Dataset Overview

- The dataset for our project is ***"Steel Industry Energy Consumption"***, gathered by the DAEWOO Steel Co. Ltd in South Korea.
- **Data Source:** https://archive.ics.uci.edu/dataset/851/steel+industry+energy+consumption
- **Response variable:** “Usage_kwh".
- **Predictors:** There are 10 features, including date, lagging/leading reactive power, lagging/leading power factor, CO_2,weekStatus,day of week.
- **Number of observations** : 35040
- The screenshot of first two samples of data is attached below.

```
> head(data,2)
              date Usage_kWh Lagging_Current_Reactive.Power_kVarh Leading_Current_Reactive_Power_kVarh CO2.tCO2.
1 01/01/2018 00:15      3.17                                 2.95                                    0         0
2 01/01/2018 00:30      4.00                                 4.46                                    0         0
  Lagging_Current_Power_Factor Leading_Current_Power_Factor  NSM WeekStatus Day_of_week  Load_Type
1                        73.21                          100  900    Weekday      Monday Light_Load
2                        66.77                          100 1800    Weekday      Monday Light_Load
```

Dataset Overview cont.

Table 2.1: Dataset feature description

Feature	Description
Date	Data collected in real time on the first of the month
Usage_kWh	Energy Consumption in Industry kWh continuous
Lagging Current	Reactive energy kVarh Continuous
Leading Current	Reactive energy kVarh Continuous
CO2	CO2 Continuous ppm
NSM	Minutes and seconds since midnight S Continuous
Week status	Weekday or Weekend
Day of week	Sunday, Monday ..etc
Load Type	Light Load, Medium Load, Maximum Load

Procedures/methods

1. Data Preprocessing
2. Data Visualization
3. Evaluation methods/criteria
4. Statistical Learning Methods

1. Data Preprocessing steps

1. **Null Values Handling**
   - Removed null values from the data
2. **Feature Extraction**
   - Extracted “Month” from “Date” feature
3. **Feature Removal**
   - Removed "Date" as similar features such as Month, weekStatus, and Day of week are present in the Dataset.
   - Removed "Load_Type" as it's similar to the response variable, "Usage_kWh."
4. **Categorical to Numerical Conversion**
   - Converted categorical features (weekStatus, Day of week, Month) to numerical representations.

4. **Train-Test splits**
   - Performed a 70-30 train-test split for model evaluation

