Housing Price Prediction

Kausik Chattapadhyay

Bellevue University

DSC 550 Data Mining

Introduction

The housing market is a crucial sector that affects individuals, communities, and the economy at large. Housing price prediction is a significant data science project that aims to forecast property prices accurately. By leveraging historical data, I can develop models that estimate the potential value of houses based on various features.

Problem Statement:

- House price prediction is crucial for driving real estate efficiency.
- The goal of this project is to examine if machine learning can reliably predict house sale prices and identify important variables for the analysis.
- Housing prices are influenced by various factors such as location, size, number of rooms, amenities, and market conditions.
- Determining the fair value of a property is challenging for individuals, homebuyers, real estate agents, and investors due to the complexity of these factors.
- Data science techniques will be used to build models that capture the intricate relationships and provide reliable predictions for housing prices.

Justification of importance/usefulness:

- Solving the housing price prediction problem empowers homebuyers to make informed decisions and negotiate fair prices.
- Accurate predictions provide insights into the factors driving property values, helping homebuyers avoid overpaying or underestimating a property's worth.
- Real estate agents benefit from accurate predictions by offering better guidance to clients and streamlining their operations.
- Investors can leverage housing price predictions to identify profitable opportunities and make informed investment decisions.
- Predictions guide developers in identifying lucrative areas for new projects.
- Accurate housing price prediction assists policymakers in understanding market dynamics for effective urban planning.
- Overall, accurate housing price prediction promotes transparency, efficiency, and stability within the housing market.

Pitch to stakeholders:

- The data science project aims to accurately predict property prices, revolutionizing the housing market.
- Valuable insights will be provided to homebuyers, real estate agents, investors, developers, and policymakers.
- The project envisions a world where homebuyers can confidently make informed decisions and real estate agents can optimize their operations.
- Investors can identify profitable opportunities, and policymakers can make data-driven decisions for effective urban planning.
- The project already has relevant housing data from trusted sources, ensuring the accuracy and reliability of the predictions.

Dataset Details

I will be using a housing dataset from Kaggle that examines data from various cities with several attributes. The data for my housing price prediction project can be obtained from reliable sources, including public real estate databases, government records, and industry-specific platforms. I have collected a comprehensive dataset that includes relevant features such as location, property size, number of rooms, amenities, and historical transaction data.

Organized and detailed summary

Exploratory Data Analysis:

Visualizing numerical predictor variables with Target Variables.

