House Price Prediction

This presentation outlines the steps involved in a house price prediction project, from exploratory data analysis to model building and evaluation. By leveraging various features, we aim to accurately forecast house prices and provide valuable insights for home buyers and sellers.

Introduction to the Project

Objective

Develop a predictive model to forecast house prices based on location, size, and other key features.

Significance

Accurate price predictions can assist homebuyers, sellers, and real estate professionals in making informed decisions.

Approach

Leverage advanced data analysis and machine learning techniques to uncover patterns and relationships in the data.

Exploratory Data Analysis

1 Data Inspection

columns contain missing values more than 60% PoolQC, MiscFeature, Alley, Fence

4]:

	col	dtype	unique values	count unique	count null	percentage null
72	PoolQC	object	[nan, Ex, Fa, Gd]	3	1453	99.520548
74	MiscFeature	object	[nan, Shed, Gar2, Othr, TenC]	4	1406	96.301370
6	Alley	object	[nan, Grvl, Pave]	2	1369	93.767123
73	Fence	object	[nan, MnPrv, GdWo, GdPrv, MnWw]	4	1179	80.753425
25	MasVnrType	object	[BrkFace, nan, Stone, BrkCmn]	3	872	59.726027
57	FireplaceQu	object	[nan, TA, Gd, Fa, Ex, Po]	5	690	47.260274
3	LotFrontage	float64	[65.0, 80.0, 68.0, 60.0, 84.0, 85.0, 75.0, nan...	110	259	17.739726
58	GarageType	object	[Attchd, Detchd, BuiltIn, CarPort, nan, Basmen...	6	81	5.547945
59	GarageYrBlt	float64	[2003.0, 1976.0, 2001.0, 1998.0, 2000.0, 1993....	97	81	5.547945
60	GarageFinish	object	[RFn, Unf, Fin, nan]	3	81	5.547945
63	GarageQual	object	[TA, Fa, Gd, nan, Ex, Po]	5	81	5.547945
64	GarageCond	object	[TA, Fa, nan, Gd, Po, Ex]	5	81	5.547945
32	BsmtExposure	object	[No, Gd, Mn, Av, nan]	4	38	2.602740
35	BsmtFinType2	object	[Unf, BLQ, nan, ALQ, Rec, LwQ, GLQ]	6	38	2.602740
30	BsmtQual	object	[Gd, TA, Ex, nan, Fa]	4	37	2.534247
31	BsmtCond	object	[TA, Gd, nan, Fa, Po]	4	37	2.534247
33	BsmtFinType1	object	[GLQ, ALQ, Unf, Rec, BLQ, nan, LwQ]	6	37	2.534247
26	MasVnrArea	float64	[196.0, 0.0, 162.0, 350.0, 186.0, 240.0, 286.0...	327	8	0.547945
42	Electrical	object	[SBrkr, FuseF, FuseA, FuseP, Mix, nan]	5	1	0.068493
0	Id	int64	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	1460	0	0.000000

Irrelevant column Id

Top correlation between features 0.88

```
GarageArea          GarageCars          0.882475
GarageYrBlt         YearBuilt           0.825667
TotRmsAbvGrd        GrLivArea           0.825489
1stFlrSF            TotalBsmtSF         0.819530
OverallQual         Property_Sale_Price 0.790982
GrLivArea           Property_Sale_Price 0.708624
                    2ndFlrSF            0.687501
BedroomAbvGr        TotRmsAbvGrd        0.676620
BsmtFinSF1          BsmtFullBath        0.649212
GarageYrBlt         YearRemodAdd        0.642277
Property_Sale_Price GarageCars          0.640409
FullBath            GrLivArea           0.630012
Property_Sale_Price GarageArea          0.623431
TotRmsAbvGrd        2ndFlrSF            0.616423
TotalBsmtSF         Property_Sale_Price 0.613581
2ndFlrSF            HalfBath            0.609707
Property_Sale_Price 1stFlrSF            0.605852
OverallQual         GarageCars          0.600671
                    GrLivArea           0.593007
YearBuilt           YearRemodAdd        0.592855
dtype: float64
```

2 Insights

Top correlated features with the house price OverallQual, GrLivArea, and GarageCars are strongly correlated with house prices

Critical features with missing data need to be handled properly before training the model

