COMP SCI 7209 Big Data Analysis and Project
Project Title: Big Data: House Price Analysis and Prediction (Part D)
Ning Ni (a1869549)
Project Coordinator: Bernard Evans
10 August 2023

1. Restatement and Summary

1.1 Restatement of Questions

In this project, there are four questions that need to be researched.

Table 1 Summary of Questions

Order	Questions	Question Type
RQ1	Describe in detail the impact of the type, size, quality and zoning of the house on the price of the house.	Visualization
RQ2	Evaluate the importance of built year on the predicted house price model.	Regression
RQ3	Utilize models to quantify the impact on prices by adding additional bathrooms, while keeping other features in the same condition.	Regreesion
RQ4	Can these key features effectively classify different price levels for houses?	Classification

1.2 Summary of Models

Based on the analysis of Part C, Adaboost Model (Schapire, 2013) was employed to analyze regression problems and Random Forest Model was used to predict the classification problem.

Table 2 Summary of Models

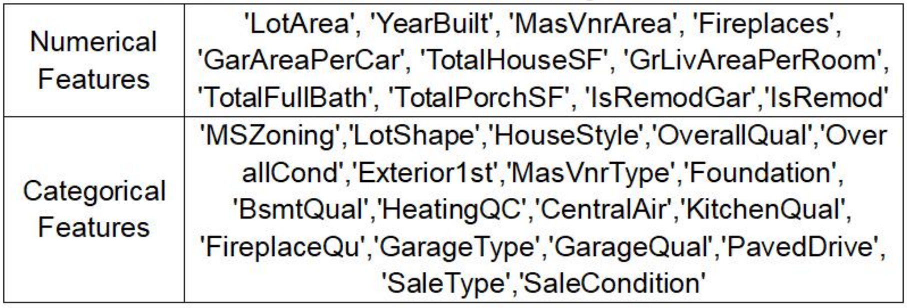
Order	Model	Apply	Predict Variable	Hyperparameter
1	Adaboost Regression Model	RQ2, RQ3	SalePrice	'n_estimators' :530, 'learning_rate' :1.2
2	Random Forest Model	RQ4	PriceLevel	'max_depth': 14, 'min_samples_leaf': 1, 'min_samples_split': 3, 'n_estimators': 201

2. Analysis and Visualization

2.1 Analysis of RQ1

In Part B and Part C, The impact of each house feature on price is explored in detail on the section of EDA and Feature Engineering. In the processing, the project extracted 29 house features from 80 valid variables, as the table 3 shown.

