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. Due to data availability, modelling challenges and ease of comparison with the analysis conducted by the IACS Team in Fall 2020, the focus will be on predicting demand for homes in the Denver market. Subsequent sections will expound the difficulties related to constructing a model for supply prediction.

Instead of census-tract based submarkets (census tracts are more granular than zipcodes and are strictly defined geographical demarcation), this project will explore a Bayesian 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 ROC-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 Bayesian hierarchical model against non-Bayesian baseline models, ranging from models without submarkets and those incorporating submarkets.

2 Literature Review

Firstly, most real estate indices are focused on prices, ranging from stock market performance of real estate companies and appraisal-based indices analyzing returns on real estate institutional investor portfolios to transaction-based prices indices. The more reliable price indices are based on actual transaction prices, such as in the Paris Housing Market by Maurer et. al. (2004) [1] and Gourieroux et. al. (2006) [2]. Instead of naive price indices based on mean or median prices, these transaction-based indices are constructed from hedonic price models with data availability and consistency in the French markets. The hedonic assumption here refers to home prices being predicted fully from the individual homes' attributes, shown to be plausibly fallible based on our preliminary developed models. Another hedonic price index by Jiang et. al. (2014) [3] in the Singapore residential real estate market uses data on single-sale and repeat-sale properties, outperforming the SP Case-Shiller price index for single-sale properties. Nevertheless, this is predicting on a city or national level rather than providing information on different submarkets.

There are two key pieces of literature that motivate aspects of our model. Previous work towards building a more granular index of real estate supply and demand was developed by the IACS REX team from the Fall 2020 semester [4]. 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 geographical 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. This can also potentially be construed as being similar to a fixed-effects regression model where census-tracts demarcation provide valuable and statistically significant information in predicting demand and supply. As mentioned earlier, we intend to test our models using the same predictive domain (the Denver metro area) with the goal of predicting identical proxies for demand and supply (number of listings/sales) for ease of comparison.

Liu et. al. [5] introduce a Bayesian framework for modeling real estate valuation using hierarchical clustering to capture distinct trends between learned submarkets. This approach can be viewed as 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. One major assumption in this Bayesian hierarchical framework is also the hedonic assumption where prices can be fully predicted based on the individual homes' attributes. Moreover, this model seeks to predict noisy price signals instead of demand and supply.

Therefore, our proposed developed model leverages the strengths from these two key literature by using a Bayesian hierarchical approach to forecast demand and supply (through proxies described in subsequent sections) in place of a property's monetary value. Our model also simplifies the structure presented in Liu et. al. (e.g. opting for a linear hedonic demand/supply model over a Gradient Boosting Tree hedonic price model) in order to allow a reasonable future 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.

The second datasource is REX, which provides 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 garage amongst others. This dataset has information of 200,000 listing properties and 1,200,000 non-listing properties.

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.

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

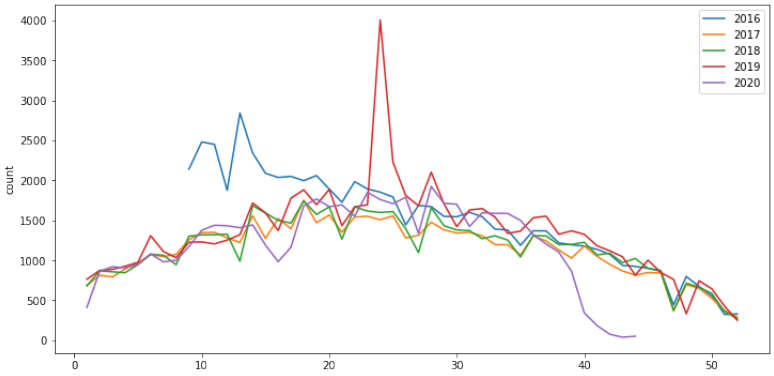

Figure 1: Number of listing each year

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 6 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.

