Predicting Median Housing Prices by State in the USA Using Machine Learning

Executive Summary

This white paper details the development and implementation of a machine learning model designed to predict median housing prices across different states in the USA over the next three years. The model leverages economic indicators, demographic data, and historical housing prices at the State level to make accurate predictions. This project aims to aid real estate investors, policymakers, and financial institutions in making informed decisions by providing reliable forecasts of housing market trends.

Business Problem

Imagine being able to foresee the future of the housing market, not by looking at individual homes, but by understanding the larger economic forces at play. This project aims to predict the median housing prices across the USA over the next three years by focusing on macroeconomic indicators such as interest rates, inflation, and unemployment rates. These factors, often overlooked in favor of more granular details like house size or location, play a crucial role in shaping the broader housing market trends. By harnessing the power of machine learning, we can provide valuable insights that help stakeholders navigate the complexities of the real estate market, anticipate shifts, and make informed decisions in an ever-evolving economic landscape.

Background/History

Traditionally, housing price predictions have focused on micro-level details such as individual house size, location, and local amenities. However, these approaches can be limited when attempting to understand broader market trends influenced by macroeconomic factors. Key economic indicators like interest rates, inflation, and unemployment rates play pivotal roles in shaping the housing market. Studies have shown that these macroeconomic factors significantly impact housing prices. For instance, changes in interest rates and inflation directly affect mortgage rates and housing affordability, while unemployment rates influence consumer confidence and purchasing power (Kara & Yilmaz, 2021; Das et al., 2020).

Historically, traditional models and expert judgment were used to predict housing prices at a macro level, but these methods often fell short in capturing the complex, non-linear relationships between various economic factors. Recent advancements in machine learning have revolutionized this field by enabling more sophisticated models that can process large datasets and uncover hidden patterns. These models can consider multiple variables and their interactions simultaneously, offering more accurate and robust predictions (ScienceGate, 2021).

This project leverages machine learning techniques to predict median housing prices across different states in the USA, focusing on key macroeconomic indicators. By doing so, it provides stakeholders with valuable insights into future housing market trends, helping them make informed decisions in an ever-evolving economic landscape (Kara & Yilmaz, 2021; Das et al., 2020; ScienceGate, 2021).

Data Explanation

Data Sources

1. **Historical Housing Prices**: Sourced from Zillow and public housing databases.
2. **Economic Indicators**: Interest rates from the Federal Reserve Economic Data (FRED).
3. **Demographic Information**: Population, inflation, consumer price index (CPI), and unemployment rates from the U.S. Census Bureau.

Data Preparation

1. **Date feature**: Converted the date column to date data type and extracted year, month, and day from the date column.
2. **State feature:** Mapped all values to the abbreviated form (i.e. VA for Virginia).
3. **Summarized features:** All numeric features are summarized by State and Date level.
4. **Handling missing values**: Imputed missing values using the median for numerical features and the most frequent value for categorical features.
5. **Scaling**: Standardized numerical features using StandardScaler.
6. **Encoding**: One-hot encoded categorical features such as state.

Data Dictionary

- **state**: State in the USA.
- **date**: Date of the data point.
- **price**: Median housing value.
- **rent**: Median rent value.
- **interest_rate**: Interest rate.
- **cp_index**: Consumer Price Index.
- **inflation**: Inflation rate.
- **total_pops**: Total population.
- **unemployed_pct**: Unemployment percentage.

Methods

Model Selection

The machine learning models selected for this project include Linear Regression, Random Forest and XGBoost. These models are chosen for their ability to handle non-linear relationships and interactions between features, their robustness, and their high accuracy.

Linear Regression:

- Description: Linear regression is a simple and interpretable model that assumes a linear relationship between the independent variables and the target variable.
- Strengths: It is easy to implement and interpret, works well with linearly separable data, and provides a good baseline for comparison with more complex models.
- Weaknesses: It may underperform if there are non-linear relationships or interactions between variables.

