

uRBAN hOUSING PRICE PREDICTION

FINAL PROJECT REPORT



Rup Kiran Boyana

Contents

[Introduction: 2](#_Toc152606017)

[Training Supervision: 2](#_Toc152606018)

[Loading Data: 3](#_Toc152606019)

[Assumptions: 3](#_Toc152606020)

[Initial Data Exploration: 3](#_Toc152606021)

[Graph Generation: 5](#_Toc152606022)

[Highlighting Findings: 8](#_Toc152606023)

[Additional Questions: 9](#_Toc152606024)

[Pre-processing: 13](#_Toc152606025)

[Cleaning: 13](#_Toc152606026)

[Handling Outliers: 15](#_Toc152606027)

[Data Partitioning: 16](#_Toc152606028)

[Ensuring Robust Data Preprocessing: 16](#_Toc152606029)

[Model Selection: 22](#_Toc152606030)

[Performance Measure: 26](#_Toc152606031)

[Learning Curve Graph: 27](#_Toc152606032)

[Regularization Techniques: 28](#_Toc152606033)

[Additional Questions: 29](#_Toc152606034)

[Hyperparameter Tuning: 29](#_Toc152606035)

[Regularization Adjustments: 31](#_Toc152606036)

[Evaluation Metrics for Fine-Tuning: 32](#_Toc152606037)

[Summary of Findings: 32](#_Toc152606038)

[Model Performance Showcase: 33](#_Toc152606039)

[Insights Derived from Data: 33](#_Toc152606040)

[Challenges Encountered: 34](#_Toc152606041)

# Introduction:

• Identify your problem.

The "Urban Housing Prices" project aims to delve into the intricate dynamics of urban housing markets with the primary objectives of predicting housing prices for prospective homebuyers and providing valuable insights for real estate investors. The dataset, curated meticulously, encompasses diverse features that influence the housing market, creating a foundation for a comprehensive exploration.

• What are the primary objectives of your project?

The primary objective is to predict housing prices, which involves predicting a continuous numerical value.

# Training Supervision:

• Is the model's learning process supervised, unsupervised, semi-supervised, selfsupervised, or reinforcement learning? Explain.

The model's learning process in this project is supervised learning task. In supervised learning, the algorithm is trained on a labeled dataset, where the input data includes both features and corresponding target labels (in this case, housing prices). The goal is for the model to learn the mapping from input features to the target labels during the training phase.

• Identify the type of task: classification, regression, or something else?

Regression is well-suited for problems where the output is not a category but a quantity, making it suitable for predicting housing prices in this context.

• Should the model employ batch learning or online learning techniques, and why?

For this project, the model should employ batch learning techniques. Batch learning involves training the model on the entire dataset at once. In the context of predicting urban housing prices, batch learning is suitable when the dataset size is manageable, and the model can be trained efficiently in one go.

Batch learning is preferable in scenarios where the dataset can fit into memory, and retraining the model with new data is not frequent. Since housing prices tend to be influenced by relatively stable and longer-term factors, the periodic update of the model using batch learning can capture the evolving patterns in the housing market.

In contrast, online learning would be more suitable for scenarios where the data is continuously streaming, and the model needs to adapt to changes in real-time. However, for predicting housing prices, where the market dynamics may not change rapidly, the efficiency and simplicity of batch learning make it a more pragmatic choice.

# Loading Data:

• What methods will you use to load the dataset into your project?

• How can you ensure the data is properly loaded and accessible for analysis?

From, the panda’s library I have used read\_csv method to load the data and used head() to display the first 5 rows

A screenshot of a computer screen

Description automatically generated

# Assumptions:

• List the assumptions you are making about the data (e.g., district prices are numerical).

I have observed the target variable, which is the housing prices, is numeric. This is a common assumption for regression tasks where the goal is to predict a continuous value.

