FINAL PROJECT

Automated Valuation Model of Land

Noah Pape[a]

[a] *Koç University, Department of International Relations, Istanbul, Turkey;*

Acknowledgements

The author has completed this work as his submission for the Spring 2021 INTL 550 Final Project assignment.

1. Introduction

Automated valuation models (AVMs) have proliferated in recent years and are a key part of the iBuyer ("instant buyer") technology that is used by companies like Zillow and Opendoor to make instant offers to residential homeowners looking to sell their residence. Researchers are now calling for the same technology to be used in the property tax assessment process to more efficiently reach accurate assessment values. Despite the progress in recent years, these models have generally remained limited to only valuing residential homes – most of these firms do not purchase condos or apartments. While firms like Offerd have created AVMs for commercial real estate in recent years, they still have not deployed them like an iBuyer; instead, the AVM is used for internal use as part of the broader investment process. In this paper, we propose an AVM for undeveloped land that uses a Random Forest.[1]

In Section 2 of this report, we briefly review other land AVM models that have been proposed in the literature and outline our hypothesis. Section 3 consists of a brief description of our data and outlines the process through which it is acquired. In Section 4, we detail our methods used to test our hypothesis before presenting results in Section 5. Finally, we conclude with a brief summary and ideas for future research.

2. Hypothesis

As we will outline later in Section 4, we believe that a Random Forest model effectively replicates the land valuation process that is undertaken by both brokers and tax appraisers. This leads us to our first hypothesis below:

H1: Using a Random Forest will improve the accuracy of our model compared to the baseline

Additionally, we feel that including features that are descriptive of the land will also improve the quality of our model. In particular, the features we focus on are the percent of the perimeter of the property that has road frontage, the percent of the perimeter of the property

[1] https://www.prnewswire.com/news-releases/offerd-launches-1st-and-only-commercial-real-estate-ibuyer-and-ai-acquisitions-platform-300939023.html

that has water frontage, and the percent of the property that is in the floodplain. We believe on theoretical grounds that each of these should have a noticeable effect on price. First, a property that has road frontage is likely to have better road access and be more convenient to reach, two things that are generally positively valued by buyers. Additionally, properties that have large amounts of road frontage can easily be subdivided and we believe the economic value of this option value is reflected in the price. Next, we have a high level of confidence that properties with water frontage will receive premium valuations. Finally, we expect that properties in the floodplain will receive lower valuations as it is difficult or impossible to develop in the floodplain in most areas in Texas and thus the economic option value of the land is lower. This leads us to our second hypothesis:

H2: Including additional GIS features will improve the performance of our model

3. Data and Descriptive Analysis

The data was sourced from LandsofTexas.com, where all active listings for undeveloped land and/or ranch land with a lot size of more than 10 acres was included. After this, the listing was algorithmically matched with a set of corresponding parcels in a geographic information system by using the geographic point that was provided in the listing information. Next, multiple data sources for roadways, waterways, and floodplain were sourced from the State of Texas' open GIS portal. After including these data sources in our geographic information system, we create GIS-related features that capture additional information about the property that is for sale. This data was directly provided by NXT Analytics who granted use of the data for academic purposes. In total, there are n=4,750 listings.

The geographic distribution of our data can be seen in the figure below. Notably, there are two gaps in the areas that would make up downtown Houston and Dallas-Fort Worth. These areas not highly represented in our dataset for the following reason: there are very few properties that remain in these very urban areas that are less than 10 acres and remain undeveloped land. Unexpectedly, we see that our data is concentrated in the greater urban

areas of the four major cities in Texas: Austin, Dallas, Houston, and San Antonio. Other cities such as Lubbock are noticeable as well. Because this map shows a 2d histogram that represents the listing counts and *not* the total listing areas, it is not at all unexpected that the rural areas