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 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 $[-105.25, -105.55]$

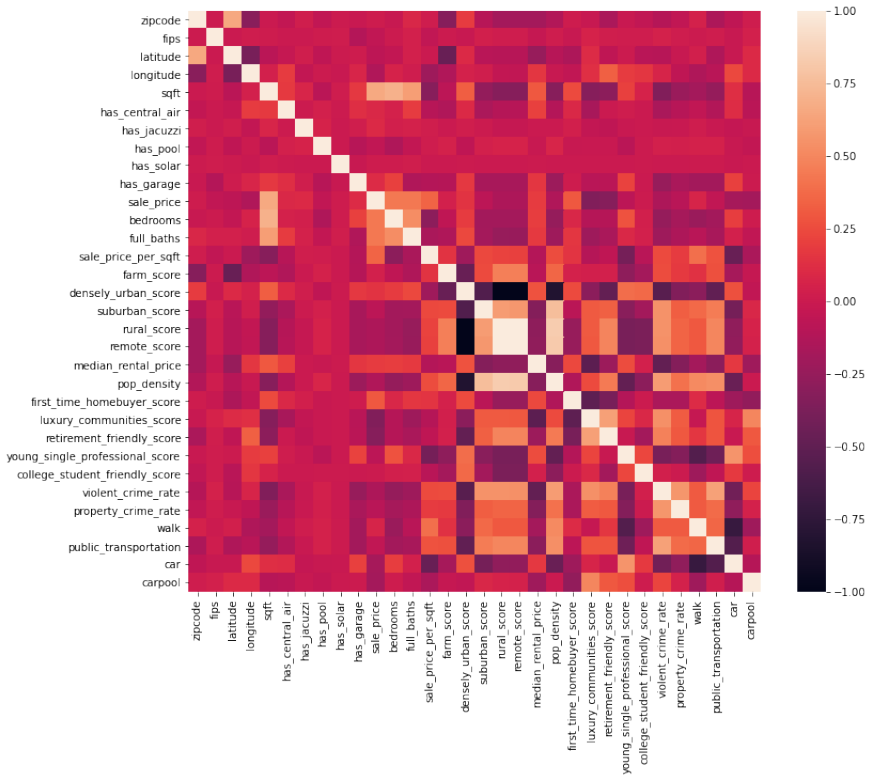

Figure 2

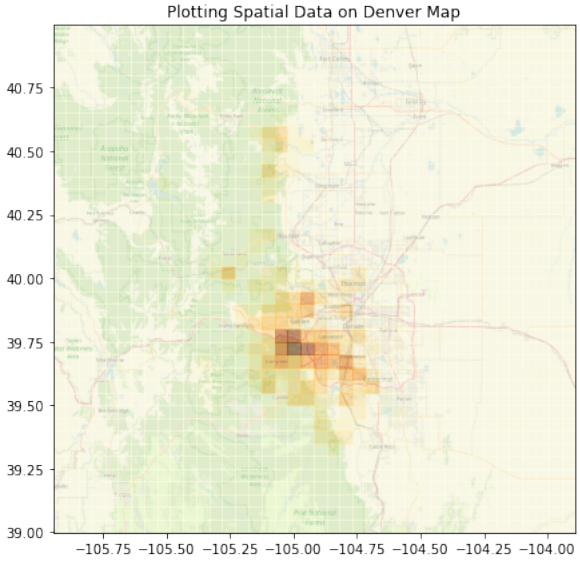

Figure 3: Geographical distribution of data

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 data point 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.