1. Note that some of the features have quite high correlation with the target. These features are really significant.

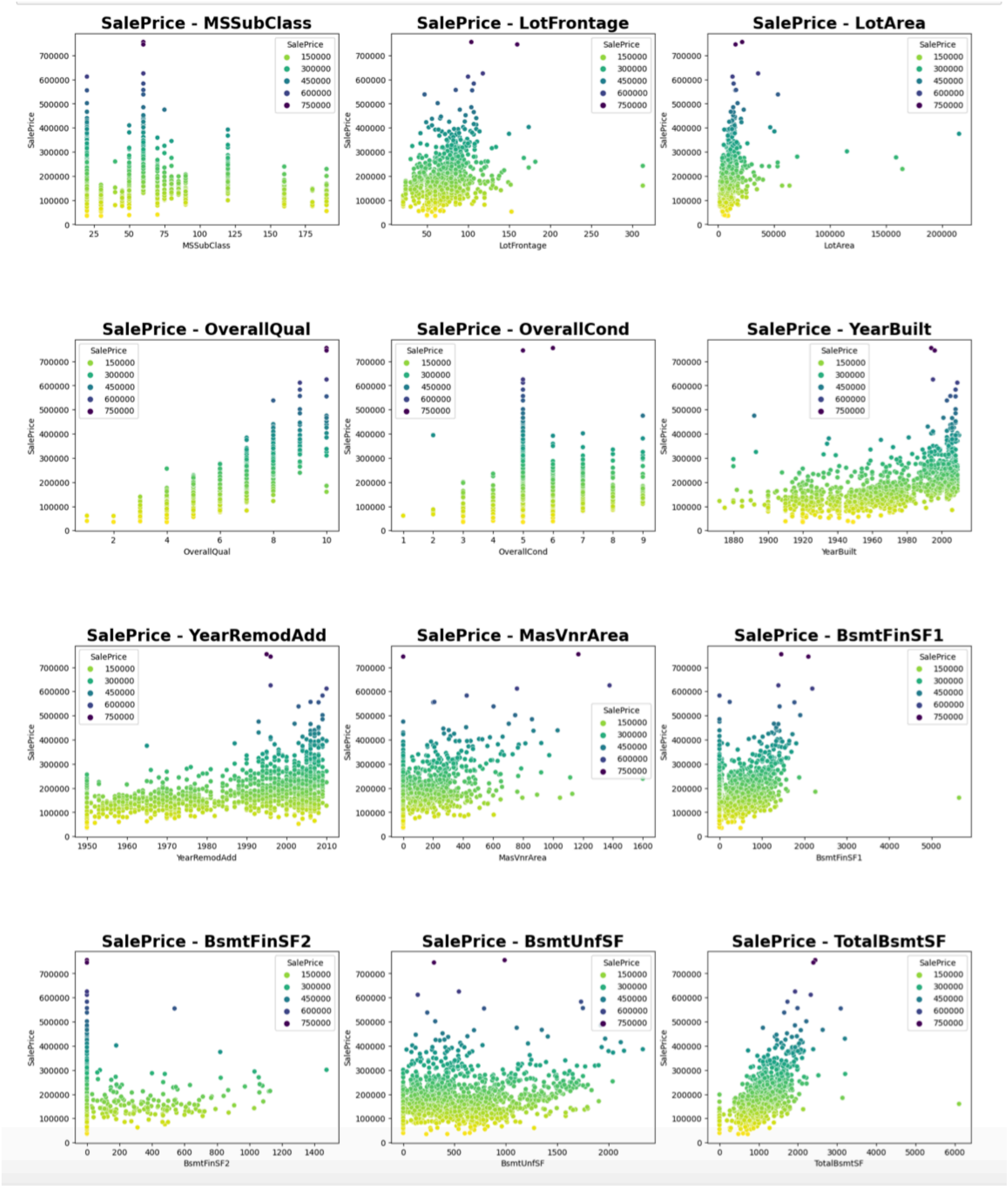
2. Of these the features with correlation value >0.5 are really important. Some features like GrLivArea etc.. are even more important.

3. We will consider these features (i.e. GrLivArea,OverallQual) etc. The Living area and Sale Price have roughly a linear relationship.