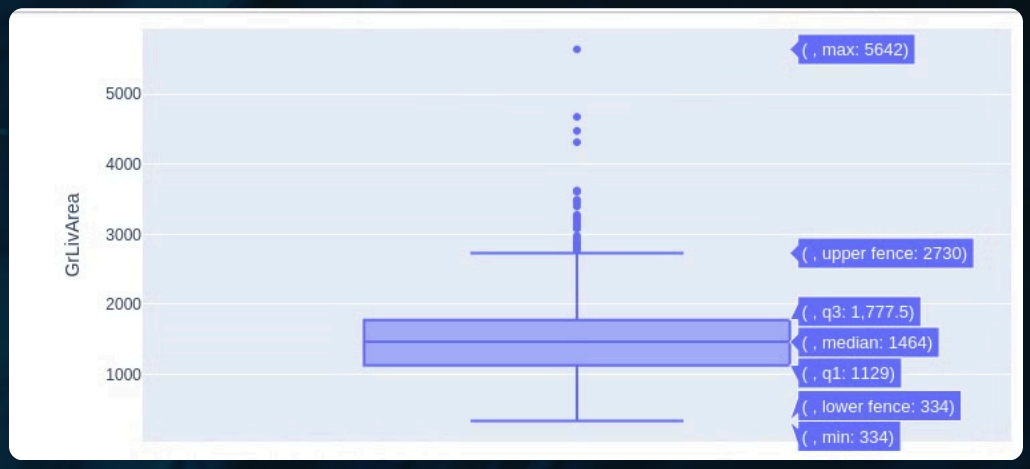
Right skew in house price and unbalance in categorical features

```
Exterior2nd
['VinylSd' 'MetalSd' 'Wd Shng' 'HdBoard' 'Plywood' 'Wd Sdng' 'CmentBd'
 'BrkFace' 'Stucco' 'AsbShng' 'Brk Cmn' 'ImStucc' 'AsphShn' 'Stone'
 'Other' 'CBlock']
Exterior2nd
VinylSd    504
MetalSd    214
HdBoard    207
Wd Sdng    197
Plywood    142
CmentBd     60
Wd Shng     38
Stucco      26
BrkFace     25
AsbShng     20
ImStucc     10
Brk Cmn      7
Stone        5
AsphShn      3
Other        1
CBlock       1
Name: count, dtype: int64
```

Data Cleaning and Preprocessing

Outlier Treatment

we handled outliers by deleting the extremes and use **IQR (Interquartile Range) Method**



Missing Values

we used imputation, dropping rows with incomplete information and deleting features with missing values more than 60%.

	col	dtype	unique values	count unique	count null	percentage null
72	PoolQC	object	[nan, Ex, Fa, Gd]	3	1453	99.520548
74	MiscFeature	object	[nan, Shed, Gar2, Othr, TenC]	4	1406	96.301370
6	Alley	object	[nan, Grvl, Pave]	2	1369	93.767123
73	Fence	object	[nan, MnPrv, GdWo, GdPrv, MnWw]	4	1179	80.753425
25	MasVnrType	object	[BrkFace, nan, Stone, BrkCmn]	3	872	59.726027
57	FireplaceQu	object	[nan, TA, Gd, Fa, Ex, Po]	5	690	47.260274
3	LotFrontage	float64	[65.0, 80.0, 68.0, 60.0, 84.0, 85.0, 75.0, nan...	110	259	17.739726
58	GarageType	object	[Attchd, Detchd, BuiltIn, CarPort, nan, Basmen...	6	81	5.547945
59	GarageYrBlt	float64	[2003.0, 1976.0, 2001.0, 1998.0, 2000.0, 1993....	97	81	5.547945
60	GarageFinish	object	[RFn, Unf, Fin, nan]	3	81	5.547945
63	GarageQual	object	[TA, Fa, Gd, nan, Ex, Po]	5	81	5.547945
64	GarageCond	object	[TA, Fa, nan, Gd, Po, Ex]	5	81	5.547945
32	BsmtExposure	object	[No, Gd, Mn, Av, nan]	4	38	2.602740
35	BsmtFinType2	object	[Unf, BLQ, nan, ALQ, Rec, LwQ, GLQ]	6	38	2.602740

```
df[['MasVnrType', 'MasVnrArea']][(df['MasVnrArea'] == 0) & (df['MasVnrType'] != "None")]
```