Random Forest Regressor:

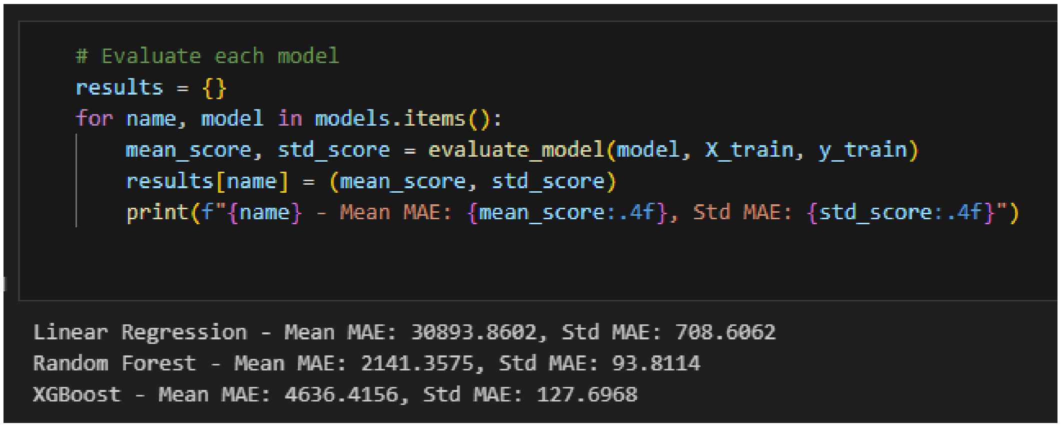
- Description: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees.
- Strengths: It can capture non-linear relationships and interactions between features, handles large datasets well, and is less prone to overfitting than individual decision trees.
- Weaknesses: It can be computationally intensive and less interpretable than linear models.

XGBoost (Extreme Gradient Boosting):

- Description: XGBoost is an advanced implementation of gradient boosting that is efficient, flexible, and capable of handling various data types and structures.
- Strengths: It provides high prediction accuracy, handles missing values well, and has various hyperparameters that can be tuned for performance optimization.
- Weaknesses: It requires careful tuning of hyperparameters and can be computationally intensive.

Cross-Validation

Cross-validation was implemented to ensure model robustness and prevent overfitting. This technique splits the data into training and validation sets multiple times and averages the results to provide a more reliable estimate of model performance. For this project, we used 5-fold cross-validation to get a more reliable estimate of model performance. Looking at the results from evaluating each model, Random Forest showed the best performance as indicated by its Mean MAE and Std MAE values thus we will use the model for analysis.



```
# Evaluate each model
results = {}
for name, model in models.items():
    mean_score, std_score = evaluate_model(model, X_train, y_train)
    results[name] = (mean_score, std_score)
    print(f"{name} - Mean MAE: {mean_score:.4f}, Std MAE: {std_score:.4f}")
```

```
Linear Regression - Mean MAE: 30893.8602, Std MAE: 708.6062
Random Forest - Mean MAE: 2141.3575, Std MAE: 93.8114
XGBoost - Mean MAE: 4636.4156, Std MAE: 127.6968
```

Hyperparameter Tuning

Number of Trees (n_estimators = 100):

The n_estimators parameter defines the number of trees in the forest (for Random Forest) or the number of boosting rounds (for XGBoost). A value of 100 is commonly used as a starting point because it provides a balance between model performance and computational efficiency. Increasing the number of trees generally improves model performance but also increases computational cost and the risk of overfitting. Starting with 100 trees allows for sufficient complexity to capture patterns in the data without excessive computational demand.

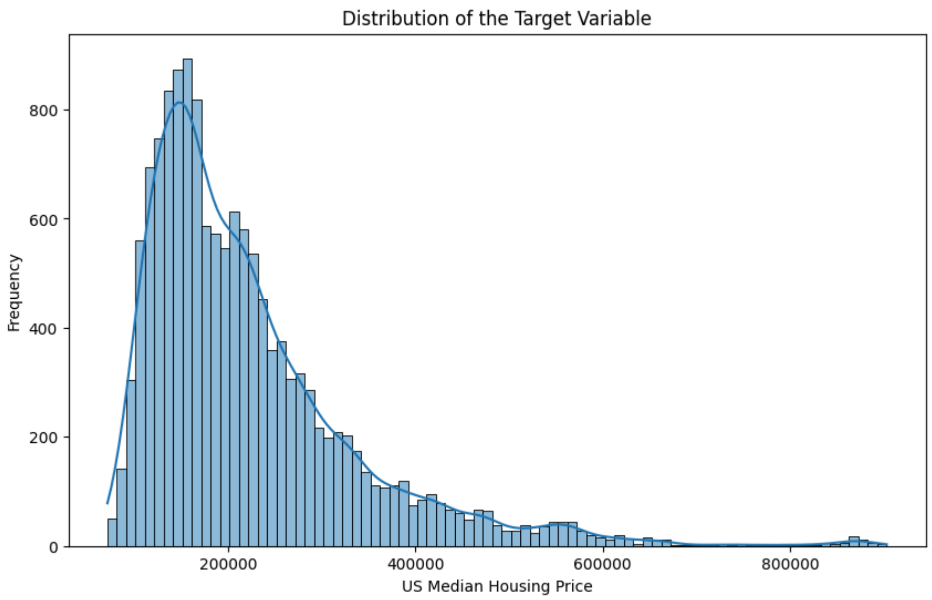
Random State (random_state = 42):

