A Technical Report on Real-Estate Rent Prediction

Setareh Rafatirad
Information Sciences and Technology Department
George Mason University

CONTENTS

LIST OF FIGURES

LIST OF TABLES

A Technical Report on Real-Estate Rent Prediction

***Abstract*—Real-estate rent prediction is sensitive to several independent parameters and has allured a lot of researchers in the past few years to constructing automated tools using (ML) commodities. However, most of the proposed solutions are limited in scope, and are only investigated on a particular locality, house type, or based on one type of machine learning algorithm. Furthermore, the past work often used synthetic data which can compromise the accuracy of the output, as it is not closely identical to real-world datasets. To address these challenges, we study a wide range of Machine Learning techniques applied to three real-estate housing types, using real-world data. Unlike prior work which attempt to develop a one-size-fits-all model with fixed set of features, our study shows that the important parameters for rent prediction depends highly on the type and locality. Further, for each property type, there is a different winning algorithm to perform rent prediction. Accordingly, we construct multiple rent prediction models using a large Zillow dataset of 50K real estate properties in the state of Virginia and Maryland. In addition to Zillow, external attributes such as walk/transit score, and crime rate are collected from online sources. Our comprehensive case study indicates that real-estate rent behavior strongly depends on the type of house and locality. As such, we deploy a two-layer clustering approach to partition data into multiple training sets based on house-type and similar zip codes. We evaluate and report the performance of the prediction models studied in this work based on two metrics of R-squared and Mean Absolute Error, applied on unseen data.**

I. Introduction

Predicting the rental price of a Real-Estate property is indispensable in estimating the Rate of Return - a salient index used to evaluate real-estate investment outcome. This topic has allured a lot of researchers due to the availability of data and machine learning commodities [**?**], [11]. In real-estate Net present value (NPV) is an investment criterion which defines the profitability of the investment based on rate of return. The NPV yields an accurate insight to real estate investors on whether they achieve a satisfactory rate of return within a certain period of time. In the equation below, CF is the cash flow generated from a rental property for each period n in the holding period N (i.e., the time period of an investment), and r is the desired investor rate of return.

$$NPV = \sum_{n=0}^{N} \frac{CF_n}{(1+r)^n}$$

Based on this equation, one of the important factors in evaluating the NPV of a real estate property investment is cash flow (CF) that is further calculated based on the following equation for a 12-month period:

$$CF = 11\rho - (12\mu + \tau + \varepsilon + \iota)$$

where ρ is the rent income (the Vacancy rate analysis is not discussed in this paper for the sake of simplicity, and the assumption is that the house has a rent income at least 11 months a year), μ is the house mortgage, τ is the annual house tax, ε is the annual house expenses, and ι is the mortgage house insurance. In the above equation, ρ is only factor that has a positive effect on cash flow. So the more accurate the rent income, the more reliable is the calculated cash flow. Therefore, it is very essential to provide an accurate rent prediction method. For many people, house is an invaluable asset. Therefore, having a safe investment is a significant task. Proper rental property investments can lead to a successful and profitable Rate of Return over time. However, such ventures can be very risky due to miscalculation or inaccuracy of algorithms used in rent prediction. Applying machine learning (ML) algorithms to perform house rent prediction is not a novel trend. However, to the best of our knowledge, this is the first work that considers the major challenges involved in real estate rent prediction, including identifying the influencing parameters based on the house type (i.e., town house, single family, or condo), and discovering the winning machine learning algorithm for the same.

We present a case study in Figures 1, 2 which indicates that there is a rent prediction model for every house type within a locality, bring a zipcode or a group of similar zipcodes. We applied principle component analysis (PCA) technique to dataset internal attributes which suggests that zip code and house type should not be included in the subset of attributes used for training the rent prediction model. As such, we deploy a two-layer clustering approach to partition data based on house type and similar rent behavior across different zip codes, where each cluster serves as the training set for our rent prediction models.

