

**Housing Project**

Submitted by:

Arpan Biswas

# INTRODUCTION

**Business Problem Framing:**

* Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy.
* A US-based housing company named Surprise Housinghas decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.
* Here, we need to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:
  1. Which variables are important to predict the price of houses?
  2. How do these variables describe the price of the house?

**Conceptual Background of the Domain Problem:**

* In this project, we have dataset of Australian houses and purpose of this dataset is to analyse the data to know more about Australian properties and decide which variables affect the price of houses. After that we will build a machine learning model to predict the price of houses.
* This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns

**Motivation for the Problem Undertaken:**

As we know, real estate is very large market and there are various companies working in the domain. We take this project because data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies.

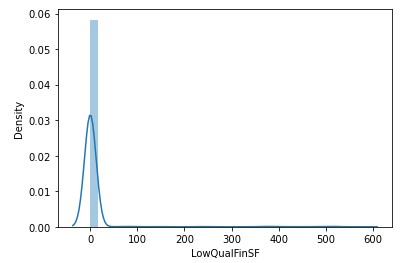
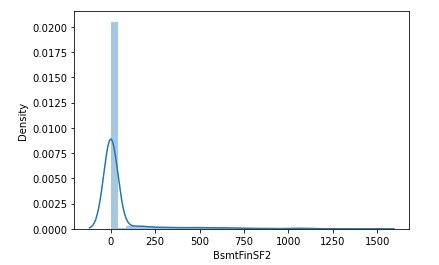
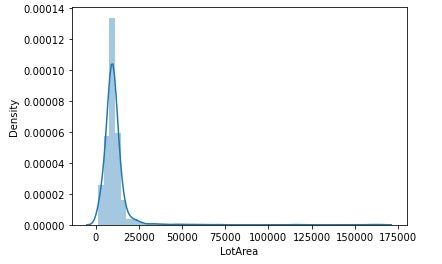
# Analytical Problem Framing

**Mathematical/ Analytical Modelling of the Problem:**

* We use some mathematical, statistical and analytics approaches in this project. Which is described below:

1) **Outliers:** Outliers are extreme values that fall a long way outside of the other observations. Here, we checked our data with two methods for removing outliers. Zscore and Inter Quartile Range.

* By using both methods, we conclude that percentage loss of data is very high. So, removing outliers is not a good idea because it impact on our result.
  + 1. **Log method:** In this dataset some attributes are skewed and skewness is affecting our machine learning model. So, it is necessary to remove skewness. For solving this issue we use log method. This method follows a normal or near normal distribution of the data. Below we can see that skewed data of some variables in form of graphical representation:



From the above distribution plot we can see that, all graph is right skewed, so we need to make this graph normal or near normal distributed. To solve this issue we use log method.

* + 1. **Standard scaler:** After removing skewness, we need to scale our data. For this we use standard scaler method. This method normalizes our data and essential for machine learning algorithms that calculate distance between data. For instance, most of the classifiers calculate the distance between two points by the distance. If one of the features has large value, then distance consider that particular feature. This method is necessary, where large and small values present in our data. This method transform our data with mean = 0 and standard deviation = 1.

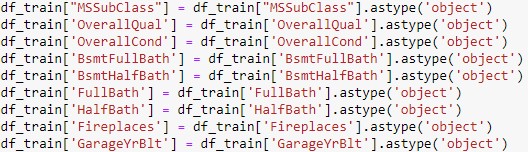
**Data Sources and their formats:**

* We obtain our data from the US-based housing company named Surprise Housing. The data contain different features of houses in Australia and we need to do analysis of that data and make machine learning model.
* This data set contains train data and test data.
* Train data have 1168 records and 81 variables and test data have 292 records and 80 variables. All the variables are related to properties in Australia.

**Data Pre-processing Done:**

For cleaning or pre-processing the data we use some techniques which are described below:

* We check null values present in the data and find many attributes have null values. So, we fill null values according to provided data description. For instance, Alley features have 1091 null values and according to data description NA means “no alley”, so we fill all null values in alley columns with “no alley”.
* Then dropped Id and MiscFeature columns. Drop id column because it is not useful for our model and drop MiscFeature column because it contains extra categories of home and have almost 96% null values.
* After that we found that some categorical variables data type is integer or float, so we convert them into object data type to treat them as a categorical data. Below is the snapshot of code:



* After analysis of the data, we remove skewness of the data using log method, and encoding categorical data into numerical data using one hot encoding method.
* At last, we scale data with mean = 0 and standard deviation = 1 using standard scaler method.

