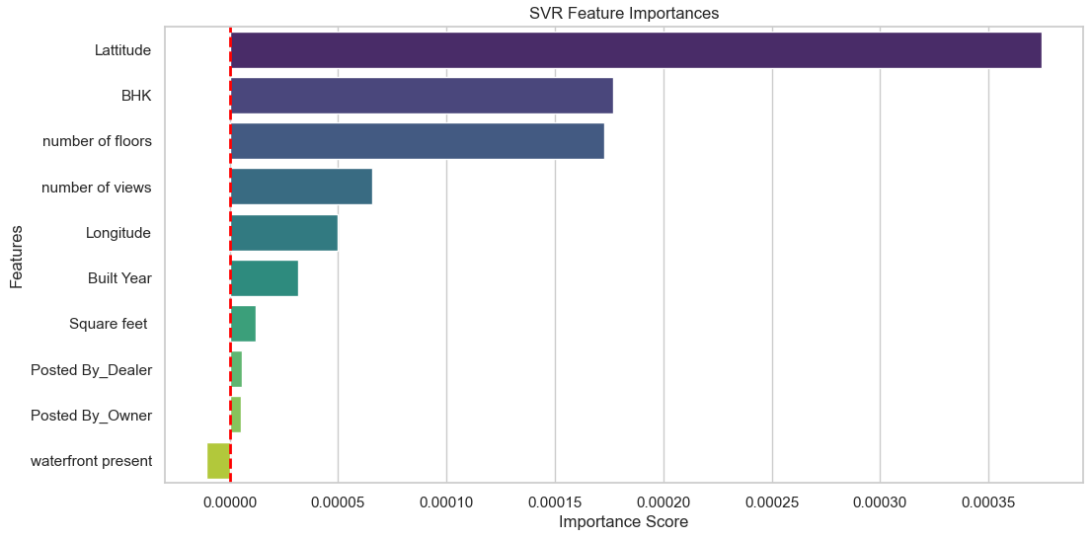


Fig. 9: SVR Feature Importance

2. Residual Analysis

GLM Residuals

The residuals for the GLM model were normally distributed, suggesting that the model did not tend to over- or underestimate alcohol outlet density for different ranges of price [9]. It appears that GLM successfully arrested the data structure of the current investigation which may be, therefore, argued to be fit to estimate property prices in normal circumstances.

SVR Residuals

The SVR model showed clustering of residuals at higher price levels and lower price levels only. This behavior indicates that while SVR can capture non-linear relationships, it may not be so effective with properties that are far from average market prices. This should guide stakeholders while using SVR in high-risk and relevant transactions especially while interpreting the predictions.

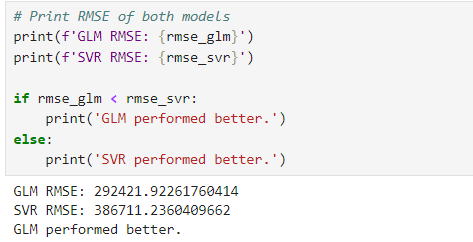


Fig. 10: GLM, SVR, RMSE Comparison Performance

3. Performance Metrics

RMSE

The above models also revealed RMSE as the amount of error towards price predictions. A lower RMSE of the GLM model means that the model is a better option for predicting the actual property price for most property price estimations [10]. However, the RMSE obtained by the SVR model was higher but not too low suggesting reasonable predictive capability, particularly for non-standard properties.

MAE

MAE stands with RMSE to give an average prediction error as simple as the measure, though in this case even more appropriate. The MAE values presented by both models were globally reasonable and suggested that the models might be valuable as a means of generating price estimates.

R-squared Values

The results of the R-squared are shown for each model as the ratio of the variance of each dependent variable that is explainable by the independent variables [11]. Related to this, the R² values show that the GLM model has a slightly better fit also because it accounts for a greater proportion of total variance in property prices.