2. Data Visualization

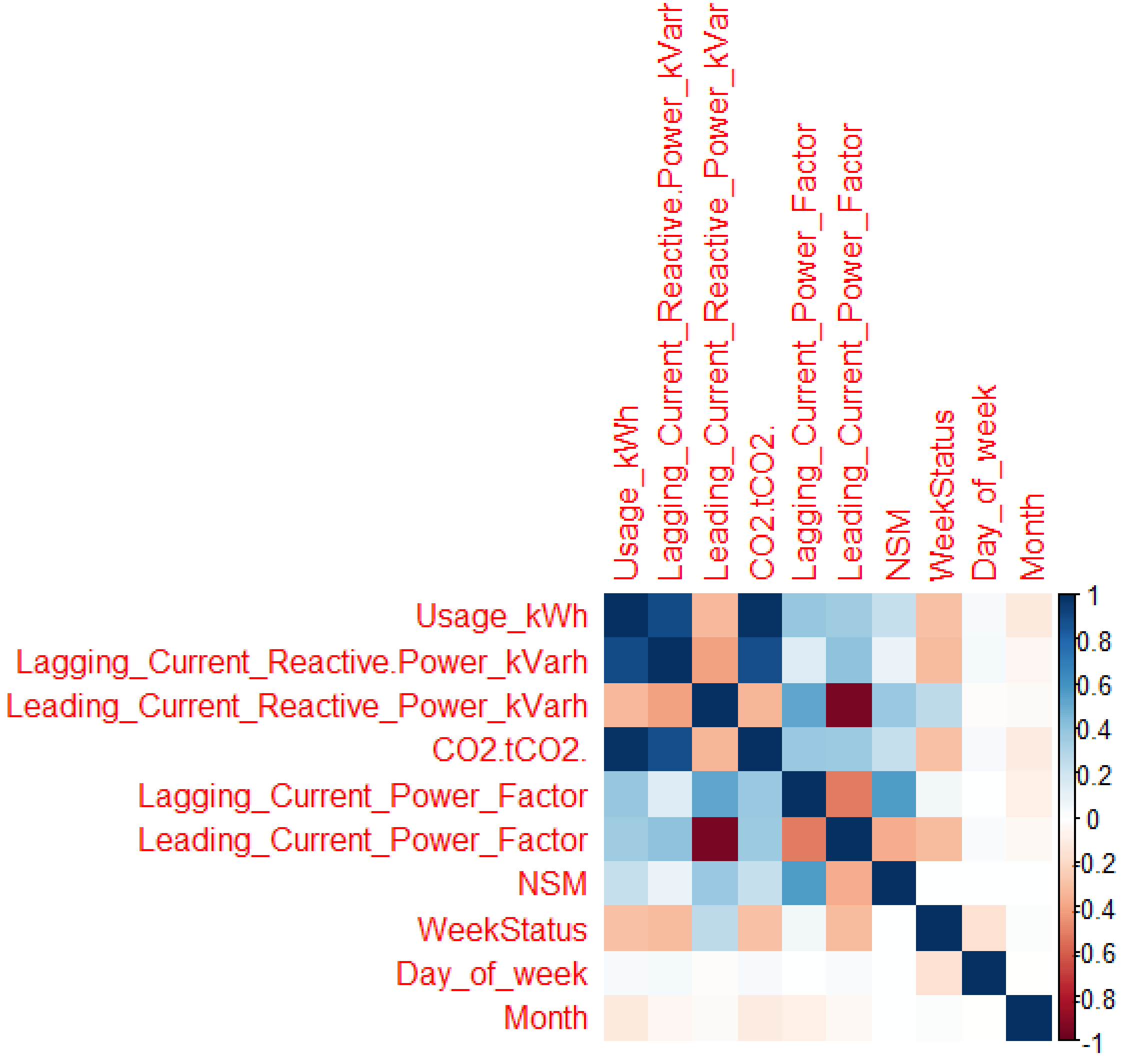


Fig. The correlation matrix

2. Data Visualization cont.

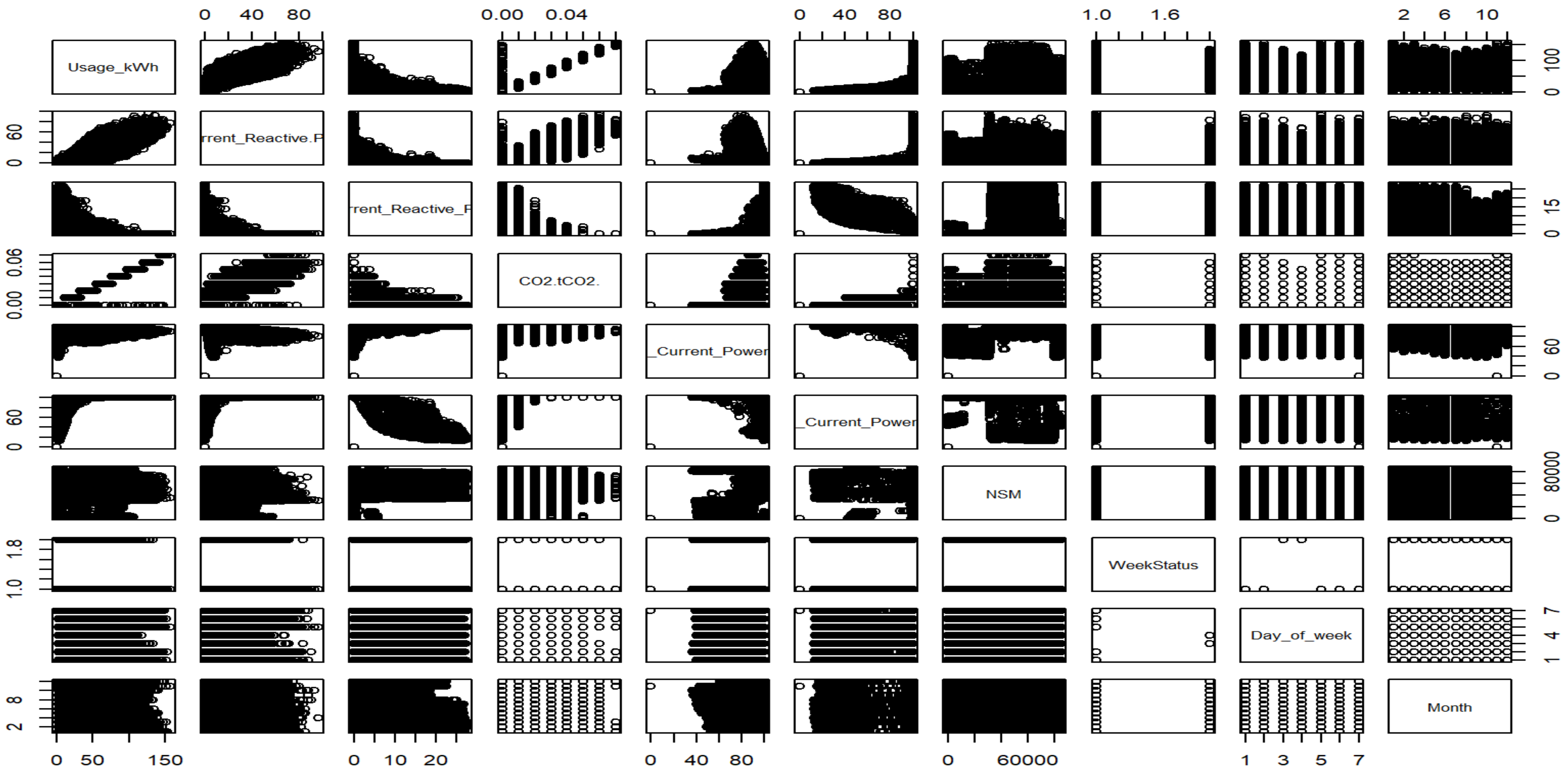
