Feature Selection and Regression

Ipshita Ghosh

20122006

Machine Learning CAC

Index

1. Introduction
2. Data
    a. Data Assumptions
3. Methodology
    a. Missing Values
    b. Scaling and Feature Selections
    c. Model Building
4. Analysis
5. Run Book
6. Reference

Introduction

The task given was to predict the target variable, y, with minimum MAE. The dataset consisted of 100,000 rows with 304 parameters. Three different scaling types were used, namely, MinMax Scaler, Robust Scaler and Standard Scaler. Preprocessing of the data included removing the null values and imputing them accordingly. The approaches for eliminating parameters were PCA, VIF, Collinearity and Kbest features. From a total of 18 models, the scaling with MinMax Scaler and Principal component analysis to get the best features shows minimum mean absolute error, which was 0.065.

Data

The data had a total of 100,000 rows with 304 parameters and one target variable. On looking at the statistical summary, it was seen that there were some substantial values and some small. The dataset also had nulls in it, which varied from column to column.

Data Assumptions

The assumptions taken in data were:

1. Linearity between the input parameters and the target variable, many of the input variables were dropped during the study.
2. The error term e_i is normally distributed with a mean of 0
3. The error terms show no patterns and are independent of each other.

To check for multicollinearity, few statistical tests were done.

Methodology

a. Frameworks used:

The frameworks used for the task as given along with the task they were imported for :-

```
For data wrangling
numpy
pandas

For model buildng
train_test_split
LinearRegression
glm
xgboost

For feature selection
PCA
variance_inflation_factor
SelectKBest
f_classif

For scaling of Dataframe
RobustScaler
StandardScaler
MinMaxScaler

For model evaluation
r2_score
mean_absolute_error

For making graph
matplotlib.pyplot

For statistical tools and steps
statsmodels.api
stats
kendalltau

Extra
```

