



Group1 Consulting

See Through the Ambiguity: Predicting Home Prices

**Samuel Martin, Cullen Flynn,
De'Angelo Pillai, Gerardo Arvizu**

Table of Contents

Our Understanding	4
Data Exploration	5
Data Cleaning	6
Modeling	7
Evaluation	11
Next Steps	12

Meet Your Delivery Leads

Your dedicated project delivery leads have worked to create a customized solution tailored to your specific needs.



Samuel Martin
Analytics Researcher

Data Scientist with 5+ years experience developing Neural Networks and analytics pipelines.



Cullen Flynn
Data Scientist

Data Scientist with 5 years of expertise in leveraging data for business insights and problem-solving.



Gerado Yanez Arvizu
Analytics Lead

Data Scientist with 6 year of expertise in IT, analytics, and modeling.



De'Angelo Pillai
Data Scientist

Data Scientist with 5+ years experience developing and analyzing linear regression models.

House Price Prediction Algorithm

FlatTech Real Estate has approached Group1 Consulting asking for a home price prediction algorithm.

- Group1 Consulting is representing FlatTech Real Estate - an investment firm interested in residential, single-family homes in Austin.
- FlatTech seeks precise home price predictions using key attributes, bypassing the need for appraisers.
- Provide a competitive edge with Group1's in-depth analysis of which homes to invest in within the Austin area.



Data Exploration

Using Kaggle, our team was able to find a great dataset for our needs; "Austin TX, House Listings"

The dataset we needed:



Austin Housing Prices: the dataset needed to have prices of houses in it as one of the features for us to predict towards.



Variety of Features: multiple features of house listings such as number of beds, baths, etc. that primarily influence home prices.

The dataset we found:




Made for Austin Homes: the dataset contained over 15,000 rows, each of which were a home in the Austin TX area.



Numerous Features: not only did the dataset contain typical home features like bed and bath, but also school district rating, number of parking places, etc.




Very Clean: Eric Pierce the creator of the dataset scraped Zillow and used this for his college capstone project. The result was a highly clean dataset.


ERIC PIERCE · UPDATED 2 YEARS AGO
41
New Notebook
Download (3 GB)

Austin, TX House Listings

Features and Images scraped in January 2021

Data Card Code (5) Discussion (0)



About Dataset

Context

This data was originally sourced in support of my capstone project at Northwestern University. The Austin Housing market is one of the hottest markets in 2021, and these listings show how that market has changed over the past couple of years.

Content

This dataset includes a (relatively) clean set of features. The original uncleaned dataset consisted over over 700 columns, and can be downloaded if you select "version 1" instead of the latest version.

I also included the first image from each home listing on Zillow. I used this data to predict home price using images in addition to the features in the datafile.

Usability ⓘ

9.41

License

[GPL 2](#)

