Progress Report Home Price Prediction in Ames, Iowa – Team 84

Background:

The United States, as of 2020, has a housing shortage of over 3.8 million dwellings and trending upwards.[1] This critical shortage can have a long-term detrimental impact on the displayed persons and the greater community. To fix this issue, a variety of industries such as building contractors, realtors, sellers, and public policy administrators all must come together to provide not only homes but also homes that buyers desire and can reasonably maintain.

Real estate is unique when compared to other industries. Houses are durable and last multiple decades. It is a necessity, and at the same time, it offers stable growth; hence, it is one of the best long-term investment options. Every house is unique in terms of style, design, features, location, and construction type, making the pricing highly complex. The real estate business can be very lucrative if you understand the key factors that drive people to buy or sell homes. It has been challenging to predict the price of a home because there are a lot of factors contributing to its value.[2] In addition to the physical and geographical aspects of a home, there is a temporal component to home pricing too. Since the pandemic, US home prices have soared compared to income growth due to multiple factors, including inflation, rising interest rates, supply chain issues, labor constraints, and the economy's overall health. All these factors play an important part in consumer confidence, which impacts the willingness of someone to invest in real estate. Going hand in hand with when to buy/sell a home is determining what a consumer wants in their home. What features of a home are a necessity versus a nice to have and a do not want to have? Essentially, when considering buying/selling real estate, it is critical to understand what aspects of a home a buyer is willing to pay for and what a buyer sees as a liability. This suggests that the ability to create a model that can predict home prices based on various factors about a home could be very beneficial.

Hypothesis:

We expect to build a model that has a minimum RMSE < 6000 and $\boldsymbol{R^2}$ > 80% when predicting against the test dataset. Iterative methods i.e., using feature engineering, dimension reduction for model building, and model selection will determine the conclusion of our analysis. We hope to find a reduced subset of attributes that can predict the sales price.

Domain focus for this project has been on real estate and the results of our analysis will benefit realtors or home buyers in looking for their perfect house or making a profit(realtors). This can also help contractors decide what to include in new developments. Some of the potential benefits include

1. Maximizing profit with minimal investment for real estate promoters
2. Easy risk assessment for the lenders
3. Buyers can be assured that they are paying a fair price
4. A promoter can decide what feature to offer in a particular neighborhood

Proposal:

The purpose of this project is to create a model that elucidates how and what physical and geographical aspects of one's home impacts the sales price. Traditionally, it is understood that having more beds and baths increases the home price. Similarly, having a home in a low crime area and better schools tend to increase home prices.[3, 4]. A study in 2016 by Ozgurs examined a set of twelve factors to predict home prices in Indiana. This study ended up showing the Home Owner's Association Fee (HOA) is the best linear predictor of sales price, demonstrated by checking the linearity assumptions. They show in this paper that even with twelve factors, price predictions are quite difficult especially with many outliers in the data. The novelty of this project is that this project examines roughly 85 factors taken from both the Ames housing dataset compiled by Dean De Cock and from *Neighborhood Scout*. The Ames housing dataset only includes physical aspects of homes while the factors that were scraped from *Neighborhood Scout* are environmental in nature like the level of crime, quality of the school district, and a variety of other attributes for neighborhoods in Ames, IA.

Planned Approach:

We have downloaded the Ames housing dataset from Kaggle, using Kaggle APIs. We have also scraped environment data from *Neighborhood Scout*, using web scraping. Both these datasets were merged to create a master dataset for our exploratory analysis and model training. We have removed the outliers, imputed missing values for dependent variables, performed numeric conversions for factor variables, and scaled and normalized the values as required. We are working on feature engineering to determine the key features that are statistically important in influencing the dependent variable (sales price).

We are currently in the process of training various regression models – Stepwise Regression, Random Forest Regression, AdaBoost Regression, Gradient Boost Regression, Linear Regression, and Multilevel Perceptrons Regression models. We also plan to use PCA decomposition of the data and build the above-mentioned models and compare their output. We will primarily use R^2 as the key factor to compare model performance. In addition, we will use model loss chart, MSE/RMSE charts for model selection, Residuals/Histogram of residuals, and QQ plots.