Table 3 The Main Features Affecting House Price

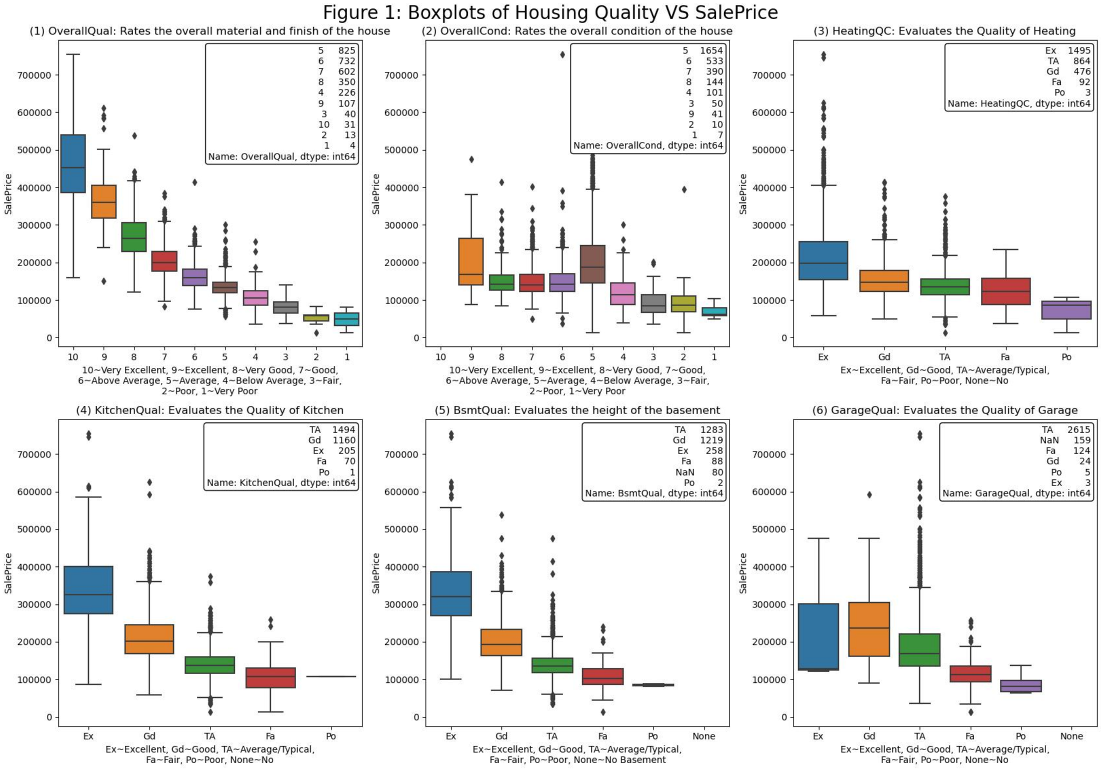


Numerical Features	'LotArea', 'YearBuilt', 'MasVnrArea', 'Fireplaces', 'GarAreaPerCar', 'TotalHouseSF', 'GrLivAreaPerRoom', 'TotalFullBath', 'TotalPorchSF', 'IsRemodGar','IsRemod'
Categorical Features	'MSZoning','LotShape','HouseStyle','OverallQual','OverallCond','Exterior1st','MasVnrType','Foundation', 'BsmtQual','HeatingQC','CentralAir','KitchenQual', 'FireplaceQu','GarageType','GarageQual','PavedDrive', 'SaleType','SaleCondition'

2.1.1 Housing Quality

Figure 1 illustrates a general trend wherein the sale price increases with improvements in overall housing quality and condition. This pattern is also observed for specific functional zones, indicating a consistent correlation between quality enhancement and higher price.

Figure 1: Boxplots of Housing Quality VS SalePrice

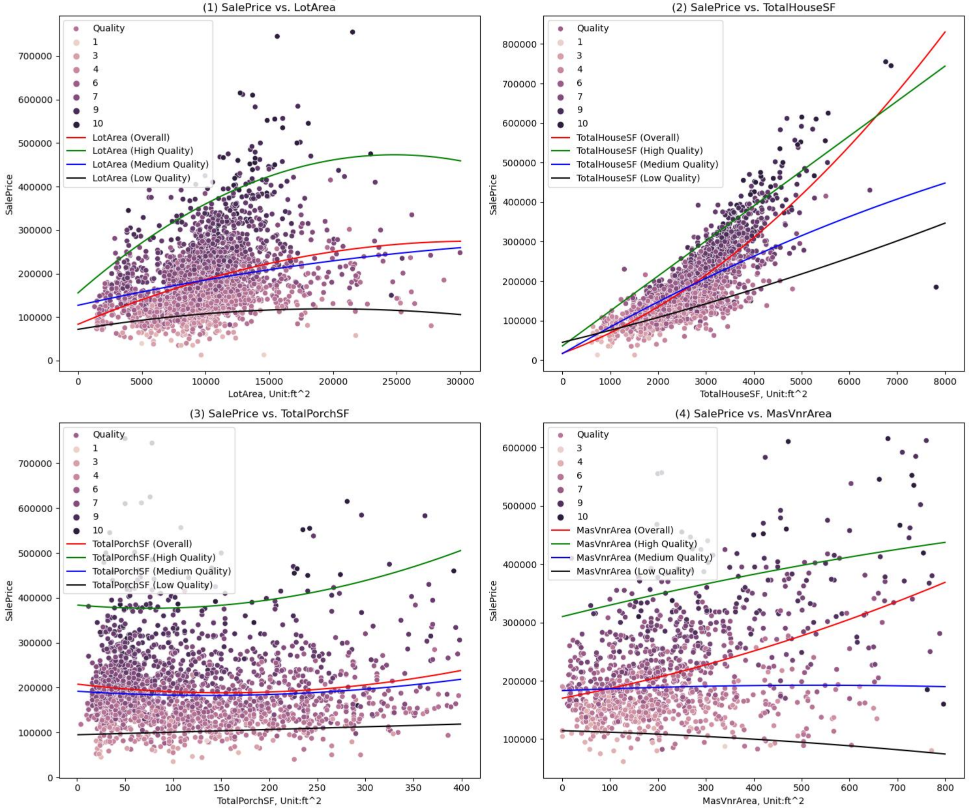


2.1.2 Housing Size

Obviously, there is a clear trend of housing prices increasing with larger lot sizes and house dimensions. Moreover, in cases where both lot size and house size remain constant, properties with higher quality levels tend to command higher prices.

When it comes to porch area and masonry veneer area, the influence on housing prices becomes more pronounced in the context of high-quality houses. However, for houses of varying quality levels, the connection between these areas and pricing is comparatively less discernible.

Figure 2 SalePrice vs. Variables of Housin Size



(1) SalePrice vs. LotArea

(2) SalePrice vs. TotalHouseSF

(3) SalePrice vs. TotalPorchSF

(4) SalePrice vs. MasVnrArea

2.1.3 Housing Type and Zoning

Properties classified as "Floating Village Residential" exhibit the highest average prices. On the other hand, properties situated in areas zoned for Agriculture, Commercial, and Industrial purposes tend to command lower prices compared to those within Residential Zones. The majority of house types gravitate toward the Low and Medium Density Residential Zones. Moreover, most of house type focus on one or two story. The more the floor, the more expensive the house is.

