Every workday morning for 30 days date, temperature, and current weather condition, as well as the time to get to work are recorded. Using this data, the goal is to predict how long the commute to work will take on a day in the future.

Which of the four basic machine learning methods are you applying to achieve the goal? Explain why. How is the data organized into feature matrix and target vector and what are their shapes?

Which -

I would go for Supervised Learning - Regression method.

Why -

We are required to predict the ETA to work. Since we have past data for the ETA as well, we can easily split our whole dataset into training and test for our model. Hence, I would go with Supervised Learning. ETA is a continuous vector, so regression makes sense.

Data Organization -

- Feature Matrix would contain date, temperature, and weather as the three features. It would have thirty rows as samples. Its shape would be (30, 3)

- Target Vector would contain the ETA. Its shape would be (30,).

A trained model is evaluated using training and validation data. Training score is 0.56, validation score is 0.55. The metric used is accuracy with a maximum of 1.0.

Answer the following questions providing reasoning:

1) Does this model exhibit high or low variance? Why?

2) Is this model over or underfitting? Why?

3) Would the next step be to increase or decrease complexity? Why?

1) This model exhibits low variance. The model performance on the training set and validation set is similar; and this performance is poor. This model has high bias and is not even flexible enough to fit the data it has already seen.

2) This model is underfitting. This model is a poor predictor both for the training data and for any previously unseen data.

3) The next step would be to increase the complexity. Increasing the complexity would mean introducing more flexibility to the model, which would make it fit more of the training data. Once the complexity is optimal, the model could potentially fit training data and validation data much better.

With a feature matrix X with shape (30, 4), what are the shapes of training X\_train and validation X\_val matrices in a 3-fold cross-validation? Explain how you arrive at your numerical answer.

A 3-fold cross-validation means that we would split our feature matrix into 3 parts, and use one-third for validation, and the rest for training. Since we have 30 samples, 10 of them would be used for validation and 20 of them would be used for training in each trial.

Hence, in each trail of the cross-validation, the shape of X\_train will be (20, 4) and the shape of X\_val will be (10, 4).

A column 'color' in the feature matrix X contains the values ['red', 'green', 'red', 'blue']. A Sklearn OneHotEncoder is used to pre-process this column. Write the rows of resulting matrix below using the column headings:

x0\_blue, x0\_green, x0\_red

+---------+----------+--------+

| x0\_blue | x0\_green | x0\_red |

+---------+----------+--------+

| 0.0 | 0.0 | 1.0 |

| 0.0 | 1.0 | 0.0 |

| 0.0 | 0.0 | 1.0 |

| 1.0 | 0.0 | 0.0 |

+---------+----------+--------+