House Sale Price Prediction using Feature Engineering Techniques and Ensemble Learning Algorithms

MSc Research Project
Data Analytics

Oyindamola Eniola Ogunbiyi
Student ID: X18149065

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name	Oyindamola Eniola Ogunbiyi
Student ID	X18149065
Programme	Data Analytics
Year	2020
Module	Msc Research Project
Supervisor	Jorge Basilio
Submission Due Date	23/04/2020
Project Title	House Sale Price Prediction using Feature Engineering Techniques and Ensemble Learning Algorithms
Word Count	6906
Page Count	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL Internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ………

Date: ………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

House Sale Price Prediction using Feature Engineering Techniques and Ensemble Learning Algorithms

Oyindamola Eniola Ogunbiyi
X18149065

Abstract

Housing is one of the fundamental essential of every living thing hence the reason for continuous research in this sector. This project simply examines a dataset, which consists 1460 observations and 80 features that contribute to the sale price of the houses. Dataset was cleaned and transformed and some explorations were done on it to answer some basic questions that anybody would like to ask about housing. Feature engineering was performed on the transformed data using Principal Component Analysis (PCA) and dummy encoding and this is to ensure our dataset is ready in the right form with the right variables to be used in the algorithms, which results in improved model accuracy.

Different ensemble algorithms were used on the dataset in this project. The overall result of this project shows that the most important variables that determine the price of a house being sold.

Keywords: Housing price, Principal Component Analysis, Dummy encoding, Ensemble Algorithms and Feature Engineering.

1 Introduction

The real estate market is one of the ever-developing markets in the world today. This can be attributed to the fact that housing is an essential need for man. There are several stakeholders interested in the prediction of housing prices in a particular area ranging from landowners, to realtors, potential buyers, as well as government authorities as this, plays a major role in the economy. The sale or purchase of a house is considered one of the crux decisions of life due to a large amount of money involved and the commitment of relocation. The pricing of houses is affected by various factors, from the location of the house, the features in the house, including the demand and supply of houses in that area (Phan, 2019).

Studies have also shown that properties also appreciate over time resulting from economic growth and development except in cases of wars, emigration and natural disasters amongst others. Therefore, there is always appreciation in the value of house prices. This has made a real estate investment a lucrative venture. Thus, the prediction of housing sale price can also be considered as an important economic index. The value of a house that increases with time requires the appraisal value to be calculated as this value is required during the sale, purchase or even mortgage of a house (Shinde and Gawande, 2018).

There could be bias in determining the appraisal values by the stakeholders especially the professional appraisals. The bias is a shortcoming from professional appraisals which would, in turn, affect actual value on the housing (Shinde and Gawande, 2018). To curb this, determining the actual value of the house sale price would require a system void of the bias. This is hence the need for introducing machine learning in predicting housing sale price. The system, in turn, would guide stakeholders and inexperienced stakeholders can, in turn, be guided by this system in order not to make losses. Machine learning trains a computer to do what comes normally to animals or human beings while learning from experience. A model from machine learning makes use of computational models to "learn" from data without depending on a predetermined equation as a model (MathWorks, 2016). Various trends can influence the prices of hosing. These trends, as well as other parameters such as the construction materials used, number of bedrooms, living area (Bagheri, 2015), location, upcoming projects, proximity (Bourassa, Cantoni and Hoesli, 2011) amongst others, are turned into raw data. The raw data are used as inputs for the model and in turn, produce a just price without any form of bias.

For the prediction, feature-engineering techniques will be used to extract features from raw data and the features extracted will be used to improving the performance of the selected machine-learning algorithm. The algorithm to be used here are ensemble algorithm and they include Random Forest, XGBoost (eXtreme Gradient Boosting), Stochastic Gradient Boosting, Stacking NNET, GLM, KNN, SVM Radial and Bagged CART. K Fold Cross validation was used for evaluation and to select the best algorithm for this research work.

XGBoost is an advanced implementation of the gradient boosting (GBM) algorithm. It has a higher speed than existing gradient boosting implementations due to its features of implementation. Finally, the Random Forest, which is considered as the most versatile machine learning methods, will also be used. After the prediction of the results of the algorithms will be evaluated.

2 Literature Review

Housing is one out of the 3 basic needs for human survival hence research about housing and all that relates to it can never be over emphasized. In this section of this research work I will examine different papers or research work that have been published previously as regards housing and prices in the past and the research gap noticed which forms the basis for my own research work.

2.1 Real Estate Price Prediction using Machine Learning

This author did a research work on how previous researchers have concluded that artificial neural networks have more influence in predicting house prices but have neglected other machine learning algorithms hence his research was based on them Aswin (2017). The dataset used in this research was downloaded from UCI data repository and contains 2000 rows of 10 attributes.

He used 6 different machine-learning algorithms, which include Random Forest, Neural Networks, Gradient Boosted, Bagging, Support Vector Machine and Multiple Regression. At the end of the research he was able to evaluate all the algorithms using the R-squared and RMSE (Root Mean Square Error) to select the best algorithm. R studio was used to perform Random Forest and Multiple Regression while Rapid Miner was used to perform SVM and Gradient Boosted Trees and finally WEKA to implement Neural Networks and Bagging.

R studio is an open source statistical programming language while WEKA (Waikato Environment for Knowledge Analysis) is an open source machine learning software that has a graphical user interface that can be used for data mining and Rapid Miner is a data science software platform that with an environment to process data and perform machine learning.

He concluded based on the results of the algorithm that Random Forest was the best as it returned the highest accuracy with R-squared value of 90% and RMSE value of 0.012.

Hujia Yu and Jiafu Wu (2016), also worked on real estate price prediction and the goal of their research is to create a regression model and a classification model that are able to accurately estimate the price of the house given the features.

They concluded that for classification models the best algorithm is SVC (Support Vector Classifier) with linear kernel, with an accuracy of 0.6740 and after PCA (Principal Component Analysis) was performed on the dataset it increased to 0.6913 while for regression problem, the best model is SVR (Support Vector Regression) with gaussian kernel, with RMSE of 0.5271.

2.2 Modelling House Price Prediction using Regression Analysis and Particle Swarm Optimization.

There were 4 researchers that worked on this paper and the aim was to predict house prices using regression and particle swarm optimization (PSO) Adyan et al (2017). PSO was used to select the most important variables and regression was used to determine the optimal coefficient in the prediction. Their research was focused in Malang City, because Malang is one of tourism and urban city in East Java. They listed the 3 different factors that affect house prices physical condition, location and concept according to Rahadi, et al (2015). Physical condition is described as attributes of the house you can see physically, location is in which geographical area the house in located in and concept is the idea of the seller being able to convince the buyer of the house what type of house it is for example, you can say it is a minimalist home, a smart home, etc.

The dataset used in this research was from 9 houses in the Malang area between 2014-2017 and this is one of the noticeable challenges they faced during the research i.e limited access to data. MAE and RMSE were used to select the best method for the models in this research. Regression analysis was performed and also 7 different models were built while using the particle test and it represented different areas. The results were tabulated with their MAE and RMSE values and the best figures were picked at 1800 particles, 700 iterations, an inertia weight of 0.4 and 0.8 with RMSE value of 14.186 which is the least error predicted out of all the models. They concluded that in future different methods should be used to perform this prediction since the errors in all the predicted models are very large and also a large data set should be used.

