Big Data Analytics

IST 718.M002

Project Proposal for

The application of PySpark and ML in the classification of Housing Market Heat Index

Submitted by

Aadit Malikayil

Bhavika Karale

Kapil Tare

Shreyas Kashyap

Under the guidance of

Prof. Dunham

Professor

Spring 2024

Table of Contents

Objective

The housing market is considered to be very volatile. As a buyer, it can be daunting to invest in such an uncertain market. It is important to ensure that the investment is made when it is a “buyers' market.” A buyers’ market is considered as such when the supply of real estate is greater than the demand for real estate. Housing prices are usually cheaper during a buyers’ market and houses are listed on the market longer [1]. On the other hand, sellers need to ensure that they are maximizing their return on investment (ROI) when selling their assets. In this application, it is important to ensure that the transaction is completed during what is considered as a “sellers’ market.” This is the opposite of a buyers’ market where the demand of real estate is greater than the supply of real estate. Sellers have leverage over buyers in the market, as houses are priced higher than during a buyer’s market and are not listed on the market for as long [1].

Calculations are made for each geographic location and a range of Heat Index Values (temperatures), are associated to each category of market conditions. Zillow applies Time-Series to capture the balance between for-sale supply and demand on a listing [2]. Zillow’s calculations rely on user engagements on house listings, share of listings with a price cut, and the share of conversions of for-sale listings to pending in 21 days (about 3 weeks). They consider a score of 70+ to be a “strong sellers’ market.”; a score of 55 to 69 to be a “sellers’ market.”; a score of 44 to 55 to be “neutral market.”; a score of 28 to 44 to be a buyers’ market.”; and a score of 27 and below to be a “strong buyers’ market” [2]. However, we believe that there are other factors that can influence the classification of market conditions.

Regional preferences are researched as an effective indicator of whether a market is a buyers’ or sellers’ market. A buyer’s location preference could add to or remove from demand and supply for a geographical location [1][3]. Another feature that is complemented by regional preferences is seasonality of a geographical location [3]. Is it more of a sellers’ market in New York during the winter compared to the summer? Are vacation homes an important contributor to market conditions? These are questions that the model will shed light on during our project. The age of houses in a geographic location is also another feature that would affect not only the price of the house but also the demand for real estate. Typically, newer homes appraise better and have more demand [4]. Economics and Interest rates are another set of features that contribute to the demand and supply of housing properties [4].

Our objective with this research is to provide sellers and buyers with predictions on the current state of the market. Our stakeholders comprise of the following demographics: real-estate companies and agents, homeowners, home buyers, and banking institutions. Future research will include tracking our predictive model's success and adding political features that could affect the market's status.

Dataset Description

Our data mainly comes from Zillow housing (Zillow). The shareable link for the same can be found here.

Big Data Project.

Zillow is a trusted real-estate company that was founded in 2014. For the scope of the project, we are going to be using data from 2018 to 2024 (tentatively). We used the Zillow API to download the data for housing price, sale price, market heat index, days to closing and days to pending. A brief description of the data are as follows:

1. **Housing Price**: Zillow has a measure called the ZHVI (Zillow Home Value Index) that measures home values and market changes across regions and housing type. The ZHVI dollar amount is designated as the "typical" home value for a region as it is calculated as a weighted average of the middle third of homes in each region.
   Rows: 896, *Columns*: 78
2. **Sale Price:** The median price at which homes across various regions were sold.
   Rows: 710, *Columns*: 78
3. **Days to pending:** This data contains the median number of days it takes homes in a region to change to pending status after first being shown as for sale. A home is considered pending when the offer has been accepted by the seller, but the deal hasn't been finalized yet. This excludes the number of days the house is on contract.
   Rows: 724, *Columns*: 84
4. **Days to close:** Median number of days between the listing going pending and the date of sale.
   Rows: 625, *Columns*: 83
5. **Market Heat Index:** As defined in the objective above, this data is basically a score that helps us categorize the market in different regions. For instance, a continuous observed increase in the score over a few months in a region would indicate that there are more potential buyers available per listing.
   Rows: 929, Columns: 85

Methods

As mentioned in our objective, we aim to predict the Heat Index of the housing markets basis the available historical data ingested through Zillow API. With the help of these predictions, we plan to understand the real-time trends, thereby helping our various stakeholders to make data-driven business decisions.

Now basis our understanding of an ML project lifecycle, we have distributed our project into 3 stages –

1. **Data Collection and Preprocessing** –

   We will collect our data through Zillow API and keep the focus primarily on the key indicator tables mentioned above in the data description. The ingestion of these tables will be done using the PySpark library. We will also apply the traditional data scaling and null imputation techniques for data cleaning. For the feature engineering portion currently, our game plan is to extract new indicators regarding seasonality, geographic regions, and economic factors such as interest and inflation rates. We will also perform an in-depth Exploratory Data Analysis using the Pandas library to identify correlations and dependencies of the features on each other and visualize the same using libraries like Seaborn and Matplotlib.

2. **Data Partitioning and Model Selection** –

   As per traditional partitioning, we are planning to split our data in the following manner - 70% will be our training data, 15% will be our validation data and the remaining 15% will be our test data. Now, since our target variable is going to be a continuous one as we are predicting the heat index value, we'll be implementing various regression techniques linear regression, tree and forest regressors.

3. **Model Evaluation and Validation** –
   Now to evaluate our regression models we'll be utilizing the mentioned-below metrics –
   - Mean Squared Error
   - Root Mean Squared Error
   - R-squared

We are also planning to perform cross-validation along with necessary hyperparameter tuning to enhance the model's prediction performance. This will also help us avoid overfitting of the model.

Risks/ Concerns

Predicting housing market index can have various challenges that might impact its accuracy and reliability. One such obstacle is overfitting. The model might learn patterns that are mainly focused on the training dataset rather than the general trends. This might result in getting high accuracy during the training phase, but the model might perform poorly when new data is introduced. Another challenge is multi-collinearity, where the features present in the data might be highly correlated. This might result in unstable predictions by discerning the individual impact of variables. At times, information might not be available at the time of prediction. This can lead to an overly optimistic model performance.

Events like variations in interest rates, natural disasters or economic recessions can cause unexpected changes in the Real Estate market. Incorporating factors like these while modeling might be challenging. Moreover, the Real Estate market shows variations depending on the geographic location, weather conditions and factors related to supply and demand. Hence, a model that performs well for one location may not perform well for another.

Citations

1. *Buyer's market vs. seller's market: What does each mean for you?* Rocket Mortgage. (n.d.). https://www.rocketmortgage.com/learn/buyers-market-vs-sellers-market

2. Research, Z. (2024, May 15). *Zillow's market heat index methodology*. https://www.zillow.com/research/market-heat-index-methodology-34057/

3. Nguyen, J. (n.d.). *4 key factors that drive the real estate market*. Investopedia. https://www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp

4. Gomez, J. (2022, June 4). *8 critical factors that influence a home's value*. Opendoor. https://www.opendoor.com/articles/factors-that-influence-home-value