Team REX Technical report

Preston Ching, Marcel Hedman, Nam Luu, Owen Schafer
Supervisor: Zona Kostic

March 2021

1 Introduction

The $9.6 trillion (as of 2019) global real estate market is one of the least technology disrupted markets today. Connections, transactions and relationships in residential real estate are often forged through traditional stakeholders and human interactions. Our partner organization, REX, leverages the power of machine learning and big data to match home owners and buyers, seeking to create a platform that significantly reduces transaction fees. Amongst the various approaches to reduce information asymmetry, REX intends to establish a more robust and fine-grained real estate index for realtors and property investors. Current solutions include price indices from the National Association of Realtors (NAR) and automated valuation machines (AVMs) implemented in Zillow or Google. However, these price indices are usually only based on state-level or city-level metrics, and do not provide clarity into more fine-grained submarkets. Moreover, AVMs ascertaining the spot price of residential real estate are noisy metrics mired with varying bid-ask spreads and mortgage approvals. Prices from sale transactions are also considered second-order effects since they are reflecting the equilibrium of supply and demand dynamics. Therefore, this project aims to create a real estate index for predicting supply and demand via home listings and sale transactions.

Instead of census-tract based submarkets, this project will explore a hierarchical approach to more accurately and effectively identify submarkets which can be a hybrid of geographical (proximity to schools, business districts), environmental (crime rates, demographics) and physical (size, apartment view) factors. To better inform realtors and property investors, one of the key goals is for our model(s) to be interpretable and allow insights into the various submarket classifications and properties. Nevertheless, there are several challenges in implementing a hierarchical model for a real estate index. Firstly, there is no clear definition of supply and demand target variables as a function of home listings and sale transactions as proxy variables. Apart from using the number of listed properties and sale transactions within a particular discretized time period, we have also explored using the number of days on market as a better proxy for demand. Secondly, the Bayesian hierarchical model involves fixing k number of submarkets which implicitly requires potentially faulty assumptions on submarket classification types. In terms of defining success metrics, we intend to use traditional machine learning metrics of accuracy and AUC on test datasets as well as the ease of identifying submarkets based on similar attributes within a particular model-generated submarket. Moreover, we will benchmark the efficacy of the hierarchical model against traditional machine learning algorithms and potentially last semester's capstone project on Denver listings and transactions, instead of other current price index models, for a more robust comparison.

2 Literature Review

There are two key pieces of literature that motivate aspects of our model:

1. Previous work towards building a more granular index of real estate supply and demand was developed by the AC297R REX team from the Fall 2020 semester [1]. The model described in their work predicts a number of listings and sales over a specified time-frame at a market level, and then predicts these same quantities within predefined submarkets (census tracts) by scaling their market prediction by the mean ratio of sale/listing rates within a particular submarket to those of the market as a whole. This approach can be viewed as "top-down", in that predictions made at a more granular submarket-level are generated in part from the model's previously determined market-level prediction. Our work differs in that it is a "bottom-up" approach, aiming to first predict at a submarket level and then generate an aggregate prediction of supply and demand at the market level. Both projects are conducted alongside our partner REX, so in order to provide an easier comparison between these two different approaches, we conduct our work using the same predictive domain (the Denver metro area) with the goal of predicting identical proxies for supply and demand (number of listings/sales).
2. Liu et. al. [2] introduce a Bayesian framework for modeling real estate valuation using hierarchical clustering to capture distinct trends between learned submarkets. We adopt a similar hierarchical structure in our work, forecasting supply and demand (through quantities described later in this document) in place of a property's monetary value. Our model simplifies the structure presented in Liu et. al. (e.g. opting for a linear hedonic supply/demand model over a Gradient Boosting Tree hedonic price model) in order to allow a reasonable adaptation for time series data.

3 Data

In this study, we use three sets of data collected from multiple listing services (MLS), REX and the Census Bureau.

The first MLS dataset, compiled from numerous cooperating real estate brokers, provides information on home listings and sale transaction data for our target variables. This dataset contains 439,427 observation points, which have the transaction data of 399,883 unique properties in the greater Denver area from 03/2016 to 11/2020. For each transaction, we have the list date, sale date, withdrawn date, expired date, status (whether the property is sold, expired or withdrawn), the property's zip code and sale price. In this study we only focus on the properties that are sold, hence we remove all properties that have expired or withdrawn status, and we reduce our dataset to 327,350 data points.