2.3 House Price Prediction Using Machine Learning

G. Naga et al (2019), built a prediction model-using machine learning algorithms such as linear regression, multiple regression, lasso (Least Absolute Shrinkage and Selector Operator) regression, gradient boosting algorithm in python. The dataset used for this research wasn't described in the report writing but a screenshot of the data been loaded into R was screenshot and put in the report. Some exploration was done between the variables such as price and living area, price and latitudes, price and area, waterfront and price, condition of the house and price. All this was done so as to better understand basic relationship and answer some questions about both the independent and dependent variables. They noted in the conclusion that future work be done on stacking to reduce time for the different algorithms running. Also the gradient boosting algorithm returned more accuracy than the lasso regression with 91.2% and 76.1% respectively.

2.4 Conclusion

In conclusion after reviewing the above papers the first thing I noticed with the dataset is that a lot of research has been done in the past on US housing sector hence I decided to look out for data for other countries and I got one for Australian houses. Also different algorithms have been used to predict housing prices but the most efficient one is regression of which my research work will perform on this dataset using different regression algorithms and present the result to know which gives the best accuracy using the R-squared and RMSE value.

3 Methodology

This chapter describes the methodology approach used in predicting the house sale price. Cross industry standard process for data mining (CRISP -DM) was used as it is in line with the objective of the project. This approach consists of six phases and they will be explained below as it relates with this project.

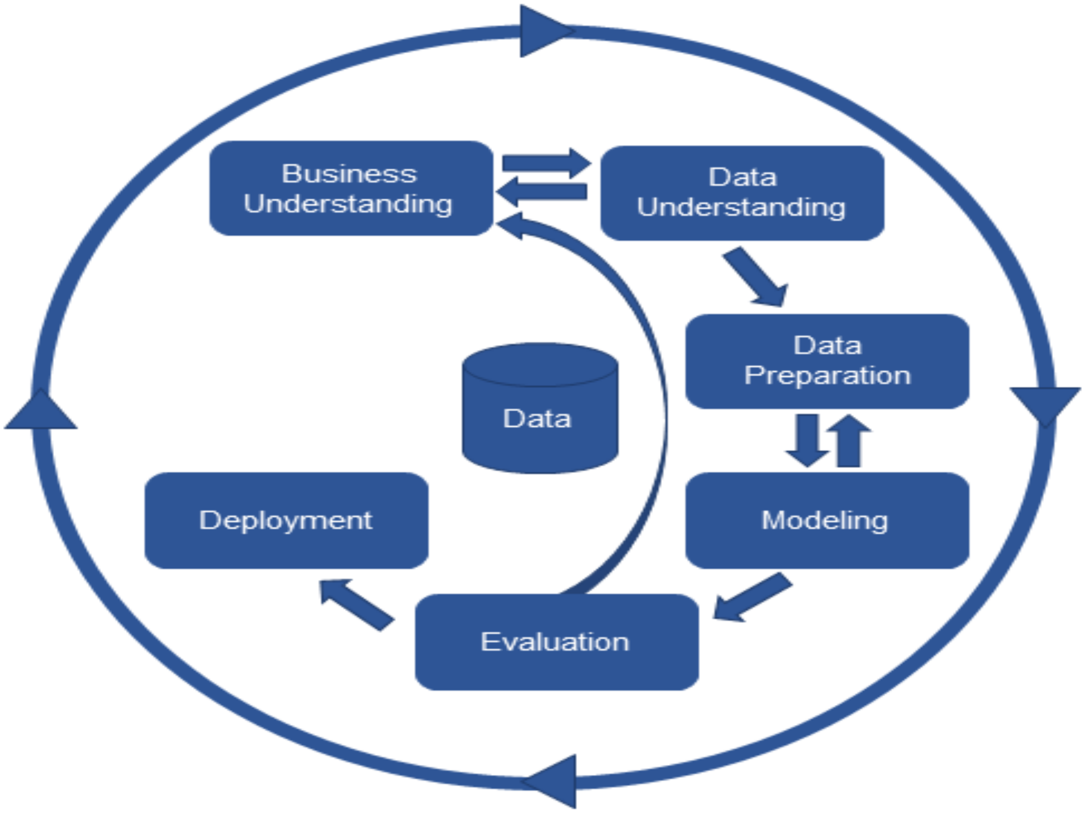


Figure 1: CRISP-DM

3.1 Business Understanding: Surprise Housing is a US based housing company that just made a decision to enter the Australian market. The company uses data analytics to buy houses at a low price mostly lower than their actual value and in turn sell it at a higher price.

The company collected data from sale of houses in Australia and made a dataset of about 1460 observations and 80 variables. They intend to perform analytics on this data to be able to select prospective houses to buy to enter the market and how each variable affect the price of housing in Australia.

A clear understanding of the business objective is a major step in creating a good model that will benefit the business. In this project, we would like to predict a good level of accuracy of the prices of houses to be able to marginally make profits. To determine the price of a house various factors can influence it. Three factors affect the price of a house, Rahadi *et al.*, (2013) classifies them into three; physical condition, location, and concept. To solve this business problem, there would be a need to create a model that best combines these features to make good predictions and thus enable the company to make better profit.

3.2 Data Understanding: To better understand the business goal, we are breaking down all the variables of this dataset to see how each of them contributes to or affect the price of the housing price in Australia.

The dataset named **train2.csv** contains 1460 observations and 80 variables of housing details in Australia, which is enough for the purpose of this project.

The dataset was loaded into R studio, data cleaning and pre-processing was done to ensure consistency before exploratory data analysis was performed to obtain some insights on the variables in the dataset.

The dataset is made up of 80 variables and 1460 observations. The variables to be used for prediction in this dataset are tabulated below:

No	Feature	Description

1.	MSSubClass	Identify the type of dwelling involved in the sale.
2.	MSZoning	Identifies the general zoning classification of the sale
3.	LotFrontage	Linear feet of street connected to property
4.	LotArea	Lot size in square feet
5.	Street	Type of road access to property
6.	Alley	Type of alley access to property
7.	LotShape	General shape of property
8.	LandContour	Flatness of the property
9.	Utilities	Type of utilities available
10.	LotConfig	Lot configuration
11.	LandSlope	Slope of property
12.	Neighborhood	Physical locations within Ames city limits
13.	Condition1	Proximity to various conditions
14.	Condition2	Proximity to various conditions (if more than one is present)
15.	BldgType	Type of dwelling
16.	HouseStyle	Style of dwelling
17.	OverallQual	Rates the overall material and finish of the house
18.	OverallCond	Rates the overall condition of the house
19.	YearBuilt	Original construction date
20.	YearRemodAdd	Remodel date (same as construction date if no remodeling or additions)
21.	RoofStyle	Type of roof
22.	RoofMatl	Roof material
23.	Exterior1st	Exterior covering on house
24.	Exterior2nd	Exterior covering on house (if more than one material)
25.	MasVnrType	Masonry veneer type
26.	MasVnrArea	Masonry veneer area in square feet
27.	ExterQual	Evaluates the quality of the material on the exterior
28.	ExterCond	Evaluates the present condition of the material on the exterior
29.	Foundation	Type of foundation
30.	BsmtQual	Evaluates the height of the basement
31.	BsmtCond	Evaluates the general condition of the basement
32.	BsmtExposure	Refers to walkout or garden level walls
33.	BsmtFinType1	Rating of basement finished area
34.	BsmtFinSF1	Type 1 finished square feet
35.	BsmtFinType2	Rating of basement finished area (if multiple types)
36.	BsmtFinSF2	Type 2 finished square feet
37.	BsmtUnfSF	Unfinished square feet of basement area
38.	TotalBsmtSF	Total square feet of basement area
39.	Heating	Type of heating
40.	HeatingQC	Heating quality and condition
41.	CentralAir	Central air conditioning
42.	Electrical	Electrical system
43.	1stFlrSF	First Floor square feet
44.	2ndFlrSF	Second floor square feet
45.	LowQualFinSF	Low quality finished square feet (all floors)
46.	GrLivArea	Above grade (ground) living area square feet
47.	BsmtFullBath	Basement full bathrooms
48.	BsmtHalfBath	Basement half bathrooms
49.	FullBath	Full bathrooms above grade
50.	HalfBath	Half baths above grade
51.	BedroomAbvGr	Bedrooms above grade (does NOT include basement bedrooms)
52.	KitchenAbvGr	Kitchens above grade
53.	KitchenQual	Kitchen quality

