Predictive Modeling for House Sale Prices

Milestone 4

Lincoln Brown

DSC550-T302

Professor Werner

Introduction

This project aims to predict the sale price of houses for sale using several features such as size and location. The data used in this project was obtained from the Kimball County Assessor's GIS website. The problem I wanted to solve with this project was to have a better understanding of the real estate market conditions. Having a predictive model that uses publicly available data to give me a better understanding of what a house is worth can guide any purchasing or selling decisions.

Predicting housing sale prices is crucial for various stakeholders including homeowners, real estate agents, and potential buyers. Accurate predictions provide stakeholders with informed decision-making so they can understand fair listing prices, negotiate fair deals, and identify economical purchases. This allows homeowners to better understand their home's value, which can help them set an agreeable price for the seller and buyer.

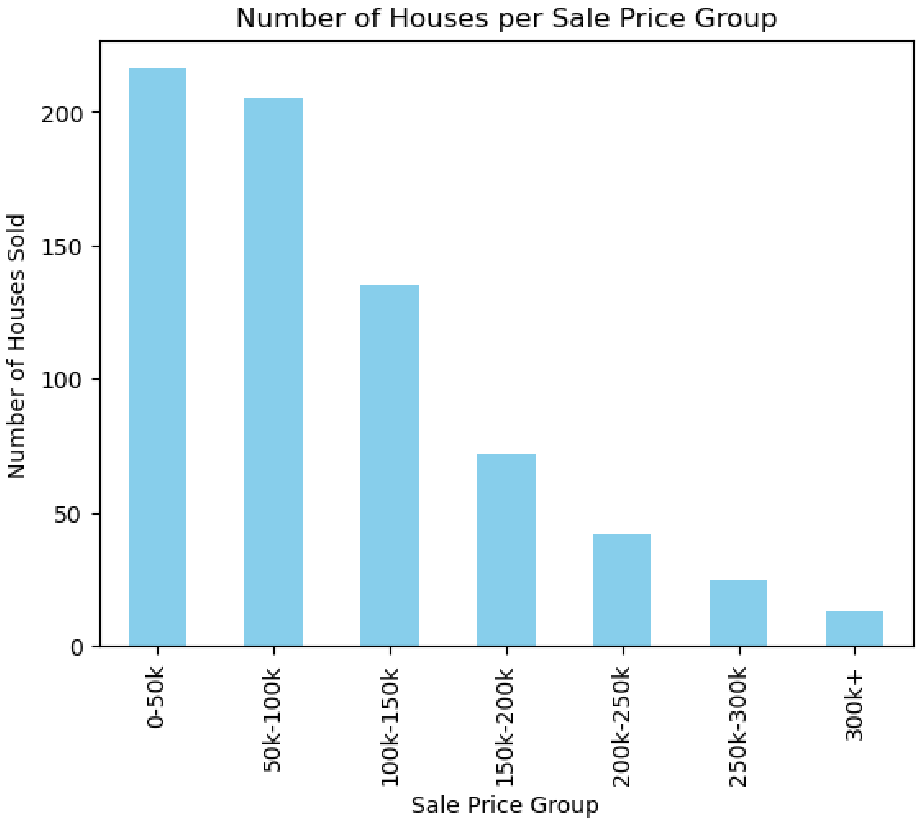
I obtained the data from a publicly available website https://kimball.gworks.com using an Application Programming Interface (API). The data on the website is uploaded by the Kimball County Assessor's office. There is a notice on the website that the data should not be used for preparing legal documentation. This disclaimer is to safeguard the assessor's office against any potentially misreported data. Despite this warning, the data is accurate enough for my purposes. Other counties in Nebraska also use similar websites, so it would be straightforward to obtain the data necessary to apply the same process for other counties in Nebraska.

Milestone 1

In Milestone 1, I started by pulling the data from the website. This involved obtaining a list of Parcel Ids for Kimball County and then contacting the API for property details for each property. Next, I used the details for each property to build a dataframe containing the information relevant to the project. Before any cleaning or exploratory data analysis began, the dataset contained 722 records and 12 features.

