Team 191 Project Final Report: **Property Tax Fighter**

Trevor Pollock, Tyler Church, Noumik Thadani, and Calvin Butson

11/22/2024

I. Introduction

Homeowners often face financial and logistical challenges when disputing their property tax appraisals, which can sometimes be set at the highest permissible values by appraisal districts. The current process requires time, specialized knowledge, and resources that many homeowners lack. While hiring attorneys is an option, their fees can reduce the tax savings. This project aims to create a website that utilizes local home pricing data and machine learning models to empower homeowners with data-driven insights, helping them contest their property taxes more effectively and affordably.

II. Problem Definition

Given a dataset of local home prices and property attributes (e.g., square footage, location, year built), the objective is to develop an automated valuation model (AVM) using supervised machine learning techniques. This model will predict fair property values based on comparable home sales and relevant property features, enabling accurate, user-friendly price estimates for homeowners. These predictions are then presented within a web application that guides users through relevant property tax appeal processes. The model should maximize predictive accuracy and minimize valuation bias to ensure fair and actionable property tax information.

III. Literature Survey

In support of the project, this literature survey explores various property valuation models, appraisal techniques, negotiation methods, and data-driven valuation systems.

Mohd et al. (2020) provide a comprehensive overview of real estate modeling techniques, including statistical, machine learning, and hybrid methods, which aid in predicting house prices.[1] This aligns closely with the project's aim to integrate machine learning for property tax valuation, as it discusses models ranging from linear regression to more complex neural networks. The study underscores the effectiveness of these models in handling diverse datasets, making it a critical foundation for understanding valuation in different market contexts.

Yavuz Ozalp and Akinci (2017) explore the hedonic pricing method, emphasizing how individual property characteristics (e.g., location, size, and amenities) influence residential prices.[2] This method's ability to isolate specific property attributes' impacts aligns with the need to pinpoint elements most influential on local property tax assessments. It also highlights the need for dimensionality reduction methods such as Principal Component Analysis to reduce correlated variables and model complexity. Pagourtzi et al. (2003) further expand on appraisal methods by examining various traditional and advanced valuation approaches, including income, cost, and comparison methods, while underscoring the need for accuracy and adaptability in real estate valuation—a fundamental concern for fair property tax assessments.[3]

Machine learning's role in real estate modeling is covered in-depth by James et al. (2013), with statistical learning and tree-based methods providing a theoretical framework for constructing robust predictive models.[4] Specifically, decision trees, k-nearest neighbors, and regression are algorithms that align well with the requirements of this project.

Voss (2016) offers insights from the negotiation perspective, which are particularly relevant to the context of contesting property taxes.[5] His principles in negotiation strategy—such as creating a "low-ball" valuation to counter an initial offer, then possibly walking the number up to a more likely value—can be applied to property tax appeals, aiding homeowners in effectively presenting their case to tax authorities. His emphasis on framing and influencing provides a strategic component that complements the project's technical approach.

For the regional policy context, the Texas Comptroller of Public Accounts (2021) outlines property tax remedies, offering guidance on the legal frameworks and steps taxpayers can take in challenging assessments.[7] This resource anchors the project within the legal procedures of property tax contests, ensuring the website adheres to and supports the statutory rights of taxpayers in Texas.

Gröbel and Thomschke (2018) introduce automated valuation models (AVMs) and the use of global versus locally weighted approaches.[8] Their work underscores the flexibility and precision needed in regional property assessments, offering valuable insights into how local pricing information might improve the accuracy and relevancy of tax appeals. Similarly, Pace (1998) and Thériault et al. (2003) discuss location-based models and attribute-specific valuation techniques, emphasizing the role of geographic factors in property value—a factor particularly pertinent to contested tax assessments.[9,11]

Lastly, Conway (2018) provides an overview of artificial intelligence and machine learning applications in real estate, noting how these technologies streamline property valuation and introduce new efficiencies.[12] Her thesis demonstrates the potential for AI-driven platforms in real estate, reinforcing the relevance of using machine learning in a project aimed at creating an accessible, data-informed tool for property tax assessment challenges.

IV. Methodology (Algorithm and Interactive Visualization)

Our approach is to provide a set of reasonable comparable properties for a user supplied residential property. The novel aspects of our approach include the cleaned dataset (see Data section), customized similarity metrics, user interactivity, and custom visualizations (see Experiments).

Comparable properties are identified using custom similarity algorithms that assess the weighted differences between the input property and others in the dataset. Weights are determined through various variable selection methods, such as linear regression, to identify impactful attributes on property prices. Our approach provides the user with appropriate comparable properties at very little cost. It also offers data driven reasoning to support the selection of comparables, utilizing weighted similarity metrics to clearly demonstrate the rationale behind each similarity measurement. This approach is unique in that it uses data and variable selection techniques to provide comparables for the user to use in their protest; this approach is better than the state of the art as it has the upsides of being data-informed like other machine learning methods used to predict housing prices and being interpretable like traditional comparable methods.

A. Workflow

1. **Input Processing**:
   a. The server.py uses flask to host and manage the Python, JavaScript, data files, and interactivity of the files for the project.