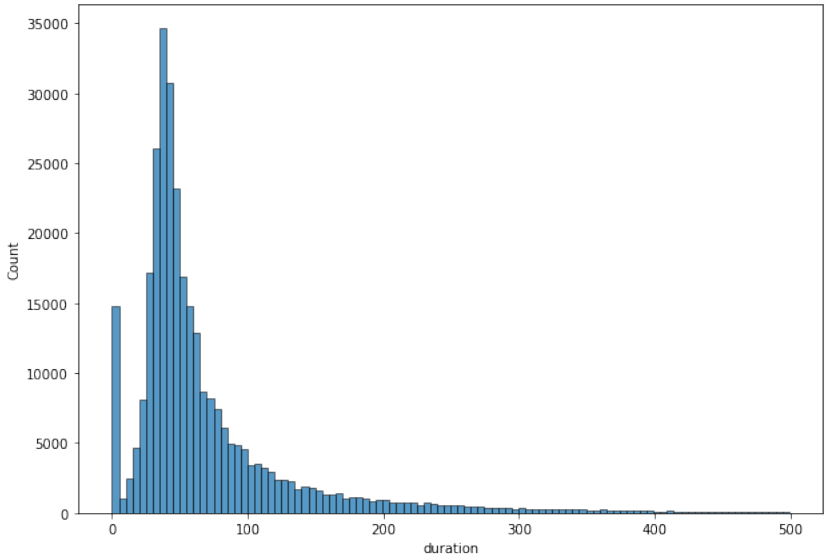

Figure 4: Distribution of days-on-market

4.2 Target variables

Initially, in our milestone 1, we created four target variables, two to predict the probability that a property is sold and listed, and two target variables to predict the days-on-market of that properties. However, from milestone 2 onward, we decided to focus on one target variable: the probability of a property being sold, given that it is listed, to keep it consistent between the baseline and developed model. We will later expand our models to predict supply if we are able to successfully address key challenges to our demand-side modelling approach.

In addition, we selected the second quarter of 2019 as the time-interval for all of our models, meaning we will predict given that a property is listed before or during the second quarter of 2019, then will it be sold during this timeframe at all. The reason that we chose 90 days because when we looked at the DOM distribution, we found out that most properties are sold within 90 days.

One challenge arises after we select this target variable, is that the number of houses that are not sold in each time frame is always much bigger than the portion of house that sold. Figure 5 shows this imbalance, where 0 in each graph indicates the number of houses that are not listed and sold, and 1 represents the portion of houses that are listed and sold, in the second quarter of 2019. We will discuss how this imbalance affects our models in the next section.

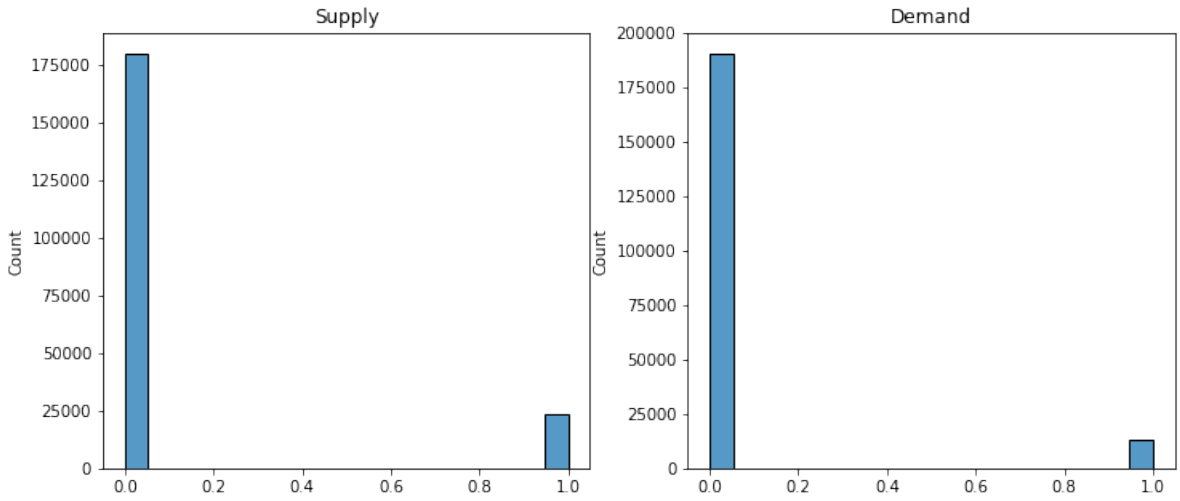

Figure 5: Distribution of supply and demand in Q2 2019

5 Baseline model

Our baseline approach falls into two categories, a submarket and non-submarket approach. The non-submarket baseline acts as a baseline for performance to ensure that our allocations of subamrkets perform better than if we had considered Denver as just one market. The submarket baseline is to ensure that the submarkets chosen in the advanced model are better than naively assigning submarkets.

5.1 Non-submarket baseline

5.1.1 Results

Methods	Testing Accuracy	ROC-AUC
Logistic Regression	0.71	0.54
Random Forest Regression	0.71	0.50
Neural network	0.82	0.51

Table 1: Non-submarket baseline models

5.1.2 Key Insights

For non-submarket baseline, beside the logistic regression model that we used in the first milestone, we also built a random forest classifier and a neural network model to compare. For the neural network model, after experimenting with different structure, we figured out that a simple model with 1 hidden layer of 256 nodes achieved the highest accuracy. The neural net converged after 9500 epochs, achieving 81% accuracy. However, the same model did not perform well when being used in conjunction with submarket classification, which is showed in table 2.

