Decision Tree Generated Real Estate Comparable Transactions

Ian Flint
New York University
if612@nyu.edu

1 Introduction

Real estate valuation is a complex task that considers many different factors to come to a conclusion about a particular property's market value. It is a common approach among machine learning practitioners in this field to try to construct a model that is as accurate as possible determining market value. However, this paper intends to explore a different approach, where machine learning models and information visualization techniques are used to aid a human decision maker. In particular, the focus will be to identify a group of properties that share similar characteristics to the property being examined, explain why they are similar, and use this information to determine a price relative to the market.

The use of comparable transactions is commonplace in the real estate market, and has been the traditional tool for relative pricing of real estate for many years. The process of finding comparable transactions is typically a very qualitative task, relying on a real estate professional's judgement of similarity. The approach explored by this paper will attempt to make this process systematic and provide a specific set of rules that determine similarity between properties.

The reason comparable transactions are so important in real estate is because they set the market price for properties similar to them. There is no intrinsic value to housing in the modern real estate market, everything is relatively valued. Thus, having the best information regarding comparable transactions is a valuable tool to someone trying to analyze the market.

2 Related Work

2.1 GAMUT

In the paper GAMUT: A Design Probe to Understand How Data Scientists Understand Machine Learning Models [1], the authors have a similar goal as this project, to assist people in understanding high dimensional data, and machine learning models built on such data. The project provides a visualization platform for generalized additive models (GAMs) using real estate valuation as an example of its capabilities.

While GAMUT takes a different apporach that this project, the authors' work provided a source of inspiration for developing a new system. While GAMUT is designed for GAMs, this project is based on tree models, and uses a different visualization approach because of it. The authors of GAMUT bring up an interesting point that models are used differently by practitioners with varying levels of experience. With that in mind, this project aimed to accommodate even the most inexperienced users by keeping the system simple and easily explainable.

2.2 Other Related Work

There are various other examples of research on visualizing real estate data, including Visualization-Aided Exploration of the Real Estate Data [2]. In this paper they also attempt to build a valuation model with a wide range of data features and display the information using maps and parallel plots. The model focuses on features like basic property information, regional characteristics, transportation and education.

In another paper, A Web-based Visual Analytics System for Real Estate Data [3], they take a slightly different approach to visualizing similar data. In this project they utilize a map view, a stacked graph to visualize data over time, a pixel bar view to visualize multiple features at a time and a treemap view to visualize the data's hierarchical structures.

2.3 Techniques Used in Popular Real Estate Tools

In addition to academic papers, the real estate industry has built many commercial tools to address related topics. Many of the most popular tools in the industry do not provide the functionality that is described in this paper and this served as one of the motivations for exploring a new approach.

Two popular tools in the US that are freely available are Zillow [4] and Trulia [5]. Each provide a geographic view, filtering options and data table access, but the tools leave it up to the user to determine their own filtering criteria if they would like to construct a universe of comparable transactions. Focusing on comparable transactions would be a potential source of improvement for these tools and this project looks to explore a possible way of doing so.

3 Method

3.1 Theory

In an effort to build an easily explained real estate valuation model, this project utilizes decision trees. Decision trees are an interesting tool for this job because of their tree structure. A simple decision tree model will construct a binary tree where every node has a splitting condition. The splitting condition divides the data provided by the parent node into two groups. By analyzing a particular leaf node's path to the root node, we can construct a list of splitting conditions that can be used to filter the complete data set to return a set that includes the leaf and other data points similar to the leaf. These similar data points are comparable transactions as they share many of the same characteristics as the original point.

A particularly useful characteristic of machine learning generated decision trees is that splitting criteria can be determined by maximizing the information gain of the split. This method results in the prioritization of attributes and potential splitting criteria that are most relevant to determining value. As a result, the attributes that provide the largest gain will be used as the splitting criteria of the nodes near the root of the tree. This method of splitting also avoids considering attributes with a large number of distinct values that may not be particularly useful in the model.

