**Prediction Property Prices**



**By: Garima Choraria**

**Introduction**

Machine Learning that works with algorithms and technologies to extracts useful information from datasets. Machine Learning helps to solve the problems algorithmically rather than using purely maths. Therefore, it is based on creating algorithms that permit the machine to learn and predict.

This project is on predicting property prices upon their needs and priorities. House Property Price always depends on area, kind of building is made on it and the year it is made. For instance, a big house may be higher price if is located in desirable area than being placed in a poor neighbourhood.

Data is used for experiment by using combination of pre-processing and machine learning algorithms so that model can predict accurate price of the property.

**Research Question**

1.what are the factors that effect the price of property

2. which machine learning algorithm is best fit and giving best accuracy on prediction of price.

**Description of data**

**Data fields**

Here's a brief version of what you'll find in the data description file.

* **PropPrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
* **PropertyClass**: The building class
* **PropertyZone**: The general zoning classification
* **PropertyFrontage**: Linear feet of street connected to property
* **PropertySize**: Lot size in square feet
* **Street**: Type of road access
* **Alley**: Type of alley access
* **PropertyShape**: General shape of property
* **Elevation**: Flatness of the property
* **Amenities**: Type of Amenities available
* **LotOrientation**: Lot configuration
* **Grade**: Slope of property
* **Neighborhood**: Physical locations within Ames city limits
* **Condition1**: Proximity to main road or railroad
* **Condition2**: Proximity to main road or railroad (if a second is present)
* **BldgType**: Type of dwelling
* **PropertyStyle**: Style of dwelling
* **OverallQual**: Overall material and finish quality
* **OverallCond**: Overall condition rating
* **YearBuilt**: Original construction date
* **YearRemodAdd**: Remodel date
* **RoofStyle**: Type of roof
* **RoofMatl**: Roof material
* **Roof1Material**: Exterior covering on property
* **Roof2Material**: Exterior covering on property (if more than one material)
* **ExteriorCladdingType**: Masonry veneer type
* **ExteriorCladdingArea**: Masonry veneer area in square feet
* **ExterQual**: Exterior material quality
* **ExterCond**: Present condition of the material on the exterior
* **PropertyFooting**: Type of PropertyFooting
* **BsmntFinish**: Height of the basement
* **BsmntMaintenance**: General condition of the basement
* **BsmntVisibility**: Walkout or garden level basement walls
* **BsmntFinRat1**: Quality of basement finished area
* **BsmntFinSty1**: Type 1 finished square feet
* **BsmntFinQual1**: Quality of second finished area (if present)
* **BsmtFinSF2**: Type 2 finished square feet
* **BsmtUnfSF**: Unfinished square feet of basement area
* **BsmntSqFtage**: Total square feet of basement area
* **Heating**: Type of heating
* **HeatingEfficiency**: Heating quality and condition
* **CentralAir**: Central air conditioning
* **Electrical**: Electrical system
* **1stFlrSF**: First Floor square feet
* **2ndFlrSF**: Second floor square feet
* **LowQualFinSF**: Low quality finished square feet (all floors)
* **GrLivArea**: Above grade (ground) living area square feet
* **BsmtBath1**: Basement full bathrooms
* **BsmtBath2**: Basement half bathrooms
* **Bath1**: Full bathrooms above grade
* **Bath2**: Half baths above grade
* **BedroomUpLev**: Number of BedroomUpLevs above basement level
* **Kitchen**: Number of kitchens
* **KitchenQual**: Kitchen quality
* **CntRmsUpLev**: Total rooms above grade (does not include bathrooms)
* **Functional**: Home functionality rating
* **CntFireplaces**: Number of Fireplaces
* **QualFireplace**: Fireplace quality
* **BasementType**: Garage location
* **BasementYrBlt**: Year garage was built
* **BasementFinish**: Interior finish of the garage
* **BasementCars**: Size of garage in car capacity
* **SquareFootage**: Size of garage in square feet
* **BasementQual**: Garage quality
* **BasementSqFootage**: Garage condition
* **PavedDrive**: Paved driveway
* **WoodDeckSF**: Wood deck area in square feet
* **OpenPorchSF**: Open porch area in square feet
* **EnclosedPorch**: Enclosed porch area in square feet
* **3SsnPorch**: Three season porch area in square feet
* **ScreenPorch**: Screen porch area in square feet
* **PoolArea**: Pool area in square feet
* **PoolQC**: Pool quality
* **BoundaryFeatures**: BoundaryFeatures quality
* **AddFeatures**: Miscellaneous feature not covered in other categories
* **AddVal**: $Value of miscellaneous feature
* **SaleMon**: Month Sold
* **YrSold**: Year Sold
* **SaleType**: Type of sale
* **SaleCondn**: Condition of sale