Interestingly, an intriguing observation emerges from these figures: properties with irregularly shaped lots tend to fetch higher prices than those with regular shapes. This phenomenon aligns seamlessly with real-world scenarios. "Floating Village Residential" areas, which are usually irregular, often boast meticulously landscaped surroundings, contributing to their premium prices. Conversely, properties situated in non-residential zones tend to face certain disadvantages such as noise, privacy concerns, and pollution, which collectively influence their lower prices relative to residential zones.

Figure 3: Barchart of Housing Type and Zoning vs. SalePrice

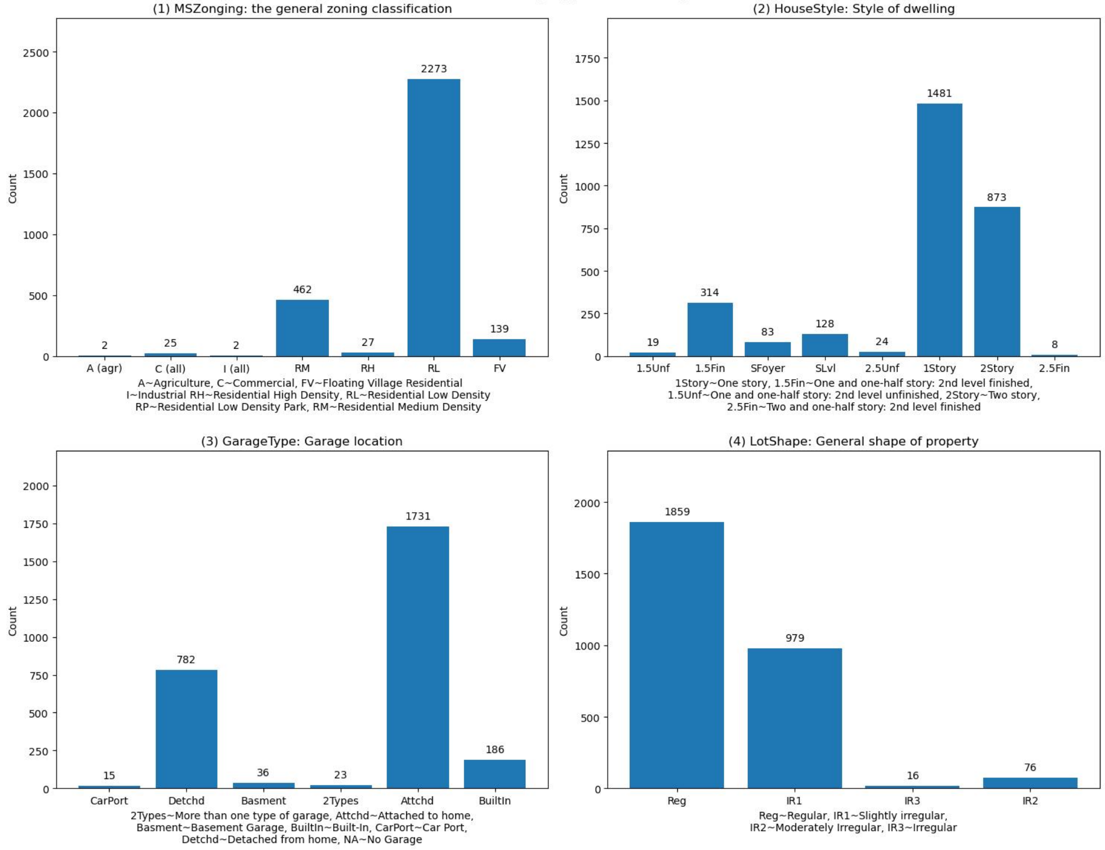
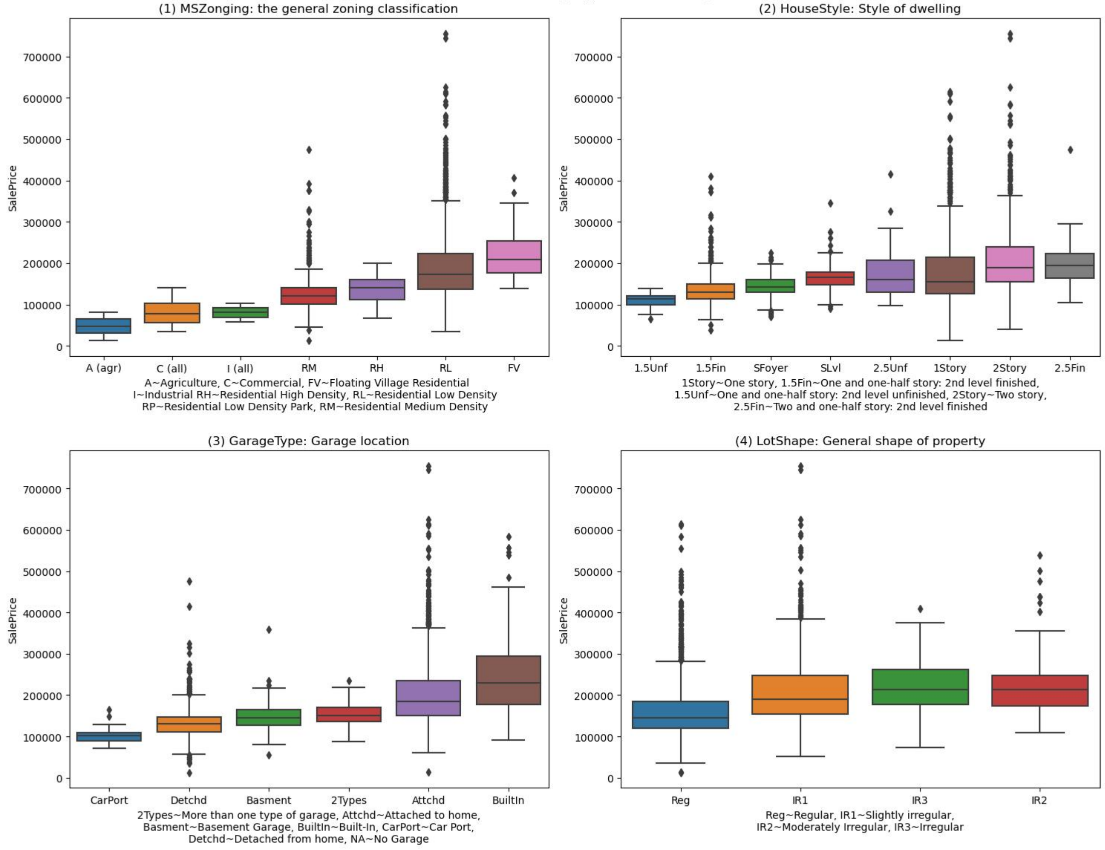


