House Price Prediction and Data Visualization

IEORE 4523 Data Analytics

Dragon's Crest Analysts

Minli Song

Botong Yuan

Qinyi Zhao

Xianrui Huang

Yunpeng Liu

Introduction

Despite the Federal Reserve's post-Covid tightening of monetary policy, housing prices persist in their upward trajectory, as evidenced by sustained public interest reflected in Google trend data. This trend underscores the increasing demand for reliable forecasting tools among economists and individuals alike. Accurate housing price forecasts are indispensable for informed financial planning, enabling individuals to make well-informed decisions regarding property transactions, be it buying, selling, or renting. Furthermore, such forecasts facilitate strategic investment decisions, allowing individual investors to optimize returns by capitalizing on emerging market trends while mitigating risks associated with market volatility.

In essence, the pursuit of dependable forecasting mechanisms is driven by the need to navigate the complexities of housing markets effectively. By providing insights into future price movements, accurate forecasts empower individuals to align their financial strategies with prevailing market dynamics, fostering stability and facilitating optimal resource allocation. Thus, a nuanced understanding of housing price dynamics and robust forecasting methodologies are essential for informed decision-making, ultimately contributing to sustainable economic growth and prosperity.

Dataset Description

The dataset, sourced from Kaggle, comprises approximately 1500 selling records, each containing 81 features. These features encompass 37 numerical columns, including 15 ordinal columns, and 44 categorical columns. Key features include LotFrontage, representing the linear feet of street connected to the property, LotArea indicating the size of the lot in square feet, Utilities denoting the type of utilities available, and YearBuilt, providing the original construction date of the properties.

Data preprocessing

Feature Understanding

Before manipulating our dataset, we conducted a detailed analysis to understand its features and identify preprocessing requirements such as normalization, outlier detection, and more complex transformations. Our initial step involved creating histograms for each numerical attribute to assess their distribution—whether normal, skewed, or bimodal. While no clear outliers were detected, we noted significant skewness in several features that we planned to address later.

We also examined the distribution of our target variable, house prices, categorizing and visualizing them by range. This step was crucial to confirm the suitability of our chosen modeling techniques, revealing a notable skewness in the target variable. To correct this and fit regression model assumptions, we applied logarithmic transformations to the price data.

Further, we explored relationships between house prices and other numerical features through scatter plots, which helped uncover trends and correlations essential for our feature selection process. This analysis guided us in identifying the most relevant variables for effective modeling, ensuring a thorough preparation of the dataset for subsequent predictive tasks.

Data manipulation

After gaining a thorough understanding of the data features and their interrelationships, we undertook multiple data cleansing and normalization procedures to enhance the quality and suitability of the data for modeling. These steps are essential in reducing potential biases and inaccuracies that could adversely affect the predictive modeling process.

Outliers can profoundly impact the performance of many algorithms, potentially leading to erroneous interpretations and predictions. Initially, our histogram inspections did not reveal any clear outliers, and to ensure accuracy, we further employed the Interquartile Range (IQR) method for verification. This technique calculates the first (Q1) and third (Q3) quartiles, defining an "outlier" as any data point that lies more than 1.5 times the IQR below Q1 or above Q3. By applying this method, we systematically checked for outliers. However, consistent with our initial observation, the IQR method confirmed that there were no outliers present in our data.

Missing data is a critical challenge in predictive modeling as it can introduce bias or diminish the statistical power of the model. In our dataset, the "LotFrontage" feature, which is crucial for predicting our target, has approximately 18% missing data. Fortunately, we have related variables such as "LotArea," "LotShape," and "LotConfig" that can aid in its prediction. To effectively address this issue, we employed Multiple Imputation by Chained Equations (MICE). This robust method treats each feature with missing values as a function of other features, using a round-robin approach to model and impute missing data. This technique ensures a comprehensive and statistically reliable way to handle missing entries, enhancing the integrity of our modeling process.

Skewness in feature distributions can significantly impair the performance of many machine learning algorithms, which typically assume that input data are normally distributed. Our initial histogram analysis revealed several skewed features within the dataset, prompting us to set a skewness threshold of 0.5 for corrective action. To address this skewness in numerical features, we applied a logarithmic transformation using the np.log1p function. This method is especially effective because it gracefully handles zero and negative values by adding one before computing the logarithm, thereby stabilizing variance and normalizing distributions. This transformation ensures that our data conforms more closely to the assumptions underlying many advanced analytical models.

Machine Learning Model Evaluation

We conducted a comprehensive evaluation of various machine learning models for predicting house prices using regression techniques. The models assessed include Linear Regression, Random Forest, TensorFlow Decision Forests, CatBoost, LightGBM, and XGBoost. The key performance metrics were Root Mean Square Error (RMSE) and R-squared score.

Model Performance

Linear Regression is a statistical method that models the relationship between a dependent variable and one or more independent variables. It serves as a baseline for comparison with more complex algorithms. The model yielded an RMSE of 0.19968 and an R-squared of 0.73664.

Random Forest is a machine learning method capable of performing regression and classification tasks. It uses ensemble learning, combining the predictions of multiple decision trees. The model achieved an RMSE of 0.13903 and an R-squared of 0.87232.

TensorFlow Decision Forests is a collection of high-performance decision forest learning algorithms integrated into TensorFlow, offering flexibility and efficiency in building models. This model delivered an RMSE of 0.13266 and an R-squared of 0.88375.

CatBoost is an open-source gradient boosting library developed by Yandex, designed for both regression and classification tasks, with strong performance and built-in handling for categorical variables. The model recorded an RMSE of 0.13319 and an R-squared of 0.88283.

