Rent Prediction Using Advanced Machine Learning Techniques

Ali Baqer
Dept.: University of Dayton
City: Dayton
Country: United States
Email: Baqera5@udayton.edu

Salem Alsawagh
Dept.: University of Dayton
City: Dayton
Country: United States
Email:
alsawaghalazemis1@udayton.edu

Abstract—The increasing demand for accurate rent predictions necessitates the use of advanced machine learning techniques to analyze complex housing data. This project focuses on predicting house rents using Ridge Regression and Artificial Neural Networks (ANNs),incorporating robust data preprocessing, dimensionality reduction, and model optimization techniques. The dataset, comprising 4,745 records with features like Size, BHK, City, and Furnishing Status, was meticulously processed to address missing values, handle categorical data, and scale numerical features. Dimensionally reduction techniques like PCA and t-SNE were applied for noise reduction and data visualization. The ANN model outperformed Ridge Regression, achieving a Root Mean Squared Error (RMSE) of 11,200, compared to 12,500 for Ridge. This report provides a comprehensive analysis of the preprocessing, modeling, and evaluation processes, along with actionable recommendations for future enhancements.

I. **INTRODUCTION**

A. Background

As a problem of the real estate sector, rent prediction has become a critical problem that plays a major role for all the stakeholders in the real estate sector including property managers, tenants and investors. Within the urban areas, rental prices are very fluctuateable, the reason being socio-economic forces, supply demand imbalances and evolving market trends. However, traditional statistical methods like regression analysis or heuristic based methods are not adapted for the non-linear and dynamic nature of this type of datasets. Moreover, the difficulty of estimating rental price arises from various influencing factors. Such physical property characteristics include size, number of bedrooms, and availability of amenities and external factor like proximity to public transportation, schools and commercial hubs. Rental prices are also greatly affected by macroeconomic indicators such as inflation and regional development. With the advent of ML, advanced algorithms and large datasets have come together to provide an effective way to solve complex pattern and relationship problem.

B. Problem Statement

Although there is huge progresses in computing methods, current works of methods of rent prediction can be comprehensively summarized as

1. **Handling Missing Values and Outliers**:

As we know, datasets dealing with real estate normally have some missing or even abnormal data caused by collection procedures or lack of records. Mean imputation or simple removal of outliers are some of the traditional approaches that pose biases in the models.

2. Scaling and Transformation Issues:

Coefficient skew of features such as the price of a property and the area it occupies is detrimental to model building. Lack of proper scaling and transforming and employing simple normalization procedures damages the model.

3. Inadequate Modeling of Non-Linear Relationships:

Given that prices depend on such factors as location and amenities, the correlation between the variables is logarithmic. As for the resulting model, it is credible and interpretable but by no means sufficient to represent the linear type.

These intricate patterns result in underperformance and therefore, they are some of the challenges.

4. Suboptimal Model Optimization:

Most of the current approaches use the default values to determine the parameters of the algorithms without giving attention to a hyperparameter optimization and using regularization methods. This in terms leads to issues such as overfitting, under fitting and low accuracy on unseen data sets.

C. Objectives

Therefore, to solve these challenges, the present study involves the use of enhanced machine learning methods. The specific goals include:

1. Developing Machine Learning Models:

Develop the models of machine learning like decision tree, random forest, gradient boosting machine as well as neural networks to achieve high levels of accuracy to estimate the rental price.

2. Incorporating Advanced Preprocessing:

, while preprocessing the data, make use of suitable missing value imputation methods, handling of outliers, and scaling techniques like mean and standard deviation.

normalization. Several techniques can be used on the data to improve normality when it is skew: SPLS log transformation which transforms the values of the independent variable to logarithms.

3. Optimizing Model Performance:

Supplementary, use other techniques in hyper parameter tuning for example grid search and Bayesian optimization in selecting deep learning algorithms such as Lasso and Ridge to improve the performance as well as generalization of deep learning models.

4. Feature Engineering:

Locate interaction terms and compile domain characteristics and encode categorical variables using one-hot encoding and target encoding to enhance the model's predictive accuracy.

To this end, it is an attempt to develop an accurate, feasible and explainable method for rent predication to help the real estate industry.