• Are there any specific features you assume to be categorical or numerical? Discuss this

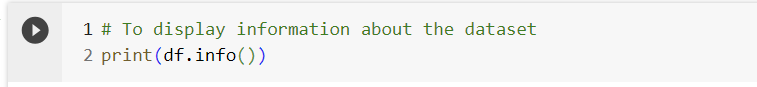
No there are no assumptions as such

# Initial Data Exploration:

• Before diving deep, what preliminary steps will you take to understand the structure of the data?

I have used head(), info() and describe() methods to understand the dimensions of the data set and with the help of the info() method I have figured out the number of columns that are categorical data and numerical data. There are a total of 2919 rows and 81 columns in this data set out of which 38 columns have numerical data and 43 columns are categorical data.

• Can you provide examples of Python functions (e.g., head(), info(), describe()) thatwould be helpful in this initial exploration?

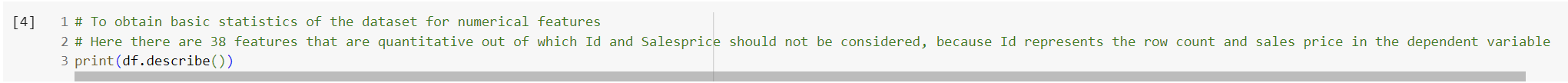


A screenshot of a computer code

Description automatically generated

A close up of a number

Description automatically generated



A screenshot of a computer screen

Description automatically generated

# Graph Generation:

• Create four types of graphs to visualize/explore the data effectively?

Bar plot to understand which features have the highest null values

A graph with blue and black text

Description automatically generated with medium confidence

Histogram for Target Variable to understand the distribution of Sale Price

A graph of distribution of sales

Description automatically generated

Correlation Heatmap to understand relationship between the numerical features and the Sale Price

A screenshot of a graph

Description automatically generated

Boxplot to understand the Sale Price by Neighborhood

A graph with different colored squares

Description automatically generated

Scatterplot to understand how OverallQual, GrLivArea, GarageCars and TotalBsmtSF are effecting SalePrice

A graph of different types of data

Description automatically generated with medium confidence

• How do these graphs contribute to gaining insights into the dataset?

Overall, I have created 5 graphs to understand the data

I have created the Bar plot to understand which features have the highest null values

Histogram help to show the distribution of the Target Variable, Sale Price

Correlation Heatmap is used to understand relationship between the numerical features and the Sale Price

Boxplot is used to understand the Sale Price by Neighborhood

Scatterplot helps to understand how OverallQual, GrLivArea, GarageCars and TotalBsmtSF are affecting SalePrice

# Highlighting Findings:

• What are four specific findings you aim to highlight through your data exploration?

There are highest NA values in features – Alley, FireplaceQu, PoolQC, Fence, MiscFeature and SalePrice

The frequency of SalePrice is high between the range 100000 and 200000

There is high correlation for features OverallQual, GrLivArea, GarageCars and TotalBsmtSF and SalePrice

The neighborhoods NoRidge, NridgHt, Timber and StoneBr have highest sale Prices compared to others

• How do these findings contribute to a better understanding of the dataset?

This interpretation helps to understand how the independent variables are affecting the depend variable or the target variable SalePrice

Statistical Analysis:

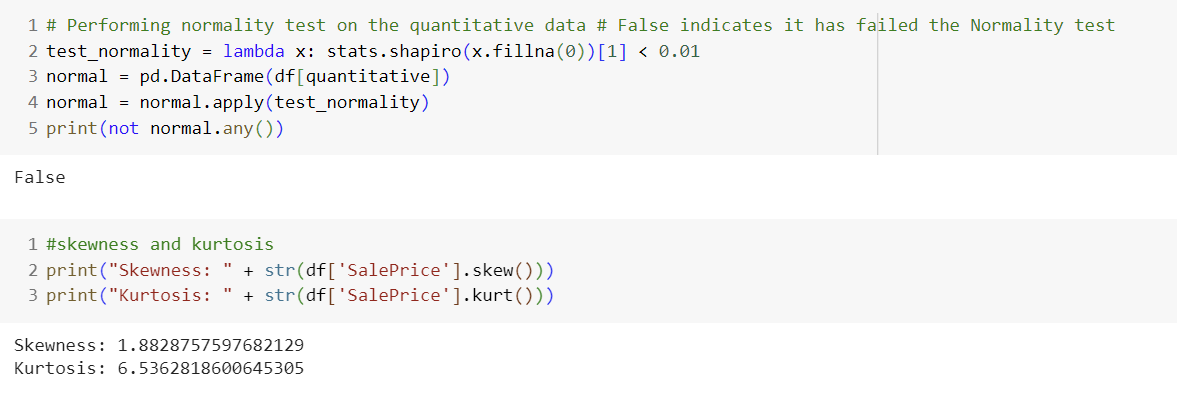
• In what ways will you perform statistical analysis on the data?

