Home.LLC Data Science Assessment Documentation

Introduction

This documentation provides an overview of the data analysis and modeling process undertaken for the Home.LLC Data Science Assessment. The primary goal of this assessment was to analyze factors influencing home prices in the United States over the last 20 years and build predictive models to explain these influences.

Data Sources

The data for this analysis was collected from various sources, including publicly available datasets and websites. The key data sources used in this assessment are as follows:

1. **S&P Case-Schiller Home Price Index**: Link
    - This dataset serves as a proxy for home prices in the United States and was used as the dependent variable in the analysis.
2. **GDP Growth (Annual %)**:
    - Data was sourced from Link.
3. **Unemployment Rates (Annual %)**:
    - Data was sourced from Link.
4. **Inflation (Annual %)**:
    - Data was sourced from Link.
5. **Interest Rates (Annual %)**:
    - Data was sourced from Link.
6. **Population Growth (Annual %)**:
    - Data was sourced from Link.

Data Preprocessing

The collected data underwent preprocessing to ensure its suitability for analysis. This preprocessing included the following steps:

1. **Data Cleaning**: Identifying and handling missing or inconsistent data points to ensure data quality.
2. **Data Integration**: Merging data from different sources to create a comprehensive dataset for analysis.
3. **Feature Engineering**: Creating additional features, such as calculating the average home price index over the years.

4. **Data Splitting**: Separating the data into training and testing sets for model development and evaluation.

Modeling Approaches

Three modeling approaches were employed to analyze the factors influencing home prices:

1. **Multiple Regression**: A linear regression model was built to understand the linear relationships between GDP growth, unemployment rates, inflation, interest rates, population growth, and the average home price index.
2. **Polynomial Regression**: A polynomial regression model was implemented to account for potential non-linear relationships between the independent variables and the dependent variable.
3. **Decision Tree**: A decision tree model was used to capture complex and non-linear relationships between the factors and home prices.

Model Evaluation

The performance of the models was evaluated using the R-squared value, which measures the proportion of the variance in the dependent variable that is predictable from the independent variables. The following R-squared values were obtained for the respective models:

- Multiple Regression: 0.7871
- Polynomial Regression: 0.9579
- Decision Tree: 0.9463

Enhancements and Insights

Achieving high R-squared values, particularly in the range of 0.94 to 0.96, is a strong indication the models are effectively explaining the variations in home prices.

The R-squared values for both the decision tree and polynomial regression models demonstrate that they are now providing a highly accurate representation of the relationships between the included factors and home prices.

Both the decision tree and polynomial regression models have performed impressively.

Conclusion

In conclusion, the analysis provided valuable insights into the factors influencing home prices in the United States. The models, particularly the decision tree and polynomial regression models, exhibited strong predictive power and can be used to understand the dynamics of the housing market.

Code and Outputs

- The code for the analysis is available in a Jupyter Notebook, which can be accessed on GitHub.
- Data visualizations and reports were created using Power BI. Visualizations and summary reports are available for reference in the GitHub repository.

Submission

The final submission for this assessment includes:

- A link to the GitHub repository containing the Jupyter Notebook and Power BI reports.