House Price Prediction

Prediction of House Prices using Linear Regression and Decision Tree

Team members : Haritha Jampani
Kiran Sri Sai Praneeth Kondreddi
Aye Chan Moe

Agenda

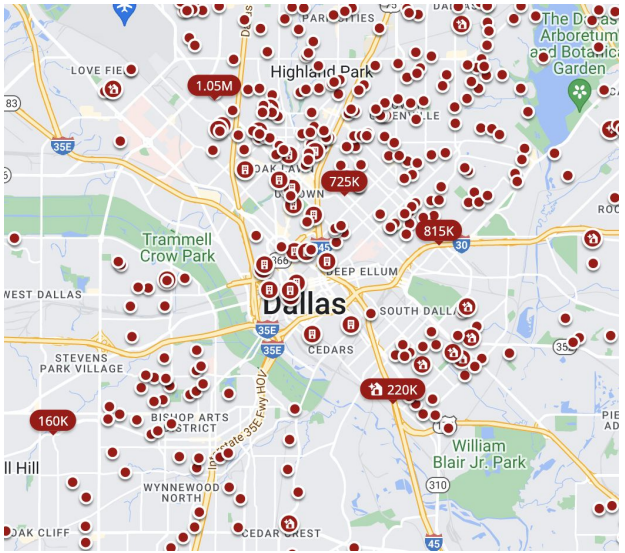
- ❑ Introduction
- ❑ Data Preparation
- ❑ Model 1 interpretation and results
- ❑ Model 2 interpretation and results
- ❑ Model 3 interpretation and results
- ❑ Comparisons
- ❑ Visualization
- ❑ Business Insights and Recommendation

Introduction

"The global Real Estate Market size was valued at USD 3.69 trillion in 2021 and is poised to grow from USD 3.88 trillion in 2022 to USD 6.13 trillion by 2030"

Objectives:

1. Predict house prices based on historical prices and variables that affect the value
2. Pinpoint factors that affect house prices
3. Identify trends and patterns in the housing markets across time



Data Preparation

- Converted ‘Type’ variable into dummy variables making the Multiple Occupancy as Referenced Variable
- Created a new dataset where we removed outliers which results in homes that are Single Family Type only
- Transform ‘Sale_date’ variable into sale month
- No further data cleaning process is needed since there are no missing values
- We also used 70/30 splits for all our models

Variable name	Description or possible values
Record	A modified ID for each house
Sale_amount	Sale price of the house in U.S. dollars
Sale_date	Sale date of the house
Beds	Number of bedrooms in the house
Baths	Number of bathrooms in the house
Sqft_home	Square footage of the house
Sqft_lot	Square footage of the lot
Type	Multiple Family Multiple Occupancy Single Family
Build_year	Year the house was built
Town	Name of the campus town
University	Name of the university

Model 1: Linear Regression

```
Residual standard error: 96660 on 6836 degrees of freedom
Multiple R-squared:  0.7709,    Adjusted R-squared:  0.7691
F-statistic: 418.3 on 55 and 6836 DF,  p-value: < 0.00000000000000022
```

- Built with a dataset with no outliers
- Works best on Single Family Type homes
- The predictor variables include the square footage of the home, the square footage of the lot, the number of bedrooms, the number of bathrooms, the year that the house was built, town variables, and the sale month
- Adjusted R-squared: 0.7691

Model 2: Linear Regression

```
Residual standard error: 174800 on 7405 degrees of freedom
Multiple R-squared:  0.7203,    Adjusted R-squared:  0.7181
F-statistic: 334.5 on 57 and 7405 DF,  p-value: < 0.00000000000000022
```

- Multicollinearity between ‘Town’ and ‘University’ variable
- Built with only Town variable which was converted as a dummy variable
- The other predictor variables include number of bedrooms, number of bathrooms, square footage of the house, square footage of the lot, type of the house, sale month, and year built.
- Adjusted R-squared: 0.7181

Model 3: Decision Tree

	ME	RMSE	MAE	MPE	MAPE
Test set	-2030.909	218398.3	98414.63	-18.89625	36.43376

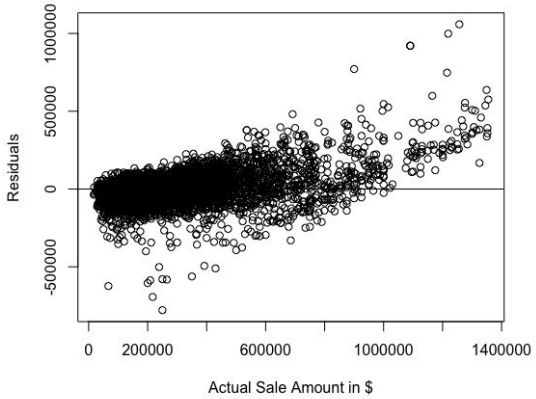
- Full tree has 7240 branches
- Best pruned-tree with 35 branches
- Has a CP value of 0.002174092
- Has predictor variables Town, Sqft_home, Baths, Build_year, Beds, Sqft_lot, Sale_date, Type, Sale_month.

Model Comparisons

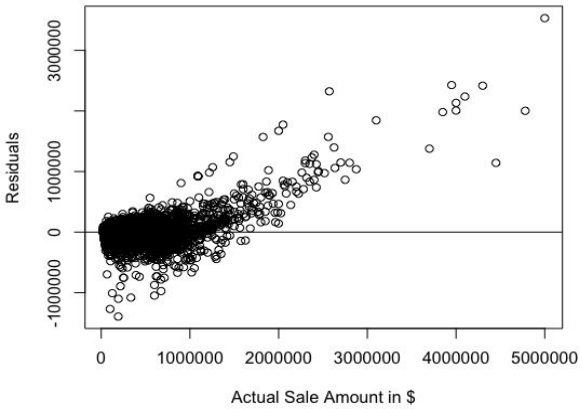
Model	ME	RMSE	MAE	MPE	MAPE
Linear Model 1	1193.2	105978.4	66261.6	-10.4	28.3
Linear Model 2	-6602.5	247984.7	95207	-8.9	34.6
Decision Tree	-2031	218398.3	98414.6	-18.9	36.4

Visualizations

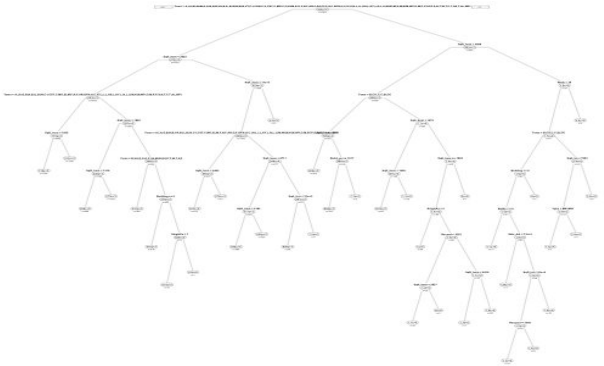
Model 1



Model 2



Model 3



Business Insights and Recommendations

- We recommend Model 1, if predicting house prices just for Single Family Homes.
- Between Model 2 and Model 3, we recommend using Model 3 which is the decision tree since it has lower RMSE
- Using a latest dataset would improve the model accuracy, and the predicted values will be less deviant from the actual ones which really helps to give recommendations.