54.	TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)
55.	Functional	Home functionality (Assume typical unless deductions are warranted)
56.	Fireplaces	Number of fireplaces
57.	FireplaceQu	Fireplace quality
58.	GarageType	Garage location
59.	GarageYrBlt	Year garage was built
60.	GarageFinish	Interior finish of the garage
61.	GarageCars	Size of garage in car capacity
62.	GarageArea	Size of garage in square feet
63.	GarageQual	Garage quality
64.	GarageCond	Garage condition
65.	PavedDrive	Paved driveway
66.	WoodDeckSF	Wood deck area in square feet
67.	OpenPorchSF	Open porch area in square feet
68.	EnclosedPorch	Enclosed porch area in square feet
69.	3SsnPorch	Three season porch area in square feet
70.	ScreenPorch	Screen porch area in square feet
71.	PoolArea	Pool area in square feet
72.	PoolQC	Pool quality
73.	Fence	Fence quality
74.	MiscFeature	Miscellaneous feature not covered in other categories
75.	MiscVal	$Value of miscellaneous feature
76.	MoSold	Month Sold (MM)
77.	YrSold	Year Sold (YYYY)
78.	SaleType	Type of sale
79.	SaleCondition	Condition of sale
80.	SalePrice	Sale Price of houses

Table 3.1: Features and description of the dataset

3.3. Data Preparation: Data is never readily available in the state that a machine-learning algorithm can use it. Data pre-processing, data normalization will be used to normalize the data to a form that is suitable for the machine-learning algorithm. In some situations, there can be situations of mission data and this has to be handled gracefully by either removing the data completely or replacing the missing data with the mean value of that feature in the dataset. In this project we replaced the missing values with NA meaning that condition doesn't exist for a particular house. Most of the Variables were visualized to get an idea of what they look like and which contributes more to the house price sale.

3.4 Modeling: In this phase modeling techniques such as SGBoost, XGBoost, Random Forest and stacking of NNET, GLM, KNN, SVMRadial were used.

3.4.1 SGBoost (Stochastic Gradient Boosting)

Proposed by Friedman (2001), it uses the gradient descent approach to build models in the negative sense of the partial derivative of the loss function concerning the set of predictions. It seeks to find an addictive model that minimizes the loss function (Persson *et al.*, 2017). This algorithm starts with the model being adjusted (fit) to the data, predictions are made and residuals are obtained. Next a new model is adjusted to the previous residuals, based on this new model new predictions are made which are in turn added to the initial prediction and then new residuals are obtained. This process continues until a final convergence criterion is reached; the final prediction is obtained by averaging the model outputs. In each iteration of this technique a new model is created by adjusting the data such that the new model compensates for the weakness of the model created in the previous iteration.

3.4.2 XGBoost (Extreme Gradient Boosting)

XGBoost is an extension of the Stochastic Gradient Boosting Algorithm with an objective to optimize available computational resource and reduce model over fitting. To achieve this a regularization term and weight are added to the loss function to control the complexity of the model (Chen and Guestrin, 2016).

3.4.3 Random Forest

Random forest's main objective is to improve the performance of regression trees by reducing their variance (Assouline, Mohajeri and Scartezzini, 2018)

The Random Forest algorithm performs as follows:

a. Given a training set of N samples for a variable response and M predictors, B samples are generated using bootstrap technique
b. B regression trees are built per sample
c. With the best subset obtained in (a), an estimate of the response variable is generated for each regression tree
d. A final prediction is generated via aggregation, considering estimates in step (c)

3.4.4 Stacking

Stacking is another ensemble approach that combines different prediction models into a single model working at different levels. This approach aims to minimize the generalization error by introducing meta-learning and representing an asymptotically optimal learning system (Wolpert, 1992; Van Der Laan, Polley and Hubbard, 2007).

The main improvement of using stacking arises when there is diversity among the models of the different levels this is because different principles of generalization tend to produce different results (Mendes-Moreira *et al.*, 2012). To introduce diversity in the modelling process, the use of models that follow different learning strategies, employing different data characteristics can be used.

An example of a stacked architecture is shown in figure 3.2 below. Using the stacking approach with a 2-level layer (l0 and l1), in l0 various models are trained and the predictions from those models are fed to the model in l1. The l1 model (meta-model) too is trained and its prediction will be the desired end result. From this approach, it is observed that each subsequent layer-model learns from the model of the previous layer where the models at a subsequent level provide the best estimate (Shamaei and Kaedi, 2016; Petropoulos *et al.*, 2017). This work will explore the use of a Stack of Regression Trees, Generalized Linear Model, K-Nearest Neighbour and Linear Discriminate Analysis.

3.5 Evaluation: R-squared is a way we statistically measure how well data fits the model in which it is been used. It also explains the proportion of the variance of the dependent variable that is explained by the independent variable. The R squared value is always between 0% and 100%. This means that the exact R squared value we get after running our code explains the particular variance of that dataset. If the R squared value is 75% it means that the data is more fitted into a straight line around the model but if it is as low as 25% we can expect that it will be dispersed away from the model. In this project we are going to select the highest R squared values for the different algorithms.

Root Mean Square Error (RMSE) is the standard deviation of the prediction errors in the dataset being used in this research. Using different algorithms for the prediction we will look through the various RMSE values and report them accordingly.

Mean Absolute Error (MAE) is the average value of the data absolute difference between the predicted observation and the actual observation in the dataset.

K Fold Cross Validation is a technique we use in predictive models whereby the data is test and train set of data. The train set is used for training the model while the test set is used to evaluate the model.

3.6 Deployment: This is the final stage of the report, which consists of a written report and the code executed in R.

4 Implementation

The figure below illustrates the process flow that guides the implementation performed in this research. R programming language was used to perform all the experiments.

