Georgia Tech®

MGT 6203

Instructor: Jonathan Fan

Group Project Proposal

Team #21

TEAM INFORMATION (1 point)

Team #: 21

Team Members:

1. **Nildip Chaudhuri, GT Id: 903951206**
   a. Education: Undergrad: BBA (2018-2023) and Computer Science (2018-2023), University of Waterloo, Canada
   b. Professional Background: Junior Data Engineer (CIBC), Data Engineer (TD Bank) - Internship, Treasury Analytics (CIBC) - Internship, Capital Markets summer analyst (CIBC) - Internship
   c. Projects: Personal project: Football event detection Model (computer vision), Work Project: Fraud analytics data ingestion into graph database (Neo4J)

2. **Erik Euler, GT Id: 903848886**
   a. Education: Undergrad: BA Statistics from Elon University (May 2017)
   b. Professional Background: Data Scientist with Booz Allen Hamilton (Aug 2017 - Present)
   c. Projects: Personal project: Fantasy Football lineup recommender, Work Project: Health Care surveillance automation

3. **Nick Sienicki, GT Id: 903004037**
   a. Education: Undergrad: BS Economics and Business Analytics from the College of William & Mary (Dec 2016)
   b. Professional Background: Product Manager/Data Analyst at Freddie Mac
   c. Projects: Personal project: Prediction of career length of NBA players, Work Project: Emulation of income calculation rules

4. **Maryam Paknejad, GT Id: 903835757**
   a. Education: BSc in Civil Engineering, K.N.T.U. Iran (March 2004), Data Analytics, Big Data, and Predictive Analytics Certificate (Toronto Metropolitan University 2022)
   b. Professional Background: Incident Analyst and EM Specialist at Toronto Metropolitan University, Civil engineer Consultant at Papila Consulting Engineers
   c. Projects: Certificate projects: Online Shoppers Purchasing Intention (click stream analysis), Winning Space Race with data Science (Predicting the success/failure of first stage of the SpaceX Falcon 9 rockets), Course project: Blackjack Simulation, Work Project: Analysis of Major Crime Indicators

5. **Giovanni Marrero, GT Id: 903845934**
   a. Education Background: Undergrad: Syracuse University, B.S. Biotechnology (2019)
   b. Professional Background: Research Associate II & Lab Manager (University of Pittsburgh - Immunology), Research Associate III & Lab Manager (Broad Institute - Fei Chen Lab - Spatial Transcriptomics)
   c. Projects: Coursework Related: Blackjack Strategy Simulator, Personal: Daily Fantasy Football Roster Generator

OBJECTIVE/PROBLEM (5 points)

Project Title

“Predicting Time on the Market for New Home Listings”

Background

Housing finance in the United States is a twelve trillion-dollar industry that touches every corner of the American economy, from a deep relationship with the federal funds rate and employment to historically low-risk government-backed securities for purchase by investors. Often, interacting with buying and selling a home is the largest financial transaction in an individual's life which makes getting the right home at the right time a critical decision.

There are several factors that go into which home to pick to make an offer on, but a key component remains the time that a home has been on the market. For a buyer, a home that has been on the market for a considerable time demonstrates the opportunity to offer a below-asking price due to the presumption of low demand. For a seller, the longer their home is on the market, it triggers more anxiety or re-evaluation of the asking-price, listing details, or even triggering delisting the home from the market.

But how long is typical for the market you are in?

Problem Statement

With the housing market as competitive as it is, it is becoming more and more difficult for buyers to know when to make their offer. By understanding how long a home is expected to be on the market within their metro area, buyers and sellers can act with more confidence in their decision making to remove listings from the market, identify opportunities to offer under-asking, or reevaluate the listing price.

Primary Research Question (RQ)

Can the length of time in which a home is placed on the market to when it is pending sale be predicted using macroeconomic data from the federal government and publicly available data from the leading real estate and rental marketplace providers?

Supporting Research Questions

1. Does median home price for a metro area interact with interest rates? Are there changes on higher levels of that range?
2. How does the short window of data (2018-present) that Zillow provides for the dependent variable impact the results?

Business Justification

Given the size of the housing finance industry, a business insight that can drive even a component of the industry to a more efficient action is worth a great deal. Realtors often operate on qualitative knowledge of a market, and buyers and sellers typically have limited to no experience at all with buying and selling a home independently. Creating metrics to get at a quantitative approach for evaluating time on the market for a specific listing is useful to many stakeholders. More knowledge about the time a specific home has been listed relative to the market could drive a buyer to identify a value investment or trigger a seller to lower asking-price to accelerate a sale. These small decisions often mean thousands of dollars of change in asking price. Being able to clearly define when these actions are probable or necessary to drive the desired impact using existing data, can lead to advantages in an often-asymmetric market.