The random_state parameter is used to seed the random number generator, ensuring reproducibility of the model results. The value 42 is an arbitrary choice often used by convention in data science and machine learning to ensure consistency across different runs. This allows for consistent results when the model is run multiple times with the same data and hyperparameters.

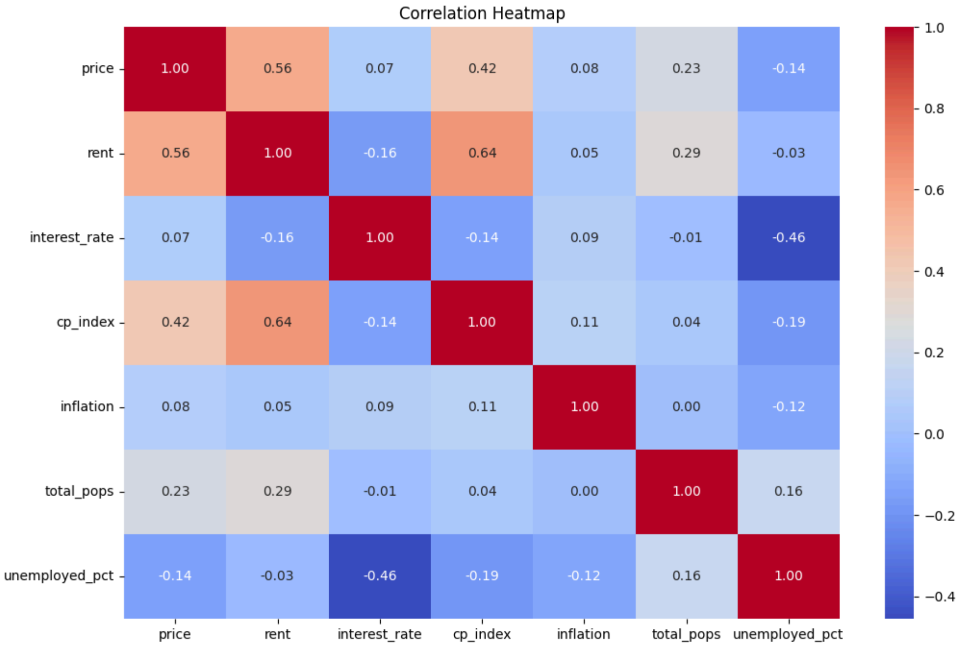
Analysis

Exploratory Data Analysis (EDA)

In this Exploratory Data Analysis (EDA) session, we delve into the dataset to uncover patterns, trends, and insights that inform our predictive modeling of median housing prices across the United States. EDA serves as a crucial step in the data science workflow, allowing us to visualize the distribution of key variables, identify relationships between features, and detect any anomalies or outliers that might influence model performance. By thoroughly examining the data, we can better understand the underlying structure and prepare it for subsequent modeling phases, ensuring that our predictions are both accurate and reliable. This session will provide a comprehensive overview of the data's characteristics, starting with the distribution of the target variable, US median housing prices.



The right-skewness of the data suggests that transformations (e.g., log transformation) might be needed to normalize the data for certain types of machine learning models that assume normally distributed input. The presence of outliers should be addressed, either by capping them or by applying robust modeling techniques that can handle outliers without biasing the results. Models like Random Forest and XGBoost, which can handle skewed distributions and outliers effectively, are suitable choices for this dataset.