Initially I performed normality test on the quantitative data using Shapiro-Wilk test for normality. The lambda function is used to update NA values to 0, here if the p value is greater than 0.01 it returns false indicating that one of the quantitative columns did not pass the normality test.

Later I have performed skewness and Kurtosis on the Target Label, SalePrice the output of the skewness is 1.88 indicating that the distribution is skewed to right resulting in right skewed distribution and kurtosis value 6.5363 is higher than the kurtosis of normal distribution which is 3.

• How will statistical measures enhance your insights into the dataset? Show results.

This statistical measure helps to decide on transformations for SalePrice to address skewness or outliers, or you may choose modeling techniques that are robust to non-normality based on these insights.



## Additional Questions:

• How will you handle outliers and anomalies during the exploration process?

Here I am only considering the outliers from the features that are highly correlated with the SalePrice

OverallQual vs SalePrice - There are no outliers when we check the above scatter plot

GrLivArea vs SalePrice – There are two outliers from the above scatter plot

GarageCars vs SalePrice – There are no outliers

TotalBsmntSF – There is one outlier

I have removed the two outliers from the GrLivArea

A screenshot of a computer code

Description automatically generated

After removing the outliers from GrLivArea

A screen shot of a graph

Description automatically generated

• What role does feature correlation play in your data exploration strategy?

It helps to understand the relationship between features in dataset, if the features are highly correlated it means that there is a strong relation between them, and we focus on those features compared to others. Also, we can figure out which features are negatively correlated.

• Can you identify potential patterns or trends that may impact your later modeling decisions?

Homoscedasticity means the variance being constant even when the variability of dependent variable does not increase as the value of the independent variable Increase.

Heteroscedasticity means the variance changes for dependent variables as there is change in the independent variables

And from the below plots we can understand that there is random scattering of points across the horizontal axis and the spread of points changes with the values of the independent variable indicating heteroscedasticity.