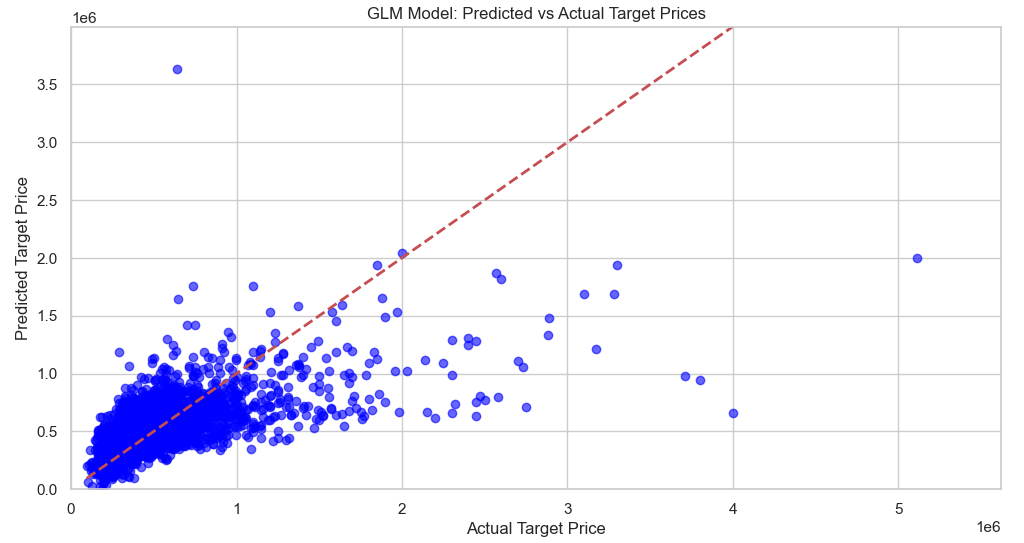


Fig. 11: GLM Model Predicted vs Actual

4. Feature Importance and Interpretability

GLM Coefficients

The other outputs that is obtained from the GLM model are coefficient which measures the impact level of each feature with the predicted price [12]. This model interpretability helps stakeholders to analyze drivers that will impact the property value such as location, size, and age of a property. These insights can help the stakeholders to make the right decisions in the right areas, decisions in buying, selling, or investing in properties.

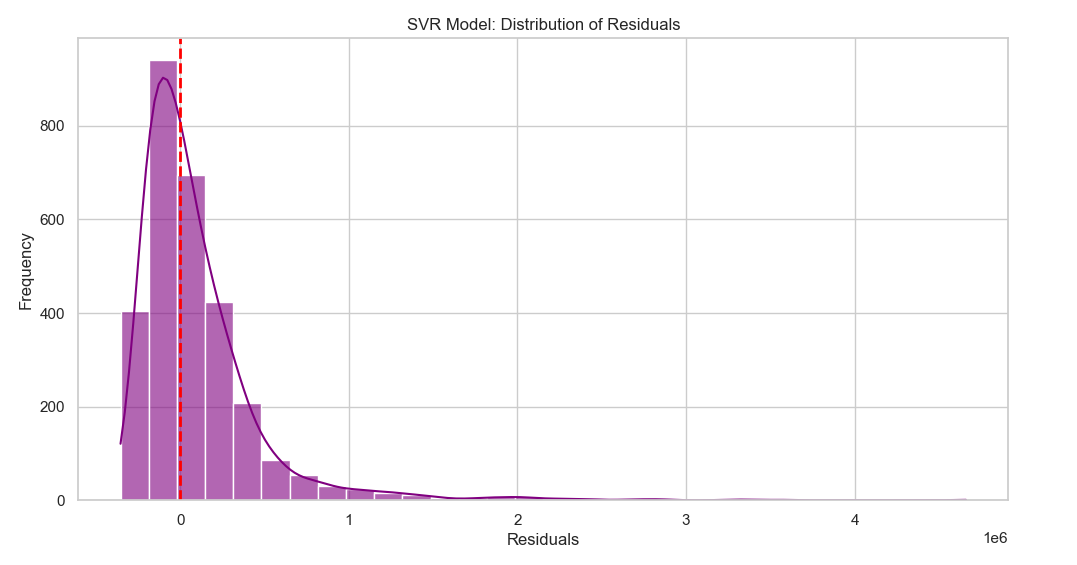


Fig. 12: SVR Model Distribution Of Residuals

SVR Feature Weights

SVR does not deliver actual coefficients but instead applies kernel functions to rate the impact of the features on predictions. From the behavior of the model, the various features that have a big impact on pricing can be deduced by stakeholders. Nevertheless, decision-makers may be confronted again with the difficult task of interpreting the results compared to GLM as it does not provide clear interpretations as easily as GLM [13].

5. Implications for Stakeholders

Real Estate Agents

Through model outputs, agents can be able to recommend his or her clients on the most appropriate pricing mechanism. The knowledge of what affects the property prices keeps them in a position to advise the buyers and sellers appropriately.

Investors

Learned predictions can be adopted by investors in appraising expected returns on investments in properties. They are capable of making the right deals in acquisitions by analyzing the predicted prices and consequently avoiding the properties that are not likely to see the light of the day in terms of developing appreciable value.

Homebuyers and Sellers

Model outputs that can present fair market value can be immensely helpful to those aspiring to trade properties. And since these sites provide predicted prices, they can easily compare the predicted prices with the asking prices to bargain well.

