RESEARCH ARTICLE OPEN ACCESS

Combination of supervised and unsupervised learning models to forecast a successful business decision

Joseph Benny[1], Maria Joy[2], Vrinda K[3]
1(Department of Computer Science & Engineering, Mar Athanasius College of Engineering, Kothamangalam
Email: josephbk99@gmail.com)
2(Department of Computer Science & Engineering, Mar Athanasius College of Engineering, Kothamangalam
Email: m21cse002@mace.ac.in)
3(Department of Computer Science & Engineering, Mar Athanasius College of Engineering, Kothamangalam
Email: vrindak@mace.ac.in)

Abstract:

The Rate of Return, a crucial measure used to assess real estate investment alternatives, is calculated in large part utilising real estate rent forecast in housing market analysis. Real estate investing may generate capital gains and ensure financial success with accurate rent projection. We conduct a thorough analysis and investigation of popular machine learning techniques for rent prediction in this work, including SVM and Tensorflow. We develop a new model that includes the single-family, townhouse, and condominium housing types. The dataset's data instances each have 18 internal properties. a portion of the gathered features that filter algorithms have chosen for the prediction models. Additionally, we apply a hierarchical clustering method to group the data based on the zip code's estimated average rent and the kind of dwelling. The empirical findings reveal that, in comparison to eager learning techniques, rent prediction models based on lazy learning algorithms lead to greater accuracy and reduced prediction error.

Keywords **—SVM, Tensorflow, Data analytics, Machine learning, Data mining.**

I. INTRODUCTION

The Rate of Return, a metric used to assess how well an investment in the housing market has performed, is calculated in large part utilising real estate rent forecast. The two key components of Rate-of-Return, Net Present Value and Future Value, can be used to assess the quality of a real estate investment over a given period of time. A good and profitable Rate of Return can be achieved over time with prudent rental property investments. Due to algorithmic errors or inaccuracies in rent projection, these undertakings, however, might be exceedingly dangerous. It's not a new practise to utilise machine learning algorithms to anticipate how much a residence will rent out for. Lambert and Greenland look into eager learning techniques like multi-layer perceptrons and bagging REP trees to determine the rental pricing for both landowners and students interested in renting a place close to a university campus. . Two different property types—apartments and condos—make up the training set. The training set's coverage is restricted to the three distant zip codes that surround a university campus. The proximity of the residence to the university campus, the size and amenities of the apartment, the length of the lease, and the construction date are all input factors considered in this work. The best algorithm for predicting rent, according to the study, is bagging REP trees. However, the skewed data set, which is all located near a university campus, can lead to a biased model being produced by the proposed global learning-based method. Social sciences can make use of machine learning models, which give a broader perspective on people's perceptions of the housing market. More insightful transfer models can offer more accurate predictions of home price trends. Health, cyber security, computer hardware, computer science, and business all benefit from sophisticated uses of machine learning models. The real estate rent/price prediction models used in earlier studies are fairly generic and don't account for house type or zip code differences. To forecast rent and home prices, for instance, a generalised prediction model is

suggested using city-level data. However, this could result in incorrect forecasts. However, this could result in incorrect forecasts. Even for real estate properties in the same city or state as one another or within close geospatial proximity to one another, rent behaviour varies. Both internal and external factors affect how much rent is on average in a certain zip code. In fact, the cost of rent is influenced by outside variables like the crime rate and the quality of the schools in a given zip code, which are deal-breakers for many real estate investors. This essay considers both internal aspects of a home as well as external factors including walkability, transportation accessibility, crime rate, and academic standing. The walk score reveals which tasks can be completed on foot and which ones require a car to reach adjacent services. The transit score reflects service frequency, accessibility to jobs, and connectivity. The crime score reveals the frequency of violent and non-violent incidents in a given zip area.

Studies in the past have looked into the effects of eager learning techniques for predicting real estate rent and price. The overall goal of lazy learning approaches, in contrast to eager learning methods, is to identify the regionally optimal solutions for each test instance. By storing the training examples, Friedman, Kohavi, and Yun and Homayouni, Hashemi, and Hamzeh in [3] postpone generalisation until a fresh instance is received. A different study by Galv'an et al. [1] contrasts memory-based (or locally) learning vs. neural network methods and finds that memory-based (i.e., lazy) learning method outperforms Neural Network approach using a variety of data sets from the UCI machine learning data set repository, including Iris, Diabetes, Sonar, Vehicle, car, and balance. In this study, eager learning methods SMO and Naive Bayes algorithm were surpassed by the two lazy learning algorithms Integer Part and Atomic Radios. These study investigations provide as inspiration for our work. Three different types of homes are included in the data set: condos, single-family homes, and townhouses. The requirement to create models for dwelling type and zip code is what drove this study. We separated the data set into home types to address the thin (i.e., sparse) data in each zip code, and then used K-means clustering to produce subsets of examples within each zip code with comparable average rent values. The clustering method calculates the separation between the data points using the average-rent similarity metric. A rent prediction model is later trained using the samples of data from each cluster. In this paper, we compare and contrast eager vs. lazy learning approaches to examine the effects of several machine learning techniques on this data set. In order to demonstrate the superiority of lazy learning algorithms over eager learning techniques in real estate rent prediction for each house type and a subset of zip codes with comparable average rent prices, we present actual results.