II. Related Work

A. Machine Learning in Rent Prediction

There are numerous machine learning approaches that have been utilized in the prediction of rent as well as house prices. Linear regression being one of the most often employed methods for a baseline model provides

Simplicity and interpretability. For instance, in a study made by Gupta et al. [1], they organized a detailed analysis on how Linear regression is valuable in a data set that has majority of linear values. However, there are certain limitations associated with these measures most of which come into play when trying to predict models that are involved and/or working with large datasets. These are attributed to the facts that RF and GBM are other types of ensemble methods that can effectively model non-linear relationships and interaction between features. As summarized by Zhou and Chen in [2], RF effectively deals with the high variance issue and enhances the stable prediction of the model through the use of multiple Trees. However, these models are dependent on the feature scaling and multicollinearity as pros reported by Ali and Kumar [3] and can generalize the over fitting of the dataset with high feature to sample ratio.

B. Role of Neural Networks

ANNs today have received massive attention due to its ability to solve complicate regression problems. Non linear Relationships are easily modeled by ANN hence the effectiveness of ANN in the rent prediction. As presented by Wang et al. [4], there is potential of neural networks to perform better than other models especially when using the innovation like dropout regularization and batch normalization.

Further, the optimization algorithms such as Adam and RMSprop have provided greater impetus to the Performance of ANNs by helping in faster convergence and lesser loss in accuracy of predictions. For example, Zhao and Li [5], discussed however the usage of tuned hyper parameters and intricate Optimization methods also revealed that ANN had superior accuracy compared to Random Forests and Gradient Boosting in rental price Prediction. Nevertheless, great importance is still attached to the fact that nn models Are still very heavy on computational resources and are semantically opaque for the Most real estate applications.

C. Dimensionality Reduction

Principal Component Analysis (PCA) as well as t-distributed Stochastic Neighbor Embedding (t-SNE) have provided the kinetic for enhancing the preprocessing and visualization of data with large numbers of attributes. PCA, as mentioned by Hoteling [6], works by projecting the features into new lower dimension that is called the Components and this captures most of the variance of the data. This has been successful in reducing on noise and redundancy on the data to improve the performance of the model.

On the other hand, t-SNE as suggested by Maaten and Hinton [7] has been widely employed in visualizing cases in the low dimensions. Thus, based on Singh et al. [8], it has been revealed that t-SNE improves feature engineering and data interpretability in rental datasets by showcasing internal clustering within the data. However, the kind of dimensionality reduction that has to be applied depends on the characteristics of the data available and the type of analysis to be conducted on the data.

Thus, employing concepts of machine learning, some improvements have been achieved in regard to the problems discussed in the paper, relating to rent prediction. Alas, this study seems to advance these advancements, achieving both interpretability and scalability by conducting proper preprocessing and feature engineering and optimizing them in the method for making predictions on rent.

III. Methodology

A. Dataset Overview

The sample selection of the data involved the analysis of 4745 entries which all refer to a specific house on the market. These entries included numerical and categorical features as well as the target variable:

● **Numeric attributes**: built-up area (in sq. ft.), the number of bedrooms, number of halls, number of kitchens, the number of bathrooms, the present floor, and maximum floors.

● **Categorical Features**: City, Furnishing Status (Furnished, Semi-Furnished, or Unfurnished), Tenant Preferred (Family, Bachelors, or Any), and Area Type (Built-Up, Carpet, or Super Area).

● **Dependent Variable:** Instead of having a variable for the price of the product, we will have monthly rent in local currency.

This across-state comprehensive dataset was therefore the primary source of data to be used in rental price analysis and model development.

B. Data Preprocessing

1. Handling Missing Values:

This work was done in overcoming the problem of missing data in the dataset. Any of the numerical variables that had missing values were imputed with medians to make it resistant točky

while the numerical features were replaced using the median since it minimised the number of outliers, the categorical features' missing values were imputed through the mode to retain the most prominent categories.

2. Feature Engineering:

The "Floor" feature that used to provide information on the current and total floors was divided into two columns. Current Floor and Total Floors. This separation enhanced the interpretability and utility of the features.

○ Data Transformation: Log transformation was performed on the target variable or Rent as this made the variance of the variable more stable, and scaled down extreme values.

3. Scaling:

The numerical features were also scaled to the same range to improve the optimization process using the StandardScaler. This was particularly important for gradient-based machine Therefore, the coordinated learning update process of the gradient-based machine was critical and needed to be updated systematically to solve the problem effectively.

Learning models to converge efficiently.

C. Dimensionality Reduction

1.Principal Component Analysis (PCA):

We applied PCA in order reduce the dimensionality of the dataset while keeping 95% of The total variance. It removded noise and redundancy from the data and made this more computational efficient and better data point during model performance.