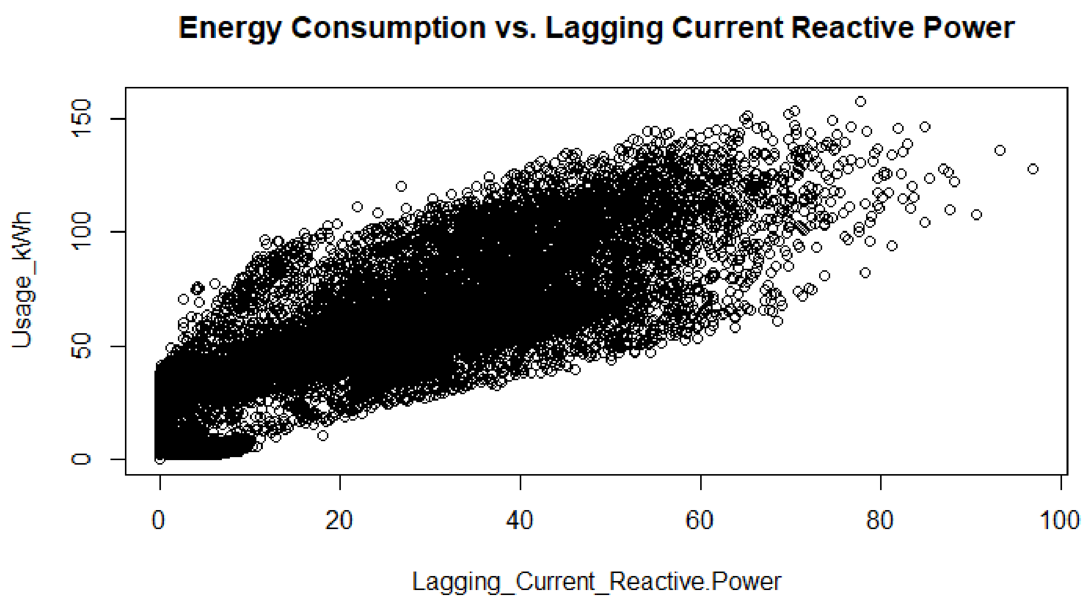
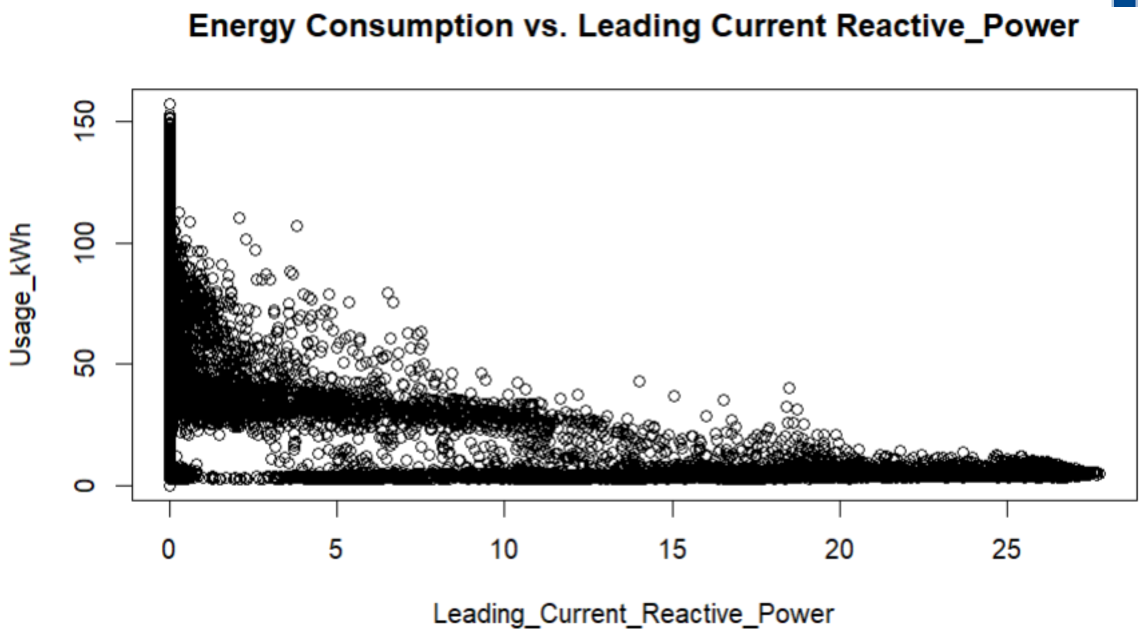
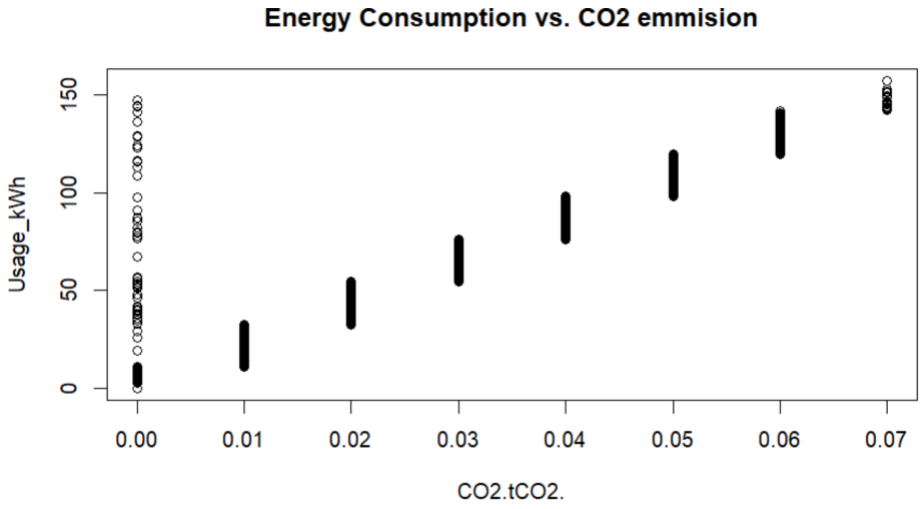