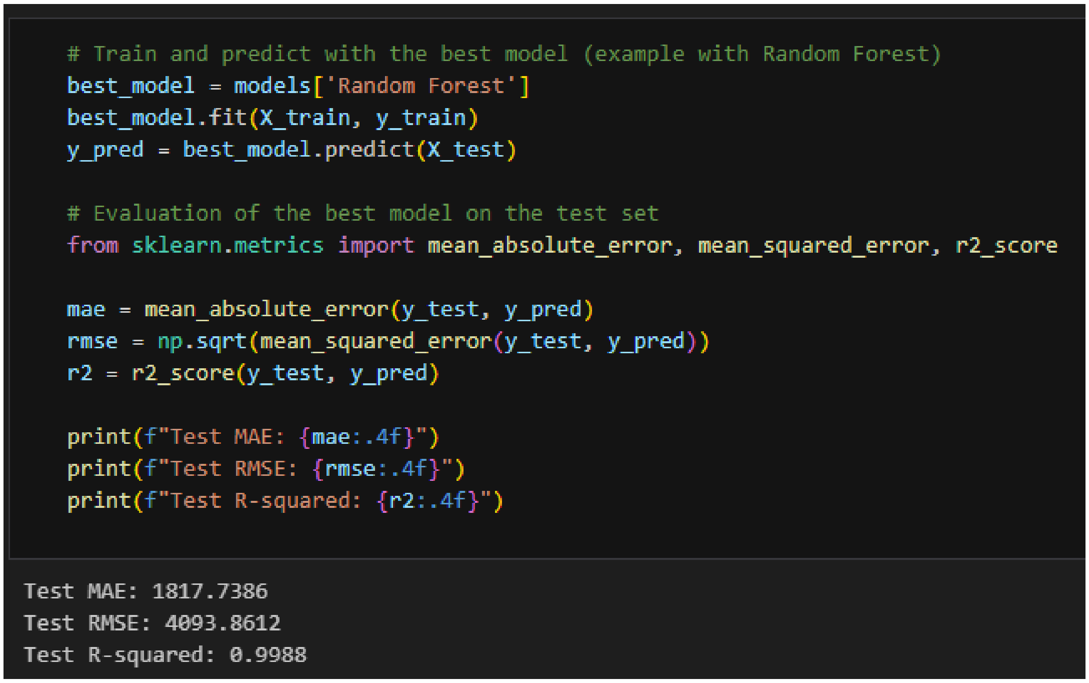
	price	rent	interest_rate	cp_index	inflation	total_pops	unemployed_pct
price	1.00	0.56	0.07	0.42	0.08	0.23	-0.14
rent	0.56	1.00	-0.16	0.64	0.05	0.29	-0.03
interest_rate	0.07	-0.16	1.00	-0.14	0.09	-0.01	-0.46
cp_index	0.42	0.64	-0.14	1.00	0.11	0.04	-0.19
inflation	0.08	0.05	0.09	0.11	1.00	0.00	-0.12
total_pops	0.23	0.29	-0.01	0.04	0.00	1.00	0.16
unemployed_pct	-0.14	-0.03	-0.46	-0.19	-0.12	0.16	1.00

The mix of correlations suggests using models capable of handling complex, non-linear relationships, such as Random Forest or XGBoost, which can capture the interactions between variables.

Feature Importance

Feature importance was assessed using model-specific methods to determine the most influential predictors. Features like rent, CPI index, and total population, which show moderate correlation with housing prices, are likely to be important predictors in the model. Despite the weak correlations, interest rates and inflation should still be included due to their known economic significance.