In addition, we trained our three new models with a different dataset compared to milestone 1's. Rather than using only 12 handpicked features, we incorporated a new dataset that REX provided that gave us an additional 80 features. However, the accuracy did not significantly change and the low ROC-AUC score indicated that the testing accuracy was a reflection of the data distribution rather than the actual prediction. As an attempt to resolve this problem, we initially used SMOTE, which generated synthetic data of the properties that are sold. However, after we used SMOTE, we saw a significant drop in accuracy in all of our models. The models' accuracy drop to 50% and there is also no increase in ROC-AUC score. We suspect that because we used SMOTE with a high-dimensional dataset, it may introduce some noise and outliers to the dataset. We are still experimenting with different techniques to handle this imbalance issue.

5.2 Submarket baseline

The submarket baseline approach was developed to allow comparison with the advanced model and to ensure that the advanced model performed better than naive classification of homes into submarkets. Therefore, we created five submarkets using k-means clustering (on solely the numerical features). We then trained five distinct models on each of the discovered submarkets in line with the hedonic assumption and found the weighted average scores to arrive at final results for test accuracy.

5.2.1 Results

Methods	Testing Accuracy	ROC-AUC
Logistic Regression	0.72	0.55
Neural Network	0.68	0.58

Table 2: Submarket baseline models

5.3 Key Insights

The major takeaway from the submarket baseline approach is that classifying homes to submarkets based on the similarity of their numerical features is not enough to drive performance boosts from our models. We therefore must use more sophisticated approaches which are explored in the developed model.

6 Developed model

Our developed model attempts to improve upon our baseline models in two ways. Firstly, we aim to have our model learn to define optimal submarket classifications on its own, rather than being limited by a predefined clustering of the data for each home. Secondly, we aim to be able to interpret the hedonic demand / supply model used within each submarket in order to understand the relative importance and effect of each feature of a given home on demand / supply.

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 (potentially stochastic) hedonic demand / supply function of its

features as determined by its submarket classification. Below, we discuss two subtly different realizations and implementations of this modelling framework. (Note that like our baseline models, these models are currently structured for demand-only prediction within a single specified time interval.)

6.1 Hedonic Demand Function #1: Bayesian Logistic Regression

6.1.1 Model Specification

For this model, we assume that our target variables for each home are functions of a stochastically-weighted linear combination of the home's features (akin to Bayesian logistic regression), and that the distribution of these weights is determined by the home's submarket membership. 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 (displayed graphically in Figure 6):

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 $k \in \{1 \ ... \ K\}$.

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 this 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 within this model, we currently assume a multivariate Gaussian distribution of home features, some of which may be discrete or boolean.

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.

6.1.2 Estimation with PyMC3

As mentioned in earlier sections, the developed model focuses on predicting demand, or the probability of listed home being sold in a given discretized period. Setting the number of submarkets parameter K to be 5, the Bayesian hierarchical model with a Bayesian logistic regression hedonic demand function (estimated by PyMC3) resulted in 3 submarkets identified. The training and testing accuracy, along with the ROC-AUC for the individual submarkets is shown in Table 3, as well as the weighted average according to the various submarket sizes.