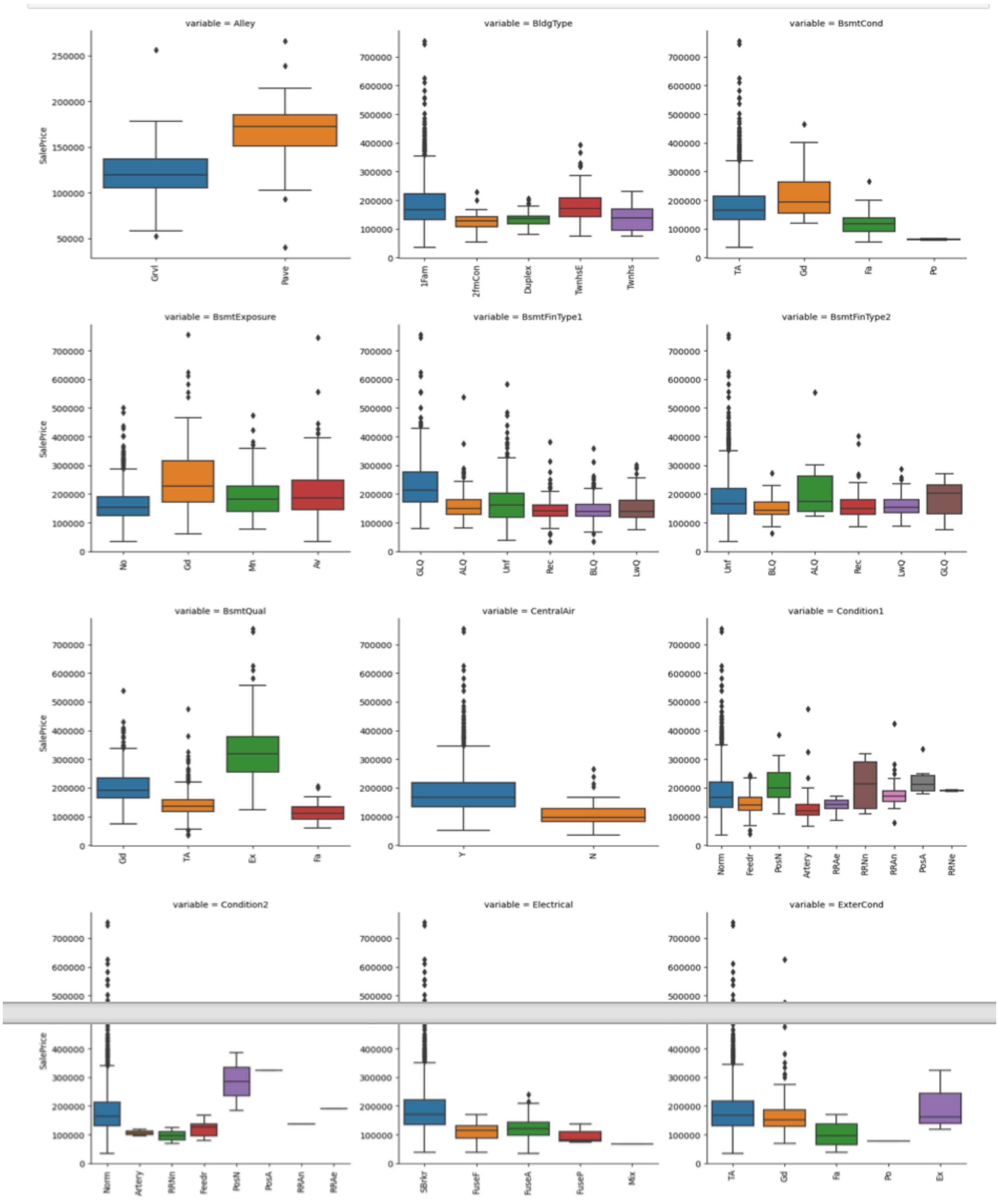
The most relevant features (numeric) for the target are:

```
SalePrice       1.000000
OverallQual     0.817185
GrLivArea       0.700927
GarageCars      0.680625
GarageArea      0.650888
TotalBsmtSF     0.612134
1stFlrSF        0.596981
FullBath        0.594771
YearBuilt       0.586570
YearRemodAdd    0.565608
GarageYrBlt     0.541073
TotRmsAbvGrd    0.534422
Fireplaces      0.489450
MasVnrArea      0.430809
```

GrLivArea’ and ‘TotalBsmtSF’ seem to be linearly related with ‘SalePrice’. Both relationships are positive, which means that as one variable increases, the other also increases. In the case of ‘TotalBsmtSF’, we can see that the slope of the linear relationship is particularly high. ‘OverallQual’ and ‘YearBuilt’ also seem to be related with ‘SalePrice’. The relationship seems to be stronger in the case of ‘OverallQual’, where the box plot shows how sales prices increase with the overall quality.

