# Surprise Housing Case Study Report

**Problem Statement:**

Houses are one of the necessary need 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. It is a very large market and there are various companies working in the domain. 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. Our problem is related to one such housing company. A US-based housing company named Surprise Housing has 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. The data is provided in the CSV file below. The company is looking at prospective properties to buy houses to enter the market. You are required 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 a variable?
2. How do these variables describe the price of the house?

**Business Goal:**

You are required to model the price of houses with the available independent variables. 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. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

**Technical Requirements:**

- Data contains 1460 entries each having 81 variables.
- Data contains Null values. You need to treat them using domain knowledge and your own understanding.
- Extensive EDA has to be performed to gain relationships of important variables and price.

- Data contains numerical as well as categorical variables. You need to handle them accordingly.
- You have to build Machine Learning models, apply regularization and determine the optimal values of Hyper Parameters.
- You need to find important features which affect the price positively or negatively.
- Two datasets are being provided to you (test.csv, train.csv). You will train on the train.csv dataset and predict on the test.csv file.

## EDA and Data Cleaning

To create a successful model significant exploratory data analysis was required, including identifying outliers, determining appropriate treatment of missing values. Further EDA efforts included identifying numerical columns that had a relationship with our target variable, sale price.

Additional EDA identified object columns that included categorical or ordinal information that could be converted to numerical amounts using feature engineering--identifying additional variables that are related to the target variable.

## Testing of Identified Approaches (Algorithms)

The algorithms used on training and test data are as follows:

- ➔ Linear Regression Model
- ➔ Ridge Regularization Regression Model
- ➔ Lasso Regularization Regression Model
- ➔ Support Vector Regression Model
- ➔ Decision Tree Regression Model
- ➔ Random Forest Regression Model
- ➔ K Nearest Neighbours Regression Model
- ➔ Gradient Boosting Regression Model
- ➔ AdaBoost Regression Model

- ➔ Extra Trees Regression Model

## Overall Evaluation Metrics

- ➔ **LinearRegression**
  - RMSE Score is: 24913.63300286765
  - R2 Score is: 88.52926905772831
  - Cross Validation Score: 74.15129303753828
  - R2 Score - Cross Validation Score is 14.377976020190033
- ➔ **Ridge Regularization**
  - RMSE Score is: 24815.189980744228
  - R2 Score is: 88.61974020249215
  - Cross Validation Score: 74.45483255058475
  - R2 Score - Cross Validation Score is 14.164907651907399
- ➔ **Lasso Regularization**
  - RMSE Score is: 24917.18385422087
  - R2 Score is: 88.52599905988446
  - Cross Validation Score: 74.1554161073105
  - R2 Score - Cross Validation Score is 14.370582952573955
- ➔ **SupportVectorRegression**
  - RMSE Score is: 76592.05128076131
  - R2 Score is: -8.413750687388166
  - Cross Validation Score: -6.214424099645246
  - R2 Score - Cross Validation Score is -2.1993265877429202
- ➔ **DecisionTreeRegressor**
  - RMSE Score is: 57727.62379648374

  - R2 Score is: 38.41366921116711
  - Cross Validation Score: 41.26696984258857
  - R2 Score - Cross Validation Score is -2.8533006314214617

- **RandomForestRegressor**
  - RMSE Score is: 40500.20926981103
  - R2 Score is: 69.68681972622807
  - Cross Validation Score: 64.69051255780563
  - R2 Score - Cross Validation Score is 4.996307168422433

- **K Neighbors Regressor**
  - RMSE Score is: 39967.560097269954
  - R2 Score is: 70.47892004289771
  - Cross Validation Score: 63.17026924423994
  - R2 Score - Cross Validation Score is 7.308650798657773

- **Gradient Boosting Regressor**
  - RMSE Score is: 34940.85742452848
  - R2 Score is: 77.43766288942612
  - Cross Validation Score: 79.92477771412683
  - R2 Score - Cross Validation Score is -2.4871148247007113

- **AdaBoostRegressor**
  - RMSE Score is: 31820.346272586143
  - R2 Score is: 81.28771728128767
  - Cross Validation Score: 79.16566313678824
  - R2 Score - Cross Validation Score is 2.1220541444994296

- **ExtraTreesRegressor**
  - RMSE Score is: 23795.64151951764
  - R2 Score is: 89.53566093796192

- Cross Validation Score: 84.77260003691624
- R2 Score - Cross Validation Score is 4.763060901045677

As we evaluated the metrics, we found ExtraTreesRegressor as the best model to choose for hyper parameter tuning.

Used Grid Search CV for hyper parameter tuning and found out that the default values were giving the best values and hyper parameters tuning didn’t work here (it happens sometimes).

## Conclusion:

Generated a strong predictive model based on thorough exploratory data analysis, feature engineering and regression methods.

Overall R2 score (using Extra trees regressor) indicates ~ 89.53% which is considered to be quite good for the model but also getting high RMSE value, which needs to be minimized.

To increase the Sale Price of a home, the most important feature to focus on is the overall quality score, along with ensuring the living area has a sufficient amount of square footage and the home is equipped with central air!