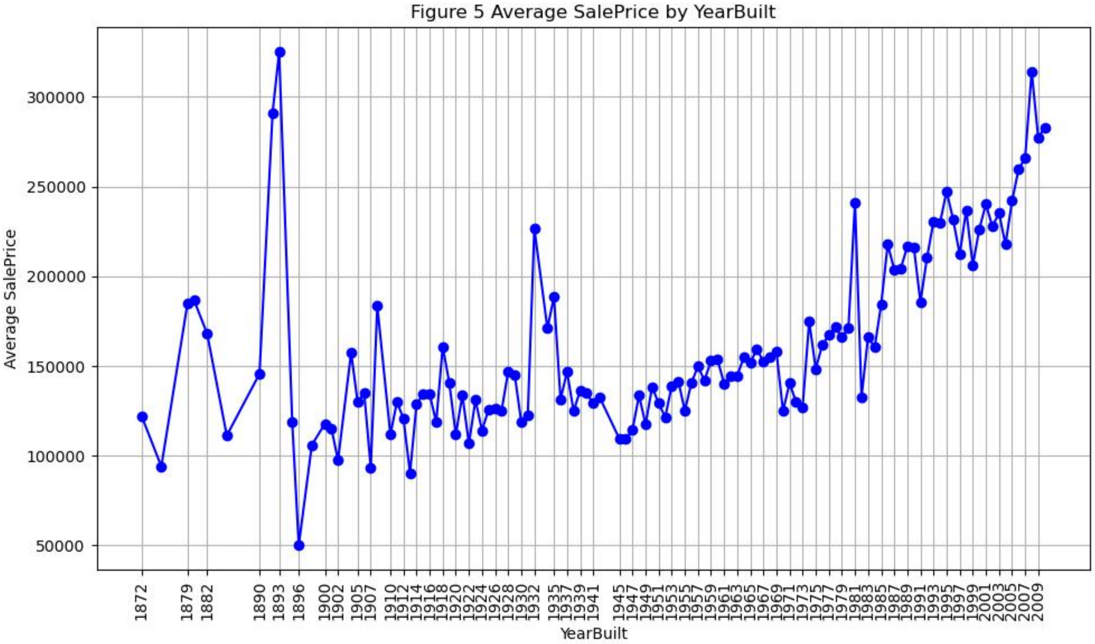
Figure 4: Boxplots of Housing Type and Zoning vs Price