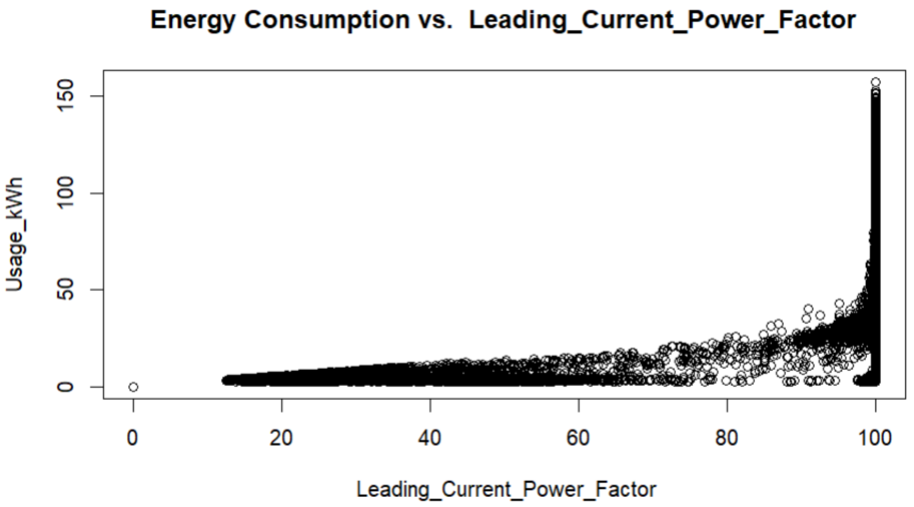
Fig. matrix of scatterplots for each pair of variables