The prioritization of high information gain splits at the top of the tree is important because we often need to limit the tree depth of the decision tree model. If the splits are not prioritized in this way, the model would leave out the splitting characteristics that it finds most useful to determining value. In practice, this is often what happens when humans construct comparable transaction universes without systematic help, important features are not considered or less useful ones are wrongly prioritized. With a systematic approach, this process can be more thoroughly studied and results would be more easily replicated.

After the decision tree model is trained, each data point's path to the root node is recorded. This path consists of an ordered list of nodes, each with a splitting condition. When visualizing the model, the list of splitting criteria is utilized as a list of cumulative filters for the original data set. With this information we can construct a comparable group of transactions at every level of the tree by applying that level's splitting condition filter in addition to every condition above it on its path to the root.

3.2 Data & Technology

This project utilizes a public domain data set from King County, Washington that includes a collection of characteristics about real estate sales that took place between May 2014 and May 2015. The data includes a number of features that are commonly used in real estate analysis and would be useful to professionals in the field. These features include:

- Price
- Bedrooms
- Bathrooms
- Square feet of living space
- Square feet of lot
- Floors
- Waterfront accessibility
- View quality
- Condition
- Grade

However, this data set has less features than what would be ideally desired by a professional in this field making large financial decisions. There are many interesting data collections that would supplement this list, however, data cleaning and merging can be a very difficult and time consuming process. This project has decided to focus on demonstrating the proposed visualization system with a simple data set rather than prioritizing additional data cleaning work. However, this is an area of potential future improvement and will be discussed later in this paper.

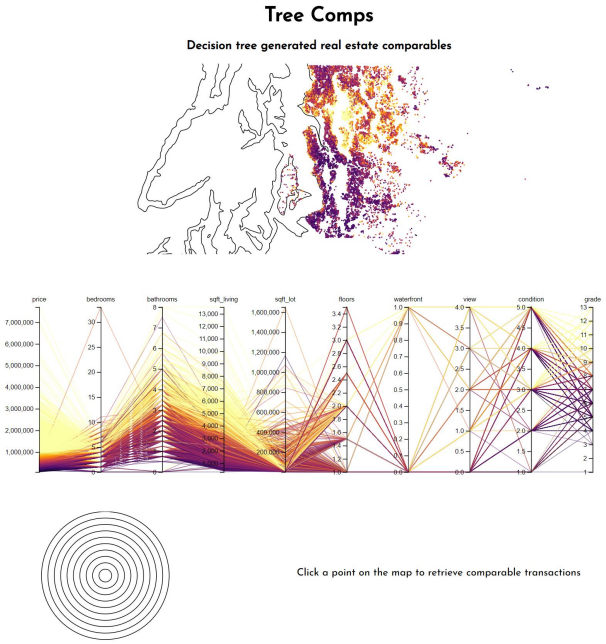


Figure 1: Overview showing the geographic, parallel coordinates and tree explorer views

The project uses scikit-learn for implementing the decision tree valuation model. Due to the nature of this project, being able to explain the model is a higher priority than optimizing model accuracy. As a result, a simple decision tree model was chosen for this task and scikit-learn provides great tools to implement this. Once the model is constructed, D3 is used to construct the data visualizations and data tables in the final output.

3.3 Visualization

Overview

The visualization in this project consists of four main areas, the geographic, parallel coordinates, tree explorer and data table views. Each view provides a different layer of information to the user of the system.

In the first view, users can visualize the geographic distribution of real estate values. The parallel coordinates view provides a visual overview of every point's characteristics, which general trends are more easily to see than a data table. The third view is the tree explorer view which includes a way of visualizing the tree model and an interface for filtering the data. In the final view, the exact values of each characteristic are listed at the bottom in a data table for every filtered comparable transaction.

Geographic View

A geographic presentation of data is very commonly included in real estate data visualizations because location is one of the most important factors in the analysis. However, two dimensional geographic representations are limited regarding how much non geographic information they can effectively convey at a single time and must be supplemented with additional views or a richer mapping experience.

While this project only provides a two dimensional map, it was in the original plan of this project to develop a three dimensional map to display more information to the user. Such a map would give the user more visual context to the physical environment of a neighborhood. This subject will be an area of future work to improve the system.