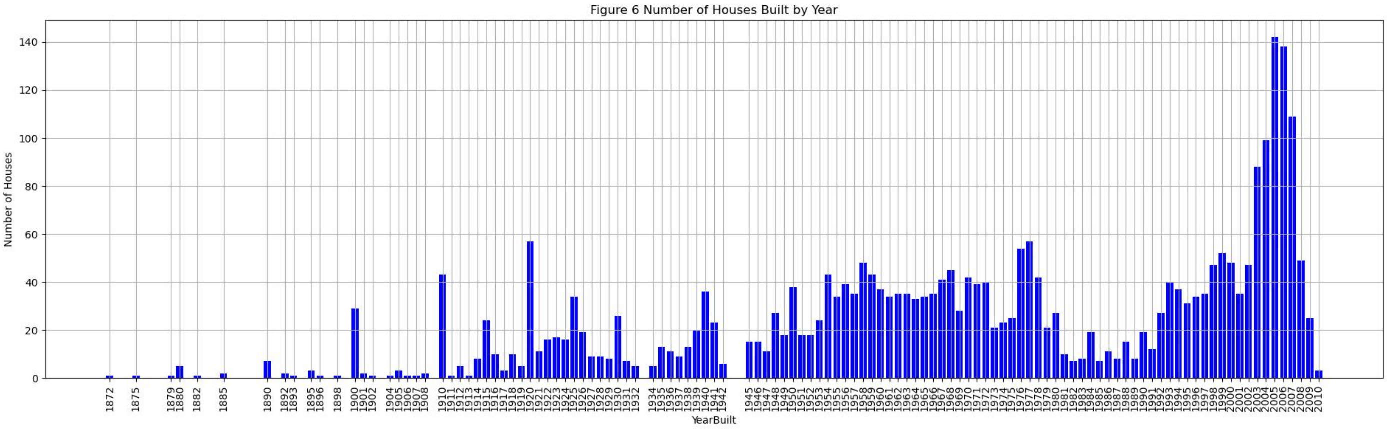
2.2 Analysis of RQ2

2.2.1 The Impact of YearBuilt on Prices

Based on the analysis of the dataset, the average house price demonstrates a period of stability until around 1955. If we exclude the substantial impact of major economic crises, specifically the periods of 1982-1983 and 1932-1935, associated with the events known as "The Panic of 1893" and the "Great Depression" respectively (Wikipedia: Panic of 1893, 2023; Great Depression, 2023). Subsequent to 1955, a noticeable trend emerges where house prices embark on a gradual upward trajectory.



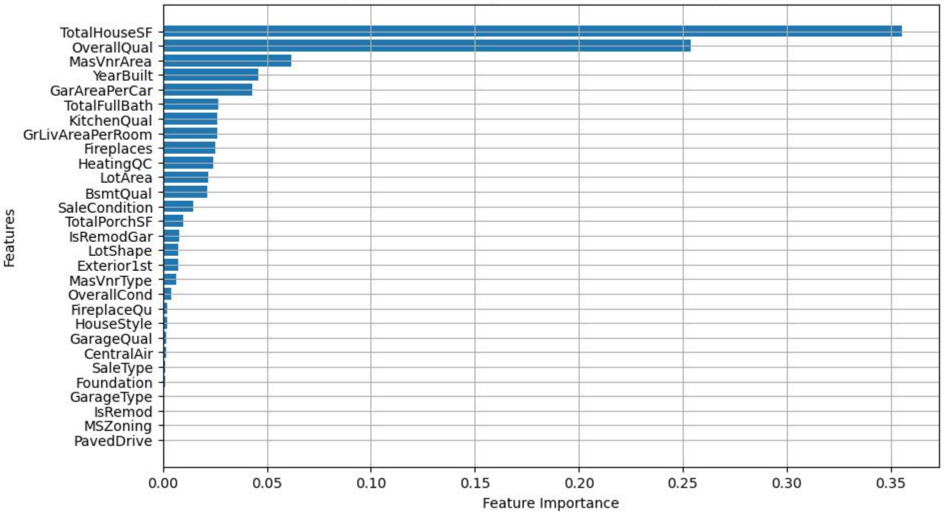
From an overall perspective, the number of newly constructed houses has exhibited an upward trend year by year. However, there is a notable decline in the number of houses built during the 1980s. This phenomenon can be attributed to the fact that there was minimal population growth in the local area during the 1980s (Wikipedia: Ames, Iowa, 2023), resulting in a lack of market demand for new housing.



2.2.2 SalePrice Prediction based on YearBuilt

In the context of Adaboost Regressor, the quantification of feature importance reveals that YearBuilt takes a prominent role in forecasting SalePrice.

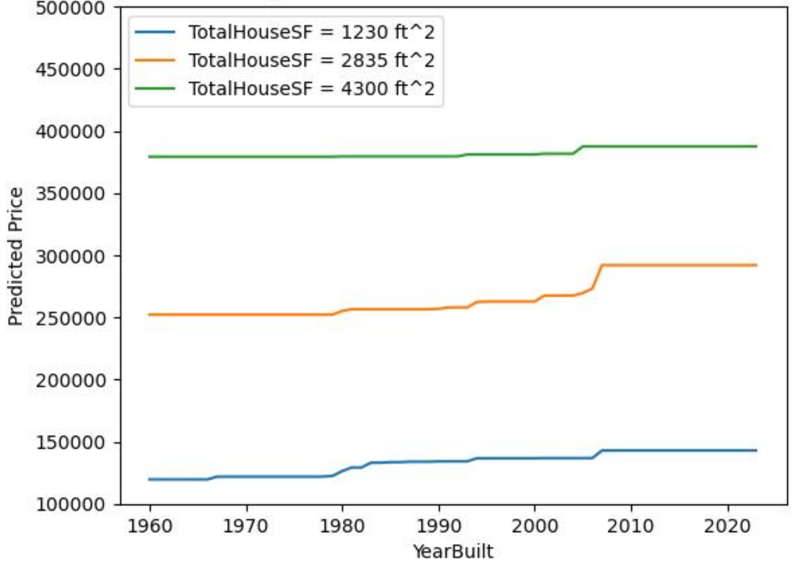
Figure 7 Feature Importance from AdaBoostRegressor



