**Inference Document for SML -CT3**

**1.Read the dataset – Using Pandas**

### 2. Data Understanding

**a. What are the number of rows; no. & types of variables (continuous, categorical etc.)**

**b. Calculate five point summary for numerical variables**

**c. Summarize observations for categorical variables – no. of categories, % observations in each category**

**1. Dataset Overview**

**1.1 Number of Rows and Columns**

* The dataset contains [number of rows] rows and [number of columns] columns.

**1.2 Variable Types**

* There are [number of numerical variables] numerical variables and [number of categorical variables] categorical variables in the dataset.
* **Numerical Variables**: [List of numerical variables]
* **Categorical Variables**: [List of categorical variables]

**2. Five-Point Summary for Numerical Variables**

* The five-number summary provides statistical insights into the distribution of numerical variables.

1. **Count: [Count]**
2. **Mean: [Mean]**
3. **Standard Deviation: [Standard Deviation]**
4. **Minimum: [Minimum]**
5. **25th Percentile: [25th Percentile]**
6. **Median (50th Percentile): [Median]**
7. **75th Percentile: [75th Percentile]**
8. **Maximum: [Maximum]**

**3. Summary of Categorical Variables**

* For each categorical variable, we present the number of categories and the percentage of observations in each category.

**3. Data Visualizations**

**Distribution of 'price'**

* The histogram with KDE shows that the majority of prices are concentrated within a certain range.

**Geographical Distribution of Listings**

* The scatter plot indicates the geographical distribution of listings based on latitude and longitude, with different colours representing neighbourhood groups.

**Count of Listings in Each Neighbourhood Group**

* The count plot illustrates the distribution of listings across different neighbourhood groups.

**Room Type Distribution**

* The pie chart provides a clear breakdown of the distribution of different room types.

**KDE Plot for Numerical Variables**

* Kernel Density Estimate (KDE) plots for numerical variables visualize their underlying distributions.

### 3. Data Preparation

#### **Check for defects in the data. Perform necessary actions to ‘fix’ these defects (5 Marks)**

**a. Do variables have missing/null values?**

**b. Do variables have outliers?**

**c. Is the data normally distributed? Is it a defect? Why or why not?**

**1. Dropping Unwanted Columns**

We dropped the following columns from the dataset:

* 'id'
* 'name'
* 'host\_id'
* 'host\_name'

**2. Data Type Analysis**

**2.1 Numerical Variables**

* The dataset includes the following numerical variables:
  + [List of numerical variables]

**2.2 Categorical Variables**

* After the removal of unwanted columns, the categorical variables are:
  + [List of categorical variables]

**3. Categorical Analysis and Encoding**

* For each categorical variable, we examined unique values and their counts.
* Columns with more than 10 unique values were dropped.
* Categorical columns were encoded using Label Encoding.

**4. Box Plot of Numerical Variables**

* A box plot was created for the remaining numerical variables to visualize their distributions.

**5. Skewness Transformation**

* Skewness of numerical variables was analyzed.
* Variables with skewness greater than 0.5 underwent a log transformation to make the data more normally distributed.

**6. Handling Empty Columns**

* Columns with missing values were dropped.

**7. Updated Box Plot of Normalized Data**

* The box plot was updated after skewness transformation and removal of unwanted columns.

**8. KDE Plot for Skewed Data**

* Kernel Density Estimate (KDE) plots were generated for the skewed numerical variables.

**4.Summarize relationships among variables (5 marks)**

**a. Plot correlation plots. Which are the variables most correlated with Target? Which independent variables are correlated among themselves? Do you want to exclude some variables from the model based on this analysis? What other actions will you take?**

**1. Correlation Heatmap for Non-Skewed Data**

A correlation heatmap was generated for the numerical variables in the non-skewed dataset (**df2**). Key observations include:

* **Positive Correlation:**
  + [Positive correlations between specific pairs of numerical variables.]
* **Negative Correlation:**
  + [Negative correlations between specific pairs of numerical variables.]
* **Strength of Correlation:**
  + Correlation coefficients range from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation.

**2. Correlation Heatmap for Skewed Data**

A correlation heatmap was created for the numerical variables in the skewed dataset (**df3**). Notable findings include:

* **Positive Correlation:**
  + [Positive correlations between specific pairs of numerical variables.]
* **Negative Correlation:**
  + [Negative correlations between specific pairs of numerical variables.]
* **Strength of Correlation:**
  + Similar to the non-skewed data, correlation coefficients range from -1 to 1.

**3. Heatmap Comparison of Skewed and Non-Skewed Data**

To compare the correlation patterns between skewed and non-skewed datasets, a combined correlation heatmap was generated:

* **Observations:**
  + Differences in correlation patterns between skewed and non-skewed datasets are evident.
  + [Specific observations regarding changes in correlation strengths or directions.]
* **Insights:**
  + [Insights into how data preprocessing, such as skewness transformation, has affected correlations.]
* **Overall Comparison:**
  + The heatmap provides a visual representation of how correlations differ between the two datasets.

