Introduction:

The housing and real estate market is a vital contributor to the global economy, impacting individuals' fundamental needs worldwide. Leveraging data science tools, companies can enhance revenue, refine marketing strategies, and adapt to evolving trends in property transactions. Predictive modeling and market analysis techniques empower housing firms to achieve their business objectives.

Problem Statement

Surprise Housing—a US-based company venturing into Australia\'s market, aiming to leverage data analytics for strategic property investments. The project's core objective is to develop a machine learning model that predicts property values, enabling informed investment decisions and offering insights into the complex dynamics of the housing market.

We 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.

Data Collection and Preprocessing:

Two datasets are being provided to us(test.csv, train.csv). We will train on the train.csv dataset and predict on the test.csv file. The “Data file.csv” and “Datadescription.txt” are enclosed with this file

The dataset initally consists of 1460 entries with 81 variables. The target feature is Salesprice. So it is a regression type of problem. The given dataset is dirty as it contains many missing value , significant outliers are present .

Exploratory Data Analysis (EDA):

For that extensive data analysis is performed to remove the null values, checking for multi-collinearity , finding out outliers.The columns exhibiting high multicollinearity is subsequently removed . By the help of skewness we understand the direction of outliers present in our dataset and by checking

correlation matrix we understood which variables are affecting the price of our target feature positively and negatively,

Then i separated both the numerical and categorical features present in our dataset. Countplot is plotted for the categorical variable to give us a better insight. Then we plotted the boxplot showing the impact of each categorical variable with our target Salesprice. For the numerical features i plotted probability distribution function to check how each one is varying visually.
Since our dataset is not free from outliers and removing outliers is not feasible in this case as it removes almost 60% of our data so we impute the outliers with the upper and lower quartile accordingly based on the interquartile range.

Finally we encoded all the categorical variables with the label encoder as using one hot encoder will increase the dimensionality too much. Since our training and testing datasets are provided in 2 different files so we repeat all the same preprocessing steps for our test dataset as well.

Model Selection and Training:

Training and test datasets should ideally both have target variables. In the training dataset, the target variable is used to train the model and in the test dataset, the target variable is used to measure the performance of the model on the dataset .
The notation usually used to represent these are as follows:
x_train- independent variable of training dataset
y_train- target variable of training dataset
x_test- independent variable of test dataset
y_test- target variable of test dataset
Since in our case for test dataset y_test is not provided so we cannot test our dataset on it . In fact according to the problem our aim is to predict the target feature ("salesprice") of the test dataset.
So we split our training dataset into training and validation sets and applied the algorithyms to check one by one which algorithym is fitting well.

ML models used & Model evaluation:

For this problem we applied Linear Regression, Random Forest, Gradient Boasting , KNN regressor and Xgboast and evaluated our performance for each of these algorithym by the help of evaluation metrics like Mean Absolute error, Root Mean Square Error and R2 score (co-efficient of determination)
The model which has less Root mean square error and a higher R2 value is the best fitting model for this dataset. I find out that Random Forest gives the best result compared to the others .

Conclusion:

After finding the best model I use that model to predict the target feature of our already preprocessing test dataset.

Business Implications:

The developed machine learning model holds substantial business implications for Surprise Housing's expansion into the Australian market. By accurately predicting house prices based on key independent variables, the company gains the strategic advantage of informed property investment decisions. This empowers the management to allocate resources effectively, focus on high-return opportunities, and tailor their market entry strategy for optimal outcomes. Furthermore, the model's insights into pricing dynamics offer valuable market intelligence, enabling the management to navigate the new market landscape with confidence and precision.