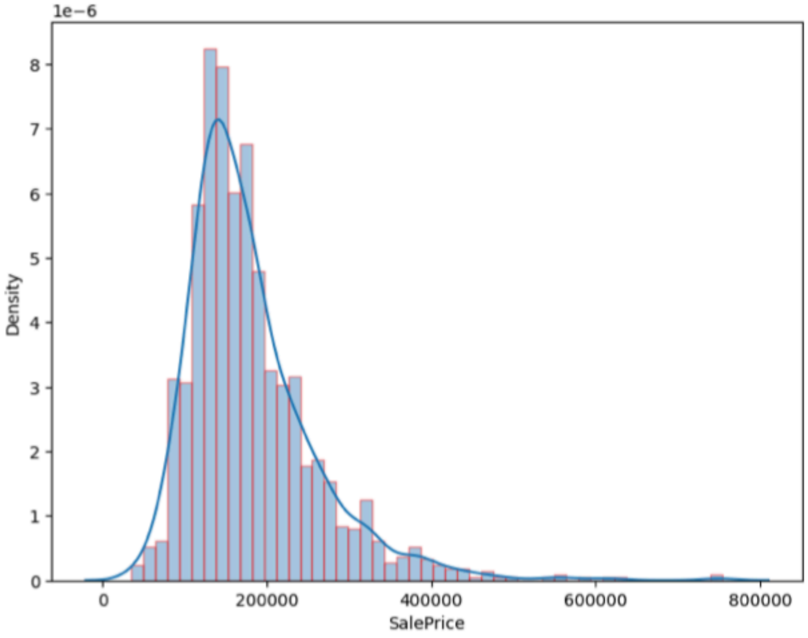

SalePrice - MSSubClass
SalePrice - LotFrontage
SalePrice - LotArea
SalePrice - OverallQual
SalePrice - OverallCond
SalePrice - YearBuilt
SalePrice - YearRemodAdd
SalePrice - MasVnrArea
SalePrice - BsmtFinSF1
SalePrice - BsmtFinSF2
SalePrice - BsmtUnfSF
SalePrice - TotalBsmtSF
SalePrice
150000
300000
450000
600000
750000
0
100000
200000
300000
400000
500000
600000
700000
25 50 75 100 125 150 175
MSSubClass
50 100 150 200 250 300
LotFrontage
0 50000 100000 150000 200000
LotArea
2 4 6 8 10
OverallQual
1 2 3 4 5 6 7 8 9
OverallCond
1880 1900 1920 1940 1960 1980 2000
YearBuilt
1950 1960 1970 1980 1990 2000 2010
YearRemodAdd
0 200 400 600 800 1000 1200 1400 1600
MasVnrArea
0 1000 2000 3000 4000 5000
BsmtFinSF1
0 200 400 600 800 1000 1200 1400
BsmtFinSF2
0 500 1000 1500 2000
BsmtUnfSF
0 1000 2000 3000 4000 5000 6000
TotalBsmtSF

Visualizing Categorical predictor variables with Target Variables



Distribution of Target variable (SalePrice)

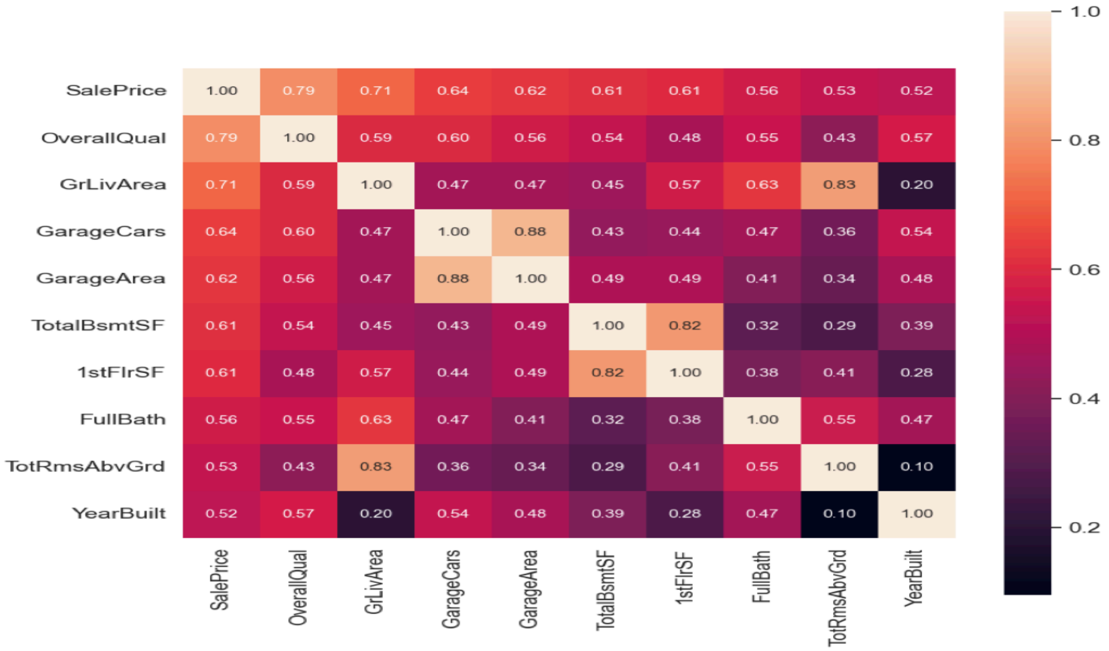
Distribution of SalePrice: Analyzing the distribution of the target variable, SalePrice, can provide insights into the pricing patterns of the housing properties. EDA can reveal whether the distribution is skewed, normally distributed, or exhibits any outliers.



