UNIVERSITI TUNKU ABDUL RAHMAN

ACADEMIC YEAR 2021/22

PRACTICAL ASSIGNMENT

UEMH 3163 ARTIFICIAL INTELLIGENCE

LAB 3 – Supervised Learning Assignment

Name	Student ID	Course
Gervin Fung Da Xuen	1801655	Software Engineering
Hwong Yu Jie	1900894	Software Engineering
Tay Ming Liang	1804612	Software Engineering

Table of Contents

1. Explanation of Dataset

The chosen dataset for this assignment is Energy Efficiency. This data is available at https://archive.ics.uci.edu/ml/datasets/Energy+efficiency. The dataset is used to study energy efficiency in buildings. The dataset contains 768 samples and eight characteristics, with the goal of predicting two energy loads in buildings.

Table 1.0 Features and Targeted Responses of The Dataset

Feature				Target	
X1	Relative Compactness	X5	Overall Height	Y1	Heating Load
X2	Surface Area	X6	Orientation	Y2	Cooling Load
X3	Wall Area	X7	Glazing Area		
X4	Roof Area	X8	Glazing Area Distribution		

The table above summarises the features and targeted responses of the dataset, the left column of feature and target is the variable name assigned in the excel or CSV file while the right column is the real name of each feature and target responses.

2. Preprocessing Data

A raw dataset is not ready to be used directly to train models. Thus, the dataset is cleaned, formatted and transformed so that the trained models can yield predictions with high accuracy.

```
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Relative Compactness       1296 non-null   object
 1   Surface                    1296 non-null   object
 2   Wall Area                  1296 non-null   object
 3   Roof Area                  1296 non-null   object
 4   Overall Height             1296 non-null   object
 5   Orientation                1296 non-null   object
 6   Glazing Area               1296 non-null   object
 7   Glazing Area Distribution  1296 non-null   object
 8   Heating Load               1296 non-null   object
 9   Cooling Load               1296 non-null   object
dtypes: object(10)
```

Figure 2.0 Description of The Raw Data

2.1 Check for missing value

Firstly, the dataset is checked if it contains data with missing values by searching for NaN and empty strings. Data with missing values are removed to improve the accuracy of the outputs.

2.2 Data formatting

Since the data is found to be in object type, they are formatted to either integer or float accordingly.