Expected update frequency

Never

Tags

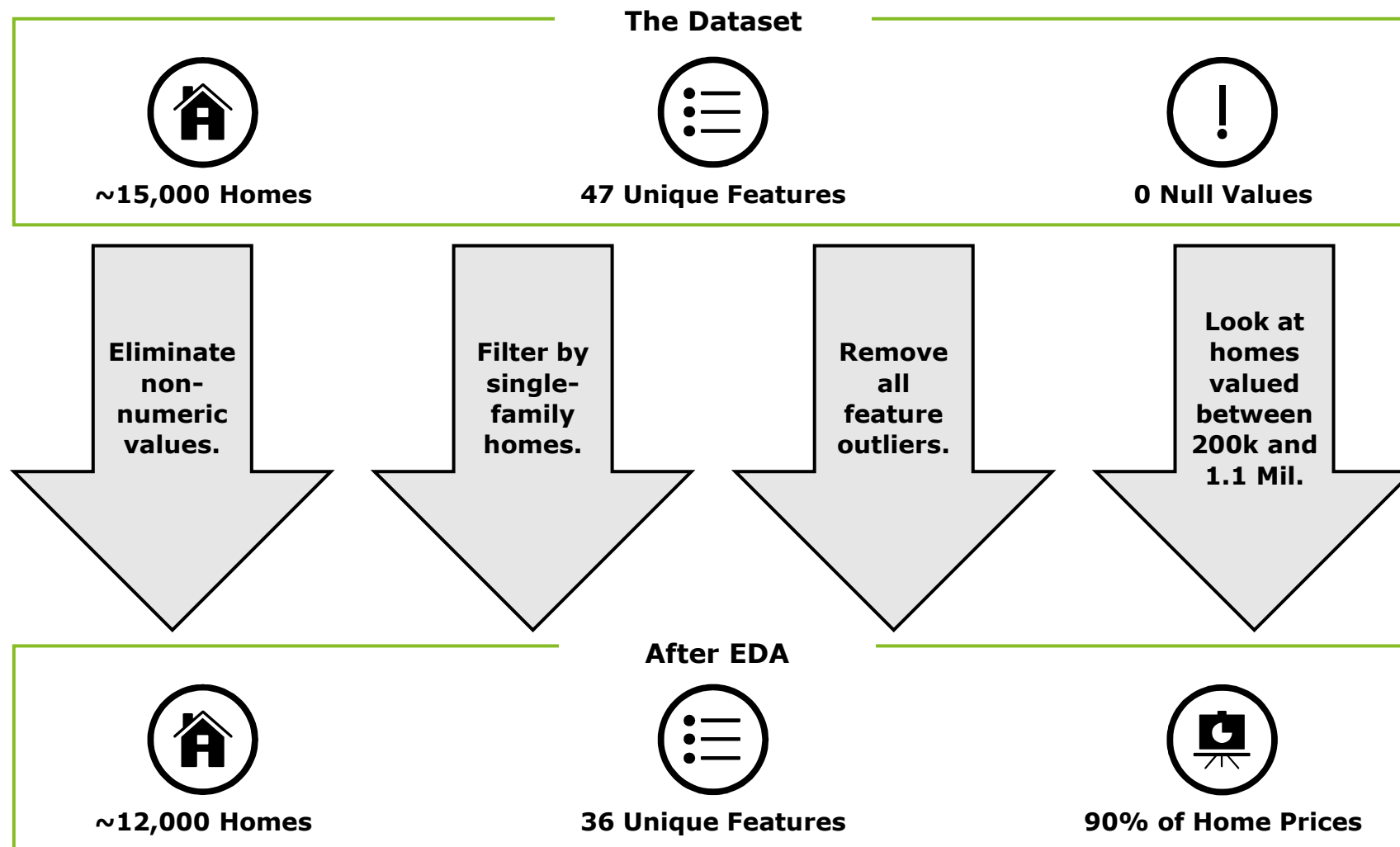
Real Estate Regression Cities and Urban Areas Housing

Data Analysis & Cleaning

The dataset our team found was very clean, however there were still additional manipulations to be made as well as points to consider.

Key Consideration

- Data is from 2018-2021 which means our results will be skewed to prices of that era
- Not a complete representation of the entire Austin real estate market



Modeling Approaches

To ensure a strong analysis, Group1 tested three different models, evaluating each on standard metrics.

Linear Regression

Used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

XGBoost

An advanced machine learning algorithm that boosts the performance of decision tree models by combining multiple weak learners into a strong predictive model.

Neural Networks

Deep learning models inspired by the human brain, capable of learning intricate patterns from data through interconnected layers of artificial neurons.

