Regression Activity

For AMES Dataset, perform the following

1. Data Understanding

1a. The dataset has 81 columns to work with which is huge. Hence our task is to eliminate variables which are unimportant. Looking at data description – list columns which are not really relevant for the business.

1b. What statistical evidence can be provided to support elimination of above variables. For example – is it correlation or scatter plot / bivariate analysis. Your evidence can be any insight, data based, visual, statistical test. For the columns which you listed in 1a, test them statistically and comment if your business understanding in 1a is validated by statistical test / visuals.

1c. Take a call and eliminate the above variables before solving 2,3.

2. Data Scaling and Pre-Processing

2a. Find if any of the columns have missing values, outliers, irrelevant data, duplicate rows etc. Based on your understanding, report the issue and treat them

2b. For numeric variables, figure out features which have to be scaled. Report any features which need not be scaled. Decide which scaling technique would you use. Add hypothesis why it's chosen over other scaler.

2c. How do you scale the variable PoolArea ? What is so special with Poolarea when compared to other numeric variables. Comment on effect of scaling Pool Area.

2d. Which categorical variables need encoding? Can we minimise the number of new columns created after one hot encoding. Come up with smart techniques on how to one hot encode with much lesser columns as output and implement them.

3. Data Modelling

3a. Which metric will you choose to optimise in the case of regression in this case.

3b. Implement Statsmodel OLS method and drop the number of variables which are insignificant. Write a code to automatically check and drop below certain insignificance.

3c. Report VIF scores for all significant variables and drop high VIF Values.

3d. Draw residual versus target column and analyse residual assumptions. Are any of them violated? If yes, pick up any transformations and run OLS again. Are the results any better?

3e. Implement gradient based regressor and compare results with OLS stats model.

3f. Run Lasso regression with all variables from original data. Keep varying the regularisation coefficient and drop variables. Compare the features dropped by lasso regression to features dropped via insignificant variables, VIF, Business understanding. Is lasso dropping same set of features or different set ?

Dataset available here :
https://drive.google.com/open?id=1sHVrbKfOjKHeV6OgfXOBCJUEVnNeacrK&authuser=prudhvi%40supervisedlearning.com&usp=drive_fs