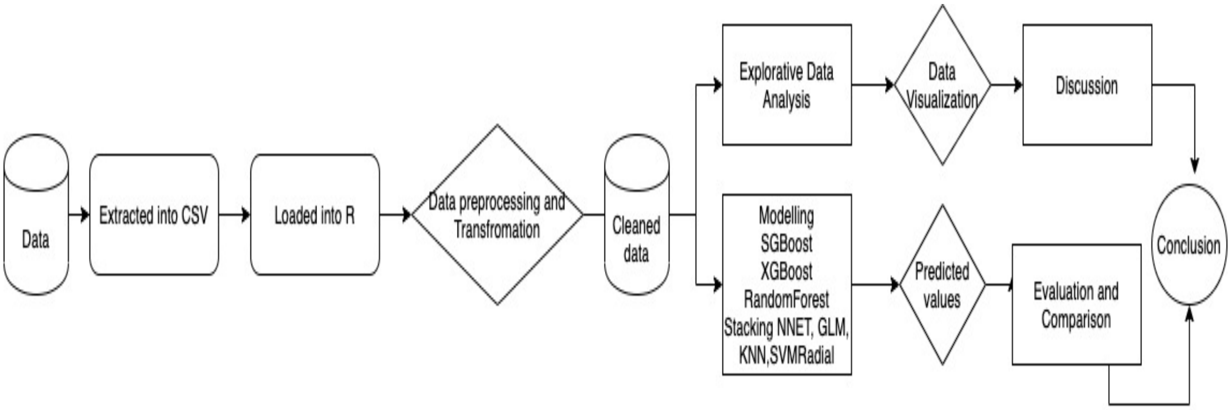

Data
Extracted into CSV
Loaded into R
Data preprocessing and Transfromation
Cleaned data
Explorative Data Analysis
Data Visualization
Discussion
Modelling SGBoost XGBoost RandomForest Stacking NNET, GLM, KNN,SVMRadial
Predicted values
Evaluation and Comparison
Conclusion

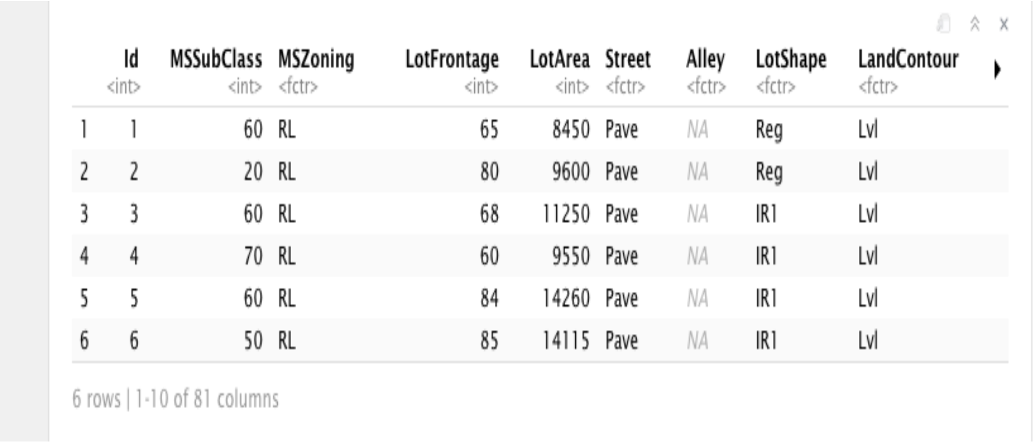
Figure 2: Flow Diagram of Sale Price Prediction

4.1 Data Collection

The data was downloaded from Kaggle. (Link: https://www.kaggle.com/srikanthladda/house-price-prediction). It is a CSV file with size 461kb after extraction from the ZIP file downloaded.

4.2 Data Pre-processing and Transformation

I. The extracted CSV file was loaded into R studio
II. Structure of the data was investigated as against the meta data provided. It was discovered that some variables were numeric instead of factors so that was converted to factors.
III. The ID column in the dataset was dropped as it isn't necessary for the prediction
IV. Missing values in the numerical variables were replaced by the mean of the column while for factor variables it was replaced with 'No_' and this is because the explanation in the metadata says where there is a missing value it is because that feature doesn't exist for that house.
V. Specificity was done whereby all the levels in the factor columns were correctly represented as it is in the meta data
VI. Standard deviation of the dependent variable was calculated and it showed that the independent variables aren't to far away from the mean of the sale price. It shows that the values in the dataset are normally distributed.
VII. There was no noticeable outlier to be worried about as this was checked using cooks distance
VIII. Three new columns were added to calculate the number of years and these are year built, year removed (same as year renovated) and year garage was built.
IX. Multicollinearity was used to check for the correlation between the independent variables using a cut off of 0.69. Total basement and Ground living area was found to be highly correlated hence they were both removed.
X. Pearson correlation was also checked for and here we removed low correlation and too high correlation with the dependent variable using a minimum of 0.3 and 0.69 maximum. The columns dropped here includes type 2 finished square feet, unfinished square feet of basement area, enclosed porch area in square feet, three season porch area in square feet, dollar value of miscellaneous feature and year Sold
XI. SMOTE was used to split the date into 80% train and 20% test data. Cross validation technique is used to check for the best tuning in terms of RMSE, RSQ and MAE.



	Id <int>	MSSubClass <int>	MSZoning <fctr>	LotFrontage <int>	LotArea <int>	Street <fctr>	Alley <fctr>	LotShape <fctr>	LandContour <fctr>
1	1	60	RL	65	8450	Pave	NA	Reg	Lvl
2	2	20	RL	80	9600	Pave	NA	Reg	Lvl
3	3	60	RL	68	11250	Pave	NA	IR1	Lvl
4	4	70	RL	60	9550	Pave	NA	IR1	Lvl
5	5	60	RL	84	14260	Pave	NA	IR1	Lvl
6	6	50	RL	85	14115	Pave	NA	IR1	Lvl

