**Data:** The dataset utilized in this assignment is named "Housing\_Iowa," which can be accessed on Kaggle through the following link: Housing\_Iowa Dataset (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview). This dataset encompasses a total of 81 variables and includes 1,460 observations. Within the set of variables, there is a numerical target variable denoted as "SalePrice," an identification variable labeled as "ID," and seventy-nine predictor variables.

The target variable "SalePrice" exhibits skewness; nevertheless, this can be a significant issue when we try to build a multiple regression model. Furthermore, typical data issues, including missing values across several predictors, skewness in numerous numerical predictors, and the potential presence of outliers, need to concern us as well since data challenges are likely to significantly affect the performance of the fitted models when utilizing multiple regression algorithms. It worth to know that the missing value symbol is “NA” in this data set.

**Assignment Goal:** The objective of this concluding examination is to assess your comprehension of data preparation challenges in constructing a meaningful and effective regression model using LASSO through appropriate data preparation to fellow these steps.

**Step 1: Get Data (20 Points)**

Incorporating data into the software system utilized to construct this model can be achieved through various methods.

**Note: Put your programming coding in Appendix 1 to demonstrate how to obtain the data in the system.**

**Step 2: Data Exploration (50 Points)**

Data exploration is a critical step in the model-building process. For this assignment, we like you to produce the following tables.

1. Table 1: This table includes “Variable Name”, “Missing Count”, “Missing Percentage”, “Skewness”, “Mean”, “Median”, “Minimum”, and “Maximum” for all numerical predictors.
2. Table 2: This table includes “Variable Name”, “Cardinality”, and “Missing Percentage” for all character predictors.

Please incorporate a summary table for each component of STEP #2 within this response. Additionally, include the corresponding code in Appendix 2.

**Step 3: Data Preparation (100 Points)**

1. Create one missing value indicator for each predictor
2. Impute all numerical predictors with missing value using simple mean imputation
3. Perform adequate transformation for all skewed numerical predictors after imputation and produce a table that has the following columns: “Variable Name”, “Transformation Power”, “Skewness Before Transformation”, and “Skewness after Transformation”
4. Prepare all character predictors with cardinality higher than 6 using smoothing mean discussed in class
5. Skip outlier detection in this assignment

Please incorporate a summary table for each component of STEP #3 within this response. Additionally, include the corresponding code in Appendix 3.

**Step 4: Data Partition (30 Points)**

Considering the relatively modest size of this dataset, comprising 1,460 cases, we opt for a five-fold cross-validation approach for the purposes of this assignment. The complete dataset, denoted as B, is randomly partitioned into five distinct subsets:

**Please put your code for this part in Appendix 4.**

**Step 5: Five Multiple Regression Model Fitting (40 Points)**

Using the data partition outlined in Step 4, build five models by using four parts for training and reserving one part for testing. This process will allow for the computation of five distinct performance metrics, as detailed in Table 1.

|  |  |  |
| --- | --- | --- |
| **Table 1** | | |
| **Model** | **Training Data** | **Testing Data** |
| **I** |  |  |
| **II** |  |  |
| **II** |  |  |
| **IV** |  |  |
| **V** |  |  |

**Report the following information all five models built by completing Table 2 below.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 2** | | | |
| **Model** | **Number of Terms** | **Training MSE** | **Testing ASE** |
| **I** |  |  |  |
| **II** |  |  |  |
| **III** |  |  |  |
| **IV** |  |  |  |
| **V** |  |  |  |

**Please put your code for this part in Appendix 5.**

**Step 6: Ensemble all five models and calculate the MSE of this ensembled model (20 Points)**

The ultimate model is an ensemble formed by combining these five models, and the performance of this ensemble model can be estimated by calculate this ensembled model’s MSE.

**Step 7: Understanding the Results (40 Points)**

Write a report to report all your findings.