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Relative Compactness    768 non-null    float64
 1   Overall Height          768 non-null    float64
 2   Orientation             768 non-null    int32
 3   Glazing Area            768 non-null    float64
 4   Glazing Area Distribution  768 non-null    int32
 5   Heating Load            768 non-null    float64
 6   Cooling Load            768 non-null    float64
dtypes: float64(5), int32(2)
```

Figure 2.1 Description of Preprocessed Data

2.4 Data Transformation

Scaling data helps in improving performance of the model. Standardization method is adopted to scale the data because it can reduce multicollinearity (Frost, n.d.). The data is standardized so that its mean and standard deviation are 0 and 1 respectively.

2.3 Check for collinearity and multicollinearity

When two features or predictors are strongly correlated, it is known as collinearity or multicollinearity when more than two features are involved (Ayuya, 2021). Collinearity should be avoided because it affects the interpretability of the model (Choueiry. n.d.). Hence, the dataset is checked for collinearity by calculating variance inflation factor (VIF). Generally, VIF score that exceeds 10 suggests collinearity (Choueiry, n.d.). To reduce collinearity, features with high VIF scores are combined or removed.

3. Data Splitting

The dataset is split into two portions, where 80 percent of the dataset will be used for training while the remaining 20 percent is used to test the trained models.

4. Hyper Parameter Tuning

To successfully train a model that performs well, it is important to use the correct hyper parameters of the model, which are parameters whose values cannot be inferred from the data. As such, grid search is implemented to fine tune and determine the best hyperparameters that

work best. Grid search tries each and every combination of the provided hyperparameters to find the best combination with the highest average cross validation score.

The regression techniques used in this assignment are 4 regression model

1. Random Forest regression
2. Ridge Regression
3. Lasso Regression
4. Linear Regression

5. Results

Summary

Table 5.0) Summary of Best Parameters

Model	Best Estimator	
	Heating	Cooling
Random Forest Regression	RandomForestRegressor(max_depth=90 , n_estimator=300)	RandomForestRegressor(max_depth=100)
Ridge Regression	Ridge(fit_intercept=False, normalize=True, tol=1e-05)	Ridge(fit_intercept=False, normalize=True, tol=1e-05)
Lasso Regression	Lasso(tol=1e-05 ,max_iter=10000)	Lasso(tol=1e-05 ,max_iter=10000)
Linear Regression	LinearRegression(normalize=True)	LinearRegression(normalize=True)
Poisson Regression	PoissonRegressor(tol=1e-05 ,max_iter=10000, warm_start=True)	PoissonRegressor(tol=1e-05 ,max_iter=10000, warm_start=True)

Table 5.1) Summary of Best Score

	Random Forest Regression		Ridge Regression		Lasso Regression		Linear Regression		Poisson Regression	
	Heating	Cooling	Heating	Cooling	Heating	Cooling	Heating	Cooling	Heating	Cooling
Train	0.9997	0.9957	0.9053	0.8795	0.8449	0.8199	0.9087	0.8808	0.9188	0.8975

score (r^2)										
Test score (r^2)	0.9981	0.9724	0.9070	0.8927	0.8434	0.8326	0.8940	0.8867	0.8915	0.8951
Adjusted r^2	0.9973	0.9768	0.8988	0.8761	0.8381	0.8629	0.8945	0.8654	0.8879	0.8915
Time elapsed (s)	2526	2796	12	9	36	33	0	0	27	25

Interpretability

R^2 is the coefficient of determination, which represents the strength between the relationship between the predictors collectively and dependent variables. The value of R^2 ranges from 0 to 1. The higher the value of r^2, the higher the proportion of variance that can be explained by a predictor. While high r^2 can be interpreted that the model is a good fit, there are other factors that can contribute to high coefficient of determination. R^2 can be affected by the number of predictors. Any increase in the number of predictors will always result in an increase in r^2 value (Grace-Martin, n.d.).

Due to the reduced interpretability of r^2 in models with high numbers of predictors, adjusted R^2 is introduced for interpretation. Adjusted R^2 is more suitable to measure the fitness of multivariate regression because the increase due to the increase in number of predictors is neutralized (Grace-Martin, n.d.).

1. **Random Forest Regression**

The Random Forest is a model consisting of many decision trees, which forms the "forest". Moreover, it also adds a certain degree of randomness into the model, hence it searches for the best feature among a random subset of features. It can be seen that it introduces high accuracy even if it takes very long to complete the algorithm, possibly due to many combinations of different parameters and huge iterations. Lastly, it's also possible that it's slow due to the "forest" generated by the algorithm.

Furthermore, the parameters for both predictions are not the same. For predicting Heating, a maximum depth (the depth of the tree) of 90 is used, hence the deeper the tree, the

more information it contains. Also, “n_estimator” is 300 which means for this particular model it has 300 leaves, more leaves means that the model will learn more about the data, however, too many leaves is detrimental as it will slow down the algorithm

.Lastly, there is only a parameter used to predict Cooling which is maximum depth of 100 and similarly it also managed to obtain a high accuracy of prediction as well. It’s worth noting that it took about 40 minutes 30 seconds to complete the model.

```
[Parallel(n_jobs=-1)]: Done 9000 out of 9000 | elapsed: 44.5min finished

Random Forest Grid Search CV Best Score 0.9967447699024256
Random Forest Grid Search CV Best Estimator RandomForestRegressor(max_depth=90, n_estimators=300)
Random Forest Grid Search CV Training Score 0.9996176357977228
Random Forest Grid Search CV Testing Score 0.9974240446585745
```

Figure 5.0: Heating Prediction of Ridge Regression

```
[Parallel(n_jobs=-1)]: Done 9000 out of 9000 | elapsed: 40.5min finished

Random Forest Grid Search CV Best Score 0.9671955359926063
Random Forest Grid Search CV Best Estimator RandomForestRegressor(max_depth=100)
Random Forest Grid Search CV Training Score 0.995404942518749
Random Forest Grid Search CV Testing Score 0.9775618910990339
```

Figure 5.1: Cooling Prediction of Ridge Regression

2. Ridge Regression

The Ridge Regression is a model that is used to analyse data with multicollinearity. When there is a problem with multicollinearity, least-squares are unbiased, and variances are significant, the projected values are distant from the actual values. The figure below shows the result of prediction (heating and cooling).