**Objectives:**

1. Import dataset in jupyter and processing data findings name of columns, shape, type of columns it has (categorical or numerical), getting the mathematical information about data.

2. For making data clean the important step is finding null or missing values and also duplicates data and fill them or delete duplicates so that best accurate outcome can be achieved.

3. performing Exploratory Data Analysis (EDA) on the collected data to identify key variables that influence property prices.

4. To find the outliers in the data and removing the outliers by using IQR technique.

5. Data Preprocessing is used for separating numerical and categorical data and after that using level encoding technic. So that model can perform best way.

6. Feature engineering is used to transfer raw data into features that are suitable for model to understand by selecting, transforming and extracting the most relevant data to build the accurate model.

7. Selecting and comparing data from sklearn machine learning and finding out best model which prediction is highest

8. Model Evaluation is done to check accuracy in percentage (%) so that we can get the conclusion which model will give high accuracy.

9.Plotting Residuals so that can get error between a predicted value and actual value and observe actual values.

**Handling Missing values:**

In given dataset it has 19 missing values and in few columns like Allay, Extrerior Cladding Type, Qual fire place, PoolQC, Boundary Features, Add Features has more than 30% null value which will affects the accuracy of model so prefer to drop this column rest null columns are filled by mean and mode.

For numerical columns used mean to full the null values and mode for categorial columns. In result achieved a clean dataset with no missing values and converted cleaned data set into CSV file.

**Exploratory Data Analysis (EDA):**

\*The average PropertySize is around 10,516 square feet, with a wide range of sizes, as indicated by the standard deviation of approximately 9,981 square feet.

\*The average OverallQual of properties is approximately 6.1, suggesting a moderate quality level. The overall condition (OverallCond) also provides an overview of the general condition of the properties.

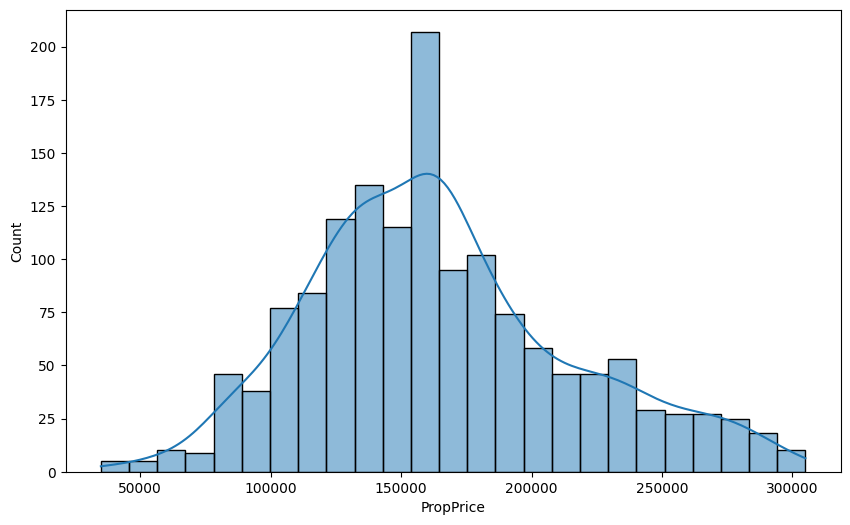
\*The dataset spans properties built from 1872 to 2010, with an average construction year (YearBuilt) of around 1971. The average year of remodelling or addition (YearRemodAdd) is approximately 1984, indicating historical development trends.

\*Features like WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, and PoolArea provide information about the presence of amenities in the properties.

\*Property prices (PropPrice) range from $34,900 to $755,000. The average price is approximately $180,921.20, with a standard deviation of $79,442.50, indicating a considerable range in property values.