2. Data Visualization cont.

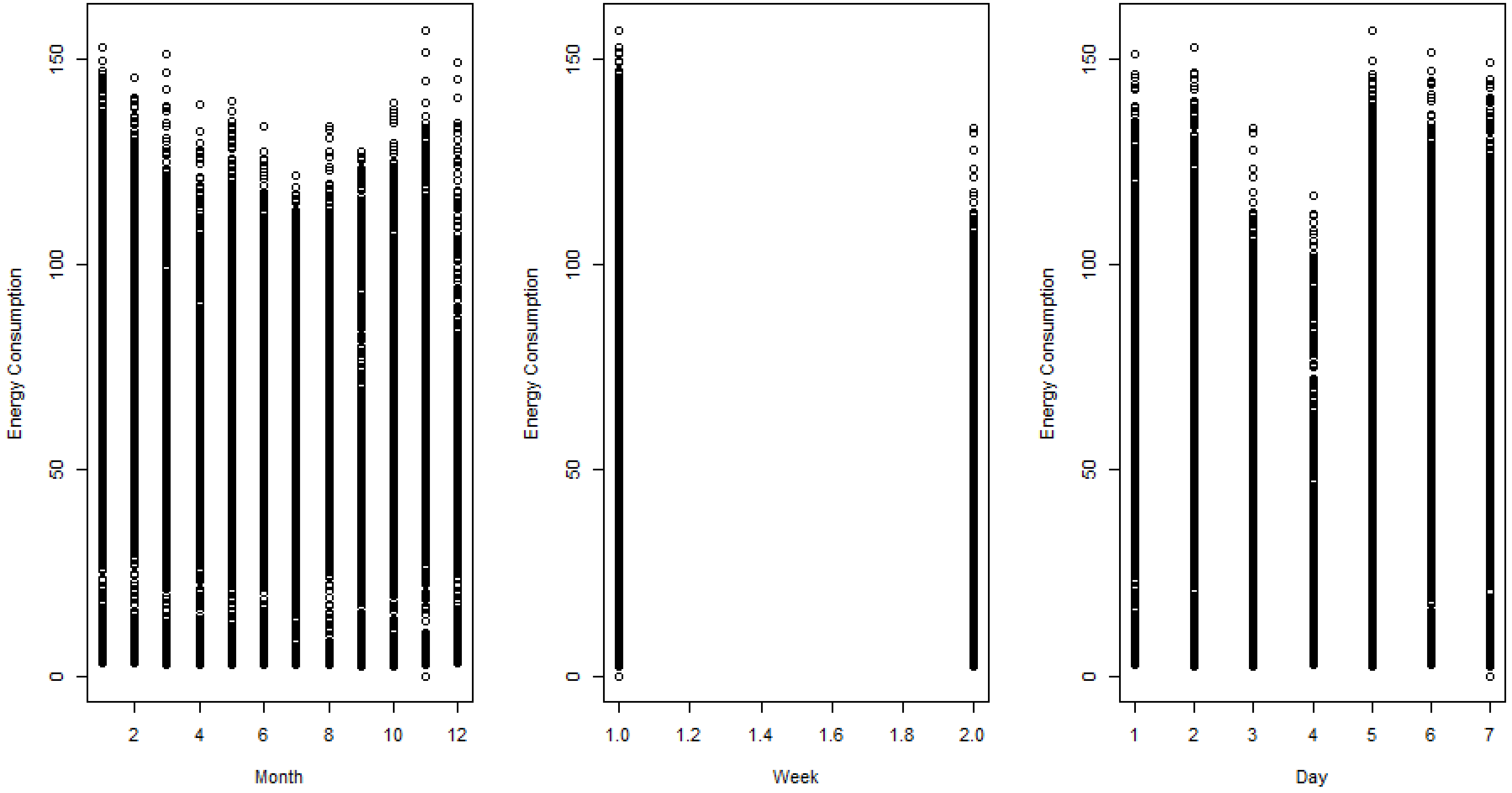






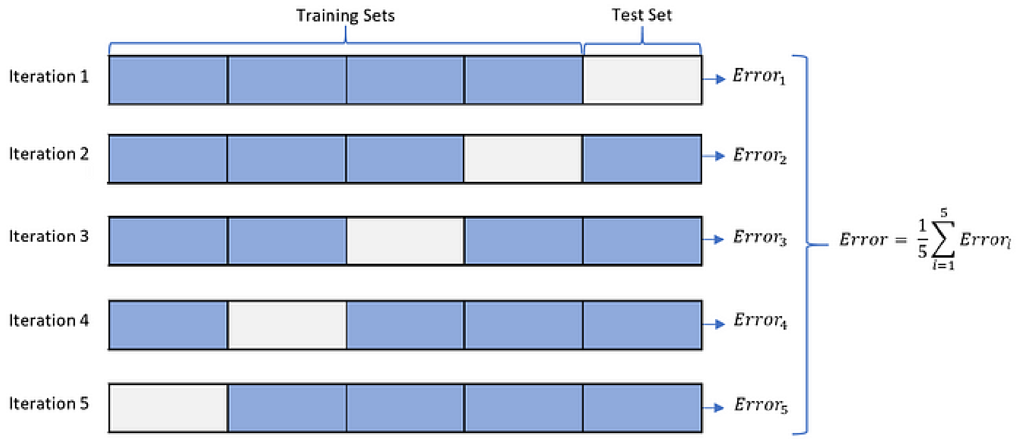


2. Data Visualization



3. Evaluation methods/criteria

1. K-folds Cross Validation



2. Test MSE

$\text{MSE} = (1/n) * \sum (yi - \hat{y}i)^2$, *where* yi is the actual observed values, and $\hat{y}i$ is the predicted values

3. Test R-squared

$R^2 = 1 - \frac{SSR}{SST}$,where SSR is the sum of the squared residuals, and SST is the total sum of squares.

4. Statistical Learning Methods

1. Linear Regression Model
2. Subset Selection
3. LASSO Regression Model
4. Ridge Regression Model
5. Principal Component Analysis (PCA)
6. Random Forest Model
7. Random Forest Model with CV (Proposed Method)

Results

1. Linear Regression Model

- From the summary table, all features are significant as their p-value is greater than 0.05.
- Achieved MSE of 22.85 and R-squared value of 0.97 on test data.

```
> summary(lm_model)

Call:
lm(formula = Usage_kWh ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.201  -0.958   0.062   1.224 118.706 

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -1.191e+01  4.749e-01 -25.075  < 2e-16 ***
Lagging_Current_Reactive.Power_kVarh  3.011e-01  4.635e-03  64.980  < 2e-16 ***
Leading_Current_Reactive_Power_kVarh  1.185e-01  1.263e-02   9.382  < 2e-16 ***
CO2.tCO2.                             1.687e+03  5.764e+00 292.705  < 2e-16 ***
Lagging_Current_Power_Factor          1.229e-01  3.024e-03  40.642  < 2e-16 ***
Leading_Current_Power_Factor          6.917e-02  3.264e-03  21.191  < 2e-16 ***
NSM                                   9.684e-06  1.476e-06   6.560 5.49e-11 ***
WeekStatus                           -1.601e-01  7.281e-02  -2.199  0.02788 *  
Day_of_week                           4.158e-02  1.505e-02   2.764  0.00572 ** 
Month                                -9.822e-02  8.880e-03 -11.061  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.655 on 24518 degrees of freedom
Multiple R-squared:  0.9807,	Adjusted R-squared:  0.9807 
F-statistic: 1.384e+05 on 9 and 24518 DF,  p-value: < 2.2e-16

> cat(" Mean Squared Error (MSE):", mse, "\n")
 Mean Squared Error (MSE): 22.85371 
> cat("R-squared:", r_squared, "\n")
R-squared: 0.979435 
```