In the previous studies, the prediction models are very generic and they don't differentiate according to the house type or locality [11]. For instance, a generalized prediction model is proposed by [2] for city-wise scope of data, to predict rent and house prices. However, this can lead to inaccurate predictions. Figure 1 shows zip code-wise variation of the rent behavior for the real estate properties in the same state/city and within a close geospatial proximity from each other. For instance, 22066 and 20190 are neighboring zip codes but they show a very different behavior in terms of the average rent price. Also, our study shows that influential parameters for each house type affects the rent price (see Figure 2). The average rent price for a zipcode depends on internal factors (such as house type, number of bedroom/bathrooms, house price, area space, HOA fee) and external factors. In fact, external factors like crime rate and school ratings corresponding to a zipcode impact the price of rent and are deal-breakers for many

Fig. 1. Average rent prices for real-estate properties across different zipcodes. The dark blue shows higher rent prices relative to the bright blue color which indicates lower rent prices.

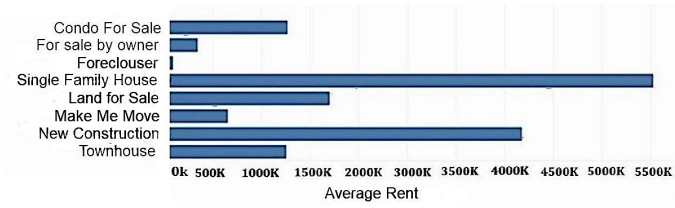

Fig. 2. Rent price distribution for different house types for multiple zip codes.

real estate investors [7]. In this paper, internal and external attributes like walk score, transit score, crime rate, and school rating are deployed. Walk score indicates the errands that can be accomplished on foot or those that require a car to nearby amenities. Transit score indicates the connectivity (i.e., proximity to metro), access to jobs, and frequency of service. Crime score indicates the rate of violent and non-violent incidents related to a zip code [19]. We collected a Zillow data set of 50K real estate properties in Virginia State. In addition, transit score, walk score, and crime rate are collected from information sources like alltransit.cnt.org, walkscore.com, and crimereports.com respectively. Our comprehensive analysis of the transit parameters entertained in the data collected from AllTransit data source clearly indicates the proximity to metro as a significant parameter in determining the transit score of a location. In Figure 3, the distribution of real estate properties across state of Virginia zipcodes is demonstrated. Our dataset consists of three house types: town house, single family, and condo. This study is motivated by the need to build models with respect to house type and locality. Exploring the dataset, it was evident that data within each zipcode is very sparse. To address this challenge, we divided the dataset according to house type, and then applied K-means clustering to generate subsets of instances within the zipcodes with similar average rent prices as illustrated in Figure 4. The clustering method uses the similarity measure of average-rent to compute the distance between the data points. The data samples in each cluster is later used to train a rent prediction model.

Fig. 3. Distribution of real estate properties across VA zip codes. The zip codes in dark orange indicate higher densities.

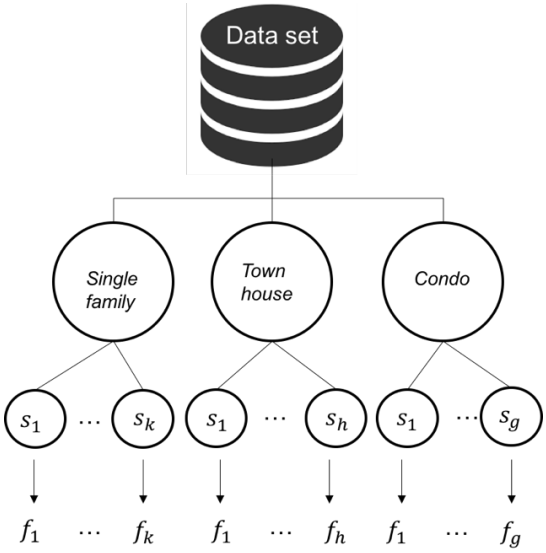

Fig. 4. Two-layer clustering technique, first according to house type, and then based on average rent price.

In this work, we study the impact of several machine learning methods on this data set by performing a comparative analysis of various lazy and eager learning methods. We identify the influencing features for each house type and discover the winning ML algorithm for the same. We examine the performance of Linear Regression (LR), SMO, Multilayer Perceptron (MLP), J48, SVM, and Random Forest (FR) algorithms (eager/globally-based learning) against KNN, ML-KNN, lazy Decision Tree, locally weighted learning (LWL)

and KStar algorithms (lazy/memory-based learning/instance-based), using two performance evaluation metrics: i) R-squared and ii) Mean Absolute Error (MAE). The target variable is the rent price and the evaluation metrics show the variance between the predicted target variable and the actual rent price. Our rent prediction algorithm uses a salient subset of data set attributes, which is determined during feature selection phase using Principal Component Analysis technique (PCA). PCA technique filters out unwanted features based on each house type (i.e., single family, town house, and condo). For imputation, we removed the observations with many missing attributes as the pro-portion of these instances to the entire data set was less than 3%.