	MasVnrType	MasVnrArea
688	BrkFace	0.0
1241	Stone	0.0

```
df.groupby('MasVnrType')['MasVnrArea'].mean()
```

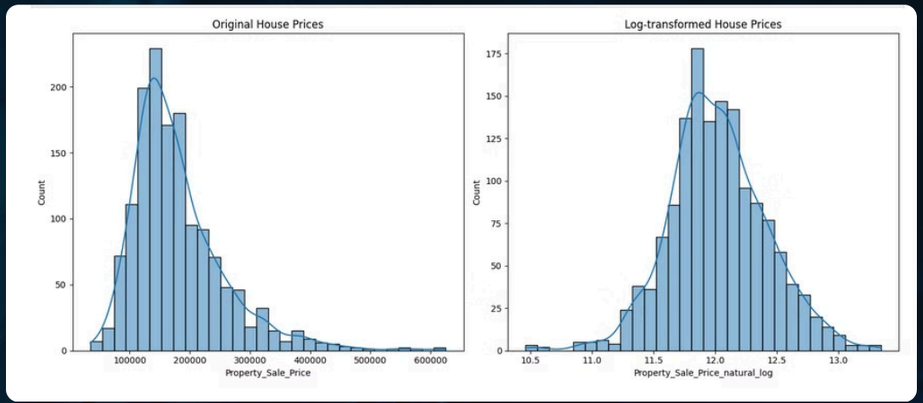
```
MasVnrType
BrkCmn     247.666667
BrkFace    256.957207
None         1.087356
Stone      239.304688
Name: MasVnrArea, dtype: float64
```

```
df.loc[688, 'MasVnrArea'] = 259.008989
df.loc[1241, 'MasVnrArea'] = 239.304688
```

Feature Engineering

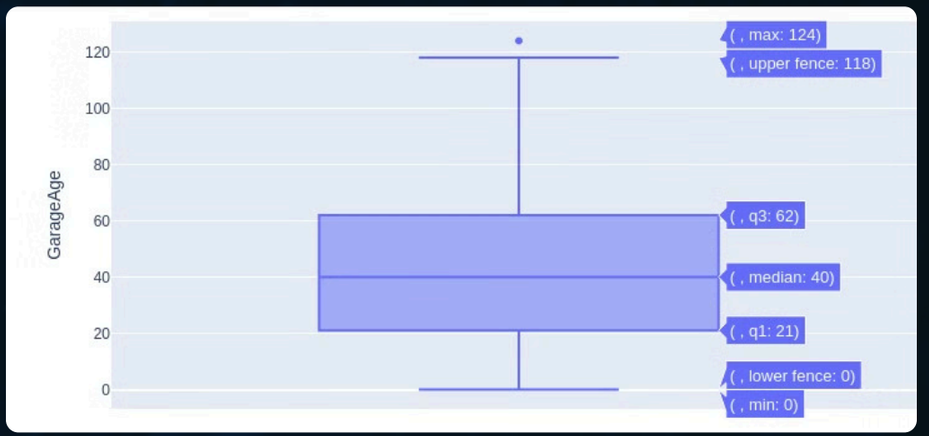
Feature Scaling

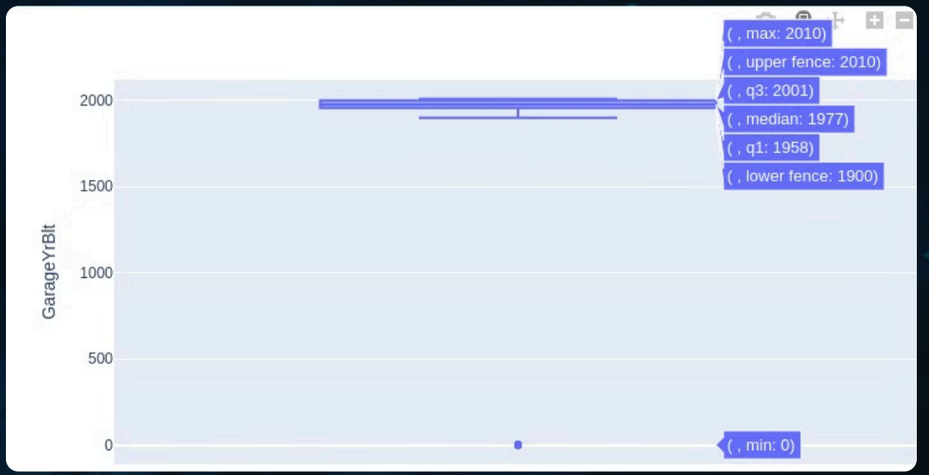
we transform the Property_Sale_Price to log format and we used StandardScaler() to scale features to avoid biases in the model training process.