The second dataset from REX provides more information about the profile of each property. This dataset provides an additional 70 features, both numeric and categorical, such as the square footage area, the number of rooms and presence of a garage amongst others. However, out of these 70 features, only 12 features have more than 80% non-missing data in the dataset. Some features that we originally deem important, such as the number of bedrooms and bathrooms, only have 30% non-missing in the dataset. To allow a more complete and robust analysis, we remove features that are missing a significant number of values in the dataset.

Other than datasets for transactions and house profiles, we believe that the socio-economical environments of the residential community where the houses are located are also important indicators for supply and demand, thereby augmenting our analysis with the census dataset. This dataset is a combination of data from NeighborhoodScout, a database of US neighborhood analytics, and other public census records, generating an additional 122 features containing information on city-level demographics as well as census-level metrics on friendliness to college students, professionals and transportation score amongst others. After merging all 3 datasets based on property identification numbers and removing duplicate entries, we obtain a final dataset of 308,698 unique properties and 192 attributes.

4 Exploratory Data Analysis

4.1 Data preparation

Since we have a dataset that is rich in features, we need to trim down the number of attributes. We first handpicked the features that we think are good indicators for our model. Of the remaining features, we removed those that have too many missing values, or have very low variance. We also modified the *sale price* attribute to *sale price per sqft*, to remove the correlation between the area of a property to its price. This way, we shortened our list of features to the following, which could be classified into 2 categories:

- House profile features: *zipcode*, *area*, *sale price per sqft*, *property type*, *has central air*, *pool*
- Residential community features:*farm score*, *median rental price*, *population density*, *home buyer score*, *retirement friendly score*, *young single professional score*, *college student friendly score*, *violent crime rate*, *walk score*, *transportation score*, *carpool score*, *densely urban score*, *suburban score*, *rural score*, *luxury communities score*, *farm score*

These features all have numerical or boolean/categorical values except for property type. There are 6 property types (single, multi-family, condominium, land, townhouse and other) and we map each type to a number from 0 to 5.

When we plot the number of listings per week each year (figure 1), we observed that the number of listings in the area follow almost the same pattern every year: the number of listings gradually increased and peak in the summer before slowing down towards the end of the year. Therefore, to account for the seasonal effect in the real estate market, we created an additional temporal feature, that records the average days on market of other properties in the same communities in the past 30 days. We called this dependent variable *mean_dom*. The intuition behind this feature is that houses that are temporally sold near each other may have similar DOM values, since they share similar market conditions.

Figure 5 show the correlation matrix of these features. After examining the correlation matrix, we further removed the following features: *transportation score*, *carpool score*, *densely urban score*, *suburban score*, *rural score*, *luxury communities score*, *farm score*, *bedrooms*, *bathrooms*, because they strongly correlated with other existing attributes in the dataset. The decision to remove the number of bedrooms and bathrooms is also partially because they have a significant amount of missing data.

To deal with outliers, we first plot the geographical distribution of all transactions (figure 3). We observe that the dataset we received not only contain data about the Denver area, but also from other