\* Explore sales trends over time by analysing the distribution of sale months (SaleMon) and years (SaleYr), providing insights into potential seasonal variations or trends in property sales.

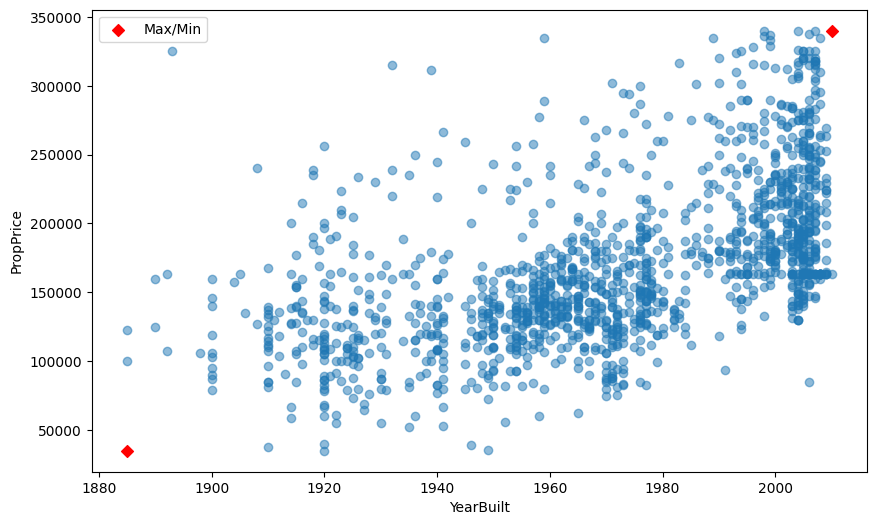
\* For finding skewness in Property Price used a histplot from seaborn library and it found that Property Price column is positive skew as its tail to right side of the distribution it means that mean and median is greater than mode.



In the dataset, the maximum property price is $340,000, while the minimum property price is $34,900. This indicates a significant range in property prices.

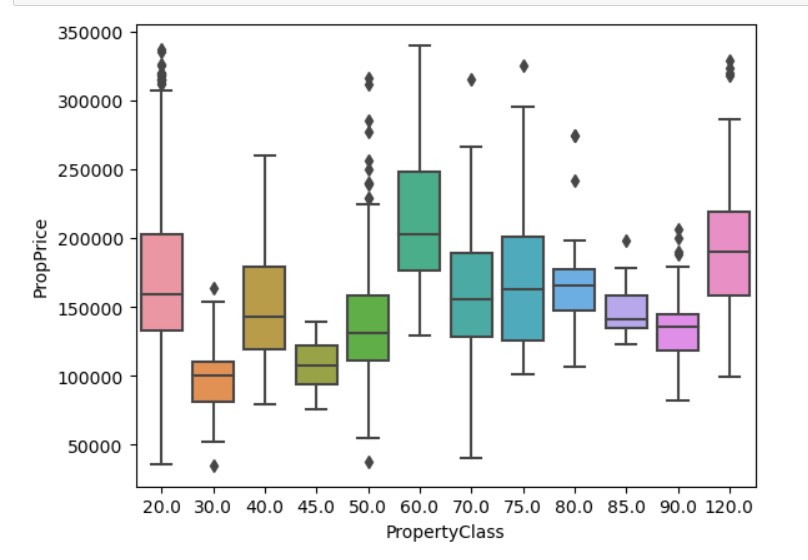
Similarly, the maximum year built for properties in the dataset is 2010, suggesting that the dataset includes relatively recent constructions. On the other hand, the minimum year built is 1885, indicating the presence of older properties in the dataset.

These insights highlight the diversity in property prices and construction years within the dataset, providing a broad perspective on the types of properties considered for analysis.