A screenshot of a computer screen

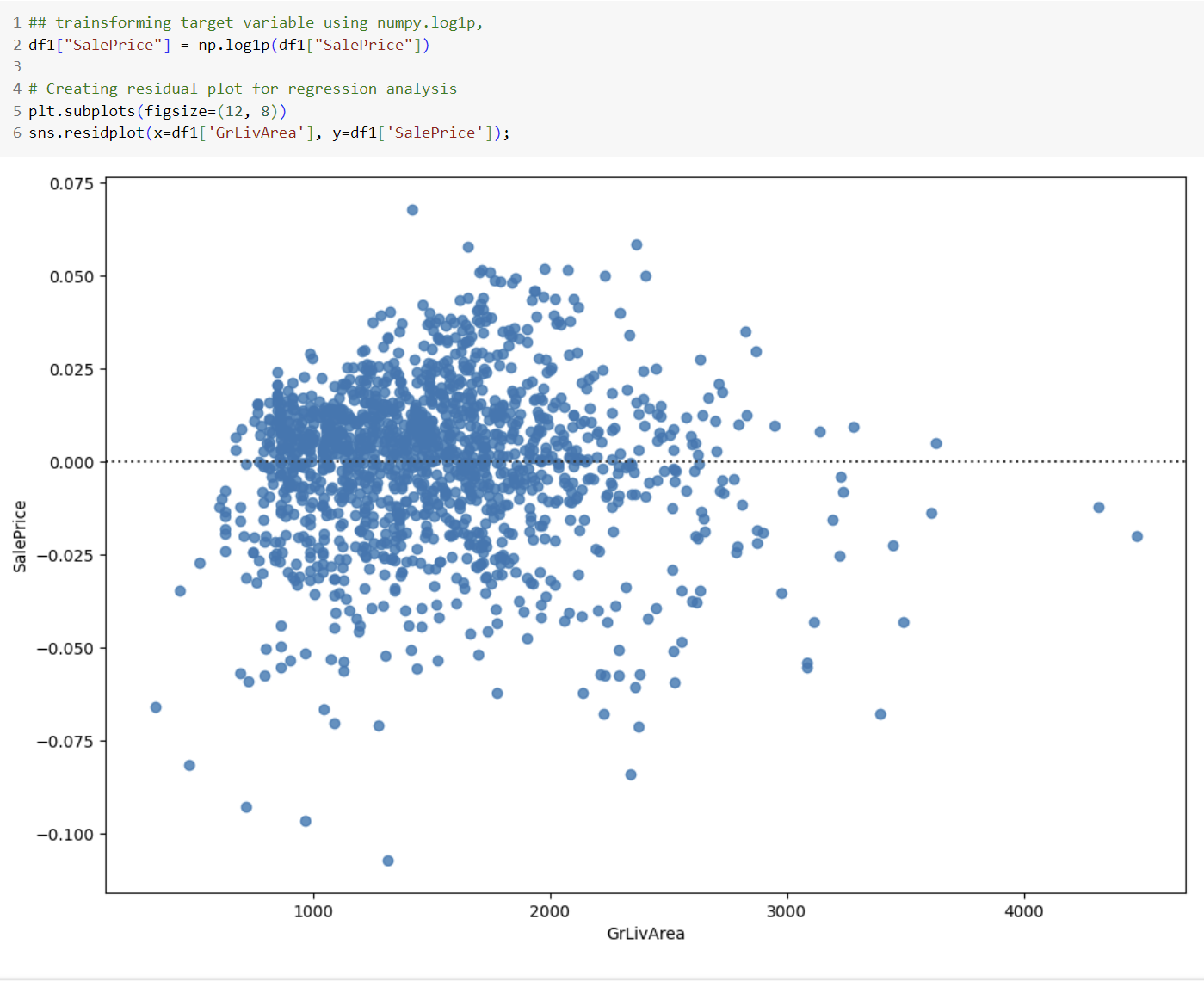
Description automatically generated

A screen shot of a graph

Description automatically generated

• How does the initial exploration guide your decisions on data preprocessing?

Here I have used log transformation on the dependent variable to normalize skewed distribution and stabilizing variance mainly. Also, checked for skewness and kurtosis



As we see from the results below the skewness is nearly equal to zero indicating the Normal distribution and the kurtosis value is less than 3 suggests a distribution that is less peaked and has lighter tails compared to a normal distribution.

A computer code with numbers and symbols

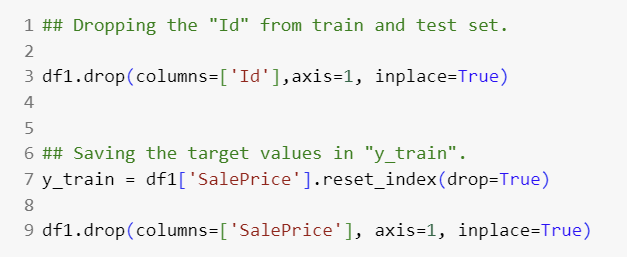
Description automatically generated

# Pre-processing:

• What specific steps will you take to pre-process the data before feeding it into machine learning algorithms?

In this step I have taken the Insights from the earlier steps, firstly figured the columns that need data preprocessing. After identifying the columns, I have worked on the columns for data quality issues by type conversion and populating the NA values with relevant data using grouping, median and mode techniques. Later worked on updating the NA values as None or 0 based on column datatype.

Before pre-processing the data, I dropped column ID from the df1 data frame. Later stored SalePrice in the y\_train as it is dependent variable. After that I dropped the target column from df1 just to work the independent variables.