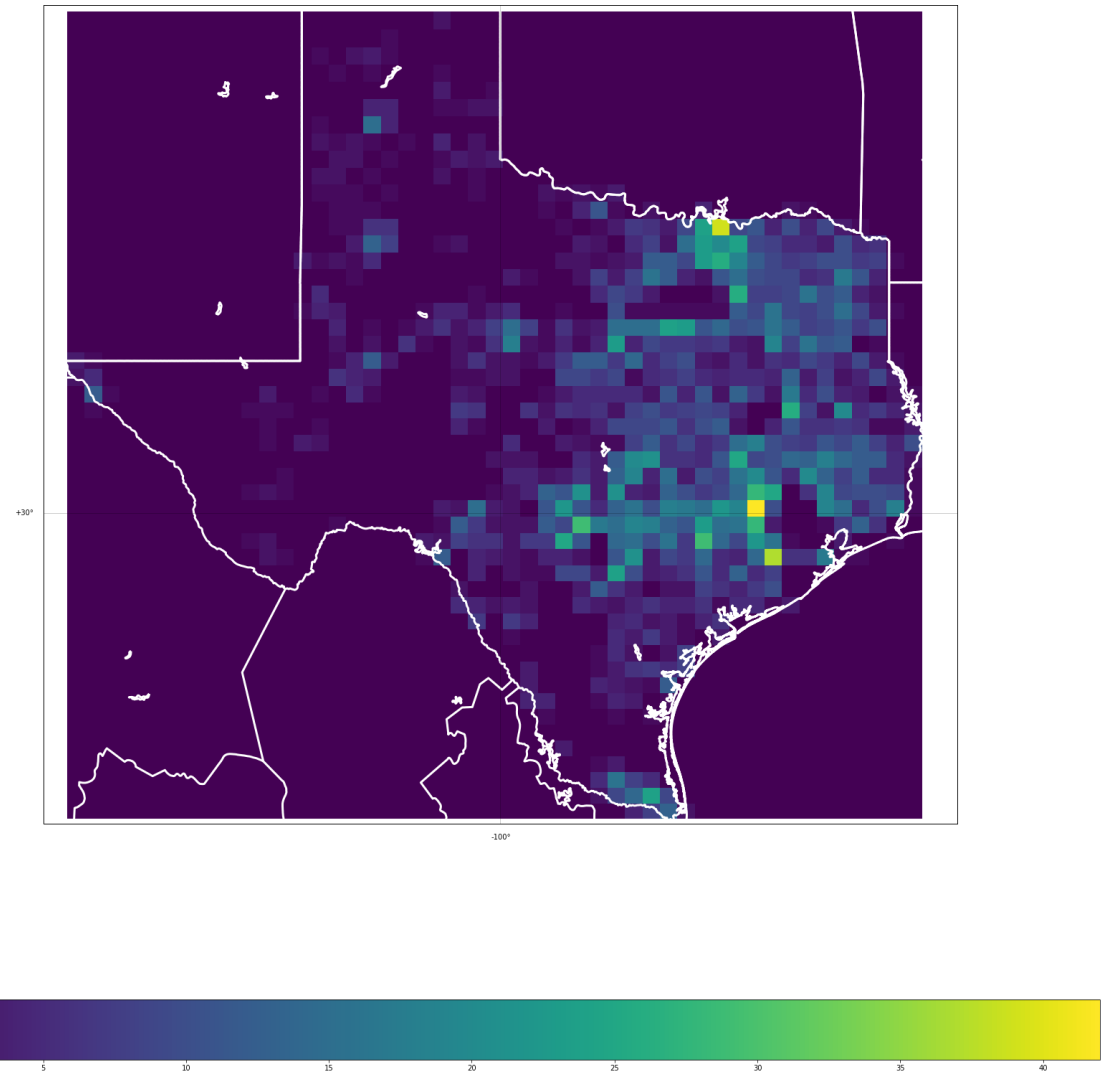

Figure 1: 2D Histogram -- Number of Listings

Next, we want to show how the average price per acre varies across geographies. This can be seen in Figure 2 below. The results are not at all surprising – the major metropolitan areas have the highest average price per acre which decays as one gets further from these areas. Particularly, the strength of the real estate market in areas like Austin and Houston are

evident, which reflects what is seen in the real world.[2] One strength of the proposed Random Forest Regressor is that it can make multiple splits on latitude and longitude which will allow the model to capture the clearly nonlinear relationship between these two variables and price.