**5. Fit a base model. Please write your key observations (5 marks)**

**a. Fit the Linear Regression Model**

**b. What is the overall R2? Please comment on whether it is good or not.**

**c. Which variables are significant?**

**d. Calculate MSE, RMSE, MAE, MAPE.**

**1. Data Preparation**

**1.1 Features and Target Variable**

* The dataset was split into features (**X**) and the target variable (**y**), with the target variable being 'price'.

**1.2 Scaling**

* Standard scaling was applied to normalize the features using **StandardScaler**.

**1.3 Train-Test Split**

* The dataset was split into training and testing sets with a test size of 20%.

**2. Linear Regression Model**

**2.1 Model Training**

* A linear regression model was trained on the scaled training data.

**2.2 Model Prediction**

* The model was used to predict the target variable on the test set.

**2.3 Model Evaluation - R-squared Value**

* The R-squared value, a measure of how well the model explains the variability of the target variable, was calculated.
* R-squared Value (R2) for the model with Standard Scaler: [R-squared value].

**2.4 Variable Significance**

* Variable significance was checked using the stats models library to understand the impact of each variable on the target variable.

**3. Evaluation Metrics**

**3.1 Mean Squared Error (MSE)**

* MSE measures the average squared difference between the predicted and actual values.

**3.2 Root Mean Squared Error (RMSE)**

* RMSE is the square root of MSE, providing a measure of the average magnitude of the residuals.

**3.3 Mean Absolute Error (MAE)**

* MAE represents the average absolute difference between the predicted and actual values.

**3.4 Mean Absolute Percentage Error (MAPE)**

* MAPE calculates the percentage difference between predicted and actual values on average.

**4. Results**

* **R-squared Value (R2):**
  + The R2 value indicates [interpretation of the R-squared value].
* **Variable Significance:**
  + [Key findings from variable significance analysis].
* **Evaluation Metrics:**
  + **MSE**: [Mean Squared Error value]
  + **RMSE**: [Root Mean Squared Error value]
  + **MAE**: [Mean Absolute Error value]
  + **MAPE**: [Mean Absolute Percentage Error value]

**6. Perform feature engineering using any of the listed techniques: Forward feature selection, backward elimination and recursive feature elimination and use the features to improvise the model and compare the results with the base model. (5 marks)**

**1. Feature Selection using Recursive Feature Elimination (RFE)**

**1.1 Data Split**

* The dataset was split into training and testing sets using an 80-20 split.

**1.2 Base Model - Linear Regression**

* A base linear regression model was trained on the entire set of features.
* Predictions were made on the test set, and the Root Mean Squared Error (RMSE) was calculated as the baseline performance metric.

**1.3 RFE Model**

* Recursive Feature Elimination (RFE) was applied to select a specific number of features (in this case, 3).
* Selected features were identified based on the RFE model.

**1.4 Model Training with Selected Features**

* A new linear regression model was trained using only the selected features from the RFE process.
* Predictions were made on the test set, and the RMSE was calculated for the RFE model.

**2. Results**

**2.1 Base Model Performance**

* **Base Model RMSE:** [Base Model RMSE value]

**2.2 RFE Model Performance**

* **Selected Features:**
  + [List of selected features]
* **RFE Model RMSE:** [RFE Model RMSE value]

**2.3 Model Comparison**

* The RMSE values were compared between the base model and the RFE model.

**7. The prediction model output reliability can be improved using regularization techniques and parameter tuning. Perform regularization techniques and compare the results with the base model. (5 marks)**

**1. Ridge Regression Model**

**1.1 Data Split**

* The dataset was split into training and testing sets using an 80-20 split.

**1.2 Standardization**

* Standard scaling was applied to the features using **StandardScaler** to ensure that all variables were on the same scale.

**1.3 Base Model - Linear Regression**

* A base linear regression model was trained on the scaled training data.
* Predictions were made on the test set, and the Root Mean Squared Error (RMSE) was calculated as the baseline performance metric.

**1.4 Ridge Regression Model**

* Ridge Regression, a form of regularized linear regression (L2 regularization), was applied to the scaled data.
* The alpha parameter, controlling the strength of regularization, was set to [chosen alpha value].
* Predictions were made on the test set using the Ridge model, and the RMSE was calculated.

**2. Results**

**2.1 Base Model Performance**

* **Base Model RMSE:** [Base Model RMSE value]

**2.2 Ridge Model Performance**

* **Ridge Model RMSE:** [Ridge Model RMSE value]
* **Alpha Value:** [Chosen alpha value]

**2.3 Model Comparison**

* The RMSE values were compared between the base linear regression model and the Ridge regression model.

**3. Conclusion**

* Ridge Regression, with L2 regularization, was applied to the dataset, offering a way to mitigate overfitting.
* The performance of the Ridge model was evaluated and compared to the base linear regression model.
* Insights gained from this analysis can guide the selection of appropriate regularization techniques for improving model generalization.

**Output :**