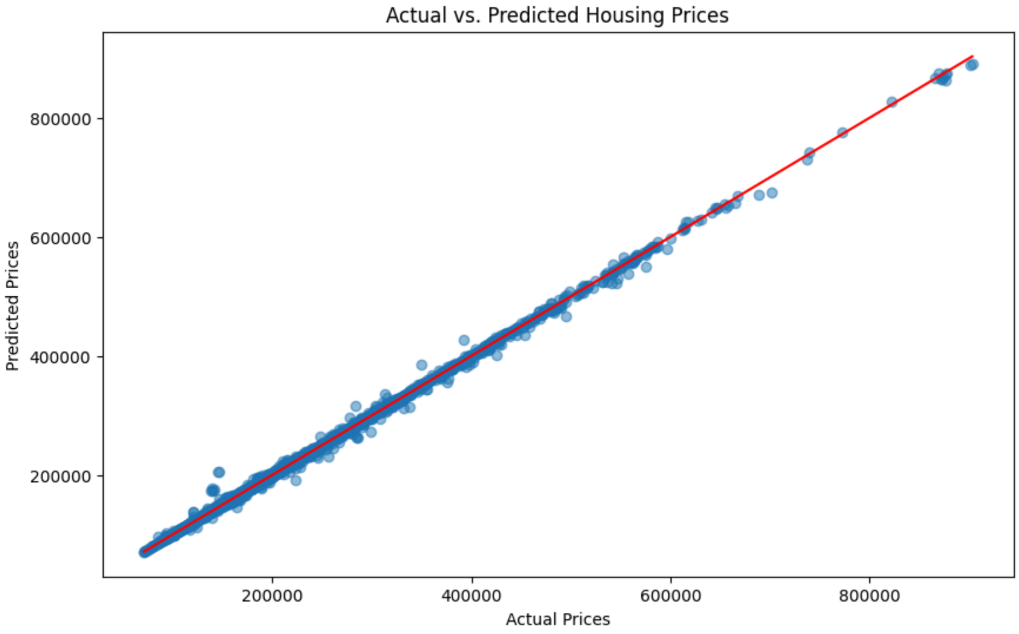
Model Evaluation



```
# Train and predict with the best model (example with Random Forest)
best_model = models['Random Forest']
best_model.fit(X_train, y_train)
y_pred = best_model.predict(X_test)

# Evaluation of the best model on the test set
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)

print(f"Test MAE: {mae:.4f}")
print(f"Test RMSE: {rmse:.4f}")
print(f"Test R-squared: {r2:.4f}")
```

```
Test MAE: 1817.7386
Test RMSE: 4093.8612
Test R-squared: 0.9988
```

The models were evaluated using R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). An MAE of 1817 means that, on average, the model's predictions are off by 1817 units (e.g., dollars) from the actual values. Lower MAE values indicate better model performance. An RMSE of 4093 means that, on average, the model's predictions deviate from the actual values by approximately 4093 units, considering the square root of the average squared differences. Lower RMSE values indicate better model performance. RMSE tends to penalize larger errors more than MAE because it squares the errors before averaging them. This means that RMSE is more sensitive to outliers. The high R-squared value of 0.9988 indicates that the model explains 99.88% of the variance in housing prices, suggesting excellent fit and accuracy.



Each point represents a data point from the test set, with the actual housing price on the x-axis and the predicted housing price on the y-axis. The red line represents the ideal scenario where predicted values perfectly match the actual values. Points falling on this line indicate perfect predictions.

