**Home LLC Assessment:**

**Problem Statement:**

Q1. Find publicly available data for key \*supply-demand\* factors that influence US home prices \*nationally\*. Then, build a data science model that explains how these factors impacted home prices over the last 20 years.

**Assessment Documentation: Factors Influencing US Home Prices**

**Prepared by: Tamilarasu P**

**Date: 6/9/2023**

**Table of Contents**

1. **Introduction**

**- Overview of the Problem Statement:**

The problem statement revolves around understanding the key supply-demand factors that influence US home prices nationally over the past 20 years. The main focus is on identifying and analyzing various economic, demographic, and market-related factors that play a significant role in shaping the housing market. The objective is to gain insights into the dynamics of the US real estate market and determine which factors have the most substantial impact on home prices.

**2. Data Collection and Preprocessing**

**- Data Sources**

**- Publicly Available Datasets**

By considering the problem statement I have decided to find and get some features that will affect the US home prices nationally. The following are the some of the features which affects the US home prices.

1) US GDP Growth 2) US Inflation 3) Crude Oil Prices 4) Property Taxes 5) Consumer Confidence 6) Construction material cost 7) Price to rent ratio 8) US Unemployment rate 9) US GDP per capita 10) US Economic growth 11) Housing Market Indices 12) National Home Ownership Rate 13) Population growth 14) Poverty Rate

So, after deciding the features I have downloaded the dataset of those feature from the following websites

1) <https://fred.stlouisfed.org/series/CUUR0000SEHA> 2) <https://data.world/finance/home-construction-price-index> 3) <https://www.nar.realtor/research-and-statistics> 4) ttps://[www.census.gov/construction/nrc/](http://www.census.gov/construction/nrc/) 5) <https://www.bea.gov/> 6) <https://www.bls.gov/> 7) <https://www.census.gov/construction/nrc/> 8) <https://www.zillow.com/research/data/>

**- Data Collection Process**

**- Data Gathering Methods**

Collected the historical dataset in the above-mentioned links.

**- Data Preprocessing Steps**

- Data Cleaning

- Handling Missing Values

- Data Integration and Transformation

**3. Exploratory Data Analysis (EDA)**

- Visualizations

- Scatter Plots

- Pair Plots

- Correlation Heatmap

- Key Insights

- Positive/Negative Correlations

- Distribution of Key Variables

- Used Lasso Regularization to address multicollinearity

**4. Feature Engineering**

- New Features

- Derived from Existing Features

- Feature Selection

- Identifying Relevant Features

**5. Model Selection**

- Chosen Algorithms

- Random Forest

- XGBoost

- Rejected Models (if any)

- Linear Regression

**6. Hyperparameter Tuning**

- Random Forest Hyperparameters

- Grid Search

- XGBoost Hyperparameters

- Randomized Search

**7. Model Evaluation**

- Evaluation Metrics

- Root Mean Square Error (RMSE)

- R-squared (R²)

- Training and Testing Performance

- Random Forest: RMSE = 0.836, R² = 0.99683

- XGBoost: RMSE = 1.374, R² = 0.99499

- Model Comparison

- Random Forest vs. XGBoost

**8. Cross-Validation**

- Cross-Validation Technique

- K-Fold Cross-Validation (5 folds)

- Cross-Validation Results

- Random Forest Mean RMSE: 2.0837

- XGBoost Mean RMSE: 2.50714

- Cross-Validation Insights

**9. Feature Importance**

- Feature Importance Visualizations

- Random Forest Feature Importance Plot

- XGBoost Feature Importance Plot

- Influential Features

- GDP Growth Rate, Unemployment Rate

**10. Conclusion**

- Main Findings

- Both Random Forest and XGBoost models perform exceptionally well.

- Key factors influencing US home prices include GDP growth rate and unemployment rate.

- Model Performance Summary

- Random Forest outperforms XGBoost with lower RMSE.