Derived Features

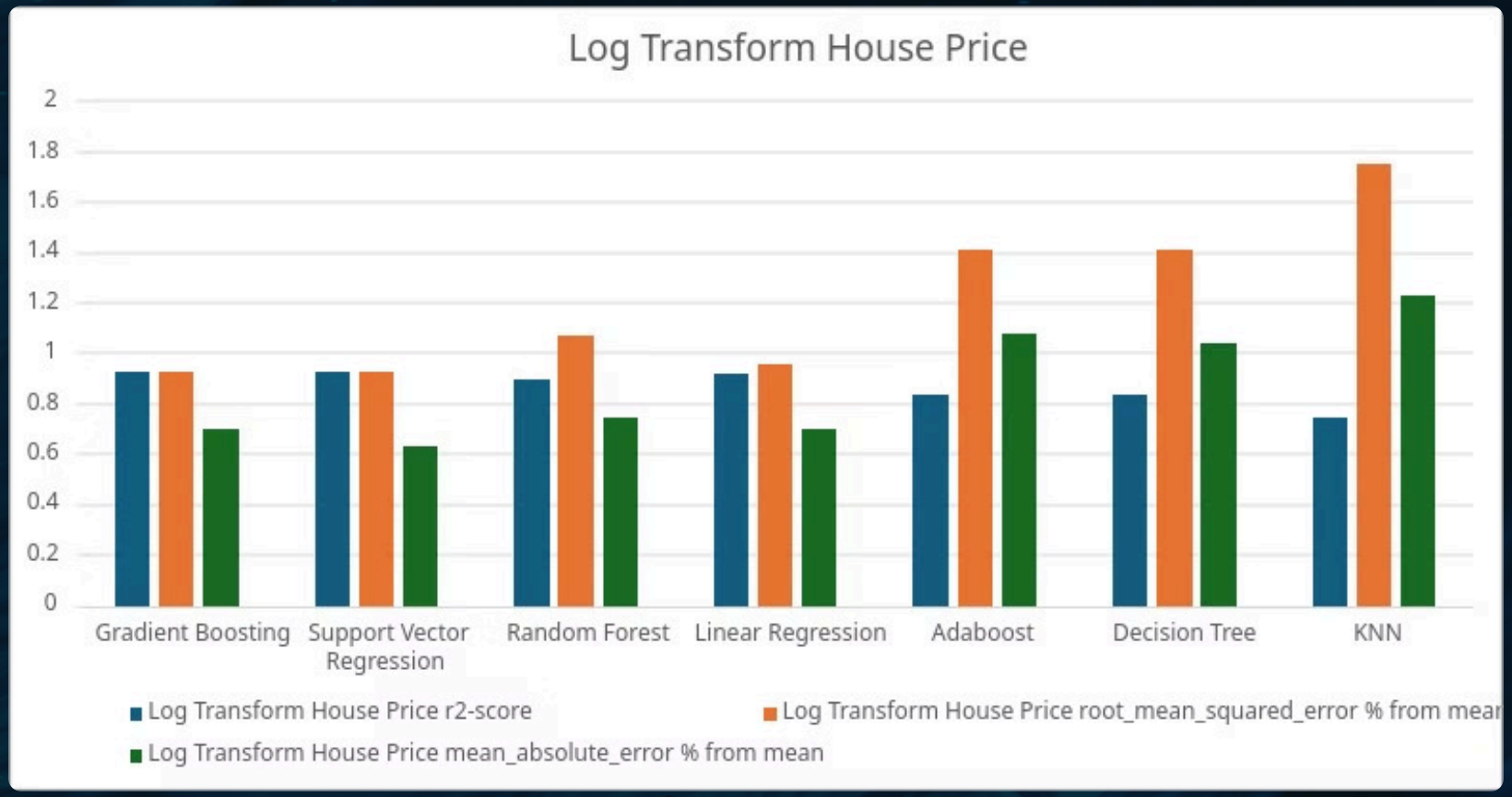
we create new feature GarageAge instead of GarageYrBlt.



