Data Science Project Team 2

Edmonton Housing Price Prediction

Mohammed Ghuzi, Rubia Martignago Mariath, and Shannon Wilson

Business challenge

New home owners deciding on an area within Edmonton where home properties are priced moderately yet valued high for long term investment purposes

Real estate investors predicting home prices for investing, holding and selling based on location.

Data Review

How much data do we have?

- Lots of data from 2012 to 2021 but few features to select from.
- Some missing values in lot size, location (latitude & longitude points), and year built.
- No missing values in assessed value, neighbourhoods, garages and zoning

What correlations can we draw?

- Is there increased property value with garages?
- Is there a correlation in the year the house was built to property value?
- Neighbourhood location increase property value?
- Certain zoning categories with higher property value?

Link to tableau dashboard: Tableau Online

Initial Findings

Narrow down data

- Due to the amount of data, we will need to scale down to an appropriate amount of years.

Property classifications

- Will require filtering out types that might skew data such as commercial, farmland and other residential types. Residential only

Missing data on certain features

- Latitude and Longitude missing values.
- Years built
- Suite
- Lot size

Factors that impact property values to consider

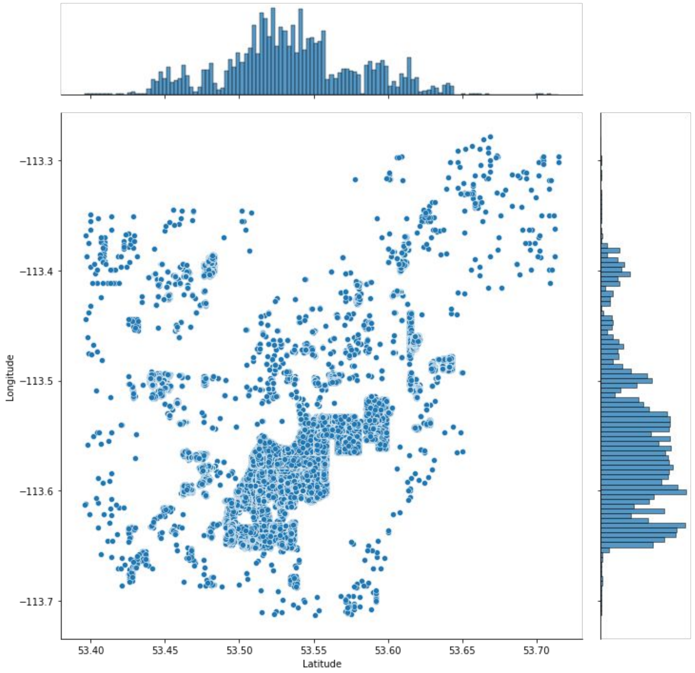
- Do garages built have an impact property values?
- Does the year the property built impact property values?
- Zoning codes to be considered.

Data Preparation

- Scaled back data by initially looking at assessment years 2019, 2020 & 2021
- Filtered down residential types by determining unique values and using residential only
- Feature engineer “Age of property”
- Missing data for longitude, latitude and age for some properties were removed instead of benign filled in
- A large amount of suite values were missing as many as the dataset. We decided to ignore the suite in our analysis.
- Finalized:
    - Target: Assessed Value
    - Features: Assessment Year, Location (Latitude, Longitude), Lot Size, Age and Garage.

This factor was reviewed once our model was run the first time

https://prod-ca-a.online.tableau.com/t/naitdatascienceproject/views/Map2021Edmonton/Dashboard1?:showAppBanner=false&:display_count=n&:showVizHome=n&:origin=viz_share_link



Modeling

3 Models used. Ridge Regression, Random Forest Regressor & Gradient Boosting Regressor

- Reason we choose is the regression model is because we were looking for a numerical value and this seemed to be the simplest model for first iterations.

1 Round Results using Ridge Regression

- The ridge regression model produced very poor results. 0.21 as a precision score.
- Changed the parameters of the model. Simplest model, without many parameters to adjust.
- We also concluded that we needed to scale back on the data to only one calendar year (2021) as the year over year data we were using could be deemed duplication/redundant.