```
[Parallel(n_jobs=-1)]: Done 3000 out of 3000 | elapsed:    7.2s finished
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.

Ridge Regression Grid Search CV Best Score 0.8953519490481904
Ridge Regression Grid Search CV Best Estimator Ridge(tol=1e-05)
Ridge Regression Grid Search CV Training Score 0.9000822948435954
Ridge Regression Grid Search CV Testing Score 0.928164338286723
```

Figure 5.2: Heating Prediction of Ridge Regression

```
[Parallel(n_jobs=-1)]: Done 3000 out of 3000 | elapsed:    3.0s finished
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.

Ridge Regression Grid Search CV Best Score 0.8767831584084261
Ridge Regression Grid Search CV Best Estimator Ridge(tol=1e-05)
Ridge Regression Grid Search CV Training Score 0.880386840401586
Ridge Regression Grid Search CV Testing Score 0.8883040634297871
```

Figure 5.3: Cooling Prediction of Ridge Regression

Both figures shown above use the same parameter to form the best result or model. The "fit_intercept" is set to *False* to indicate that there will be no intercept calculation. The "normalize" is set to *True* which means it will normalize "x" regressors. Lastly, "tol" is the precision of the solution. Hence, Ridge Regression works best for this dataset when interception will not be calculated with the x regressors normalized and precision of 0.00001.

3. **Lasso Regression**

The Lasso Regression is another regression model that performs regularization to prevent overfitting. It can effectively reduce the number of features used in regression. Hence, the Lasso Regression is well-suited for models with high levels of multicollinearity. However, since the multicollinearity is reduced, therefore the Lasso Regression may obtain a lower accuracy compared to other regression model.

```
[Parallel(n_jobs=-1)]: Done 9000 out of 9000 | elapsed:    9.2s finished

Lasso Regression Grid Search CV Best Score 0.8400643820577038
Lasso Regression Grid Search CV Best Estimator Lasso(max_iter=10000, tol=1e-05)
Lasso Regression Grid Search CV Training Score 0.8448710527811366
Lasso Regression Grid Search CV Testing Score 0.8434378503364446
```

Figure 5.4: Heating Prediction of Lasso Regression

```
[Parallel(n_jobs=-1)]: Done 9000 out of 9000 | elapsed:    9.2s finished

Lasso Regression Grid Search CV Best Score 0.8166124679756956
Lasso Regression Grid Search CV Best Estimator Lasso(max_iter=10000, tol=1e-05)
Lasso Regression Grid Search CV Training Score 0.8199159419922368
Lasso Regression Grid Search CV Testing Score 0.8326050429755372
```

Figure 5.5: Cooling Prediction of Lasso Regression

Referring to the two figures shown above, it's evident that the Lasso Regression has a score of 0.80 to 0.85 which is relatively low as compared to that of other regression models. Still, the parameters needed for Lasso Regression for both predictions are the same whereby the maximum number of iteration, denoted by "max_iter" is ten thousand (10 000) and the "tol" which is the tolerance for optimization. Hence, when the tolerance of optimization is 0.00001 and the maximum number of iteration is 10 000, Lasso Regression can make the best prediction for this dataset.

4. Linear Regression

Linear regression is a linear approach to model the relationship between dependent and independent variables and the amount of independent variables they employ. Lasso and Ridge Regression are regularization methods based on Linear Regression.

```
Fitting 5 folds for each of 4 candidates, totalling 20 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  20 out of  20 | elapsed:    0.0s finished


Adjusted r squared: 0.826949807940927
Fitting 5 folds for each of 4 candidates, totalling 20 fits
Linear Regression Grid Search CV Best Score 0.9042164564262677
Linear Regression Grid Search CV Best Estimator LinearRegression(normalize=True)
Linear Regression Grid Search CV Training Score 0.9066665486093857
Linear Regression Grid Search CV Testing Score 0.9028902434398559

