Abstract:

This paper develops and helps in understanding the house market and how significant are other features when it comes to selling and buying the house. For example, comparing this model with a human – a bunch of features make up a person, there a bunch of features that make up a house more valuable. The detailed analysis talks about what feature co-relate with the price and make it a more valuable asset for an individual that is buying or even selling the house. The paper likewise talks about different categories that are not important for this experiment and were eliminated after the feature selection was done, four models that would help predicting the housing prices: KNN, SVR, Decision Tree and Random Forest, the hyper tuning that was done with each specific model and lastly the results that each model gave. The paper also explains why which model performed better. Lastly, the paper gives an overview about the housing market in general, the rise and fall of it and explains how a benefit to this study is that we have two clients at the same time – the seller and the buyer. The Data Set includes data that was gathered from 2007-2010 in Ames, Iowa.

Introduction:

The 2008 Financial crisis affected the lives of everyone in the United States – and there is a lot of talk and research that seeks to identify which sector of the economy was hit the hardest and how did it affect the others. That one thing that comes to attention is housing as the financial crisis was characterized by a rise in mortgage prices.

Since the housing prices are very important and are a reflection of the economy, the prices ranges are of great interest for both the buyers and the sellers. In this report we explore building a predictive model for the sale price of residences in Ames, Iowa using the Ames housing dataset. The applicability of predicting property sale prices is widespread in today's market. For example, many consumers rely on online websites such as Zillow to inform both their buying and selling of real estate. The focus was mainly on variables that consumers commonly ask about when they are looking for a house, such as the square forage, the year the house was built, the year the house was remodeled etc. This analysis gave us results that the total living area about grade, year built and the mean price per square foot by neighborhood have a very positive correlation with the Sale Price which makes a lot of sense.
The house prices will be predicted with various techniques like SVR, KNN, Decision Tree and Random Tree Regression. The main goal of this project is to create a regression models that are able to accurately estimate the price of the house with given features. The data set consists of large number of categorical variables associated with the data set. They range from 2 to 28 classes with the smallest being STREET (gravel or paved) and the largest being NEIGHBORHOOD (areas within the Ames city limits).

Literature Review:

Within the next two decades, the world's population is going to increase 60% of the total population (Alberti, 2010; Pickett et al., 2010; Rees and Wackernagel, 2008). Current urban land areas which are almost around 6% of the total global land surface will have to be expanded to meet this growth and already expansive urban footprint that influence the climactic systems within and around the cities. This expansion will bring in extension, loss and disruption of the ecosystem (McKinney, 2002, 2006, 2008). As these changes start to occur, communities will have to make choices about the land that is not being commercially used but will be used (Alberti et al. 2003). This further relates with the housing prices and features that are being discussed in this paper.

In the recent years, studies that are dealing with house pricing models have increased more in number. This is because more people are investing in buying houses according to what features compliment their style. The

increasing availability of data sets from different cities and areas about housing data has contributed to the study of housing pricing and the understanding of what is important to people and what is not.

Generally, the percentage increase of a housing price happens to due to the price increment in the raw material that is used to build the houses as well. Looking at different journals and articles, there is a lot more that leads to the factors of house price determinations. One of the studies shows that income (demography trends) and nominal interest rates are some of the main factors that are highly dependent on the housing prices. Another factor that really matters is the equity returns which area a big influential factor (Pages & Maza, 2003).

Since this data set only gives us details about the Housing market in Iowa and specifically in Ames, the housing market can also be compared with various other regions to analyze and compare. Tan, in his study about the housing marketing in Malaysia, worked on a regression analysis showed that, income, unemployment rate, total loans played a huge role in the Malaysian housing prices. Which clearly means that the economic and financial factors played an important role – even though the paper talks about the physical features that the customer wants complimenting their style, it is interesting to see what other factors are important in different parts of the world.