6 rows | 1-10 of 81 columns

Figure 3: Dataset loaded into R

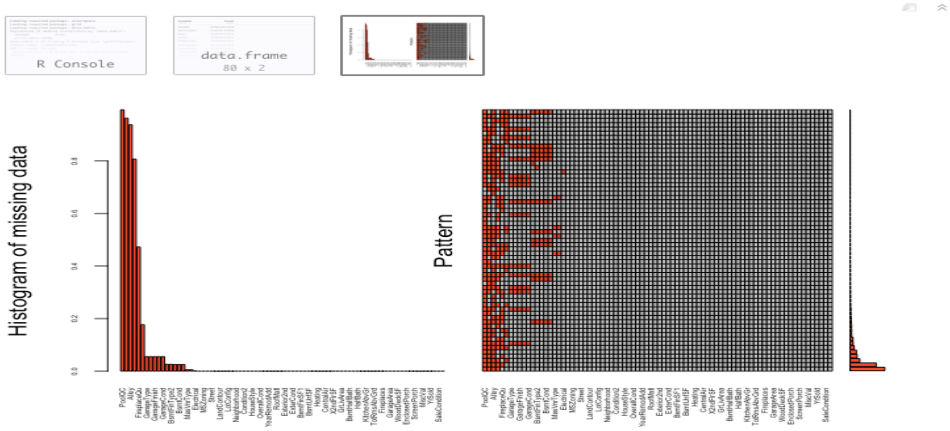


Figure 4: Histogram of missing values

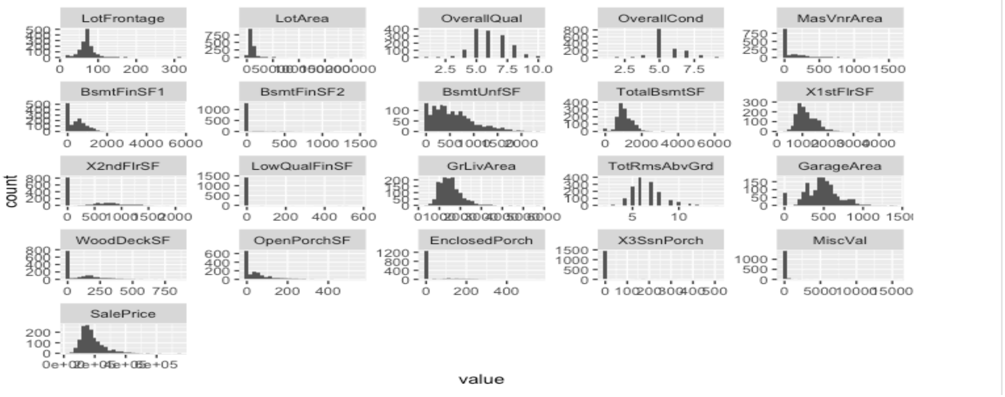


Figure 5: Visualization of the numerical columns

4.3 Exploratory Data Analysis

This section shows the exploration done on the dataset, which is what motivated the use of the algorithm. The following are the questions explored in this project and for the sake of writing I will only show some of the visuals here while I will provide the codes that shows the full visualization of all the questions explored.

I. Is there a significant relationship between sale price and building's age? Anova was used to check for this and we can see that there is a relationship between how much old the building is and how much it was sold for.

II. What is the average sale price based on overall condition of the house, year it was built, condition1 - proximity to social amenities and sale condition. For overall condition we have levels 1-9 with 1 been the lowest and 9 been the highest and the average of each level is shown. For the others the result is in the code provided.

III. What is the sale price distribution based on the overall quality of the house?

IV. What categories of house (based on age built) have the highest sale price? Here it was obvious that houses built below 50 years have sales price higher than $700,000

V. Sale Price versus Month it was sold. From here we saw that house price increases more during winter than autumn, spring and summer.

VI. What sale type has the highest sale price? There are 9 different types of sales type that was considered against the sale price.

VII. Price distribution and season. The bar charts shows there is higher percentage of people buy houses across all seasons at less than 200k.

VIII. At what price will people buy more even with garage attached? Density of garage type has a high peak at claim size about 160k$. It tells us that people are liable to buy houses at that point regardless of the sale price as long as a garage is attached to the house

IX. Sales per seasons. The probability of people buying houses is higher in summer and spring is more than autumn and winter.