Adjusted r squared: 0.8996095084209321
Fitting 5 folds for each of 4 candidates, totalling 20 fits
Linear Regression Grid Search CV Best Score 0.8772062510206935
Linear Regression Grid Search CV Best Estimator LinearRegression(normalize=True)
Linear Regression Grid Search CV Training Score 0.8790934379756637
Linear Regression Grid Search CV Testing Score 0.8940478008047992
```

Figure 5.6: Heating (top) and Cooling (bottom) Prediction for Linear Regression

The Figure above shows the R^2 is above 0.75 and thus has a strong relationship between dependent and independent variables. Both scores for prediction of Heating and Cooling are considerably high and there is only one parameter required, that is the data has to be normalized. Hence this shows that even if there are 768 data in this dataset, and has already been normalized due to its massive quantities, normalization should also be performed for this model because normalizing data to reduce redundancy and to organize the data more easily. Thus Linear Regression will perform the best prediction for this dataset if the data is normalized.

5. Poisson Regression

Poisson Regression is a linear model for Poisson Distribution. Although Poisson distribution is best suited for count data, it can be used in discrete or continuous data as well and the dataset chosen contains both discrete and continuous data as well. To further strengthen the point, this dataset contains 768 data and thus has already been approximately normalized due to its massive quantities, and a normalized data is very suitable for the Poisson Regression Model.

```
[Parallel(n_jobs=-1)]: Done 9000 out of 9000 | elapsed:   26.5s finished
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.

Poisson Regression Grid Search CV Best Score 0.9160405753564682
Poisson Regression Grid Search CV Best Estimator PoissonRegressor(max_iter=10000, tol=1e-05, warm_start=True)
Poisson Regression Grid Search CV Training Score 0.9187779380519988
Poisson Regression Grid Search CV Testing Score 0.8915224505437771
```

Figure 5.7: Heating Prediction for Poisson Regression

```
[Parallel(n_jobs=-1)]: Done 9000 out of 9000 | elapsed:   25.1s finished
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 4 concurrent workers.

Poisson Regression Grid Search CV Best Score 0.8941448387383278
Poisson Regression Grid Search CV Best Estimator PoissonRegressor(max_iter=10000, tol=1e-05, warm_start=True)
Poisson Regression Grid Search CV Training Score 0.8975495416617187
Poisson Regression Grid Search CV Testing Score 0.8950873940697218
```

Figure 5.8: Cooling Prediction for Poisson Regression

The figure shown above proves that Poisson Regression also works for this model and has obtained a very high accuracy. Also, both Poisson Model used the same parameters, which has a maximal number of iterations for the solver of ten thousand, stopping criteria of 0.00001 and "warm-start" equal to True which means that it will reuse the solution of the previous call. This Poisson Regression Model should be best used for counting data, however, it's also suitable for data that are normalized as well as continuous and discrete data.

Conclusion

As a conclusion, it's evident that normalizing more often than not is not required even if Lasso and Ridge can perform normalization to obtain a higher accuracy for the dataset. This is due to the large number of data in this dataset, and thus had already been normalized. Furthermore, from all of the five regression models implemented, it's found that RandomForestRegressor obtained the highest score but took the longest. On the other hand, Linear Regression is the fastest to be computed and managed to obtain a decent score as well. Lastly, it's important to try multiple combinations of different parameters to obtain the best parameters for a particular model.

Reference

Engineering Education (EngEd) Program | Section. 2021. *How to Detect and Correct Multicollinearity in Regression Models*. [online] Available at: <https://www.section.io/engineering-education/multicollinearity/> [Accessed 6 September 2021].

Choueiry, G., 2021. *Correlation vs Collinearity vs Multicollinearity – Quantifying Health*. [online] Quantifyinghealth.com. Available at: <https://quantifyinghealth.com/correlation-collinearity-multicollinearity/> [Accessed 6 September 2021].

Frost, J., 2021. *When Do You Need to Standardize the Variables in a Regression Model? - Statistics By Jim*. [online] Statistics By Jim. Available at: <https://statisticsbyjim.com/regression/standardize-variables-regression/> [Accessed 6 September 2021].

Grace-Martin, K., 2021. *Assessing the Fit of Regression Models - The Analysis Factor*. [online] The Analysis Factor. Available at: <https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/> [Accessed 6 September 2021].