• How do these steps contribute to improving the quality of the dataset?

These steps helped to avoid Inc

## Cleaning:

• Identify potential data quality issues. How will you address them during the cleaning process?

Converted MSSubClass from int to type string as it represents category

There are also NA values in the SalePrice stored in the y\_train I have also dropped those rows

Lastly Converted year sold and month sold to strings as we do not perform any numerical operations on this data and these act as labels



A computer screen shot of a program code

Description automatically generated

A close up of words

Description automatically generated

• How will you handle missing values to ensure the completeness of your dataset?

In this step I have considered the columns with highest number of NA values and updated the Missing values as None from the categorical features and for Numerical columns updated the Missing values to 0 as there is a reason for Null.

Now I have considered the columns with least number of NA values and updated columns Functional, Utilities, KitchenQual and Electrical to Categories based on the dataset description

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A computer code on a white background

Description automatically generated

# Handling Outliers:

• How do outliers impact machine learning models, and how will you decide on an appropriate approach to handle them?

There are many aspects of outliers impacting machine learning, mainly it affects the model evaluation and performance. Considering this during the explore and visualization step I have eliminated the potential outliers from the data set from column GrLivArea.

A close-up of a number

Description automatically generated

• What techniques can be employed to detect and address outliers effectively?

I have used visualization techniques and plotted scatterplot to detect any patterns or trends in order to identify outliers. Also performed log transformation the dependent variable for skewness and kurtosis correction and handle outliers.

A diagram of a number of blue dots

Description automatically generated A diagram of a graph

Description automatically generated

# Data Partitioning:

• Why is it crucial to partition the data into training and testing sets?

Partitioning data into training and testing sets is essential for developing reliable and generalizable machine learning models. It helps to identify if there are overfitting and underfitting issues when the test data is provided. Also, when using multiple models this partitioning helps unbiased evaluation.

• How will you determine the optimal ratio for data partitioning?

This is a trial an error process but in general it is advised to consider 80% for training data and 20% for testing data or 70% for training and 30% for testing.

A screen shot of a computer code

Description automatically generated

# Ensuring Robust Data Preprocessing:

• What role does feature scaling play in the preprocessing phase, and why is it important?

Feature scaling is used to convert all the independent variables into a single scale. For the categorical variables I have used pd.dummy() to perform one-hot encoding on the categorical columns of the DataFrame df1. It creates binary columns for each category and assigns 1 or 0 based on the presence of the category in the original data.

Later I used standardscaler method to standardizes or normalizes the range of independent variables or features of the dataset.

We perform feature scaling because Algorithms that rely on distances between data points can be affected by these differences in scales. Feature scaling ensures that all features contribute equally to the model's learning process.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

• How do you ensure that the preprocessing steps are consistent across training and testing datasets?

I have initially checked for skewness and kurtosis for the numerical variables and worked on decreasing the skewness of the features with features of skewness greater than 0.5.

After performing pd.dummy() there is a chance of duplicate variables, so I have deleted dummy columns from the df2 dataframe

Creating training (X\_train), testing (X\_test) sets, and printing the shapes of these sets along with the shape of the target variable (y\_train).

Created Overfit reducer method for the features if the percentage of the most frequent value is greater than 99.94%, the feature is considered overfitted. I have identified MSSubClass feature as overfitted and removed it from the X\_train and X\_test.

A screenshot of a computer program

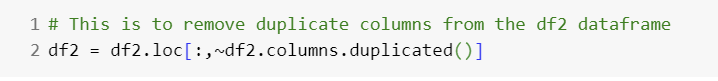
Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated



A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

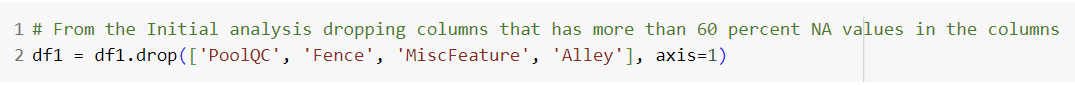
• Can you explain the trade-offs associated with different methods of handling missing data?