Correlation Analysis: Conducting correlation analysis between numerical features and SalePrice can help identify which features have a strong positive or negative correlation with the sale price. This can assist in understanding the factors that significantly influence the pricing of the houses.

1. OverallQual, GrLivArea, GarageCars, GarageArea, TotalBsmtSF, 1stFlrSF, FullBath, TotRmsAbvGrd, YearBuilt, YearRemodAdd have more than 0.5 Pearson correlation with SalePrice.
2. OverallQual, GrLivArea, GarageCars, YearBuilt, GarageArea, FullBath, TotalBsmtSF, GarageYrBlt, 1stFlrSF, YearRemodAdd, TotRmsAbvGrd and Fireplaces have more than 0.5 Spearman correlation with SalePrice.
3. OverallQual, GarageCars, GrLivArea and FullBath have more than 0.5 KendallTau correlation with SalePrice.
4. EnclosedPorch and KitchenAbvGr have little negative correlation with target variable.

These can prove to be important features to predict SalePrice.



Data preparation

Outlier Remover:

- Outliers are data points that deviate significantly from the normal range and can have a disproportionate impact on the model's performance.
- Statistical methods like z-score or interquartile range, as well as machine learning algorithms like Isolation Forest or Local Outlier Factor, can be used to detect and remove outliers from numerical features.
- The purpose of removing outliers in the preprocessing step is to enhance the model's robustness and accuracy.

Dropping Features:

- In the dataset, certain features may not significantly contribute to the prediction task or have a high number of missing values.
- These features with more than 80% missing can be dropped during the preprocessing stage.
- By removing irrelevant or missing features, the dataset becomes more focused and reduces dimensionality, potentially improving the model's performance and efficiency.

- There are 19 features in the dataset with missing values, which need to be handled.
- The data description indicates that "NA" refers to "No Pool", so the missing values in those features can be replaced with "None" based on the data dictionary.
- For other features with missing values, the approach is to replace them with the median since there are outliers present.

Handling Skewness:

I will be removing Skewness from our model and predictor variables due to following reasons:

- For coefficients to be interpretable, linear regression assumes a bunch of things.
- The errors our model commits should have the same variance and error terms should be normally distributed.
- Following the linear regression assumptions is important if we want to either interpret the coefficients and can be used in business goals.
- When the dependent variable is as skewed as our data is, the residuals usually will be too. Hence, we are handling skewness in our data.

- This model will then be used to understand how exactly the prices vary with the variables`.

One-Hot Encoding with Categorical Features: Categorical features need to be encoded into numerical values for the model to understand them. One-hot encoding is a common technique used in data preprocessing, where each category of a categorical feature is converted into a binary column.

Model building and evaluation

The goal is to predict a numerical outcome, specifically the sale price of houses. Based on this requirement, the author intends to apply linear regression models such as Support Vector Regressor (SVR), Gradient Boosting Regressor, and StackingCVRegressor.

Linear regression is chosen because it assumes a linear relationship between the independent variable(s) and the dependent variable, which aligns with the type of prediction the author is aiming for. Linear regression is a common statistical regression method used for predictive analysis, particularly when examining the relationship between continuous variables.

Applying linear regression models will be suitable for their dataset. It's likely expected that the linear nature of the models will capture the underlying relationships in the data and provide accurate predictions for the sale price of houses.

I plan on using the following metrics to evaluate the results generated: explained variance score metric and the r2 squared metric. Both of these metric functions are provided by the Scikit-learn package that python offers. The explained variance score will explain the dispersion of errors of the dataset I chose to use. The R squared is a score that measures how well the dependent variable of my dataset explains the variance of the independent variable. This metric is commonly used and accepted for most regression models.