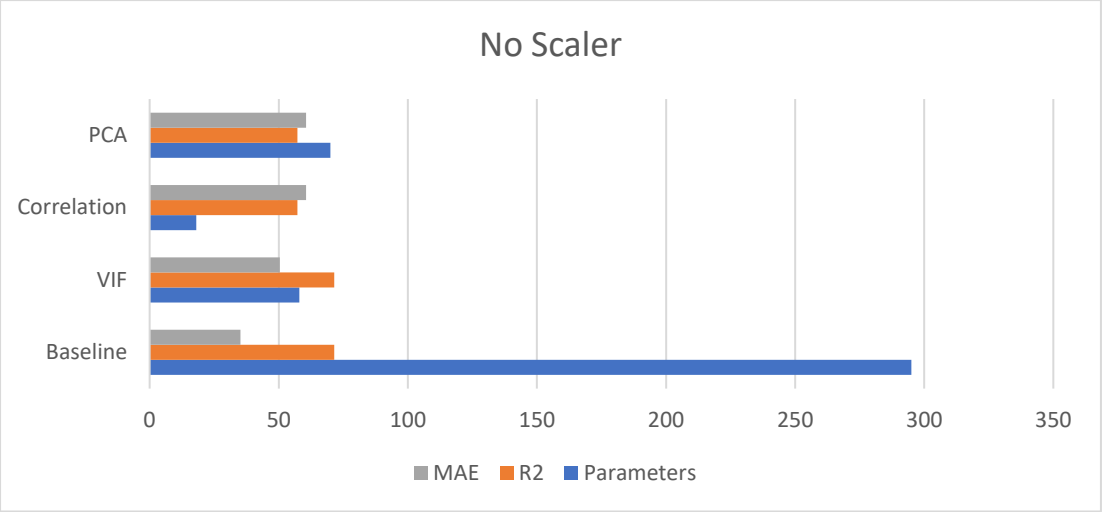
```
warnings
```

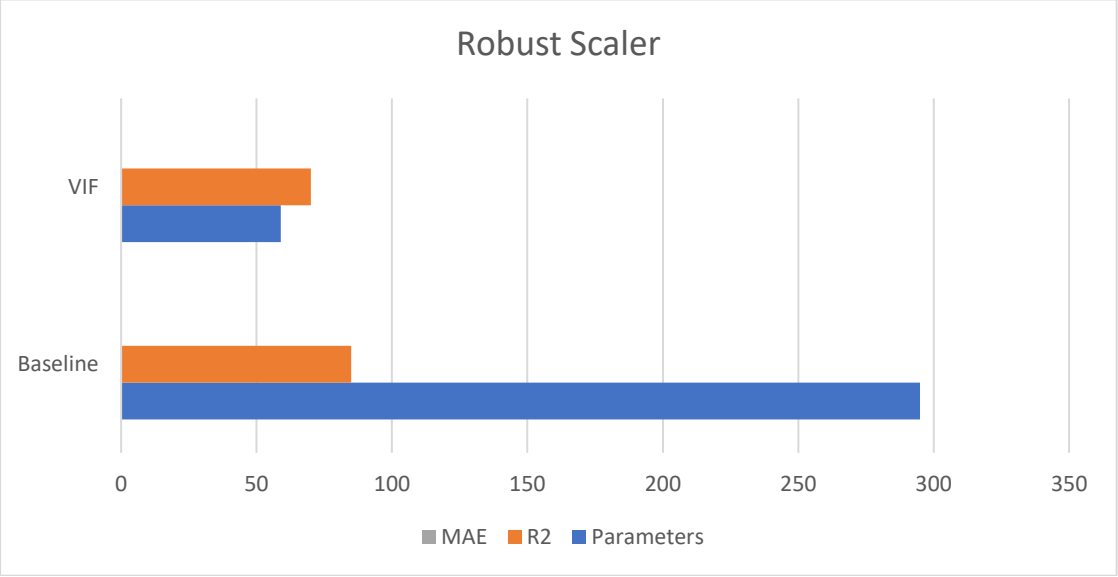
b. Missing Values

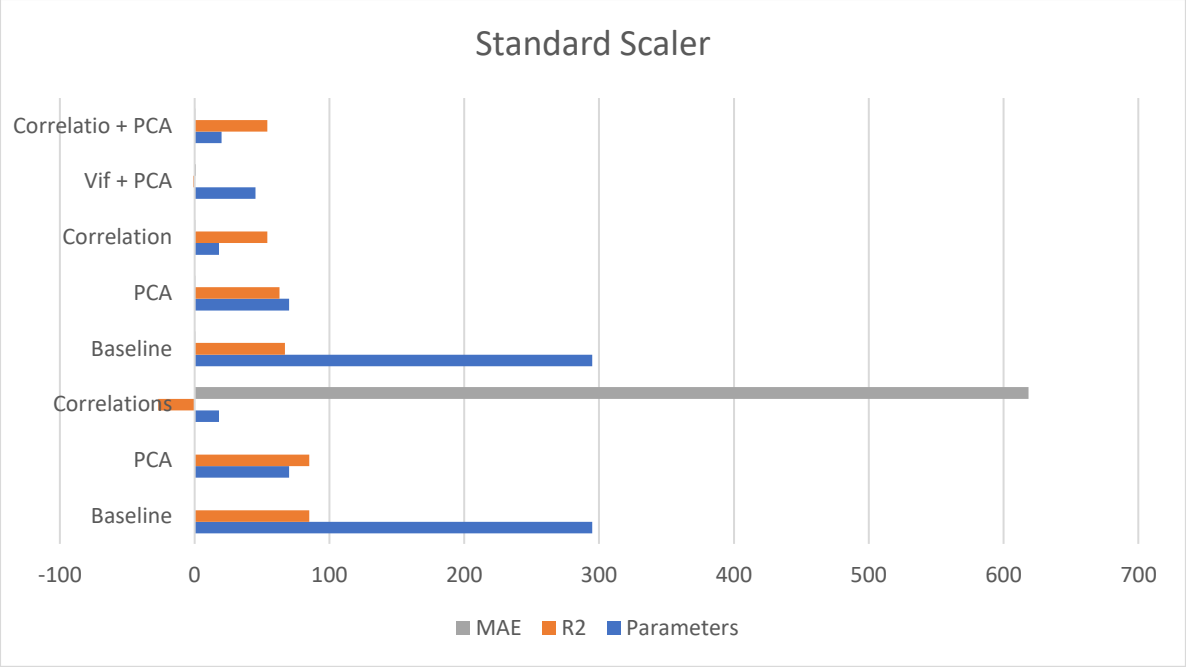
First, the percentage of missing values were looked and the columns having 75% or more of their data as null were dropped. For the data having less than 75% null, the data means were used as imputation.

