Department of Electronics and Telecommunication Engineering

EN3150 Assignment 01

Learning from data and related challenges and linear models for regression

Name: S.M.S.M.B.Abeyrathna
Index No: 210005H

Date: 2024/09/10

1) Data Preprocessing

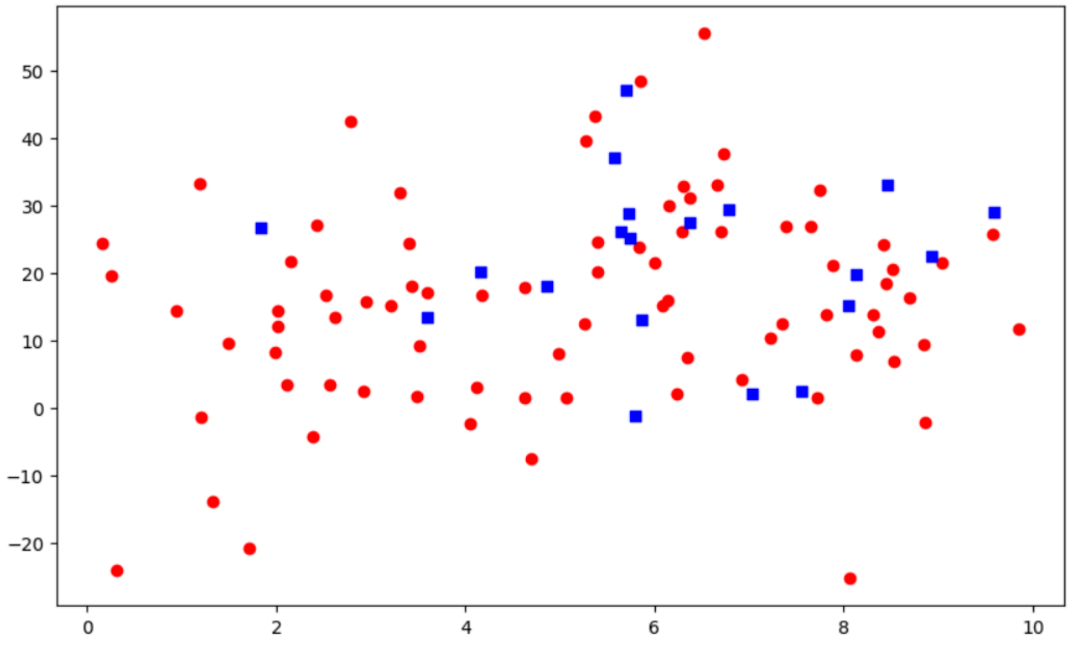
1)

- **Feature 1 -** The values of Feature 1 are mostly concentrated around 0, with a few significant outliers. Since the data is centred around zero and has outliers, **max - abs scaling** is likely the best choice. This method will scale the values based on the maximum absolute value, preserving the structure and sparsity of the feature while keeping the influence of outliers minimal.

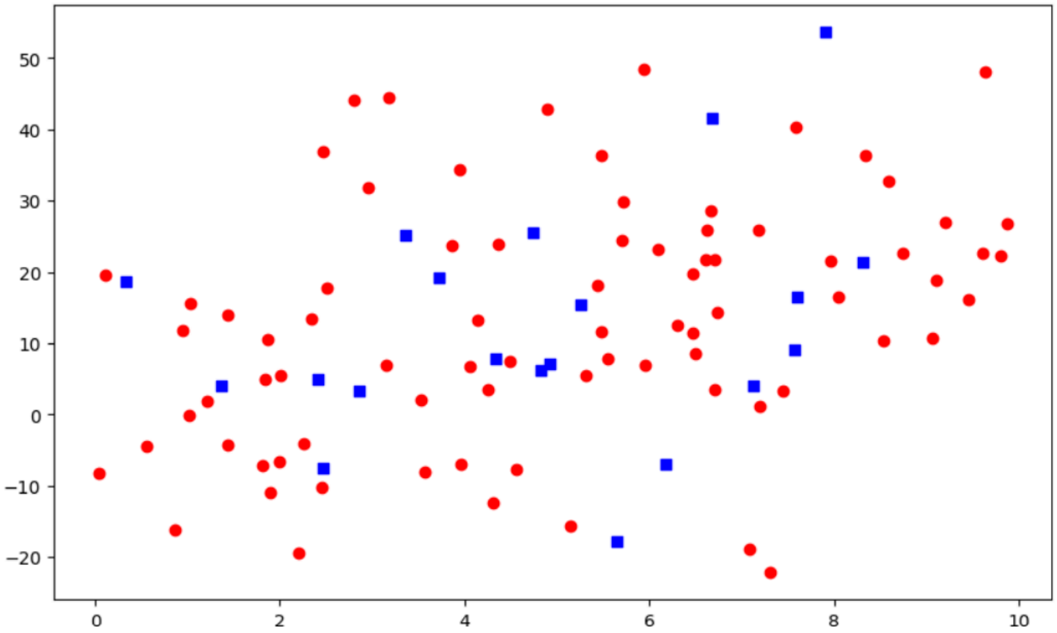
- **Feature 2 -** The values of Feature 2 vary significantly, with a wide range from around -40 to 30. This feature does not appear to be centred around zero and has a broad spread of values. **min-max scaling** would be appropriate here to scale the data to a standard range, preserving the relative distances between data points while compressing the wide range.