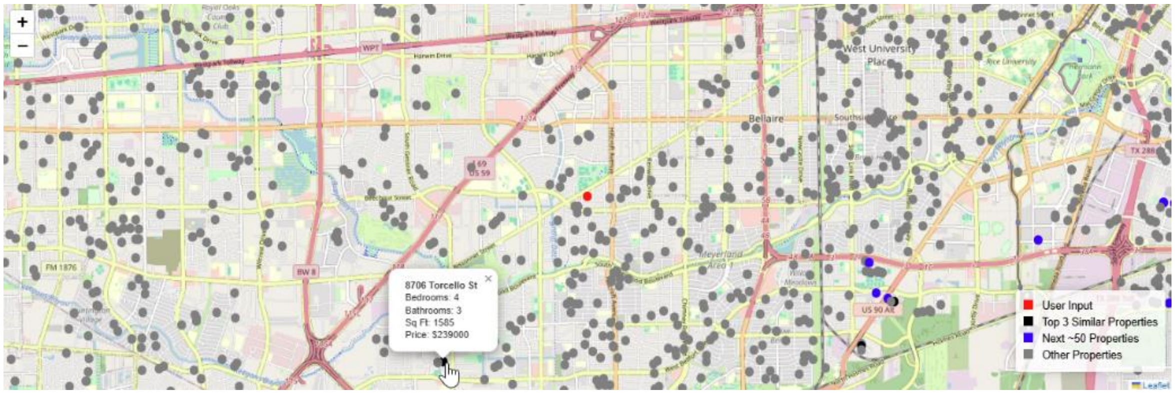
    b. On the website, the user specifies features of a target house (e.g., size, location, number of rooms) via the landing page: index.html.
    c. The `run_comparisons.py` script either matches the input to an existing house in the dataset or geocodes the address to fetch location data.

2. **Similarity Calculation**:
    a. Using functions from `distance_metrics.py`, the script calculates feature-wise similarities for all houses in the dataset as compared to the user input house.
    b. Each feature similarity is scaled by a weight and combined into a `total_distance`.
3. **Ranking**:
    a. Houses are ranked based on `total_distance`, with lower similarities indicating higher similarity.
    b. The ranked results and detailed metrics are returned to the user in the result.html. Here, various interactive visualizations and analytics are displayed and utilized by the user.
    c. Once the desired comparable homes are selected, the user clicks the "Generate Report" button and a new tab opens to a customized report in protest.html.

B. User Interface

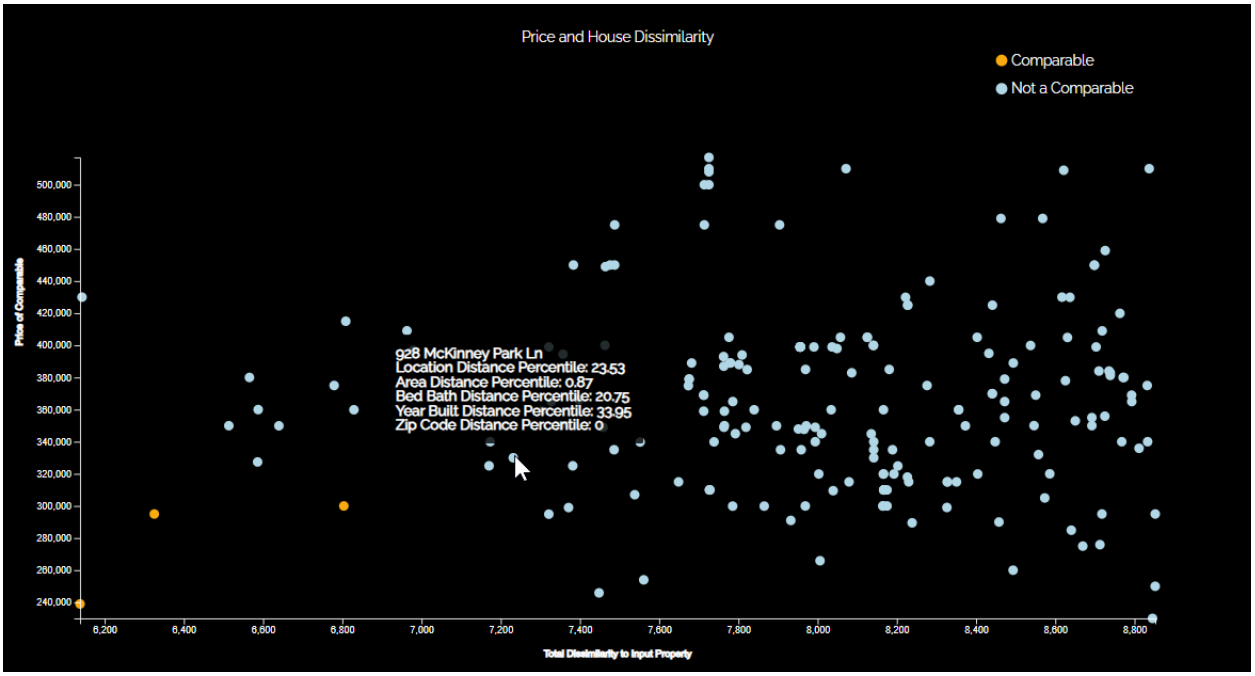
Our user interface needs to do several things. It needs to provide a way for the user to input information about their house, allow the user to browse comparable houses, provide the user with the ability to select houses to include in their set of comparables, and it needs to show the user relevant information about the similarities of their current set of comparables and possible comparables.