DATASET/PLAN FOR DATA (4 points)

Data Sources

- Zillow core data: https://www.zillow.com/research/data/

  - The dependent variable of exploration along with several key predictors, Zillow's log of metropolitan statistical areas' (MSA) median days to pending will be compared to others available from zillow.com. Each is available by CSV download or API access.

- Zillow to fed mapping: https://data.world/zillow-data/crosswalk-between-zillow-and-federally-defined-regions
  - Zillow's RegionID field will be mapped to federally used and recognized identifiers for regional level analysis, if applicable.

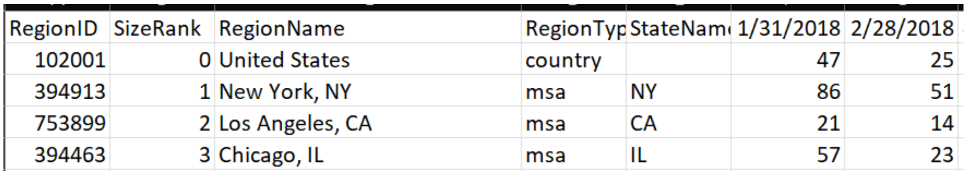
- Census bureau data:
  - American Community Survey data: https://www.census.gov/programs-surveys/acs
  - 2020 Census Demographic and Housing Characteristics: https://www.census.gov/data/tables/2023/dec/2020-census-dhc.html#by-topic
  - Aggregated datasets accessed using https://data.census.gov/
    - To repurpose the survey data and to numerically represent existing variables, county averages or rates per 100k households will be calculated and merged with Zillow data.

- Consumer confidence index: https://data.oecd.org/leadind/consumer-confidence-index-cci.htm
  - Although this data is representative of the entire United States of America as opposed to having select regional data, there is a potential for CCI to impact the time a property stays on the market.

- GDP and Personal Income - Regional Data: https://www.bea.gov/data/economic-accounts/regional
  - This data is given for the United States of America & also state-by-state. This data also exists in its raw form, without any adjustment for the regional buying power factor.

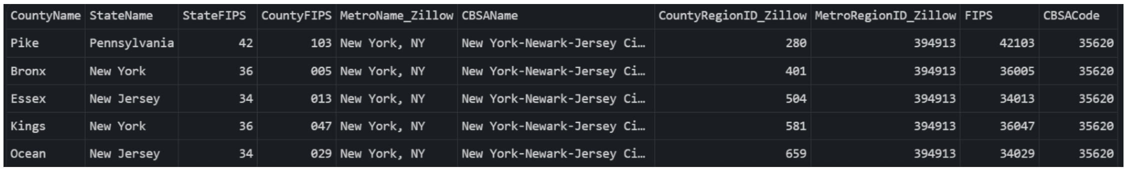
Data Description

Zillow's data comes sheet by sheet, which can be combined and manipulated to get from longitudinal format for placement alongside other aforementioned sources of data. It's available monthly or weekly divided by MSA:



RegionID	SizeRank	RegionName	RegionTyp	StateNam	1/31/2018	2/28/2018
102001	0	United States	country		47	25
394913	1	New York, NY	msa	NY	86	51
753899	2	Los Angeles, CA	msa	CA	21	14
394463	3	Chicago, IL	msa	IL	57	23

Zillow's RegionID MSAs will be linked to federally recognized MSAs using this mapping document:



CountyName	StateName	StateFIPS	CountyFIPS	MetroName_Zillow	CBSAName	CountyRegionID_Zillow	MetroRegionID_Zillow	FIPS	CBSACode
Pike	Pennsylvania	42	103	New York, NY	New York-Newark-Jersey Ci…	280	394913	42103	35620
Bronx	New York	36	005	New York, NY	New York-Newark-Jersey Ci…	401	394913	36005	35620
Essex	New Jersey	34	013	New York, NY	New York-Newark-Jersey Ci…	504	394913	34013	35620
Kings	New York	36	047	New York, NY	New York-Newark-Jersey Ci…	581	394913	36047	35620
Ocean	New Jersey	34	029	New York, NY	New York-Newark-Jersey Ci…	659	394913	34029	35620

Key Variables

- Dependent variable:
  - Days to Pending (Zillow)

- Independent Variables:
  - For-Sale Inventory: The count of unique listings that were active at any time in a given month. (Zillow)

- New Listings: Indicates how many new listings have come on the market in a given month. (Zillow)
- Percent of Sales under/over List: Ratio of sales where Sale Price below/above the final list price; excludes homes sold for exactly the list price (Zillow)
- Days to Close (mean/median): Number of days between the listing going pending and the sale date. (Zillow)
- Median List Price: The median price at which homes across various geographies were listed. (Zillow)
- Median Sale Price: The median price at which homes across various geographies were sold. (Zillow)
- Price Cuts: The mean and median price cut for listings in a given region during a given time period, expressed as both dollars ($) and as a percentage (%) of list price. (Zillow)
- Share of Listings with a Price Cut: The number of unique properties with a list price at the end of the month that's less than the list price at the beginning of the month, divided by the number of unique properties with an active listing at some point during the month. (Zillow)
- Real Personal Income (State & Country)
- Consumer Confidence Index (Country)

The federal funds rate impacts the housing market substantially. The expectation is that this predictor will have a significant impact on the time that a house stays on the market in all US markets. Within the Zillow data and beyond, there is likely to be a significant amount of correlation between explanatory variables. Final selection of predictors through variable selection and introduction of interaction terms has to be controlled rigorously. The anticipation is that many of the Zillow predictors will not be included in the final model.

APPROACH/METHODOLOGY (8 points)

Planned Approach

In all the models mentioned below training, validation and test sets will be used with 60-15-25 percent split respectively. The subsets with respect to all the features will be stratified to lessen the effects of overfitting. The planned optimization for each model has been explained below.

1. **Multivariate model**: The objective is **not** to find a casual interoperation of a variable. Instead, the focus will be on a model fit to get the closest predictor to the actual time the house is on the market. Polynomial and logistic transformations will be experimented based on the observations from Exploratory Data Analysis (EDA). In terms of hyper-parameters, Lasso and Ridge Regression techniques will be tested using the scikit-learn library to optimize the regression model and reduce overfitting. A significant amount of time will be dedicated to trimming variables through Lasso and Ridge Regression, as well as analysis of covariance between explanatory variables.

2. **Regression Tree and Random Forest models:** Regression Tree is a derivative of the multivariate regression model. As the dependent variable is continuous, a regression tree will be used as opposed to a vanilla decision tree. Then, the same logic will be used to create the Random Forest model. This should reduce some of the overfitting that would be caused by the decision tree. In terms of hyper-parameters, various depths, minimum sample split, number of estimators and number of trees will be tested. The latter parameters will be used to improve model fit. The data would not have to be further transformed as the scikit-learn library would automatically select the root node and recursively split the trees.

3. **Vanila ANN:** For this model, the input would be a vector of N features and the output would be one node. This would be a continuous value. The hyper-parameters would be the epochs, activation functions, optimizers (adam, SGD, etc.), number of layers, dropping rate, and learning rate. The data would need to be transformed as each vector would have all the required features (NumPy array). This would be the input layer.

Each of the above models has their respective pros and cons. Intuitively, the random forest and ANN models would be over-fitted. Moreover, even with stratifying the train and test sets, getting an exact match of a continuous variable would be difficult. Hence, there is a need to create a range of correctness when evaluating each model. The latter means a value X would need to be selected where the predicted output would be f(x) +/- X. Also, these categories would be converted into buckets, but that would lead to other biases. As a result, the model with the highest number of predictions within the thresholds would be the model of choice. However, the model performance can also be measured by the lowest average prediction error. Intuitively, this method should yield the same result, but it would be interesting to see otherwise.

Anticipated Conclusions/Hypothesis

The expectation is to determine several economic factors that contribute to the increased duration of a house remaining on the market. Another hypothesis is that there will be a positive correlation between the length of time on the market & the cost of the house. Other personal economic factors may have various impacts on the time a house stays on the market, which will be investigated when conducting exploratory data analysis & building models. The goal is to be able to create a strong model, that when taken into consideration several key factors such as region, cost, time, macroeconomic factors, and economic status of regional populations, will be able to predict how long a home will be on the market. Once an appropriate model is selected for this purpose, the model will be tested to determine how accurately it can forecast the time on the market for homes by using the test dataset. The objective is to have an accurate model with a confidence interval that yields meaningful predictions and conclusions.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

Predicting the time that a house will be on the market will result in better predicting how the market will behave, thus giving buyers better information about the timing of their offers to buy a home. Also, if there are outlier values for how long a house has been on the market, it is reasonable to assume that there is either a large discrepancy in the listing price to the value of the home, or that there is an extremely limited demand to purchase those properties.

PROJECT TIMELINE/PLANNING (2 points)

October 2-7: Finalize project proposal

October 9-28: Compile all relevant data and begin EDA, prepare necessary cleaning scripts for the selected data sources, finalize dataset, and begin exploratory data analysis, begin early-stage model development

October 30 – November 4: Writeup progress report deliverable for submission covering data compilation process, details on finalized dataset, and any relevant details from the beginning stages of model building.

November 6 – 15: Create models and begin model tuning stage

November 15 – December 4: Finalize all models and determine which model to select as a part of the group submission, cleanup groups GitHub page, writeup the corresponding report for submission, submit final project report and package code files.