R.L DATA

For - REAL - Estate

Team: Hallie Bruno, Mateo Restrepo

Advisor: Jonathan Johnson

Introduction

The real estate market often lacks transparent and accessible pricing tools. Our project addresses this gap by developing a machine learning model and platform that predicts property values using physical and geographical features. Designed for users ranging from homebuyers to policymakers, the tool aims to democratize access to accurate pricing insights and reduce market inefficiencies through data-driven decision-making.

Aim and Objectives

Develop an accurate pricing model and a user-friendly tool to predict property prices using machine learning, enhancing transparency and supporting informed real estate decisions.

Method

- Collected and preprocessed real estate data (size, location, amenities, etc.).
- Engineered features, dealt with null values, and normalized inputs for consistency.
- Tested multiple regression models for baseline accuracy.
- Implemented a neural network for advanced comparison.
- Selected Gradient Boosting Regressor as top performer.
- Optimized model with grid search for hyperparameters.
- Analyzed feature importance to refine predictions.
- Developed a simple user interface to demonstrate predictions interactively.

Technologies

Python was used for model development with libraries such as Scikit-learn, TensorFlow, and Pandas. Google Colab served as the primary development environment. The frontend was built using React, and a SQLite database was implemented for backend data management. The dataset we used is from Kaggle.

Results

- Machine learning algorithm

Initial experiments with Linear Regression and Decision Trees provided baseline performance but lacked precision. A neural network improved pattern recognition but required extensive tuning. Gradient Boosting Regressor ultimately achieved the highest accuracy (86%) after hyperparameter optimization, making it the most effective model.

```
Model Name:  DecisionTreeRegressor
Score of Model: [0.6518832781799049]
-------------------------
Model Name:  GradientBoostingRegressor
Score of Model: [0.8673377295646235]
-------------------------
Model Name:  RandomForestRegressor
Score of Model: [0.8516466407384201]
-------------------------
Model Name:  XGBRegressor
Score of Model: [0.8647006248183509]
```

```
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - loss: 0.0444 - mae: 0.1588
Test MAE: 0.16
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
Test RMSE: 209001.73
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 0.0444 - mae: 0.1588
Test Loss (Mean Squared Error): [0.045726049691438675, 0.16158951818943024]
Test Loss (RMSE): [0.2138365  0.40198199]
10/10 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step
0.7577372035733991
```

- Feature importance

Analysis revealed location, square footage, and number of bedrooms as key predictors.