The user interface provides a way for the user to input information about their house with a form that cycles through various questions. It also allows users to browse comparable homes in several ways. First, through an interactive map that they can browse, where comparable homes and their features are shown:



8706 Torcello St
Bedrooms: 4
Bathrooms: 3
Sq Ft: 1585
Price: $239000

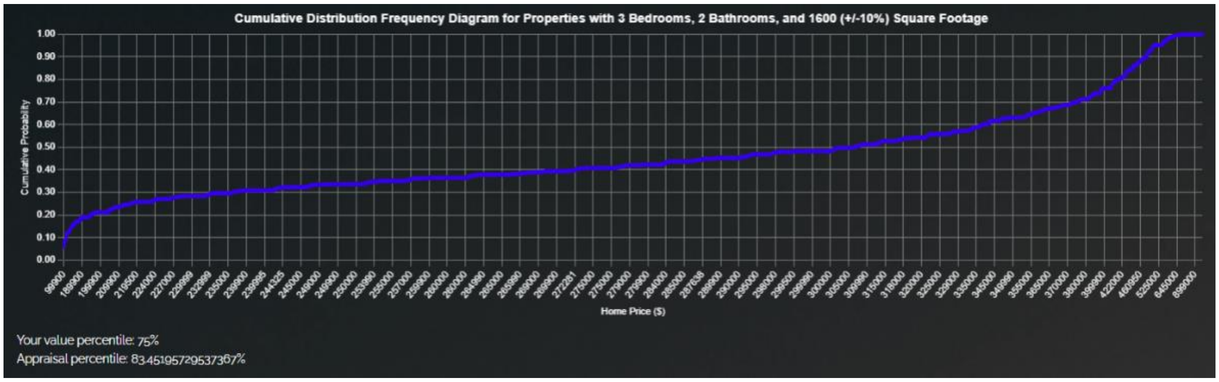
Users can also browse comparables using a Dissimilarity vs Price scatter plot, which also has tooltips informing the user about where the dissimilarity between the two houses comes from:



928 McKinney Park Ln
Location Distance Percentile: 23.53
Area Distance Percentile: 0.87
Bed Bath Distance Percentile: 20.75
Year Built Distance Percentile: 33.95
Zip Code Distance Percentile: 0

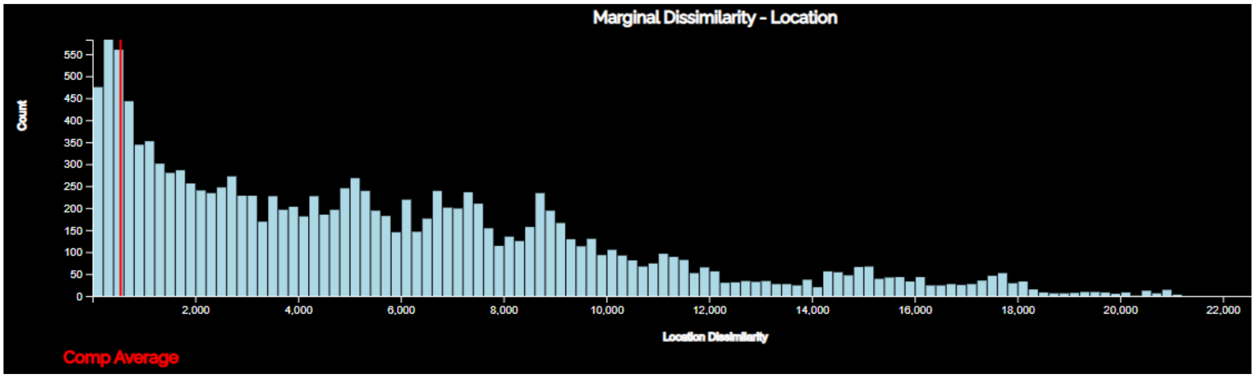
Our user interface allows users to change their selected properties by simply clicking on any of the dots that represent houses in the previous two visualizations. This also updates all other graphs which rely on information from the selected properties.

Our user interface provides the users with multiple ways to understand their house as well as their comparable homes and the reason the houses are similar or dissimilar. All these visualizations (except the CDF) change depending on the selected set of similar properties. Firstly, high-level information regarding the distribution of the pricing the users should expect to see is shown with a CDF generated from similar housing:



Your value percentile: 75%
Appraisal percentile: 83.45195729537367%

Next, high level property details are given to the user in the form of a table. Finally, information about the similarity, by home feature, between the set of houses and the user's house is provided in a series of histograms.