2) Learning from data

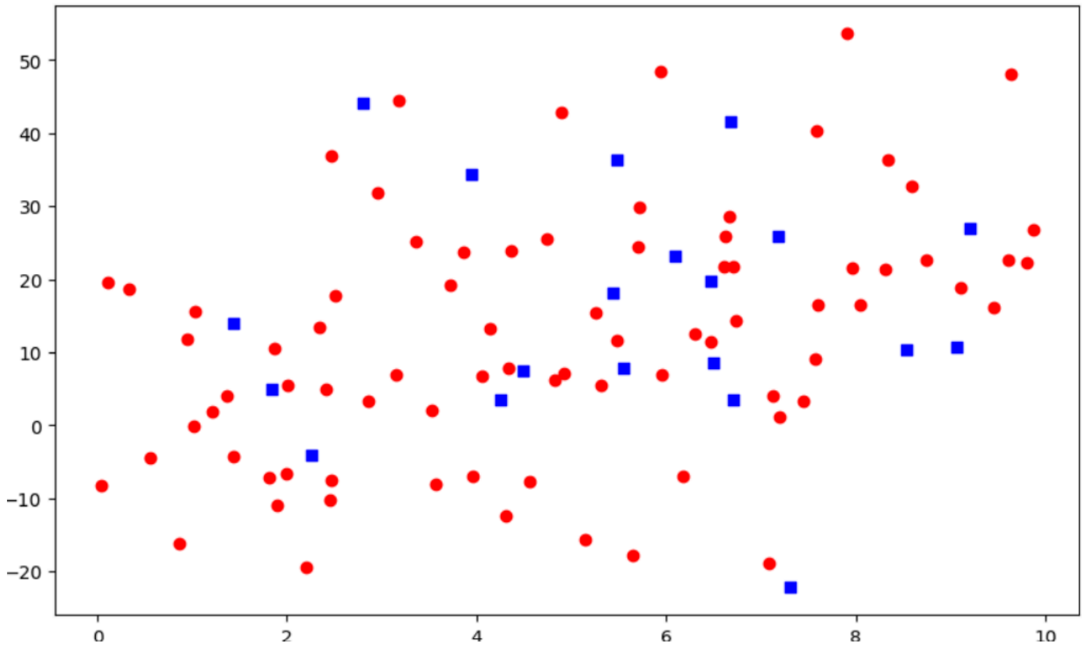
2)