LightGBM and XGBoost are efficient and speedy gradient boosting algorithms, they generated RMSE: 0.13347, 0.14316 and R-squared: 0.88232, 0.86462 respectively.

We focused on tree-based models for several reasons:

1. Handling Non-linear Relationships: These models partition the feature space into smaller regions and fit separate models in each, effectively handling non-linear relationships.
2. Categorical Features: Tree-based models manage categorical features well, especially when appropriately transformed or encoded.
3. Feature Importance: These models provide insight into feature importance, helping in understanding which features are most influential.

Among all the models evaluated, the TensorFlow Decision Forest model achieved the lowest RMSE and the highest R-squared score, making it the best-performing model for predicting house prices in the dataset.

Feature Selection

For the best performing model TensorFlow Decision Forest, perform feature selection to improve the model accuracy. First determine the importance of each feature by the *SUM_SCORE* attribute, which essentially evaluates how frequently a feature is used in the tree split. A higher frequency means the feature contributes more to the tree structure and is therefore more important. With this importance score, perform the following logic:

For $i = 1, 2, ..., n$ (where n is the number of all features), extract the most important i features to train the TFDF; record the model accuracy.

With this approach, we can understand how the model performs as we include more and more features and we use the subset of features with the best accuracy score. Note that this approach does not necessarily guarantee the most optimal feature subset to use because it does not attempt to test every possible subset. It is likely that the combination of some top important features with some other relatively less important features will also do well (for example, the combination of top 20 features and 30-40 features). However, exhausting all possible subsets of features requires $O(2^n)$ models to train, while our approach only requires training $O(n)$ models. Because of computation considerations, we try to achieve a suboptimal solution and conclude that including around the top 30% features would yield the best performance.

Conclusion

In this study, we meticulously evaluated various machine learning models to predict house prices, placing a strong emphasis on robust data preprocessing. Our preprocessing involved addressing missing values through Multiple Imputation by Chained Equations (MICE), managing outliers with the Interquartile Range (IQR) method, and normalizing skewed distributions using logarithmic transformations. This careful preparation of the data was pivotal in optimizing our models for accurate predictions.

The TensorFlow Decision Forest emerged as the most effective model, demonstrating superior performance in terms of both RMSE and R-squared scores. We further enhanced this model's accuracy through a strategic feature selection process, identifying and utilizing the top 30% of the most influential features based on their contribution to model splits.

These results underscore the crucial role of data preprocessing and feature selection in enhancing the predictive power of machine learning models. They also lay a foundation for future research, suggesting that further advancements could be achieved by exploring new preprocessing techniques and experimenting with more sophisticated models. This approach not only bolsters the accuracy of predictions in real estate but also extends to broader applications within the field of predictive analytics.