When considering three distinct house sizes while keeping other housing features constant for each case, the predictions from Adaboost depict a trend in which the prices of houses of all sizes gradually decrease as the age of the houses increases. This finding aligns with Leguizamon's perspective (2010, p.518), which suggests that the size of the house and the yard positively impact house prices, while the age of the house negatively affects them.

However, it is noteworthy that larger houses showcase a stronger tendency for maintaining their value over time, while houses with medium and small dimensions experience a more pronounced depreciation as their age increases.

Figure 8 Predicted Price vs YearBuilt



This may suggest a fact that larger houses tend to have more amenities and features that can contribute to their value retention. They might offer more living space, larger yards, and higher-end finishes, which are appealing to potential buyers and can help maintain their value over time.

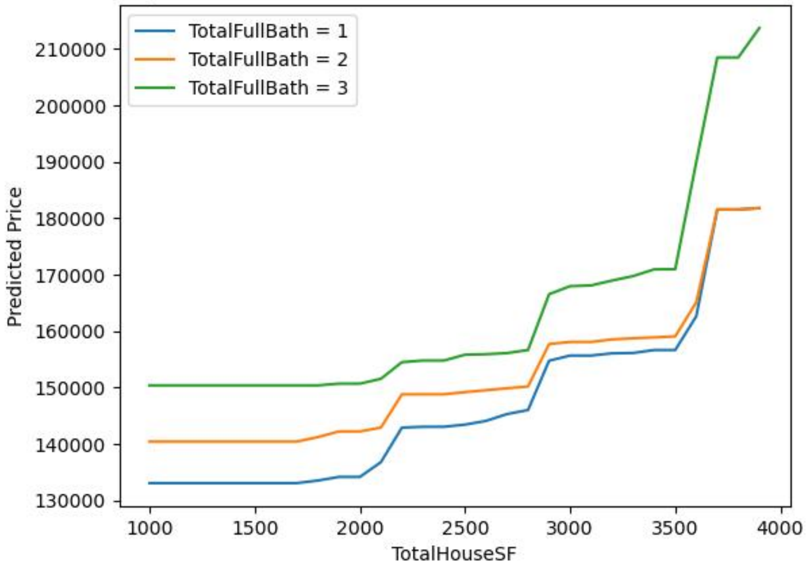
While smaller and medium-sized houses might be more affected by wear and tear over the years due to a potentially higher occupancy rate. Maintenance costs could be relatively higher for smaller houses, and as they age, potential buyers might factor in these costs when determining their value.

2.3 Analysis of RQ3

As the Figure 7 shown, TotalFullbath is also a significant factor in predicting SalePrice.

When considering houses of the same size, it's evident that those with a higher number of full bathrooms are worth a higher price. Notably, for houses larger than 3600 square feet, properties equipped with 3 or more full bathrooms exhibit greater value compared to those with only 1 or 2 full bathrooms. This observation suggests that houses with 1 or 2 bathrooms might not adequately cater to the requirements of larger households that may accommodate more family members.

