Predicting House Price Using Machine Learning

Team leader Name: KOTA POLI SIVA SUNEEL

Team Leader Register No: 211521104076

Team Members:

Member 1:

Name: BHEEMAVARAM PAVAN KUMAR

Register No: 211521104022

Member 2:

Name: THATIPARTHI HEMANTH SAI

Register No: 211521104171

Member 3:

Name: THATIPARTHI SARATH CHANDRA REDDY

Register No: 211521104168

Member 4:

Name: VISHAL. M

Register No: 211521104180

Problem Statement

The housing market is an important and complex sector that impacts people's lives in many ways. For many individuals and families, buying a house is one of the biggest investments they will make in their lifetime. Therefore, it is essential to accurately predict the prices of houses so that buyers and sellers can make informed decisions. This project aims to use machine learning techniques to predict house prices based on various features such as location, square footage, number of bedrooms and bathrooms, and other relevant factors.

Design Thinking:

1. Data Source: Choose a dataset containing information about houses, including features like location, square footage, bedrooms, bathrooms, and price.

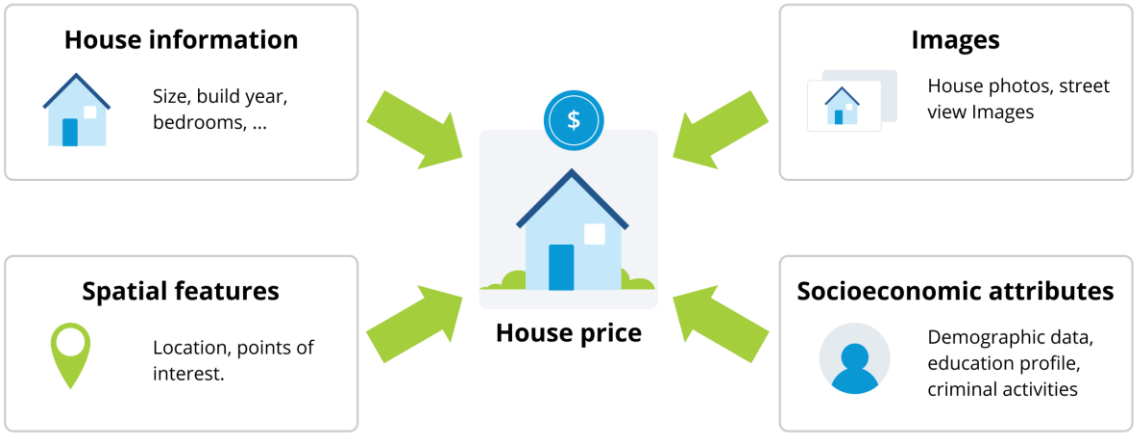


Figure 1: Data Source

2. Data Preprocessing: Clean and preprocess the data, handle missing values, and convert categorical features into numerical representations.