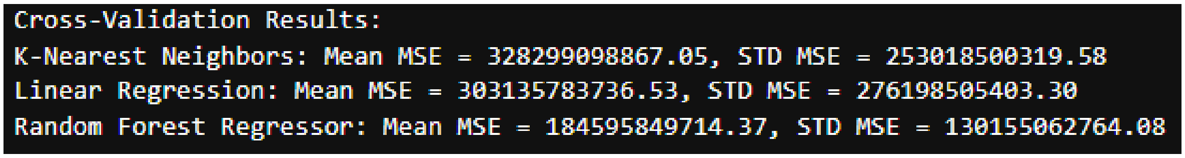


V. Data

The project uses Zillow data from Kaggle, specifically homes for sale in Houston, Texas, in 2024. We extracted relevant details from a 6.6 GB JSON file into a 45.6 MB CSV file using Python. After cleaning the dataset, it consists of 15,157 rows and 723 columns, focusing on pricing prediction and visualization. Key property features include the number of bedrooms and bathrooms, living area, lot size, year built, appliances, interior finishes, security features, and construction materials. One issue working with the data was that at first calculating these similarity metrics took a long time, upwards of 80 seconds to compare one house to the rest of the dataset. We were able to substantially reduce the time it took to make these comparisons by changing from looping through the rows and doing the comparison individually to using matrices to calculate the marginal similarity from a house to the whole dataset at once.

VI. Experiments and Results

To evaluate our approach, we benchmarked our method of comparable selection and price prediction with other machine learning models, such as k-nearest neighbors, linear regression, and random forest regression.[1,4] Additionally, we prepared plots such as difference in similarity vs difference in price to see how well our similarity metrics capture the variance in price. We tested the three machine learning algorithms with default hyper parameters and 10-fold cross-validation. The results (below) show that random forest regression performed the best (lowest mean square error), followed by linear regression, and finally k-nearest neighbors.



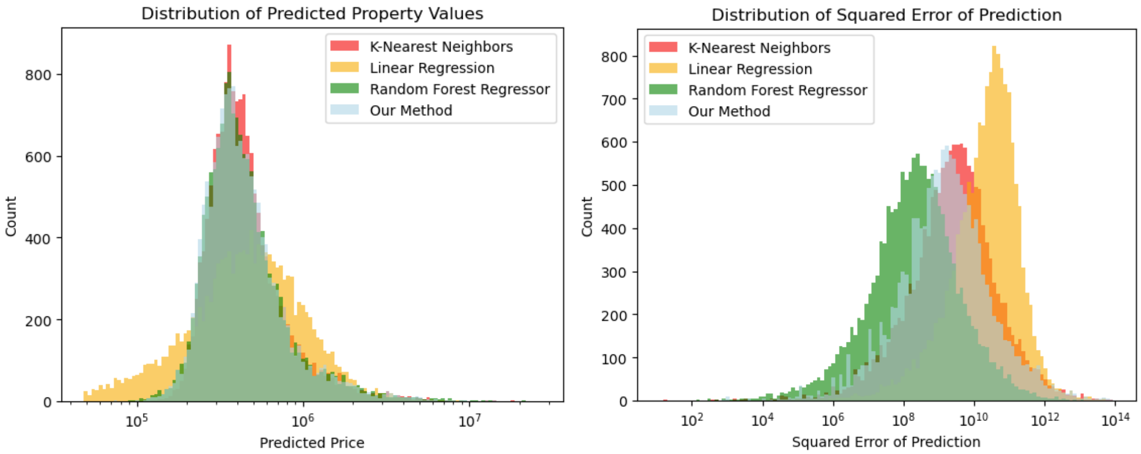
```
Cross-Validation Results:
K-Nearest Neighbors: Mean MSE = 328299098867.05, STD MSE = 253018500319.58
Linear Regression: Mean MSE = 303135783736.53, STD MSE = 276198505403.30
Random Forest Regressor: Mean MSE = 184595849714.37, STD MSE = 130155062764.08
```

Cross-validation results of 3 machine learning algorithms showing random forest performed best (lowest MSE).

Experiments were run to test the predictions of three randomly selected houses in our dataset (see the Appendix). Our current similarity metric algorithm is outperforming KNN and linear regression, but underperforming random forest. Our algorithm is also consistently underestimating the price of the user input houses. More work needs to be done to fine tune the weighting parameters for our distance metric. However, our algorithm remains preferable due to its simplicity and ability to pinpoint comparable houses, which are vital to a successful tax protest.

The appendix displays the cumulative distribution function (CDF) plots that were generated for the three machine learning algorithms. Each CDF plot was created with user input data for a specific house. The dashed blue line displays the predicted house price from each model. The red line is the predicted price using our custom similarity metric algorithm and is the same in each plot.

We also found the prediction error for every row in our data using our method with the closest three houses and each of the three different models. These experiments show that our model predicts values in a distribution very similar to the K-Nearest Neighbors and Random Forest Regressor models and outperforms all models except the random forest regressor.



VII. Conclusions and Discussion:

To summarize, we have created an application that allows homeowners to more easily fight increases in their property taxes by providing them with a tool that they can use to find good comparables that have a low price. Users are able to find these comparables with the use of a custom similarity metric that we defined, and interactive visualizations we provide to choose homes with a low price that aren't too dissimilar from their input home.