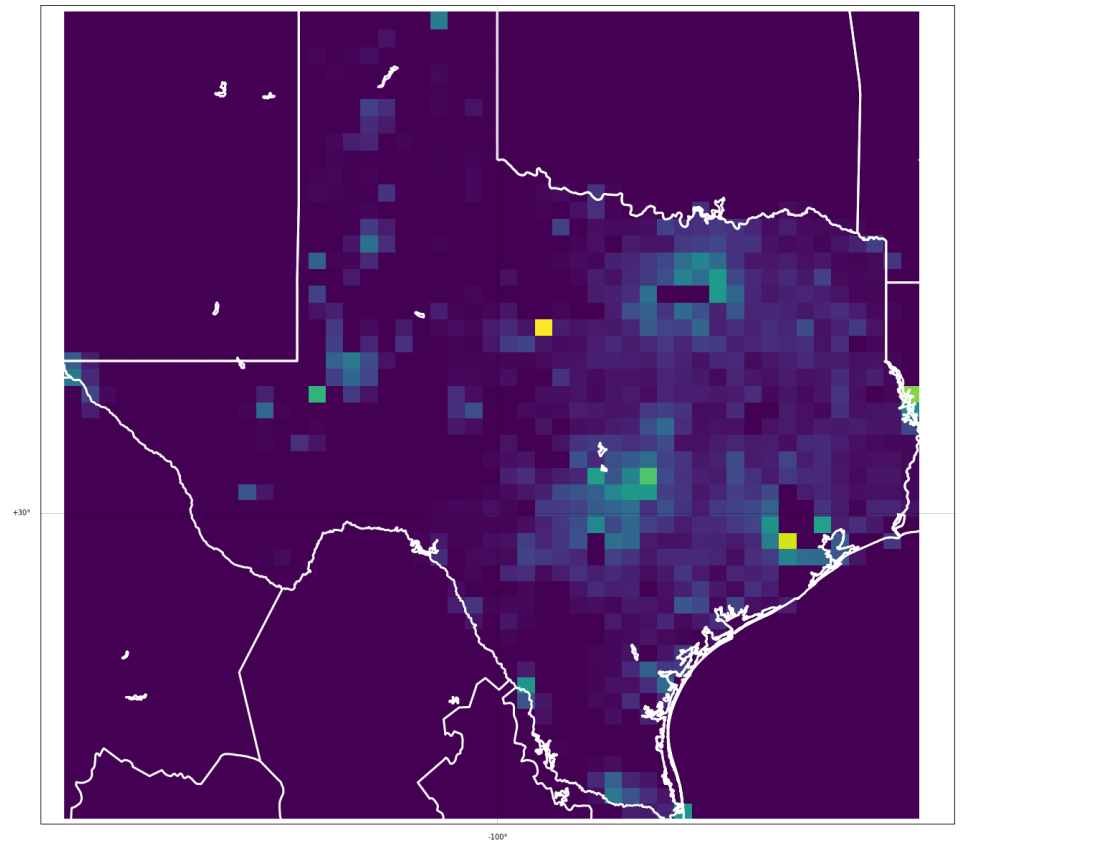

Figure 2: Average Price per Acre

Finally, we provide a map of the number of acres for sale. Unsurprisingly, this map does not have the same spikes in urban areas that is seen in the other maps. This map is more balanced for the following reason: while there are more land listings in the areas that surround cities, these listings are usually for smaller properties; on the other hand, while rural areas have

[2] https://www.kvue.com/article/money/economy/boomtown-2040/austin-housing-market-home-sales/269-64540b3d-4d3b-4acb-9627-3f9486f94ed3

fewer listings, the properties for sale are much larger. In Figure 4, we show some of the large listings in West Texas that contribute to the large number of acres for sale in these areas.