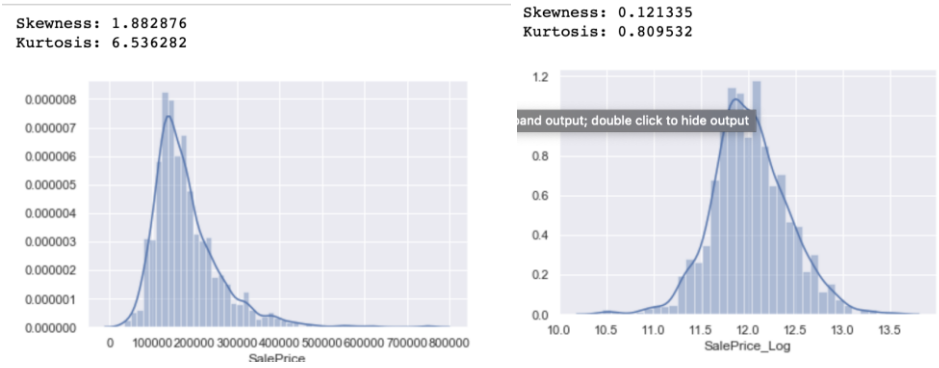
Real estate is one of the most important production factors of the housing business. When the house prices increase, the cost of rent or purchasing a house also increases with it (Gao Bo, 2012), proportional to these factors, this also increases the price of the land and all the materials that are being used (Liu, Lin, 2013), which also increases the cost of plant construction and other industrial factors that come along with this business. The main point is, that since one factor is so dependent on the other, the higher the prices are, there are many other factors with it that are increasing. For example, other than the factors that were listed above, in order to meet the increasing money supply for the demand, it also increases the living expensive of the labor force, resulting in a further rise in labor wages since the industry is growing. If more people want houses, each side of the industry starts growing vastly.

In the long run, the price of the houses is going to be influenced by the land, where is the land, how strong is the structure of the house and the financial system. However, in the short run, the changes in the land will then further effect on the user costs, income, what features would they want to see in their houses, how much are they willing to spend for them and how much mortgage loan can they take from banks to be able to fulfil the wanting of a house.

Methodology:

The data set contained 79 explanatory variables which described almost every aspect of residential homes in Ames, Iowa and the idea was to predict the final price of each home. Since the data set included 79 variables, this would help to see which ones of them actually mean a lot more as compared to the others when buying a house or selling it; or what features are not that important and can be eliminated before the Models are being run for prediction. There were two methods used to see which variables were important – Correlation Matrix and Feature Selection (wrapper method).

If somebody was to glance at the data set, there is a lot of information that is probably related to the Sale Price, for example the area, the neighborhood and the condition and quality. Maybe some of the features are not so important for predicating the target, also there might be a strong correlation for some of the features like Garage Cars and Garage Area.



As we see from Graph on the Left side, it shows that the Target variable **SalePrice** is not normally distributed, which means that this can reduce the performance of the ML Regression Models because some of them assume Normal Distribution. To change this, Log Transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patters in the data more interpretable and for helping to meet the assumptions of the inferential statistics. Log transformations are necessary to reduce the skewness of continuous data when plotted against each other.

Missing Data and Variable Selection:

Looking at the data in the detail for missing values, there were a few columns that showed NaN entries. However, after reading through the data description in detail, these were not the values that were missing but this meant that the features were missing from the houses. For example, if it listed NaN in front of the Pool, Fence or Fire Place etc. it didn't mean that the data was not recorded correctly, but it meant that the particular house didn't have these features.

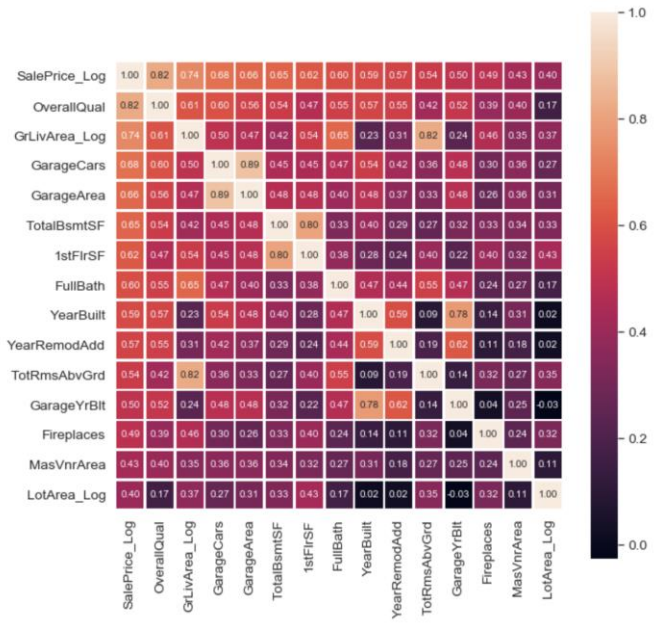
	Total	Percent
LotFrontage	259	0.177397
GarageYrBlt	81	0.055479
MasVnrArea	8	0.005479
SalePrice_Log	0	0.000000
ExterCond	0	0.000000

This highest percentage of the missing values was 17% approximately and the lowest was 0.5% - and this was taken care by taking the mean of the values.

