House Simple

Khaled Al Najjar

Farukh Jabeen

Saad Moinuddin

Bhavnil Patel

Anna Sak

Chongli Zhao

# Introduction:

Data - recordable and packaged information - is perhaps the most useful commodity in the world today. A positive feedback loop of computational advancements as well as storage technologies throughout the last century have forged a massive industry out of the collection and analysis of data. The abundance of data - especially open source - has allowed the sharing of information on unprecedented levels and evolved the area of solution development. Large corporations have been thriving off utilizing big data and machine learning - improved automated performance through repetitive tasking - to produce new services on a massive scale. However, machine learning is a technology that even small institutions can adopt to develop solution to improve quality of life and provide niche services.

Knowledge based services - such as providing expert advice to consumers based off of precise data analytics are becoming more and more common. Consumers are looking to gain accurate answers to their questions and there are more questions than ever before. Examples of niche services that utilize AI could be using facial recognition to identify soft skills to potential human bias in interviewers, using natural language processing to identify grammatical errors in speech writing, or even using an accelerating tools to sift through massive stores of data and determine the optimal algorithm for model building.

Ontarians - especially residents in the GTA - are especially aware of housing market trends in the province. While citizens enjoyed a thriving housing market in the early 2010s, a recent near-halt in the value of homes in most parts of Ontario have potential home buyers and sellers concerned. Home sellers might be concerned about the current value of their home. How do you determine the value of your home? How can you standardize the process of home value determination in an agreeable manner? What features matter to potential home buyers? What tools would help a potential home buyer to identify whether they are overpaying for a home? What tools would help a potential home seller to list their home for a fair price and not miss out on higher sale price? Could decentralizing the information that is often restricted to real estate agents for decision making and sharing it with consumers assist in a personalized manner provide valuable insights?

Extensively overlisting a home’s value is risky, especially in a stagnant market. If the listing price is too high, it can deter home buyers from even considering to purchase the home. Even if the property is promising and garners traction, a challenging listing price will only draw the attention of “window shoppers.” Attracting attention - but no buyers - for extended periods of time brings two other risks with it and raises the following questions for others: Why is this property not selling? Is there something wrong with this property?

Over listed properties that do not sell for a significant period of time fall into a death spiral. The longer a home is listed without sale serves as a negative feature. When the owner subsequently drops the price of the home, despite gaining visits, it serves as a negative feature again, as buyers are wary as to what caused the home to drop in value. Is the market declining? Is the home a bad buy? This can in turn cause the homeowner - or advising real estate agent - to drop the listing price of the home even lower into margins of significantly lower profit. An event like this can be a heavy price to pay for a homeowner and oftentimes prevented with better planning on when to list a home and initial asking price. . We discuss “duration of time listed” and its relevance to buyers further in the later portions of this report.

Conversely, under listing the value of a home can lead to losing out on potential sales. When negotiating the sale price with buyers, sellers will rarely ever raise the price of their home on the bargaining table. Even if a seller feel’s that they have underlisted their property, raising the price of their home can lead to the buyer feeling cheated and leaving the table. Hence, keeping the listing price within a respectable range to the actual value of the home is vital for both buyers and sellers.

But what is the true value of your home?

# Data Acquisition:

The true value of a home is difficult to pin down, as it falls somewhere between the perception of the buyer and seller. For the sake of this project, we argue that the true value of a home is the final sale price, as it is the only documented price of a home that both parties agree on.

Data on home sales is often restricted to professionals within the industry - such as brokers and real estate agents - and thus access to listing sites like the Multiple Listings Service is limited. However, from the assistance of a very helpful contact who decided to remain anonymous, we were able to extract all home sales in Southern Ontario from 2015 to 2017 from the MLS database. Extraction of the data from the web source was simple and requires no coding. The MLS website allows for advanced filtering options of their database and downloading data from the resource is only a matter of a few well placed clicks.

The data was extensive and contained information on over 150,000 homes sold in Southern Ontario, from St. Catharines to Whitby. As it was available in CSV, we were able to view the data in Microsoft Excel.