II. RELATED WORK

Machine learning approaches have already been researched in various papers for real estate rent/price prediction. [2]– [5]. In PSO-SVM algorithm is used for real- estate price prediction. In [10], [11], spatiotemporal dependencies between housing transactions is used to predict future house prices. However, this approach is limited by spatial autocorrelation, since the degree of similarity between observations is not solely based on the distance separating them.

Some of the previous work focus on hedonic price models as a method of estimating the demand and value in the hous- ing market and determination of house prices. In these studies, rather than internal and external house features, economic submarkets are used in the prediction model which are defined in terms of the characteristics of neighborhoods or census units. In a sample size of 200 houses of all house types were used in a hedonic price model and an artificial neural network (ANN) model, and shows that the eager method ANN outperforms the hedonic model. The problem with the hedonic approach is disregarding the differences between the properties in the same geographical area. According to, ardent learning techniques might occasionally result in predictions that are less accurate than desired since they derive a single model that aims to minimise the average error

over the whole data set. In contrast, lazy learning techniques can increase the accuracy of predictions. While we extend our research analysis by contrasting the effects of eager and lazy learning algorithms on the predictive accuracy of the models that were produced with regard to each house type and a subset of zip codes with comparable average rent prices. Using a selection of criteria related to internal and exterior real estate property, we cluster data using a two-layer approach.

In the real estate industry, there are several sources of housing data. Zumper API provides house information for the Kerala neighbourhood, such as historical data on sales prices, year of sales, tax information, number of beds/baths, etc. For the purpose of this study, we gathered a collection of residential housing data for the Kerala region using the ZumperAPI. This data set's 2000 housing property records (containing townhome, single-family, and condo units) have a total of 18 characteristics. Additionally, data on external characteristics like the crime rate, transit score, and walk score is gathered.

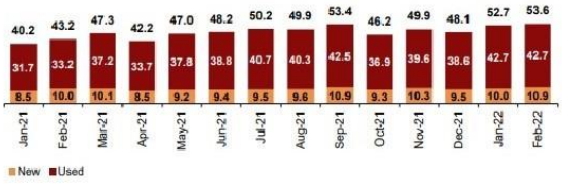

Fig. 1 Number of registered houses for sale

III. METHODOLOGY

Details of the methodology of this project is given as follows:

A. Pre-processing of Data

Re-sampling the data to balance them is one of the fundamental ideas of calibrating machine learning models when dealing with biased data. Compared to other places, some of the locations have much greater densities. We re-sampled the data in zip codes with higher home prices due to their crowded density compared to the zip codes with lower house prices in order to standardise the data. We employed SVM and Tensorflow to impute the missing values of external characteristics. Data points' separations are determined in relation to each cluster centroid. We do feature selection by running PCA on each of the data set's 18 characteristics in order to decrease the dimensionality of the data set and improve the generalisation of the model.

To determine the co-linearity between the variables, we examined the correlations between the different variables in the data set. Finding co-linearity between the variables in the data set and the desired outcome provides important information about the dependent factors that influence the rent. The number of homes listed for sale in the data set is shown in Figure 1. There is no collinearity between zip code and rent price since zip code is a nominal property. Additionally, there is a significant correlation between rent and internal factors like the number of bedrooms and bathrooms, the area size, the year, and the sale price. The metrics like walkability, transit accessibility, and crime rate used in urban planning also correlate with the average rent price. The overall pattern shows a favourable association between the average rent and walk/transit score, and a negative correlation between the average rent and crime rate across multiple zip codes.