Figure 9 Predicted Price vs TotalHouseSF and TotalFullBath

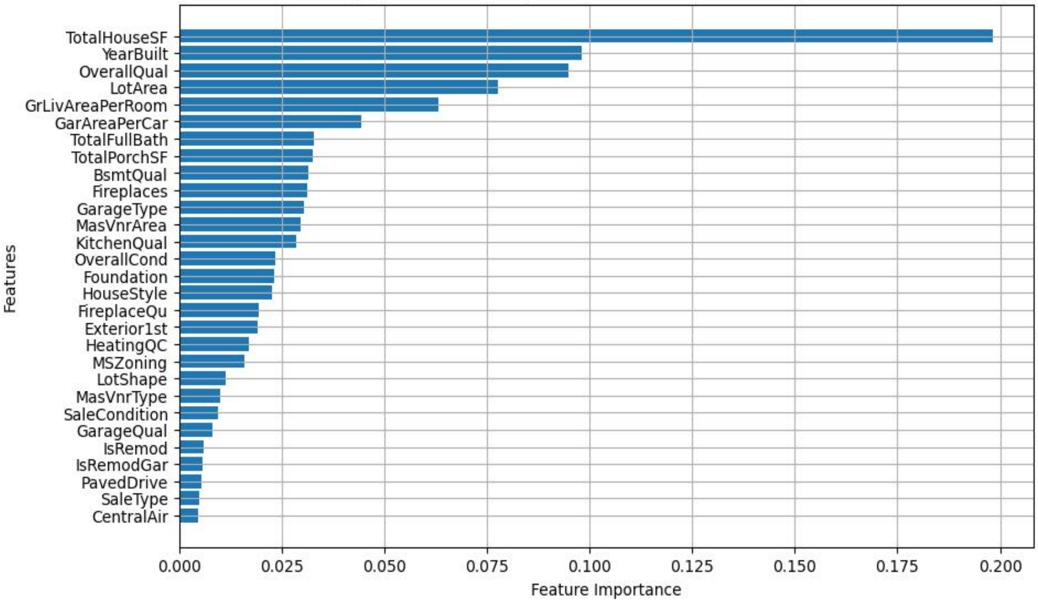


Therefore, when the house size is below 3600 square feet and remains constant, the addition of each extra bathroom corresponds to an average price rise of around $5,000 to $10,000 for the house. However, when the house size exceeds 3600 square feet, properties with 3 bathrooms command a price premium of nearly $30,000 compared to those with 1 or 2 bathrooms.

2.4 Analysis of RQ4

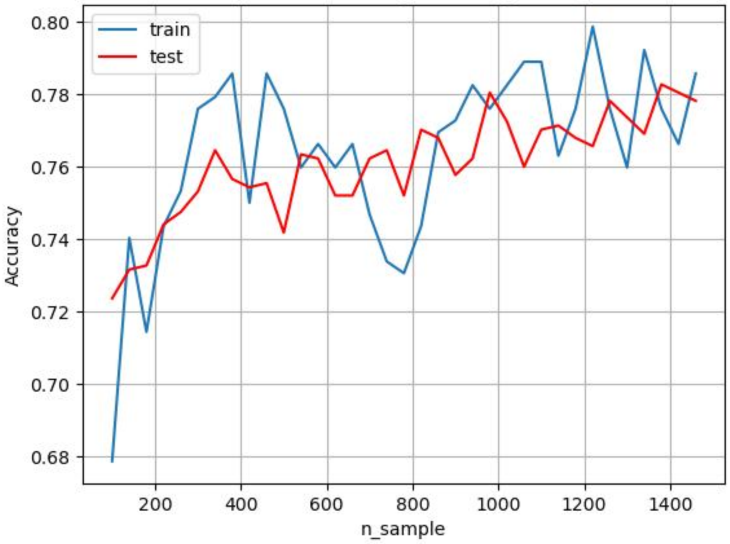
Similar to AdaboostRegressor, the RandomForestClassifier highlights TotalHouseSF, YearBuilt, and OverallQual as the top three influential features for categorizing the range of house prices.

Figure 10 Feature Importance from RandomForestClassifier



Domingos (2012) provides three points for evaluating a classification model, including Representation, Evaluation and Optimization. In Part C, the RandomForestClassifier exhibits commendable performance in both the training and testing datasets, achieving accuracy rates of 79.76% and 78.95%, respectively. While the accuracy might not reach an ideal value, like above 90%, it's important to note that perfection isn't always attainable in real-world scenarios. Furthermore, Figure 11 visually illustrates the excellent fit of the Random Forest Classifier, indicating its strong generalization capabilities.

Figure 11 Learning Curve by Fitting



The recall on the training set is 77.18%, and on the testing set, it is 76.43%. Recall refers to the proportion of positive samples that the model correctly predicts out of the actual positive samples. The relatively high recall on the testing set indicates the model's effective ability to identify samples from different house price ranges.

The AUC on the training set is 0.939, and on the testing set, it is 0.935. AUC represents the area under the ROC curve, which is a crucial metric for evaluating a model's classification performance. Higher AUC values suggest that the model performs well across different threshold values.

In summary, these performance metrics, including Accuracy, Recall, AUC, and the consideration of Bias-Variance trade-off, provide evidence that the RandomForestClassifier can proficiently classify the price range based on these crucial features.

3. Improvement of Solution

3.1 Regression Model

While the Adaboost model's RMSE of around $30,000 in Part C might seem reasonable, especially considering the substantial value of house prices that can extend into the hundreds of thousands of dollars, there's an opportunity to enhance predictive accuracy further.

Notably, certain key features exhibit robust linear correlations with the SalePrice variable. The exploration of alternative linear regression models that capitalize on the pronounced linear relationships within the data could potentially yield even more precise predictions, such as Polynomial Regression or Spline Regression.

Moreover, Feature Engineering shouldn't be limited to the pre-model establishment phase; rather, it should also be carried out during the model-building process itself. This dynamic approach can enhance the effectiveness of the specific model.

3.2 Classification Model

The RandomForestClassifier demonstrates an accuracy rate of approximately 80%, indicating that there is room for potential improvement through various methods.

