Machine Learning to Predict Housing Prices

Jiayue Liu, Yi-Ju Chao, Isaiah Harris

December 2023

1 Introduction

1.1 Problem Statement: Predicting Residential Home Prices

In the realm of real estate, determining the optimal price for a residential property involves a multitude of factors beyond the conventional considerations such as the number of bedrooms and the presence of a white-picket fence.

This dataset contains explanatory variables that capture nearly every aspect of residential homes in Ames. The variables encompass structural attributes, amenities, location-based characteristics, and more, offering a rich tapestry of information to unravel the nuanced dynamics affecting property valuations. The challenge set forth is to develop a robust predictive model that can effectively estimate the final price of homes in Ames. Innovative approaches, feature engineering, and advanced machine learning techniques are required to uncover hidden patterns, relationships, and dependencies that contribute to the intricacies of real estate pricing.

1.2 Motivation

Predicting housing prices holds substantial importance in real estate and economic contexts. The ability to accurately forecast residential properties values offers benefits for multiple aspects. For instance, it can serve as investment strategy and economic indicators as housing prices are highly related to education resources and urban planning. In addition, academia or policymakers can utilize accurate predictions to study trends, assess housing market, and develop effective housing policies. Furthermore, it helps buyers and sellers to make informative decision while purchasing or selling real estates.

By addressing this problem, we contribute to fostering a more transparent, and equitable real estate landscape, empowering stakeholders with the knowledge to make sound financial decisions and or urban development.

2 Preprocessing the Data

In analyzing the problem, we implemented 4 different models: linear regression,

ridge regression, lasso regression, and a neural network. For the training and test split of the data, we used 75% of the data for training while the remaining samples were used for testing, randomly generated using the train_test_split function, evaluating the accuracy of the models with the mean absolute error and the R^2 confidence score.

2.1 Data Exploration

Before working on the data set, we explored the data to address the presented problem statement properly. We identified the features and drop the unnecessary features. Additionally, we checked for missing values and considered alternative ways to present the missing data. For instance, if the housing data's alley value is NA, indicating the property has no alley, we replaced it with 0.

2.2 Data Preprocessing

The data set includes 79 features, 46 of those being categorical with the rest being numerical. We identified 25 features with a correlation lesser than 10% with the Sale-Price and dropped these features, leaving us with 54 features in total. We initially tried One Hot Encoding. However, it did not work when the test and training data have an inconsistent number of categories for a feature. For instance, if the training data set has A, B, C for feature x, but testing data set only has A and B, it will result in an inconsistent number of columns for our data. Hence, we chose to use label encoding instead which preserve the meaning for most categorical features. For instance, categories such as Excellent, Good, Poor, are assigned values 3,2,1.

For each model we trained and tested with and without normalization to compare the results. When normalizing, we tried three common ways of normalization, namely the StandardScaler, RobustScaler and MinMaxScaler. We also ran each model 10 times and obtain the mean absolute error and accuracy to better evaluate the models.

3 Approaches and Result

model	MAE	accuracy
linear regression without norm	19525	0.8413
ridge regression without norm	19353	0.8421
lasso regression without norm	19397	0.8427
linear regression with norm	19570	0.8411
ridge regression with norm	19347	0.8410
lasso regression with norm	19497	0.8418
neural network	37039	0.4578

3.1 Linear Regression

The basic linear regression ended up being the most inaccurate model of the data. Using data as is resulted in a mean absolute error of 19525 with a confidence of 84.13% and using normalized data resulted in a mean absolute error of 19570 with a confidence of 84.10%.

3.2 Ridge Regression

To train our ridge model, we tried an array of alphas ranging from 0.5 to 100 and chose the model that had the lowest error. Using the data as is resulted in an alpha of 95 being chosen with a mean absolute error of 19353 and confidence of 84.21%. Using normalized data resulted in an alpha of 5 being chosen with a mean absolute error of about 19347 and confidence of 84.10%.

3.3 Lasso Regression

We trained our lasso model similarly to the ridge model, using an array of alphas ranging from 0.5 to 100. Using data as is resulted in an alpha of 95 being chosen with a mean absolute error of 19397 and a confidence of 84.27%. Using normalized data resulted in an alpha of 60 being chosen with a mean absolute error of 19497 and a confidence of 84.18%. Both ended up using all the features in the data.

3.4 Neural Network

We also tried to train using convolutional neural network. We used 1d convolution since our input is a consecutive list of features, instead of images with shape. The input shape is (100, 1, 54) since we have 100 samples per batch, 1 channel and 54 features. In each layer, we double the number of channels and halve the number of features through max pooling. After passing through four layers, the output has shape (100, 128, 3) since there is 100 samples per batch, 128 channels and 3 features. It is then compressed and passed through the fully connected layers where the final output is the predicted housing price. Further adding additional number of layers and channels and increasing the number of epoch will not improve the result significantly.

However, the accuracy for Neural Network remained around 45%, lower than all regressions while the mean absolute error was around 37000, higher than all regressions. Hence, we chose not to use Neural Network as the final model.

4 Conclusion

In conclusion, the exploration of various predictive models for residential home prices, revealed distinctive performances across different methodologies. Basic linear

regression proved to be the least accurate, demonstrating limitations in capturing the intricate patterns within the data set. Ridge and Lasso regressions, on the other hand, showcased improvements in predictive accuracy, with carefully chosen alpha values enhancing model performance, especially when applied to the raw data. Nevertheless, the results are pretty similar among different regression, and perhaps other methods can be considered for better predicting the housing price. Moreover, it is interesting to note that normalizing the data did not significantly improve the accuracy and mean absolute error for all three forms of regression.

Exploration into the realm of Neural Networks, provided an alternative approach. However, the Neural Network's accuracy plateaued at approximately 40%, accompanied by a higher mean absolute error compared to the regression models. Consequently, the Neural Network was deemed less effective for this particular predictive task.

In summary, the Lasso Regression achieves favorable mean absolute errors, and emerges as a strong candidate. However, the decision ultimately hinges on the specific goals, interpretability requirements, and computational considerations of the modeling task at hand.

5 Self-Evaluation and Future Development

The evaluation of predictive models for residential home priceshas provided valuable insights into various regression methodologies and Neural Networks. The report highlights the strengths and weaknesses of each model, emphasizing key performance metrics and rationalizing decisions. However, there is room for improvement, perhaps a more exhaustive exploration of hyperparameter choices, and a comprehensive comparison of computational efficiency can be implemented to enhance the robustness of the prediction.

For the future, we could focus on exploring ensemble methods for mixture models, and incorporating time-series analysis for a better understanding of housing market dynamics. Additionally, we can continue to refine and explore dynamic model updating strategies which would ensure adaptability to changing real estate trends, ultimately leading to more accurate predictions and actionable insights in the field.

6 References

1 Data set
`https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview`

2 Help with linear models
`https://www.kaggle.com/code/apapiu/regularized-linear-models/`

notebook

3 Github
https://github.com/imharris2702/comp-562-final-project