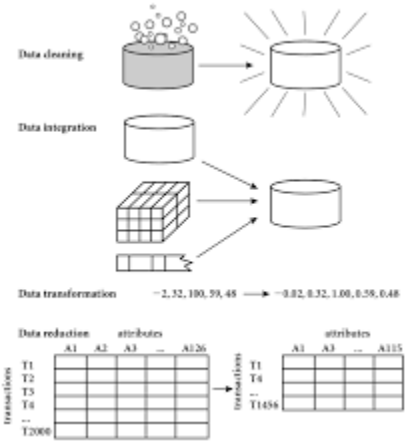


Figure 2: Data Preprocessing

3. Feature Selection: Select the most relevant features for predicting house prices.

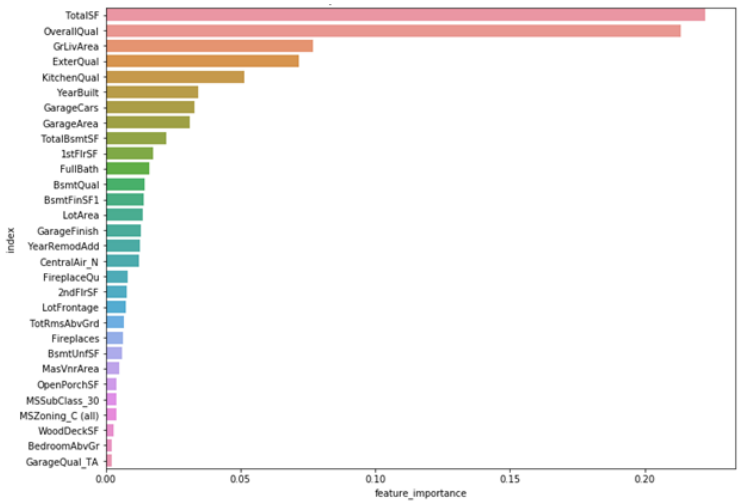


Figure 3:Feature Selection

4. Model Selection: Choose a suitable regression algorithm (e.g., Linear Regression, Random Forest Regressor) for predicting house prices.

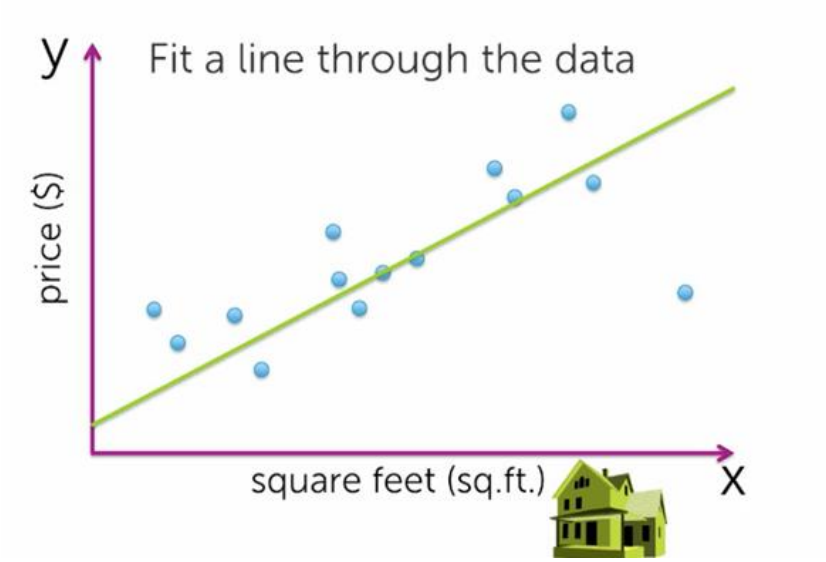
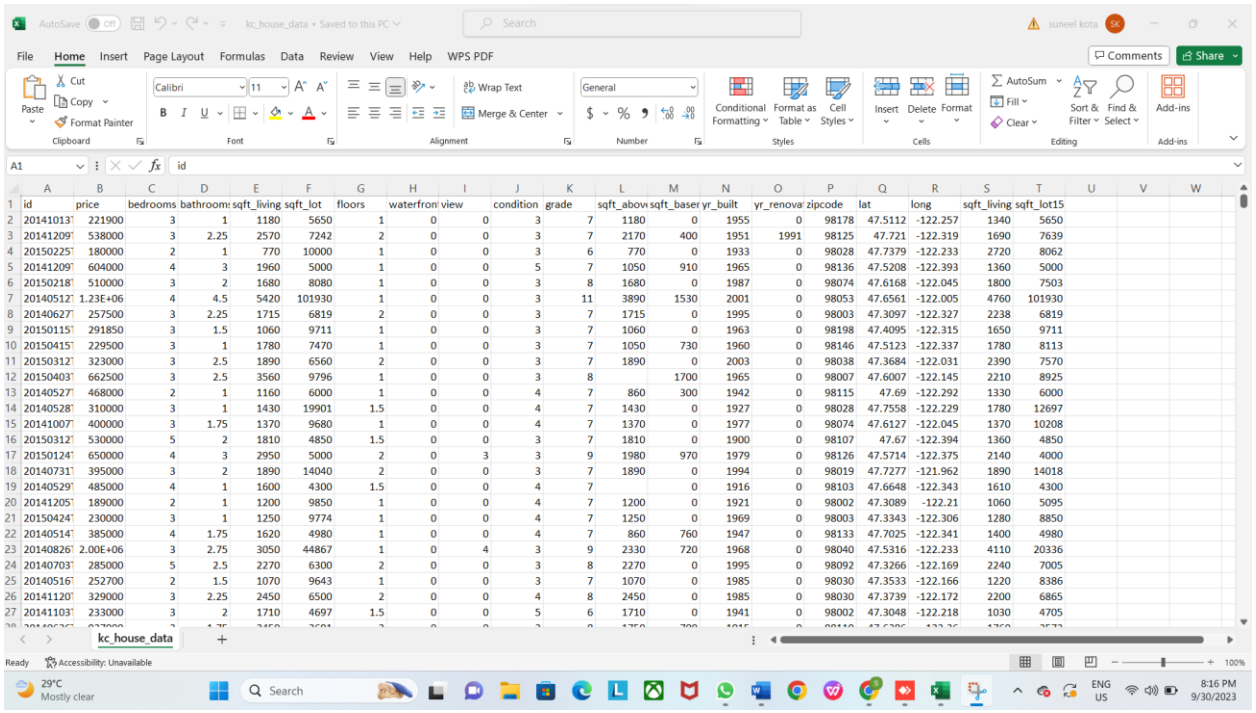