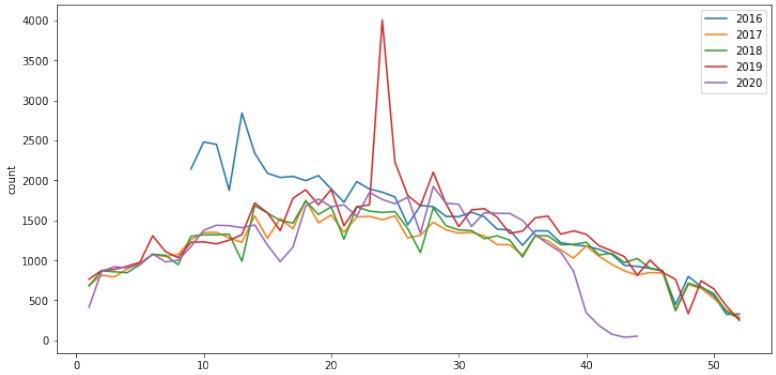

Figure 1: Number of listing each year

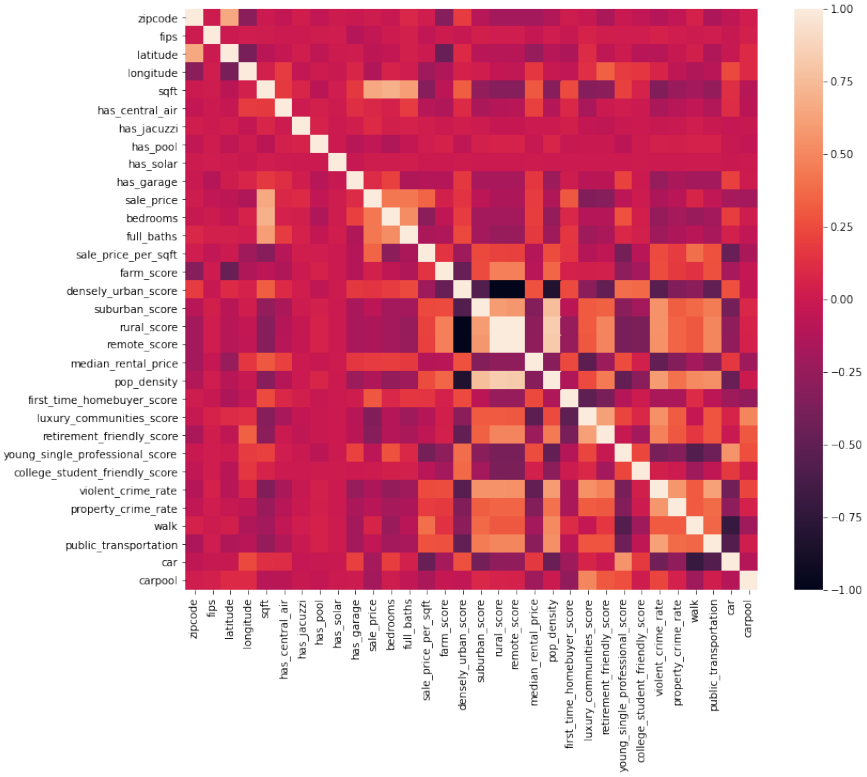

Figure 2

nearby cities as well. Since we only focus on the supply and demand of the Denver metropolitan area, we only consider properties that have latitude inside the range of $[39.5, 40]$ and longitude in $[-105.25, -105.55]$

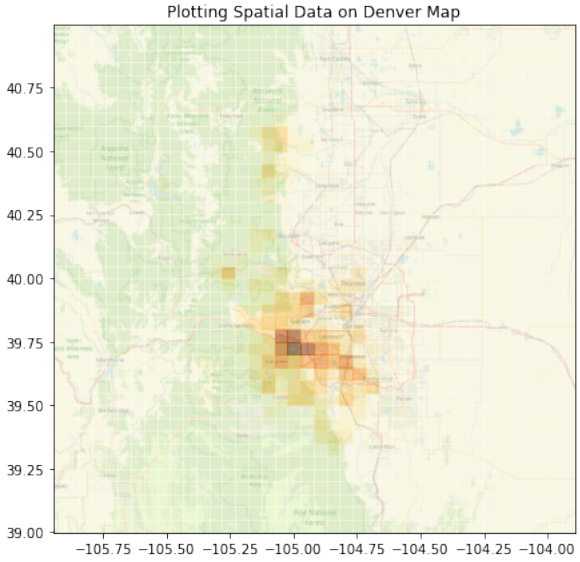

Figure 3: Geographical distribution of data

We also detected outliers by plotting the distribution of each features. In figure 4, for example, when plotting the distribution of days on market, we decided to remove the datapoint that have days-on-market equal to 0, or list date and sale date are the same, because we speculate that there are some agreements being made before the houses were listed, and the numbers do not truly reflect the days-on-market of the properties. The visualization of other features' distributions can be found in our GitHub repository.