Appendix

Figure 1: Sale Price Distribution

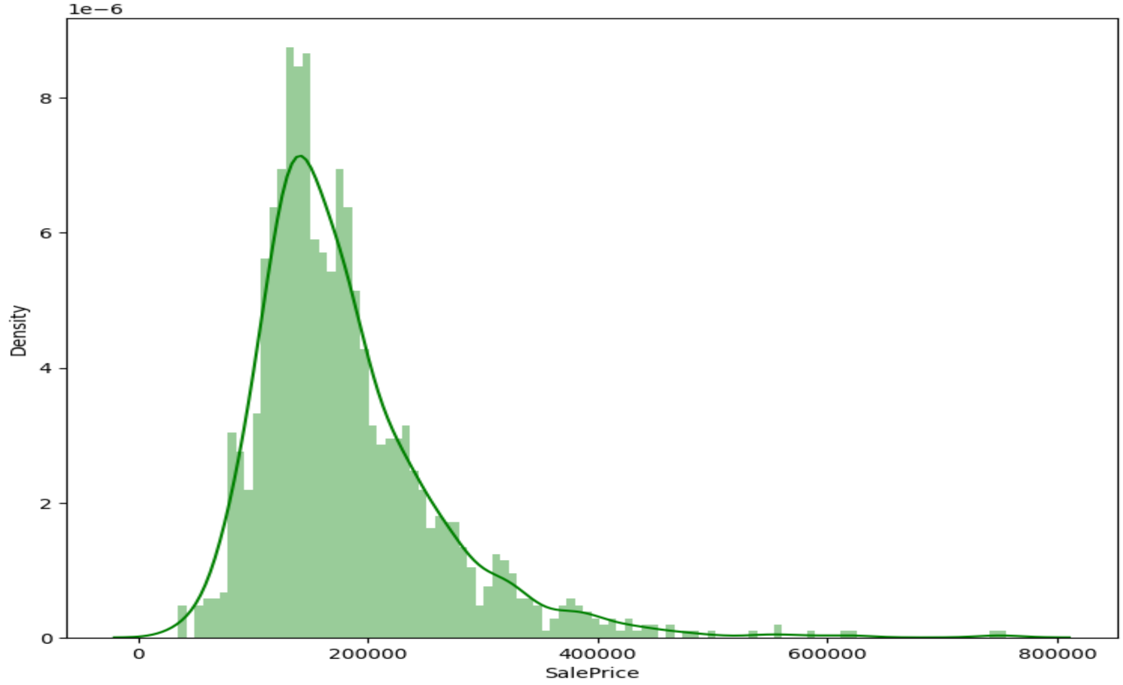


Figure 2: Sale Price Range Distribution

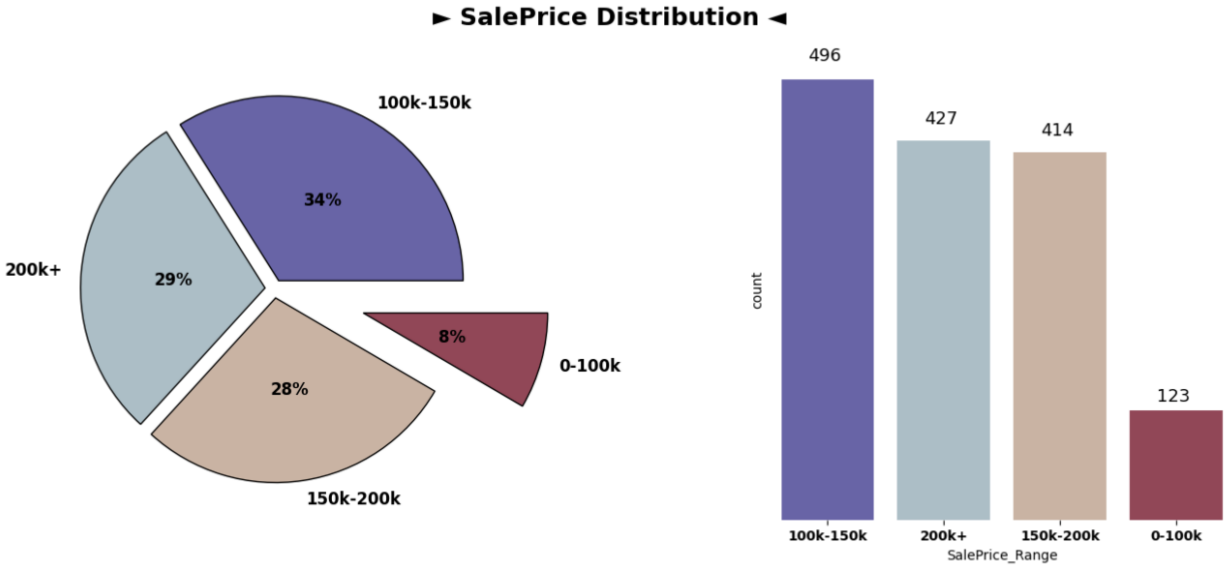


Figure 3: Numerical Feature Distribution

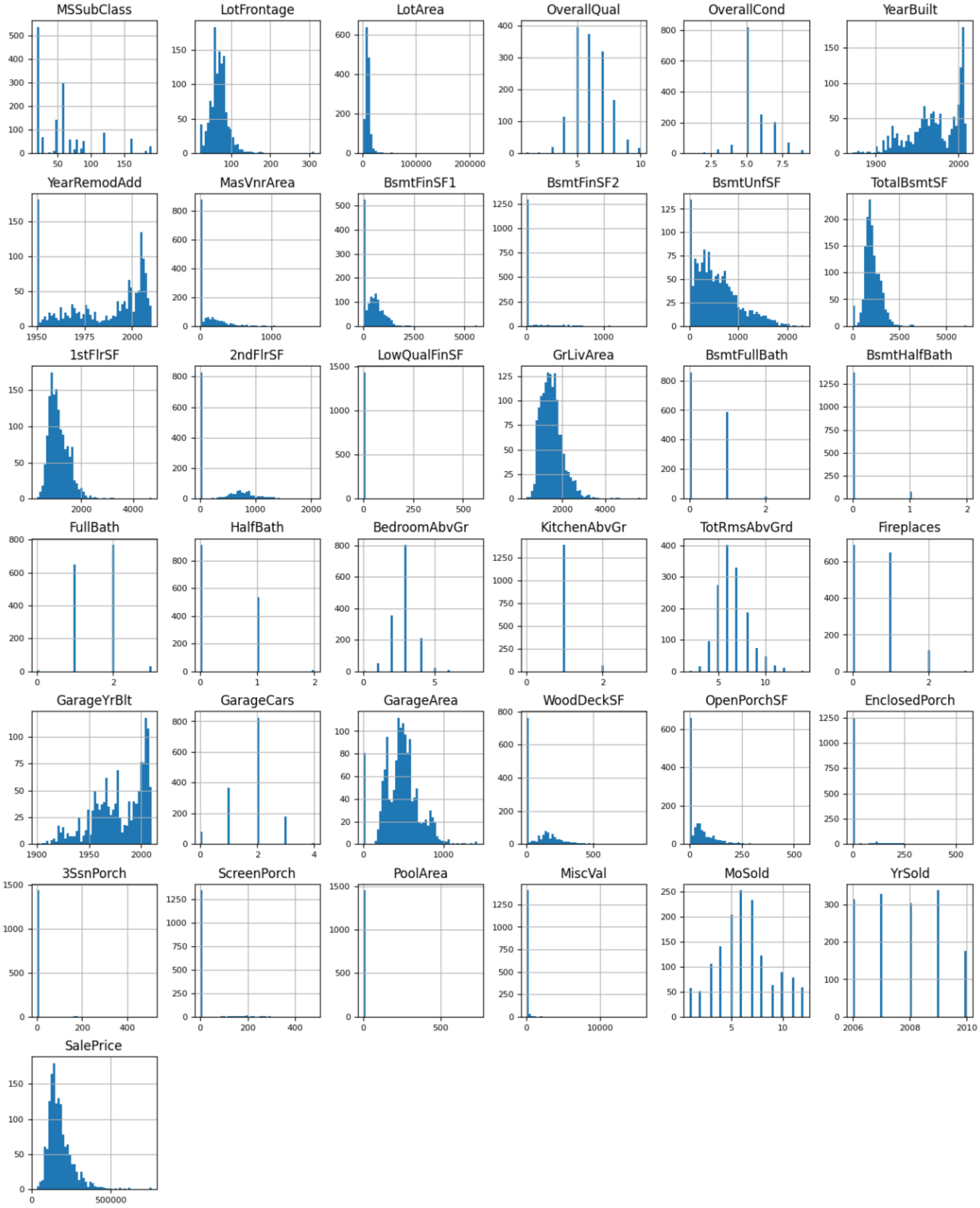


Figure 4: Target Variable Transformation plot

