RUSSIAN HOUSING PRICE PREDICTION

PROBLEM STATEMENT

- While considering the real estate sector, it has been observed that the housing properties are either underpriced or overpriced i.e., they are being bought and sold at prices well above or below the market standard prices which leads to unfair profit or loss among the concerned parties involved (buyers/sellers).
- The requirement of the project is to accurately predict the housing price based on the historical housing sales data which will help to clear the uncertainty related to the housing prices and help generate awareness among customers regarding the true market value of the housing.
- When the customers are aware of the true market value, there is less chance of being exploited and it will also ensure that both the parties involved get the maximum benefit out of the deal.
- The analytic problem that needs to be solved is to predict the housing price based on a certain set of predictor variable values with a high level of accuracy.

CHOICE OF MODEL

- The regression model is used to accurately predict the outcome i.e., the housing price based on the predictor variables like number of floors, living area etc.
- The data science pipeline of the solution will be as follows:
  - i. Extract and Load the data.
  - ii. Transform the data (i.e., Clean and enrich the data).
  - iii. Split the data into training and validation sets.
  - iv. Build a regression model based on the training set.
  - v. Evaluate the efficiency of the model on the validation set.
  - vi. Deploy the model to predict housing prices based on real-world data.

DATA COLLECTION

- The Sberbank Russian Housing Market dataset is used for the regression analysis Link: https://www.kaggle.com/competitions/sberbank-russian-housing-market/data
- Each record in the Sberbank Russian Housing Market dataset corresponds to housing property transactions.
- The target variable is called price_doc.

DATA CLEANING PROCESS

- By initially previewing the dataset of housing_df, we found that this dataset has 292 Columns and some of them are not so relevant for our analysis. So, we first created a new dataset by selecting the 26 columns that are closely related to our project for the next step of our analysis.

1. HANDLING MISSING VALUES

- First, we checked the number of null values in the dataset and found that the column "build_year" has many missing values (13605 null values), because there is a lot of data and it is not appropriate to impute it with median or mode values, so we decided to delete the records with missing values in the "build_year" attribute. After this step, we had 2 more

columns with less than 800 missing values.

CHALLENGES FACED AND ADDRESSED

- The approach of removing missing values is effective, but the potential impact of missing data on the analysis must be considered. Since our dataset has more than 30,000 rows of data, the impact of removing these data is within acceptable limits.

2. IMPUTATION OF MISSING VALUES

- We addressed the remaining missing values by using different strategies based on skewness. For columns with missing values, we calculate the skewness of their distribution. If the skewness is greater than 1 (indicating significant skewness), we impute the missing values with the median of the column. If the skewness is less than or equal to 1 (indicating a relatively normal distribution), we impute the missing values with the mean of the column (Kumar,2023). After this step, we no longer have missing values.

CHALLENGES FACED AND ADDRESSED

- We carefully considered the distribution of each column before deciding on an imputation strategy. The median is less sensitive to outliers and thus provides a more reliable estimate of the central tendency for skewed distributions. The mean is aligning with the assumption of normality and allows us to handle missing values in a way that preserves the distribution's characteristics.

3. DUPLICATE DATA

- In the continuation of our data-cleaning process, we checked for duplicate rows in the subset of the housing dataset. Fortunately, we found that there were no duplicate values present.

4. LOW VARIATION DATA

- We also examined categorical variables to determine if any have only one unique value. Since we didn't find any categorical variables with a single unique value, it indicates that there is no low variation data present.

5. IRRELEVANT DATA

To further refine the dataset, we focused on identifying and addressing irrelevant data related to the "sub_area" attribute: We calculated the distinct values of the "sub_area" attribute and their respective occurrence counts within the dataset. We noticed that some "sub_area" values had very few occurrences, with counts less than 50 (Cloud, 2023). Recognizing the potential impact of such infrequent occurrences on biasing our analysis, we decide to exclude records where the number of occurrences of the "sub_area" attribute was less than 50. This approach ensures that our analysis is not influenced by rare and potentially biased data points.