The remainder of this paper is structured as follows: In section 2 related work is discussed. We describe the data set used in this paper for analysis in section 3. Section 4 describes our methodological framework including data preprocessing, data exploration and feature selection, building prediction models, and model evaluation. In section 5, experiments and results are discussed. Finally, section 6 gives the conclusion.

II. Related Work

Real estate rent/price prediction using machine learning techniques has been recently studied in several works [12], [13]. Lambert and Greenland [6] investigate eager learning methods like MLP and bagging REP trees to estimate the rental rate for both the land-owners and students interested in renting a place close to a university campus. The training set contains two property types: i) apartment and ii) condo. The coverage area of the training set is limited to three distant zip codes surrounding a university cam-pus. The input features entertained in this work include proximity to university cam-pus, apartment appliances (like Cable TV) and dimensions, the length of the apartment con-tract, and the date of the residence's constructions. The study reports bagging REP trees as the best rent prediction algorithm. However, the proposed global learning-based solution leads to a biased model due to the skewed data set, all located surrounding a university campus. In [15], [16], 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 housing market and determination of house prices [17], [18]. In these studies, economic sub-markets are used in the prediction model which are defined in terms of the characteristics of neighborhoods or census units. The problem with the hedonic approach is disregarding the differences between the properties in the same geographical area.

In the past few decades, machine learning techniques have been widely used to perform prediction and classification tasks in various domains like real-estate rent/price prediction [4], [5]. Khamis and Kamarudin [12] compared the efficacy of the eager learning method Neural Network (NN) against the hedonic model Multiple-Linear Regression (MLR), and showed that NN outperforms MLR. However, Galvan et al. in [3], reports the superiority of lazy learning methods over NN. According to Webb [14], eager learning methods can lead to suboptimal predictions because of deriving a single model that seeks to minimize the average error over the entire data set, whereas lazy learning can help improve prediction accuracy. In this work, we study distinct feature/ML algorithms performance for different house types which is not addressed in the past literature.

III. Dataset Description

Zillow API delivers home details including historical data on sales prices, year of sale, tax information, number of bed/baths, so forth, for the US. In fact, Zillow is tied to various sources like real estate agents, homeowners, tax assessors, public records, and Multiple Listing Service (MLS). Normally, rent prices in real-estate housing do not change abruptly within a very short time window. For ex-ample, rent price of a real estate property is not subject to change every day. We analyzed the real-estate rent prices in different zip codes provided by Zillow, and did not find any drastic changes within a period shorter than 4 months, which suggests that the listed rent prices are reasonably reli-able. We also developed a framework to automatically collect real estate housing data every 4 5 months to ensure the accuracy of our classifier. In this paper, we used the Zillow API to collect a data set of residential housing data for the state of Virginia. The size of this data set contains about 4000 housing property records (including townhouse, single-family, and condo) with 21 attributes. The attributes consist of ZipID (a unique id for each house in the Zillow API), Number of bed/baths, floor size (the area of the house based on SQF), Lot (lot size), 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. In addition, external attributes, namely walk score, transit score, and crime rate are collected. Figure 5 shows the crime rate for violent and non-violent crime incidents in different zip codes. The description of walk score is illustrated in Table 1. Transit score data was collected from All-Transit data source: their dataset is collected from 824 agencies, and it includes 662K stop locations and 13K routs. Transit and walk scores are collected per household, while crime rate is obtained for each zipcode, normalized by the number of people living in that area using Selenium tool with Python. Crime score data was normalized using Dickson method [20] indicated in the following equation:

$$\Gamma = \frac{\chi * 1000}{\Phi}$$

, where Γ is the normalized incident, χ is the number of crime incidents and Φ reflects population.

We obtained zip code-wise population by collecting data from www.moving.com.

Walk Score	Description
90-100	Highest walkability.
70-89	Very walkable.
50-69	Somewhat walkable.
25-49	Car-dependent for most errands.
0-24	Car-dependent for all errands.

