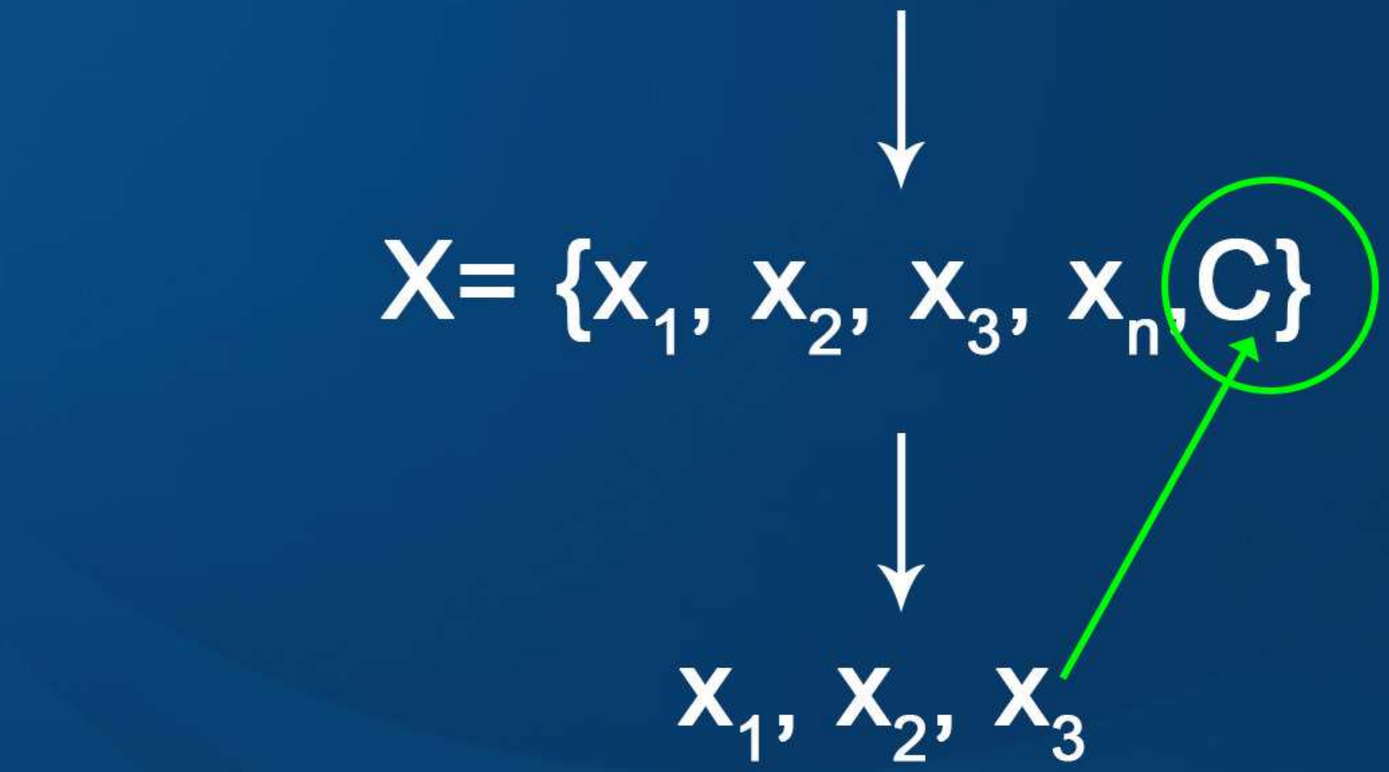
DATA REDUCTION



DATA REDUCTION II

Data Selection Methods

- Data selection methods falls into following three categories:

1. Univariate Techniques

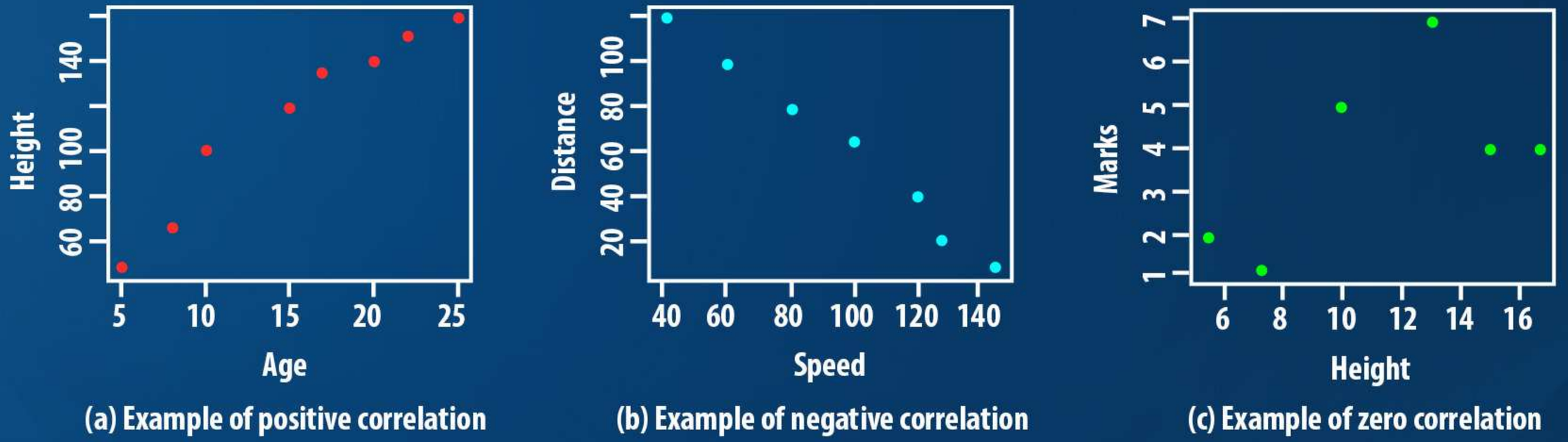
Pearson correlation coefficient

- It is the method to find relationship between two numerical features.
- In order to use Pearson correlation coefficient for data selection, the relationship is measured between each feature present in the data set with the target variable.
- Once the Pearson correlation coefficient is calculated between each pair of feature and the target, the features are arranged in descending order of the score obtained by the test. The top m features are selected to form new feature space.

Chi-Square

- Like Pearson Coefficient, Chi-square test also measures the relationship strength between two variables.
- However, unlike Pearson Coefficient, it is applicable only if two set of variables are discrete in nature.

DATA REDUCTION III

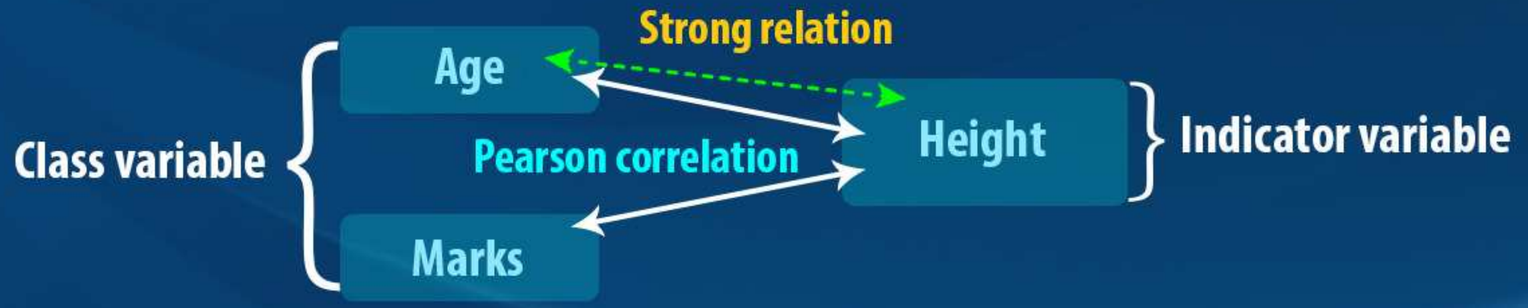


(a) Example of positive correlation

(b) Example of negative correlation

(c) Example of zero correlation

Figure 21: Understanding positive, negative and zero correlation



DATA REDUCTION IV

2. Multivariate Analysis

- Study involving more than one variable is called multivariate analyses.

