MGT 6203 Spring 2024
Group Project Proposal
Team #60

TEAM INFORMATION (1 point)

Team #: 60

Team Members:

1. Joy Kakkanad; jkakkanad3
   . Working for Wells Fargo as a Risk Analytics Consultant. Bachelor of Science in Information Technology. This is my second semester, and I took CSE 6040 and MGT 8803 in my first semester. I have worked in the analytics space for about 10 years involving SQL databases, data warehousing, Tableau, Power BI and SSRS amongst other tools.
2. Russell Dawkins Jr. ; rdawkins6
   . Research Scientist at the Georgia Tech Research Institute. Bachelor of Science in Computer Engineering from Georgia Tech. I have a strong software background but no data analytics background. This is my second semester in the program. Last semester I took ISYE 6501 and MGT 8803.
3. Marc Presume; mpresume3
   . Food Scientist at Conagra Brands. Master's of Science in Food Science from Auburn University. No analytics background. This is my second class after ISYE 6501.
4. Mohamed Abdallahi; mabdallahi3
   . Previously in operations at Amazon. I have a double major in Chemistry and Geosciences from the University of Calgary (Canada). I'm relatively new to the field of analytics and to the OMSA program. So far I've completed ISYE 6501 and CSE 6040.
5. Evan Swain; eswain7
   . Formally Director of Research in head trauma litigation; BA Philosophy, University of Pennsylvania; analysis of medical state of the art in new diseases, jury pool analysis and polling

OBJECTIVE/PROBLEM (5 points)

Project Title:

Combining Disparate Approaches to Better Predict Future Residential Housing Prices

Background Information on chosen project topic:

The United States housing market plays a crucial role in the economy, with a market cap of $47 trillion (about $140,000 per person in the US) and anticipated growth of 4.5% per year in the coming decade (Rosen, 2023). Given its significant contribution to economic activity and wealth generation, precise forecasting of residential real estate prices is vital for stakeholders.

Accurately forecasting residential property prices is pivotal in the real estate sector, unlocking numerous business opportunities and guiding critical decision-making processes. From assisting first-time homebuyers in securing fair deals to enabling large-scale investments in residential properties, the implications of precise price prediction are manifold.

Traditionally, price prediction in real estate has been approached through two primary methodologies. The first involves forecasting real estate prices in aggregate across regions or markets over time, while the second focuses on predicting individual property prices by comparing them to recent sales within the same market. These approaches have historically relied heavily on historical sales data, which, while informative, often overlooks crucial factors such as neighborhood amenities and market sentiment.

Factor-based models offer an alternative approach, aiming to predict the price of a particular property within a specific market or neighborhood. These models utilize attributes like lot size and number of bedrooms to identify "comparable properties" and make predictions based on this comparison. While effective in the short term, factor-based models tend to lack predictive power beyond the immediate future, providing limited insights into pricing trends over three, six, or twelve months. This restricts their usefulness for stakeholders requiring comprehensive, long-term price forecasts.

Recent studies have explored innovative approaches to price prediction, including the use of aggregated internet search data. Authors have demonstrated the potential of such methods in predicting housing prices on a large scale across time (Choi & Varian, 2012), (Wu & Brynjolfsson, 2015), (Veldhuizen, Vogt, & Vogt, 206). However, while these approaches have shown promise in forecasting trends in major markets and broad indices, their applicability to smaller markets and individual properties remains limited. Further research is needed to address these challenges and develop more robust predictive models capable of providing accurate and actionable insights into residential property prices.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

In this project, we attempt to combine these approaches to better predict the sales price of individual residential properties. By adding lagged aggregated internet search data across time to a factor-based model, we aim to improve on the prediction of the original model.

State your Primary Research Question (RQ):

Can a combined model significantly outperform a time-series or attribute-based approach?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. What is the best factor-based model we can create using this data? (model selection, factor selection)
2. How can the time-series element be incorporated into the factor based model?
3. How does the combined model compare in terms of interpretability and robustness to outliers compared to the original factor-based model?
4. Has the housing market really increased since 2020?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

Accurate prediction of single residential real estate prices, if successful, present a multitude of business opportunities. Point of sale and continuous monitoring of accurate property market value has implications for price speculation for buyers and sellers, tax collection for governments, and mortgage/bridge loan evaluation for financiers, renovators, and builders.

DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

Our real estate sales data is aggregated data on single-family home sales in the Houston metro area for the past five years. It was downloaded from the agent section of the Houston Association of Realtors' website (har.com) on February 10, 2024 and spans the period December 08, 2018 to February 09, 2024.

https://www.har.com/

Our internet search data is aggregated statistics on real estate-related searches on Google for the same period. We consider a variety of search terms typically used by retail house-buyers, and consider search statistics at the country-wide, Texas, and Houston levels geographically.

https://trends.google.com/trends/explore?date=all&geo=US-CT&q=housing%20prices&hl=en-US

Data Description (describe each of your data sources, include screenshots of a few rows of data):

The real estate data from har.com has around 363 variables and more than 120,000 rows. The raw data is more than 400 MB in size. This data would need to undergo the preparation steps to bring it to a stage where it can be utilized to generate good models, do research and for further steps in the data modeling process. A snapshot of the raw data given below:

Acres	AnnualMaintDesc	Area	MasterBathDesc	BathsFull	BathsTotal	BedRoomDescription	BedsTotal	BedsMax	SqFtTotal	CDOM	Census Tract	City	CloseDate	ClosePrice	ClosePriceAdj
0.1277	No	2		1	1.1	All Bedrooms Down	3	3	1032	0	2125	Houston	2/4/21	40000	4000
0.1148	No	4		1	1		3		864	26	3135	Houston	3/23/21	55000	5500
0.1643	No	2	Primary Bath: Tub/Shower	1	1	All Bedrooms Down	3	3	1128	13	2336	Houston	2/4/21	56000	5600
0.2296	No	2		1	1		2		792	2	2336	Houston	2/8/21	59000	5900
0.1515	No	3		1	1	Primary Bed - 1st Floor	3	3	1380	0	3322	Houston	3/3/21	59000	5900
	Mandatory	34		1	1		2		1136	5	2225	Houston	3/4/21	59900	5990
0.233	No	3		1	1		4		1559	0	3329	Houston	3/16/21	64000	6400
0.1148	No	3		1	1		2		875	1	3311	Houston	2/4/21	67150	6715
0.1504	No	3		2	2	All Bedrooms Down, Primary Bed - 1st	4		2143	0	3317	Houston	2/1/21	70000	7000
0.1653	No	11	Disabled Access, Secondary	1	1	All Bedrooms Down	3		1100	45	5333	Houston	2/4/21	70000	7000
	Mandatory	2		2	2	All Bedrooms Down	3		1188	0	2323	Houston	3/18/21	74000	7400

Fig. 1 Snapshot of the housing market data in Houston

	A	B
1	Category:	All categories
2		
3	Month	house price: (Connecticut)
4	2004-01	0
5	2004-02	45
6	2004-03	0
7	2004-04	0
8	2004-05	0
9	2004-06	0
10	2004-07	0
11	2004-08	32
12	2004-09	0
13	2004-10	0
14	2004-11	30
15	2004-12	90
16	2005-01	0
17	2005-02	79
18	2005-03	27
19	2005-04	0
20	2005-05	0

Fig. 2 Google trends data

For the internet search data, we will use Google Trends data for a variety of search terms relating to home sales (This may involve additional research and analysis regarding the time frame since the home buying process takes more time than a typical consumer product). The search terms will include topics a typical home buyer may use in home buying from the start of the home buying process to its completion.

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

Dependent variable: “sale amount” in real estate data

Variable selection: The predictors that we think would be important for this data, at this initial stage would be Total sq feet, number of bathrooms, bedrooms, garage, appliances, heating, cooling, neighborhood, school district, swimming pool, and year built. This is our initial hypothesis of the predictors to be important to predict selling prices. We may have to create new variables that we foresee in two distinct kinds of situations. First, there are some predictors, for example, that have comma separated values. We may have to separate these into their own predictors. We will also have to create dummy variables for some predictors which have categorical data.

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

We will approach this problem in two steps. First, we will create a factor-based model using only the real estate data from the Houston Association of Realtors. Then, we will add in smoothed time series values from the Google Trends data as additional factor(s) to see whether they improve our predictive model.

In the first step, we will begin with cleaning the data and dealing with missing values. The real estate data has columns with missing values, many of which we may discard. In cases where columns we judge to be important are missing values, we will select an approach to impute them, such as using the mean or estimating by regression. We will judge the column's importance based on our reading of the literature and by soliciting advice from subject experts (realtors).

We will then proceed to model selection. We expect to find a good predictive model through experimentation with linear regression and clustering models. We will concurrently refine variable selection through elastic net or its component approaches.

In the second step, we will begin by query selection. We will create a list of terms related to single-family home purchases and obtain the data from Google Trends. Next, we will examine the correlation between the search metrics for these terms as we expect them to be highly correlated. We will then decide how to aggregate this data into one time series. Choosing one representative term or using the mean are two approaches we are considering.

Once we have our time series data, we will incorporate the data into our factor-based model as an additional factor. We expect this to be an iterative process, as we must find the best delay between the search data and sales date. We may also try transforming the data through exponential smoothing or incorporating momentum indicators to improve the search data's contribution to the model.

Finally, we will compare our first model (only real estate data) with our second model (with internet search data added). Statistically, we will compare R-squared values and other metrics. Practically, we will consider accuracy in back-testing. We may also add a classification model (logistic regression or support vector machines) if we judge it to improve ease of use in the business case.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement

We anticipate our factor-based model based on real estate data will perform in line with similar models in predicting the sale prices of single-family homes. We further anticipate adding internet search data will improve the model by capturing more of the variance in sales price.

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

This analysis can lead to improved decision making, increased profitability and help level the playing field to a certain extent.

- Real estate companies can gain a competitive edge by leveraging accurate price predictions to position their properties effectively in the market. By pricing properties competitively, they can attract more buyers and secure higher sales volumes.
- Real estate agents and sellers can accurately estimate the sales price of individual residential properties, enabling them to set listing prices that reflect market demand and maximize profitability. Moreover, agents can better explain the value of a property to clients by referring to the model.
- Investors and property developers can make informed decisions about buying or selling residential properties by predicting their future sales prices. This enables them to optimize their investment portfolios, minimize risks, and maximize returns on investment

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

- Data cleaning by 02/25
    - Real estate data
- Data exploration by 02/25
    - Real estate data
    - Google Trends
- Model creation by 03/03
    - Factor-based
    - Combined model
- Back testing by 03/10
    - Factor-based
    - Combined model
- Output analysis by 03/10
    - Statistical
    - Write up
- Progress report ready 3/17
- Final submission ready 4/14

REFERENCES

Choi, H., & Varian, H. (2012). Predicting the Present wwith Google Trends. *Economic Record*, 2-9.

Rosen, P. (2023, 08 11). *Business Insider*. Retrieved from https://markets.businessinsider.com/news/commodities/housing-market-inventory-shortage-home-prices-value-real-estate-property-2023-8

Veldhuizen, S. v., Vogt, B., & Vogt, B. (206). Internet searches and transactions on the Dutch housing market. *Applied Economics Letters*, 1321-1324.

Wu, L., & Brynjolfsson, E. (2015). The Future of Prediction: How Google Search Foreshadow Housing Prices and Sales. In A. Goldfarb, S. M. Greenstein, & C. E. Tucker, *Economic*

Analysis of the Digital Economy (pp. 89-118). Chicago: National Bureau of Economic Research.