TABLE I. CATEGORICAL DESCRIPTION OF WALK SCORE (WWW.WALKSCORE.COM/METHODOLOGY.SHTML).

zipcode	Population	Incidents	ViolentInc
20105	15,021	91	9
20106	5,187	0	0
20109	37,332	17	3
20110	48,019	3	1
20111	34,000	4	1
20112	26,867	0	0
20115	6,551	0	0
20117	2,530	1	1
20119	4,345	0	0
20120	41,180	185	57

Fig. 5. Crime score in the zip code level, including violent and non-violent crime incidents.

IV. RESEARCH METHODOLOGY

A. Data Preprocessing

One of the rudimentary principles in calibration of machine learning models when dealing with a biased data is to resample the data to balance them [1]. As shown in Figure 3, some areas have much higher densities compared to other areas. To normalize the data, we re-sampled the data in zip codes with higher house prices due to their crowded density relative to the zip codes with lower house prices. For imputing the missing values of external attributes, we used K-means clustering and KNN. Figure 6 shows the result of this clustering to impute transit score data. The distances between data points is calculated with respect to each cluster centroid. To reduce the dimensionality of the dataset and enhance the generalization of the mod-el, we perform feature selection by applying PCA (principle component analysis) to all 21 attributes of the data set. However, before applying PCA, attributes are normalized based on Min-Max Normalization, based on the following equation:

$$x_{norm} = \frac{X - X_{min}}{Xmax - Xmin}$$

B. Data Exploration

We analyzed the correlations between various variables of the data set to identify the co-linearity between the variables. Discovering co-linearity between the data set variables and the target variable yields valuable insights about the dependent variables that affect the rent price. While Figure 7 shows the correlations between the internal dataset attributes and the class variable that is average rent price, Figure 8 illustrates the correlations between external attributes - urban planning parameters like walk/transit-score and crime-rate, and the class variable. The general trend in Figure 8 indicates a positive correlation between average rent and walk/transit score, and a negative correlation between the average rent and crime rate across multiple zipcodes.

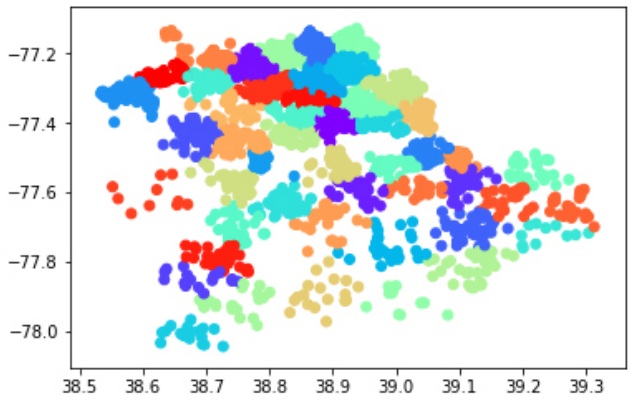

Fig. 6. K-means clustering of transit score data generated 48 clusters with maximum distance 6 miles.

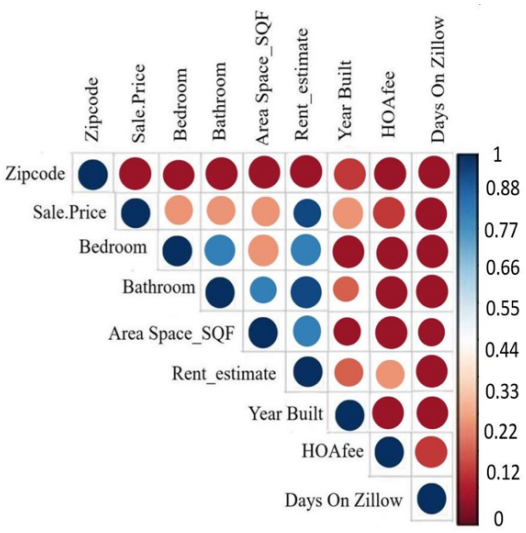

Fig. 7. Correlation matrix for internal attributes of dataset.