2. Subset Selection

- Applied Forward subset selection
- Based on adjusted R-squared, the minimum number of variables is 8.

```
> coef(best_subsets, 8)
          (Intercept) Lagging_Current_Reactive.Power_kVarh Leading_Current_Reactive_Power_kVarh
        -1.230485e+01                         3.023797e-01                         1.222028e-01
             CO2.tCO2.          Lagging_Current_Power_Factor          Leading_Current_Power_Factor
         1.686720e+03                         1.230232e-01                         7.064882e-02
                  NSM                           Day_of_week                                Month
         9.862332e-06                         4.662592e-02                        -9.775445e-02
```

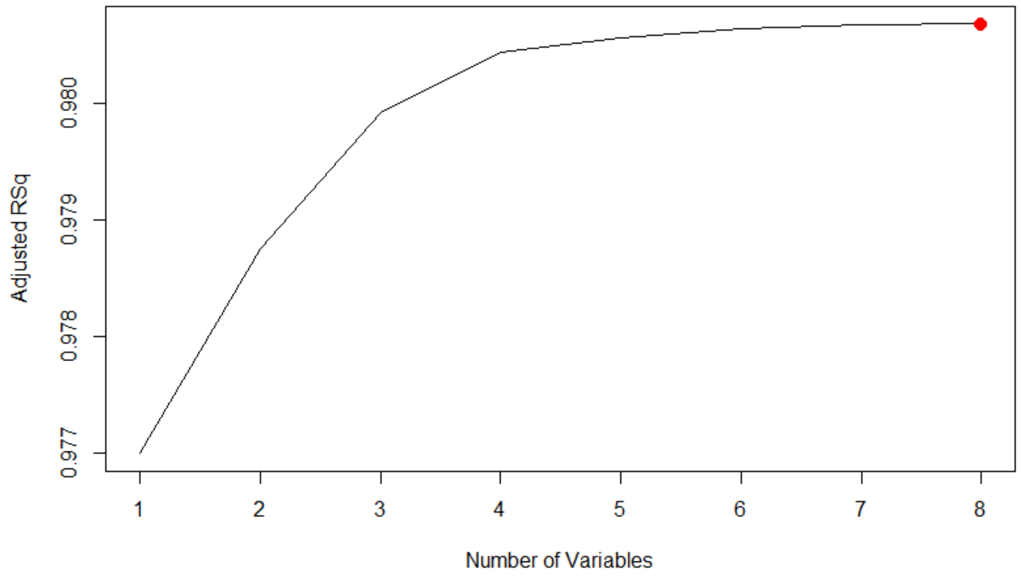


Fig. Adjusted R-squared curve

- Achieved MSE of 22.85 and R-squared value of 0.97 on test data using LR on selected subsets.

```
> # Calculate Mean Squared Error (MSE)
> mse_subset_lr <- mean((actual_Y - Y_pred)^2)
> mse_subset_lr
[1] 22.85903
```

```
> R_squared_subset_lr
[1] 0.9794302
```

3. LASSO and Ridge Regression Models

- Achieved MSE of 21.124 using Lasso model with best lambda of 0.078
- Got Zero coefficients for two features using Lasso

```
> lasso.coef1[lasso.coef1==0]
Leading_Current_Reactive_Power_kVarh                          Day_of_week
                                   0                                    0
```

- Achieved MSE of 30.145 using Ridge model with best lambda of 3.3096

```
> lasso_best_lambda            > ridge_best_lambda
[1] 0.07825612                 [1] 3.309641
> mse_lasso <- mean((y_test - lasso_predictions)^2)
> cat("MSE for LASSO Regression:", mse_lasso, "\n")
MSE for LASSO Regression: 21.12467
> rmse_ridge <- sqrt(mean((y_test - ridge_predictions)^2))
> cat("RMSE for Ridge Regression:", rmse_ridge)
RMSE for Ridge Regression: 5.490532
```