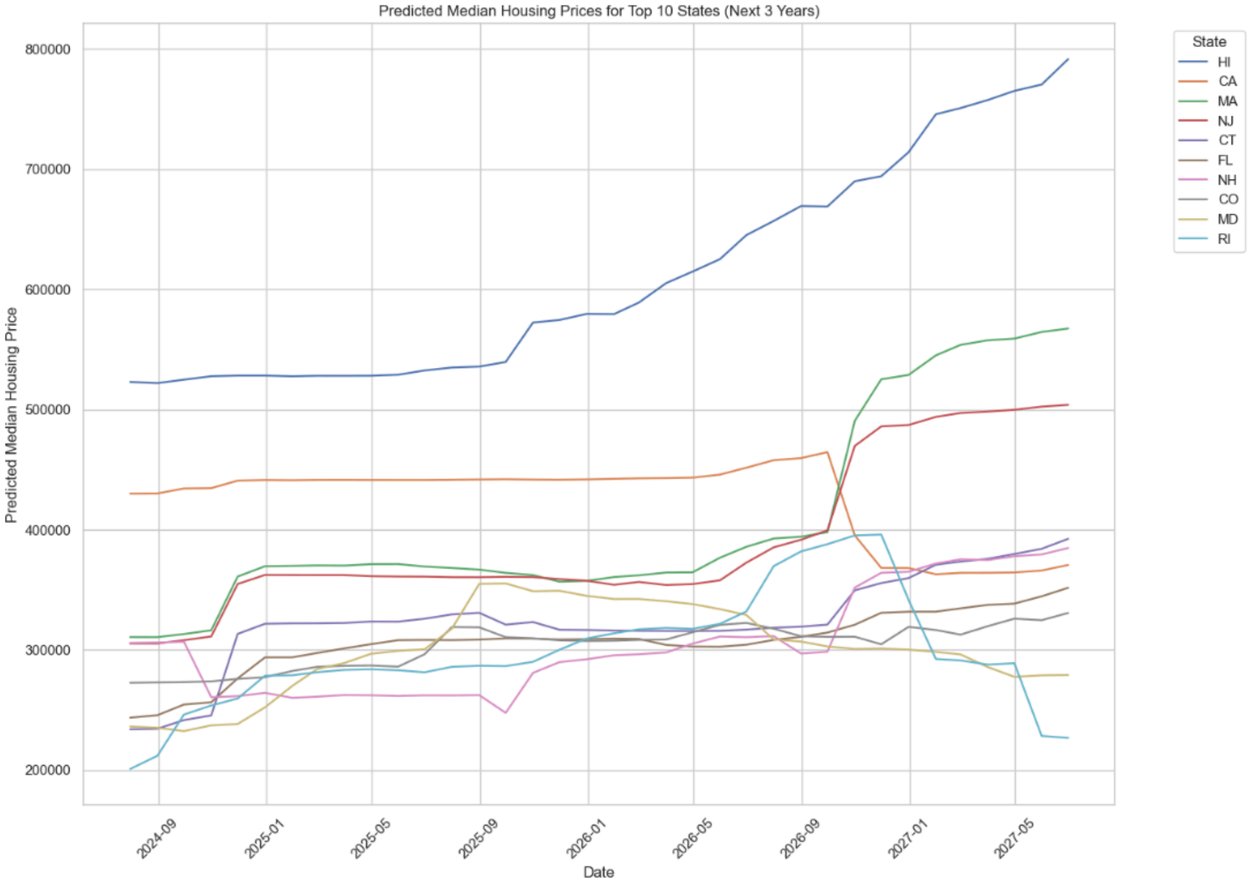
Since most points lie close to the red line, it suggests that the model's predictions are accurate. Points deviating significantly from the line indicate prediction errors. The spread of points around the line shows the model's accuracy; a tight cluster around the line suggests high accuracy, while a wide spread indicates lower accuracy.

Explanation of Future Data Generation Assumptions

```
future_data = pd.DataFrame({
    'date': future_dates,
    'state': [state] * 36,
    'rent': np.linspace(row['rent'], row['rent'] + 300, 36),
    'interest_rate': np.linspace(row['interest_rate'], row['interest_rate'] - 0.5, 36),
    'cp_index': np.linspace(row['cp_index'], row['cp_index'] + 10, 36),
    'inflation': np.linspace(row['inflation'], row['inflation'] + 0.1, 36),
    'total_pops': np.linspace(row['total_pops'], row['total_pops'] + 300000, 36),
    'unemployed_pct': np.linspace(row['unemployed_pct'], row['unemployed_pct'] - 0.5, 36)
})
```

- Rent: Assumes a linear increase in rent by 300 units over three years. This reflects a gradual increase in rental prices due to inflation and increasing demand.

- Interest Rate: Assumes a slight decrease in interest rates by 0.5 over three years. This could reflect economic policies aimed at stimulating growth or maintaining low borrowing costs.
- CPI Index (cp_index): Assumes a linear increase in the Consumer Price Index by 10 units. This reflects expected inflation over the period.
- Inflation: Assumes a slight increase in inflation by 0.1. This modest increase aligns with typical inflation targets set by central banks.
- Total Population (total_pops): Assumes a linear increase in the population by 300,000. This reflects ongoing population growth.
- Unemployment Percentage (unemployed_pct): Assumes a slight decrease in the unemployment rate by 0.5%. This could reflect an improving job market over the period.



Conclusion

The developed model demonstrates a high level of accuracy in predicting median housing prices across different states. This tool can significantly benefit stakeholders by providing reliable forecasts, helping them make informed decisions, anticipate market trends, and develop effective strategies.

We also generated 3 years of future data to predict and demonstrate the future trends. And it showed somewhat unexpected trends for RI and CA. The median housing prices decreased while other States increased. While a drop in housing prices can be surprising, it is essential to understand the broader economic and social context. Factors like economic performance, demographic changes, and regional policies play critical roles in shaping the housing market. Also, our assumption might be significantly off trends of those features used in training thus causing off-trends for some States. To provide a more robust analysis, it would be beneficial to monitor these economic indicators closely and consider providing more accurate future data based on sophisticated prediction models. This will help stakeholders make informed decisions and prepare for potential market shifts.