**Data Inputs- Logic- Output Relationships:**

* We used regression machine learning models because our target variable is saleprice. So, we need to find price of the houses.
* There are many regression models but here we used some of them models.
* First we split our training dataset into two segments: training and testing. We take 75% data for training and 25% data for testing. For splitting data we use train test split method. Below is the code for splitting the data:



* + 1. 75% of the observation as training set--> x\_train
    2. The associated target for each observation in x\_train --> y\_train 3) 25% of the observation as test set--> x\_test 4) The target associated with the test set--> y\_test.
* After splitting data we passed training data to machine learning models. The fitted model will first be used to generate prediction on the test set (x\_test). Next, the predicted class labels are compared to the actual observed class label (y\_test) to see the difference between them.

**Libraries Used:**

* We used many libraries used in this project, which is described below:
  1. Numpy: This library is used for scientific computing. It supports multidimensional arrays and matrices.
  2. Pandas: This library is used for data analysis and modelling convenient in python. Pandas simplify analysis by converting CSV, JSON, and TSV data files or a SQL database into a data frame with rows and columns.
  3. Matplotlib and Seaborn: Both libraries are used for data visualization.
  4. Scikit-learn: The Python library, [Scikit-Learn,](https://scikit-learn.org/) is built on top of the Matplotlib, Numpy, and SciPy libraries. It has wide range of algorithms.

**Models Development and Evaluation**

**Identification of possible problem-solving approaches (methods):**

* + 1. Data reading and understanding
    2. Data cleaning
    3. Data analysis
    4. Handling outliers
    5. Handling skewness
    6. Encoding data
    7. Scaling
    8. Train test split
    9. Machine learning algorithms

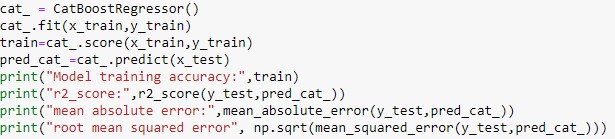
**Testing of Identified Approaches (Algorithms):**

The regression algorithm that we used is:

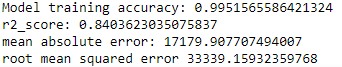
* 1. Linear regression
  2. K-Neighbors regressor
  3. Decision tree regressor
  4. Extra tree regressor
  5. Lasso regression
  6. Ridge regression
  7. Support vector regressor
  8. Random forest regressor
  9. Adaboost regressor
  10. Gradient boosting regressor
  11. Xgboost regressor
  12. Catboost regressor

**Building machine learning models:**

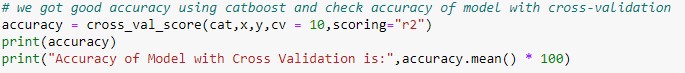
* We use many algorithms to find best model, but here we describe only best model.
* We find catboost regressor as a best model. Catboost builds upon the theory of decision trees and gradient boosting. The main idea of boosting is to sequentially combine many weak models and thus through greedy search create a strong competitive predictive model.
* Below is the code of our model with catboost regressor:



Output:



* We get 99% training model accuracy and 84% test data r2\_score accuracy; also we get good mean absolute error (17179.90) among all other algorithms. Now, we have to confirm that model is not going through underfitting or overfitting.
* So, we check catboost regressor training model accuracy using cross validation to confirm that our model is not going through underfitting or overfitting. Below is the code for cross validation:



* Here, we use CV = 10, that means our training set is divided into 10 parts and provide mean accuracy of those 10 parts. Output:



* We get 87% r2\_score accuracy using cross validation that means our model is not underfitted or overfitted.