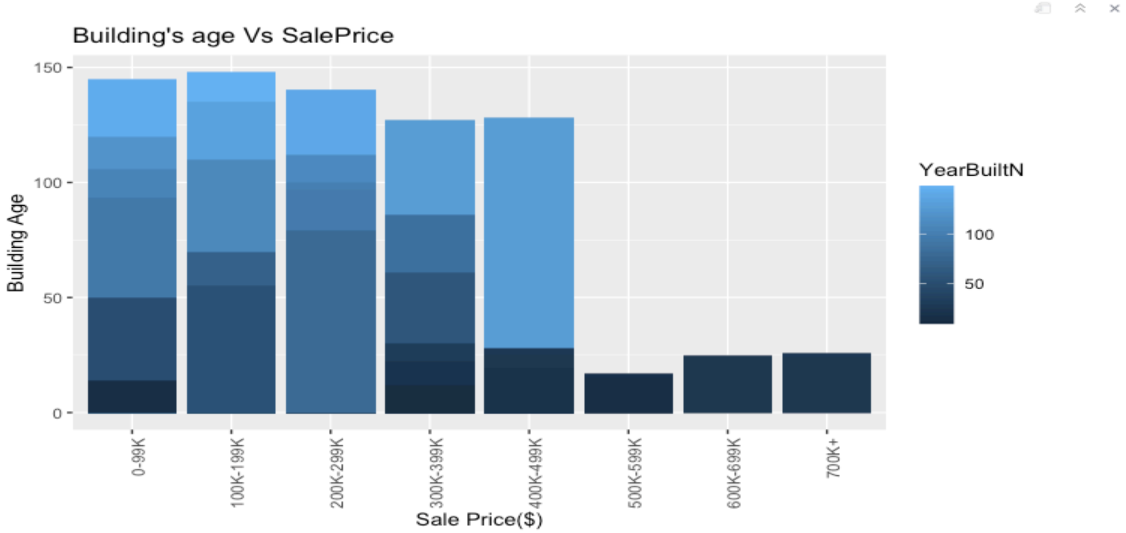


Figure 6: Building’s age VS Sale price

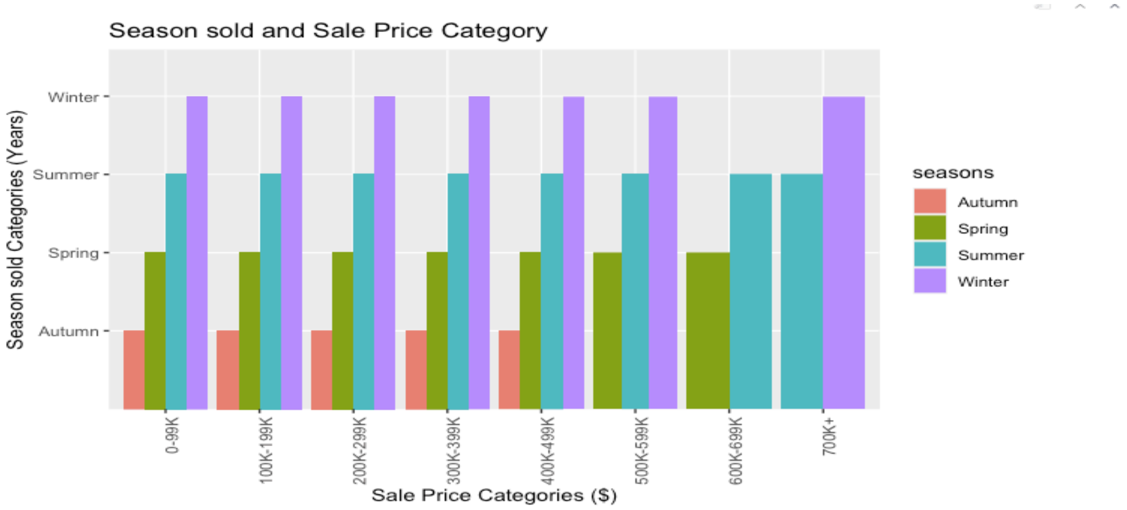


Figure 7: Season sold VS Sale price

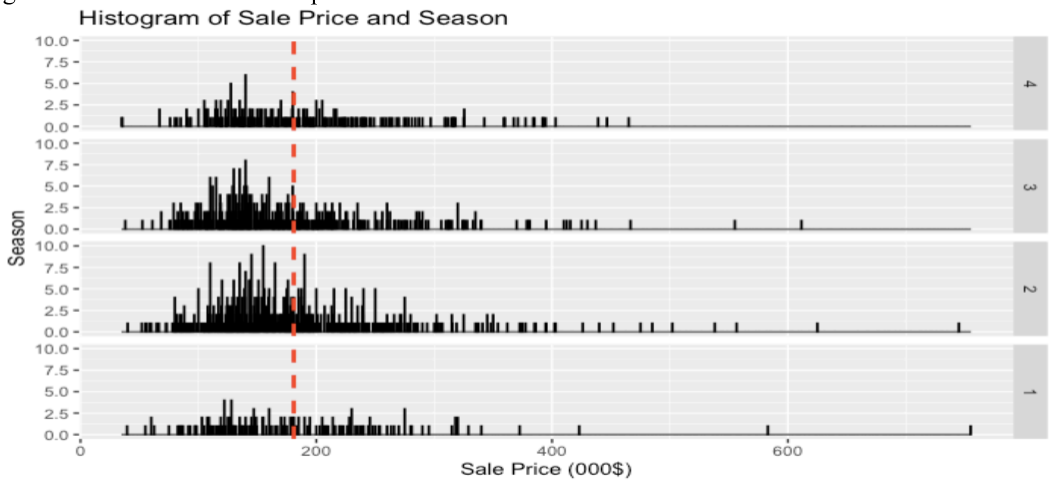


Figure 8: Sale price VS Season

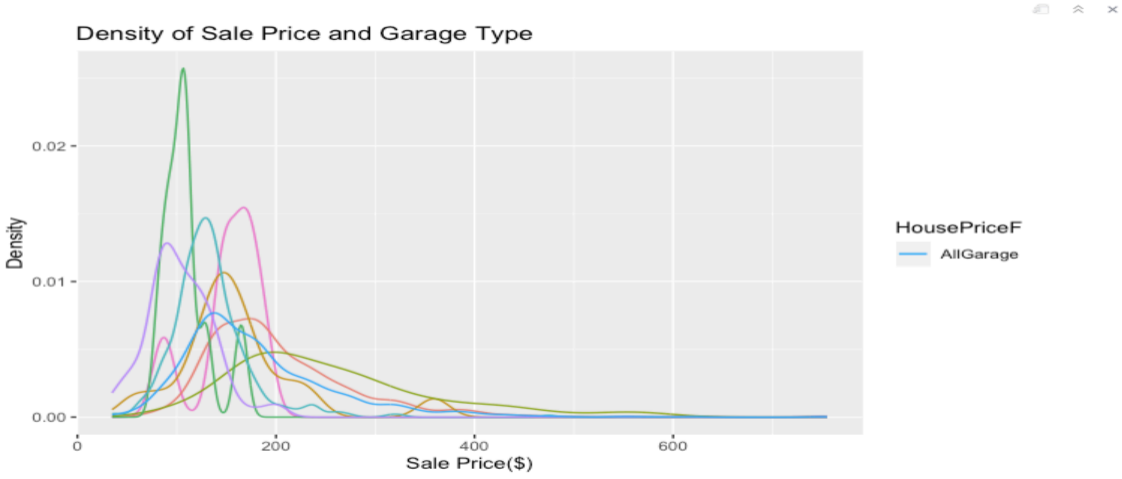


Figure 9: Density of Sale type and Garage type

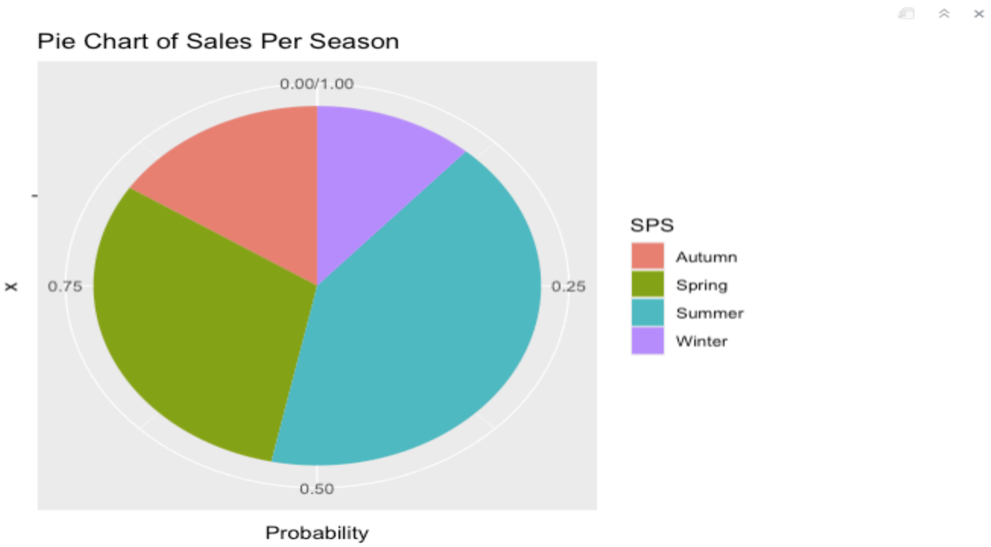
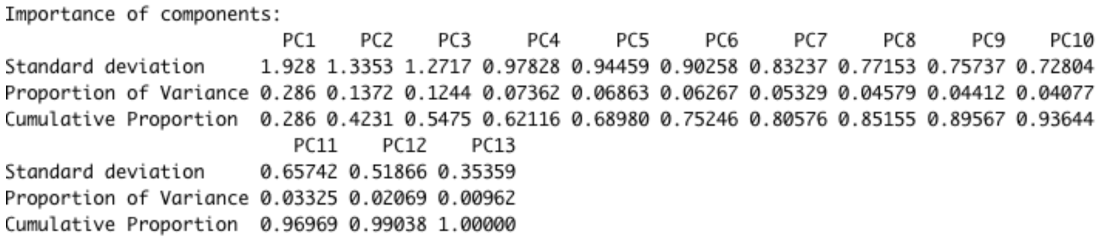


Figure 10: Pie chart of Sales per season

4.4 Feature Engineering

I. Dummy Encoding was used for factors, so factors with multiclass variables were dummified
II. Principal Component Analysis (PCA) was used numeric variables.
III. PC1 accounted for about 30% of the data and the first 9 PC's accounted for about 90%.
IV. Every PC that has Eigenvalue >0.7 was considered.
V. Then both engineered data was joined together.

Below is a screenshot of the PCA analysis.



```
Importance of components:
                         PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     1.928 1.3353 1.2717 0.97828 0.94459 0.90258 0.83237 0.77153 0.75737 0.72804
Proportion of Variance 0.286 0.1372 0.1244 0.07362 0.06863 0.06267 0.05329 0.04579 0.04412 0.04077
Cumulative Proportion  0.286 0.4231 0.5475 0.62116 0.68980 0.75246 0.80576 0.85155 0.89567 0.93644
                          PC11    PC12    PC13
Standard deviation     0.65742 0.51866 0.35359
Proportion of Variance 0.03325 0.02069 0.00962
Cumulative Proportion  0.96969 0.99038 1.00000
```