CHALLENGES FACED AND ADDRESSED

- Our challenge was to decide how to appropriately treat these rare occurrences to ensure the robustness of our analysis. To address this challenge, we inferred that if sub_area with very low occurrences were to be included in our analyses, the associated records might not provide a representative sample, which could lead to biased or inaccurate results. To minimize this bias, we decided to include only records with a minimum of 50 occurrences. This ensures that we

retain a more balanced and representative dataset for our analysis.

6. HANDLING INCORRECT DATA

In this step, we focused on identifying and rectifying incorrect data present in the dataset. When calculating summary statistics for the "build_year" variable, we found that certain values were not allowed, such as 0, and that the maximum value was unrealistically high, suggesting that there may have been a data entry error. To correct this issue, we removed records with "build_year" values below 1500 or above 2023. We recognized that certain numeric attributes (e.g., "price_doc," "full_sq," "life_sq," "floor," "num_room," and "kitch_sq") should not have a value of zero, and we eliminated records with zero values for these attributes.

CHALLENGES FACED AND ADDRESSED

The challenge we encountered is to identify values that were clearly incorrect or outside the allowable range, as incorrect data can seriously affect the validity of our conclusions. By correcting unrealistic or unallowable values in the variables, as well as eliminating some zero values that should not exist, we effectively improved the accuracy of our dataset.

7. HANDLING CATEGORICAL DATA

- Encoding the categorical data involves converting categorical attributes into a numerical format, specifically the "product_type" and "sub_area" variables. These attributes contain non-numeric values, which need to be converted into a numerical format.

CHALLENGES FACED AND ADDRESSED

- "product_type" only has 2 categorical values, we can simply replace them with 1 and 2, the challenge is "sub_area". We processed the unique values within the "sub_area" attribute to create a consistent and meaningful encoding. By iterating through each unique value, we assigned a numerical code (incrementing by 1) to represent each "sub_area." This encoding scheme maintains the categorical distinction while enabling us to work with numerical representations.

8. HANDLING OUTLIERS

- We created box plots for each relevant numerical attribute to visually identify potential outliers. Based on the Interquartile Range (IQR) method (Gulati,2022), identifies and returns a list of outliers and handles outliers by removing "build_year" records with outlier values, as outliers in this context could potentially represent data entry errors. For other numerical attributes, we replaced outlier values with the median of the respective attribute. This approach minimizes the impact of outliers on analysis results.

CHALLENGES FACED AND ADDRESSED

- Outliers can arise for a variety of reasons, such as data entry errors or truly extreme observations. There are different ways to deal with different causes. We learned the IQR method to determine which numbers are outliers and what to do about them.

9. FEATURE SCALING

- We implemented feature scaling on the numerical attributes of the housing dataset.

CHALLENGES FACED AND ADDRESSED

- The challenge was to ensure that numerical attributes are on a common scale to prevent certain attributes from dominating others due to their inherent differences in magnitude. To address the challenge, we created a StandardScaler object, which will be used to standardize the numerical attributes and applied the standardization process to all numerical attributes.

EXPLORATORY DATA ANALYSIS & VALIDATION SPLIT

- After cleaning the data, we use a plot histogram to identify the distribution of values and prepare the data for regression analysis by defining the predictor and outcome variables, and then split the data into a training set and a validation set.

MACHINE LEARNING MODEL DEVELOPMENT

- The solution to the business problem involves the prediction of a numerical outcome variable (housing price) based on a set of housing predictor variables. The following machine learning algorithms are being deployed on the training data and their efficiencies are evaluated.
  - Linear Regression
  - K Nearest Neighbours