* Now, check accuracy of all used algorithms:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algorithms** | **Model**  **training accuracy** | **R2 score** | **Mean**  **absolute error** | **Root mean squared error** |
| Linear regression | 0.95 | 0.05 | 33051.43 | 80929.32 |
| K-Neighbors regressor | 0.71 | 0.46 | 36024.52 | 61316.07 |
| Decision tree regressor | 1.0 | 0.59 | 29494.64 | 53267.45 |
| Extra tree regressor | 1.0 | 0.80 | 20568.65 | 36507.99 |
| Lasso regression(HP) | 0.95 | 0.65 | 25354.73 | 48813.56 |
| Ridge regression(HP) | 0.93 | 0.80 | 21722.44 | 37157.29 |
| Support vector regressor | -0.04 | -0.05 | 58389.21 | 85591.81 |
| Random forest  regressor(HP) | 0.97 | 0.77 | 20854.60 | 39300.30 |
| Adaboost regressor | 0.87 | 0.77 | 25080.52 | 39428.95 |
| Gradient boosting regressor(HP) | 0.97 | 0.83 | 19191.08 | 34041.79 |
| Xgboost regressor(HP) | 0.99 | 0.80 | 20111.17 | 36692.31 |
| Catboost regressor | 0.99 | 0.84 | 17179.90 | 33339.15 |

**Note**: In above table HP means with Hyperparameter Tuning.

**Key Metrics for success in solving problem under Consideration:**

* We use three types of metrics for solving problem. Which is described below:
  + 1. R2\_score: R2 score is the percentage of variation explained by the relationship between two variables. Range of the r2 score is varies from 0 to 1. Mathematical formula of the r2 score is as below:

**R2= 1- SSres / SStot**

Where,

SSres is the sum of squares of the residual errors.

SStot is the total sum of the errors.

* + 1. Mean absolute error (MAE): The Mean Absolute Error*,* also known as MAE, is one of the many metrics for summarizing andassessing the quality of a machine learning model.
* Mean absolute error subtract the predicted value from actual value as below:

Prediction Error **→** Actual Value - Predicted Value

* This prediction error is taking for each record after which we convert all error to positive. This is achieved by taking Absolute value for each error as below:

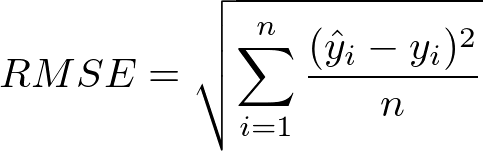
Absolute Error **→** |Prediction Error|

* Finally we calculate the mean for all recorded absolute errors (Average sum of all absolute errors). Below is the formula of MAE:



3) Root mean squared error (RMSE): Root Mean Square Error is the measure of how well a regression line fits the data points. RMSE can also be construed as Standard Deviation in the residuals.

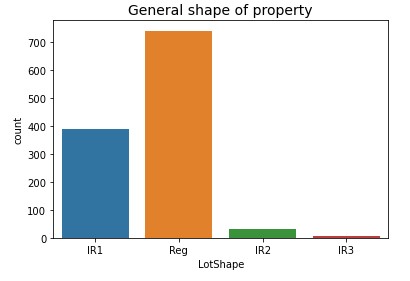
* It is a standard way to measure the error of a model in predicting quantitative data. Formally it is defined as follows:



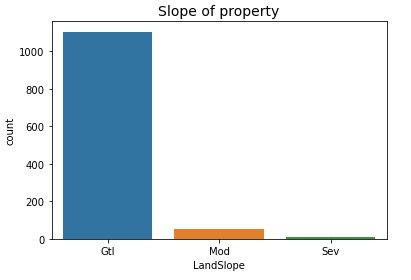
* ŷ1, ŷ2,… ŷn are predicted values
* y1,y2,…yn are observed values
* n is the number of observations

**Visualizations:**

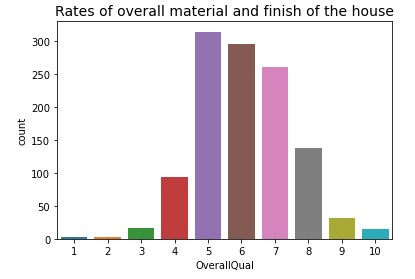
* Now, let’s look at the analysis of the data.



Out of 1168 properties, Approximately 750 properties are in regular shape, while around 400 properties are slightly irregular. (Reg=Regular, IR1=slightly irregular, IR2=Moderately Irregular, IR3=Irregular)

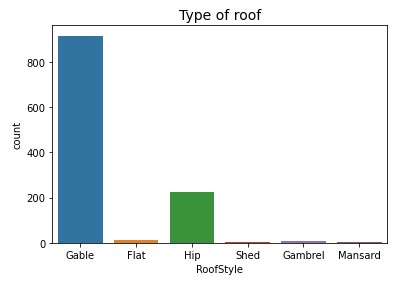


From the above plot we can see that, most of the properties land slope is gentle. While, very low number of properties land slope is moderate and severe. (Gtl=Gentle slope, Mod=Moderate Slope, Sev=Severe Slope)