- Data selection using multivariate analysis comes in two main flavours:

1. Regression

- Regression is a supervised machine learning technique. However, it can also be used for feature selection.
- In simple terms, regression technique represents the data set provided by the user into a linear equation. Where each feature is given a weight that indicates its contribution in deciding the class label. The higher the weight, higher is its importance to the class label.
- Using regression line, we can select top n features with high weights.

2. Wrapper method

- Wrapper method makes use of classification models in Machine learning to find subset of features that improves the end result.

DATA REDUCTION V

Regression for Data Selection- Example

- Suppose for data set in Figure 22, we use regression to discover top two best features out of three given.
- Regression technique on data set produces a regression line of the form

  $C = b + 0.6\,x_1 + 0.9\,x_2 + 0.2\,x_3$.

  Where, b is the slope of the line, quantities in red small circle associated with features represents the weights or contribution of feature in deciding the variable C.
- For the objective of selecting the most important features, we may select x_2, x_1 for their higher weights.

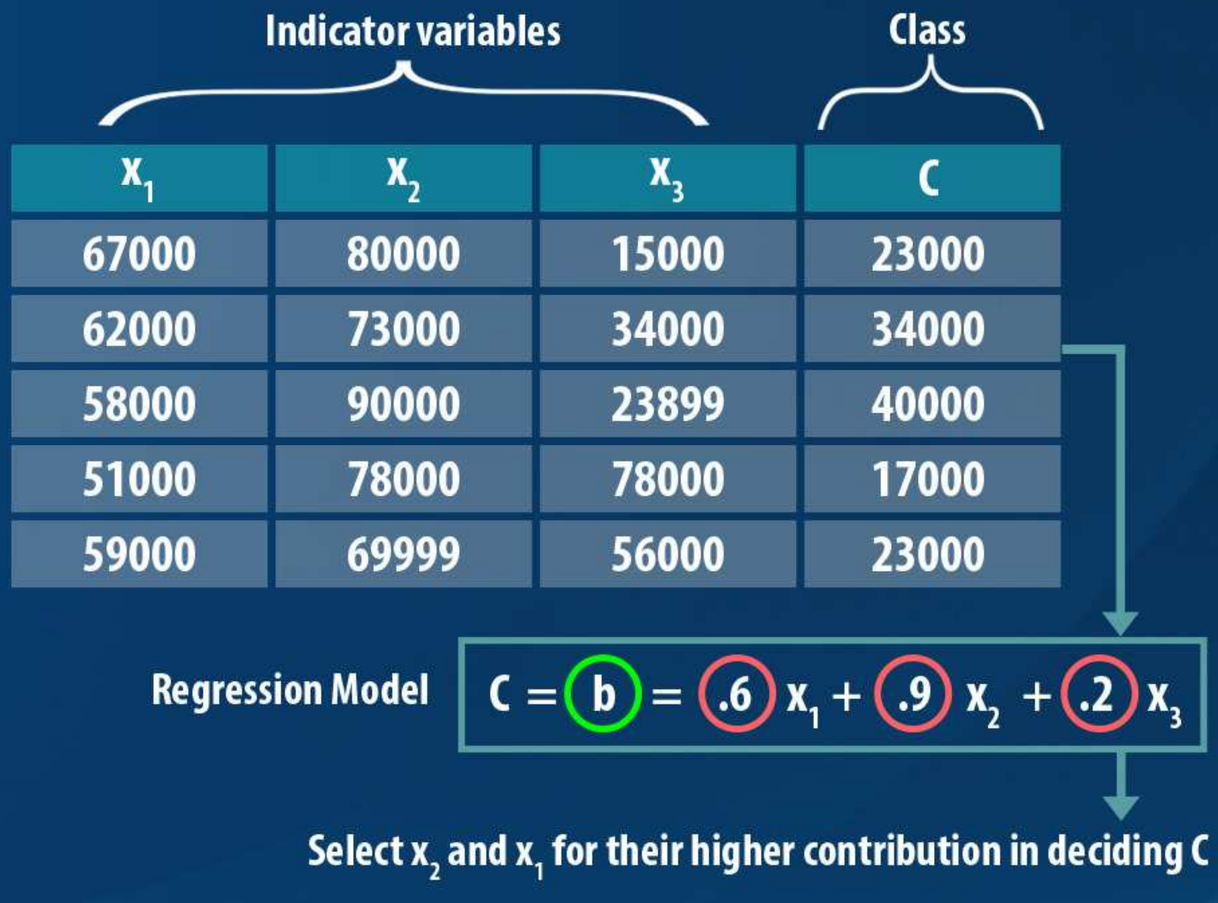


Figure 22: Regression for feature selection.

DATA REDUCTION VI

Wrapper method

1. Backward:

- Here we start with entire feature space, apply Machine Learning model and check its performance.
- In the next step, we delete first feature from the feature space and retest the model performance.
- If the performance is better, we keep the feature else we delete it permanently.
- This process is repeated for subsequent features till we discover a stable performance by the model.

2. Forward:

- In this method, we do not start with entire feature space rather, we test the model performance by picking one feature at a time.
- More features are added to the space if the performance of the model is improving.
- The moment model performance goes down by adding any of the feature, we delete that feature permanently.

DATA REDUCTION VII

Example Illustration-Forward method

- In Figure 23 forward method is explained. Where, suppose we are given three indicator variables (x_1, x_2 and x_3) and one class variable (C).
- As illustrated in the figure, we randomly start by picking one indicator variable say x_1 and test the model performance.
- As in second step, we add x_2 to the feature space and test the model performance. Here in step 2, if the performance of the model goes down in comparison to what we achieved in step 1 then, it indicates that recent addition of variable to the feature space is the cause. In this case, x_2 is deleted and, same process is repeated with other indicator variables given.
- This process is repeated until all indicator variables are exhausted.