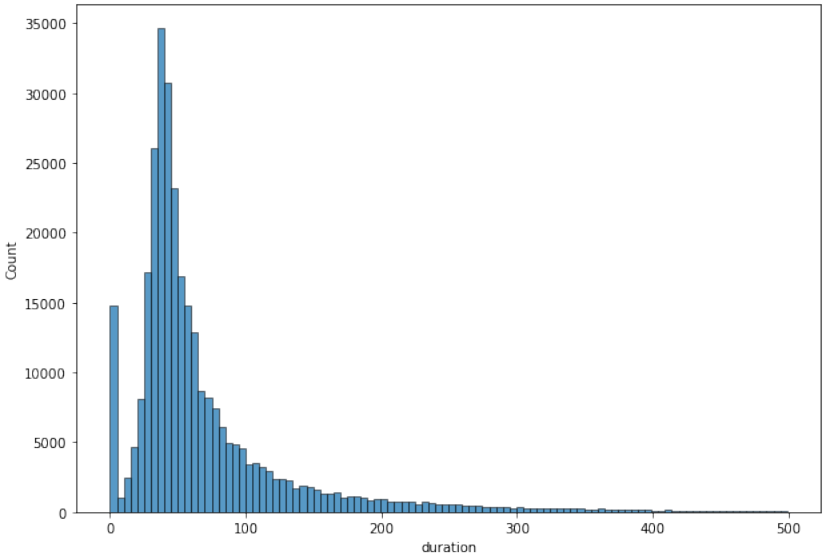

Figure 4: Distribution of days-on-market

4.2 Target variables

We create 4 target variables for our prediction models. Since we want to predict the supply and demand of an individual property, we construct two binary values to indicate whether the property will be listed or sold within a fixed time period. For the baseline models, the time period that we chose is April 2019.

While building the baseline model, we realized that these two binary values may not be sufficient to produce a reasonable index for supply and demand. The reason is that the number of houses being listed/sold during a particular period is very small compare to the entire dataset, thus the models can achieve high accuracy by simply returning a constant value. Thus, we introduced a new target variables: *DOM* (shorthand for days-on-market). This variable was calculated by subtracting the list date from the sale date. However, our regression model struggled to return a reasonable result to predict *DOM*, thus we construct a slightly easier version, by discretizing *DOM* into 4 buckets, to indicate whether a listing property will be sold in 1,2,3 or more than 3 months. The reasoning for this selection that most properties are sold within 90 days from the time they are listed, which is showed in figure 4. Using this discretization, we can build a classification model to predict *DOM*.

5 Baseline model

Armed with the above prepared data, we developed our baseline models. However, due to constraints on the data readily available, our models currently make predictions purely on demand. This will be iterated in the future to additionally predict on supply.

As defined in section 4.2, we have both the binary target variable defining whether a house was sold in a time interval and we also have the *DOM* target variable. We built baseline models for both of these variables. It is worth noting that neither of these baseline approaches incorporates any submarket effects.

5.1 Time interval approach

In this approach, we predicted the probability that a house would be sold, given that it was listed within a 90 days time interval. The time interval we predicted on in our baseline was the 90 days following April 2019. We used a logistic regression model to obtain probabilities and used 50% as the threshold for classification.

5.1.1 Results

Model	Test Accuracy	Test AUC
Logistic regression	73%	0.54

Table 1: Predicting demand with time interval approach

This approach returns an artificially high accuracy as when taking into the account the AUC score we realise that there are many false classifications being made. This high accuracy and low AUC is an indication that the data is imbalanced and we intend to explore methods to resolve this in a way that can also be expanded to multiple time periods.

Alongside the probabilities at a per-home level, we also agglomerated the classifications to get a predicted number of total homes sold in the time period. We found that the ratio of predicted number of sold homes to actual sold homes in the 90 day period was just 0.17. As we address the imbalanced class problem, we expect that the low ratio will be resolved as well.

5.2 Discretized DOM

In this case, the proxy for demand is the discretized DOM target variable. We chose three different models to use for the baseline (as described in table 2) and made no attempt to tune, or regularise the models as the aim was to provide a simple baseline to act as a benchmark.