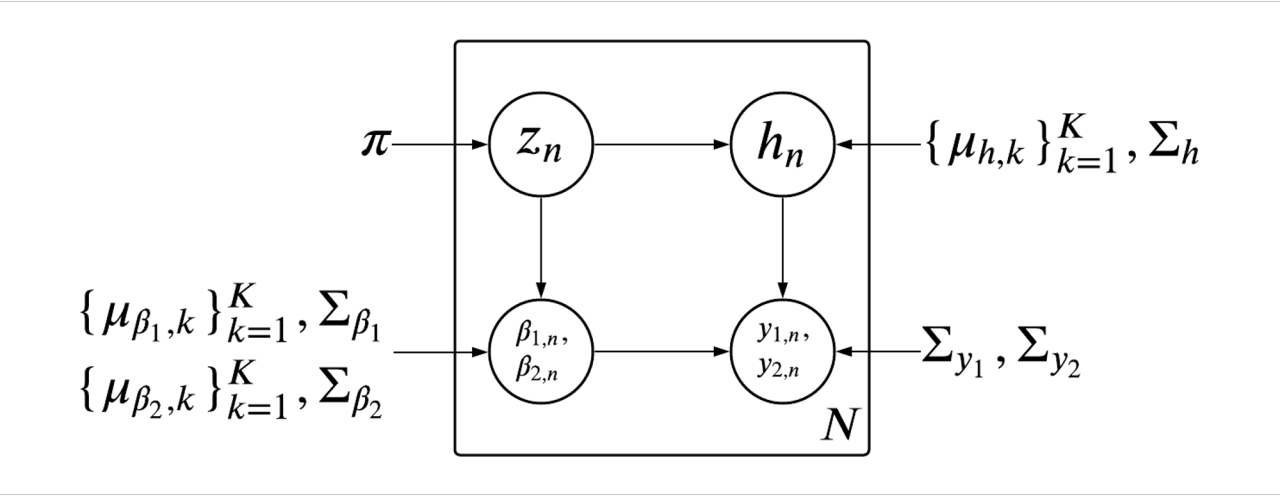

Figure 6: Proposed Hierarchical Model #1

Submarket	Number of Homes	Training Accuracy	Testing Accuracy	ROC-AUC
1	124	69.8%	43.2%	0.404
2	7402	72.6%	72.1%	0.538
3	41	62.5%	85.7%	0.788
Weighted Average	-	72.5%	71.7%	0.537

Table 3: Submarket Metrics using PyMC3

It is observed that the weighted average testing accuracy and ROC-AUC of this model have similar performance compared to the submarket baseline models using logistic regression. This could be due to the PyMC3 model not fully converging or an inappropriate submarket initialization. The geographical representation of the various submarkets are depicted in Figure 7, where there are significant overlaps and no clear demarcation between the submarkets.

6.2 Hedonic Demand Function #2: Frequentist Logistic Regression

6.2.1 Model Specification

For this model, we assume that our target variables for each home are functions of a deterministically-weighted linear combination of the home's features (akin to logistic regression), and that these weights are determined by the home's submarket membership. 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 (displayed graphically in Figure 8):

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 $k \in \{1 \ldots K\}$.

2. $\mathbf{h}_n \sim P_h(\Theta_{z_n})$: 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

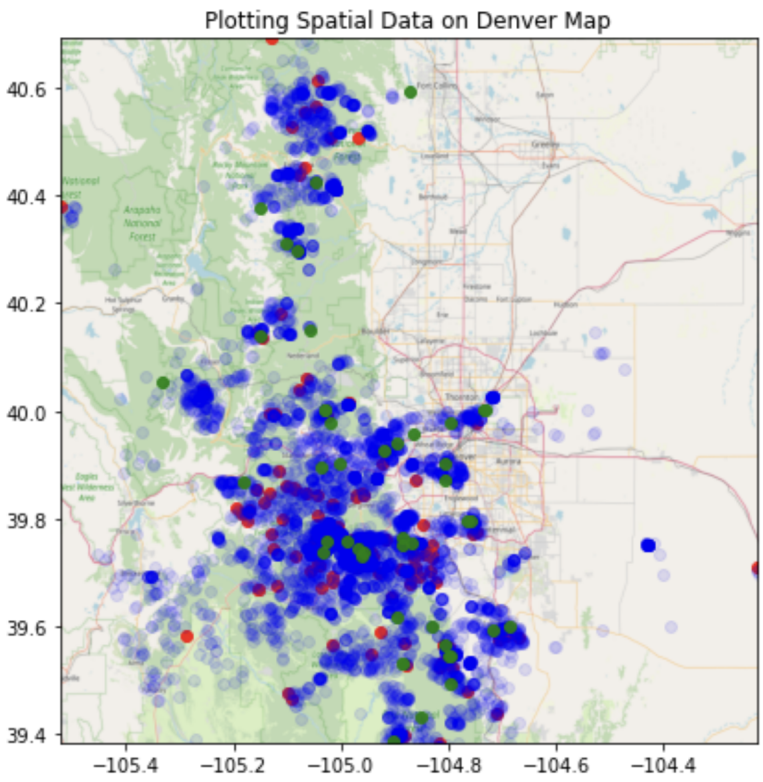

Figure 7: Locations of 3 Submarkets Generated by PyMC3

by Θ_{z_n}, a set of parameters characterizing the multivariate-normal, Poisson and Bernoulli distributions of continuous, discrete and boolean home features (respectively). Note that we no longer assume a multivariate Gaussian distribution of *all* home features within this model, nor do we assume that the covariance matrix of continuous home features is shared across submarkets.