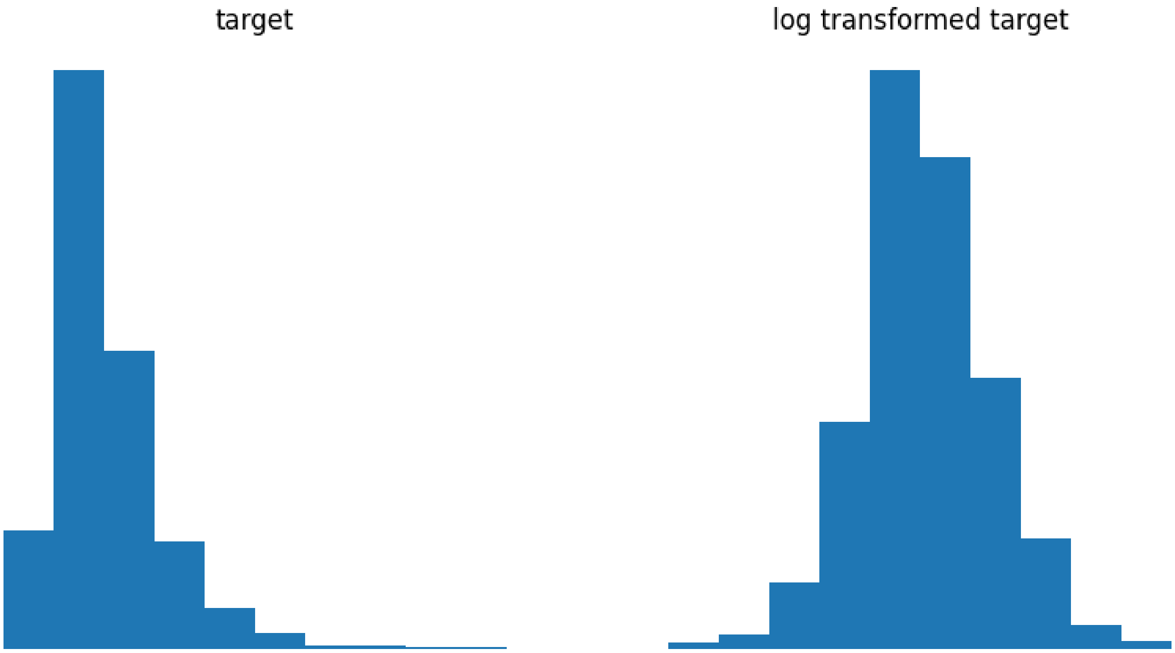


Figure 5: Linear Regression Actual vs Predicted Prices

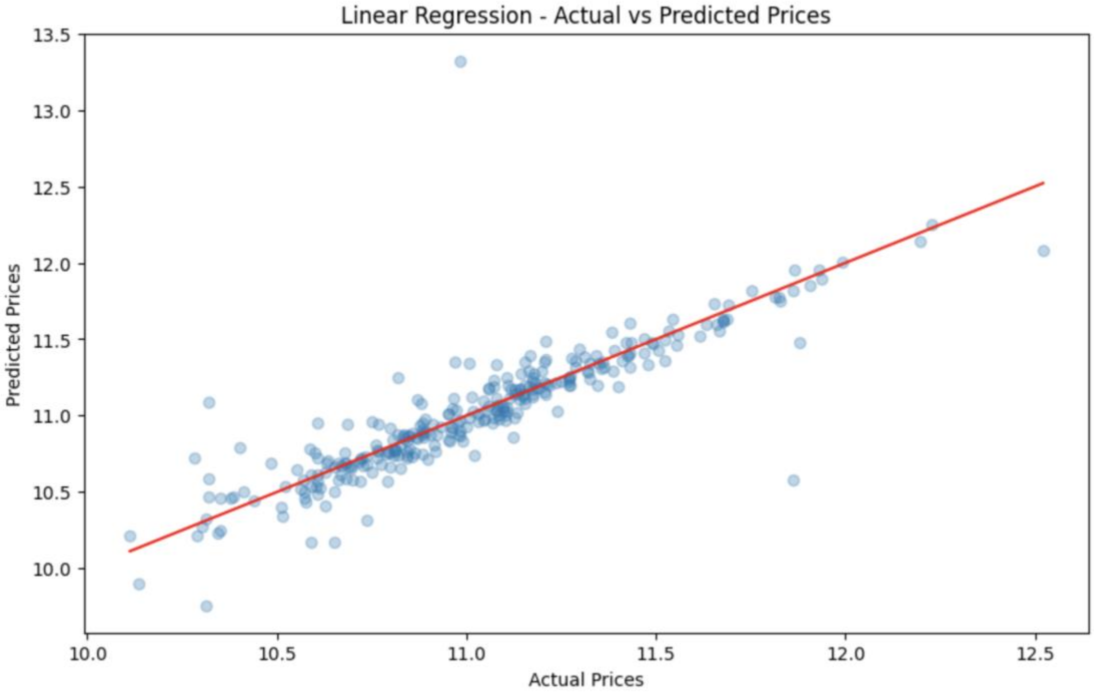


Figure 6: Random Forest Actual vs Predicted Prices

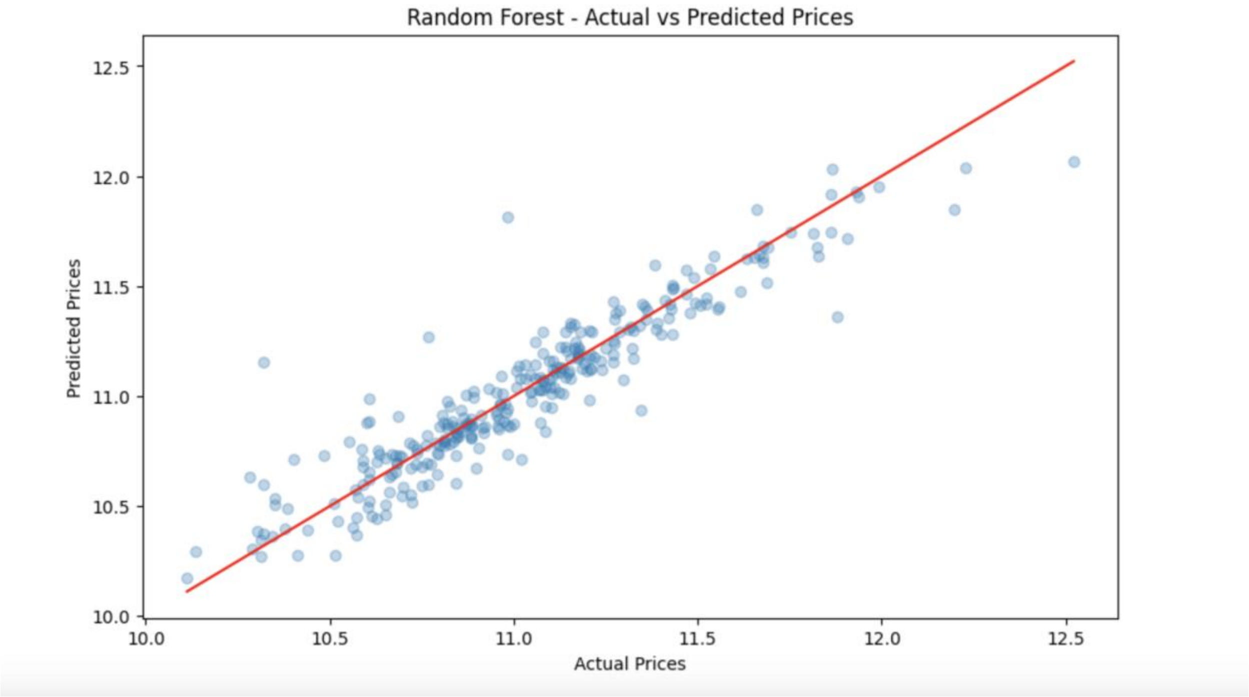


Figure 7: CatBoost Actual vs Predicted Prices

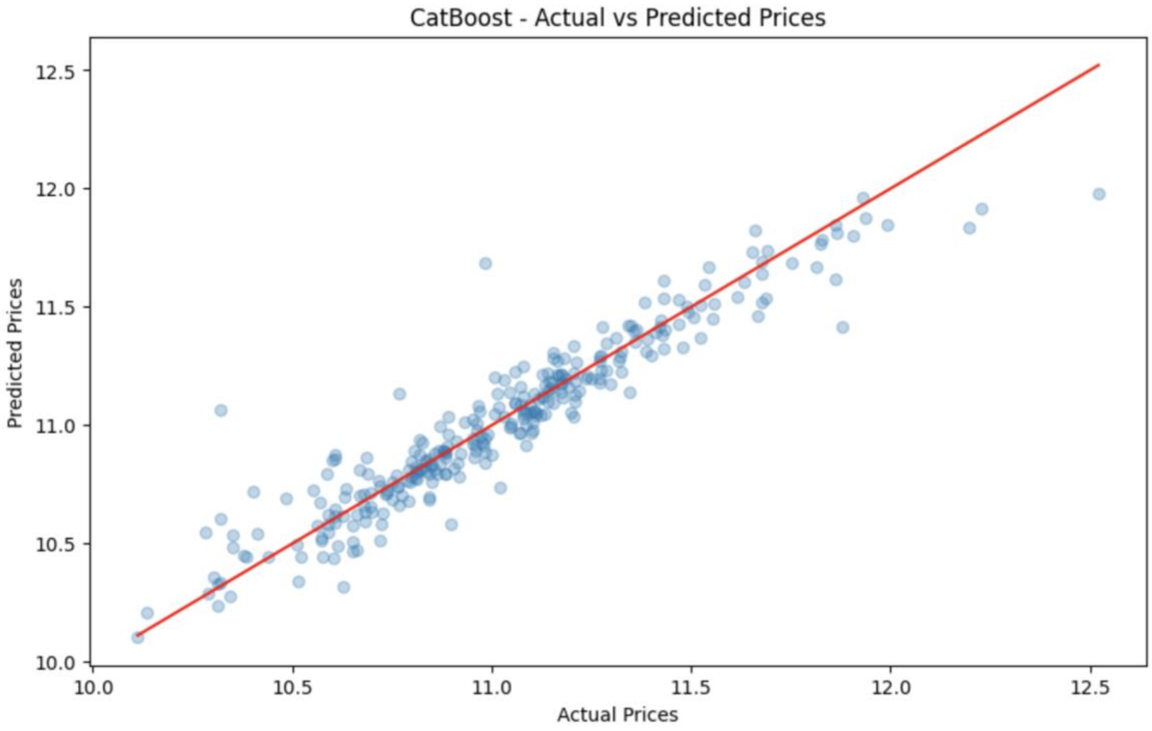