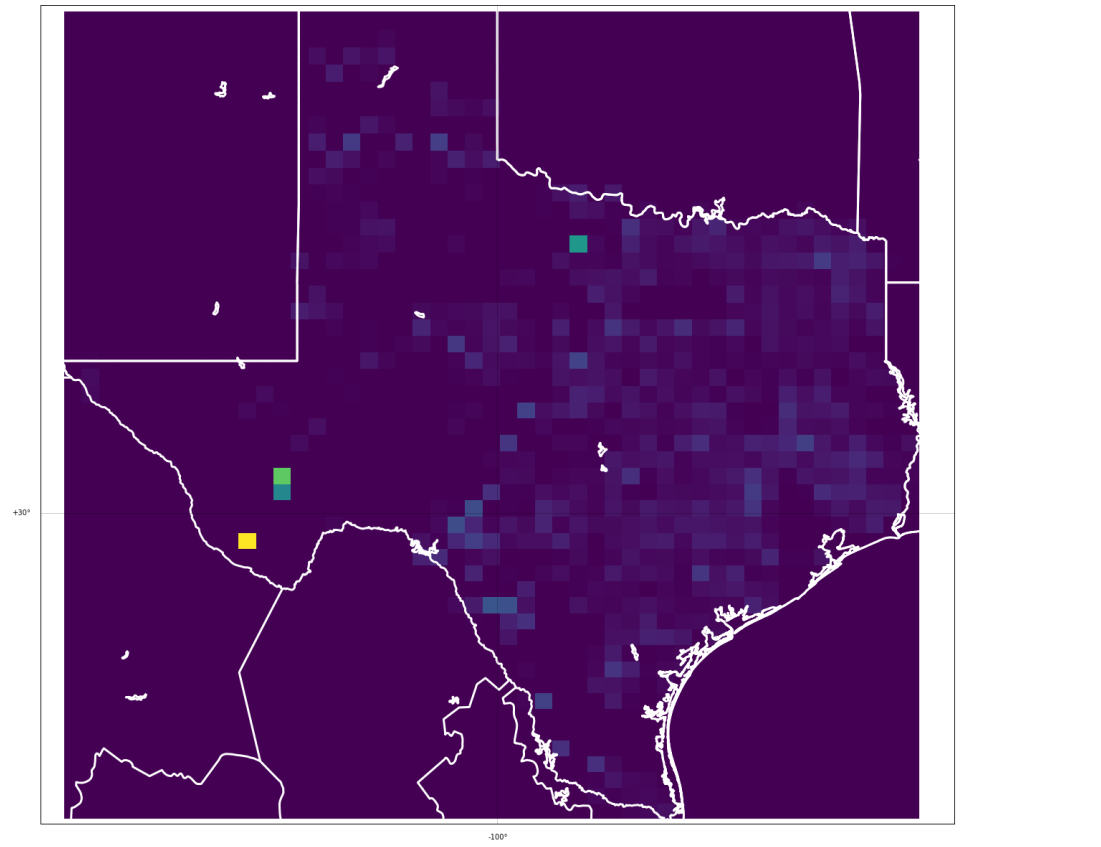

Figure 3: Number of Acres for Sale

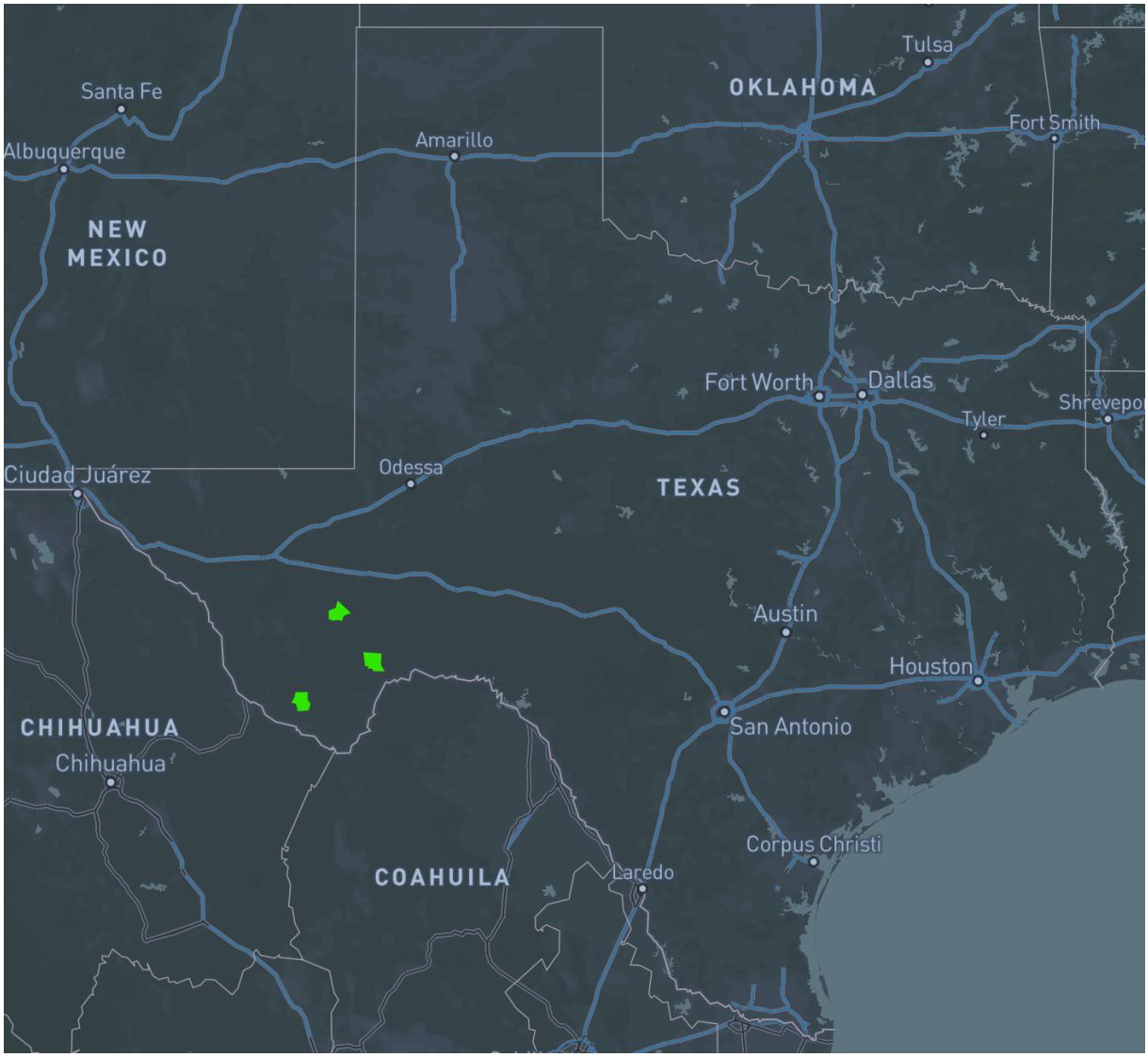

Figure 4: Large Listings in West Texas

We also provide a brief description of our other features. Figure 5 contains scatterplots versus price per acres for four of our features. We first note that for each of these features, there are many listings that do not contain them at all; this is particularly true for river and water frontage and is reflected by the concentration along the x-axis at the value 0. It is hard for us to directly identify a trend from these scatterplots and the relationship is clearly

nonlinear. This provides further justification for our Machine Learning approach to this problem.