In this project, the prices are categorized into four levels using quartiles to maintain balance among different classes. However, this approach may result in unreasonable in the original price classification ranges. To address this, it could be beneficial to employ appropriate binning techniques for assigning price ranges. Examples of such techniques include Equidistant Binning (Heyong, 2014), Clustering Binning (Jang, 2021), and Adaptive Binning (Meyer, 2018).

4. Conclusion and Future Work

4.1 Conclusion

This project has established a comprehensive data pipeline to delve into various aspects of house pricing. The process encompasses formulating pertinent questions, gathering relevant data, conducting EDA, performing feature engineering, constructing both regression and classification models for price or its range prediction.

All four questions were fully analyzed and answered. The results of the study show that:

- The type, size, quality and zoning of the house affects the price of the house in different ways.The size of the house is the most significant feature variable in the model, and various types of houses follow the pattern that larger house sizes correspond to higher prices. A similar trend is observed in Lot size. Furthermore, the overall quality of the house and the quality of various functional areas show a clear trend where better quality leads to higher prices. The quality of the house is also an important predictive variable. Lastly, houses located in residential zones, especially low-density residential areas, have higher average prices compared to those in medium to high-density residential, industrial, and commercial zones. Floating Village Residential exhibits the highest average house price, showcasing the impact of house zone and type on price disparities.

- In both regression and classification models, "YearBuilt" consistently stands out as the one of most influential features for prediction. It's noteworthy that houses with a younger age tend to command higher average prices. When considering the interaction with house size, it becomes evident that larger houses tend to exhibit a stronger inclination to preserve their value over time. On the other hand, houses with medium and small dimensions experience a more pronounced decline in value as they age.

- An intriguing phenomenon emerges: the number of bathrooms has a favorable effect on house prices. When considering houses of equal size, the presence of additional bathrooms leads to an increase in the average price by approximately $5,000 to $10,000 per bathroom. Notably, for houses with a size exceeding 3600 square feet, those featuring three bathrooms command prices that are $30,000 higher than houses with just one or two bathrooms.

- The RandomForestClassifier proves its effectiveness in predicting price ranges, achieving an average accuracy rate of nearly 80%. This conclusion is drawn by evaluating its performance across various metrics, including Accuracy, Recall, AUC, and the balance between bias and variance.

4.2 Future Work

Based on the limitations and expectation of this project, some works may be extended by following:

- Further in-depth exploration of data features can be conducted by incorporating specific models for feature selection. Techniques like Embedded Feature Selection Methods and Wrapper Feature Selection Methods (Liu, 2019; A. ElDahshan, 2023) can be employed to enhance the selection of relevant features.

- While the dataset of this project encompasses numerous features, the relatively limited sample size could be a significant factor contributing to the lack of precision in prediction outcomes. Moving forward, acquiring larger datasets from major international cities like Sydney and New York, and expanding the dataset with more samples from such urban centers might provide further insights into the model's performance.

- From the application point of view, cloud computing technology might be essential to actualize and deploy the outcomes of extensive big data analytics. This encompasses facets like real-time data processing, setting up data lakes, creating data pipelines, implementing machine learning algorithms, and enabling API interactions. While this aspect goes beyond the scope of this project's objectives, but as an extension it is worth trying.

Reference

Schapire, R.E., 2013. *Explaining AdaBoost, Empirical Inference Festschrift in Honor of Vladimir N*. Vapnik, Springer Berlin Heidelberg , Berlin, Heidelberg.

Wikipedia: The free encyclopedia, 2023. *Panic of 1893*. [online] Available at: <https://en.wikipedia.org/wiki/Panic_of_1893> [Accessed 6 August 2023].

Wikipedia: The free encyclopedia, 2023. *Great Depression*. [online] Available at: <https://en.wikipedia.org/wiki/Great_Depression> [Accessed 6 August 2023].

Wikipedia: The free encyclopedia, 2023. *Ames, Iowa*. [online] Available at: <https://en.wikipedia.org/wiki/Ames,_Iowa#Demographics> [Accessed 6 August 202 3].

Leguizamon, S. 2010. The Influence of Reference Group House Size on House Price: Reference Group House Size and Price. *Real Estate Economics*, 38(3), pp.507–527.

Domingos, P. 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10), pp. 78–87.

Heyong, W, Ming, H & Shyu, M. 2014. Study on the Use of Equidistant Binning on Residential Hedonic Price Discretization. *Information Technology Journal,* 13(13), pp. 2121–2121.

Jang, J, Oh, H, Lim, Y & Cheung, Y.K., 2021. Ensemble clustering for step data via binning. *Biometrics*, 77(1), pp. 293–304.

Meyer, D.W. 2018. Density estimation with distribution element trees. *Statistics and Computing*, 28(3), pp. 609–632.

Liu, H, Zhou, M & Liu, Q, 2019. An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica*, 6(3), pp.

703–715.

A. ElDahshan, K, A. AlHabshy, A & Thamer Mohammed, L, 2023. Filter and Embedded Feature Selection Methods to Meet Big Data Visualization Challenges. *Computers, Materials & Continua*, 74(1), pp. 817–839.