Hyper-parameter tuning/optimization will be done in the regression setting once we finalize the model. We will add various weights to minimize the RMSE and maximize $\boldsymbol{R^2}$.

Environmental DataSet Challenges:

We hoped to include a dataset on environmental factors that pertain to specific neighborhoods, which in theory could help us create more granular models. This data was scraped from *Neighborhood Scout*. Figure 1, shows the breakdown of neighborhoods by color, there are 18 in total. The Ames housing dataset included 25 neighborhoods. Unfortunately, the breakdown of neighborhoods and their names were different and so a third table was needed to join them. This table was created by manually assessing where the Ames Housing dataset neighborhoods "fit" into the *Neighborhood Scout* neighborhood division. Surprisingly, only one neighborhood from the Ames Housing dataset did not cleanly map over. To resolve this, the Ames Housing dataset neighborhood was assigned to

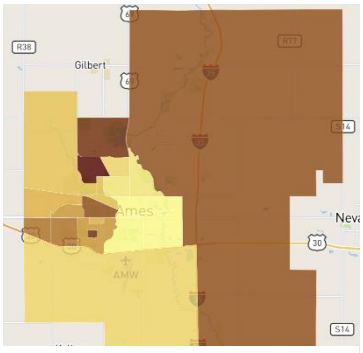

Figure 1: Neighborhood Scout's Neighborhood Break Down for Ames, IA

the neighborhood, in Figure 1, that seemed to contain most of the residential housing. When creating models and looking at correlations, discussed later, it turns out that each Ames, Iowa neighborhood cannot be broken down well into its respective environmental factors, due to many of the Ames Housing neighborhoods fitting into the same divisions in the *Neighborhood Scout* dataset.

Data Cleaning:

We have used a two-prong approach to doing data cleaning and analysis, one in R and the other in python. For analysis in R, factors have made data cleaning very trivial. Many of the variables, in the housing dataset, use a "NA", implying "Not Applicable". An example of this could be not having an "Alley" next to your home. R can easily turn the character "NA" strings into a "NA" factor. Three other variables that are numeric in nature, 'Lot Frontage', 'GarageYrBlt', and 'MasVnrArea' include a 'NA' when they are not part of the home. It was decided, as a first attempt, to treat 'NA' as zero. This assumption is intuitive for "Lot Frontage" and 'MasVnrArea' (Masonry veneer area in square feet) but not so much for GarageYrBlt, which debatably could be categorical even though garages have been created in many different years. It should be noted that there are other variables that do capture a garage not being built by showing the number of cars spots as 0 or a value of "NA" for GarageFinished. Our python analysis used a similar approach of setting missing data to 0 and including dummy variables where needed.

Type of Models Implemented:

As a first attempt to understand the data, a correlation diagram of the numeric variables was created, as shown in Figure 2. This chart has some satisfying results such as 'OverallQuality' and 'GLivArea' are strongly positively correlated to 'SalesPrice', which is intuitive. It was a bit surprising that '1stFlrSF' was strongly correlated with 'SalesPrice' while '2ndFlrSF' was only weakly correlated with 'SalesPrice'. This analysis again only includes numeric variables but at least is a good first attempt.