During this analysis, I hope to find out if the models I've selected will actually achieve an explained variance score above 70 percent once applied. I also am hoping to learn, if possible, what variables of the dataset that are particularly correlated to the housing prices, negatively or positively.

Conclusion:

The analysis and model building in the US house prediction data science project provide valuable insights into predicting house prices. The models, such as linear regression, Support Vector Regressor (SVR), Gradient Boosting Regressor, and StackingCVRegressor, have been applied to capture the relationships between independent variables and the sale price.

Model Readiness:

While the models have been developed, it is important to assess their readiness for deployment. This involves evaluating their performance metrics, such as accuracy, precision, and recall, against benchmark models or industry standards. Additionally, thorough testing and validation are necessary to ensure the model's reliability and generalizability.

Ethical Implications:

Fairness and Bias: It is essential to ensure that the regression model does not exhibit any biases or discriminate against certain groups of people based on protected attributes such as race, gender, or ethnicity. Care should be taken to avoid perpetuating or amplifying existing biases in the housing market.

Privacy and Data Protection: The dataset may contain sensitive information about individuals, including their addresses and personal details. It is crucial to handle this data with care and adhere to privacy regulations to protect the privacy and confidentiality of the individuals involved.

Transparency and Explainability: The regression model's predictions should be interpretable and explainable, allowing stakeholders to understand the factors influencing the predicted housing prices. This transparency helps build trust and ensures accountability in the decision-making process.

Impact on Housing Market: Machine learning models trained on housing data have the potential to impact the real estate market. It is important to consider the potential consequences and unintended effects of deploying the model. The model's predictions should not artificially inflate or deflate housing prices or contribute to housing market volatility.

Responsible Data Collection: When collecting and using the housing dataset, it is important to ensure that the data is obtained through ethical means and in compliance with applicable laws and regulations. Data should be obtained with proper consent and in a manner that respects individuals' privacy and rights.

Mitigating Displacement and Gentrification: Regression models predicting housing prices can indirectly impact communities by influencing investment decisions and urban development. It is important to consider the potential effects on local communities, such as displacement and gentrification, and take measures to mitigate any negative impacts.

Addressing these ethical implications ensures that the regression machine learning problem is approached with fairness, transparency, and social responsibility, mitigating potential harm and promoting ethical practices in the housing domain.

Recommendations:

Based on the analysis, it is recommended to further fine-tune and optimize the models to improve their predictive performance. This can involve feature selection or engineering, hyperparameter tuning, and cross-validation techniques. Furthermore, it is crucial to evaluate the models' robustness by testing them on unseen data or conducting out-of-sample validation.

Challenges and Additional Opportunities:

Some potential challenges in the US house prediction project may include handling missing data, dealing with outliers, and selecting the most relevant features for modeling. Additionally, incorporating temporal aspects, such as capturing market trends or seasonality, could enhance the accuracy of predictions. Exploring the integration of alternative models, ensemble methods, or advanced techniques like neural networks may also provide further opportunities for improving model performance.

In summary, the analysis and model building in the US house prediction data science project offer valuable insights. However, the readiness of the model for deployment needs to be carefully assessed, and further refinement and optimization are recommended. Addressing challenges and exploring additional opportunities will help enhance the accuracy and applicability of the model in the real estate market.

References

Begum, A., Kheya, N. J., & Rahman, Z. (2022, January). Housing Price Prediction with Machine Learning. Retrieved from https://www.ijitee.org/wpcontent/uploads/papers/v11i3/C97410111322.pdf Goyal, C. (21, May 2021). Importance of Cross Validation: Are Evaluation Metrics enough?

Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/05/importance-of-cross-validation-areevaluation-metrics-enough/ Shin, T. (2019, December 12). Exploratory Data Analysis — What is it and why is it so important? (Part 1/2). Retrieved from medium.com: https://medium.com/swlh/exploratory-data-analysis-what-is-it-and-why-is-it-soimportant-part-1-2-240d58a89695

https://github.com/ajayarunachalam/RegressorMetricGraphPlotURL.