# Data Preparation:

The database contained 21 variables, with relatively few rows containing a single blank. Blanks were only recorded for two variables. However, due to the redundant nature of the variables, they were excluded. The variables were as follows:

1. Mlsno: Each home possessed a unique MLS identifier, which would serve as the indexing column for the DataFrame once the data is imported in Python
2. Stat: 100% of the homes in our database had the status of “sold”. This variable was dropped when constructing our DataFrame in Python
3. Stno: Street number of the property. Address identifiers were dropped in favour of geographical coordinates
4. Stname: Street name of the property. Address identifiers were dropped in favour of geographical coordinates
5. Sttype: Street type of the property (blvd, road, way, circle, etc). Address identifiers were dropped in favour of geographical coordinates
6. City: City that the property falls under.
7. Area: Notable regions in each city, further used to cluster location of homes
8. Askprice: Asking price of the seller
9. Inputdate: Date the home was entered into the MLS. For the sake of this project, we assumed the input date and listing date were identical
10. Soldprice: Selling price of the property
11. Solddate: Date the property was sold
12. Type: This variable contained several types of homes such as (Semi-detached, detached, cottage, duplex, fourplex, condo, apartment, semi-detached condo, condo apartment, detached condo, townhome-condo, etc). For the sake of this report, we filtered this variable to only retain the following:
    * 1. Detached
      2. Semi-Detached
      3. Townhome
13. Style: This variable contained manly styles of homes from an aesthetical perspective. However, it contained over 20 styles of homes and their definitions were unavailable. This variable was dropped.
14. Bdrm: Number of bedrooms was deemed as a key indicator
15. Wshrm: Number of bedrooms was deemed as a hey indicator
16. Latitude: The geographical coordinates of the homes was deemed a key indicated and favoured over the conventional address of the home. Using coordinates, we are able to measure distance between two points
17. Longitude: The geographical coordinates of the homes was deemed a key indicated and favoured over the conventional address of the home. Using coordinates, we are able to measure distance between two points
18. House\_condo: Identifies whether the home is classified as a “house" or a “condo”. Criteria for this identification is not provided within the database.
19. Maint: Maintenance fees indicate the holding of the home, and whether it is free-hold or not
20. Aptno: Apartment number of the property - applies only to apartments.
21. Address. A variables that represents the concatenation of the house number, street name, street type, and city.

The purpose of this project is to predict the monetary value of the houses located in the Mississauga area. In this study, we applied basic data analysis techniques on data collected for housing prices in the Mississauga area to predict the selling price of a new home. First step of the analysis was feature engineering and cleaning of data to obtain important features and descriptive statistics about the dataset. Next, step was to properly split the data into test and training sets The housing price in Mississauga real estate market is highly competitive and you want to be the best real estate agent in the area.

For this project, the dataset has been preprocessed as follows:

1. Transform categorical data into numerical data (house type, house fee, and house area)
   * Rationale: categorical data needs to be quantified to be eligible as model inputs.
2. Adjust sold prices by CPI index
   * Rationale: under our cross-sectional regression model, the independent variables are time invariant (i.e. the distance to GO station and the number of bedrooms do not change with the passage of time). However, sold prices could change due to inflation. Adjusting the sold prices by CPI index ensures that the prices are measured by the same dollar value across the entire period from July 1st 2015 to June 30th 2017.
3. Incorporate real estate selling seasonality based on sold dates
   * Rationale: seasonality is an observable pattern in real estate market where its business tends to prosper in summer but cool down in winter.
4. Model real prices for selective house types (town house, detached, and semi-detached)
   * Rationale: the raw file contains other property types out of the scope of our designed model including vacant land, mobile trailer etc.
5. Scale the minimum distance to GO station and hospital to the range of 0 to 100 with the shortest distance to GO station and hospital close to 0 and the furthest close to 100.
   * Rationale: recalling the minimum distances to GO station and hospital can make them comparable to other categorical variables, which ranges from 0 to 34. This can also avoid having extreme estimated parameters / betas.
6. Transform the real sold price using log function
   * Rationale: the histogram of real sold prices indicates a pattern of log normal distribution. Using log function to transform the real sold prices can make them behave as if a normal distribution. In addition, log transformation ensures that the predicted real sold price is always positive, for example, exp(X) > 0.

While there was an abundance of variables, the redundant nature of some of them (for example address) was a determining factor on whether the variable should be dropped from our pandas DataFrame or not. To reduce the load on the system running the code, we filtered our data and dropped certain variables. How factors influence the price of a home vary significantly from city to city. We decided to only analyze data on the 22,472 homes in Mississauga. As we used the unique geographical coordinates as the location identifier for each home, address relative variables such as street name, house number, street type, and address were dropped from the DataFrame.