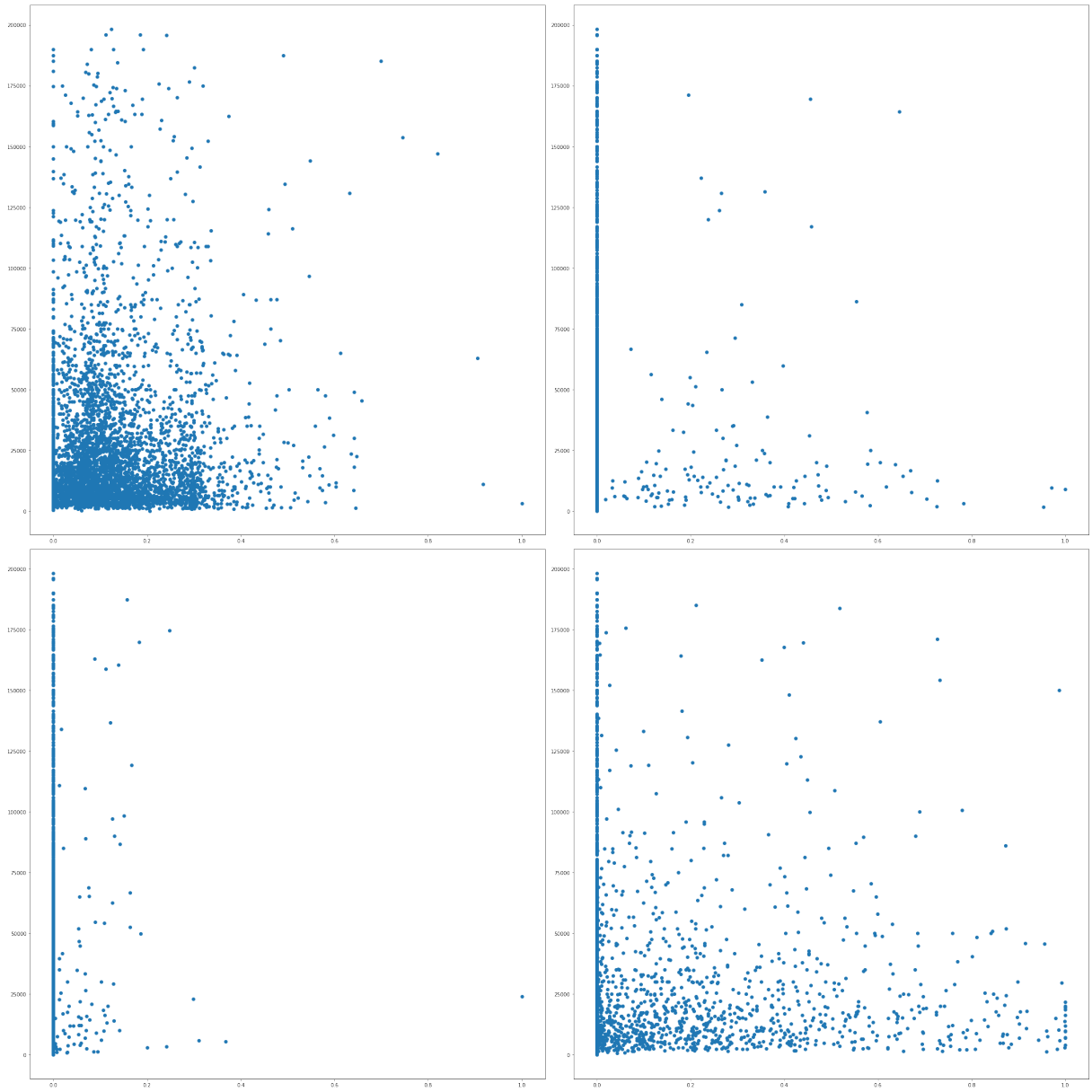

Figure 5: Left to Right. Top Row: Road % of Perimeter vs PPA, River % of Perimeter vs PPA. Bottom Row: Lake % of Perimeter vs PPA, Floodplain % of Acres vs PPA.