Attempt 1



Attempt 2

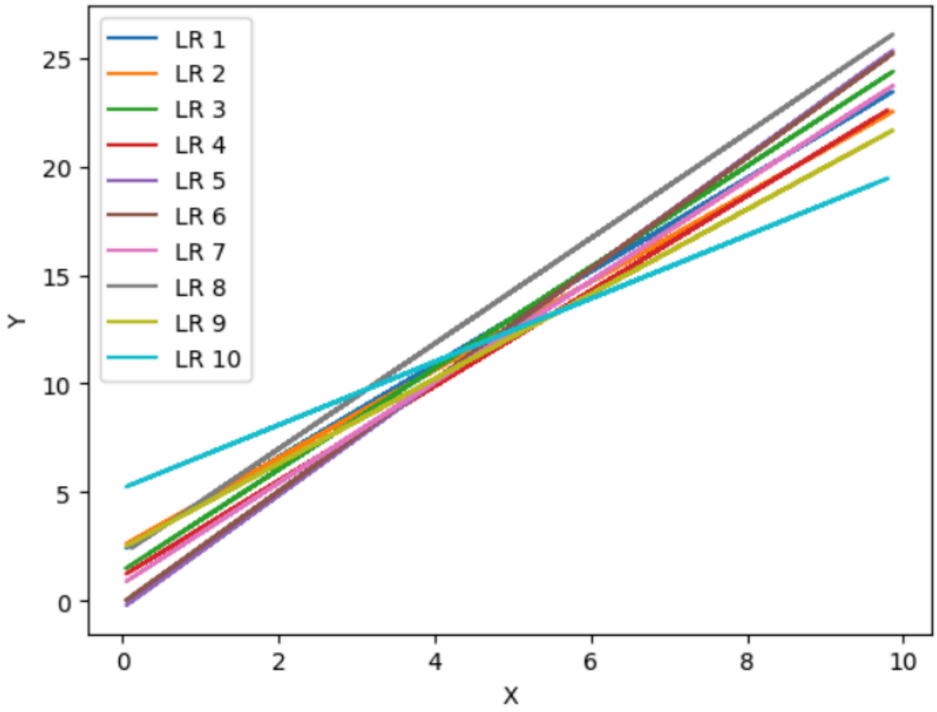


Attempt 3

- **Observations -** When I run the code in Listing 2 multiple times, the training and testing data points will be different in each run.

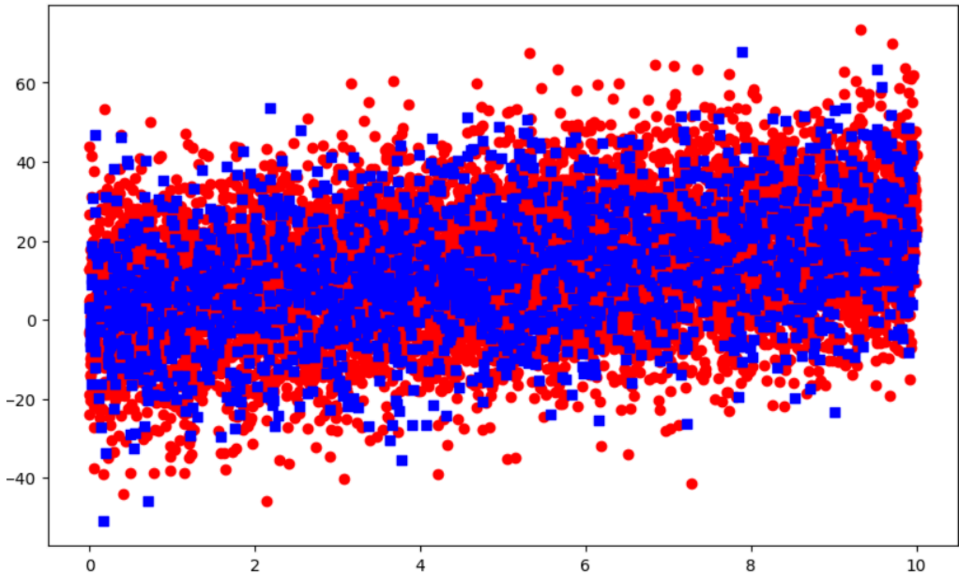
- **Reason for the Variation -** This variation occurs because the random_state is being set using a randomly generated integer (r = np.random.randint(104)), which changes in each execution. The random_state parameter controls the shuffling of the data before splitting it into training and testing sets. Since r changes in each run, the split is different every time, leading to different subsets of data being allocated to training and testing.