We hypothesize that there is a rent prediction model for every house type within a zip code/similar zip codes. To test this hypothesis, we carry out the following analysis:

Prior to attribute selection, to obtain a suitable represen- tation of the data set, we apply PCA (principle component analysis) to the 18 data set attributes. The attributes consist of ZipID (a unique id for each house in the Zumper API), Number of bed/baths, floor size (the area of the house based on SQF), latitude and longitude (geographical location of each house), year built (the year of house construction), status (house type), zip code, house features (facilities in a house described by owner), estimated rent (basic amount of rent price for each house used as a class label in the prediction task), so forth. Since zip code and house type are nominal attributes, there is no collinearity between the rent price and these two attributes.

B. Feature Selection

To identify important attributes, we apply PCA (principle component analysis) - which is a well-known and studied method- on three subsets of data samples, each subset covering a different house type across all zip codes in the state of Kerala. Unlike town house and condo, features like HOA fee, walk score, and transit score show a very low variance for single family instances. The higher variance of transit score, especially for condos, explains the outpacing of median appre- ciation rates of condos compared to single-family detached- houses in large metropolitan areas [12]. Number of bedrooms is found to be an important feature only when house type is single-family or town-house. Next, unlike town-house in- stances, average school rating is discovered to be an important feature for both single-family and condo instances. This can be explained due to sparsity of school rating for town-house instances in our data set. In our future work, we will employ data mining techniques to obtain this information for town- house instances.

We validated the above mentioned strategy by training our models based on a data set including all house types, and then based on each house type. We discovered that the latter approach leads to relatively higher accuracy and lower prediction error.

C. Data Clustering

The housing data set is first clustered according to home type and zip code attributes, and then a model is finally learned for each cluster. The ability to train the prediction models may be significantly impacted by some of the clusters, which have a very low instance count of under 100. We used a different approach, dividing the data set into three groups depending on the home type feature in order to boost the density of the training samples and facilitate the accuracy of prediction models at the same time. We call these teams status-clusters. The average rent was then determined for each zip code within each status-cluster.

D. Model Evaluation

The three measurements utilised in this section's major comparison measure for regression analysis and model evaluation are: Area Under the Receiver Operating Characteristics Curve (AUC), Mean Absolute Error (MAE), and R-squared (AUROC or AUC). Over the test data set, MAE calculates the prediction models' accuracy. A quadratic statistical scoring formula called R-squared (also known as the coefficient of determination) indicates how closely the actual target data match the fitted regression line. The variation between the anticipated target variable and the actual rent price is displayed in the study using R-squared. The better our model matches the data, the lower the MAE and the higher the R-squared.

A graphical method for visualising prediction model performance and choosing the optimal model based on that performance is the ROC curve. A prediction model's accuracy is measured by AUROC. We compute and analyse the effectiveness of the generated rent prediction models for seven machine learning algorithms MLP, RF, LR, SMO, LWL, KStar, and KNN based on the hierarchical clustering based on these assessment measures. We utilised straightforward unweighted voting for K=3 based on Euclidean distance for the KNN combination function. Figures 2 and 3 compare MAE and R-squared to provide an example.

Fig. 2. Model performance for family house

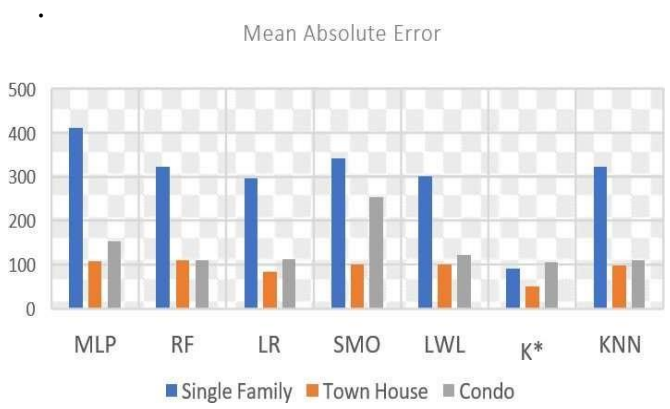

Fig. 3. Comparison of Model performance

IV. RESULTS

This section covers the outcomes of the tests we ran to assess and contrast the effectiveness of the MLP, RF, LR, SMO, LWL, KStar, and KNN algorithms. The KStar algorithm outperforms the competition. Among the algorithms examined for this study, KStar exhibits the lowest variability (R-squared) and the best accuracy (MAE). We contrast the best eager approaches with those of lazy methods based on the overall measure of the model's fit. While KNN and KStar have the lowest variance (R-squared) among the lazy techniques, LR and RF exhibit the best accuracy among the eager methods examined.