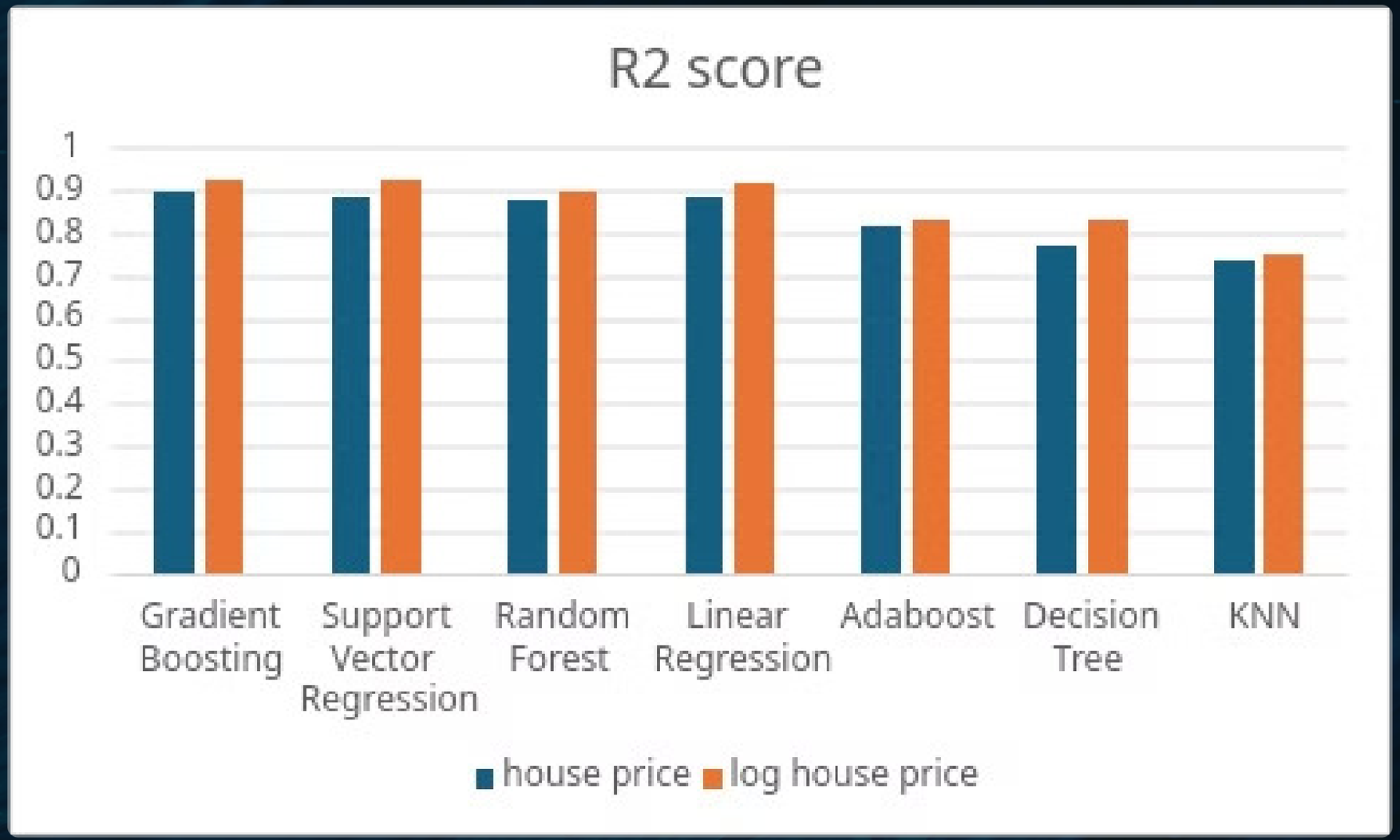


Comparing the Regression Models

Comparing the Regression Models using log transformation of target variable



Comparing R2 Score for all Regression Models



Conclusion

Accurate Predictions

The Gradient Boosting model achieved the highest performance, making it a reliable tool for house price forecasting.

Valuable Insights

The exploratory data analysis revealed key factors influencing house prices, such as location and property size.

Future Opportunities

The project can be expanded to include more data sources and explore additional machine learning techniques.