3. $y_{i,n} = f_{i,z_n}(\mathbf{h}_n),\ i \in \{1, 2\}$: 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. The functions f_{i,z_n} are hedonic demand / supply functions of $\mathbf{h}_n$ determined by the home's submarket membership. Note that f_{i,z_n} corresponds to a logistic regression of $\mathbf{h}_n$ in our current implementation, although this implementation allows us to easily explore other functional forms in our later work.

6.2.2 Estimation with Expectation Maximization (EM)

For this model, we once again focus on predicting demand, or the probability of listed home being sold in a given discretized period. Setting the number of submarkets parameter K to be 5, the Bayesian hierarchical model with a frequentist logistic regression hedonic demand function (estimated by 5000 iterations of EM) resulted in 5 submarkets identified. The training and testing accuracy, along with the ROC-AUC for the individual submarkets is shown in Table 4, as well as the weighted average according to the various submarket sizes.

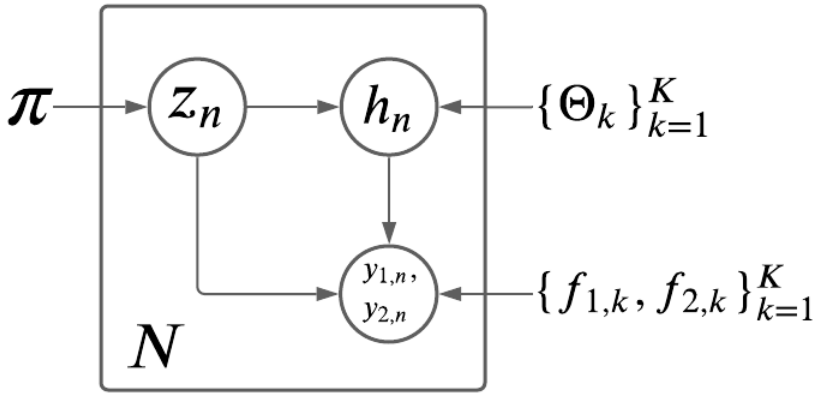

Figure 8: Proposed Hierarchical Model #2

Submarket	Number of Homes	Training Accuracy	Testing Accuracy	ROC-AUC
1	1523	64.7%	63.7%	0.617
2	801	89.5%	88.4%	0.548
3	887	70.3%	68.5%	0.518
4	2470	75.4%	75.3%	0.560
5	1886	71.4%	69.8%	0.518
Weighted Average	-	73.1%	72.2%	0.574

Table 4: Submarket Metrics using EM

It is observed that the weighted average testing accuracy and ROC-AUC of this model indicate slightly better performance compared to the submarket baseline models using logistic regression, although these improvements are not substantial. This could be due to the EM algorithm not fully converging or an inappropriate submarket initialization (similar to the issues cited alongside the model implemented using PyMC3), and can potentially be improved upon through more iterations of EM. The geographical representation of the various submarkets are depicted in Figure 9, where there are significant overlaps and only some demarcation between the submarkets.

6.3 Key Insights

We find that both of the above variations of our developed model provide (at best) marginal improvements in test accuracy and ROC-AUC. As previously stated, this can be caused by a lack of convergence or poor initialization of our samplers. However, the near-uniform performance of all of our submarket-based models seems to stem from two fundamental challenges to our approach:

1. **Class Imbalance:** Within a specified time frame, the number of homes that are listed but not sold is much greater than the number of homes that are sold. This imbalance allows classifiers to achieve an artificially high predictive accuracy by predicting the majority class for most inputs, and is evidenced

