Final Project – Ames House Price Prediction

David Zang

CIS 4321

Professor Batareshe

Introduction

This project is about building a predictive regression model to predict future home price in Ames, IA. I have chosen to make my model about home prices because everyone needs a place to live. After all, everyone needs a place to live, whether by buying, renting, or loaning a property. But not everyone will make the smartest financial decision when purchasing or renting a place to live. So, by doing this, not only will I be making a smart choice when buying or renting the property at a low price, but I will also be making a smart financial decision. In addition, this is also a huge business opportunity given that we can buy property at a low price and sell or rent them out. And this is why I have decided to build my predictive regression model based on homes.

Problem Statement

The main problem I wanted to explore is about home prices. To be more specific, I had to split home prices into smaller problems. So, things such as what will affect a property price, how will it affect property price, outliers, etc. To explain, when you purchase a property, a lot of things that factor into the cost – how big is the property, history of the property, its neighborhood, how much the seller wants to sell if for, state of property, etc. And these factors are obliviously important given that they affect price, which is what I’m trying to predict. Because of this, there were many factors that I had to consider when building this model.

Speaking of model, I have also chosen to build my model using regression techniques. As we have covered, although both, regression and classification are used to predict, regression is better for this model. I say this because while classification techniques are used to generally predict/classify discrete value, regression techniques are used to predict continuous values (which in this case, is price).

Understanding the Dataset

So, as explained in the problem statement, because I had to be aware of all of factors that contributes to pricing, I had to find a dataset that contained all of those factors beside of those that only had address and price. And I got lucky. Because I never had a specific location, I was simply looking everywhere that had a database regarding an area of housing info. And I found that information on the Ames County website.

This dataset has 81 different categories with 1483 records. We have categories that contain info from the range of zoning to backyard size to number of rooms to type of heating/cooling system to if the property had a garage (and the garage size), and everything in between and more. But most importantly, the dataset has the history price of the property - year sold, sales type, sales condition, and especially, sale price, which was my target variable.

Because the dataset has 81 different categories, the dataset also had all types of data type. We had int64 for categories such as year built, number of rooms, area size, etc., float64 for things such as sale price, building type, addresses, etc., string for neighborhood, whether if the house contained something (yes or no), type of foundation, etc.

And because of all of that, I honestly thought that I would have a hard time preprocessing my data. But, surprisingly, I did not. The dataset didn’t have any NaN values which is nice. There were no problems with getting the different data types to work together. Basically, everything went smoothly, which was pleasantly unexpected.

Data Analytic and Result

I had help with the data analytic. The learning center first helped me create a heat map to determine which categories I should drop, keep, and use. From there, they helped me create feature selection and scalar for standard deviation. This was important because the heat map helped me filter out the important categories to use while the feature selection will be used to help build the different regression techniques. And scalar was used to help prevent categories with large variances (such as overall quality) from affecting the training and testing. After that, I started to build the different regression techniques, which I will skip (like, skip in terms of me building the different techniques). The final result is the RSME cross-validation, which is calculated using MAE, MSE, RMSE, and R2.

The techniques I chose to use was – linear, ridge, lasso, and random forest. I first chose linear regression because I remember that this is one of the most used algorithms that tries to predict continuous values (which is price in this case). And when comparing the result with the other 3 techniques, linear provides the highest cross-validation score, which was a good sign.

The next thing I chose was ridge and lasso. I chose to use these two because just like what our Regularization and Shrinkage collab states, my database suffers from overfitting of features to the target variable (such as if the bathroom had a full bath or half bath to the price of house – to be more specific, the database had full bath and half bath in two different categories. If the bathroom had a full bath, it was a 1, else it was a 0. If half bath or full bath had a 0, that category becomes useless and overfitted). In addition, I also had few predictors with large coefficients that can overpower the whole prediction (acers, year sold, and total square feet of all floors combines). Because of this, I had to use ridge and lasso to regularize and shrink. And the result is that lasso scored the second highest cross-validation score (only smaller then 11 compared to linear) while ridge scored the third highest (46 smaller compared to lasso).

The last one is random forest. I chose random forest because I simply wanted to build the tree visualization again. But I couldn’t get the visualization code to work, so I simply give up on the visualization. And I suspect that it was because I copy pasted the linear regression code and changed it to work with the tree instead of building it from scratch (I did the same thing with ridge and lasso). But although the visualization didn’t work, the tree itself still worked. And surprisingly, although the cross-validation score was the lowest, it was 4000 points lower compared to ridge. And this was surprising to me because I though that tree would have the second highest score given that the tree is the other most used technique in this kind of situation. But because the tree was still in range compared to the other 3 technique results, I’m not that bothered by it. I was simply surprised that it scored a lower score.

Conclusion

As seen, I have built a predictive regression model to predict future home price in Ames, IA. By using house price as the target variable, using two most commonly used techniques for this situation, and two techniques to regularize and shrink, I am quite confident that I have built an effective model that can predict future home price in Ames. And I say this because the results for all 4 techniques are quite close to each other. And because the scores are quite close to each other, it shows that all the techniques arrived to similar result, which is great for us (cross-validation score for linear is 35933, lasso is 35922, ridge is 35887, and 31495 for ridge).