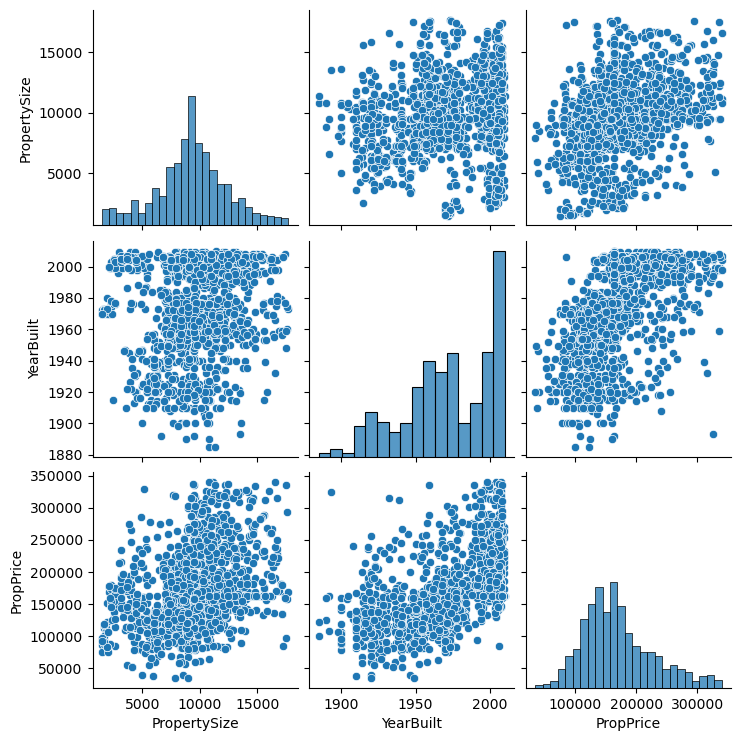


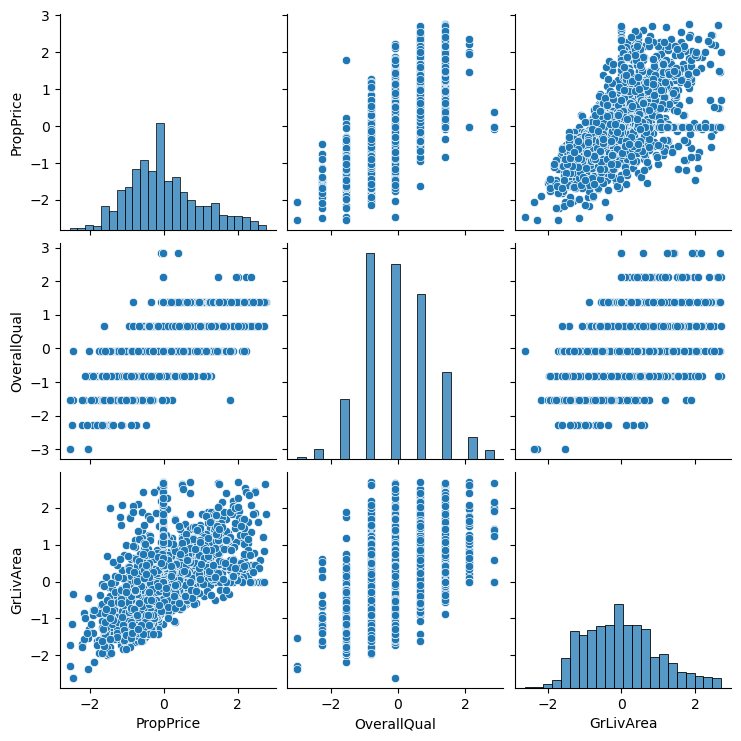
\*For calculating the object distributions, I choose to use countplot for all object column by using for loop.