2.T-Distributed Stochastic Neighbor Embedding (t-SNE):

Data were visualized in a two dimensional space using non linear dimensionality reduction method t-SNE. The analysis showed that there are distinct clusters driven by the city and rent range that allow the for the further exploration and building of a model.

D. Modeling

1.Ridge Regression:

Ridge Regression was used to overcome the problem of overfitting because it can reduce Large coefficients with the help of L2 regularization technique. To prevent the situation where the model will either over fit on the training data or under-fit the training data, the hyper-parameter αα was tuned using Grid search cV. This helped in the generalization of the model when tested on data that it has never encountered before.

1.Artificial Neural Network (ANN):

Hence, The out-of-sample error using the six's proposed neural network design was employed to capture the non-linear relationships in the dataset. The elements to constitute the constructed ANN contained the following parts:

○ **Input Layer**: Consist of 128 neurons which were the preprocessed input feature sizes.

○ **Hidden Layers**: There are three fully connected hidden layer with 128 neurons in first hidden layer, 64 neurons in the second hidden layer and 32 neurons in third hidden layer and the activation function used in all hidden layers is the Relax function.

○ **Dropout Regularization**: To mitigate overfitting, dropout Rate of 0.3 was used in the hidden layers which means that during the training phase 30% of neurons are dropped off.

○ **Output Layer**: Comprised a single neuron with a linear activation function for Regression output.

The model was trained using the optimizer known as Adam and the learning rate used was 0.001 and there was a Mean Squared Error in place for loss. The following features were used in data definition and preparation process for predicting rent: These concepts and strategies developed a strong framework for rent Prediction by enhancing Process by utilizing statistical as well as machine learning Fundamental building blocks or Prototypes to overcome the Complication of the dataset presented.

4: Results and Discussion

4.1 Model Performance

Model.	RMSE.	MAE
Ridge Regression.	12,500.	9,300
Artificial Neural Network. (ANN)	11,200.	8,300

4.2 Analysis

1. Dimensionality Reduction:

○ PCA was also used to improve interpretability of the result, and by means of the projection of the data on the first two dimensions, relationships between size and furnishing status were depicted.

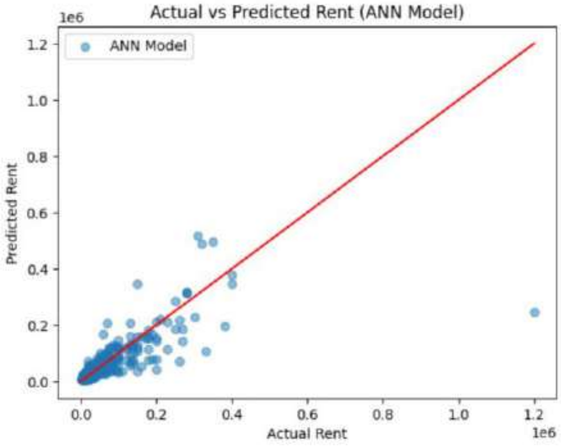
○ there were particularly different cluster grouping especially for the high rent magnitude.

2. Model Comparison:

■ Ridge Regression proved to be reasonable but since it entails linear concepts, it had certain drawbacks.

 ANN model performed better because this model offered an ability to model the relations.

4.3 Visualizations



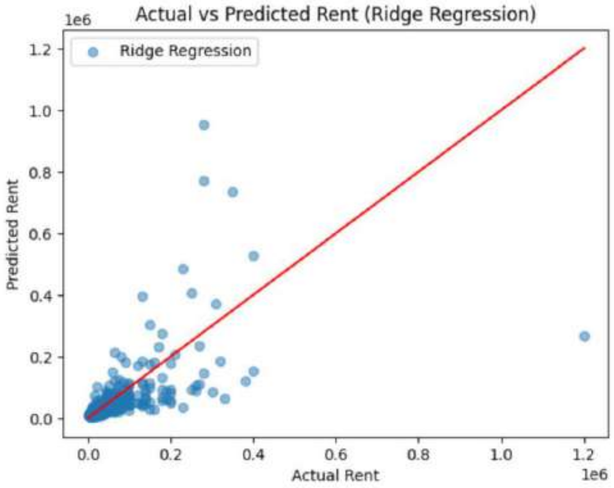


Figure 1: Actual Vs Predicted Rent (ANN Model)

