**Summarized Report on House Price Prediction Model**

**Built And Tuned**

**By**

**Jibril Yahaya Jibril**

August 2024

**Table of Contents**

1. One – Introduction
2. Two – Datasets Information
3. Three – Preliminary Checking
4. Four – Data Cleaning and Preprocessing
5. Five – Feature Selection, Engineering and Model Building
6. Six – Prediction and Competition Submission

*Links*

*Reference*

**One –** **Introduction**

**Objective:**

The primary goal of this project is to develop a predictive model using multiple linear regression to estimate housing prices based on various features and attributes of the properties.

**Importance:**

Predicting housing prices is a crucial task in the real estate industry as it helps buyers, sellers, and real estate agents make informed decisions. By leveraging data science techniques such as multiple linear regression, we can identify the key factors influencing housing prices and provide valuable insights for pricing strategies, investment decisions, and market trends.

**Background:**

Multiple linear regression is a statistical method used to model the relationship between a dependent variable (housing prices) and multiple independent variables (housing features). By fitting a regression model to the data, we can estimate the coefficients of each feature and understand how they contribute to the overall variability in housing prices.

**Two –** **Datasets Information**

**Source:**  
 The Ames Housing Dataset dataset is publicly available on Kaggle: "House Prices: Advanced Regression Techniques"

**Dataset Structure:**

* **Training Set:** 1460 observations (rows) x 81 features (columns)
* **Testing Set:** 1459 observations (rows) x 81 features (columns)

**Variables:**

1. **Target Variable:**
   * SalePrice: Continuous variable, representing the sale price of each house
2. **Feature Variables:**
   * **Nominal Variables** (43):
     + Categorical features like neighborhood, building type, roof style, etc.
   * **Ordinal Variables** (7):
     + Ranked features like overall quality, condition, etc.
   * **Continuous Variables** (31):
     + Numerical features like area, number of rooms, etc.

**Additional Files:**

* data\_description.txt: Detailed description of each feature
* submission.csv: Sample submission file for testing predictions in the Kaggle housing price prediction competition

**Three - Preliminary Checking**

* Some potential anomalies in certain rows of the dataset may cause the column data type to become an 'object'. This may lead to an error in distinguishing between numerical and categorical columns. This was checked efficiently.
* It is possible that there are numerical columns that have data in the form of discrete, and limited number of values. Such columns may also be interpreted as categorical data.
* Since I’m using Multiple Linear Regression, I log-transformed the target column (SalePrice) to minimize the skewness and improve prediction accuracy.
* Based on a first viewing of the scatter plots against SalePrice, there appears to be:
  + A few outliers on the LotFrontage (say, >200) and LotArea (>100000) data.
  + BsmtFinSF1 (>4000) and TotalBsmtSF (>6000)
  + 1stFlrSF (>4000)
  + GrLivArea (>4000 AND SalePrice <300000)
  + LowQualFinSF (>550)
* With reference to the target SalePrice, the top 15 correlated attributes are:

OverallQual

GrLivArea

GarageCars

GarageArea

TotalBsmtSF

1stFlrSF

FullBath

TotRmsAbvGrd

YearBuilt

YearRemodAdd

GarageYrBlt

MasVnrArea

Fireplaces

BsmtFinSF1

* Missing Values:

Missing values were addressed efficiently such as the

1. LotFrontage (Median Imputation)
2. GarageYrBlt which is highly correlated with YearBuilt, so as an after-note, it is discarded before the machine learning step. Hence no action required.
3. MasVnrArea has 8 missing values, the same number as missing MasVnrType values. Likely not to have masonry veneer. Hence, fill with 0.
4. I assume that PoolQC to Bsmt attributes are missing as the houses do not have them (pools, basements, etc.). Hence, the missing values were filled in with "None".

**Four - Data Cleaning and Preprocessing**

* Missing values were addressed as stated above.
* Outliers were addressed based on observations from box plots and scatterplot visualized therein in the dataset.
* Some variables with high correlation were removed to address multicollinearity.
* Categorical data was encoded using One-Hot Encoding
* The Multiple Linear Regression model was initialized, trained and evaluated using Root Mean Squared Error.

**Five - Feature Selection, Engineering and Model Building**

* The test data was preprocessed using the same methods applied to the train data.
* The test data was used for prediction and the result was stored to a CSV file for submission.
* The output file was submitted to the competition board and the result was a 15591.77662 public score.

***Links***

[https://www.kaggle.com/code/jibrilyahayajibril/house-price-prediction-model-using-mlr#Addressing-outliers](https://www.kaggle.com/code/jibrilyahayajibril/house-price-prediction-model-using-mlr" \l "Addressing-outliers)

***Reference***

https://www.kaggle.com/code/cheesu/house-prices-1st-approach-to-data-science-process