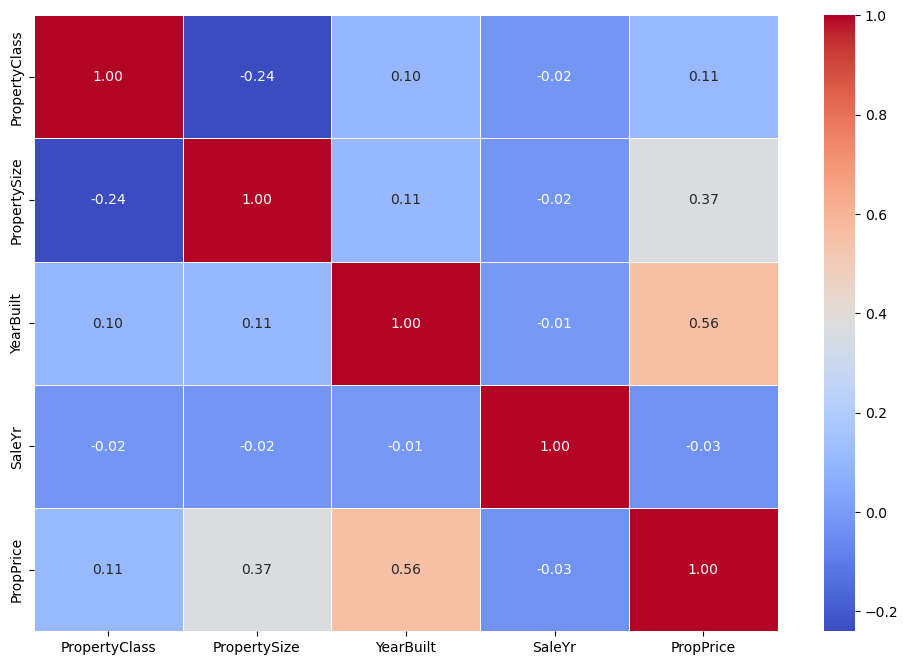
\* For finding outliers in data used box plot and found outliers and value beyond 3 standard deviations were considered as outliers and where either removed or transformed.

\*Displays the distribution of 'PropPrice' for different 'PropertyClass'. It helps identify the central tendency and spread of property prices across different classes.

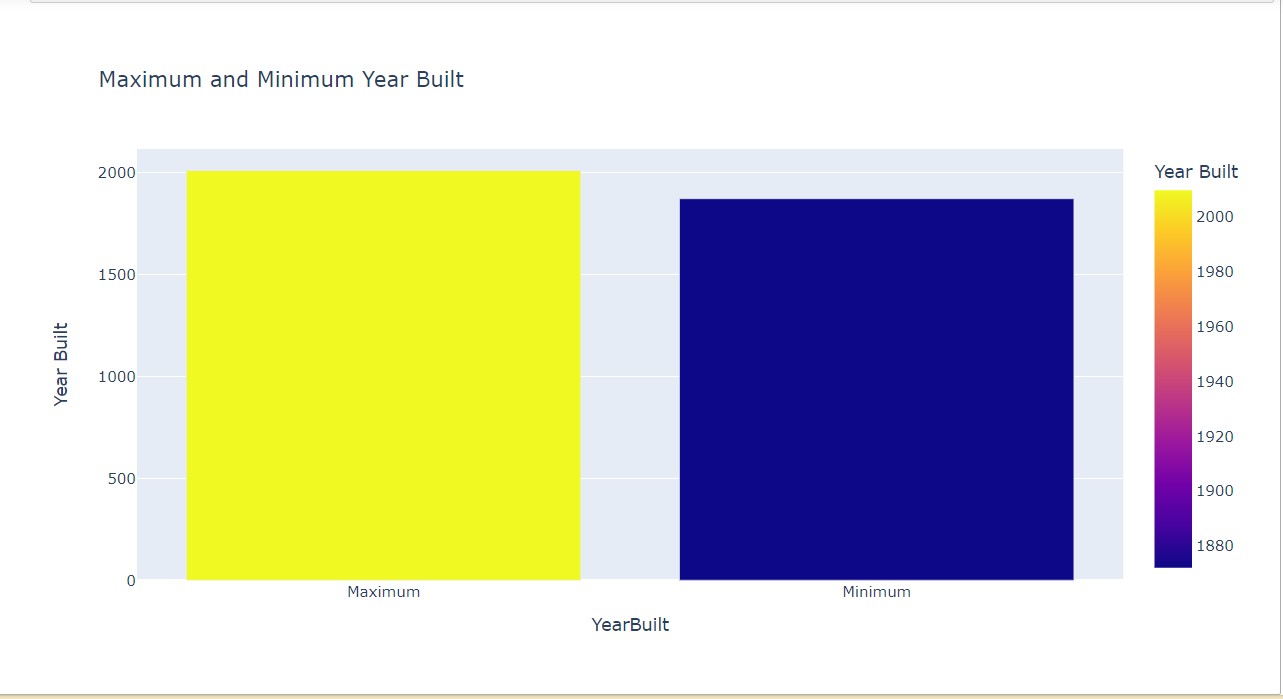
\*For finding relation between PropertySize,YearBuilt,PropPrice and Prop Price, Overall Qual, Gr Liv Area, used pair plot This plot shows scatterplots for 'PropertySize' vs 'PropertySize', 'YearBuilt' vs 'YearBuilt', and 'PropPrice' vs 'PropPrice', ‘OverallQual’ vs ‘OverallQual’, ‘GrLivArea’ vs ‘GrLivArea’. Diagonal histograms display the distribution of each variable.