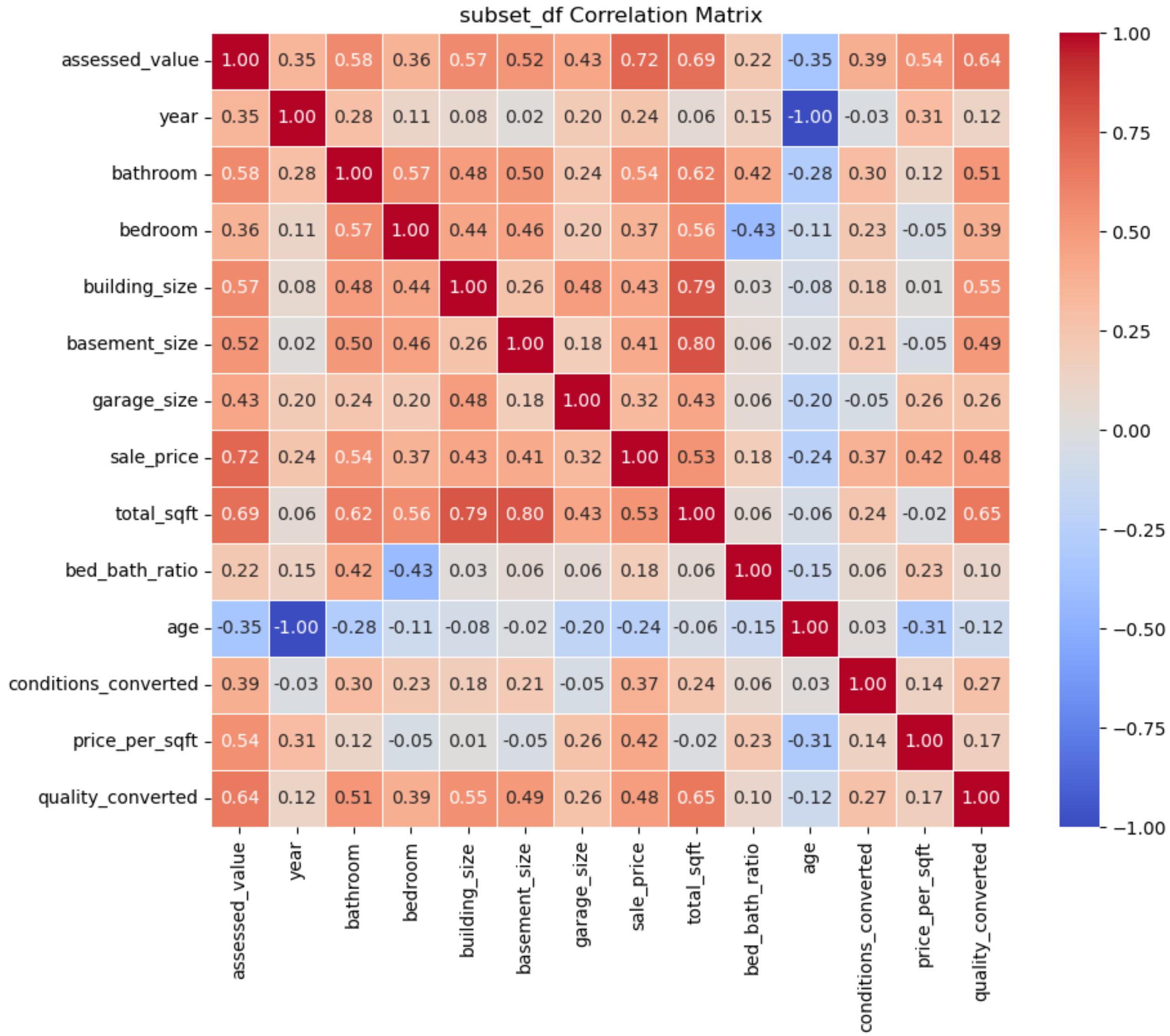
In this milestone, I began my exploratory data analysis of the data. I started by checking for duplicate records, looking at the features available for analysis, understanding their correlation with my target variable (sale price), and eliminating outliers. After cleaning, eliminating outliers, and feature creation, the dataframe consists of 708 records and 23 features.

Let us examine the target feature ‘sale_price’ with some visualizations:



From the bar graph above, we see that most of the houses are sold for less than $100,000.