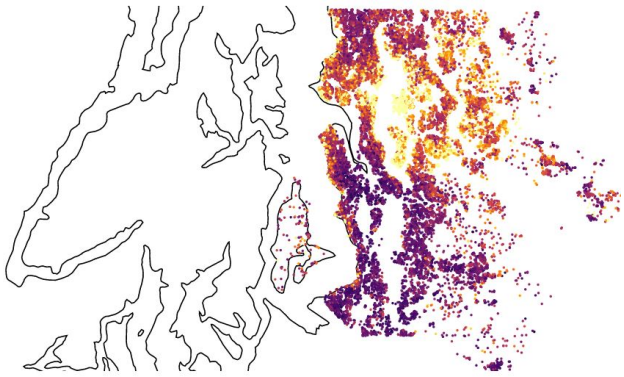


Figure 2: A closer view at the geographic view portion of the visualization. In this example, color is used to show property value per square foot, the characteristic that the decision tree model is trained to predict.

Parallel Coordinates View

Parallel coordinates are a very effective tool for visualizing high dimensional data and this format is often underutilized in visualizations in the real estate industry. In the parallel coordinates view of this project, the user will be able to see how each property compares to the others in all of the features that are considered in the model.

When many properties are being considered at the same time, the plot is difficult to use to track an individual line due to the clutter. Thus, when a smaller group of properties is selected, this section is filtered to only display the relevant line paths.

Tree Explorer View

The tree explorer view is the core interactive segment of the visualization and provides a way for the user to interact with the tree model. After selecting a point on the map, the relevant tree path is loaded and the filtering criteria are displayed in the tree explorer. The left side of the tree explorer consists of a number of concentric circles representing the layers of filtering at each node of the tree model. These are used to show which condition is currently selected, highlighting it in red. The concentric circles also visually reinforce the idea that each layer of filtering is a subset of the previous layer.

To the right of the circles, each layer's splitting condition is listed. The criteria are listed in order from the root node splitting condition on the left to the leaf node splitting condition at the far right. Selecting a filtering condition from this list will apply the selected filter to the data set, as well as every condition listed to the left of it. For example, in Figure 3, the leaf node is selected on the far right of the list and therefore the middle circle is highlighted red.

The selection of the leaf node is the most selective filtering option because it includes all the filters on that particular property's path according to the decision tree model. As the user moves to the left and selects the filtering conditions of nodes higher up the tree, the number of comparable transactions increases as it is being filtered less. This allows the user to define how specific they want to get in defining the comparable transaction universe.

This filtering system differs from the traditional filtering tools in that the user does not pick particular characteristics to filter. Instead the user picks the level of similarity they are looking for defined by how many levels of the tree they wish to filter with. Future work may consider a more customized selection process, but in the interests of simplicity this feature was left out of this initial exploration.

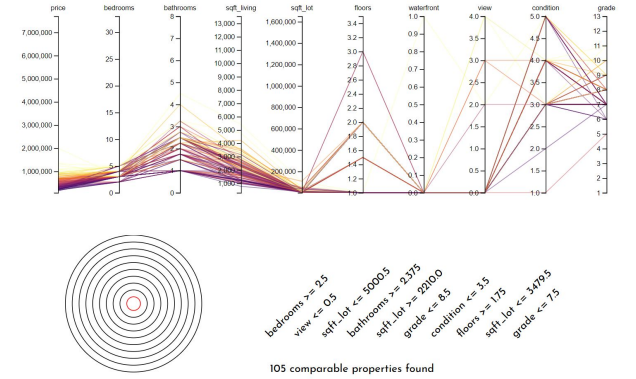
Data Table View

The other views are accompanied by a traditional data table section, where the filtered comparable transactions can be displayed for the user. This is particularly useful if the user needs to look up the exact values of a property's characteristics.

