Project Write Up

Step1: Loading and exploring data

1. Importing library pandas and NumPy for analysis.
2. Load Train and Test data set.
3. Describing the data to see different statistical parameters.

Stpe2: Cleaning the data

1. Drop duplicate rows from the train data set.
2. Check if any column in train data has got null values.
3. Check the number of unique values in each column in train data.
4. Check the index having zero variance in the train set.
5. Drop columns having zero variance. Drop the same columns from the test dataset.

Step3: Split the train data into dependent variables and target variables.

Step4: Encoding categorical columns

1. Concat train (having only dependent variables) and test data sets.
2. Import Label Encoder modules from sklearn for performing label encoding.
3. Encode categorical data to numerical.
4. Split back train (having only dependent variables) and test data sets.

Step5: Perform Scaling and Dimensionality reduction on train data set

1. Import Minmax Scaler and PCA modules.
2. Scale the train dataset (having only dependent variables).
3. Perform dimensionality reduction on the train dataset (having only dependent variables).

Step6: Build Model

1. Import XGBRegression from xgboost library.
2. Import train\_test\_split module from sklearn.model\_selection.
3. Split the data in train and test.
4. Create an instance of XGBRegressor with learning rate of 0.05
5. Fit the model and predict the test output.
6. Check the accuracy using R2 score.

Step7: Predict values for test data set.

1. Scale the test data set using the same scaler instance used for the train data set.
2. Perform PCA on test data using same transformation as train data set
3. Predict output for the test data set.

**Conclusion:** Using Xgboost algorithm we got an R2 score of 0.55.