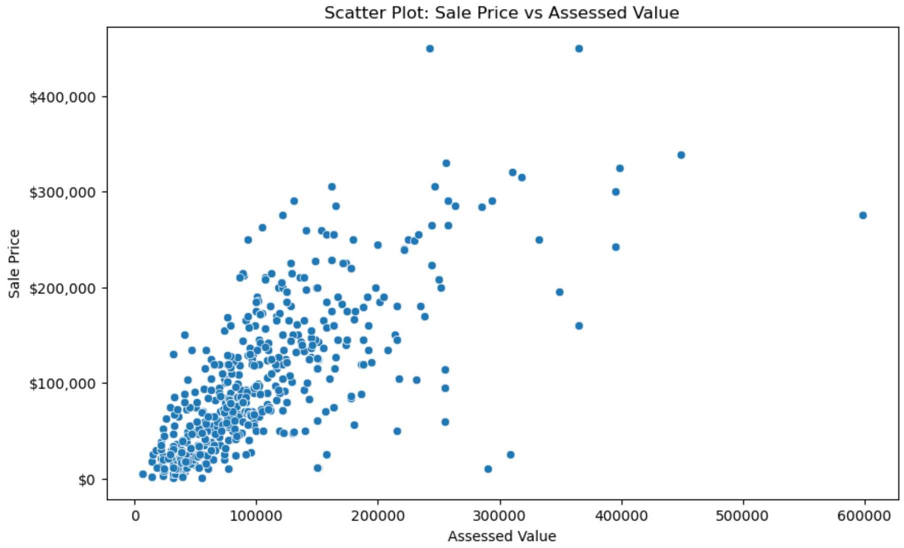
Below is the correlation matrix which shows the features that have the strongest correlation with the sale price.



	assessed_value	year	bathroom	bedroom	building_size	basement_size	garage_size	sale_price	total_sqft	bed_bath_ratio	age	conditions_converted	price_per_sqft	quality_converted
assessed_value	1.00	0.35	0.58	0.36	0.57	0.52	0.43	0.72	0.69	0.22	-0.35	0.39	0.54	0.64
year	0.35	1.00	0.28	0.11	0.08	0.02	0.20	0.24	0.06	0.15	-1.00	-0.03	0.31	0.12
bathroom	0.58	0.28	1.00	0.57	0.48	0.50	0.24	0.54	0.62	0.42	-0.28	0.30	0.12	0.51
bedroom	0.36	0.11	0.57	1.00	0.44	0.46	0.20	0.37	0.56	-0.43	-0.11	0.23	-0.05	0.39
building_size	0.57	0.08	0.48	0.44	1.00	0.26	0.48	0.43	0.79	0.03	-0.08	0.18	0.01	0.55
basement_size	0.52	0.02	0.50	0.46	0.26	1.00	0.18	0.41	0.80	0.06	-0.02	0.21	-0.05	0.49
garage_size	0.43	0.20	0.24	0.20	0.48	0.18	1.00	0.32	0.43	0.06	-0.20	-0.05	0.26	0.26
sale_price	0.72	0.24	0.54	0.37	0.43	0.41	0.32	1.00	0.53	0.18	-0.24	0.37	0.42	0.48
total_sqft	0.69	0.06	0.62	0.56	0.79	0.80	0.43	0.53	1.00	0.06	-0.06	0.24	-0.02	0.65
bed_bath_ratio	0.22	0.15	0.42	-0.43	0.03	0.06	0.06	0.18	0.06	1.00	-0.15	0.06	0.23	0.10
age	-0.35	-1.00	-0.28	-0.11	-0.08	-0.02	-0.20	-0.24	-0.06	-0.15	1.00	0.03	-0.31	-0.12
conditions_converted	0.39	-0.03	0.30	0.23	0.18	0.21	-0.05	0.37	0.24	0.06	0.03	1.00	0.14	0.27
price_per_sqft	0.54	0.31	0.12	-0.05	0.01	-0.05	0.26	0.42	-0.02	0.23	-0.31	0.14	1.00	0.17
quality_converted	0.64	0.12	0.51	0.39	0.55	0.49	0.26	0.48	0.65	0.10	-0.12	0.27	0.17	1.00