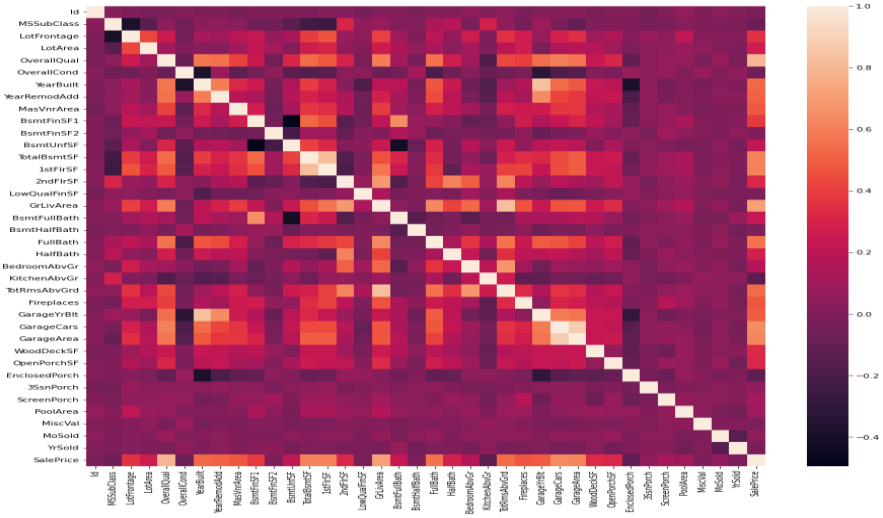

Figure 2: Correlation Diagram of the Numerical Variables

After gaining some insight into what the correlation between 'SalesPrice' and other factors the team then pursued modeling from two avenues. A portion of the team explored model design by fitting and testing various models including: BSplines reduction with Linear Regression, BSpline reduction with Group Lasso, Lasso Regression with scaled data, Ridge Regression with scaled data, Linear Regression following PCA (Principal Component Analysis) with scaling, and Linear Regression with various reduced parameters. The focus was to use both a training and test set with a 80/20 split in order compare the resulting R^2 and mean square error (MSE) of the various models against the test set. The models with the best performance will then be used against the chosen set of important features.

A preliminary analysis of feature engineering was also started. The initial attempt was to create a linear relationship between 'SalesPrice' and all other independent features. All variables that had a confidence of over 99.9% were saved and shown in the Appendix Table 1. A variety of features, including all environmental features, were not able to be fit due to the lack sufficient differences in their values between records. This suggests that the best we can do with the environmental data is to essentially accept that the differences in neighborhoods are encapsulated by their respective categorical dummy variables. Figure 3 shows some of the critical plots for this linear regression. Figure 3a indicates linearity in the model. However, Figure 3b indicates the presence of heteroskedasticity. Figure 3c indicates the presence of outliers while Figure 3d indicates the presence of leverage points. The various Figure 3 plots suggest that it will be critical to evaluate the outliers and leverage points to see if they are candidates for elimination. Another interesting aspect of this fit is that feature NeighborhoodStoneBr otherwise known as the Stone Brook Neighborhood seems to be highly correlated with 'SalesPrice'. It may be the case that Stone Brook is a very wealthy neighborhood and that is why it is more correlated to 'SalesPrice' than other neighborhoods, but this will be part of the further investigation. One other consideration that will be important to include is the presence of multicollinearity, which was not in the feature engineering above.

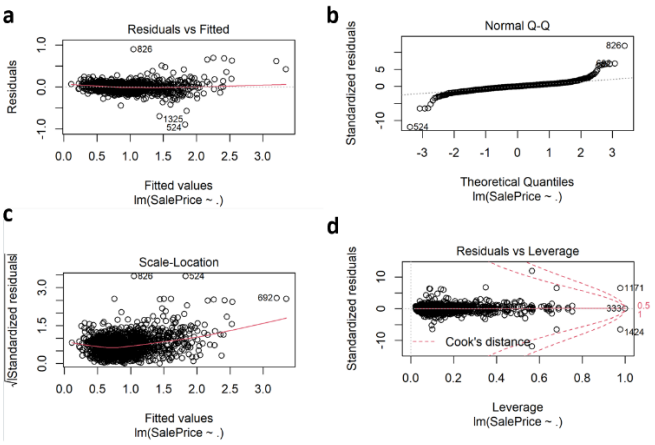

Figure 3(a) A plot of Residuals and Fitted Values (b) A plot of Standardized residuals and Normal Q-Q. (c) A plot of the square root of standardized residuals and Fitted values. (d) A plot of Standardized residuals and Leverage.