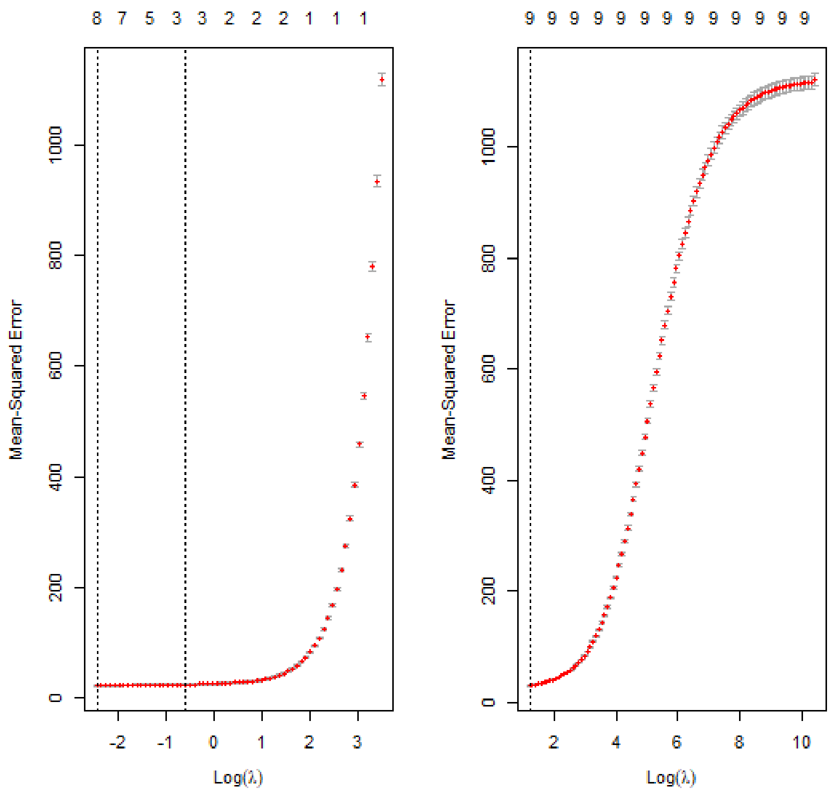


Fig. MSE Vs Log(lamda) for Lasso and Ridge model

4. Principal Component Analysis (PCA)

- From summary table, all features sorted by their contribution using PCA.
- Selected top 7 features and applied LR model.
- Achieved MSE of 22.86914 adjusted R-squared of 0.9794 on test data.

```
> summary(model_pcr)

Call:
lm(formula = Usage_kWh ~ ., data = selected_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.201  -0.958   0.062   1.224 118.706 

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         -1.191e+01  4.749e-01 -25.075  < 2e-16 ***
NSM                                  9.684e-06  1.476e-06   6.560 5.49e-11 ***
Leading_Current_Power_Factor         6.917e-02  3.264e-03  21.191  < 2e-16 ***
Lagging_Current_Power_Factor         1.229e-01  3.024e-03  40.642  < 2e-16 ***
Leading_Current_Reactive_Power_kVarh 1.185e-01  1.263e-02   9.382  < 2e-16 ***
Lagging_Current_Reactive.Power_kVarh 3.011e-01  4.635e-03  64.980  < 2e-16 ***
Month                               -9.822e-02  8.880e-03 -11.061  < 2e-16 ***
CO2.tCO2.                            1.687e+03  5.764e+00 292.705  < 2e-16 ***
Day_of_week                          4.158e-02  1.505e-02   2.764  0.00572 ** 
WeekStatus                          -1.601e-01  7.281e-02  -2.199  0.02788 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.655 on 24518 degrees of freedom
Multiple R-squared:  0.9807,	Adjusted R-squared:  0.9807 
F-statistic: 1.384e+05 on 9 and 24518 DF,  p-value: < 2.2e-16
```

Fig. Summary table for PCA model

```
> cat("MSE:", mse, "\n")
MSE: 22.86914
> # Print the R-squared value
> cat("R-squared:", R_squared_pcr, "\n")
R-squared: 0.9794211
```

5. Random Forest

- We applied RF model on train data and evaluated trained model on test data.
- We achieved MSE of 2.7521 and R-squared of 0.9975 on test data.

```
> summary(rf_model)
                Length Class  Mode
call                3  -none- call
type                1  -none- character
predicted       24528  -none- numeric
mse               500  -none- numeric
rsq               500  -none- numeric
oob.times       24528  -none- numeric
importance          9  -none- numeric
importanceSD        0  -none- NULL
localImportance     0  -none- NULL
proximity           0  -none- NULL
ntree               1  -none- numeric
mtry                1  -none- numeric
forest             11  -none- list
coefs               0  -none- NULL
y               24528  -none- numeric
test                0  -none- NULL
inbag               0  -none- NULL
terms               3  terms  call
```

```
> cat(" Mean Squared Error (MSE):", mse, "\n")
 Mean Squared Error (MSE): 2.752177
> cat("R-squared:", r_squared, "\n")
R-squared: 0.9975234
```