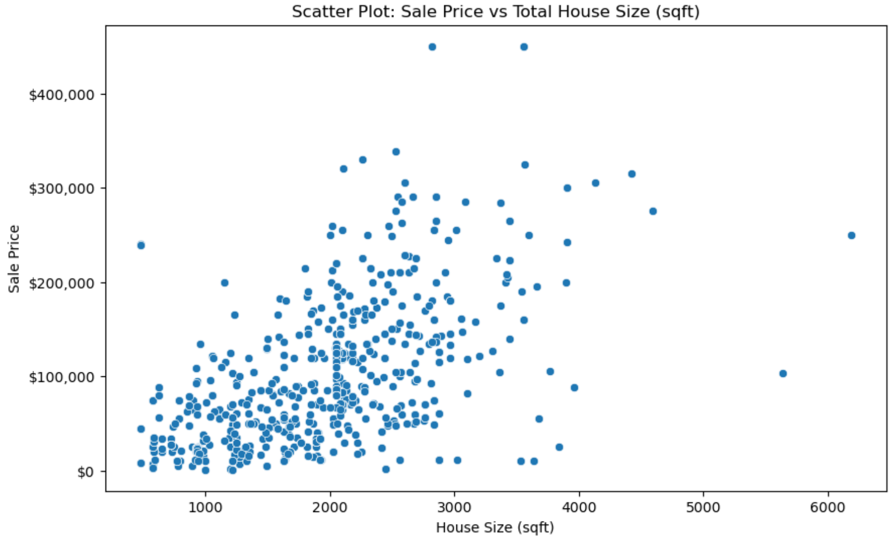
From the visualization above, we can see that ‘assessed_value’ has the strongest correlation with ‘sale_price.’

Using a scatterplot, this relationship is easier to visualize:



From the scatter plot above, we can see there is a slight linear correlation between assessed value and sale price. As the assessed value increases, the sale price increases.

The feature ‘total_sqft’ was the second feature that had the highest correlation with ‘sale_price,’ Let us look at this relationship more closely:



From the visualization above, we can see that the linear relationship between ‘total_sqft’ and ‘sale_price’ is not easily discernible. With such weak correlation between the two most correlated features, I concluded that a Linear Regression model was not likely going to be the most performant model.

Milestone 2

In Milestone 2, I focused on featured engineering and comparing different models. To start, I created eight additional features like ‘total_sqft’ and ‘bed_bath_interaction’ to improve the model’s predictive capabilities. Additionally, I created features like ‘year_group’ to better visualize the data and ‘sale_price_zscore’ to eliminate outliers. Not all houses have garages, so I had to clean up the ‘garage_size’ feature by imputing zeroes for houses that do not have a garage.

After the feature engineering, multiple regression models were trained and evaluated, including Linear Regression, Decision Tree, and Random Forest. Feature selection was guided using the data’s correlation matrix and feature importance analysis. The creation of the ‘bed_bath_ratio’ aimed to capture the relationship between the number of bedrooms and bathrooms. Furthermore, the ‘age’ feature was created to provide a more intuitive understanding of the age of a house.