4 Evaluation

4.1 Comparison to Baseline

The baseline for this project was a tool similar to what is widely available for use in the real estate industry. This



Price	Bedrooms	Bathrooms	Sqft_living	Sqft_lot	floors	waterfront	view	condition	grade
464000.0	6	3.0	2300	3404	2.0	0	0	3	7
266000.0	3	2.5	1805	3402	2.0	0	0	3	7
300000.0	3	2.5	1320	2614	2.0	0	0	3	7
315000.0	3	2.5	1570	2865	2.0	0	0	3	7
238000.0	3	2.5	1650	2807	2.0	0	0	3	7
299000.0	4	2.5	1950	3000	2.0	0	0	3	7
465000.0	4	2.5	2200	3141	2.0	0	0	3	7
250000.0	3	2.5	1530	2890	2.0	0	0	3	7
320000.0	3	2.5	1530	2890	2.0	0	0	3	7

Figure 3: Visualization is shown with the leaf node level filter selected (grade greater than or equal to 7.5), as indicated by the red highlighted inner circle. This means that this condition and all of the criteria listed to the left of this condition are used to filter the original data. In this example, 105 comparable transactions were found, and these are shown in the filtered parallel coordinates view as well as the filtered data table view

tool consists of a geographic view and some form of data table view. Most tools also have filtering functionality, where you can explicitly define each of the characteristics you want in the filtered group.

Examples of popular tools similar to this baseline are the interfaces provided by companies like Zillow and Trulia. These examples work well for simple analytical tasks and for basic exploration of data, however, they are not particularly optimized for this project's goal. Despite this, these visualizations served as a good baseline to improve upon in this project and future work.

4.2 Future Work and Potential Improvements

This project outlines a process for achieving its primary goal, the implementation and visualization of filtering characteristics based on a decision tree algorithm. However, some objectives were not achieved, and will be an area of future work to improve the system.

In particular, the data set used in this project did not include rich geographic features for displaying in the geographic view. The original plan was to incorporate 3 dimensional building geometries, and well labeled points of interest throughout the geographic view. Switching to the current data set limited the scope of this project, but will be the focus moving forward to improve the visual experience of the system.

Another opportunity for improvement would be to redesign the tree explorer view to be more compact or better integrated with the other views. In this project the view was left as its own large section to explain the filtering process clearly. However, there are opportunities to redesign this feature such that it could be non obtrusively added to other visualizations as a filtering tool.

Adding functionality to allow the user to retrain the model based on their desired specifications would be an important improvement as well. The current model is limited to a tree depth of 10 in the interests of simplicity and computational performance. Larger, more sophisticated analyses may require a deeper tree to reach the filtering level desired by the user and it would be ideal in the future to provide that capability.

4.3 Challenges of Traditional Evaluation Metrics

In many projects related to utilizing machine learning models in real estate valuation, the focus is optimizing the accuracy of the model in predicting price. In this scenario, evaluation of performance is somewhat straightforward using traditional statistical techniques. However, this project aims to be a systematically constructed, qualitative aid to the human decision making process and thus is more ambiguous to evaluate.

In the paper GAMUT [1], a more comprehensive study was conducted on how data science practitioners interact with tools of this nature. It would be a logical next step to use their framework as a model for evaluating the ultimate effectiveness of this data visualization system. The paper's conclusions and suggestions were used to help guide this project, however, a study would need to be conducted specifically for this project to come up with similarly detailed conclusions.

5 Conclusions

This paper attempts to build a framework for visualizing decision tree generated real estate comparable transactions. The approach of using decision trees to generate filtering characteristics looks promising and could add value to existing visualization systems. The project provides a useful new dimension to existing real estate data analysis tools and with future improvement could exist as a standalone tool for professional analysis. Future work regarding a more immersive geographic visualization would greatly improve the system for practical uses.

6 References

[1] Fred Hohman, Andrew Head, Rich Caruana, Rob DeLine, Steven M. Drucker. 2019. GAMUT: A Design Probe to Understand How Data Scientists Understand Machine Learning Models

[2] Li, Mingzhao and Bao, Zhifeng and Sellis, Timos and Yan, Shi. 2016. Visualization-Aided Exploration of the Real Estate Data.

[3] Sun G D, Liang R H, Wu F L, et al. 2013. A Web-based visual analytics system for real estate data. Sci China Inf Sci

[4] Zillow. Real Estate Comps: How to Find Comparables for Real Estate. https://www.zillow.com/sellers-guide/real-estate-comps/

[5] Trulia. The Home Seller's Guide to Comparable Sales. https://www.trulia.com/blog/sellers-guide-to-comparable-sales/