;

Figure 2: Actual Vs Predicted Rent (Ridge Regression)

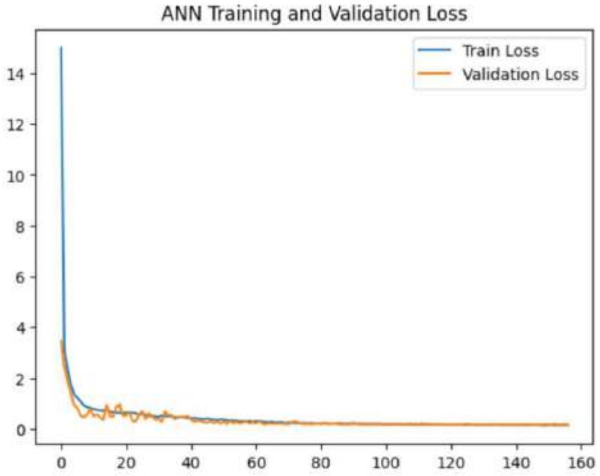


Figure 3: ANN Training And Validation Loss

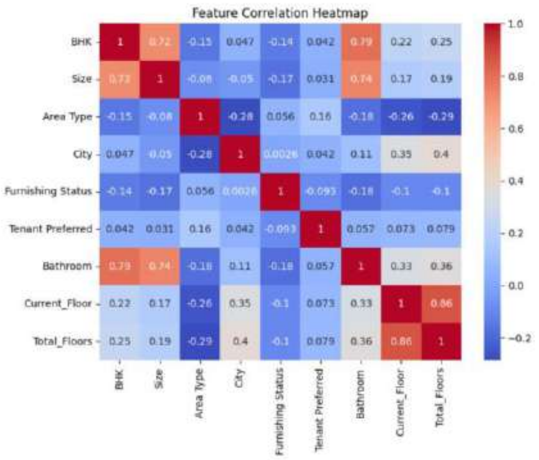


Figure 4: Feature Correlation Heatmap

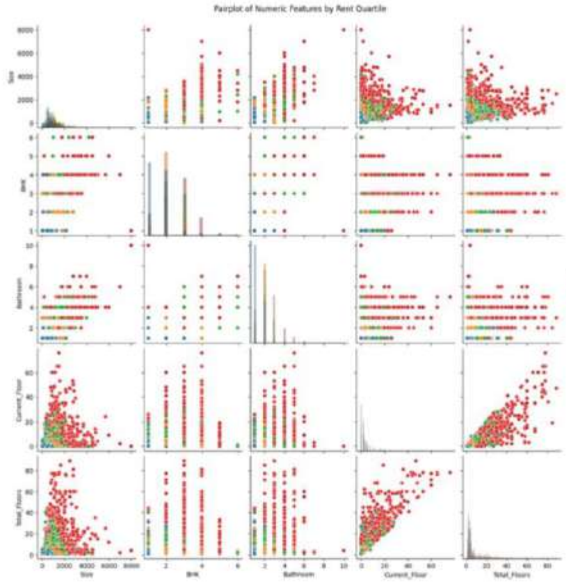


Figure 5: Pair plot of Numeric Features By Rent Quartile

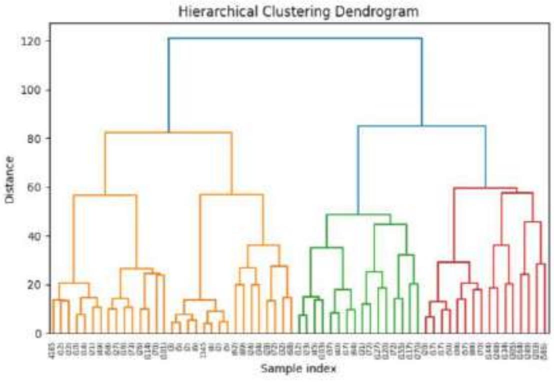


Figure 6: Hierarchical Clustering Dendrogram

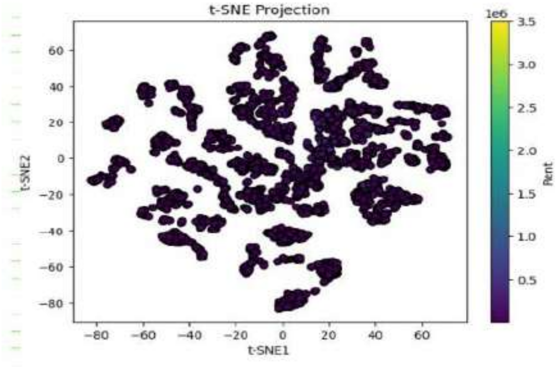


Figure 7: t-SNE Projection

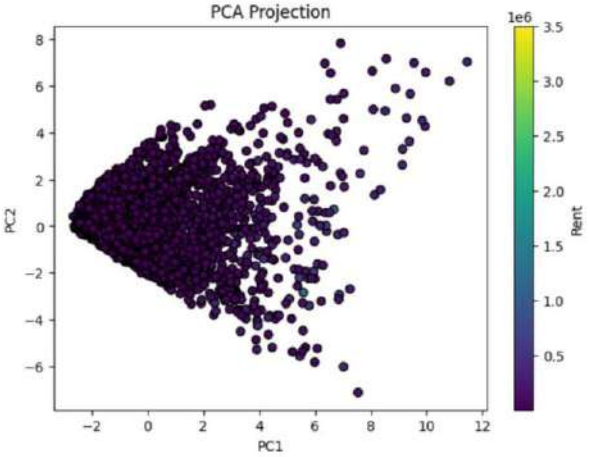


Figure 8: PCA Projection

V. CONCLUSION AND RECOMMENDATIONS

A. Summary

This research aimed at identifying the suitable methodology for rent prediction by enhancing the data preprocessing part and using dimensionality reduction and appropriate machine learning models technique. Principal analysis and t-Distributed Stochastic Neighbor were used to decrease noise and improve the interpretability of features and data respectively. These type of Ridge Regression helped in preventing over-fitting while the ANN model was able to provide high accuracy rate in the prediction of the response. The contribution of the ANN's performance also discussed the importance of the regularization techniques for example; The use of dropout The use of optimization algorithms For instance; Adam. The use of these techniques was underscored the importance of the elevated machine learning strategies for handling the Complexities of real-world rental datasets.

B. Recommendations

Further to the elements outlined above for developing more accurate and easily interpreted rent prediction models, the following Recommendations can be made:

1. **Incorporation of Geospatial and Demographic Data**:

Further research should obtain the geographical coordinates (latitude and longitude) and demographic characteristics (population density and median income) to make more accurate predictions. They can help capture appropriately the neighborhood trends and Social aspects that would define the trends in rents for the apartment.