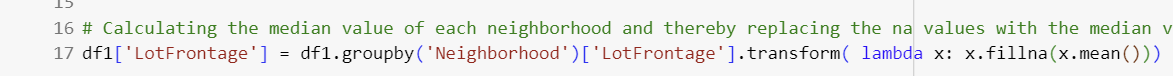
Dropped columns 'PoolQC', 'Fence', 'MiscFeature', 'Alley' which has NA values more than 50 percent from the initial analysis.

Grouped MSZoning with MSSubclass and for each group, it fills missing values in the 'MSZoning' column with the mode (most frequent value) of that specific group.

Updated NA values in columns Exterior1st, Exterior2nd and SaleType to mode value.

Calculated the median value of each neighborhood and thereby replacing the NA values with the median values in the LotFrontage, I have done this for better prediction and accuracy.





A screenshot of a computer code

Description automatically generated

A close-up of a computer code

Description automatically generated

• How will you validate the effectiveness of your data preprocessing steps in improving model performance?

Based on the data we can understand that the dependent variable is a continuous numerical variable. Hence, I have proceeded to check all the regression models with the available data.

# Model Selection:

• How will you decide on the machine learning model to use for your predictive modeling task, and why is it suitable? Explain.

Here we have the SalePrice as dependent variable, when we have dependent variables, we use supervised learning techniques. Here we are using regression to predict a continuous numeric value. There are many regression techniques as we are not sure which models obtains the best results, I have checked all the regression techniques.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

• Are there alternative models considered, and what influenced your final choice?

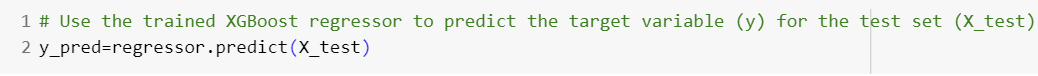
Yes, I have used xgboost ensemble learning technique and it provided me with a better rmse value than all the regression techniques used.

A close-up of a white background

Description automatically generated

A screenshot of a computer program

Description automatically generated



A screenshot of a computer error

Description automatically generated

# Performance Measure:

• Define the performance measure you will use to evaluate your model's

I have obtained root mean square error value for all the regression techniques and the ensemble learning technique. It is a common metric used to evaluate the accuracy of a regression model. It measures the average magnitude of the errors between predicted values and actual values.

effectiveness.

• How does this chosen metric align with the objectives of your project?

Considering the objective of the project to predict SalePrice. RMSE is a measure of prediction accuracy, indicating how well the predicted sale prices match the actual sale prices. Minimizing RMSE ensures that the model is making accurate predictions, and the errors are reflective of the actual differences between predicted and observed sale prices.

# Learning Curve Graph:

• Generate learning curves graph.

A screenshot of a computer program

Description automatically generated

A graph of a graph of a number of different colored lines

Description automatically generated with medium confidence

• What insights can be gained from the learning curve graph, and why is it a valuable tool in model evaluation?

From the learning curve graph, we can understand if the model is underfitting or overfitting.

• How will you interpret the learning curve in the context of overfitting and underfitting? Explain.

If there is a large gap between the training and validation errors the model may have high variance, indicating overfitting. Collecting more data can help generalize the model better, reducing overfitting.

Overfitting/Underfitting Discussion:

• Define overfitting and underfitting. How will you identify and address these challenges during the training process?

Overfitting means the model learns to well from the training data and performs well on the training data but fails to perform on the test data.

Underfitting means when the model is too simple to capture the underlying pattern. This does not perform well on both the training and the test data

We can identify this issue from the learning curve. If the validation error is below the median line, then we can say it is underfitting if the distance between the validation error and the training error is high we can say it is overfitting.

We can address these issues by removing unwanted columns, standardization, decreasing the skewness and the kurtosis values of the features.

• What adjustments might be necessary if overfitting or underfitting is observed in the learning curve?

As mentioned earlier we can do regularization, cross validation and feature selection are some of them

# Regularization Techniques:

• Did you employ any regularization techniques in your model training? If so, why were they necessary?