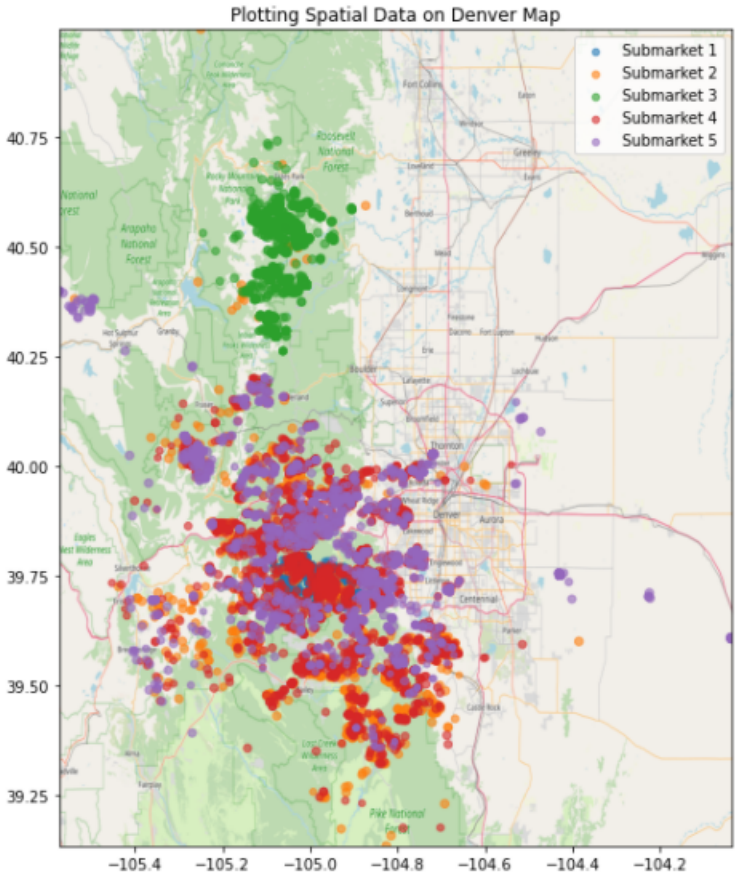

Figure 9: Locations of 5 Submarkets Generated by 5000 EM Iterations

by the low ROC-AUC of our results. In order to make informed classifications, we must find an effective means for addressing this imbalance.

2. **Hedonic Assumption**: A key assumption of our current framework is that a listed home's probability of being sold in a specified time period (as well as an unlisted home's probability of being sold) is given by a (possibly submarket-dependent) function of the home's features. It is difficult to evaluate the validity of this assumption when our model's performance is largely dictated by the previously mentioned class imbalance problem. However, as macroeconomic data has been shown to be a good predictor of market-wide metrics of supply and demand [4], we are considering relaxing this assumption. That is, we will instead assume that a listed home's probability of being sold in a specified time period (as well as an unlisted home's probability of being sold) is given by a (possibly submarket-dependent) function of the home's features *and other macroeconomic factors*.

7 Further work

Our future work ahead of the next project milestone will be focused on addressing the challenges cited in the "Key Insights" of our developed model discussion. Some key goals that we will work towards include:

- 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 models obtain an artificially high accuracy with poor AUC scores. We have previously explored the use of SMOTE for generating synthetic data to address this class imbalance, although we were unable to appropriately approximate the conditional distribution of home attributes given that a home is sold within a specified time period. We plan to attempt various alternative methods for addressing this imbalance, including but not limited to:

  - The use of variational autoencoders for generating synthetic data.
  - Stratified sampling of training data.
  - Class-weighted logistic regression.
  - A new loss function (such as loss measured at an aggregate level, comparing the total number of homes sold to our model's predicted / expected number of homes sold within a specified time period).

- Should we decide to relax our hedonic demand / supply function, we will 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] Maurer, R., Pitzer, M., Sebastian, S. (2004). Hedonic price indices for the Paris housing market. Allgemeines Statistisches Archiv, 88(3), 303-326.

[2] Gouriéroux, C., Laferrère, A. (2009). Managing hedonic housing price indexes: The French experience. Journal of Housing Economics, 18(3), 206-213.

[3] Jiang, L., Phillips, P. C., Yu, J. (2014). A new hedonic regression for real estate prices applied to the Singapore residential market.

[4] 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

[5] 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.