We constructed the following variables in Python for our analysis:

1. Season: We split the year into 3 seasonal times - winter, summer, and shoulder. The split was based on the following monthly basis:
   1. Summer: June, July, August
   2. Winter: December, January, February, March
   3. Shoulder: April to May, September to November
2. mindist\_GO\_score: Distance from each home to the nearest Go Station
   1. From geopy.distance import vincenty
3. mindist\_Hospital\_score: Distance from each home to nearest major Hospital
   1. From geopy.distance import vincenty
4. Log\_real\_prices
   1. By comparing CPI index values - an economic indicator - between sold dates of homes, we can determine inflation of change in the value of currency over time. By doing so, we are able to determine the true value of the home across multiple years in order to make more accurate comparisons

# Predicting a fair price with given features:

After through cleaning and preparation of data, the robust and valid regression models were developed which has capacity to predict the value of houses, we split the dataset into independent and dependent variable. The independent variable selected are mindist\_GO\_score, mindist\_Hospital\_score, Season, house\_area, house\_fees, bdrm, wshrm while the target variable is log\_real\_price (Real house price). We can spot a linear relationship between house price, number of bedrooms, washrooms and also the minimum distance from the Go station, from the exploratory analysis. We can also infer from the histograms that the variables seem to be normally distributed but contain several outliers. Exclude seasonality as an independent variable because the P value for seasonality is high, indicating that it is not affecting the dependent variable statistically and significantly.

# Result Explanation

We built the regression model to predict the real house price in Mississauga market. After fitting the model and plotting we can see that the blue dots in plot represents the test data displaying relationship between the house price and independent variables. The R- squared is 89.2 % which means that higher R-square percentage of the dependent variable represents better regression model and it can help in determining the real house price and dependent variables (log real sold price) can be explained by the independent variables (minimum distance to GO/Hospital, number of bedroom etc.).The High F statistic from the results indicates that all the coefficients are jointly significant in explaining the dependent variable (in other word, it is unlikely all the coefficients are zero). While high individual T statistics indicates that all the selective coefficients are individually significant in explaining the dependent variable. The value of skewness is positive and indicates that the tail is on the right side of distribution is longer means mean and median of the data is greater than mode.

# Conclusion:

Our data was extensive and very detailed and that shows in the minimal coding that was required to clean and manipulate it. However, we did experience challenges with the lack of data regarding square footage. For this reason, an analysis on bedroom numbers and washrooms number had to be conducted. These two variables would serve as indicators of house size. Keeping this in mind, we were critical of the how of home type (condo vs detached). An identical number of room and washrooms in a house and condo could yield a vastly different home price.

Appendix image four details the correlations between the variables. As expected, we observe a positive correlation between the three constructed variables and log\_real\_price.

# Appendix:

### Packages Used:

#!pip install geopy

*#!pip install python-google-places*

%matplotlib inline

**import** **pandas** **as** **pd**

**from** **datetime** **import** datetime

**from** **geopy.distance** **import** geodesic

**import** **geopy.distance**

**from** **geopy.distance** **import** vincenty **as** VIN

**import** **pandas** **as** **pd**

**import** **numpy** **as** **np**

**from** **matplotlib** **import** pyplot **as** plt

**import** **seaborn** **as** **sns**

**import** **pylab** **as** **plot**

**from** **sklearn.linear\_model** **import** LinearRegression

**from** **sklearn** **import** datasets

**from** **sklearn** **import** preprocessing

**from** **pandas.plotting** **import** scatter\_matrix

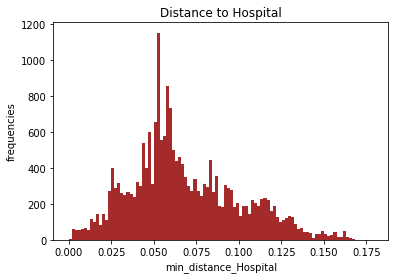
**from** **sklearn.model\_selection** **import** train\_test\_split

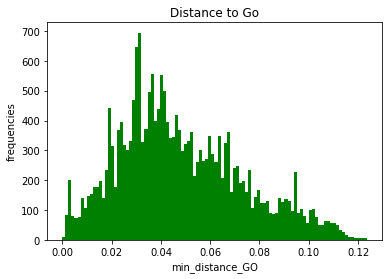
**from** **sklearn.linear\_model** **import** LinearRegression

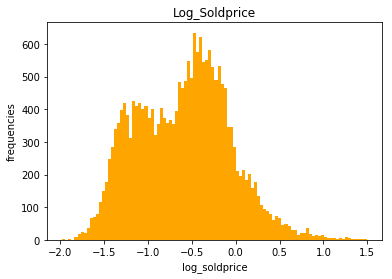
**from** **sklearn** **import** metrics

**import** **statsmodels.formula.api** **as** **sm**

### Distributions of Constructed Variables:







### Correlations Between Variables:

