Housing Price Forecasting Engine

Blake Bormes, Ben Ohno, Ben Mok, Jonathan Moges, Don Irwin

WHY

Problem Statement

- **Describe the problem you are solving:**

  Housing is the single largest line item on most working, and middle class Americans' monthly budgets.

  Buying a home at the wrong time within the housing market cycle can expose an individual or family to unrecoverable financial loss.

  For example, an individual buying a condo in Los Angeles in the beginning half of 2007, would have had to wait until approximately 2017 to simply break-even on their purchase price. Selling before that 10 year period elapsed would have necessitated realizing a loss – i.e. losing a down payment, or potentially paying out of pocket the balance on an "upside down" mortgage, or walking away from a house, causing a bad mark on their credit score.

  By contrast, an individual who bought a condo in 2012, in the same Los Angeles market, would have realized home value appreciation of approximately 40-60% by 2017.

  One may be tempted to believe that the cycle above was or is unique to the 2008 housing crisis. However, this cycle repeats itself.   A home buyer in Los Angeles' San Fernando Valley who purchased a house in 1989, would have needed to wait close to a decade to break even on their purchase price.

  Many individuals do not have the staying power, or will experience a life event which can cause a forced sale of a home during a down market.

  We aim to provide buyers and investors with a tool which will indicate the "direction of travel" within a given market which will provide a recommendation as to whether a purchase is prudent, or renting is more prudent.

- **Why is this a compelling and impactful problem to solve? Why is this a big opportunity?**

  For most Americans purchasing a house is one of, if not, the, largest financial investment and commitment that they will make.

  Housing is not only important to the individual, but it is disproportionately important to the American economy. Spending on housing and housing services accounts for approximately 13-19% of yearly GDP in the United States.

  Since records have been kept, there have been several recessions which were not accompanied by housing market declines. That is to say while the economy was in recession the housing market did not decline.

  By contrast, every sustained decline in the housing market has been accompanied by a recession.

  Put simply; we have had recessions without housing market declines, but we have not had housing market declines without recessions.

  No tool can prevent housing market downturns. However, our tool aims to protect home buyers and investors from the consequences, or opportunity cost of buying in a market trending lower, or neglecting to buy in a market trending higher.

- **What assumptions are you making about the problem / opportunity? (these assumptions would directly influence the key elements and features you would build in the minimal viable product (MVP)**

  We accept Yale professor, and Nobel laureate, Robert Schiller’s work which suggests that the housing market is cyclical.

  Regional housing price data supports our key assumption that while the housing market as a whole is cyclical, the scale, and in some cases frequency is unique to regions.

  Put in plain terms; some markets are more volatile. Some markets’ price appreciations are more drastic. In many cases their price retreats are equally dramatic.

We assume basic economic theory such as supply and demand, apply.

We assume that consumers have finite monthly housing budgets, and an monthly upper price threshold which is static. From this assumption, it necessarily follows that there exists an inverse relationship between interest rates and home prices.

A consumer with a monthly budget of $1,500, making a 20% downpayment, can afford a home with a price of $385,000 if the rate on a 30 year fixed mortgage is 2.5%.

The same consumer, with the same monthly budget of $1,500 can only afford a home with a price of $254,000 if the rate on a 30 year fixed mortgage is 6%.

The assumptions above provide us with requirements for our data. Our data must be regional, we must have data about supply (total listings) relative to demand (total population), within a region.

Our assumptions also provide us with requirements of our model, or models.

Our assumptions also point to an educational requirement which our project must fulfill.

Our minimum viable product must contain data to make forecasts and predictions, and it must provide an interface which explains to the user, in simple terms, these economic model drivers and guides them in the use of the product.

Impact and market opportunity

- **How would you go about quantifying impact, market opportunity for this particular problem you are trying to solve by building a data science product?**

  There are three market opportunities for this product:

  The individual market opportunity; An individual seeking to purchase a home or to rent a home can consult the product for a recommendation, forecast and rationalization of that forecast.

The investor market opportunity; An investor seeking to build a rental portfolio, or utilize residential real estate as an investment vehicle, may utilize this product to assist in their decision making process about capital allocations in residential real estate.