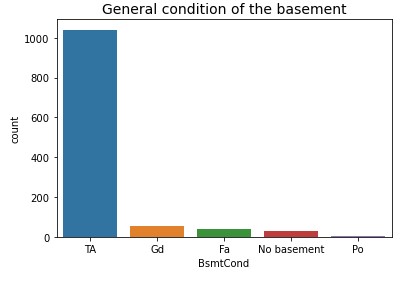


From the above plot we can see that, out of 1168 houses 314 houses overall quality is average and 295 houses overall quality is above average.(10=Very Excellent,9=Excellent,8=Very Good,7=Good,6=Above

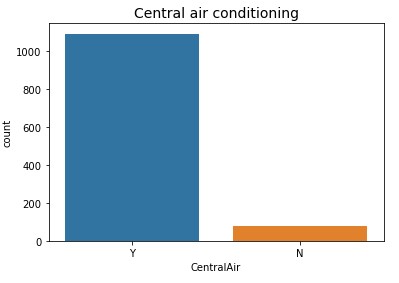
Average,5=Average,4=Below Average,3=Fair,2=Poor,1=Very Poor)



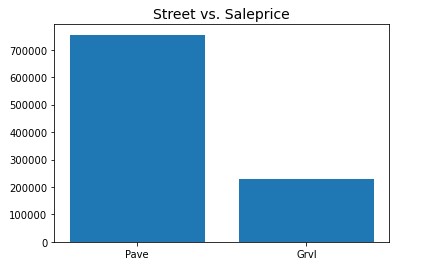
Out of 1168 houses, 915 houses roof style is Gable and only 2 houses roof style is shed.



Out of 1168 houses, 1041 houses basement condition is typical, while 30 hoses are without basement. (Gd=Good, TA=Typical - slight dampness allowed, Fa=Fair - dampness or some cracking or settling, Po=Poor - Severe cracking, settling, or wetness)



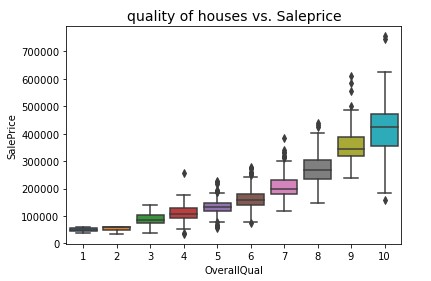
From the above count plot we can see that, most of the properties have central air conditioning system. (Y=yes, N=no)



Properties price is very high where street is paved, while price is low where street is gravel.

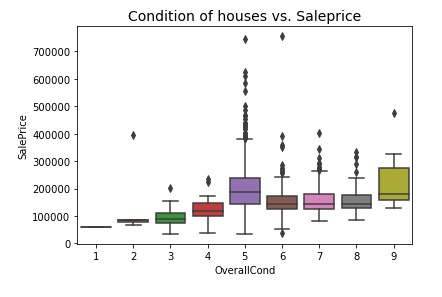


From the above bar plot we can see that, properties price is high where shape of the properties is moderately irregular, while price is low where shape of the properties regular compares to other categories. (Reg=Regular, IR1=Slightly irregular, IR2=Moderately Irregular, IR3=Irregular)



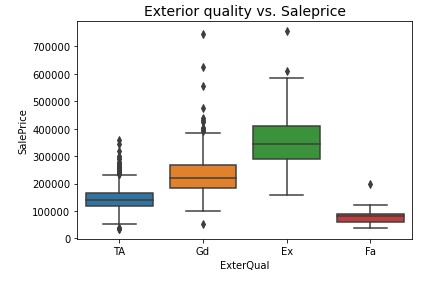
From the above box plot we can see that, price of the houses is very high where overall quality of the houses is very excellent and price decrease from going 10 to 1. (10=Very Excellent, 9=Excellent, 8=Very Good, 7=Good, 6=Above Average, 5=Average, 4=Below

Average, 3=Fair, 2=Poor, 1=Very Poor)