Figure 8: LightGBM Actual vs Predicted Prices

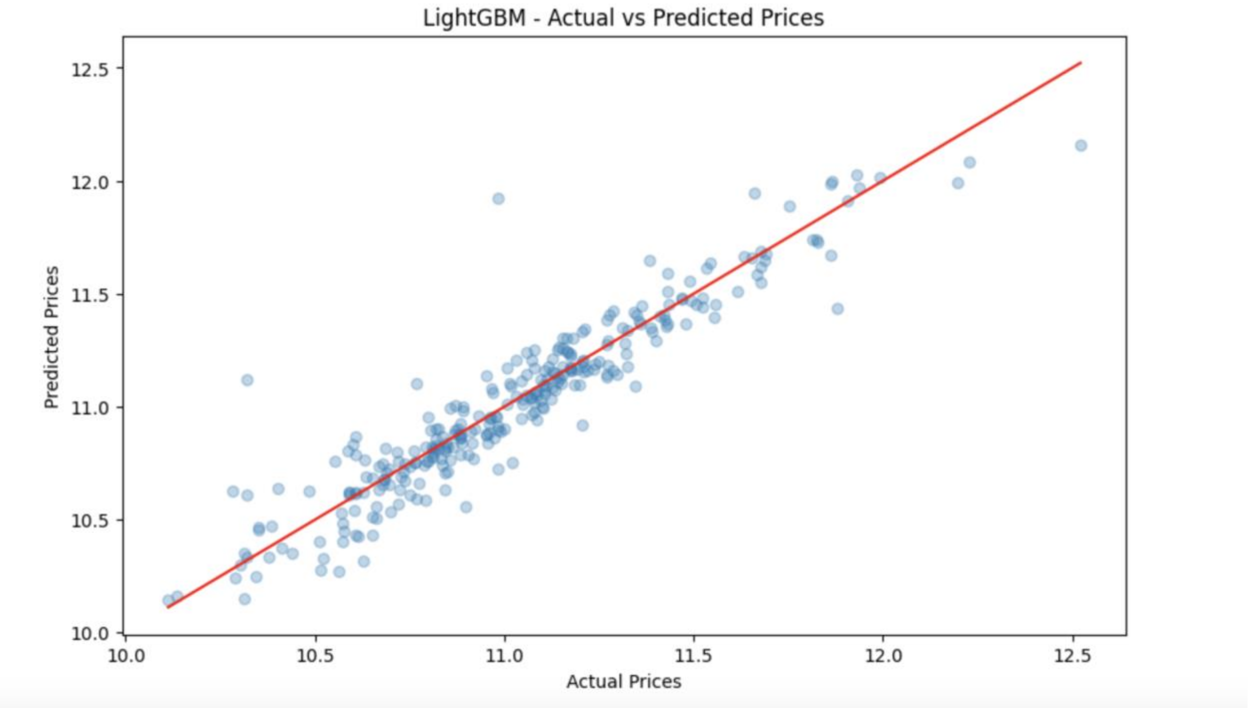


Figure 9: XGBoost Actual vs Predicted Prices

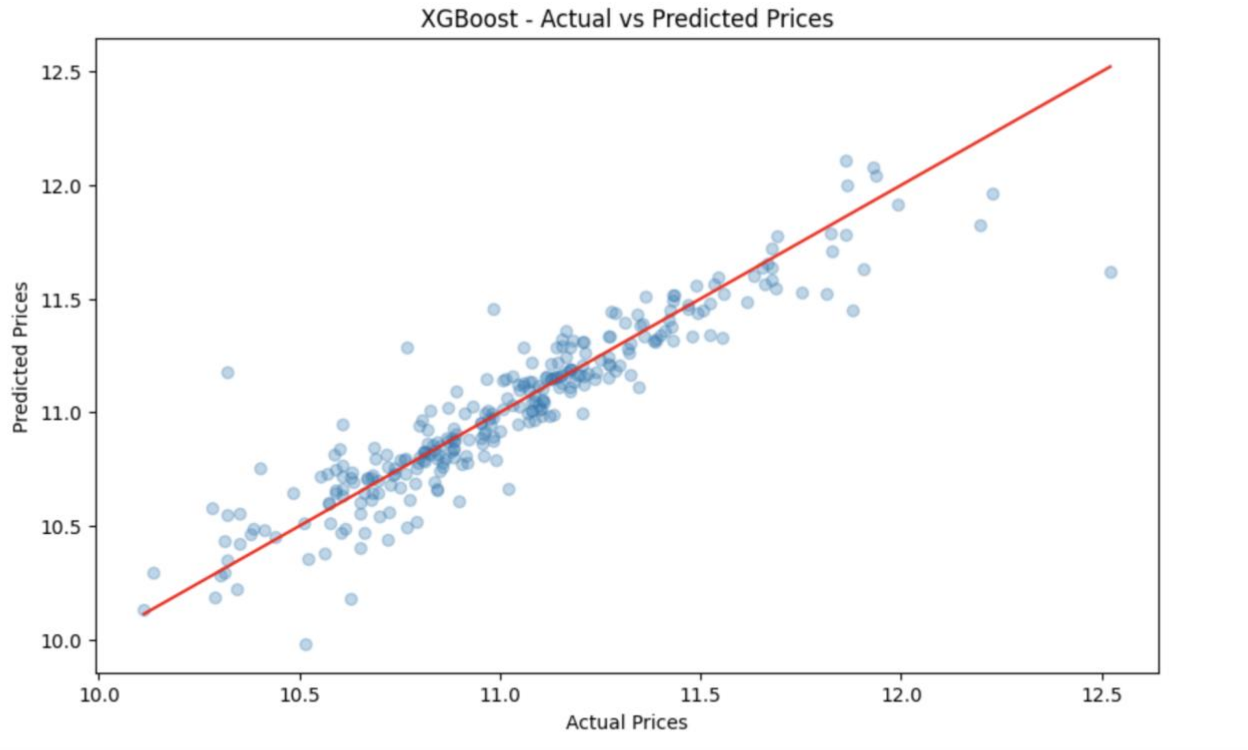


Figure 10: TFDF Actual vs Predicted Prices

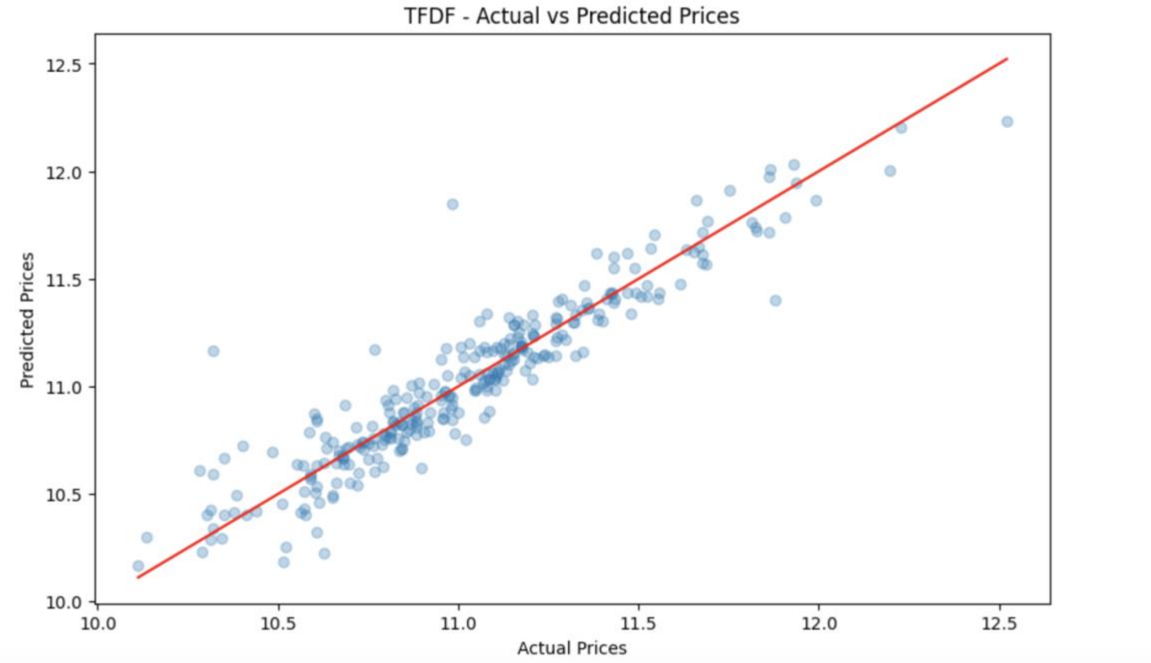


Figure 11: TFDF Feature Importance