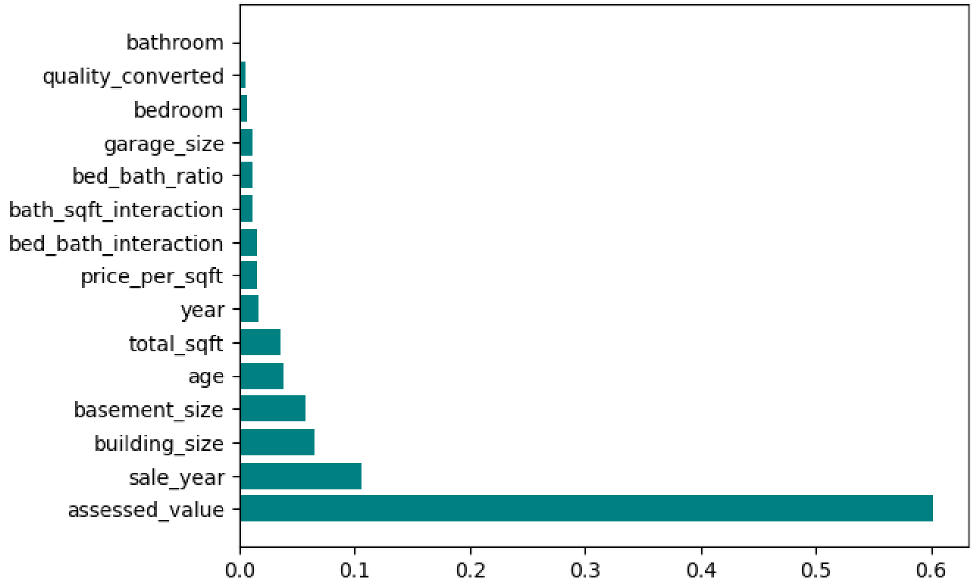
Additionally, I addressed the messy ‘condition’ and ‘quality’ features by mapping them to numerical categories and creating new columns (‘conditions_converted’ and ‘quality_converted’) for better representation. The need for these transformations arose from the inconsistent categorical nature of these features, which needed to be standardized before using them when training a predictive model. Furthermore, I eliminated the ‘year_group’ and ‘sale_price_group’ columns which were only used for data visualization purposes and ‘sale_price_zscore’ which was created for outlier detection.

The resulting dataset, named ‘features_encoded’ underwent thorough cleaning and preparation for the subsequent modeling phases. The resulting dataset contained 708 records and 16 features. Cleaning and preparation included handling missing data, and imputing zeroes for the ‘garage_size’ feature, to ensure the dataset was ready for training and testing. Additionally, I performed regression model training to assess the impact of feature engineering on predictive performance. I employed techniques like Principal Component Analysis (PCA) and Analysis of Variance (ANOVA) for feature selection. However, these efforts were not fruitful in improving model performance. Given the lack of improvements in model performance using feature selection, I decided to focus on Hyperparameter Tuning to improve performance in Milestone 3.

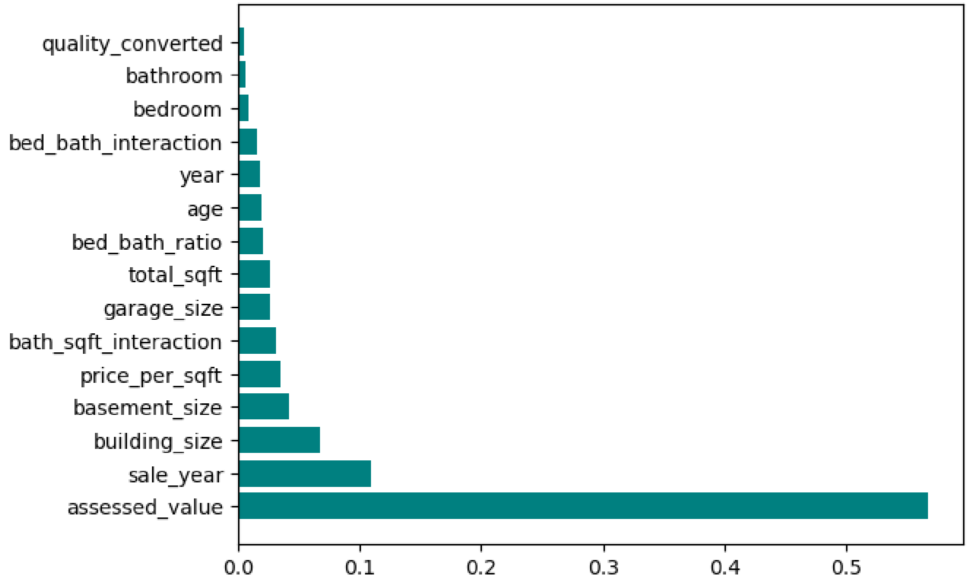
Milestone 3

Milestone 3 consisted of establishing a baseline model using a Dummy Regressor. The Decision Tree and Random Forest models performed the best, so I focused on fine-tuning these models using GridSearchCV hyperparameter optimization. Evaluation metrics such as R-Squared (R2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) were used to assess and compare the performance of these models. Feature importance analysis was conducted for both the Decision Tree and Random Forest models.