To understand the variables, more in detail – there are two types of features in the housing data – categorical and numerical. Categorical data is not always necessarily linear but it follows some kind of pattern. For example, take a feature of "Downtown" – the response in the dataset is either "Far", "Near", "yes", "no". We are not sure what either of these could mean specifically and neither can we change the responses to be better or worse than the other. Numerical data is in the number form – these features are in a linear relationship with each other. For example, a 2000 square foot place is 2 times bigger than a 1000 square foot place. The main idea is that with 81 features, how could you possibly tell which feature is most related to house prices? To manage this and to have a solution about this, there can be a correlation matrix that can be built.

The most correlated feature to the Sale price is Sale price, which makes the most sense. For the other variables, below is a short description for each one of them to have a better understanding and how they relate to the pricing of the house.

1. OverallQual: Rates the overall material and finish of the house (1 = Very Poor, 10 = Very Excellent)
2. GrLivArea: Above grade (ground) living area square feet
3. GarageCars: Size of garage in car capacity
4. GarageArea: Size of garage in square feet
5. TotalBsmtSF: Total square feet of basement area
6. 1stFlrSF: First Floor square feet
7. FullBath: Full bathrooms above grade
8. TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
9. YearBuilt: Original construction date
10. YearRemodAdd : Remodel date (same as construction date if no remodeling or additions)
11. TolRmsAbvGrd:
12. GarageYrBlt: Year garage was built
13. Fireplaces: Number of fireplaces
14. MasVnrArea :Masonry veneer area in square feet
15. LotArea_Log: Lot size in square feet



Model Generation:
For the Model Building part, there were four models that were built, and were compared to each other in the end to see which one did better. KNN, Decision Tree Regressor, Random Forest Regressor, SVR.

SVR provide another method of studying linear relationships while also being adaptable to nonlinear relationships. It can also be used as a regression method, maintaining all the main features that

characterize the algorithm. It uses the same principles as the SVM for classification with only a few minor differences.

KNN is a Supervised Machine Learning Algorithm as the target variable is known. Non-parametric as it does not make assumption about the underlying data distribution pattern. All the data points are used only at the time of the prediction, with no training step but prediction is costly. Used for both Classification and Regression, it uses feature similarity to predict the cluster that the new point will start falling into.

Decision Tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into small subsets while at the same time an associated decision tree is incrementally developed. The final result, results in nodes and leaf nodes.

Random Forest is a Supervised learning algorithm which uses ensemble learning method for classification and regression. Random Forest is a bagging technique and not a boosting technique – the trees in random forests are running in parallel when there is no interaction between these trees while building the trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classification or mean prediction of the individual trees.

Model validation and results with and without Feature Selection:

Looking forward, the two models that performed the best were Decision Tree and Random Forest. The results below that are discussed are with and without feature selection for both the models. The other models also performed well, but these two models had results very close to each other, which would make the most sense looking at them in more detail.

Without Feature Selection:

Random Forest	**Explained Variance : 0.8439** **RMSE : 0.15390** **CV Runtime : 0.05412**
Decision Tree	**Explained Variance : 0.71395** **RMSE : 0.2047** **CV Runtime : 0.042133**

With Feature Selection:

Random Forest	**Wrapper Select: Selected [Sale Price, Overall Quality, GrLivArea, GarageCars , GarageYrBlt , OpenPorchSF, BsmtFullBath,** **FullBath]** **Features (total/selected): 41 8** **Random Forest RMSE:: 0.22 (+/- 0.03)**

	Random Forest Expl Var: 0.71 (+/- 0.06) **CV Runtime: 0.0612**
Decision Tree	**Wrapper Select: Selected [Sale Price, Overall Quality, GrLivArea, GarageCars , GarageYrBlt , OpenPorchSF, BsmtFullBath]** **Features (total/selected): 41 7** **Decision Tree RMSE:: 0.23 (+/- 0.02)** **Decision Tree Expl Var: 0.68 (+/- 0.04)** **CV Runtime: 0.04237**

To look at the results in more detail, Decision Tree Model did a little better when compared to Random Forest. The CV Run time was less for Decision Tree and the RMSE value was also pretty low but higher when compared to Random Forest.
When looking at the results for Feature Selection, the RMSE was a little higher for Decision Tree, but the difference was not that huge and it selected less variables when compared to Random Forest as well. The Run time was also less for Decision Tree when the Feature Selection method was implemented on the model.
The standard deviation values for the Decision Tree model also did better when the Feature Selection was implemented, in the range of 0.02-0.04. Whereas the standard deviation values for the Random Forest model were in the range of 0.03-0.06.