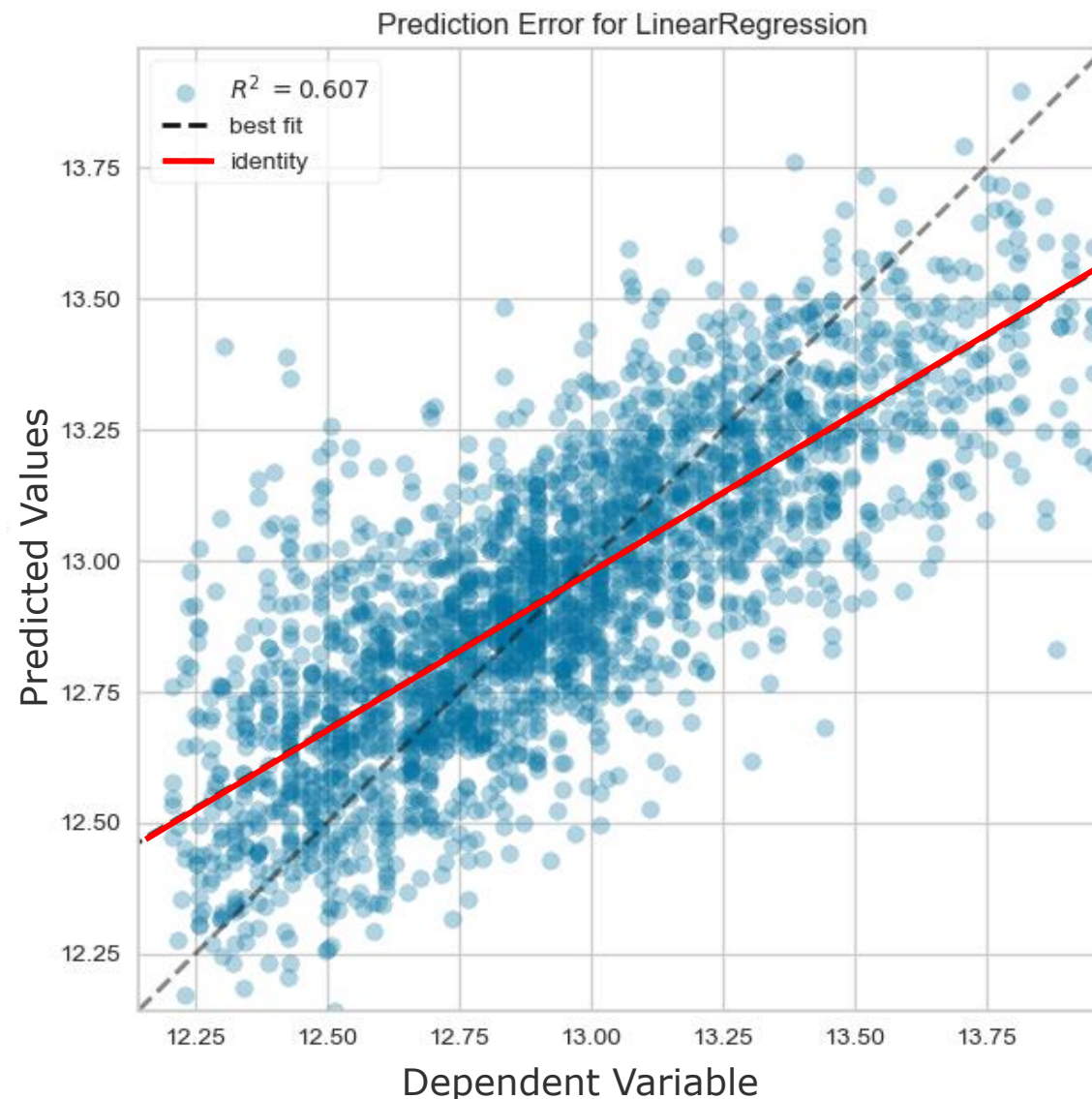
Our team will use **MAE (Mean Absolute Error)** as the core evaluator to compare our models. This evaluation metric provides a great +/- for a range to expect which each prediction.

Modeling Approaches: Linear Regression

A linear regression model is simple to enact so it became a great baseline model to evaluate potential and compare to more complex models.

Analytical Highlights

- Determination coefficient is 0.60 in other words, this model is explaining only the 60% of data and we expected at least 75%
- The MAE shows that our model is off by about 82,041 dollars in each prediction, in other words we could sell a house 82,000 more or less.
- This technique it's easy to apply but wasn't enough for our case, we needed reduce MAE.