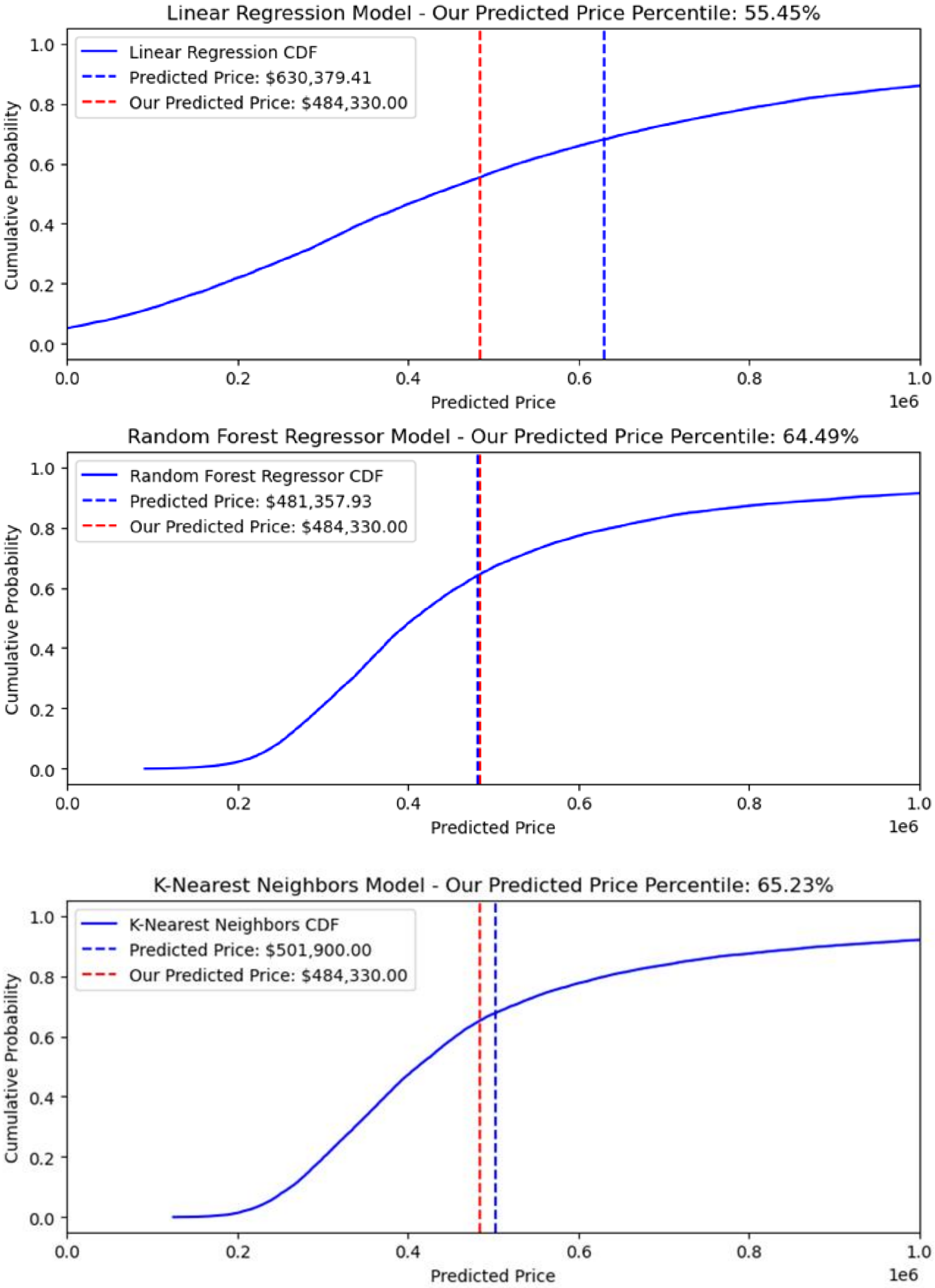
There are some limitations to our method. Firstly, we only have data for homes in the Houston area. Secondly, although we tried to scrape data about the user's input house from Zillow, we were unable to. This inability means that all information about the user's home must be provided by the user, allowing for incorrect data to be inputted, and limiting the variables we can use for our similarity metric to things that are easy for the user to input (it would be unreasonable to ask the user to input information about 100+ features of their house). Potential future expansions include expansion of the dataset we use, expansion of the features used, and the ability to gather information about the user's house given just the address.

Distribution of Effort: Trevor was the team leader, contributed the majority of the project coding, and also contributed to report writing. Calvin cleaned the data, designed the algorithms, python and d3 coding, and supported the report. Noumik contributed to the coding and the report writing. Tyler setup the github repository, made the poster, and contributed to the written reports.