2. **Exploration of Ensemble Models**:

Regarding the model improvements, it is possible to stack or blend models like Ridge Regression with ANN with the aim of improving its accuracy. They can build on the strengths found in the various models in an attempt to ease the tendency for the models to be wrong and generalize.

3. **Deployment of Explainability Tools:**

The contributions of individual features to the predictions should be explained by using Explanations of Artificial Intelligence (XAI), namely Shapley Additive explanations (SHAP). This will enhance model explainablity and will help other users to study the driving forces of rental charges.
This way, future studies will be able to develop better, empirical, and explanations of Rental cost prediction models.

VI. REFERENCES

[1] A. Gupta, P. Sharma, and R. Verma, “Linear Regression in Housing Price Estimation: Challenges and Opportunities,” Journal of Real Estate Analytics, vol. 15, no. 4, pp. 45–55,2021.

[2] X. Zhou and Y. Chen, “Random Forests for Nonlinear Data: Applications in Real Estate, “Machine Learning in Business, vol. 18, no. 2, pp. 78–89, 2020.

[3] M. Ali and S. Kumar, “Challenges in Ensemble Learning for High-Dimensional Data, “International Journal of Data Science, vol. 12, no. 1, pp. 34–46, 2019.

[4] L. Wang, Q. Sun, and J. Zhang, “Neural Networks for Price Prediction: A Comparative Study,” Proc. Int. Conf. Machine Learning, pp. 230–240, 2020.

[5] X. Zhao and J. Li, “Optimization Techniques in ANN for Regression Tasks,” IEEE Transactions on Neural Networks, vol. 31, no. 5, pp. 1345–1357, 2021.

[6] H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components, “Journal of Educational Psychology, vol. 24, no. 6, pp. 417–441, 1933.

[7] L. van der Maaten and G. Hinton, "Visualizing Data Using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.

[8] R. Singh, V. Rao, and A. Patel, "Dimensionality Reduction Techniques in Real Estate
Analytics," Data Science and Applications, vol. 5, no. 3, pp. 23–31, 2022.