5.2.1 Results

Model	Test Accuracy	Test AUC
Logistic regression	63%	0.57
Random Forest	63%	0.59
Extra Trees Classifier	55%	0.55

Table 2: Predicting DOM

Logistic regression and Random Forest classifier performed comparably on the test set and Extra Trees Classifier performed poorly on the test set. Across all of the models we see that the test AUC scores are low and this is likely due to imbalance in the dataset. *(We used the OVR method with the weighted average score to find the AUC.)* As we iterate our version of the baseline model we will do more to address class imbalance in the dataset perhaps using methods such as SMOTE (SMOTE is a method of generating synthetic data of minority classes to resolve class imbalance).

5.3 Key Insights

From the baseline models, we achieve mediocre performance. The major drawback beyond model performance is that the models cannot flexibly cluster houses into sub-markets which means they all follow the same average relationship to the target variable. Furthermore, for the tree based predictors we can't get a true sense of the relationship between the predictors and target variable. We hope to address this with the Bayesian approach in the developed model.

It should also be noted that one of the reasons we chose to explore Denver was because a previous Capstone project team had attempted to forecast number of sales and number of listings in this region. Working in the same area means that we can also use their model as a comparison [1].

6 Developed model

Our developed model attempts to improve upon our baseline model in the following ways:

- We aim to have our model learn to define optimal submarket classifications on its own, rather than be limited by a predefined clustering of the data for each home.

- We aim to be able to interpret the hedonic supply/demand model used within each submarket in order to understand the relative importance and effect on supply/demand of each feature of a given home.

In order to accomplish these goals, we introduce a Bayesian hierarchical model where the submarket classification of a home into one of K submarkets is a latent variable, and each predicted quantity for that home is distributed according to a linear combination of its features as determined by its submarket classification. That is, for a particular draw $n \in \{1, ..., N\}$ from our hierarchical structure (corresponding to a single home), our model assumes the following structure and dependencies amongst variables (note that like our baseline model, this model is currently structured for prediction within a single specified time interval):

1. $z_n \sim \text{Cat}(\pi)$: This variable represents the submarket classification of a home. The parameter $\pi \in \mathbb{R}^K$ represents the prior probabilities of a home belonging to each of the K submarkets, where $\pi_k = 1/K$ for $k \in \{1 \ \dots \ K\}$.

2. $\beta_{i,n} \sim \mathcal{N}(\mu_{\beta_i,z_n}, \Sigma_{\beta_i}),\ i \in \{1, 2\}$: These random vectors represent the weights of each home feature used in the separate hedonic models of supply ($i = 1$) and demand ($i = 2$). These weights are drawn from two of $2K$ total distinct distributions (for each of supply and demand for each submarket), and each distribution is parameterized by μ_{β_i,z_n} and Σ_{β_i}. Note that for these distributions and each of the following distributions in the hierarchy, we are currently making the simplifying assumption that prior (co)variance parameters are identical across all submarkets.

3. $\mathbf{h}_n \sim \mathcal{N}(\mu_{h,z_n}, \Sigma_h)$: These random vectors represent the features of each home. These features are drawn from one of K total distinct distributions (for each submarket), and each distribution is parameterized by μ_{h,z_n} and Σ_h. Note that we currently assume a multivariate Gaussian distribution of home features, some of which may be discrete or boolean/categorical.

4. $y_{i,n} = \sigma(x_{i,n}),\ x_{i,n} \sim \mathcal{N}(\beta_{i,n}^T \mathbf{h}_n, \Sigma_{y_i}),\ i \in \{1, 2\}$: The variable $y_{1,n}$ represents the probability that a home is either currently listed at the start of the specified time interval or will be listed during this interval. The variable $y_{2,n}$ represents the probability that a home will be sold during this interval, given that it is listed. Each of these variables is a function of another random variable $x_{i,n}$ with variance Σ_{y_i} and mean given by our linear hedonic supply/demand model $\beta_{i,n}^T \mathbf{h}_n$. The sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ is used to ensure that the predicted quantities $y_{i,n}$ can be interpreted as probabilities.