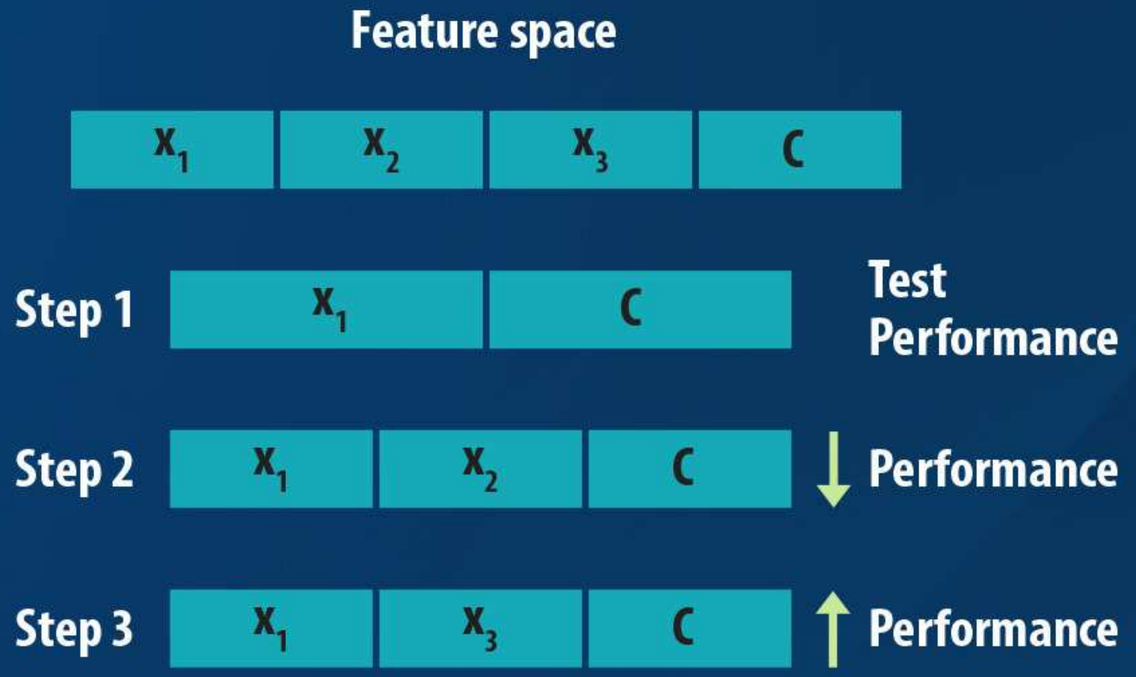


Figure 23: Process of Forward method

DATA REDUCTION VIII

Example Illustration-Backward method

- In Figure 24, backward method is explained. Unlike forward method, here we start with entire feature space and keep deleting the indicator variables at each step in such a way that performance of the model improves.
- As illustrated in the figure, at each step, we are testing the model performance with new combination of indicator variables. For any combination of indicator variables if performance goes down, we delete that variable.

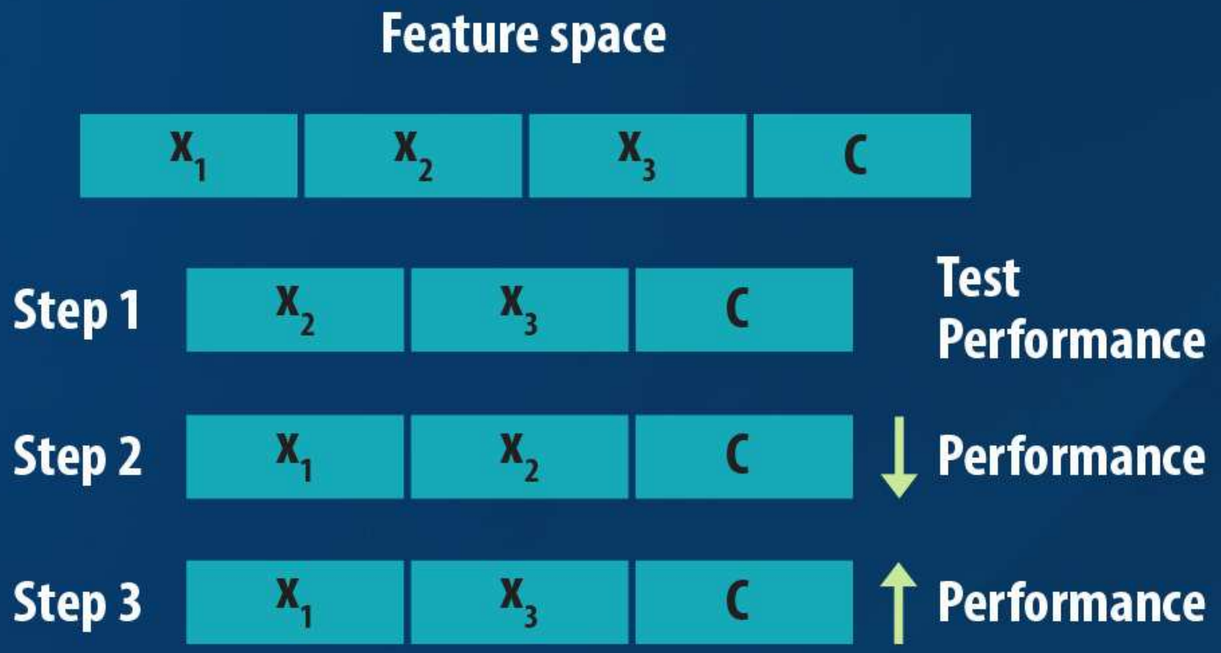


Figure 24: Process of Backward method