The developer or institution market opportunity. Software developers, land developers, and institutions may seek to utilize a form of this application for the following purposes:

1. Use of our API endpoints for integration into larger financial applications, web applications, forecasting or prediction engines, blogs, etc. ...

2. Planning building projects in specific regions.

3. Parts of the application could be licensed to financial institutions for use within their existing systems.

- **Market Size (How big is this market, based on your research?)**

  On average from 2012 until 2022 about 6 million homes are sold per year.

  It is difficult to generalize what the homebuying process is. It varies from individual to individual. According to the NAR the average buyer looks at at least six homes in person before buying.

Target Customer and user/customer discovery

Who is the primary customer/user? What is the use case? What are key assumptions you are making about the primary customer/ user and use case?

- **Identify targeted user/customer segment for the MVP.**

  The primary customer is an individual or family, between the ages of 25 and 65, who are looking to either purchase a home or move to a new city.

  The customer is at a decision point about whether to buy, or rent, and potentially what area they intend to buy or rent in.

- **Define the primary use case validated by this target user/customer.**

  User consults the application to decide whether to buy in a given county in the US.

- **What might be other key assumptions that are important to validate with the target user/customer?**

  The user's primary motivation for purchasing a property: to live in, to invest. The length of time a user intends to carry the property.

- **How would you go about validating these additional key assumptions?**

  Conducting interviews with subject matter experts.

- **Who will you contact to conduct initial user research and feedback?**

  We will contact Danielle Hale, the primary economist at realtor.com and a former colleague of one of the team members.

- **What is the user journey and UI/UX for this data product?**

  1. User is presented with splash screen which explains the core concepts.
  2. User selects county from drop-down.
  3. County-specific information is displayed (e.g., average house price, average 2B rental, average 1B rental price) along with Information "Prices likely to increase", "Prices likely to remain static", "Prices likely to decrease".
  4. In an advanced user option; a user may change core features of county information I.E. unemployment rate, median income, etc. … and re-engage the forecast engine.

Market Landscape / Competitive Landscape / Existing companies solving the same / similar problem

Who are the major players and main vendors in the space? What are the existing solutions?

Company Name	**Stage (startup, enterprise)**	**Product / Solution overview**	**Who is the primary customer?**	**Key differentiation vs your proposal (based on your understanding/ research)**
Zillow	**Enterprise**	**Market overview: hot/cold. Buyer / Sellers.**	**Property searcher.**	**No direction of travel indicator. No macroeconomic considerations.**
Realtor.com	**Enterprise**	**Zipcode facts**	**Property searcher**	**No quick facts about the market, no direction of travel indicator, not macroeconomic considerations.**
Corelogic	**Enterprise**	**Case Schiller HPI.**	**Subscriber.**	**No user-accessible UI, no county macroeconomic consideratioins.**

Relevant readings, market research, white papers, academic research (share title and link)

- **ARIMAX**: https://pyflux.readthedocs.io/en/latest/arimax.html
- **Multivariate TS/LSTM (torch)**: https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/
- **BDLP**: https://towardsdatascience.com/forecasting-with-bayesian-dynamic-generalized-linear-models-in-python-865587fbaf90
- **Pooled OLS:** https://timeseriesreasoning.com/contents/pooled-ols-regression-models-for-panel-data-sets/
- **VAR model (extension of ARIMA):** https://towardsdatascience.com/prediction-task-with-multivariate-timeseries-and-var-model-47003f629f9

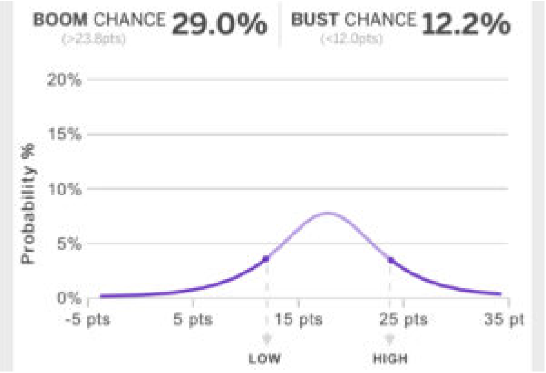
Research:

- Optimal buying and selling of a home (Lorig, Suaysom) https://arxiv.org/pdf/2203.05545.pdf
- Housing Market Forecasting using Home Showing Events (Zha, Parker, Foster, Sokolov) https://arxiv.org/pdf/2201.04003.pdf
- House Price Modeling over Heterogeneous Regions with Hierarchical Spatial Functional Analysis (Liu, Mavrin, Niu, Kong) https://arxiv.org/pdf/1803.00919.pdf

WHAT

Minimal Viable Product (MVP)

- **What is the minimal viable product that you are building that will specifically test the fundamental assumptions you have about the problem and the value of your solution?**
  - What are the main features and why?
    - A frontend website that a user can easily navigate. A backend ML engine that calls different models and chooses the ‘best’ (closest) forecast
    - Allowing the user to input data that may not be readily available on many other platforms that also perform housing prediction (unemployment rates, going/average interest rates)
    - Outputting a nice visualization, perhaps a curve like
    - Or this visualization feeds the red yellow green traffic light



  - **What is the value delivered to your user/customer?**
    - A transparent forecasting of housing price
    - The ability to tell a user **when** to buy

        - A forecast based on many other factors other than time and seasonality
    - **What is the key question or questions (max 3) that your target user/customer will be able to answer using your Capstone product?**
        - Should I buy in this neighborhood?
        - Should I buy in this neighborhood given these unemployment rates in this area?
        - Should I buy at this time given external factors that are independent of a home value?
- **What data science approach would you intend to use for the MVP? (this is NOT UI / UX but technical discussion)**
    - EDA to determine variables in models
    - Model ideation: **ARIMAX, Pooled OLS, LSTM, Bayesian Dynamic General Linear Models**
    - ML prediction engine build: FastAPI
    - Data Viz (MPL, streamlit)
- **How would you potentially test the efficacy of the MVP? When would you start testing?**
    - Have potential homebuyers or non-potential (control) homebuyers test the application and website to see its performance and use when compared to homes they would like to buy given by websites like Zillow, Redfin

What is the key **differentiation** between your MVP and the existing solutions and/or approaches?

- Our MVP allows potential homebuyers to consult a tool prior to buying a home. The purpose of this tool is to tell a homebuyer if it is the right time to buy this property. Currently, there are no such tools available on the market. The closest prodcutionized existing solution is the Zillow zestimate, which has proved to be unreliable[1] in current market conditions.

Value Proposition (what value/utility does your project/product provide to your intended users?)

- **State the value that your MVP brings to the target customer segment.**

    - An indication of whether prices are likely to increase or decrease in a given market.

    - The ability for a user to modify features of the forecast in order to explore “What if” scenarios:

    For Example:

        - Increase / decrease the unemployment rate in their market.
        - Increase / decrease listing count in a given market.
        - Increase / decrease interest rate.

    - The ability to compare two different markets

- The value proposition should indicate why your solution is better and/or more differentiated.
  At the present time there does not exist a good prediction engine which is publicly facing. Our product auto self-differentiates because of this.

Mission Statement

Purchasing a residential property is the single largest financial decision most working and middle class Americans will make.

Purchasing at a time when housing prices are likely to increase within a specific market, may be a prudent financial decision that rewards homebuyers with property appreciation and good options if and when they need to move.

Purchasing at a time when housing prices are likely to decline within a specific market, exposes buyers to financial loss, and lack of mobility.

Our prediction engine will give users insight into whether or not a market in which they are considering purchasing a home, is likely to experience price appreciation or price depreciation.

Our prediction engine provides a user three simple feedback prompts “prices likely to increase”, “prices likely to decrease”, “prices likely to remain the same”. Providing these prompts outperforms rival solutions within the online real estate space which provide the mean value within a zipcode but no prediction of the market’s “direction of travel”.

HOW

Data sets

- What datasets do you intend to use?
  - Unemployment data from the Federal Reserve Economic Data (FRED) - source
  - Realtor.com monthly inventory data - source
  - Fannie Mae lending data - source
  - Freddie Mac lending data - source
  - Bureau of labor statistics - source

- Are the datasets public?
  - All datasets we’re considering at this time are publicly available
- What are the datasets attributes / metadata that could make the exploratory data analysis easier / harder?