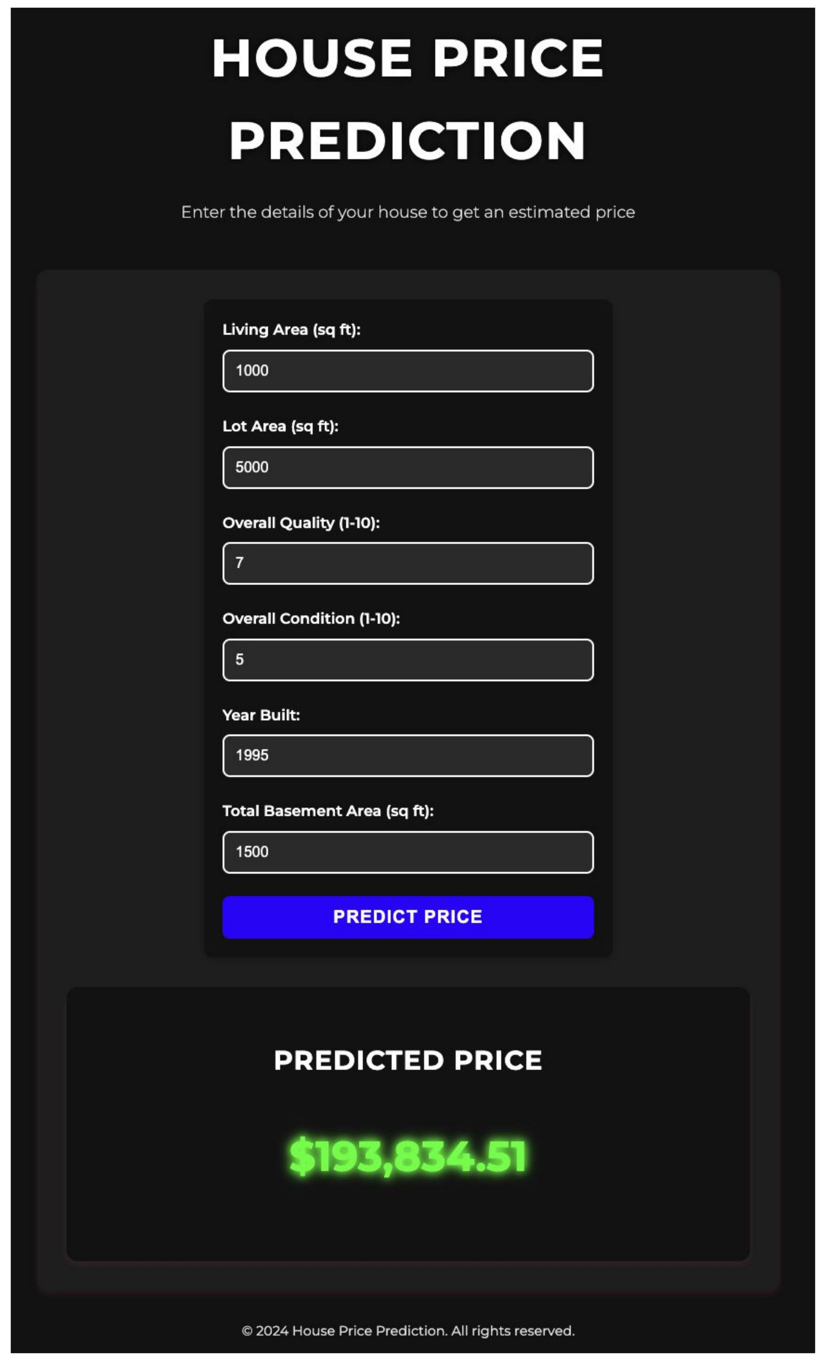
```
[ ] gbr = GradientBoostingRegressor(
        n_estimators=700,        # Number of boosting stages
        learning_rate=0.05,      # Lower values improve stability
        max_depth=5,             # Controls tree depth (complexity)
        min_samples_split=20,    # Minimum samples to split an internal node
        min_samples_leaf=2,      # Minimum samples required in a leaf node
        subsample=0.9,           # Fraction of samples used for training each tree
        max_features='log2',     # Number of features considered for best split
        random_state=42
    )

    gbr.fit(X_train, y_train)

GradientBoostingRegressor
GradientBoostingRegressor(learning_rate=0.05, max_depth=5, max_features='log2',
                          min_samples_leaf=2, min_samples_split=20,
                          n_estimators=700, random_state=42, subsample=0.9)
```

- Frontend

A user-friendly interface was developed using React, allowing users to input property details and receive price predictions. The design prioritizes simplicity and accessibility across a wide range of user types.



Significance

Our platform democratizes access to accurate real estate pricing by offering unbiased, data-driven predictions. It empowers buyers, sellers, investors, and policymakers to make informed decisions, helps track market trends, and promotes greater transparency and stability in the housing market.

Future

Future development includes integrating real-time, geographically detailed data via APIs (e.g., Zillow) to enhance prediction accuracy. Planned improvements also involve expanding platform features, increasing model scalability, and exploring the development of a mobile application for broader accessibility.

Conclusion

High prediction accuracy was achieved despite the absence of detailed location data, demonstrating the model's robustness. Future integration of geographic information is expected to further improve accuracy and extend the platform's usefulness in real estate decision-making.

Acknowledgments

We thank Professor Johnson for his invaluable guidance throughout the project, and the Trinity College Computer Science Department for their support and resources.