4. Methods

Our baseline model will be a K-Nearest Neighbors (KNN) regression where the outcome variable is the *PPA* and the input variabls will be the latitude and longitude of the centroid of the property. We will then use leave-one-out cross-validation to find the optimal value of *k*

and to determine whether to use a uniform or distance weighting scheme in the model as well as the distance formula used in the model. We will also construct a KNN model that uses GIS features as well. Next, we construct a Random Forest model and only include the same latitude and longitude input variables. We then include the GIS variables to construct a Random Forest model. Note that our GIS variables also include a categorical variable which indicates if the property has any significant improvements. For these Random Forest Models, we use *k*-fold cross-validation to find optimal hyperparameters.

Before constructing the models, we split our data into a train and test dataset where 10% of the data is in the test set. We do not use the test-set for any cross-validation and use it only at the final stage to evaluate our model. The performance of each of our models on this dataset will provide evidence for us to accept our hypotheses. We use four metrics to evaluate our models: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). In our cross-validation steps outlined early, we choose the best model on the basis of cross-validation MSE.

Note that the outcome variable of our regressors is different. For our KNN Regressors, we predict price per acre which we then multiply by the number of acres to get our predicted price. In this specification, acres are not used as an input variable at all in either KNN model. On the other hand, our Random Forest Regressor predict price directly and do include acres in their specification as an input variable. More details on the specification can be found below:

Model	**IVs**	**DV**
KNN	Centroid Lat and Long	Price per Acre
KNN-with GIS	Centroid Lat and Long, Road % of Perimeter, Lake % of Perimeter, River % of Perimeter, Floodplain % of	Price per Acre

	Acres, Significant Improvements	
RF-no GIS	Centroid Lat and Long, Acres	Price
RF	Centroid Lat and Long, Road % of Perimeter, Lake % of Perimeter, River % of Perimeter, Floodplain % of Acres, Significant Improvements, Acres	Price

4.1. KNN Regression

We provide a brief overview of the KNN Regression technique for readers who are unfamiliar with the method. This technique is a local method that is used to estimate continuous variables by looking for feature similarity. For each observation, we look for the most similar observations based on a set of observations and a selected distance formula. For example, we could calculate the distance between two points using a Euclidean or L2 norm distance where we use the square root of the sum of the squared differences between the two observations for the given features. We could also use a Manhattan Distance formula where we use the sum of the absolute differences between the two points. Regardless of the distance formula we select, we identify the *k* most similar observations (or observations with the lowest distance to our observation of interest). This *k* value is often selected by testing a range of values and looking for the one that minimizes error. Finally, after identifying the *k*-most similar observations, we calculate our prediction for the observation of interest's output variable by either using a simple average of the output variables for our *k* observations or by using a weighted average

of the output variable for the *k* observations where they weight is proportional to the calculated distance. We also note that input variables are generally normalized so that the distances are not dominated by the distance of a single feature. We prefer KNN for our baseline model over a linear regression because we have high confidence that the latitude and longitude of a property do not have a linear relationship with price per acre and because the method allows us to replicate the process that is used by appraisers in the first place.

4.2. Random Forest Regression

We also provide a brief overview of the Random Forest Regression technique for readers who may not be familiar. This is an ensemble method that is used to estimate predict continuous variables by combining the predictions of many weak learners, which, in the case of the Random Forest Regression, are decision trees. This a bagging technique which means that there is no interaction between the learners (as opposed to boosting techniques). We often utilize Random Forest methods over decision tree methods as decision trees can be computationally expensive to train and are prone to overfitting. Further, they are quite susceptible to finding local optima. However, using a Random Forest helps us with these issues by training many shallow decision trees on samples of the training data that are generated with replacement. The randomness introduced by sampling and the regularization technique that is introduced by limiting the depth of the decision trees help to solve the overfitting problem and the problem of finding local optima. Additionally, even though we are training many decision trees, because the trees are not fully 'grown,' the computational cost of the technique is quite manageable. The main disadvantage versus decision trees is that we lose the interpretability of decision trees and the measures of feature importance for random forest methods are not always useful.