VIII. References

1. Mohd, T., Jamil, N.S., Johari, N., Abdullah, L., Masrom, S. (2020). An Overview of Real Estate Modelling Techniques for House Price Prediction. In: Kaur, N., Ahmad, M. (eds) Charting a Sustainable Future of ASEAN in Business and Social Sciences. Springer, Singapore. https://doi.org/10.1007/978-981-15-3859-9_28
2. Yavuz Ozalp, Ayse, and Halil Akinci. "The use of hedonic pricing method to determine the parameters affecting residential real estate prices." Arabian Journal of Geosciences 10.24 (2017): 535.
3. Pagourtzi, Elli, et al. "Real estate appraisal: a review of valuation methods." Journal of property investment & finance 21.4 (2003): 383-401.
4. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R: Chapters 2 Statistical Learning, 3 Linear Regression, and 8 Tree-Based Methods. Springer, New York.
5. Voss, C. (2016). Never Split the Difference: Negotiating as if Your Life Depended on it, Chapter 6: Bend Their Reality. HarperCollins, New York.
6. Voss, C. (2016). Never Split the Difference: Negotiating as if Your Life Depended on it, Chapter 9: Bargain Hard. HarperCollins, New York.
7. Texas Comptroller of Public Accounts. (2021). Texas property tax: Taxpayer remedies (Publication No. 96-297-21. https://comptroller.texas.gov/taxes/property-tax/docs/96-297-21.pdf.
8. Gröbel, S., & Thomschke, L. (2018). *Developing automated valuation models for estimating property values: A comparison of global and locally weighted approaches*. *Annals of Operations Research*. https://doi.org/10.1007/s10479-018-3125-2
9. Pace, K. (1998). *Appraisal using generalized additive models*. *Journal of Real Estate Research, 15*(1), 77-99.
10. Wang, K., & Wolverton, M. (2002). *Real Estate Valuation Theory*. Springer.
11. Thériault, Marius, et al. "Modelling interactions of location with specific value of housing attributes." Property Management 21.1 (2003): 25-62.
12. Conway, Jennifer (2018). Artificial Intelligence and Machine Learning: Current Applications in Real Estate." MIT Master's Thesis. https://dspace.mit.edu/handle/1721.1/120609

IX. Appendices



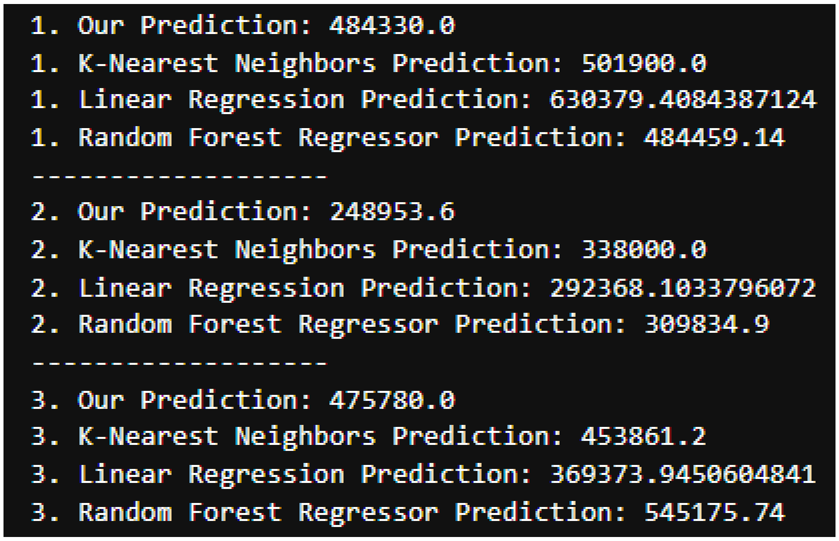
Linear Regression Model - Our Predicted Price Percentile: 55.45%

Random Forest Regressor Model - Our Predicted Price Percentile: 64.49%

K-Nearest Neighbors Model - Our Predicted Price Percentile: 65.23%

CDF plots for the machine learning algorithms showing how close our algorithm (in red) is from the model predicted price (dashed blue line).

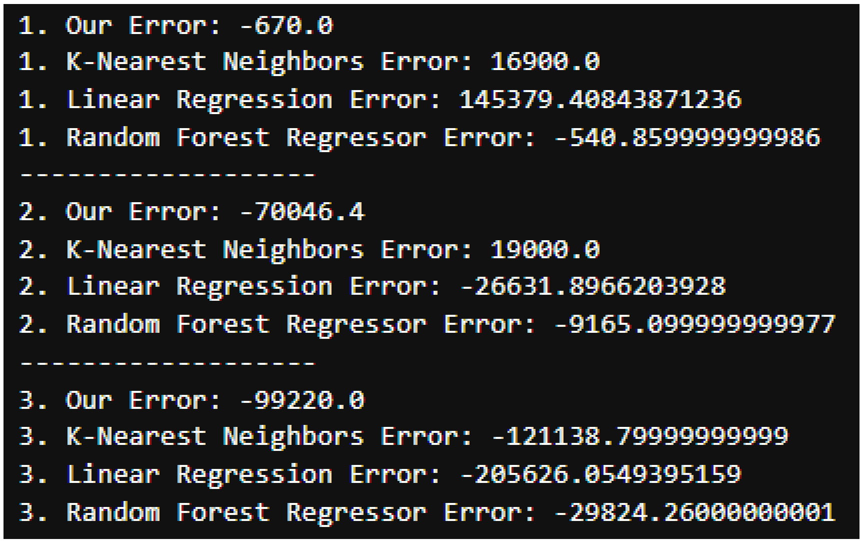
Experiments were run to test the predictions of three randomly selected houses in our dataset, houses with index 1955, 1825, and 19018 from our processed dataset (note that the indexes were not reset after dropping NA values, which is why we have an index of 19018, even though 19018 > 15157). From the results, we see that our current similarity metric and method combination consistently underestimates the price of the house (Appendix-Figure 3).



```
1. Our Prediction: 484330.0
1. K-Nearest Neighbors Prediction: 501900.0
1. Linear Regression Prediction: 630379.4084387124
1. Random Forest Regressor Prediction: 484459.14
-------------------
2. Our Prediction: 248953.6
2. K-Nearest Neighbors Prediction: 338000.0
2. Linear Regression Prediction: 292368.1033796072
2. Random Forest Regressor Prediction: 309834.9
-------------------
3. Our Prediction: 475780.0
3. K-Nearest Neighbors Prediction: 453861.2
3. Linear Regression Prediction: 369373.9450604841
3. Random Forest Regressor Prediction: 545175.74
```

1Experiments to test the machine learning algorithms on 3 houses.