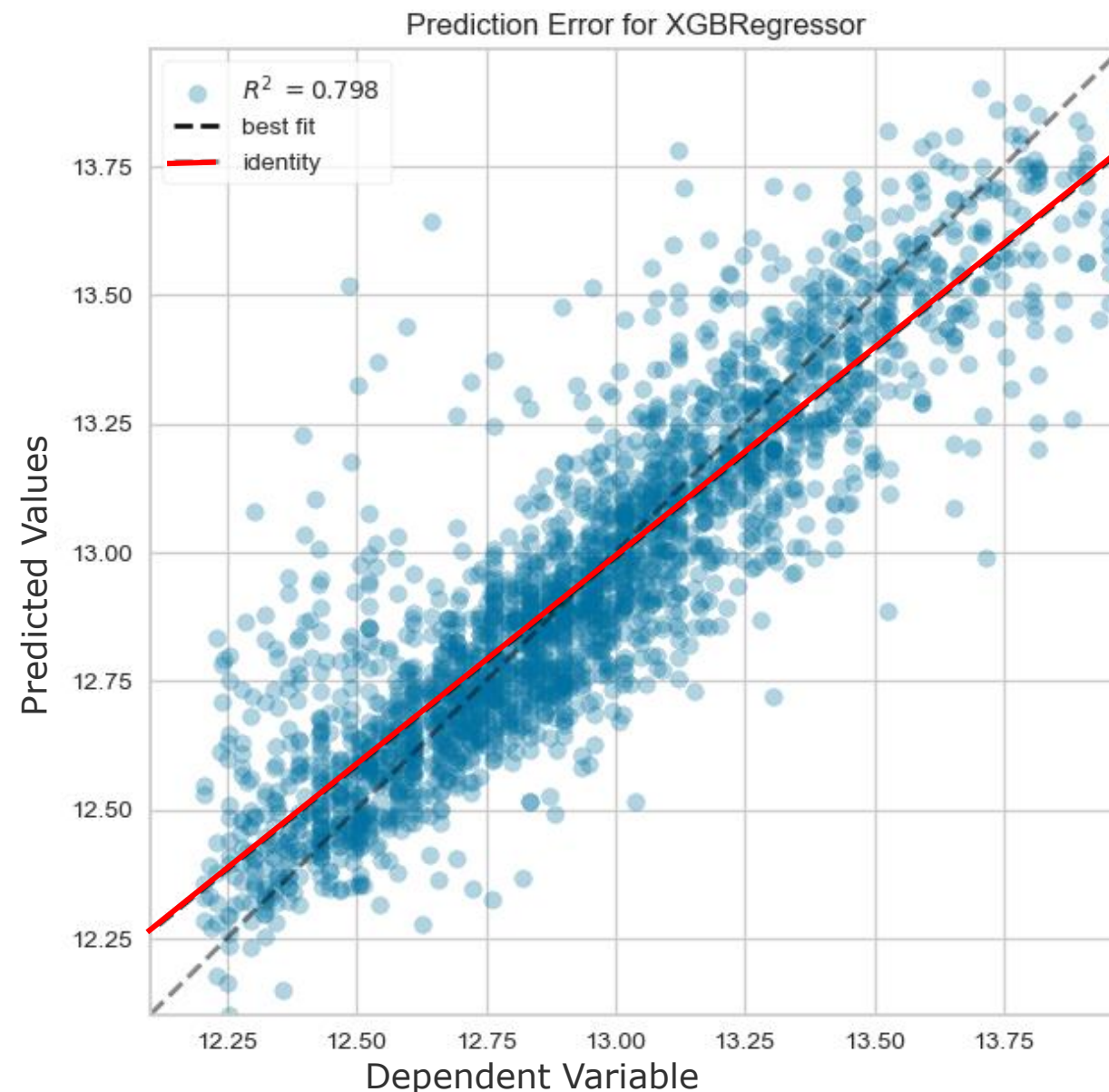


Modeling Approaches: XGBoost Regressor

XGBoost typically excels when trying to create a model with a high-dimensional dataset like the one that we have.

Analytical Highlights

- Determination coefficient is 0.798 in other words, this model is explaining the 80% of data.
- The MAE shows that our model is off by 55,456 dollars in each prediction, in other words we could sell a house 55,000 more or less.
- This technique gave us the smallest MAE
- We applied another evaluation called "cross validation" and we got the same determination coefficient.