2 Round Results using Random Forest Regressor

- Only 2021 data used
- Changed the model to random forest regressor. 0.49 was the initial score
- Changed the parameters of the model. Modified max_depth=8, n_estimators=100 and result was 0.71
- Upon review of our results and features used results seemed skewed, further review required

3 Round Results using Gradient Boosting Regressor

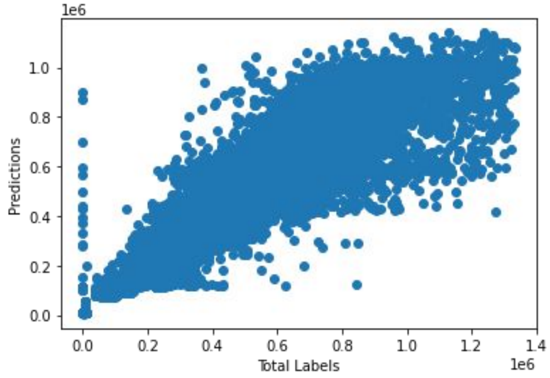
- Again, only 2021 data used
- Changed the model to gradient boosting regressor kept max_depth=8, n_estimators=100 and added more parameters learning_rate = 0.1, loss = 'ls'. The results was a more realistic 0..72
- Further refinements to future iterations to increase score required

Modeling

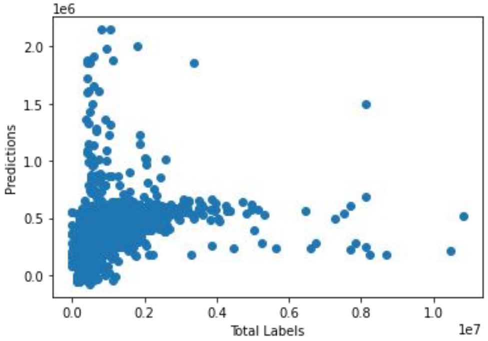
How did we determine the measurement quality of the model used?

Tried to ensure the highest r2 score possible (above .80) to ensure quality.

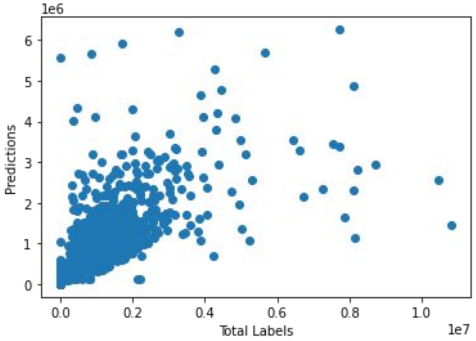
Further refinements were made on removing more outliers, and reducing lot size range.



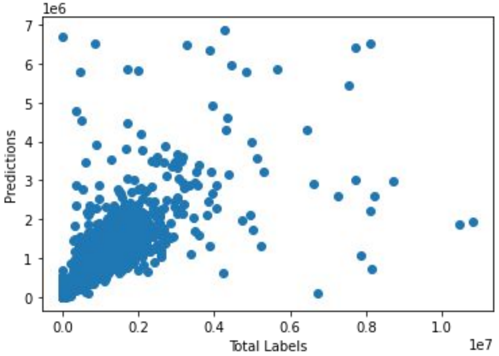
Random Forest Model after refinement
(r2 = 0.86) (MSE = 4873951301.05)



Ridge Model (r2 = 0.21)



Random Forest Model (r2 = 0.71)



Gradient Boosting Model (r2 = 0.72)

Future work

Merge data from crime data statistics by neighborhood to determine property assessment impacts.

Merging data from census dwelling unit by structure type neighborhood data.

Refinement of outliers removal and verify the influence of Latitude and Longitude in the models by removing them.

Spend time on year built feature as there were a significant amount of missing values and work on determining best way to add.

Spend more time filtering zoning of properties to nail down targeted properties.