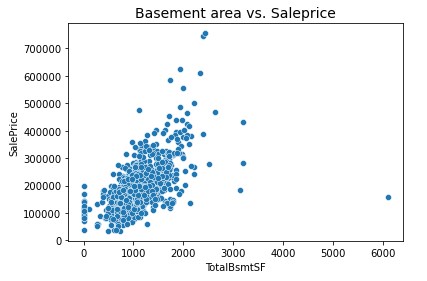


* From the above box plot we can see that, houses price is high where overall condition of the houses is excellent, while price of the houses is almost same where condition of the houses is above average, good and very good.
* Average condition houses price is higher than above average, good and very good condition houses.
* (9=Excellent, 8=Very Good, 7=Good, 6=Above Average,

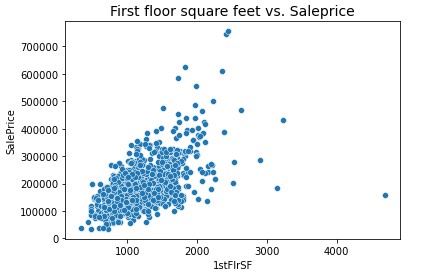
5=Average, 4=Below Average, 3=Fair, 2=Poor, 1=Very Poor)



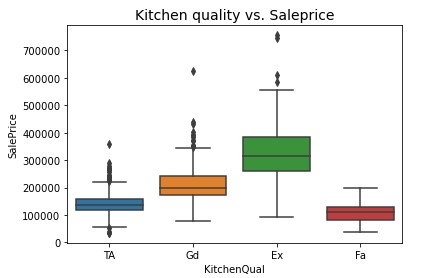
Properties price is high where quality of the material on the exterior is excellent, while price is low where quality of the material on the exterior is fair. (Ex=Excellent, Gd=Good, TA=Average/Typical, Fa=Fair, Po=Poor)



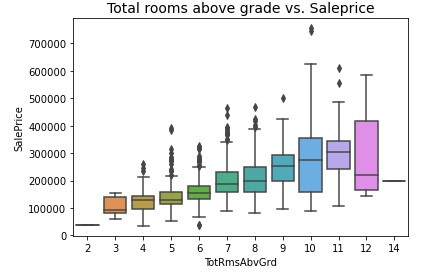
From the above scatterplot we can see that, total square feet of basement area and saleprice are positively correlated to each other. That means houses sale price is increase with increases of basement area size.



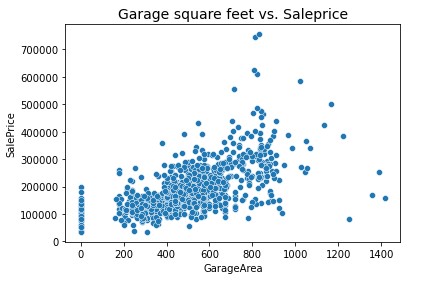
From the above scatterplot we can see that, First floor square feet and houses price are positively correlated to each other.



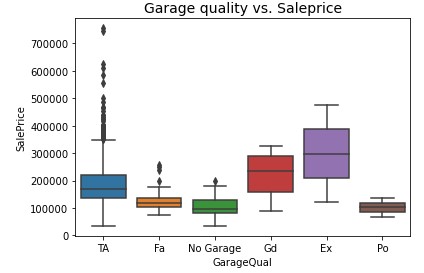
From the above box plot we can see that, houses price is higher where kitchen quality is excellent and price is lower where kitchen quality is fair compare to other categories. (Ex=Excellent, Gd=Good, TA=Typical/Average, Fa=Fair)



From the above plot we can see that, properties price is higher where total rooms above grade is 12 but price is lower where total rooms above grade is 14.



Square feet area of garage and properties price is positively correlated to each other.



Where garage quality is excellent, properties price is around 400000 dollar and price is around 100000 dollar where garage quality is poor. (Ex=Excellent, Gd=Good, TA=Typical/Average, Fa=Fair, Po=Poor, NA=No

Garage)

# CONCLUSION

**Key Findings and Conclusions of the Study:**

* The purpose of this article was twofold: to understand the pattern of Australian real estate market and make predictive model, which is able to effectively predict the price of houses in Australia.
* We use many algorithms to find best model and best result were observed of the catboost regressor with 84% r2 score accuracy.
* There are many variables important to predict the price of houses. Like quality of houses, exterior quality, basement area, kitchen quality, total rooms above grade and many more.
* In order to increase profit of surprise housing company, the company should start using of machine learning model.
* By using machine learning model company can decide whether to invest in properties or not.