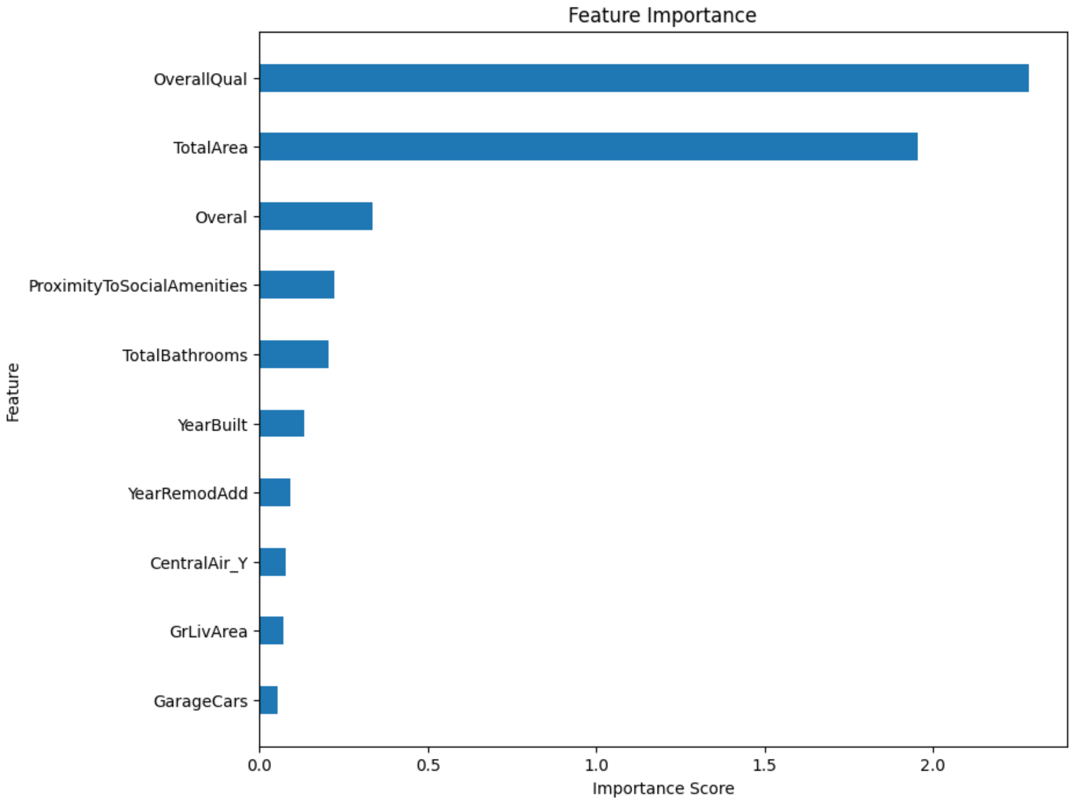


Figure 12: Feature Heatmap

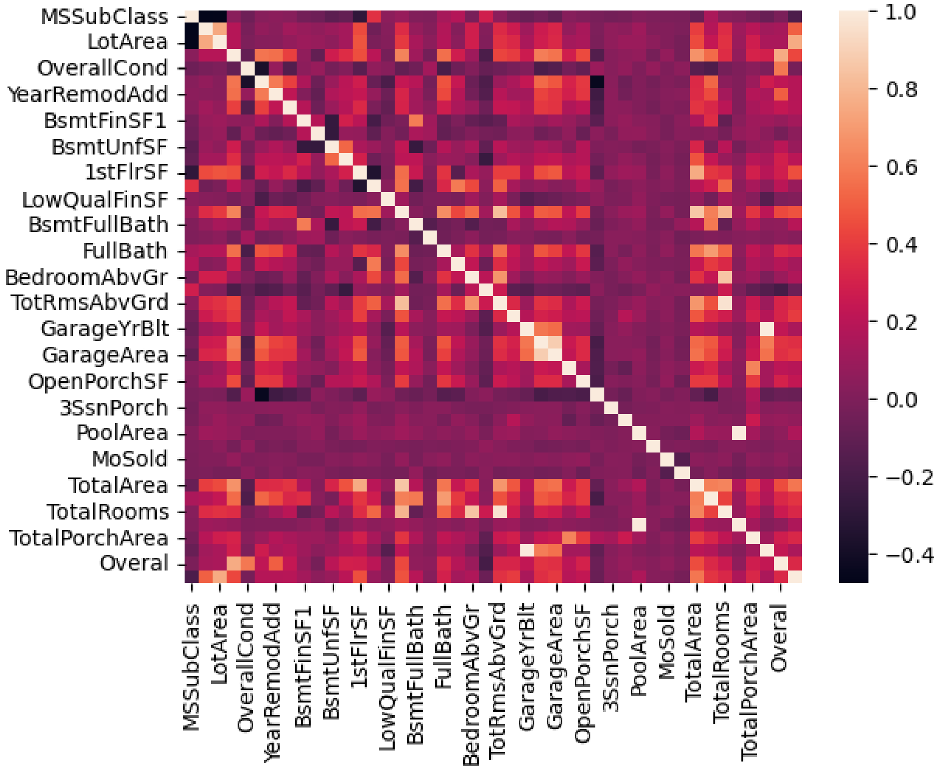
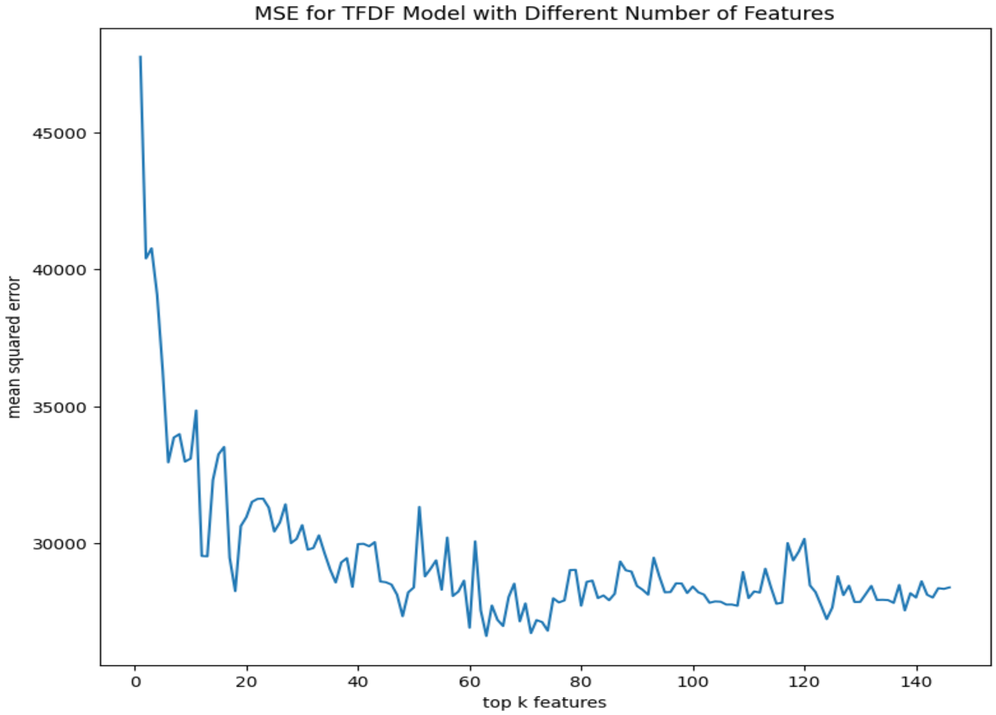


Figure 13: MSE with Different Numbers of Features



References

Module: tfdf | TensorFlow Decision Forests. (n.d.). TensorFlow. https://www.tensorflow.org/decision_forests/api_docs/python/tfdf

House Prices - Advanced Regression Techniques | Kaggle. (n.d.). https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

Soni, B. (2023, February 27). Topic:9 MICE or Multivariate Imputation with Chain-Equation. *Medium*. https://medium.com/@brijesh_soni/topic-9-mice-or-multivariate-imputation-with-chain-equation-f8fd435ca91#:~:text=MICE%20stands%20for%20Multivariate%20Imputation,produce%20a%20final%20imputed%20dataset.

3.2 - Identifying outliers: IQR Method | STAT 200. (n.d.). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat200/lesson/3/3.2