\*Finding correlation between 'PropertyClass', 'PropertySize', 'YearBuilt', 'SaleYr', 'PropPrice' as it shows that Property Price is highly correlated with Yearbuilt 

\*By plotting minimum and maximum built shows that as the year increase price of property is also raising.



\*Using Label encoding for transferring categorial column into numerical value so that apply machine learning model will be easy and will get more appropriate outcome.

**Feature Engineering:**

Feature Engineering refers to process of using to transform the most relevant variables from raw data before creating a predictive model using machine learning. Feature Engineering helps to improve the performance of machine learning algorithms. For this project created a feature for total square footage of the basement and age of the property at the time of sale.

**Model Training and Test and Model Selection:**

Selecting the most appropriate model in machine learning is Model selection. In process of selecting model first step is to spilt the data into part

1. Train Data
2. Test Data

So we can check the prediction done by the model. Various machine learning algorithms, including linear regression, decision trees regression, and random forests regression and mean squared error (r2 square) will be evaluated to determine the best model for predicting property prices.

**Model Evaluation:**

The performance of the model has evaluated by linear regression model, decision tree model, random forest and after evaluating found that the Random Forest model is performing the best among the three evaluated models based on the lower Mean Squared Error and higher R-squared values. The outcome from all models are:

|  |  |  |
| --- | --- | --- |
| **Linear Regression Model** | **Decision Tree Model** | **Random Forest Model** |
| **Mean Squared Error:**  0.2994184572877736  **R-Squared:**  0.7207087592098551 | **Mean Squared Error:**  0.47798836899824754  **R-Squared:**  0.5541425005323833 | **Mean Squared Error:**  0.2434525206442352  **R-Squared:**  0.7729126080598732 |

The Random Forest model is performing the best among the three evaluated models based on the lower Mean Squared Error and higher R-squared values

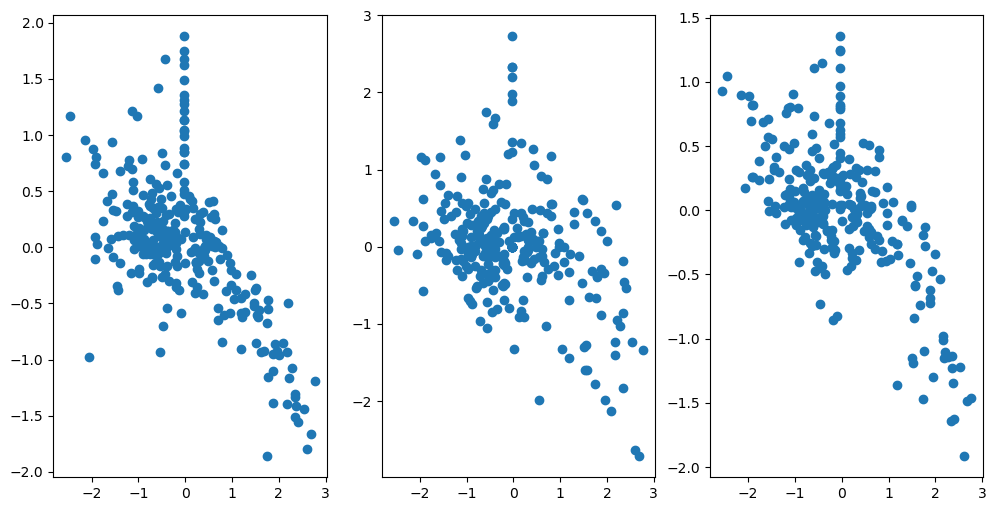
**Plotting Residuals:**

By Plotting residual, we can find how far away a point is vertically from the regression line and by plotting so findings were:

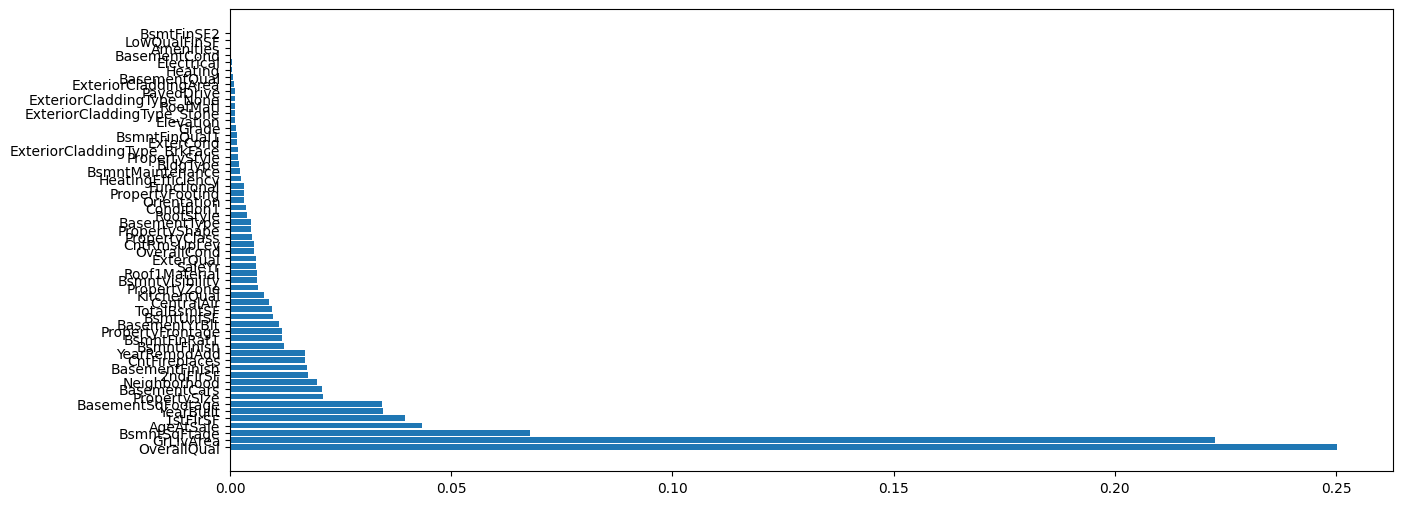
The residuals for the Linear Regression Model are concentrated between -0.3 and 0.03, indicating that the model is relatively low dispersion in its prediction.

The residuals for the Random Forest Regression Model the residual fall within the range of -0.2 to 0.2, indicating that the model is balanced spread around Zero.

The Decision Tree Regression Model exhibits a wide range of residuals, varying from -1.8 to -5.4, suggesting higher dispersion and potential inconsistency in its prediction.



The feature importance analysis reveals that attributes such as 'OverallQual', 'GrLivArea', 'YearBuilt', and 'TotalBsmtSF' significantly contribute to the predictive power of the model. Notably, certain exterior cladding types, including 'BrkFace', 'None', and 'Stone', also exhibit high importance. On the other hand, 'BsmtFinSF2' seems to have minimal impact on the model's predictions, as indicated by its importance score of 0. This information guides feature selection and highlights key factors influencing property price predictions.



**Conclusion:**

Data cleaning and Exploration:

Addressed missing values and duplicates to ensure data integrity. Explored the dataset comprehensively to gain insights into its features and distribution.

Data Preprocessing:

Employed label encoding for categorical variables to make them compatible with machine learning models. Managed outliers using a combined approach of dropping and replacing values with medians.

Feature Engineering:

Engaged in feature engineering, excluding unnecessary columns and introducing new ones where possible.

Model Building and Evaluation:

Trained various models, including Linear Regression, Decision Tree, and Random Forest, for property price prediction. Model evaluation based on Mean Squared Error and R-squared scores. Identified Random Forest as the most promising model due to superior performance.

Outliers Handling with IQR:

Implemented the Interquartile Range (IQR) technique for robust handling of outliers.

Challenges Encountered:

Successfully managed a dataset with 81 columns, overcoming computational complexities Challenges related to the dataset size were addressed effectively, demonstrating a comprehensive understanding of the dataset.