Figure 11: PCA Analysis, Scree Plot and Clustering

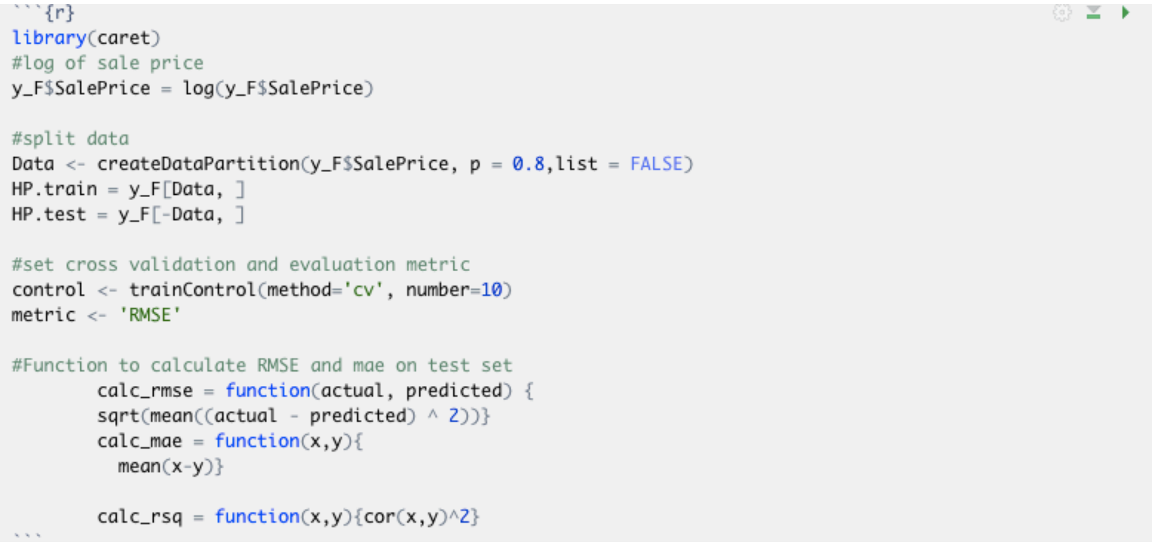
4.5 Modelling

To avoid over fitting, the caret library was used to split the dataset into train and test sets using a ratio of 80:20 before training the models on the train set and testing using the test set. RSQ, RMSE and MAE been calculated on the test set of data which is 20% of the data in this case. Below is a picture showing the implementation on R.

Random Forest is the first machine-learning algorithm that was implemented in R in this research, rf was used here and the method ranger to implement it on the two experiments, which are the original dataset and the most important variables.

Extreme Gradient Boosting popularly known as XGBoost is a machine learning algorithm that uses gradient boosted decision trees for the purpose of speed and the optimization of model performance. XGBoost package was used in R for the two experiments performed that are on the main dataset and the most important variables. xgbTree was used in the implementation of the XGBoost model after the data has been balanced using caret.

Stochastic Gradient Boosting also known as SGBoost is another machine learning algorithm that was used in R for this research work.
Stacking NNET, GLM, KNN,SVMRadial and Bagged CART algorithm were also implemented in R.



~~~
```{r}
library(caret)
#log of sale price
y_F$SalePrice = log(y_F$SalePrice)

#split data
Data <- createDataPartition(y_F$SalePrice, p = 0.8,list = FALSE)
HP.train = y_F[Data, ]
HP.test = y_F[-Data, ]

#set cross validation and evaluation metric
control <- trainControl(method='cv', number=10)
metric <- 'RMSE'

#Function to calculate RMSE and mae on test set
        calc_rmse = function(actual, predicted) {
        sqrt(mean((actual - predicted) ^ 2))}
        calc_mae = function(x,y){
          mean(x-y)}

        calc_rsq = function(x,y){cor(x,y)^2}
```
~~~

Figure 12: Test and Train of the dataset

5 Evaluation

K-fold cross validation with a K value of 10 was applied to the dataset to estimate the accuracy of the models on various holdouts of the test. These steps were performed to ensure that the obtained results were accurate without any form of over fitting and the caret library was used to compute the RSQ, RMSE and MAE of the models across the two experiments.

5.1 Model Performance on the Dataset

Model	RMSE	MAE	R-squared
Random Forest	0.141	0.006	87%
XGBoost	0.14	0.09	88%
SGBoost	0.145	0.012	87.3%
KNN	0.17	0.026	83%
SVM	0.14	0.0079	87.8%
NNET	0.26	0.05	66%
GLM	0.155	0.01	86%
BAGGED CART	0.19	0.01	78%

5.2 Model Performance on the Dataset with the 7 Most Important Features

Model	RMSE	MAE	R-squared
Random Forest	0.08	0.003	91%
XGBoost	0.08	0.006	93%
SGBoost	0.081	0.009	89%
KNN	0.10	0.013	87%
SVM	0.09	0.0031	90%
NNET	0.16	0.04	70%

GLM	0.11	0.009	90%
BAGGED CART	0.13	0.008	84%

5.3 Discussion

In this research, two experiments were performed, the first experiment was conducted using all the variables available in the dataset after pre-processing, while the second experiment was conducted using 7 most important variables and the goal of this is to be able to improve the model's performance using fewer variables. From the first experiment, xgboost had the highest R-squared and lowest RMSE value.

The second experiment was conducted to determine if utilizing fewer important variables can trigger an increase in model performance. The result showed that xgboost still performed better than other algorithms used in the research.

In fitting the model we need low RMSE and high R-square hence, we select XGBoost, which indicates that the model explains 93% of the variability of the dependent variable that is explained by the top most 7 independent variable while it has 0.08 RMSE, which signifies the least error amongst all the algorithms used.

In fitting the model we need low RMSE and high R-square hence, we select XGBoost, which indicates that the model explains 88% of the variability of the dependent variable that is explained by the independent variable while it has 0.14 RMSE, which signifies the least error amongst all the algorithms used.

Also another noticeable thing is that the accuracy increased from 88% to 93% after PCA was performed on the main dataset and the top 7 variables were selected.

6 Conclusion and Future Work

This project examined a housing dataset, prepared the data and loaded into R workspace. The data was then cleaned to meet the requirement of the prediction to be done. This data consists of 1 dependent variable and 79 other independent variables. Different visualizations was done to understand some basic questions one might have concerning sale price, season and other factors contributing to it as provided in the dataset.

The main goal of this project is to determine the most important features that determines the sale price of a house so when Surprise housing is about to purchase a house they know the features to consider first and seriously so as to make profits when reselling. This was done using different machine learning algorithms such as random forest, extreme and stochastic gradient boosting, stacking of neural networks, generalized linear models, K Nearest Neighbor and support vector machine and finally Bagged CART.

The evaluation metric used to select the best algorithm was R-squared, RMSE and MAE. The algorithm for with the highest R-squared and lowest error i.e. RMSE, on the two experiments conducted (which are on the real dataset and the top 7 dataset) was XGBoost.

After comparing the results of the model we conclude that XGBoost is the best algorithm in this project because it has the highest accuracy of 88% and 93% on the first and second model accordingly with the lowest error i.e RMSE of 0.14 and 0.08 accordingly.

In conclusion we can see that the most important 20 variables in descending order that determine the house price according to this dataset. The top 5 are quality of the material on the exterior, full bathrooms, Kitchen quality, foundation type and Size of garage in car capacity.

Hence, when the company surprise housing wants to buy a house for resale they should consider those features more importantly than the other features because that is what affects the decision of the buyers in that area.

Future work can be done on this dataset by investigating further the reason there is a cluster of sales price between $100k - $199k as we saw in the clustering diagram above. Also instead of using XGBoost the next person can use genetic algorithm to predict the importance of the independent variables in this dataset.