Our method had a smaller error on the predicted values for some models and houses but was greater on others.



```
1. Our Error: -670.0
1. K-Nearest Neighbors Error: 16900.0
1. Linear Regression Error: 145379.40843871236
1. Random Forest Regressor Error: -540.859999999986
-------------------
2. Our Error: -70046.4
2. K-Nearest Neighbors Error: 19000.0
2. Linear Regression Error: -26631.8966203928
2. Random Forest Regressor Error: -9165.099999999977
-------------------
3. Our Error: -99220.0
3. K-Nearest Neighbors Error: -121138.79999999999
3. Linear Regression Error: -205626.0549395159
3. Random Forest Regressor Error: -29824.26000000001
```

Error measures for the machine learning algorithm tests on 3 houses. Random forest has the lowest error.

Literature Survey Appendix:

Mohd, T., Jamil, N.S., Johari, N., Abdullah, L., Masrom, S. (2020). An Overview of Real Estate Modelling Techniques for House Price Prediction. In: Kaur, N., Ahmad, M. (eds) Charting a Sustainable Future of ASEAN in Business and Social Sciences. Springer, Singapore. https://doi.org/10.1007/978-981-15-3859-9_28

a. This paper reviews various real estate modelling techniques and lays out the advantages and disadvantages that these various modelling techniques have shown in housing pricing prediction. The modelling techniques are varied, including Neural Networks, Support Vector Machines, Random Forest, and Ridge Regression.

b. This paper will be useful for our project, as it will help guide our decision making for the specific model that is used to predict the actual value (or range of values) that a certain property is worth. Additionally, it introduces the idea that under certain conditions (say, when there are very few comparables) we may want to have different models to predict prices.

c. The paper lists shortcomings for each modelling technique but does not propose mitigation techniques. As such, we will attempt to find specific solutions to the problems listed as disadvantages for each of the models.

Yavuz Ozalp, Ayse, and Halil Akinci. "The use of hedonic pricing method to determine the parameters affecting residential real estate prices." Arabian Journal of Geosciences 10.24 (2017): 535.

a. This paper builds a hedonic pricing model of residential real estate in Artvin, Turkey. A hedonic pricing model is one in which the price of the house is determined through the sum of individual characteristics of the house, usually in a linear regression model. The paper finds that of the 18 attributes looked at, only 5 of them have a statistically significant impact on price in the hedonic model.

b. This paper is beneficial to the project, as we may pursue a hedonic pricing model for our project, and the paper demonstrates potential shortcomings with this type of model (such as correlation between attributes and lack of data about things such as building materials used).

c. The model they used may suffer from high rates of correlation between the various attributes. Transformation methods, such as Principal Component Analysis, may be used to get rid of this correlation.

Pagourtzi, Elli, et al. "Real estate appraisal: a review of valuation methods." Journal of property investment & finance 21.4 (2003): 383-401.

a. This paper compares and reviews various valuation methods, including methods that use models such as Artificial Neural Networks, Spatial Analysis Methods, and hedonic pricing methods and more traditional valuation methods, such as the comparable method.

b. This paper is beneficial to the project, as it lays out the steps usually taken to perform each valuation method. For example, when looking at the comparable method of pricing, the paper suggests finding the distance between different properties based on the difference of their attributes.

c. The paper does not go into how different models and methods may be used in conjunction with one another to gain a more accurate prediction. We may attempt to combine various pricing techniques to achieve a more accurate prediction.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R: Chapters 2 Statistical Learning, 3 Linear Regression, and 8 Tree-Based Methods. Springer, New York.

a. These chapters cover regression, k-nearest neighbors, and regression tree models and how to implement them.

b. The benefit to the project is that there is an example of using regression to estimate housing prices and building confidence intervals. Also of interest in the book are model testing methods.

c. The shortcomings are that the book is intended to be used with R, whereas we will likely be using Python. Since we will need to find comparable properties, k-nearest neighbors or classification may be more appropriate models.

Voss, C. (2016). Never Split the Difference: Negotiating as if Your Life Depended on it, Chapter 6: Bend Their Reality. HarperCollins, New York.

a. This chapter involves “anchoring”, “fairness”, and deadlines.

b. It’s applicable to the project in supporting the wording of final deliverables and shaping the analysis (potentially creating a “low-ball” valuation to counter the appraisal district’s initial offer, then possibly walking the number up to a more likely value).

c. The shortcoming is that the chapter is designed more for conversations, so it may be better for the customer to read the book and apply the principles in the dynamic meetings, rather than sticking to a fixed script.

Voss, C. (2016). Never Split the Difference: Negotiating as if Your Life Depended on it, Chapter 9: Bargain Hard. HarperCollins, New York.

a. This chapter is on haggling during negotiating, how to frame the discussion, and tactics to use (and not use).

b. There is an example of a tenant negotiating with their landlord’s agent and successfully fighting a rise in rent, something that is directly applicable to the project. Part of the service in our project is to help homeowners negotiate lower property home valuations (translating into lower property taxes). This chapter will be helpful in wording and coaching the deliverables of our project.

c. Many of the points, such as not making the first offer, in the text are not applicable in our case, which is a shortcoming of the source for our purposes.