For the Data Processing, the most complex machine learning models do better if we convert factors to integers when we run models, which is the reason we converted all the categorical variables to integers. Furthermore, created cross fold validation to be; 5 so that it is easy to compare the out of sample RMSE for the models.

For all these models listed above, **CV= 5** folds were used and not changed or varied for any of the models. Comparing the models, since the Accuracy is supposed to be high and the RMSE is supposed to be the lowest, Decision Tree and Random Forest almost performed the best as compared to the other models. The RMSE was very high for the other models when compared to the two models that were discussed above.

Conclusion:

In conclusion, some of the variables that perfomed the best were: Sale Price, Overall Quality, GrLivArea, GarageCars, GarageYrBlt, OpenPochSF, BsmtFullBath – which makes a lot of sense. Sales price being the highest, explains that this is something that customers looking before they are going to make such a big investment. Followed by that some of the main features that the customers are looking are some of the main structures that are important, like the overall quality of the house, the year it was built, the number of parks that can be parked on that property. Etc.

The housing market in the United States continues to draw much attention as always. This paper attempts to summarize the key changes in housing prices in the recent years in a particular area of the US. By summarizing the most important features that related to the housing prices, the reader is able to understand what the main structures are the person trying to sell or buy is looking at. Even though It was a little complicated to understand the data before the analysis and model building was done – for example some of the variables said NaN – which did not mean that the data was missing but it meant that the houses did not have those particular features in them. Another example would be, it was interesting to see how square footage, quality/condition, quality of materials, neighborhood, and the number of bathrooms are very important factors for sale prices.

Another interesting factor that can make the research and analysis more details is if there was more data coming in. The data used for this experiment was from 2007 to 2010, looking at the models that would have been created with data which is current would have been more interesting to see how the trends would have changed over the years or they remained the same.

Future Work:

For future work, there could be further study done on the data set, which would result in improved models and improved performances. For future work, the models could be hyper tuned to see if they show different results and compare them to what we have currently.
There can be additional models done as well, for example XG Boost, SGD, Ridge Regression etc.
The Data set is also complied from Ames Iowa – it would be also interesting to see how do the house price varies in other parts of the country. The models can be compared to each other to see if they have bring out the same trends and to see if there are more important features in other parts of the cities. Sales price would probably be the top one for most of the places, but it would be interesting to see if people want different features in their houses. Since Fire Places are really common in this data – this is because that part does get cold, which further explains the need or want of that feature. Places like Florida would probably not want a fireplace and would want different features that would complement their lifestyle or choice.

Citations:

- Teck-Hong, Tan. “Housing Satisfaction in Medium- and High-Cost Housing: The Case of Greater Kuala Lumpur, Malaysia.” Habitat International, Pergamon, 12 July 2011, www.sciencedirect.com/science/article/pii/S019739751100049X.
- Pagés, Martínez, et al. “Analysis of House Prices in Spain.” EconPapers, 24 Aug. 2013, econpapers.repec.org/paper/bdewpaper/0307.htm.
- Gao, B., Chen, J. and Zou, L.H. (2012) Regional Housing Price Differences, Labor Mobility and Industrial Upgrading. Economic Research, No. 1, 66-79.
- Liu, L. and Liu, H.Y. (2003) Economic Analysis of the Relationship between Land Price and House Price. Journal of Quantitative & Technical Economics, No. 7, 27-30.
- Rees, W. and Wackernagel, M. 2008. Urban ecological footprints: Why cities cannot be sustainable — and why they are a key to sustainability. Urban Ecology. 537-555.
- McKinney, M. 2002. Urbanization, biodiversity, and conservation. Bioscience. 52, 10, 883-890.
- McKinney, M. 2006. Urbanization as a major cause of biotic homogenization. Biological Conservation. 127, 3, 247-260.
- McKinney, M. 2008. Effects of urbanization on species richness: A review of plants and animals. Urban Ecosystems. 11 (2): 161-176.
- Alberti, M. 2010. Maintaining ecological integrity and sustaining ecosystem function in urban areas. Current Opinion in Environmental Sustainability.