a. LINEAR REGRESSION

- **SUITABILITY OF MODEL**: The aim is to predict a continuous numerical variable based on values of a list of categorical and numerical variables hence linear regression seems to be an ideal model for the solution.
- Assuming a linear relationship between the outcome variables and the predictor variables, the linear regression model is built using the training dataset and the model performance is evaluated against the validation dataset.
- Extreme values in the correlation matrix of predictor variables clearly indicate that there is multicollinearity among predictor variables and suitable predictor reduction techniques must be used to optimize the regression model.
- Forward selection, Backward selection and Stepwise selection techniques are used to obtain the best list of predictor variables and thus prevent the redundancy associated with multicollinearity.
- The regression models are built and evaluated using the different sets of variable lists obtained from different selection techniques and the one with minimum value of root mean square error and mean average error is selected.
- The values of variables and coefficient of the regression model are as follows:

```
The intercept of regression model: 2161738.046263988
The coefficients of regression model:
               Predictor    coefficient
0               full_sq   69985.888890
1                 floor   45683.974609
2              num_room  350518.729327
3              kitch_sq   44543.727310
4              sub_area   -1536.334934
5   public_healthcare_km  -73830.439466
6         university_km   23128.228666
7    shopping_centers_km  -95870.750365
8      church_count_3000   20596.380777
9     leisure_count_3000   52530.454909
10      sport_count_3000   28740.171902
11    nuclear_reactor_km  -63364.270065
12          radiation_km  129637.406574
13             fitness_km -123183.116362
14        metro_min_walk  -11611.942088
15  bus_terminal_avto_km  -19864.630825
```

MODEL EVALUATION

- The model evaluation results for the linear regression model on the validation set are as follows:

Root Mean Squared Error (RMS):	2444548.762
Mean Absolute Error (MAE):	1748620.08

- As an attempt to improve the performance of our regression model, the predictor variables are feature scaled to a standard scale and regression model performance is recalculated. However, it has been observed that feature scaling does not significantly improve the model performance results.

STRENGTHS

- Less complexity and easier to implement.
- Coefficients and statistical results associated with regression are easy to interpret.

LIMITATIONS

- It assumes that the attributes are independent and there is no degree of correlation between them.
- Sensitive to outliers and prone to underfitting.

b. K NEAREST NEIGHBOURS

- **SUITABILITY OF MODEL**: Since the KNN regressor predicts continuous outcome variables using feature similarity, there is a good chance that model results will have high levels of accuracy.
- The dataset is split into validation and test data then predictor variables are feature scaled for better model performance.
- A KNN regressor model is built and based on the feature similarity of the nearest neighbours the outcome variable is predicted for validation data.

MODEL EVALUATION

- The model evaluation results for k nearest neighbours regressor validation set are as follows:

Root Mean Squared Error (RMS):	2052690.469
Mean Absolute Error (MAE):	1429034.598

STRENGTHS

- Ideal for non-linear data.
- Easy to interpret and understand.

LIMITATIONS

- Multiple iterations are required to determine the ideal k-value.
- Sensitive to irrelevant features.

MODEL COMPARISON

- Based on the RMSE and MAE values of linear regression and the KNN regressor model, we can arrive at the conclusion that the KNN regressor model performs better than the linear regression model.

SCOPE

The proposed solution currently solves the ambiguity associated with housing prices. The solution can be enhanced, and K-means clustering can be performed to provide extended services like providing a recommendation for housing properties for a particular price and input data requirements which will enhance the customer experience.

REFERENCE

1. Kumar, Ajitesh. “Python - Replace Missing Values with Mean, Median & Mode.” Data Analytics, 26 Mar. 2023, https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/#:~:text=Mean%20imputation%20

is%20often%20used,to%20outliers%20than%20the%20mean

2. Cloud, Saturn. “How to Remove Entries Based on the Number of Occurrences Using Python Pandas | Saturn Cloud Blog.” Saturncloud.io, 19 June 2023, https://saturncloud.io/blog/how-to-remove-entries-based-on-the-number-of-occurrences-using-python-pandas/

3. Gulati, Aman Preet. “Dealing with Outliers Using the IQR Method.” *Analytics Vidhya*, 13 Sept. 2022, www.analyticsvidhya.com/blog/2022/09/dealing-with-outliers-using-the-iqr-method/

4. GeeksforGeeks. (2023). ML Advantages and Disadvantages of linear regression. GeeksforGeeks. https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/

5. Ellis, C. (2022, December 16). When to use linear regression. Crunching the Data. https://crunchingthedata.com/when-to-use-linear-regression/

6. Singh, A. (2023). KNN algorithm: Introduction to K-Nearest Neighbors Algorithm for Regression. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

7. Joby, A. (n.d.). K Nearest Neighbor (KNN): the most used ML algorithm. https://learn.g2.com/k-nearest-neighbor