Figure 4: Model Selection

5. Model Training: Train the selected model using the preprocessed data.
6. Evaluation: Evaluate the model's performance using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.

Dataset Used:

The dataset used here is kc_house_data.csv from kaggle. This dataset has features like No. of bedroom, No. of bathroom, lat, lon, etc. but in our model we are using some features like No. of bedroom, No. of bathroom, zip code, Sqft living.



id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
2014101370	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
2014120900	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400	1951	1991	98125	47.721	-122.319	1690	7639
2015022500	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
2014120900	604000	4	3	1960	5000	1	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
2015021800	510000	3	2	1680	8080	1	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503
2014051200	1.23E+06	4	4.5	5420	101930	1	0	0	3	11	3890	1530	2001	0	98053	47.6561	-122.005	4760	101930
2014062700	257500	3	2.25	1715	6819	2	0	0	3	7	1715	0	1995	0	98003	47.3097	-122.327	2238	6819
2015011500	291850	3	1.5	1060	9711	1	0	0	3	7	1060	0	1963	0	98198	47.4095	-122.315	1650	9711
2015041500	229500	3	1	1780	7470	1	0	0	3	7	1050	730	1960	0	98146	47.5123	-122.337	1780	8113
2015031200	323000	3	2.5	1890	6560	2	0	0	3	7	1890	0	2003	0	98038	47.3684	-122.031	2390	7570
2015040300	662500	3	2.5	3560	9796	1	0	0	3	8		1700	1965	0	98007	47.6007	-122.145	2210	8925
2014052700	468000	2	1	1160	6000	1	0	0	4	7	860	300	1942	0	98115	47.69	-122.292	1330	6000
2014052800	310000	3	1	1430	19901	1.5	0	0	4	7	1430	0	1927	0	98028	47.7558	-122.229	1780	12697
2014100700	400000	3	1.75	1370	9680	1	0	0	4	7	1370	0	1977	0	98074	47.6127	-122.045	1370	10208
2015031200	530000	5	2	1810	4850	1.5	0	0	3	7	1810	0	1900	0	98107	47.67	-122.394	1360	4850
2015012400	650000	4	3	2950	5000	2	0	3	3	9	1980	970	1979	0	98126	47.5714	-122.375	2140	4000
2014073100	395000	3	2	1890	14040	2	0	0	3	7	1890	0	1994	0	98019	47.7227	-121.962	1890	14018
2014052900	485000	4	1	1600	4300	1.5	0	0	4	7		0	1916	0	98103	47.6648	-122.343	1610	4300
2014120500	189000	2	1	1200	9850	1	0	0	4	7	1200	0	1921	0	98002	47.3089	-122.21	1060	5095
2015042400	230000	3	1	1250	9774	1	0	0	4	7	1250	0	1969	0	98003	47.3343	-122.306	1280	8850
2014051400	385000	4	1.75	1620	4980	1	0	0	4	7	860	760	1947	0	98133	47.7025	-122.341	1400	4980
2014082600	2.00E+06	3	2.75	3050	44867	1	0	4	3	9	2330	720	1968	0	98040	47.5316	-122.233	4110	20336
2014070300	285000	5	2.5	2270	6300	2	0	0	3	8	2270	0	1995	0	98092	47.3266	-122.169	2240	7005
2014051600	252700	2	1.5	1070	9643	1	0	0	3	7	1070	0	1985	0	98030	47.3533	-122.166	1220	8386
2014112000	329000	3	2.25	2450	6500	2	0	0	4	8	2450	0	1985	0	98030	47.3739	-122.172	2200	6865
2014110300	233000	3	2	1710	4697	1.5	0	0	5	6	1710	0	1941	0	98002	47.3048	-122.218	1030	4705

