ISyE 6740 - Fall 2023
Project Proposal

Team Member Name:	Seung Woo Choi
Project Title:	Evaluating Assessed and Sale Values to Determine Tax Implications in Real Estate
Please include (at least) the following sections.	Problem Statement, (Optional) Data Source, Methodology, Evaluation and Final Results

Introduction

In real estate, the sale price of a home is one of the most important figures for prospective buyers and sellers. Before a home can be transacted, however, a local government assessor must provide an assessed value for a home. The assessed value is a value that assessors provide to calculate property taxes (Kagan). Homeowners are responsible for paying property taxes every year as long as they maintain ownership of a property. Thus, the property tax associated with a home is an important factor when deciding on buying a property, and a home where the sale price exceeds the assessed value is desirable since the homeowner owes property taxes based on the home's assessed value.

Predicting the sale price of real estate tends to be the subject of many projects. This project, however, will instead focus on determining which of the two - i.e. assessed value or sale value - is higher or lower to help prospective buyers and sellers understand the initial tax burden of real estate. While property taxes can change over time, this attribute is an important consideration in the home purchasing process and will be the focus of this project.

Problem Statement

In this project, my goal is to build binary classification models that can predict whether the sale price of a home is higher (i.e. assigned a value of 1) or lower (i.e. assigned a value of 0) than the assessed value. A response variable called great_value was created to reflect this relationship between the assessed value and sale price. I plan to utilize the accuracy score to evaluate the performance of the models and aim to develop models that can outperform the baseline established by the majority class (i.e. 86.6% accuracy score or higher). The accuracy score represents the number of correct predictions divided by all predictions.

Through this project, I hope to perform feature engineering, create insightful visualizations, and build strong predictive models to enable prospective home buyers and sellers to make informed decisions when transacting on homes and understand potential tax implications.

Hypothesis

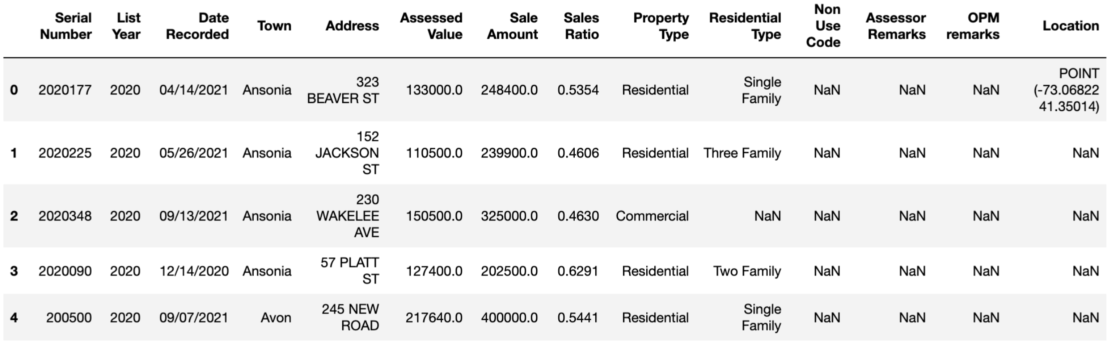
The primary hypothesis that this project seeks to answer is the following: **Can a binary classification model be constructed to predict the label of a property transaction with an accuracy score of 86.6% or higher?** Other areas of interest include determining the features that play important roles in training the model and visualizing the decision boundary of each model to understand how the model is making binary predictions.

(Optional) Data Source

The dataset used in this project is titled "Real Estate Sales 2001-2020 GL" and was found on data.gov. The data was downloaded as a single csv file, last updated on August 12, 2023, and was published by data.ct.gov. The csv file consists of 14 features, which includes serial_number, list_year, date_recorded, town, address, assessed_value, sale_amount, sales_ratio, property_type, residential_type, non_use_code, assessor_remarks, opm_remarks, and location.

The original dataset was unlabeled and required creating a response variable to make sense of the data. To convert the problem into a supervised learning problem, the assessed_value and sale_amount variables were combined to create the response variable, great_value. A row was labeled as "1" where sale_amount was greater than assessed_value, whereas a row was labeled as "0" where sale_amount was less than or equal to the assessed_value. The creation of a binary response variable transformed the analysis into a binary classification problem.

The original data set contained 5 numerical features and 9 categorical features, so it was imperative to perform feature engineering to design features that can improve the results of predictive models. Techniques like creating dummy variables, encoding frequencies, and encoding binary labels, were applied to expand the feature space. The variables, sale_amount and sales_ratio, were dropped from the test data since they provide information that can directly be used to determine the response variable. The data was unbalanced with 13.4% labeled as 1 and 86.6% labeled as 0, so re-sampling methods may be needed to strengthen the predictive capabilities of the models. The original data can be seen below.



