Appendix: Explanation of Key Formulas

Author: Heider Jeffer
Date: 2025 January 8

This appendix provides a detailed explanation of the formulas used in the code. It includes descriptions of the formulas used for feature scaling, linear regression, evaluation metrics, and more.

1. Feature Scaling with StandardScaler

Formula:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

Where:

- X is the feature data (e.g., "OffsettingSchemeEffectiveness" and "Year").
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

The purpose of scaling is to standardize the features to have a mean of 0 and a standard deviation of 1, which helps machine learning algorithms perform better, especially when features have different units or magnitudes.

Example: Let's say we have the following feature values for "OffsettingSchemeEffectiveness":

$$X = [3.5, 4.2, 3.8, 4.5, 4.0]$$

First, compute the mean (μ):

$$\mu = \frac{3.5 + 4.2 + 3.8 + 4.5 + 4.0}{5} = 4.0$$

Then, compute the standard deviation (σ):

$$\sigma = \sqrt{\frac{(3.5 - 4.0)^2 + (4.2 - 4.0)^2 + (3.8 - 4.0)^2 + (4.5 - 4.0)^2 + (4.0 - 4.0)^2}{5}} = 0.324$$

Finally, apply the scaling formula to each value:

$$X_{\text{scaled}} = \frac{X - 4.0}{0.324}$$

This results in the following scaled values:

$$[-1.54, 0.62, -0.62, 1.54, 0.00]$$

2. Linear Regression Model

Formula: The linear regression model is based on the following equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Where:

- y is the dependent variable (e.g., "PolicyGrowth").
- x_1, x_2 are the independent variables (e.g., "OffsettingSchemeEffectiveness" and "Year").
- β_0 is the intercept.
- β_1, β_2 are the coefficients (weights) for each feature.

Linear regression aims to find the best-fitting line that minimizes the difference between the predicted and actual values.

Example: Suppose the fitted model gives the following coefficients:

$$\beta_0 = 2.5, \quad \beta_1 = 0.8, \quad \beta_2 = 0.05$$

For a data point where "OffsettingSchemeEffectiveness" is 4.0 and the "Year" is 2020, the predicted "PolicyGrowth" is:

$$y = 2.5 + 0.8 \cdot 4.0 + 0.05 \cdot 2020 = 2.5 + 3.2 + 101 = 106.7$$

Thus, the predicted value for Policy Growth in 2020 is 106.7.

3. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

Formula for MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y^i_{\text{true}} - y^i_{\text{pred}})^2$$

Where:

- n is the number of test samples.
- y^i_{true} is the actual value of the i-th test sample.
- y^i_{pred} is the predicted value of the i-th test sample.

Formula for RMSE:

$$RMSE = \sqrt{MSE}$$

Where:

- RMSE is the square root of the MSE and provides the error in the same units as the target variable.

Example: Let's consider the actual values for the test set:

$$y_{\text{true}} = [6.0, 6.3, 6.5]$$

And the predicted values:

$$y_{\text{pred}} = [5.8, 6.4, 6.3]$$

First, compute the squared differences:

$$(6.0 - 5.8)^2 = 0.04, \quad (6.3 - 6.4)^2 = 0.01, \quad (6.5 - 6.3)^2 = 0.04$$

Then, calculate MSE:

$$MSE = \frac{0.04 + 0.01 + 0.04}{3} = 0.03$$

Finally, compute RMSE:

$$RMSE = \sqrt{0.03} = 0.173$$

4. R-squared (R²) Score

Formula:

$$R^2 = 1 - \frac{\sum(y_{\text{true}} - y_{\text{pred}})^2}{\sum(y_{\text{true}} - \bar{y}_{\text{true}})^2}$$

Where:

- y_{true} is the actual values.
- y_{pred} is the predicted values.
- $\bar{y}_{\text{true}}$ is the mean of the actual values.

The R² score measures how well the model fits the data. It ranges from 0 to 1, where 1 means the model perfectly explains the variation in the data, and 0 means it does not explain the variance at all.

Example: Given the actual values:

$$y_{\text{true}} = [6.0, 6.3, 6.5]$$

And predicted values:

$$y_{\text{pred}} = [5.8, 6.4, 6.3]$$

The mean of the actual values is:

$$\bar{y}_{\text{true}} = \frac{6.0 + 6.3 + 6.5}{3} = 6.27$$

Now, calculate the total sum of squares (SS_tot):

$$SS_{\text{tot}} = (6.0-6.27)^2+(6.3-6.27)^2+(6.5-6.27)^2 = 0.0729+0.0009+0.0529 = 0.1267$$

The residual sum of squares (SS_res) is:

$$SS_{\text{res}} = (6.0 - 5.8)^2 + (6.3 - 6.4)^2 + (6.5 - 6.3)^2 = 0.04 + 0.01 + 0.04 = 0.09$$

Finally, compute R²:

$$R^2 = 1 - \frac{0.09}{0.1267} = 0.29$$

This means the model explains 29% of the variance in the target variable.