Next Steps:

The next steps will be to remove outlier followed by settling on a set of features. These features will then be used for training and testing our few best models. We hope that the resulting model will result in a high R^2 and low MSE. Ideally, the model will incorporate as few features as possible to retain the high metrics.

References:

1 https://www.freddiemac.com/perspectives/sam-khater/20210415-single-family-shortage, accessed 07/04/2022 2022.

2 Ozgur, C., Hughes, Z., Rogers, G., and Parveen, S.: 'Multiple Linear Regression Applications in Real Estate Pricing', Business Faculty Publications, 2016, 61.

3 Ceccato, V., and Wilhelmsson, M.: 'Do crime hot spots affect housing prices?', Nordic Journal of Criminology, 2019, 21, (1), pp. 84-102.

4 Davidoff, I.A.N., and Leigh, A.: 'How Much do Public Schools Really Cost? Estimating the Relationship between House Prices and School Quality', Economic Record, 2008, 84, (265), pp. 193-206.

Appendix A:

Table 1: Below are the independent variables that all have a greater than 99.9% confidence for a linear model of SalesPrice to all other features.

Column Name	Pr(>|t|)
LotArea	5.64E-11
LandSlopeSev	0.00023
NeighborhoodStoneBr	2.66E-06
Condition1Norm	9.51E-05
Condition2PosN	< 2e-16
OverallQual	2.69E-11
OverallCond	5.10E-11
YearBuilt	3.69E-05
RoofMatlCompShg	< 2e-16
RoofMatlMembran	< 2e-16
RoofMatlMetal	< 2e-16
RoofMatlRoll	< 2e-16
RoofMatlTar&Grv	< 2e-16
RoofMatlWdShake	< 2e-16
RoofMatlWdShngl	< 2e-16
MasVnrArea	0.000355
ExterQualGd	1.64E-05
ExterQualTA	0.00016
BsmtFinSF1	9.33E-13
BsmtFinSF2	0.000518
BsmtUnfSF	2.09E-05
X1stFlrSF	5.69E-15
X2ndFlrSF	< 2e-16
KitchenQualGd	2.01E-11
KitchenQualTA	1.05E-08
GarageQualEx	6.55E-05
GarageCondEx	0.000956

Appendix B:

Below table 2 gives the status of the Project.

Proposal	Team Members	Start	Complete	Status
Explore Data Sets	ALL	6/6/2022	6/10/2022	☑
Choose Data Set	ALL	6/13/2022	6/17/2022	☑
Data download APIs	Jonathan, Vignesh	6/15/2022	6/20/2022	☑

Data Exploration	Siddharth, Reena	6/15/2022	6/20/2022	☑
Project Timeline creation	Vignesh, Siddharth	6/20/2022	6/22/2022	☑
Proposal Document	ALL	6/20/2022	6/22/2022	☑
Development/Progress Report	**Team Members**	**Start**	**Complete**	**Status**
Generate Dataset and Samples	Jonathan, Reena	6/23/2022	7/8/2022	WIP
Initial Software Setup	Siddharth, Vignesh	6/23/2022	7/8/2022	WIP
Write Progress Report	ALL	7/4/2022	7/6/2022	☑
Progress Video	ALL	7/4/2022	7/6/2022	☑
Hyperparameter tuning	Jonathan, Reena	7/11/2022	7/15/2022	On-Track
Visualization	Siddharth, Vignesh	7/11/2022	7/15/2022	On-Track
Final Report	**Team Members**	**Start**	**Complete**	**Status**
Final Video Presentation	ALL	7/18/2022	7/20/2022	On-Track
Code and Data package	Siddharth, Reena	7/18/2022	7/24/2022	On-Track
README-User Guide	Jonathan, Vignesh	7/18/2022	7/24/2022	On-Track
Final Report Slides	Jonathan, Vignesh	7/20/2022	7/24/2022	On-Track
Final Report	Siddharth, Reena	7/20/2022	7/24/2022	On-Track