**1. Introduction:**

The main aim of this project is to predict the sales price using machine learning models with various features of independent variables. This report will give you the more information about the insights of house for the home buyers.

**2. Dataset Overview:**

1. The dataset includes 79 explanatory variables describing different aspects of houses.
2. Saleprice is the target variable, which represents the selling price of the house.

**Attribute Information:**

* SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
* MSSubClass: The building class
* MSZoning: The general zoning classification
* LotFrontage: Linear feet of street connected to property
* LotArea: Lot size in square feet
* Street: Type of road access
* Alley: Type of alley access
* LotShape: General shape of property
* LandContour: Flatness of the property
* Utilities: Type of utilities available
* LotConfig: Lot configuration
* LandSlope: 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
* HouseStyle: 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
* Exterior1st: Exterior covering on house
* Exterior2nd: Exterior covering on house (if more than one material)
* MasVnrType: Masonry veneer type
* MasVnrArea: Masonry veneer area in square feet
* ExterQual: Exterior material quality
* ExterCond: Present condition of the material on the exterior
* Foundation: Type of foundation
* BsmtQual: Height of the basement
* BsmtCond: General condition of the basement
* BsmtExposure: Walkout or garden level basement walls
* BsmtFinType1: Quality of basement finished area
* BsmtFinSF1: Type 1 finished square feet
* BsmtFinType2: Quality of second finished area (if present)
* BsmtFinSF2: Type 2 finished square feet
* BsmtUnfSF: Unfinished square feet of basement area
* TotalBsmtSF: Total square feet of basement area
* Heating: Type of heating
* HeatingQC: 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
* BsmtFullBath: Basement full bathrooms
* BsmtHalfBath: Basement half bathrooms
* FullBath: Full bathrooms above grade
* HalfBath: Half baths above grade
* Bedroom: Number of bedrooms above basement level
* Kitchen: Number of kitchens
* KitchenQual: Kitchen quality
* TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
* Functional: Home functionality rating
* Fireplaces: Number of fireplaces
* FireplaceQu: Fireplace quality
* GarageType: Garage location
* GarageYrBlt: Year garage was built
* GarageFinish: Interior finish of the garage
* GarageCars: Size of garage in car capacity
* GarageArea: Size of garage in square feet
* GarageQual: Garage quality
* GarageCond: 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
* Fence: Fence quality
* MiscFeature: Miscellaneous feature not covered in other categories
* MiscVal: $Value of miscellaneous feature
* MoSold: Month Sold
* YrSold: Year Sold
* SaleType: Type of sale
* SaleCondition: Condition of sale

**3. Data** **Summary**:

Number of records: 1460

Number of features: 79

**Data Types:**

1. Categorical Features - 43
2. Numerical Features- 36

**Null Values:**

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Null Values** | **Type** |
| LotFrontage | 259 | Category |
| Alley | 1369 | Category |
| MasVnrType | 873 | Category |
| MasVnrArea | 869 | Numerical |
| BsmtQual | 37 | Category |
| BsmtCond | 37 | Category |
| BsmtExposure | 38 | Category |
| BsmtFinType1 | 37 | Category |
| BsmtFinSF1 | 467 | Numerical |
| BsmtFinType2 | 38 | Category |
| BsmtFinSF2 | 1293 | Numerical |
| BsmtUnfSF | 118 | Numerical |
| TotalBsmtSF | 37 | Numerical |
| Electrical | 1 | Category |
| 2ndFlrSF | 829 | Numerical |
| LowQualFinSF | 1434 | Numerical |
| BsmtFullBath | 856 | Numerical |
| BsmtHalfBath | 1378 | Numerical |
| FullBath | 9 | Numerical |
| HalfBath | 913 | Numerical |
| BedroomAbvGr | 6 | Numerical |
| KitchenAbvGr | 1 | Numerical |
| Fireplaces | 690 | Numerical |
| FireplaceQu | 690 | Category |
| GarageType | 81 | Category |
| GarageYrBlt | 81 | Category |
| GarageFinish | 81 | Category |
| GarageCars | 81 | Numerical |
| GarageArea | 81 | Numerical |
| GarageQual | 81 | Category |
| GarageCond | 81 | Category |
| WoodDeckSF | 761 | Numerical |
| OpenPorchSF | 656 | Numerical |
| EnclosedPorch | 1252 | Numerical |
| 3SsnPorch | 1436 | Numerical |
| ScreenPorch | 1344 | Numerical |
| PoolArea | 1453 | Numerical |
| PoolQC | 1453 | Category |
| Fence | 1179 | Category |
| MiscFeature | 1406 | Category |
| MiscVal | 1408 | Numerical |

**4. Handling Missing Values:**

1. Identified missing values in various columns such as LotFrontage, Alley, FireplaceQu, etc.
2. Strategies used: Mean/median imputation for numerical values, mode imputation for categorical values, and marking missing categories.

**5. Data Cleaning and Preprocessing:**

* 1. Converted categorical variables into numerical representations using label encoding.
  2. Removed outliers using statistical methods.

**6. Exploratory Data Analysis:**

**Correlation Analysis**

* Identified strong correlations between SalePrice and features like OverallQual, GrLivArea, and TotalBsmtSF.

**Data Visualization**

* Distribution of Saleprice (Histogram, Boxplot)
* Feature relationships (Scatter plots, Heatmaps)
* Categorical feature analysis using bar charts

**7. Feature Engineering**

* Created new features such as TotalSF (sum of basement, first, and second floor area).
* Transformed skewed features using log transformation.
* Selected important features using feature selection techniques.

**8. Modeling & Prediction**

**Model Selection**

* Implemented various models:
  + Linear Regression
  + Random Forest
  + Gradient Boosting (XGBoost)

**Model Evaluation**

* Performance measured using:
  + Root Mean Squared Error (RMSE)
  + Mean Squared Error (MSE)
  + R-squared Score
* Best performing model: XG Boosting model shows the highest accuracy point.

**9. Model Comparison Report**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **RMSE** | **MSE** | **R-squared** |
| Linear Regression | 24574.732692872516 | 603917486.9261372 | 0.8770851987244149 |
| Random Forest | 25714.541099164937 | 661237623.9406427 | 0.8654188810522974 |
| Gradient Boosting | 23169.099570050355 | 536807174.88690764 | 0.8907441022110982 |

* **Best Model Recommendation**: Gradient Boosting

**10. Challenges Faced and Solutions**

**Data Quality Issues**

* Missing values were handled using imputation methods.
* Outliers were removed using statistical methods.

**Model Performance**

* Hyperparameter tuning was performed to improve accuracy.
* Feature selection was used to avoid overfitting.

**11. Conclusion**

* Key insights on factors influencing house prices.
* Suggestions for home buyers based on location, price, and features.
* Future: Using deep learning model to predict the house price more accurate.