KStar also has the best accuracy compared to the other ML techniques examined in this paper. In fact, when compared to the LR technique, the KStar algorithm reduced the prediction error for single-family, townhouse, and condo properties by 69%, 41%, and 8%, respectively. Additionally, compared to RF technique, KStar reduced the prediction error for single-family, townhouse, and condo properties by 71%, 55%, and 5%, respectively. Finally, compared to the KNN technique, KStar reduced the prediction error for single-family, townhouse, and condo properties by 71%, 49%, and 6.8%, respectively. According to figure 5, the skewness of the data set is the reason why there are fewer fluctuations in the MAE measure for townhouse records than for single-family and condo records. This is because single-family residences and then condos predominate over townhouse records in the data set. In addition, Table I reveals that the AUROC for KNN and LWL are relatively similar, whereas the eager version of SMO has the lowest AUROC among the algorithms that were tested. The findings demonstrate that the KStar regression model offers the greatest fit and that overall lazy learning approaches beat eager ones. Regarding the townhouse data, all methods do rather well.

This outcome can be explained by a lack of data in terms of internal/external characteristics and observations. In towns like Kochi and Trivandrum, for example, single-family homes with rent prices over 5000 Indian rupees are extremely scarce and are sometimes offered as "home-offices" by the owner for physicians to rent; these homes are rented as "home-offices" with the medical equipment within the rental property. For townhouse records, we found that the performance of the learning technique LR is fairly comparable to the KNN lazy learning approach.

V. CONCLUSIONS

Eager learning techniques use the complete training set to create a prediction model, which is then tested on the test cases to assess the model's performance. Lazy learning approaches choose the most apt learning samples, minimise local error, and extract the general characteristics of the data, whereas eager learning methods tend to minimise the global error and extract the general properties of the data.

Known Lazy learning techniques can outperform eager learning algorithms in the rent prediction problem for each type of property (single family, townhouse, and condo) in our housing data set, according to our experimental investigation in the real estate industry. As well-known examples of lazy learning techniques, we looked at SVM & TensorFlow and compared them to the eager learning algorithms MLP, LR, SMO, and Random Forest. According to the findings, eager learning techniques may behave in a way that impairs these models' ability to generalize.

Given that high dimensional data is a good fit for deep learning techniques, we intend to use natural language processing technologies to extract metadata from property owner comments on the Zumper website and add more attributes. Additionally, we intend to look at the predictive power of deep learning models like neural networks.

REFERENCES

1. *S. Voghoei, N. Hashemi Tonekaboni, D. Yazdansepas, and H. R. Arabnia, "University online courses: Correlation between students' participation rate and academic performance," in 2019 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 772–777, 2019.*

2. *S. Zad and M. Finlayson, "Systematic evaluation of a framework for unsupervised emotion recognition for narrative text," in Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events, pp. 26–37, 2020.*

3. *N. H. Tonekaboni, S. Kulkarni, and L. Ramaswamy, "Edge-based anomalous sensor placement detection for participatory sensing of urban heat islands," in 2018 IEEE International Smart Cities Conference(ISC2), pp. 1–8, IEEE, 2018.*

4. *N. Etemadyrad and J. K. Nelson, "A sequential detection approach to indoor positioning using rss-based fingerprinting," in 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP),pp. 1127–1131, IEEE, 2016.*

5. *J. K. Nelson and N. Etemadyrad, "Prioritizing goals in cognitive sonar: Tracking multiple targets," in 2018 21st International Conference on Information Fusion (FUSION), pp. 1–6, IEEE, 2018.*

6. *G. I. Webb, Lazy Learning, pp. 571–572. Boston, MA: Springer US, 2010.*

7. *D. S, "Crime in the united states 2011 : Violent crime."http://crimeanalystsblog.blogspot.com/2014/02/how-to-calculate-crimerate.html, 3 February 2014*

8. *A. B. Sanju´an, Model Integration in Data Mining: From Local to Global Decisions. PhD thesis, 2012.*

9. *T. Fawcett, "Roc graphs: Notes and practical considerations for researchers,"pp. 1–38, 03 2004.*

10. *K. R. Harney, "Condos may be appreciating faster than single-family houses." http://wapo.st/2OZc33S, 18 April 2017.*

11. *X. Li, C. X. Ling, and H. Wang, "The convergence behavior of naïve bayes on large sparse datasets," ACM Trans. Knowl. Discov. Data, vol. 11, no. 1, pp. 10:1–10:24, 2016.*

12. *G. Adomavicius and J. Zhang, "Stability of recommendation algorithms,"ACM Transactions on Information Systems, vol. 30, pp. 1–31, Nov. 2012.*

13. *M. Heidari and S. Rafatirad, "Using transfer learning approach to implement convolutional neural network model to recommend airline tickets by using online reviews," in 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA, pp. 1–6, 2020.*