Texas Comptroller of Public Accounts. (2021). Texas property tax: Taxpayer remedies (Publication No. 96-297-21. https://comptroller.texas.gov/taxes/property-tax/docs/96-297-21.pdf.

a. The document "Texas Property Tax: Taxpayer Remedies" outlines the procedures for disputing property tax appraisals and the steps taxpayers can take if they disagree with the assessed value of their property. It also details the appeals process, including how to request corrections and pursue refunds for errors or overpayments in property taxes.

b. This document is valuable for our project because it outlines the exact steps individuals need to take to dispute their property tax assessments in Texas. With this information, we can better guide people through the process of challenging their appraisals, filing appeals, and seeking corrections or refunds, making it a useful resource for helping taxpayers navigate the system more effectively.

c. A potential shortcoming of this document for our property tax fighter project is its use of legal and procedural language, which may be difficult for individuals to understand without additional guidance.

Gröbel, S., & Thomschke, L. (2018). *Developing automated valuation models for estimating property values: A comparison of global and locally weighted approaches*. *Annals of Operations Research*. https://doi.org/10.1007/s10479-018-3125-2

a. This paper explores the effectiveness of using automated valuation models (AVMs) in real estate by comparing global models to locally weighted models. It highlights that locally weighted approaches can provide more accurate property value predictions by incorporating regional variations in the data, which global models might overlook.

b. This paper is valuable for our Houston-centric property tax fighter project because it emphasizes the benefits of using locally weighted models, which can account for regional variations in property values. This approach will help us improve the accuracy of our valuation models as we expand the project's focus to other areas beyond Houston.

c. A shortcoming of this paper is that we have not seen AVMs in our coursework currently. This would require some self-study on the implementation of AVMs and the feasibility of using them with the data we have available.

Pace, K. (1998). *Appraisal using generalized additive models*. *Journal of Real Estate Research, 15*(1), 77-99.

a. This paper explores the use of generalized additive models (GAMs) in real estate valuation, focusing on their flexibility in capturing non-linear relationships between property attributes and their market values. It highlights how GAMs can improve the accuracy of property appraisals compared to traditional linear models by allowing for more nuanced interpretations of the factors that influence prices.

b. This will be useful for our property tax fighter project because it demonstrates how generalized additive models (GAMs) can handle the non-linear relationships often found in real estate data, leading to more precise home value estimates. This improved accuracy in appraisals will strengthen our efforts to identify over-assessed properties in Houston, allowing us to better advocate for fair property tax adjustments for homeowners.

c. To address the shortcomings of the paper, we can improve our project's approach by integrating diverse data sources to enhance the accuracy and consistency of our property valuations in Houston. Additionally, we can create clear visualizations and explanatory tools to simplify the interpretation of generalized additive models (GAMs), making the results more accessible for homeowners seeking to understand their property tax assessments.

Wang, K., & Wolverton, M. (2002). *Real Estate Valuation Theory*. Springer.

a. This book provides a breadth of information covering both traditional methods like comparable sales and advanced approaches like statistical models. It focuses on the application of these techniques to understand market dynamics and emphasizes data analysis in modern valuation practices, making it ideal for scalable uses like property tax assessments.

b. This book would be highly relevant to our project because it offers a solid grounding in the theories that underlie modern real estate valuation models, which are essential for developing accurate and defendable property tax assessments.

c. A shortcoming is that it would require a significant amount of time and effort to fully grasp the underlying theories, which may be challenging for us without a strong background in real estate valuation. This learning curve could slow down the implementation of practical valuation techniques.

Thériault, Marius, et al. "Modelling interactions of location with specific value of housing attributes." Property Management 21.1 (2003): 25-62.

a. This article goes over ways to handle the interactions between spatial attributes and attributes of a specific house when using a hedonic pricing model. These interactions must be handled as the hedonic pricing models (which are usually based on some sort of linear regression model), as these models assume homoskedasticity in their predictions as well as multicollinearity and multicollinearity.

b. This article will be useful for us as we may try to implement a hedonic pricing model as an estimator for housing prices, and thus we may have to use these techniques.

c. One shortcoming of this article is that it only solves these issues for regression models and does not go into solving issues for other hedonistic pricing models, such as those created by Artificial Neural Networks.

Conway, Jennifer (2018). Artificial Intelligence and Machine Learning: Current Applications in Real Estate." MIT Master's Thesis. https://dspace.mit.edu/handle/1721.1/120609

a. This thesis provides a discussion on how machine learning is being used by a number of companies. It breaks out various data science and real estate tasks (such as analytics, valuation, risk, etc.) and charts which companies are doing them.

b. The thesis is useful for its analysis and categorization of real estate machine learning and AI companies. We can check the websites of these companies to get an idea for how they display their analyses. The machine learning algorithms section may be helpful to generate ideas for which algorithms to use in our project.

c. The shortcoming of this thesis is that it is largely untechnical and mostly provides general theory about how machine learning is useful in the commercial real estate market.