C. Feature Selection

To identify important attributes to train an accurate rent prediction model. First, we partition our dataset into clusters of matching house types. Further, the data across each cluster is fed into the Principal Comoinent Analysis method (PCA) to identify indispensable features. PCA method is a class of dimensionally reduction techniques which identifies the most variations in data by rotating the original data to a

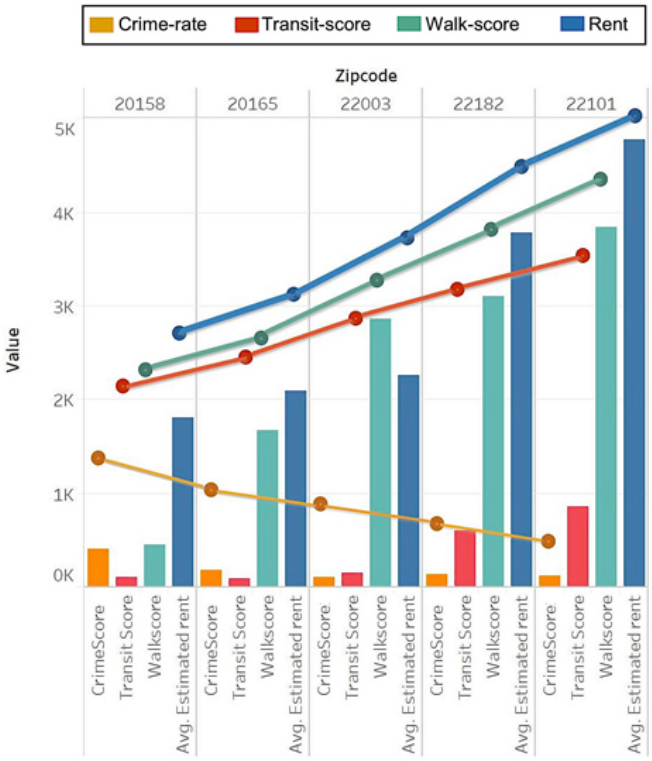

Fig. 8. Correlations between average rent price across multiple zip codes and external attributes.

new variable in a new dimension, known as the Principal Components (PC) [21], [22]. PCs are uncorrelated dimensions and are a linear combination of the original features of data. As a result, for each house type, a number of influencing features are identified. According to Table II, we observed that while some of the features like `price` and `area space` are common across all house types, other influencing features vary depending on the type of house.

TABLE II. SELECTED FEATURES DETERMINED BY PCA TECHNIQUE FOR 3 HOUSE TYPES ('SF' FOR SINGLE FAMILY, 'TH' FOR TOWN HOUSE, AND 'CO' FOR CONDO.

CO	TH	SF	Attributes
yes	yes	yes	price
	yes	yes	bed/bath
yes	yes	yes	area
yes	yes	yes	views
yes	yes	yes	price per SQFT
yes	yes	yes	year
yes		yes	school rating
yes	yes		days-on-zillow
yes	yes		HOA
yes	yes		walk/transit score
yes	yes	yes	crime rate

D. Clustering

In light of the above exploratory data analysis, we partitioned our dataset according to house type and locality. We clustered the dataset on the basis of house type and zipcode attributes to further learn a model for each cluster. However, we ran into a problem and observed that some of the clusters are very sparse with the number of instances below 100, which can immensely impact the ability the training and lead to underfitting which is one of the biggest causes for poor performance of machine learning models [10], and leads to inaccurate results. To deal with this problem, we increased the density of the training samples by first, dividing the dataset into three groups based on the house types; we refer to these groups as status-clusters. Next, we calculated the average-rent for every zipcode inside each status-cluster. Then, we applied K-means clustering to partition the content of each status-cluster based on the average-rent. Using this technique, we increased the density of the training sets which are further used to train the prediction models in this work. We compared the accuracy of the trained models based on two evaluation metrics including R-squared and Mean-Absolute-Error to report the winning machine learning algorithm for each house type.

E. Building Prediction Models

We build rent prediction models with respect to house type and a subset of zipcodes with similar rent prices, using six eager and five lazy learning algorithms, selected from a wide range of popular ML classes including feed-forward artificial neural network, regression, tree-based, and ensemble learning. We used WEKA with a customized setting to carry out the implementation. During the implementation, the dataset is split by 70:30 into train and test sets. In our experiments, 10-Fold Cross-Validation was used to partition the training data set into 10 equal parts. During each round of 10 iterations, we repeat the prediction by using one of the 10 parts as test data and the other 9 parts as training data to create a prediction model. Next, we select the model with the best accuracy. Next, we evaluate the trained rent prediction models on the test data that covers 30% of the entire Virginia housing dataset collected from Zillow website. The important attributes (common and distinct) identified for each house type during feature selection phase are used as input of the models to predict the target variable rent price.