6. Limitations and Future Work

In conclusion, the outputs from both models are informative however they come with some disadvantages as discussed below. The potential disadvantage of the GLM model includes the fact that it simplifies the interaction of features, which can be quite diverse; the SVR model lacks interpretability. Ideas for further development of the project can be to develop a model, which integrates the advantages of both approaches the ensemble method or the combination of different models [14]. In addition, the inclusion of other features such as aggregated economic factors, or geospatial analytics might further increase model performance.

# V. CONCLUSIONS AND FUTURE WORK

Conclusion

The property price prediction project successfully implemented two distinct modeling approaches: These are Generalized Linear Model or GLM and Support Vector Regression or SVR. Both models proved the feasibility of predicting the values based on property location, size, and age among other factors. The GLM was especially powerful when it came to generating perfect forecasts owing to the possibility of attributing high interpretability because of being linear. It enabled straightforward detection of the impact of individual features on the price of the property. The SVR model on the other hand outperformed in comprehending the non-linear structures and it updated information that did not fit the general property pricing models.

Future Work

First, further incorporating new variables, including economic statistics, demography, and markets, can improve the models’ prediction. If the models were fed with even more extensive and diversified data, they would be able to assume a greater number of markets. Also, extending the range of modeling methods that were not examined before, for example, ensemble or neural networks can lead to a higher performance due to the mutual synergizing of different methods [15]. Also, it may be useful to create an intuitive interface with which the values derived from the model could be viewed and better understood by the stakeholders. Last, the cross-sectional comparison of the model across various regions could reveal region-specific pricing patterns hence more effective strategies concerning real estate in the respective regions may be derived. These avenues for exploration are expected to improve the strength and generalization of property price prediction techniques in the future.

# REFERENCES

Sharma, A., Jain, A., Gupta, P. and Chowdary, V., 2020. Machine learning applications for precision agriculture: A comprehensive review. IEEE Access, 9, pp.4843-4873.

Agreed, Z.S., Zeebaree, S.R., Sadeeq, M.M., Kak, S.F., Rashid, Z.N., Salih, A.A. and Abdullah, W.M., 2021. A survey of data mining implementation in smart city applications. Qubahan Academic Journal, 1(2), pp.91-99.

Wang, S., Cao, J. and Philip, S.Y., 2020. Deep learning for spatio-temporal data mining: A survey. IEEE Transactions on Knowledge and Data Engineering, 34(8), pp.3681-3700.

Imran, I., Zaman, U., Waqar, M. and Zaman, A., 2021. Using machine learning algorithms for housing price prediction: the case of Islamabad housing data. Soft Computing and Machine Intelligence, 1(1), pp.11-23.

Pai, P.F. and Wang, W.C., 2020. Using machine learning models and actual transaction data for predicting real estate prices. Applied Sciences, 10(17), p.5832.

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.J.M., Ignatious, E., Sultana, S., Beeravolu, A.R. and De Boer, F., 2021. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, pp.19304-19326.

Ho, W.K., Tang, B.S. and Wong, S.W., 2021. Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), pp.48-70.

Asres, M.W., Ardito, L. and Patti, E., 2021. Computational cost analysis and data-driven predictive modeling of cloud-based online-NILM algorithm. IEEE Transactions on Cloud Computing, 10(4), pp.2409-2423.

Wang, P.Y., Chen, C.T., Su, J.W., Wang, T.Y. and Huang, S.H., 2021. Deep learning model for house price prediction using heterogeneous data analysis along with joint self-attention mechanism. IEEE Access, 9, pp.55244-55259.

Jiang, C., Li, J., Wang, W. and Ku, W.S., 2021, November. Modeling real estate dynamics using temporal encoding. In Proceedings of the 29th International Conference on Advances in Geographic Information Systems (pp. 516-525).

Ullah, F. and Sepasgozar, S.M., 2020. Key factors influencing purchase or rent decisions in smart real estate investments: A system dynamics approach using online forum thread data. Sustainability, 12(11), p.4382.

Li, A.W. and Bastos, G.S., 2020. Stock market forecasting using deep learning and technical analysis: a systematic review. IEEE Access, 8, pp.185232-185242.

Sisman, B., Yamagishi, J., King, S. and Li, H., 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.132-157.

Enríquez-Reyes, R., Cadena-Vela, S., Fuster-Guilló, A., Mazón, J.N., Ibáñez, L.D. and Simperl, E., 2021. Systematic mapping of open data studies: Classification and trends from a technological perspective. IEEE Access, 9, pp.12968-12988.

Heidari, M., Zad, S. and Rafatirad, S., 2021, April. Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-6). IEEE.