Data preprocessing steps:

- Import the required libraries and modules, including pandas for data manipulation, scikit-learn for machine learning algorithms, and Linear Regression for the linear regression model.
- Loading the required dataset with pd.read_csv and select the features we want to use for prediction (e.g., bedrooms, bathrooms, sqft_living, sqft_lot, floors, and zip code), as well as the target variable (price).
- Split the data into a training set and a test set using the train_test_split function, with 80% of the data used for training and 20% for testing.
- Create an instance of the linear regression model using LinearRegression(). We then perform the model training by calling the function fit() with the training data.

Linear Regression:

Linear regression is a mainly used technique for the prediction of house prices due to its simplicity and interpretability. It assumes a linear relationship between the independent variables (such as how many bedrooms, number of bathrooms, and square footage) and the dependent variable (house price). By fitting a linear regression model to historical data, we can estimate the coefficients that represent the relationship between the target variable and the features. This enables us to make predictions on new data by multiplying the feature values with their respective coefficients and summing them up. Linear regression provides insights into the impact of each feature on the house price, enabling us to understand the significance of different factors and make informed decisions in the real estate market.

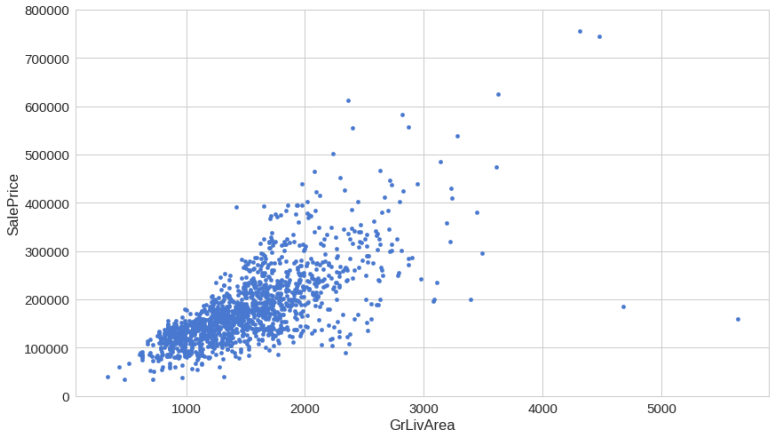


Figure 5: Linear Regression

Data Processing Steps:

Exploratory Data Analysis (EDA)

i. Summary Statistics
ii. Data Visualization
iii. Key Insights

Feature Selection and Engineering

i. Feature Importance Analysis
ii. New Feature Creation

Model Selection

i. Algorithm Selection
ii. Model Justification

Model Training

i. Train-Test Split
ii. Hyperparameter Tuning

Model Evaluation

i. Evaluation Metrics
ii. Model Performance
iii. Interpretability

Deployment

i. Deployment Environment
ii. API/Interface Description

Monitoring and Maintenance

i. Performance Monitoring
ii. Scheduled Retraining

Documentation of Code and Scripts

i. Directory Structure
ii. Code Overview
iii. Dependencies

Results and Findings

i. Model Insights
ii. Visualizations
iii. Key Takeaways

Recommendations

i. Future Improvements
ii. Potential Enhancements

Conclusion

i. Project Recap
ii. Achievements

References

i. Data Sources

ii. Literature

Appendices

i. Additional Information

ii. Glossary