F. Model Evaluation

The key comparison measure used for regression analysis and model evaluation in this section is based on two different metrics: 1) Mean Absolute Error (MAE) and 2) R-squared (R2). MAE measures the accuracy of the prediction models over the test dataset. R-squared (or the coefficient of determination) is a quadratic statistical scoring rule which shows how close the actual target data are to the fitted regression line. R-squared is used in the paper to show the variance between the predicted target variable and the actual rent price. As such, the lower MAE and the higher R-squared, the better our model fits the data. For KNN's combination function, we used simple unweighted voting for K=3, based on Euclidean distance. The comparison of MAE and R-squared is illustrated in Table III.

G. Experimental Results

Even though `single family` data has higher density relative to town house data, it is very imbalanced. To further explain this , we discovered a few samples under this category to have very high rent prices which can impede the model's ability to learn effectively as there is no clear separation in the data. For instance, the houses with the rent price above $9K in cities such as Fairfax and Gainsville are very sparse and listed as home-office by the owner to rent to doctors. We investigated these records and learned that these houses are rented as home-office with the medical equipment inside the rental property.

Based on the overall measure of the fit of the model, we evaluate the ML models deployed in this work and report the top two winning algorithms for each type of house. According to Table III, lazy-KStar algorithm outperforms the other algorithms for Town House (TH) and Single Family (SF) data. Based on our analysis, KStar improved its performance in the presence of noisy and imbalanced attributes. It also shows good capability in dealing with sparsity. KStar is based on the concept of clustering and functions with entropic distance to find the similar instances [**?**]. In contrast with SF and TH data, for CO data, SVM-eager outperforms other algorithms which can be explained due to the balanced nature of the data across its attributes. SVM is a machine learning method that is used for both classification and regression. Based on Table III, SVM generates better generalization and accuracy compared to other methods for CO data. Identifying the second-best algorithm is not as straightforward as finding the best algorithm since both accuracy and variance -indication of how much a model generalizes, are entertained. For TH and SF types, ML-KNN lazy algorithm shows a better trade-off between variance and accuracy compared to LR-eager and KNN. Although KNN algorithm shows a higher accuracy compared to ML-KNN, it does not match its generalization power. For CO type, RF-eager algorithm shows a better trade-off between the deployed evaluation metrics compared to KStar-lazy algorithm. Random Forest [8] is an ensemble learning method which brings extra randomness into the model by searching for the best feature among a random subset of features, it generally results in a model with high accuracy. We also compared the accuracy of the classifiers when distinct features where deployed versus when they were trained based on the common set of features across all house types data, and discovered around 15% performance boost.

	Single Family		Town House		Condo	
Algorithm	R2	MAE	R2	MAE	R2	MAE
MLP-eager	0.58	410	0.91	105.4	0.88	152.4
RF-eager	0.68	322.7	0.78	109.7	0.90	103.3
LR-eager	0.79	294.1	0.92	89.37	0.89	112.7
J48-eager	0.70	280.7	0.87	110.48	0.87	150.48
SVM-eager	0.60	300.2	0.84	120	0.91	101.2
SMO-eager	0.70	342.02	0.90	98.02	0.80	254.1
LWL-lazy	0.86	299.2	0.95	98.1	0.88	121.4
Kstar-lazy	0.95	91.7	0.97	49.3	0.92	109.2
lazy-DT	0.81	399.6	0.95	83.6	0.87	121.6
ML-KNN-lazy	0.82	289.6	0.80	100.6	0.93	108.6
KNN-lazy	0.93	321.065	0.92	97.15	0.92	110.78

TABLE III. COMPARISON OF EAGER VS. LAZY LEARNERS FOR RENT PREDICTION USING VA HOUSING DATA SET. THE VALUES SHOW THE AVERAGE EVALUATION MEASURES R^2 AND MAE. HIGHER R-SQUARED (R2) VALUES SHOW LOWER VARIANCE, AND LOWER MEAN ABSOLUTE ERROR (MAE) SHOWS HIGHER ACCURACY.