4.3. Model Justification

We believe that the Random Forest Regression using GIS features is the closest representations of the underlying process that is used to set prices by brokers. Thus, we feel

that the choice of this model is valid on theoretical grounds. Before choosing a listing price for a property, the broker usually identifies similar properties on the market by first looking for similar properties for sale or that have recently sold, and then among that group weighting the price in favor of properties that are similar with regard to their road access and water features, among others. This process could be hypothetically modeled by a decision tree where the first splits are based on the geographic location and then subsequent splits take into account geographic features. This similarity is why we feel the Random Forest Regression is a good choice for modeling our data.

Essentially, brokers price the land by finding the most similar pieces of land in the local area. However, this is a two-step process where the first step is finding the proximal properties solely by geographic location, and the second is finding the most similar based on other features among those properties. KNN, as a local method, does have some validity, but, in its out-of-the-box construction, it is not entirely suitable for recreating this process. If we include the GIS features as input features when training constructing our KNN model, we cannot ensure that selected neighbors will actually be within a suitable geographic proximity to our property of interest. Even if we are to use a weighted distance formula that emphasizes geographic proximity, we cannot be sure that the properties will be within a geographic range that makes sense. We could alleviate this by not including the GIS features, but then we would miss out on the second step of the process. It might be hypothesized that we could create a two-step KNN process, where the first model is trained with a large *k* and only geographic location features, while the second model is trained on only the subset of resulting properties from the first model and includes non-geographic location features as well. We propose this model as a direction for future research.

5. Results

Our results can be found in the table below. Overall, we find clear evidence to support our first hypothesis and rather inconclusive evidence for our second. Both Random Forest models

clearly outperform their KNN peers on all of our metrics. There are three statistical reasons that this might be the case: first, the Random Forest model is essentially able to act locally as a nearest neighbor model with a dynamic *k* so the model is responsive to the concentration of data in the area. Second, there is strong evidence that larger properties trade at an increasing discount on a price per acre basis.[3] By using acres as an input variable in our Random Forest Regressor, we are able to account for this while the KNN model is not able to account for this trend. Finally, we expect that using the latitude and longitude variables, the model is better able to reflect the urban land assets that have very price per acre values.

On three of our four metrics, the RF-GIS outperforms our baseline RF, but not by a convincing amount. Our KNN-GIS model only outperforms on two of the metrics, and again only slightly. We suspect that the latitude and longitude variables are dominating our models because the data is so spread out. It might be the case that if more geographically concentrated data was used that we would see a different result. Regardless, we see that the errors are quite large and we argue that much more work needs to be done in order to create an AVM that is useful in practice.

	MAE	MSE	RMSE	MAPE
KNN	1,637,931	3.46E+13	5.88E+06	127.1%
KNN-GIS	1,739,143	2.02E+13	4.49E+06	145.5%
RF	613,161	1.13E+12	1.06E+06	**77.3%**
RF-GIS	**608,650**	**8.53E+11**	**9.24E+05**	88.5%
Average	*1,149,721*	*1.42E+13*	*3.09E+06*	*109.6%*

[3] https://www.reonomy.com/blog/post/national-vacant-land-sales-report

6. Conclusion

Our results provide a starting point for the future of automated land valuation; however, there is still a long way to go. Our analysis shows that a Random Forest model might be an approach that is justifiable on both theoretical and empirical grounds. We argue that the process used by brokers and tax assessors is similar to a decision tree rather than a nearest neighbor process and our results indicate that the Random Forest model significantly outperforms the Nearest Neighbors model.

We propose a few different directions for future research. First, instead of constructing our valuation model of market prices, we might consider using actual sales data. This may be acquired by going directly to counties in states which are disclosure states. Next, we would consider adding additional GIS features; for example, we might consider breaking our road feature into the actual percentages for different types of roads as highway frontage might be seen as more valuable than other road types. We could do the same for rivers to note if a property has frontage on a major river. Finally, we would also want to construct our feature for significant improvements in a different way. Rather than simply using a categorical variable for whether or not there are significant improvements, we would use the underlying assessed improvement values in our models. We might also consider looking for data that is more concentrated geographically, particularly for testing the importance of GIS features.