    - Some of the Bureau of labor statistics/Freddie Mac/Fannie Mae datasets have treat county name differently. This will affect data integration because county name and year/month are primary keys used to integrate the datasets.
    - Inflation data will prove to be especially difficult due to the way inflation data is reported and organized. Inflation data is faceted and reported in many different ways, depending on the modeling team's needs, we'll need to mung/integrate multiple inflation sources that are aggregated and reported at different levels (i.e. national median CPI, regional personal consumption expenditures)

Project Management

- What is the role of each member (who will do what specifically)? See below!
- Who is the project manager and chief facilitator (and tie breaker)? - Don/Blake, tie breaker (Ben M)
- Who is the resident SME? - Don, previous experience with real estate markets and macroeconomic data from FRED
- Who is the product manager? - Don/Blake
- Who is the lead on infrastructure and data engineering? Don/Ben M. for infrastructure, Blake/Jonathan for data engineering
- Who is the lead on EDA? - Blake/Jonathan
- Who is the lead on model evaluation? Ben O./Ben M.
- Who is the lead machine learning engineer? Ben O./Ben M.
- Who is the lead MVP application developer? Don/Ben M. (assuming MVP app is the web app)
- Who are the backups to key roles?

  Each task has one other person assigned to it.

- Data Modeling/Training: Ben M/Ben O
- Data Engineering: Blake/Jonathan/Don
- Data Visualization: Ben O/Jonathan
- EDA: Blake/Jonathan
- ML API: Ben M/Ben O
- PM: Don/Blake
- Web App: Don/Ben M

	Role and responsibilities (immediate)	Role and responsibilities (long term) / Alternate or additional role / pair *subject to change as project dictates
Ben M.	Data Modeling/Training	Web app/ml api backend
Ben O.	Data Modeling/Training	Ml api/data visualization
Blake	EDA/Data engineering	PM/Project Deliverables
Don	PM/Data engineering/Web App	PM/Web app
Jonathan	EDA/Data engineering	Data visualization/Project deliverables

**Some teams have used pairing principles to assign two members of the team to main tasks.

- What are the strengths and weaknesses of each team member?

	Strengths	Weaknesses
Ben M.	MLE, backend, fastAPI	Front end
Ben O.	ML, fastAPI, data visualization	Slide decks
Blake	PM experience from consulting, data visualization, EDA	Web app, fastAPI
Don	Full stack, tech lead experience	
Jonathan	Data visualization, presentation to stakeholders, EDA	Front end, modeling

- Submit the Team Process Agreement.

Technical Approach and Planning

- What methodologies would you use for initial data exploratory analysis to ensure your datasets are sufficient and meaningful?
    - Pandas profiling - package in pandas for nice EDA
    - Determining data coverage - finding which counties/FIP codes we are missing information for
    - Data ranges/Distributions - determine that there is variability in the data and values for interest rate, vacancies, unemployment, etc. are within reason
- What data science algorithms are you intending to develop and build for the project? What challenges do you potentially foresee?
    - Data science algorithms:
        - LSTM neural network model for time series forecasting
        - ARIMAX model for time series forecasting
        - Pooled OLS model for time series forecasting
        - Bayesian dynamic linear model for time series forecasting
        - Fitting normal distribution to data to compare past home prices to forecasted home price from previous models
        - Not quite a model algorithm, but we do need scripts to scrape data from FRED and merge data together by FIP/county name

    - Challenges:
        - There's a possibility that our model doesn't have very much accuracy. Modeling on this kind of data hasn't been done before
        - Compute power for the neural network model (we have access to a machine with GPU thanks to Don)
        - Deploying our model/web app
- What help do you need?
    - We'll probably need assistance with deploying our app in AWS
    - AWS creds :]

Note: there is a weekly team check in at the beginning of each class starting week 4 and during non-presentation weeks. The team check-in is usually 3-5 min. Please cover: major milestone(s) achieved in the past week, major milestones to be achieved this week. One key learning to share with the class. And help needed.