To illustrate, the relationships between each variable in the hierarchy are displayed graphically in figure 5. We are currently implementing this model using PyMC3. We expect to complete a functioning beta of this model and begin analysis and tuning in the coming days.

7 Further work

Our future work ahead of the next project milestone will be focused on the development and exploration of our developed model. Some key goals that we will work towards include:

- REX has provided a new dataset that contains information about more properties, including those that have not been listed during the observation period of our current dataset. We plan to explore

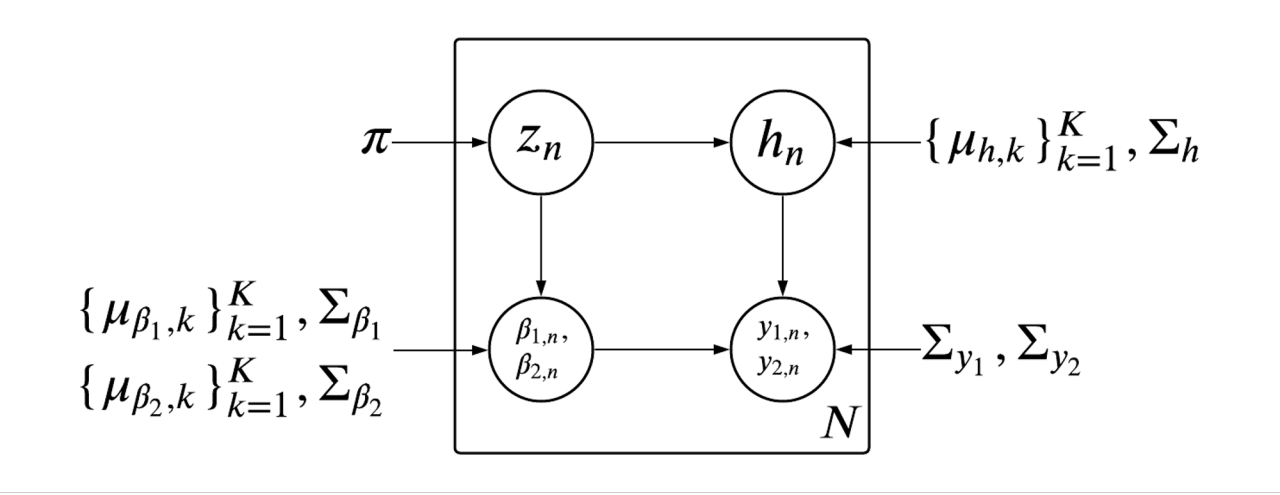

Figure 5: Proposed Hierarchical Model

this new data and merge it with our current home features datset, such that our training data is representative of the entire set of homes in the Denver metro area.

- Our exploratory data analysis revealed that the distribution of DOM values is highly skewed and is likely a manipulated figure. We plan to explore ways to determine a true DOM value for each home, and provide further EDA.
- As there is significant class imbalance within our data (e.g. more homes can be classified as ”not sold” than ”sold” within a particular time frame), our baseline models obtain an artificially high accuracy with poor AUC scores. We will explore the use of SMOTE for generating synthetic data to address this class imbalance.
- We aim to complete the implementation of the developed model described above, and evaluate its performance. To be concrete, we will focus on the model's predictive capabilities within a single time step, and will begin relaxing the distributional assumptions made in the above description of the developed model. We will also explore ways to augment our model with macroeconomic factors, and will determine the effective impact of this augmentation.
- As a stretch goal, we hope to begin our adaptation of the model for time series data. This process will begin once once we are satisfied with our model's structure and performance within the context of a fixed time frame.

References

[1] Will Fried, Jessica Wijaya, Shucheng Yan and Yixuan Di, ”Towards a Revamped Real Estate Index Towards a Revamped Real Estate Index”, Accessed 2021, https://towardsdatascience.com/towards-a-revamped-real-estate-index-c48ae27b33c5

[2] Z. Liu, J. Cao, R. Xie, J. Yang and Q. Wang, ”Modeling Submarket Effect for Real Estate Hedonic

Valuation: A Probabilistic Approach," in IEEE Transactions on Knowledge and Data Engineering, Accessed 2021, doi: 10.1109/TKDE.2020.3010548.