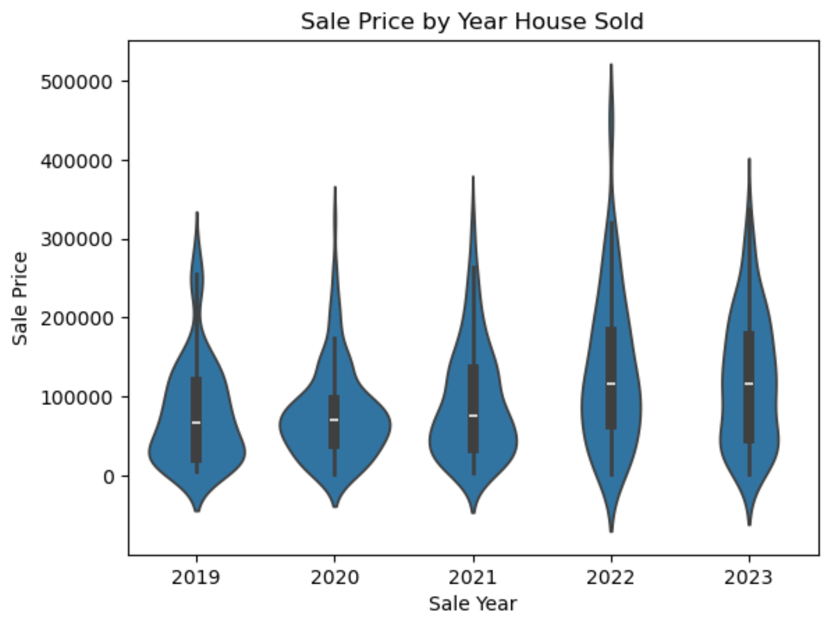
Here is a graph illustrating the feature importance for the Tuned Decision Tree:



Similarly, here is the feature importance for the Tuned Random Forest:



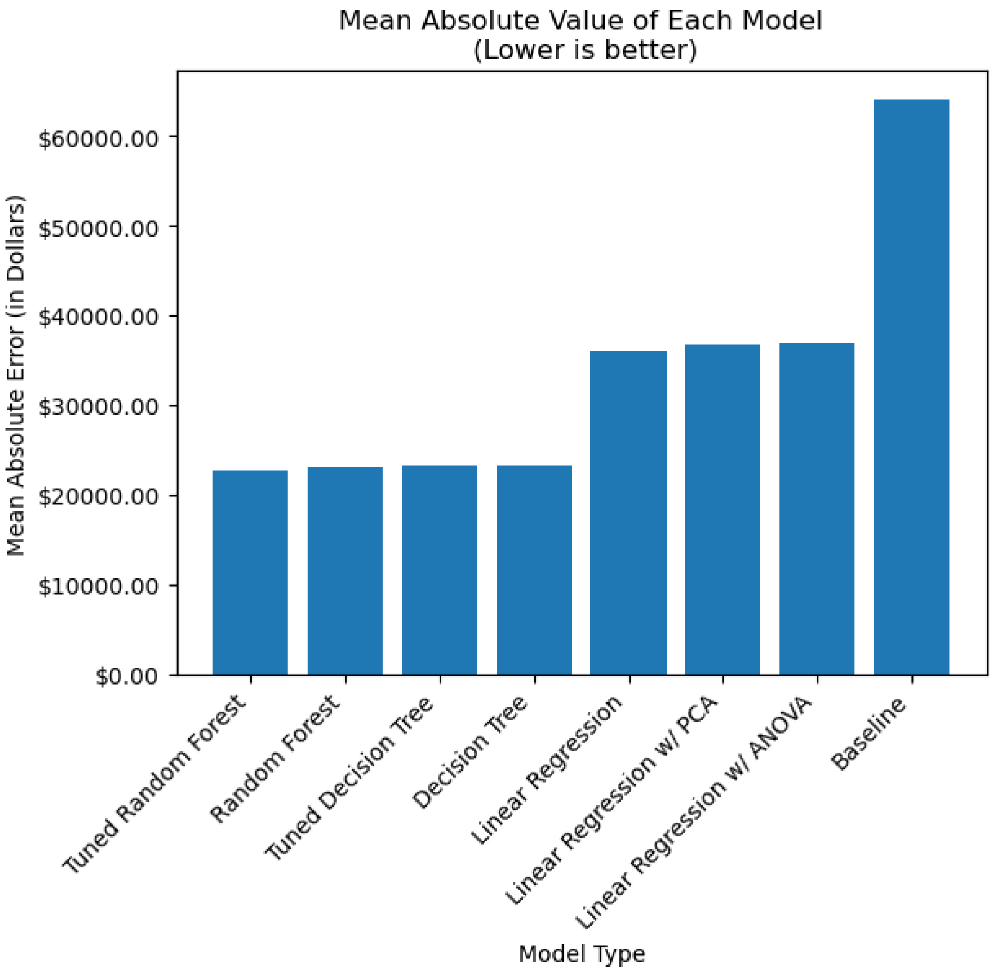
Both models have ’sale_year’ as the second-most important feature. We can observe the distribution of that feature with a violin plot:



This violin plot shows the distribution of house sale prices by the year the house was sold. The bodies of the violin show the distribution of the data, larger bodies indicate a higher probability of the population will take on that value. Additionally, the length of the body increases as the years progress. This is an indication that the maximum sale price increases as the years progress. The year 2022 has the longest tip, indicating that this year saw more houses sold for higher amounts, although these were not the norm, as evidenced by the shape of the body.

Eight total models were created in this project: Baseline (Dummy Regressor), Linear Regression, Linear Regression with PCA, Linear Regression with ANOVA, Decision Tree, Tuned Decision Tree, Random Forest, and Tuned Random Forest. Of these eight models, the Tuned Random Forest achieved the best evaluation metrics. I chose to use Mean Absolute Error to quantify model performance because it can be interpreted as the average absolute dollar amount by which the predicted model deviated from the actual sales price. Therefore, it is easy to understand in the context of the problem being addressed.

In the graph below, we can see the performance of all the different models created. It is important to keep in mind that lower MAE values indicate better performance.



Conclusion

In conclusion, the Tuned Random Forest achieved the lowest Mean Absolute Error (MAE), with a MAE value of $22,667.12. This means that the model's average prediction deviated $22,667.12 from the actual sale price. I am happy with this performance and did not intend for this project to be used as a sole determining factor to be used in the sale or purchase of a house. Rather, this is a tool to better understand what the projected value of a house is. Using this information, it is easier to decide what a house might be worth on the market or if a house that is currently on the market is over or under priced.

I do not consider this model to be ready to deploy. There are several aspects of this model that I would like to investigate further, including:

- Increasing the dataset size – I believe that the model would benefit from an increased amount of data.
- Understanding the relationship between sale year and sale price – This was the second most important feature to the final model, and I only had five years' worth of data to compare.
- Understanding the impact of Geography on sale price – Breaking the houses into neighborhoods could reveal sale price influences that are not immediately apparent given the current data. Using street names or other geographic information might be a useful tool for understanding how geography affects sale prices.

The above recommendations could improve the model's robustness and potentially improve model accuracy. Using information from nearby counties might be more beneficial than seeking previous years of sales information.

The biggest challenge for this model is the limited availability of data. Kimball is not a large town and does not have a lot of house sales in any given year. Since the markets tend to fluctuate year by year, using previous years' sales information may not be as beneficial as a larger number of contemporary sales would be. This is why I would choose to look at nearby counties for additional data before I began the process of looking for data older than five years for Kimball.

In summary, the Tuned Random Forest model has demonstrated promising results, achieving the lowest Mean Absolute Error (MAE) of $22,667.12. While this performance aligns with the initial project goal of creating a tool for better understanding the housing market, it has limitations that need to be acknowledged.

The project’s goal was to create a tool to support an understanding of the current housing market and provide better insight into any purchasing or selling decisions. This is not designed to be used as a sole determinant in any financial decisions. The predictive capability of this model could be improved by increasing the size of the dataset. Further areas for model refinement include exploring the relationship between sale year and price and the geographical location of houses.

Despite the encouraging performance of the model, caution is warned against deploying this model in its current state. Further investigation into the model’s intricacies and external factors, such as geography, is warranted. Working with markets like Kimball that have small datasets is always a challenge. Using data obtained from counties near or similar in size to Kimball may enrich the model’s dataset and improve the model’s robustness.