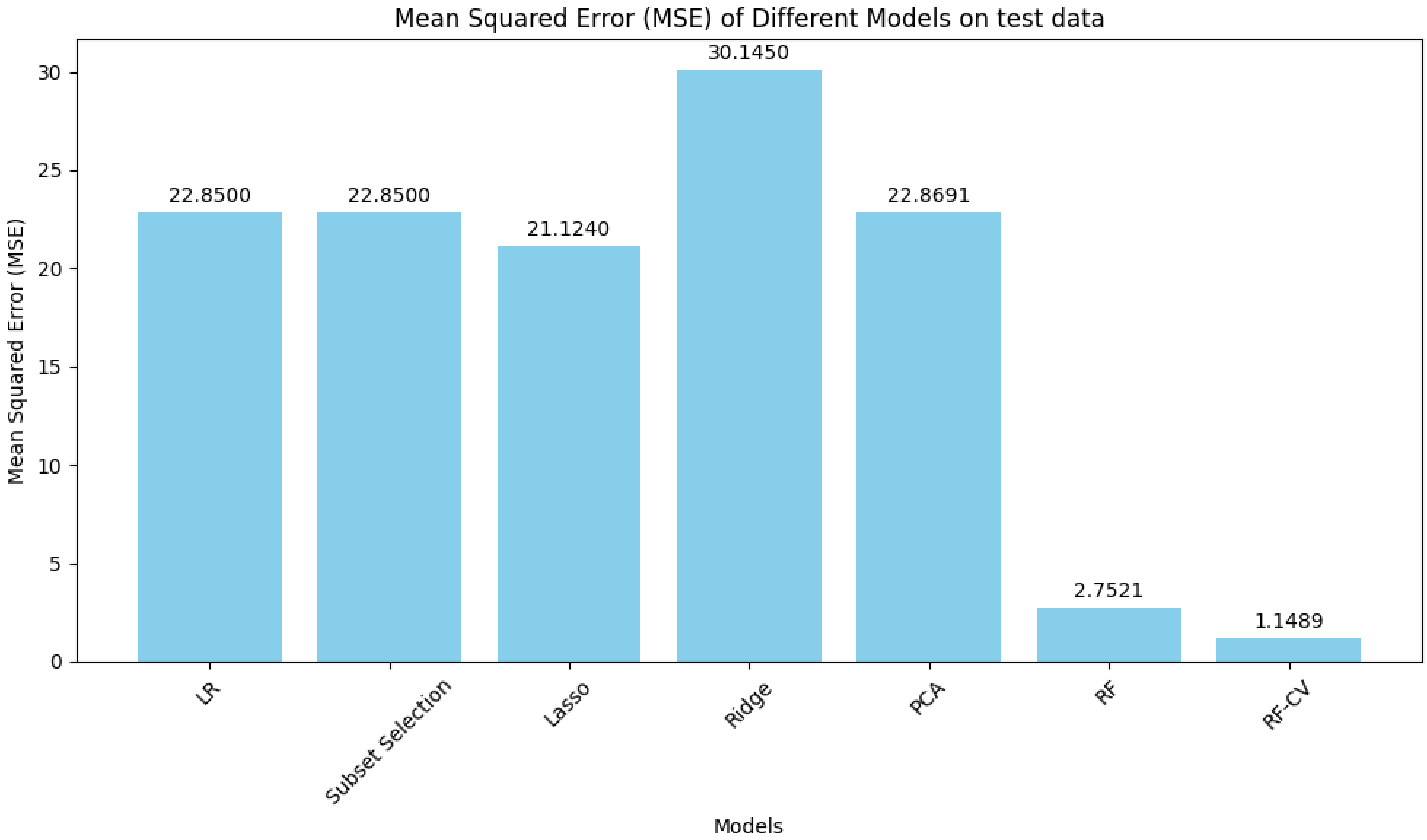
6. Random Forest with CV

- Applied 5-folds CV on training dataset to train RF model.
- Evaluated cross-validated model on test data.
- We achieved MSE of 1.148884 and R-squared of 0.999082 on test data

```
> print(rf_model_cv)
Random Forest

24528 samples
    9 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 19623, 19622, 19621, 19623, 19623
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
  2     2.508023  0.9945965  1.3459579
  5     1.412755  0.9982218  0.5911387
  9     1.250788  0.9985849  0.4118057

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 9.
```

```
> # Print the metrics
> cat("MSE for RF-Cv model on test data:", mse_cv, "\n")
MSE for RF-Cv model on test data: 1.148884
> cat("R-squared for RF-Cv model on test data on test data:", r_squared_cv, "\n")
R-squared for RF-Cv model on test data on test data: 0.9989662
```

Comparison


Mean Squared Error (MSE) of Different Models on test data
Mean Squared Error (MSE)
0
5
10
15
20
25
30
22.8500
22.8500
21.1240
30.1450
22.8691
2.7521
1.1489
LR
Subset Selection
Lasso
Ridge
PCA
RF
RF-CV
Models

Conclusion

- In this project, we have successfully explored various statistical learning methods on the *Steel Industry Energy Consumption dataset*. our proposed model emerged as the top-performing solution.
- Among the various models used for detecting power consumption, RF with CV achieved higher R-squared of 0.999082 and lower test MSE of 1.148884.
- Hence, the success of the model in accurately predicting energy consumption signifies a significant step toward achieving enhanced efficiency and cost reduction in energy consumption within the steel industry.