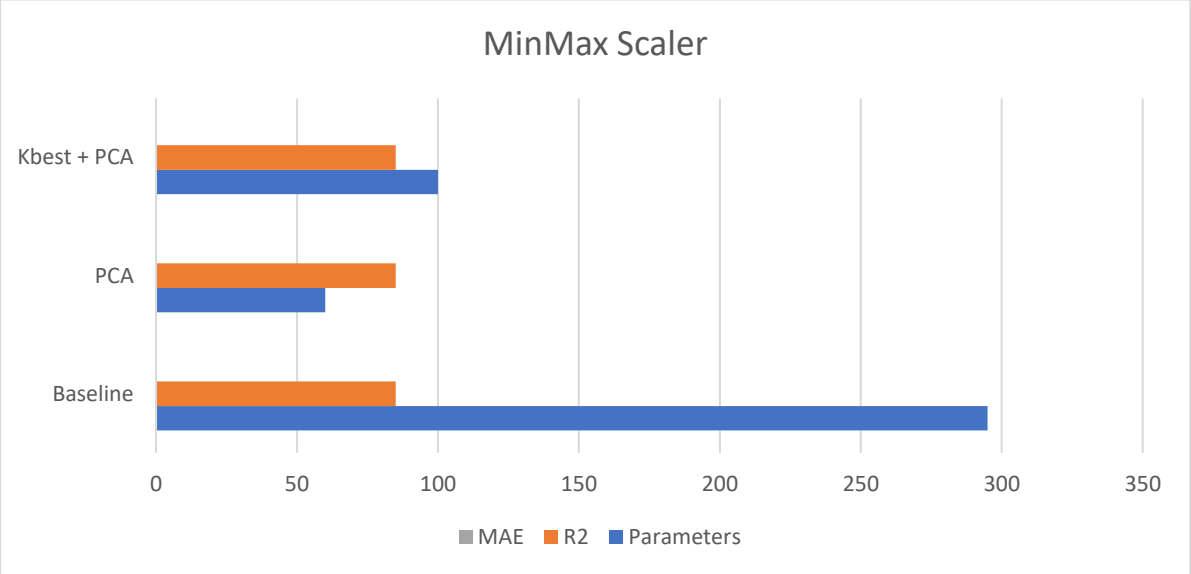
c. Scaling Techniques

The models were built first without any scaler, then with Robust Scaler, Standard Scaler and then MinMax Scaler.



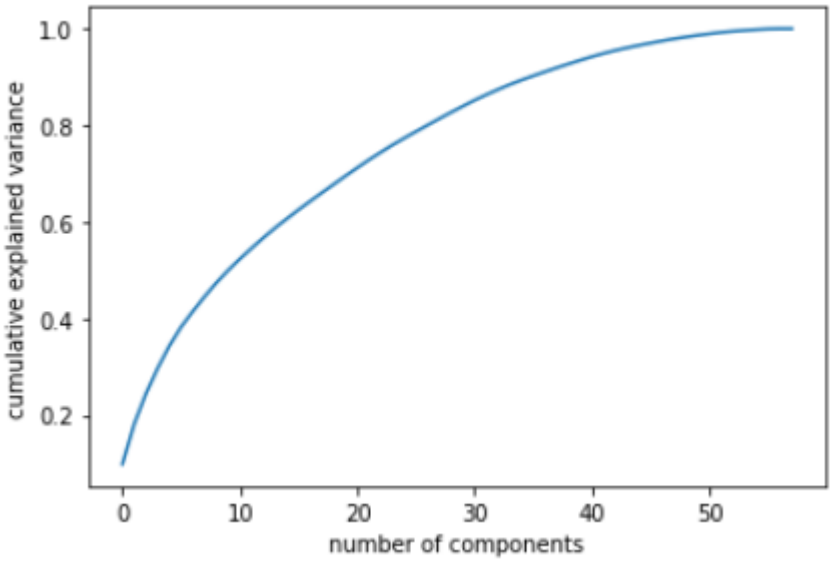






The lowest MAE recorded was 0.065, with MinMax Scaler. The most reliable model was using Minimax scaler and PCA as feature selector.

For selecting n components, the explained variance graph was used in all the cases. Several components showing 95% of explained variance was selected.



Model Building

Most of the models were built using Linear Regression, with different approaches for reducing the dimensions and Mean Absolute Error, as can be seen in the table of Analysis section.
For some models xgBoost was used, but it did not show any such improvement in decreasing the error. The lowest error achieved by xgBoost was 0.479 on the Baseline model. The lowest error by Linear regression was 0.065, which is also used as the final model.

Analysis

Parameters	R2	MAE	Method
295	71.55	35.222	Baseline
58	71.55	50.367	VIF
18	57.20	60.639	Correlation
70	57.20	60.639	PCA
295	-1062966.00	618.44	PCA
295	85.00	0.18	Baseline
59	70.00	0.26	VIF
295	85.00	0.297	Baseline
70	85.00	0.297	PCA
18	-27.00	618.599	Correlations
295	67.00	0.479	Baseline
70	63.00	0.509	PCA
18	54.00	0.565	Correlation
45	-1.00	0.868	Vif + PCA
20	54.00	0.565	Correlation + PCA
295	85.00	0.065	Baseline
60	85.00	0.065	PCA
100	85.00	0.065	Kbest + PCA

Run Book

For the run book, the best model was chosen and along with it, there are two functions :-

1. MissingValue -> This function takes in the dataframe along with the threshold null value percentage and cleans the dataset

2. LinearModelBuilder -> This function runs the regression model with the n components chosen using the graph by the user.

Reference

1. Feature Selection For Machine Learning in Python (machinelearningmastery.com)
2. How to Write a Data Science Report | by Yashraj Nigam | Medium
3. StackOverflow
4. An Introduction to Statistical Learning, Chapter 3
5. Data Preparation for Machine Learning, Chapter 7,11,12,13,14
6. www.statology.com for VIF

Code:

1. MDS271-ML/ML_CAC.ipynb at main · ipshitag/MDS271-ML (github.com) Main
2. MDS271-ML/ML_CAC_Run_Book.ipynb at main · ipshitag/MDS271-ML (github.com) Run Book