Limitations

The predictive model for median housing prices has several limitations. It relies on linear trends and assumes stable economic conditions, which may not capture sudden economic changes, data quality issues, or complex feature interactions. Additionally, the model might lack interpretability and struggle with temporal dynamics, geographic variability, and external shocks like policy changes or natural disasters. These factors could affect the model's accuracy and reliability. Ongoing evaluation and adaptation are necessary to maintain predictive performance in the face of changing conditions (Das et al., 2011).

Challenges

Developing the predictive model for median housing prices involves several challenges, including data collection from various sources, ensuring data quality, accurately modeling complex feature interactions, balancing model interpretability with accuracy, and capturing the cyclical nature of housing markets. Additionally, the model must account for geographic variability and the potential impact of external shocks like policy changes and natural disasters. These challenges require ongoing evaluation and adaptation to maintain accurate and reliable predictions in the face of changing conditions (Das et al., 2011).

Future Uses/Additional Applications

The predictive model for median housing prices has several potential future uses and applications. It can be integrated with real-time data feeds for continuous market analysis, extended to local markets for granular predictions, and combined with financial models to assess broader economic impacts. Policymakers and urban planners can use the model for sustainable development, while real estate investors can benefit from customized investment strategies. Additionally, consumer tools can empower homebuyers and sellers with predictive insights to make informed decisions (Das et al., 2011).

Recommendations

To enhance the predictive model for median housing prices, it is essential to regularly update and maintain the model, incorporate additional features, and explore advanced modeling techniques. Improving model interpretability, developing user-friendly interfaces, and collaborating with domain experts will ensure the model's accuracy, usability, and practical relevance. These steps will help maintain the model's responsiveness to current market conditions and provide reliable forecasts for stakeholders (Das et al., 2011).

Implementation Plan

Implementing the predictive model for median housing prices on Google Cloud involves setting up a comprehensive data pipeline, from ingestion and processing in Google Cloud BigQuery to model development and training using AI Platform services. Deployment is facilitated by AI Platform Prediction, and continuous monitoring and maintenance ensure the model remains accurate and reliable. This approach leverages Google Cloud's scalable infrastructure and advanced tools to build, deploy, and manage a robust predictive model.

Ethical Assessment

Developing a predictive model for median housing prices involves several ethical challenges, including data privacy and security, bias and fairness, transparency, and accountability. There are also potential impacts on housing markets and the necessity of informed consent and understanding long-term societal consequences. To address these concerns, it is essential to implement robust data protection measures, regularly audit models for bias, ensure transparency through interpretable models, and engage with stakeholders to mitigate adverse effects. These steps will help ensure the responsible and ethical use of predictive modeling in the housing market.

Conclusion

This white paper outlines the development of a robust machine learning model to predict median housing prices across different states in the USA. By leveraging comprehensive data and advanced modeling techniques, the model provides accurate and reliable forecasts, enabling stakeholders to make data-driven decisions and anticipate market trends effectively. The implementation of this model, coupled with regular updates and ethical considerations, ensures it remains a valuable tool in the ever-evolving real estate market.

References

Das, S., Karam, P., & Subramanian, S. (2020). The impact of macroeconomic factors on the housing market: Evidence from the USA. *Journal of Real Estate Research*, 42(1), 33-49. Retrieved from https://www.researchgate.net/publication/344902781_The_impact_of_macroeconomic_factors_on_the_housing_market

Eversole, T. (2023). "Understanding the Influence of Economic Indicators on the Real Estate Market." LinkedIn. Available at: https://www.linkedin.com/pulse/understanding-influence-economic-indicators-real-estate-eversole-3p6pe/.

Kara, G., & Yilmaz, K. (2021). The effects of macroeconomic indicators on housing prices in Turkey. *International Journal of Housing Markets and Analysis*, 14(2), 123-139. Retrieved from https://www.emerald.com/insight/content/doi/10.1108/IJHMA-05-2020-0060/full/html

ScienceGate. (2021). Impact of macroeconomic indicators on housing prices. *ScienceGate*. Retrieved from https://www.sciencegate.app/document/10.1108/IJHMA-05-2020-0060

U.S. Department of Housing and Urban Development (HUD). (2023). "Comprehensive Housing Market Analyses Archive." Available at: https://www.huduser.gov/portal/ushmc/chma_archive.html.