	Serial Number	List Year	Date Recorded	Town	Address	Assessed Value	Sale Amount	Sales Ratio	Property Type	Residential Type	Non Use Code	Assessor Remarks	OPM remarks	Location
0	2020177	2020	04/14/2021	Ansonia	323 BEAVER ST	133000.0	248400.0	0.5354	Residential	Single Family	NaN	NaN	NaN	POINT (-73.06822 41.35014)
1	2020225	2020	05/26/2021	Ansonia	152 JACKSON ST	110500.0	239900.0	0.4606	Residential	Three Family	NaN	NaN	NaN	NaN
2	2020348	2020	09/13/2021	Ansonia	230 WAKELEE AVE	150500.0	325000.0	0.4630	Commercial	NaN	NaN	NaN	NaN	NaN
3	2020090	2020	12/14/2020	Ansonia	57 PLATT ST	127400.0	202500.0	0.6291	Residential	Two Family	NaN	NaN	NaN	NaN
4	200500	2020	09/07/2021	Avon	245 NEW ROAD	217640.0	400000.0	0.5441	Residential	Single Family	NaN	NaN	NaN	NaN

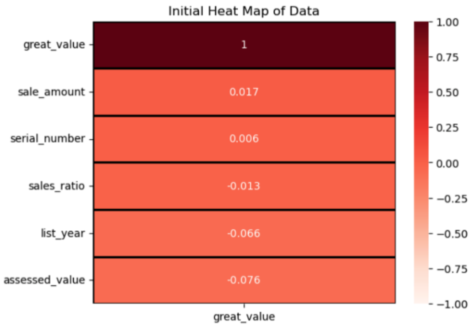
Methodology

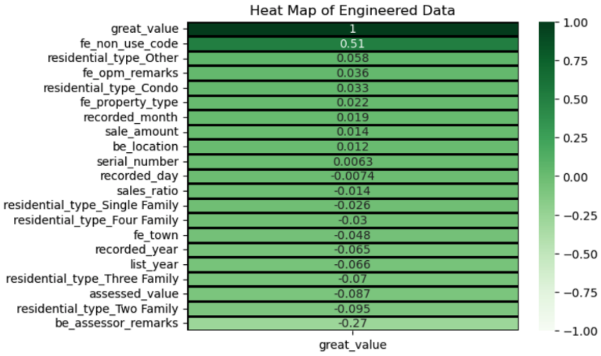
The methodology involves the following steps: (1) data collection, (2) data cleaning and preparation, (3) feature engineering, (4) exploratory data analysis, (5) pre-processing and modeling, and (6) evaluation.

The data collection step simply involved reading in the csv file from data.gov.

The data cleaning and preparation required more steps, such as renaming the columns for easier access, performing data type conversions to reformat variables into the appropriate data types, and handling missing values by removing rows where less than 1% of the data was missing and imputing null values with indicator values. The response variable, great_value, was also created using two other variables.

Various techniques were utilized to engineer new features. To begin, a stratified 70:30 train/test split was performed to prevent data leakage during the feature engineering step and to split the response variable evenly given the imbalance in the data. I employed techniques such as variable expansions, frequency encoding, getting dummies for categorical variables, binary encoding, and dropping unneeded variables to create numerical inputs for the machine learning models. I removed sale_amount and sales_ratio from the testing data since they are directly related to the response variable. Below are two visuals illustrating the impact of feature engineering on the feature space in the data set.





I performed exploratory data analysis (EDA) only on the training data since the testing data should technically be unseen data. I create multiple visualizations, such as heat maps before and after feature engineering, 1-D KDE plots, 1-D histograms, box plots, and scatter plots. I also reviewed summary statistics from each numerical variable to see if any unusual patterns could be detected.

For the preprocessing steps, I plan to scale data as necessary, apply re-sampling methods to see if they improve the accuracy score, and reformat the data into the required inputs for each model. For the modeling portion, I plan to utilize 3 to 5 different types of machine learning models to solve this binary classification problem.

To evaluate the performance of the statistical models, I plan to compute the accuracy score, plot the confusion matrix, and visualize the decision boundary.

Evaluation and Final Results

Once the modeling process is complete, I plan to assess the models utilizing three different evaluation criteria: (1) accuracy score, (2) confusion matrix, and (3) decision boundary. Together, these evaluation metrics and plots should provide enough insights to help me decide on the best performing model to capture the primary hypothesis.

I will also discuss the final results once the modeling process is finished.

References

Kagan, Julia. \Assessed Value: Definition, How It's Calculated, and Example".
\item Investopedia,
\item https://www.investopedia.com/terms/a/assessedvalue.asp#:~:text=Assessed%20value%20is

\Real Estate Sales 2001-2020 GL". Data.gov,
\item https://catalog.data.gov/dataset/real-estate-sales-2001-2018. Accessed 11
\item Oct. 2023.