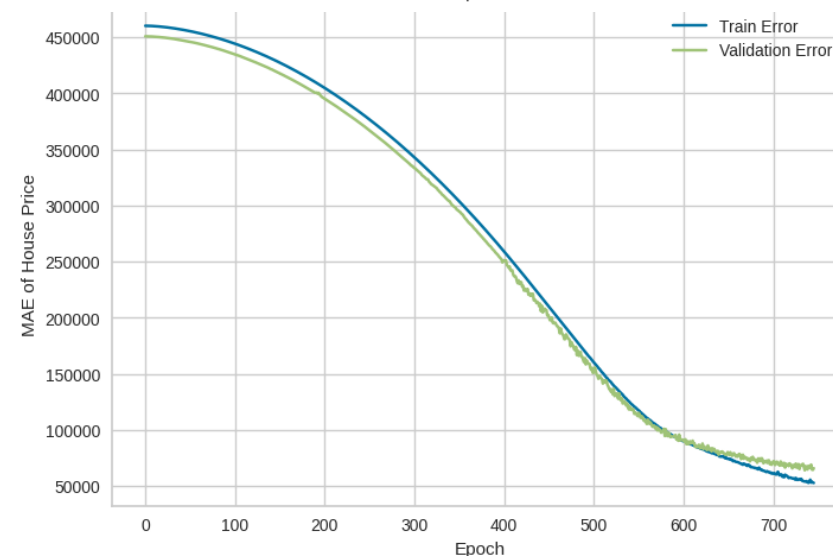
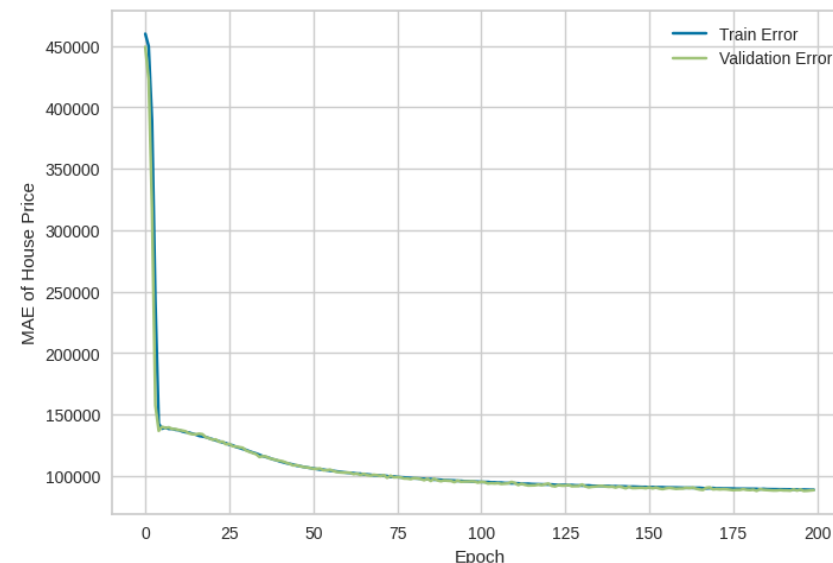


Modeling Approaches: Neural Networks

To create a more accurate algorithm our team created a neural network to create a more accurate home prediction.

Analytical Highlights

- Neural networks are complex, take more resources to create, but can often provide more accurate results.
- Multiple approaches and features were changed to attempt to optimize the Neural Networks.
- Many of the models trained had similar patterns of results, leveling off on MAE close to the 100,000 mark.



Modeling Evaluation & Outcomes

Using Mean Absolute Error (MAE) our team was able to evaluate the three separate models and determine the ideal model for FlatTech's purposes.

Linear Regression

MAE: 82,000

**Our model is off by roughly
\$82,000 dollars in each
prediction**

XGBoost Regressor

MAE: 55,000

**Our model is off by
roughly \$55,000 dollars in
each prediction**

Neural Networks

MAE: 70,000

**Our model is off by
roughly \$70,000
dollars in each prediction**

Based on our analysis, we are recommending additional exploration into the **XGBoost Model**.

Limitations and Next Steps

With an MAE of 55,000 results are promising, but additional work is needed to optimize our model.

Limitations

- 1** | 2018-2021 data
- 2** | Not entirely representative of Austin
- 3** | Large variance in prediction

Next Steps

- 1** | Collect additional data
- 2** | Tune the model using various hyperparameters
- 3** | Validate and present optimized model

Group1 Consulting

Thank you!
Questions?