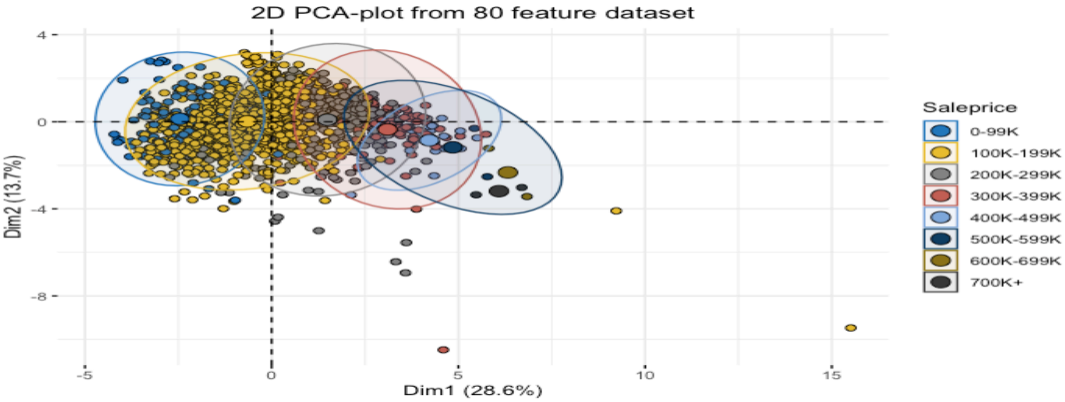


Figure 13: Clustering

7 Acknowledgement

I would like to thank my supervisor for his help and input in making this project a success even with the challenges faced during this pandemic he was available, accessible and very helpful via different online platforms. A special appreciation to my parents, family and friends for all the support they gave me from paying my fees to feeding and every other support that was rendered during my course of study in NCI. I give all the glory to God for the success of this degree.

Figures

References

Assouline, D., Mohajeri, N. and Scartezzini, J. L. (2018) ‘Large-scale rooftop solar photovoltaic technical potential estimation using Random Forests’, *Applied Energy*. doi: 10.1016/j.apenergy.2018.02.118.

Adyan N., Ruth E., Hilman T., and Wayan F., (2017), ‘Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization’ (IJACSA) journal DOI: Vol. 8, No. 10, 2017

Aswin S. R (2017), ‘Real Estate Price Prediction Using Machine Learning’

Bagheri, A. (2015) ‘Sample size impacts on high leverage collinearity-enhancing observations’, *Economic Computation and Economic Cybernetics Studies and Research*.

Bourassa, S. C., Cantoni, E. and Hoesli, M. (2011) ‘Predicting House Prices with Spatial Dependence: Impacts of Alternative Submarket Definitions’, *SSRN Electronic Journal*. DOI: 10.2139/ssrn.1090147.

Breiman, L. (2004) ‘Bagging predictors’, *Machine Learning*. doi: 10.1007/bf00058655.

Chen, T. and Guestrin, C. (2016) ‘XGBoost: A scalable tree boosting system’, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. doi: 10.1145/2939672.2939785.

Erdal, H. and Karahanoğlu, İ. (2016) ‘Bagging ensemble models for bank profitability: An emprical research on Turkish development and investment banks’, *Applied Soft Computing Journal*. doi: 10.1016/j.asoc.2016.09.010.

Fan, G. Z., Ong, S. E. and Koh, H. C. (2006) ‘Determinants of house price: A decision tree approach’, *Urban Studies*. doi: 10.1080/00420980600990928.

Friedman, J. H. (2001) ‘Greedy function approximation: A gradient boosting machine’, *Annals of Statistics*. doi: 10.2307/2699986.

Gang-Zhi, F., Seow Eng, O. and Hian, K. (2006) ‘Determinants of House Price: A Decision Tree Approach’, *Urban Studies*.

Gu, J., Zhu, M. and Jiang, L. (2011) ‘Housing price forecasting based on genetic algorithm and support vector machine’, *Expert Systems with Applications*. doi: 10.1016/j.eswa.2010.08.123.

G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu., (2019), 'House Price Prediction Using Machine Learning' DOI: (IJITEE) ISSN: 2278-3075, Volume-8 Issue-9, July 2019

Hujia Yu and Jiafu Wu., (2016), ' Real Estate Price Prediction with Regression and Classification'

Jonathon S. (2005) 'A Tutorial on Principal Component Analysis',

MathWorks (2016) ‘Introducing Machine Learning What is Machine 1’, *Perspectives on Ontology Learning*.

Mendes-Moreira, J. *et al.* (2012) ‘Ensemble approaches for regression: A survey’, *ACM Computing Surveys*. doi: 10.1145/2379776.2379786.

Neloy, A. A., Sadman Haque, H. M. and Ul Islam, M. M. (2019) ‘Ensemble learning based rental apartment price prediction model by categorical features factoring’, in *ACM International Conference Proceeding Series*. doi: 10.1145/3318299.3318377.

Nur, A. *et al.* (2017) ‘Modeling House Price Prediction using Regression Analysis and Particle Swarm

Optimization Case Study : Malang, East Java, Indonesia', *International Journal of Advanced Computer Science and Applications*. doi: 10.14569/ijacsa.2017.081042.

Park, B. and Kwon Bae, J. (2015) 'Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data', *Expert Systems with Applications*. doi: 10.1016/j.eswa.2014.11.040.

Persson, C. *et al.* (2017) 'Multi-site solar power forecasting using gradient boosted regression trees', *Solar Energy*. doi: 10.1016/j.solener.2017.04.066.

Petropoulos, A. *et al.* (2017) 'A stacked generalization system for automated FOREX portfolio trading', *Expert Systems with Applications*. doi: 10.1016/j.eswa.2017.08.011.

Phan, T. D. (2019) 'Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia', *Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018*. IEEE, pp. 8–13. DOI: 10.1109/iCMLDE.2018.00017.

Rahadi, R. A. *et al.* (2013) 'Attributes Influencing Housing Product Value and Price in Jakarta Metropolitan Region', *Procedia - Social and Behavioral Sciences*. doi: 10.1016/j.sbspro.2013.07.211.

R. A. Rahadi, S. K. Wiryono, D. P. Koesrindartotoor, and I. B. Syamwil, ―Factors influencing the price of housing in Indonesia,‖ Int. J. Hous. Mark. Anal., vol. 8, no. 2, pp. 169–188, 2015.

Ribeiro, M. H. D. M. and dos Santos Coelho, L. (2020) 'Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series', *Applied Soft Computing Journal*. doi: 10.1016/j.asoc.2019.105837.

Schapire, R. E. (1990) 'The Strength of Weak Learnability', *Machine Learning*. doi: 10.1023/A:1022648800760.

Shamaei, E. and Kaedi, M. (2016) 'Suspended sediment concentration estimation by stacking the genetic programming and neuro-fuzzy predictions', *Applied Soft Computing Journal*. doi: 10.1016/j.asoc.2016.03.009.

Shinde, N. and Gawande, K. (2018) 'Valuation of House Prices Using Predictive Techniques', *International Journal of Advances in Electronics and Computer Science*, 5(6), pp. 34–40. Available at: http://iraj.in.

Van Der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007) 'Super learner', *Statistical Applications in Genetics and Molecular Biology*. doi: 10.2202/1544-6115.1309.

Varma, A. *et al.* (2018) 'House Price Prediction Using Machine Learning and Neural Networks', *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*. IEEE, pp. 1936–1939. doi: 10.1109/ICICCT.2018.8473231.

Wolpert, D. H. (1992) 'Stacked generalization', *Neural Networks*. doi: 10.1016/S0893-6080(05)80023-1.
Yang, X. *et al.* (2019) 'Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery', *Chemical Reviews*. doi: 10.1021/acs.chemrev.8b00728.