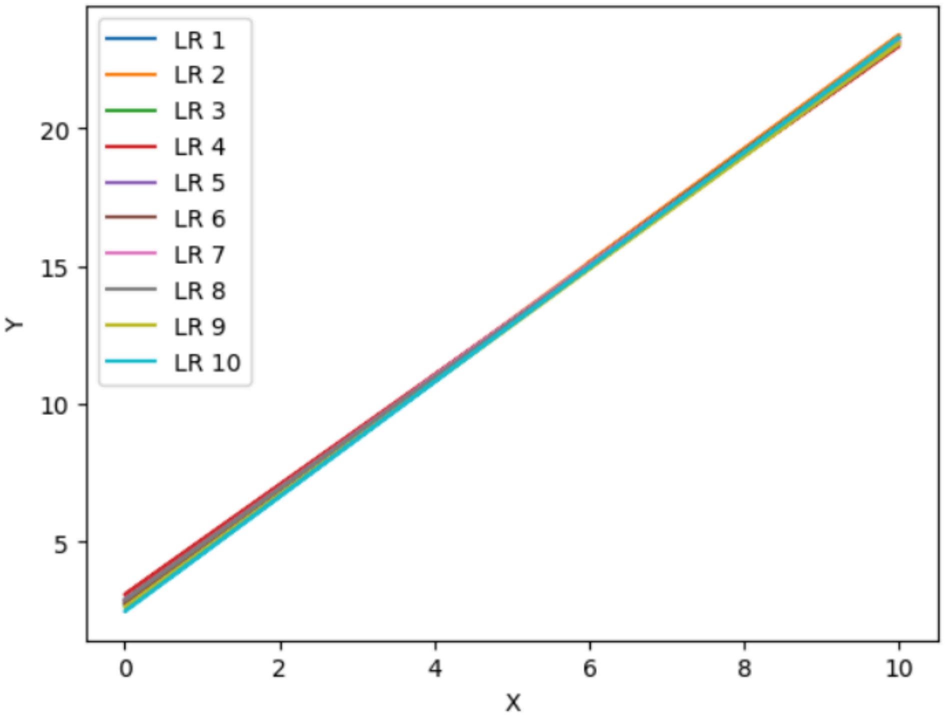
3)



The linear regression model differs across instances because the random split of data into training and testing sets changes each time the code is run. This variability in training data leads to different model parameters, resulting in different regression lines.

4)





Observations:

- With 10,000 samples, the regression lines from different instances are almost identical, showing very little variation.
- In contrast, with 100 samples, there was noticeable variability in the regression lines across different instances.

Reason:

- The larger dataset provides a more stable and representative sample of the underlying data distribution. This reduces the impact of random variations in the training data split, leading to more consistent model parameters and predictions. With more data, the model has a better chance to learn the true underlying relationship between the variables, making it less sensitive to the specific random split of the data.

3) <u>Linear Regression on real world data</u>

2)

- Independent Variables (Features): 33
- Dependent Variables (Targets): 2

3)

Yes, it is possible to apply linear regression on this dataset. But it can not be directly applied.

Steps to Follow Before Applying Linear Regression:

- Encode Categorical Variables: Convert categorical variables like “Gender”, “Age” and “Ethnicity” into numerical form using techniques like One Hot Encoding or Label Encoding.

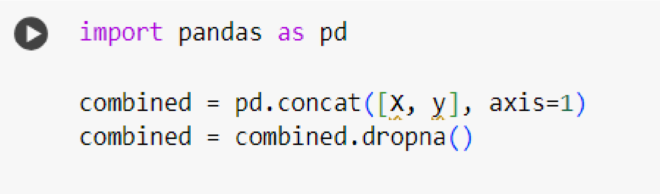
- Check for Multicollinearity: Identify and handle multicollinearity, where independent variables are highly correlated. This can distort the regression model.
- Feature Scaling: Scale the numerical features using techniques like Standardization or Normalisation. Linear regression models can perform better when the data is on a similar scale.
- Split Data into Training and Testing Sets: Split the dataset into training and testing sets to validate the performance of the linear regression model.

- Feature Selection or Dimensionality Reduction (if necessary): If the dataset has too many features, consider techniques like Principal Component Analysis (PCA) or feature selection methods to reduce the number of features, especially if the model suffers from overfitting or high variance.