V. CONCLUSION

Predicting the rental price of a real-estate property using machine learning classifier is a challenging problem. The selection and training of a suitable machine learning model for this purpose depends on many factors including but not limited to the type of data, influencing features, accuracy and classifier's structure. Our study shows that the influencing parameters for rent prediction highly depends on the type of a housing property. KStar lazy learning showed the best performance in dealing with imbalanced and biased attributes across townhouse and single-family instances while SVM eager learning demonstrated to be the best for condo data in terms of accuracy and dealing with smaller size of the selected features. While this applied machine learning research using real-world real-estate data, is shed-ding light on several key questions in this filed, there are yet important challenges that needs to be addressed. Our future work will investigate these challenges among them predicting future rental price of a property and study it as a time-series problem.

REFERENCES

[1] Antonio Bella Sanjuán, Cesar Ferri Ramırez, José Hernández Orallo, and Marıa José Ramırez Quintana. Model integration in data mining: from local to global decisions, Universitat Politècnica de València, 2012.

[2] Hujia Yu, Jiafu Wu, Real Estate Price Prediction with Regression and Classification, 2016.

[3] Inés M. Galván, Joś M. Valls, Miguel García, and Pedro Isasi. A lazy learning approach for building classification models. International journal of intelligent systems 26, no. 8 (2011): 773-786.

[4] Jerome H. Friedman, Ron Kohavi, and Yeogirl Yun. "Lazy decision trees." In AAAI/IAAI, Vol. 1, pp. 717-724. 1996.

[5] Haleh Homayouni, Sattar Hashemi, and Ali Hamzeh. A lazy ensemble learning method to classification. IJCSI (2010): 344.

[6] John Lambert, Jessica Greenland, Is the Price Right? Prediction of Monthly Rental Prices in Provo, Utah, 2015.

[7] RealEstate-US News Homepage: https://realestate.usnews.com/real-estate/articles/how-homicide-affects-home-values

[8] Vrushali Kulkarni, Manisha Petare, P.K. Sinha. Analyz-ing random forest classifier with different split measures. Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Springer, New Delhi, 2014.

[9] Dayana C. Tejera. An Experimental Study of K* Algorithm. International Journal of Information Engineering Electronic Business 7.2 (2015).

[10] Xiang Li, Charles X. Ling, Huaimin Wang. The convergence behavior of naive Bayes on large sparse datasets. ACM Transactions on Knowledge Discovery from Data (TKDD) 11.1 (2016).

[11] Visit Limsombunchai, Christopher Gan, Minsoo Lee. House price prediction: hedonic price model vs. artificial neural network. New Zealand Agricultural and Resource Economics Society Conference. 2004.

[12] Azme Bin Khamis, Nur K.K.B. Kamarudin. Compara-tive Study On Estimate House Price Using Statistical And Neural Network Model. International Journal of Scientific Technology Research 3.12 (2014): 126-131.

[13] Sabyasachi Basu, Thomas G. Thibodeau. Analysis of spatial autocorrelation in house prices. The Journal of Real Estate Finance and Economics 17.1 (1998): 61-85.

[14] Geoffrey I. Webb. Lazy Learning. Encyclopedia of Ma-chine Learning. Springer US, 2011. 571-572.

[15] Xiaolong Liu. Spatial and temporal dependence in house price prediction. The Journal of Real Estate Finance and Economics 47.2 (2013): 341-369.

[16] Michael Kuntz, Marco Helbich. Geostatistical mapping of real estate prices: an empirical comparison of kriging and cokriging. International Journal of Geographical In-formation Science 28.9 (2014): 1904-1921.

[17] Zeynep Onder, Vedia Dökmeci and Berna Keskin. The impact of public perception of earthquake risk on Istan-bul's housing market. Journal of Real Estate Literature 12.2 (2004): 181-194.

[18] Evren Ozus, Vedia Dokmeci, Gulay Kiroglu, Guldehan Kiroglu. Spatial analysis of residential prices in Istanbul. European Planning Studies 15.5 (2007): 707-721.

[19] FBI: UCR, "Violent Crime".Retrieved: May 6,2018. **https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/violent-crime/violent-crime**

[20] Scott Disckson. How to calculate crime rate.

[21] SvanteWold, Kim Esbensen Paul Gelad. Principal component analysis, Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1, pp. 37 – 52, 1987. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

[22] Carlos Oscar Sánchez Sorzano, Javier Vargas, A. Pascual Montano. A survey of dimensionality reduction techniques, ArXiv e-prints, Mar. 2014.

[23] Setareh Rafatirad, Vivian G. Motti, HomeRun: Assessing Real-Estate Mobile Application to Support Real-Estate Search. Technical Report.