Yes, I have used Lasso, Ridge, Elasticnet regularization techniques on the training split data to find the RMSE value. Please refer to the earlier code snippets attached.

• Explain how regularization contributes to improving the model's generalization performance.

For me even when I have used these techniques ensemble learning xgboost has obtained the optimal RMSE value. Please refer to the earlier code snippets attached.

## Additional Questions:

• How will you optimize hyperparameters to fine-tune your model?

I have used randomcv method to optimize the hyperparameters

• What considerations guided your decision on the model's complexity?

I have considered conditions such as Number of boosting rounds, Maximum depth of a tree, Step size shrinkage to prevent overfitting, Minimum sum of instance weight needed in a child, Type of boosting model (tree or linear), and Initial prediction score for all instances

• Can you discuss the impact of the chosen performance measure on the interpretability of your results?

Before considering the hyperparameters the RMSE value for the xgboost model is 0.145232 and after considering the hyperparameters and using randomcv method the RMSE value for the xgbosst is 0.12067347

# Hyperparameter Tuning:

• How will you approach the optimization of hyperparameters to enhance your model's performance?

As there are 81 columns which are independent variables to predict one variable sale price, I have used random search optimization technique. I did not use grid search as it is exhaustive to check each and every possible combination for the optimization which will require more computational time.

• Which hyperparameters are considered critical for fine-tuning, and why?

Number of Boosting Rounds (n\_estimators): It represents the number of trees (weak learners) added to the model. A higher number of rounds can improve performance, but there's a risk of overfitting if set too high.

Maximum Depth of a Tree (max\_depth): It's crucial to tune this parameter to find the right balance between model complexity and generalization.

Step Size Shrinkage (learning\_rate): Determines the contribution of each tree to the final prediction and it helps prevent overfitting and makes the optimization process more robust.

Minimum Sum of Instance Weight Needed in a Child (min\_child\_weight): It helps control the partitioning of nodes during tree construction. Tuning this parameter can influence the balance between tree complexity and overfitting.

Type of Boosting Model (tree or linear): Tree models are powerful for capturing complex relationships, while linear models are useful for linearly separable problems.

Initial Prediction Score for All Instances (base\_score): It plays a role in determining the starting point for the boosting process. It is often set to the meaning of the target variable.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

# Regularization Adjustments:

• If regularization was initially applied, will you make any adjustments during the fine-tuning phase?

After obtaining the learning curve I have used the Lasso and Ridge techniques but there is no change in the RMSE value. Hence, I have proceeded to consider the ensemble learning technique which is xgboost.

A screenshot of a computer code

Description automatically generated

• How do regularization parameters influence the model's generalization to unseen data?

In XGBoost, regularization is controlled by the alpha parameter for L1 regularization and the lambda parameter for L2 regularization.

For example, setting alpha=0 and lambda=0 implies no regularization, while increasing these values introduces regularization.

Increasing the values of alpha and lambda introduces regularization by adding penalties to the loss function, influencing the optimization process during training.

# Evaluation Metrics for Fine-Tuning:

• Define the evaluation metrics you will use during the fine-tuning phase.

I have used RMSE evaluation metric. RMSE measures the average magnitude of the errors between predicted and actual values, giving more weight to larger errors. Lower RMSE values indicate better model performance.

• How will you interpret these metrics to assess improvements in model performance?

RMSE is calculated between y\_test and y\_pred where y\_test is the dependent variable and the y\_pred is the regressor model prediction output made on the X\_test. So RMSE value defines how well the model has predicted. Lower RMSE indicates better model performance.

# Summary of Findings:

• What key findings will you include in the summary of your project?