4)

The given code can be used to remove NaN or missing values separately. However, this approach is not entirely correct because dropping missing values from X and y independently can lead to a mismatch between the features (X) and the target values (y). When a row containing a missing value is removed from X, the corresponding y value might not be removed correctly.

To resolve this issue, X and y should first be combined into a single dataframe, ensuring that both X values and their corresponding y values are removed together when dropping missing data. This prevents any mismatches. The corrected code is as follows:



```
import pandas as pd

combined = pd.concat([X, y], axis=1)
combined = combined.dropna()
```

5)

After executing the code, I obtained an output displaying the features ‘Age’, ‘T atm’, ‘Humidity’, ‘T Max1’ and ‘T offset1’ from the dataset. The resulting data consists of 1,018 rows with these specified columns.

6) After splitting

The Training Set has 814 and the Testing Set has 204 samples.

7)

- Intercept: 5.350271904448654
- Coefficients: [0.00106708, -0.06026008, 0.00099853, 0.91390548 ,0.09962315]

8)

- T _Max1
  The significance of each independent variable is reflected by the magnitude of its coefficients. The feature with the largest absolute coefficient holds the greatest influence in predicting the target variable, aveOralM.

9)

- Intercept: 7.1436575396059006
- Coefficients: [-0.05575003, 0.60320041, -0.05115736, 0.34035037]

10)

- Residual Sum of Squares : *20.30989897585181*
- Residual Standard Error : *0.31946798563940887*
- Mean Squared Error : *0.09955832831299906*
- R^2 Squared: *0.6428739418600805*
- Standard Error of Coefficients:
  const : 1.56946255
  T_OR_Max1 : 1.80438807
  T_FHC_Max1 : 0.09491936
  T_FH_Max1 : 0.08663473
  T_OR1 : 1.79706711

- t-values:
  const : 5.51634926
  T_OR_Max1 : 0.76840282
  T_FHC_Max1 : -0.46550164
  T_FH_Max1 : -1.15883401
  T_OR1 : 3.70343768

- p-values:
  const : 1.06823189 x 10^{-7}
  T_OR_Max1 : 4.43158815 x 10^{-1}
  T_FHC_Max1 : 6.42081296 x 10^{-1}
  T_FH_Max1 : 2.47912556 x 10^{-1}
  T_OR1 : 2.75451148 x 10^{-4}

11)

The p-value analysis suggests that features like T_OR_Max1 and T_FHC_Max1 can be discarded, as their high p-values indicate they are not significant in predicting the target variable.

4) Performance Evaluation of Linear Regression Models

2)

Model A:

$$\text{RSE}_A = \sqrt{\frac{9}{10000-1-2}} \approx 0.03$$

Model B:

$$\text{RSE}_B = \sqrt{\frac{2}{10000-1-4}} \approx 0.01414$$

3)

Model A:

$$R_A^2 = 1 - \frac{9}{90} = 0.9$$

Model B:

$$R_B^2 = 1 - \frac{2}{10} = 0.8$$

Model A has a higher R², indicating that it explains a greater portion of the variance in the response variable.

4)

Residual Standard Error (RSE) is generally a better metric for comparing models as it considers both the number of data points and the number of parameters in the model. It penalises models that introduce unnecessary complexity, offering a standardised measure of prediction error. In contrast, **R-squared (R²)** measures the proportion of variance explained by the model but does not account for the number of predictors. Therefore, RSE is preferred when comparing models with varying numbers of parameters, as it provides a more balanced evaluation of model performance.

5) Linear regression impact on outliers

2) When $a \rightarrow 0$:

For L1(w), the loss function becomes highly sensitive to outliers since the penalty for large residuals increases steeply. For L2(w), the loss function also grows more sensitive to outliers because the exponential term escalates rapidly as $|r_i|$ increases.

3)

To minimise the influence of data points with residuals $|r_i| \geq 40$, different approaches are required for L1(w) and L2(w). For L1(w), selecting a relatively small value of a reduces the impact of large residuals since the function increases less rapidly. In contrast, for L2(w), choosing a larger value of a makes the function less sensitive to large residuals due to the dampening effect of the exponential term in the denominator. In summary, using a larger a for L2(w) and a smaller a for L1(w) helps minimise the influence of large residuals, with the specific choice depending on the level of robustness required.