From the dataset we can understand that there are 81 columns overall out of which 43 are categorical and 38 are numerical

I have figured out there are more than 6 columns which are Alley, FireplaceQu, PoolQC, Fence, MiscFeatures and SalePrice have more than 50% NA values

Distribution of the SalePrice is normal

Correlation of the numerical columns OverallQual, GRLivArea, GarageCars, GarageArea, TotalBsmtSF, 1stFlrSF, FullBath, TotRmsAbvGrd, YearBuilt, YearRemodAdd with SalePrice is more than 0.50 on a scale of 1

I have plotted a box plot for the categorical variable the neighborhood column and noticed NridgHt and StoneBr have highest SalePrice among all the other Neighborhoods

From the scatter plots I have interpreted that GRLivArea has outliers which are not following the regular trend.

Removed ID column from the data set as it has no value for any sort of Interpretation

Calculating the median value of each neighborhood and thereby replacing the na values with the median values in the LotFrontage, we do this for better prediction and accuracy

Converting year sold and month sold to strings as we donot perform any numerical operations on this data and these act as labels

• How do these findings align with the initial objectives of your predictive modeling task?

As we are predicting the SalePrice value based on the 81 columns which are independent variables it is important to gain insights on the data to perform data cleaning, data preparation, deleting unwanted data, update NA values, to normalize the data, standardize the data, convert the categorical data into the numerical data in order to improve the model prediction and gain optimized results.

# Model Performance Showcase:

• How will you effectively communicate the performance of your final model to your audience?

I have obtained RMSE value for xgboost model, Ridgeregression model, Lasso regression model, Elastic net model, SGD regression model, Decision tree regressor model, Random Forest model and SVR model among all this model I have noticed optimal results for ensemble model xgboost and RMSE value is 0.12067347234319216

I have plotted learning cure for the xgboost regressor model which helps to understand the overfitting and the underfitting issue of the model’s performance for the training and validation data.

• Create visualizations or show metrics that are essential in conveying the model's predictive capabilities?

In the earlier snippets provided RMSE value for all the regression models used and for the ensemble model xgboost is shown. Kindly refer the RMSE metric value above.

# Insights Derived from Data:

• In presenting your solutions, what insights from the data exploration phase will you emphasize?

As explained earlier I have understood that are lot of discrepancies in the data, so I have removed unwanted columns, checked skewness and kurtosis of the SalePrice column and performed log transformation on the data to transform into Normal distribution. For all the numerical features I have created a function to reduce the skewness of the features with features with skewness greater than 0.5, Converted categorical columns to numerical columns by using method pd.dummy(), for some of the columns based on the understanding I have updated NA values using median, mode techniques, converted year sold and month sold to strings as we do not perform any numerical operations on this data and these act as labels finally used overfitted reducer function on each and every column and figured MSSubClass is overfitting. So removed the column from the dataset.

• How do these insights contribute to a deeper understanding of the factors influencing your model's predictions?

I have mainly considered visualizations, from the Insights gained from the visualization I have worked on skewness and kurtosis to normalize the numerical data, as all the data provided to the machine learning model has to be numerical, I have worked on converting categorical data to the numerical data using relevant methods. By checking correlation matrix understood which columns have high correlation with the SalePrice column. All the steps I have performed formed the gained insights have resulted in reaching optimal results of the model and generating minimum RMSE value.

# Challenges Encountered:

• Can you discuss any challenges faced during the project and how they were overcome?

When decreasing skewness for the numerical variables I have got exceptions I have then used try and except to handle the situation. As there are many columns it took me a while to understand the data and work on it.

• What lessons have you learned that might be beneficial for future projects?

I have learnt about different Machine learning models and understood where I have to exactly use them for different data and what metrics that I have to consider for this type of projects to evaluate the machine learning models performance.

• Can you discuss the limitations of your model and potential areas for future improvement?

For the machine learning models we have to always provide numerical data and not categorical data. Also, hyperparameter tunning has increased the model’s performance.

Recommendations and Future Work:

• What recommendations will you provide based on your model's predictions?

As I have used multiple machine learning models. Ensemble learning techniques have obtained the optimal results considering the RMSE metric. For huge dataset I suggest we use multiple models together for the optimal values.

• How might your project be extended or refined in future work to improve its applicability?

There is only 50% of the data available in the SalePrice column I am looking forward to predict the SalePrice for the columns using the independent variables.