

| **Title: Feature engineering in machine learning** |
| --- |

**Aim:** To understand & implement data pre-processing techniques in machine learning

**Expected Outcome of Experiment:**

CO1: Describe the basics of machine learning..

**Books/ Journals/ Websites referred:**

<https://towardsdatascience.com/feature-engineering-techniques-9a57e4901545>

**Feature Engineering Techniques:**

1. Imputation:

- Imputation involves filling in missing values in a dataset. This can be done using statistical measures like mean, median, or mode, or more advanced methods like regression imputation or machine learning algorithms. Example we can fill the empty spaces with the mean value of the values.

2. Handling Outliers:

- Outliers can significantly affect model performance. Techniques include removing outliers, transforming them, or using robust statistical measures that are less sensitive to extreme values. Percentile values can be used. Example top5 and bottom 5 percentile data can be removed.

3. Binning:

- Binning involves grouping a continuous variable into discrete bins or intervals. This can help handle non-linear relationships and reduce the impact of outliers. Eg the dataset can be divided into n parts. Each part can be assigned into n bins. (eg low, medium, high etc)

4. Log Transform:

- Logarithmic transformation is used to reduce the impact of extreme values and make the distribution of a variable more symmetric. It is commonly applied to skewed data. Eg age. difference of 5-10 is more impactful than the age difference of 70-75

5. One-Hot Encoding:

- One-hot encoding is used to convert categorical variables into binary vectors. Each category is represented by a binary digit (0 or 1) in a separate column. This is particularly useful for machine learning algorithms that require numerical input. one column with n different values can be converted to n-1 columns with 1 or 0 as the values

6. Grouping Operations:

- Grouping operations involve aggregating data based on certain criteria. This is often done using functions like mean, sum, count, etc., to create new features that capture the aggregated information.

7. Feature Split:

- Feature split involves breaking down a single feature into multiple features. For example, splitting a date feature into day, month, and year can provide more granular information for analysis.

8. Scaling:

- Scaling ensures that numerical features are on a similar scale. Common methods include Min-Max scaling, Standard scaling (Z-score normalization), and Robust scaling. Eg values like proce in range of thousands can be converted into range in ones.

9. Extracting Date: - Extracting information from date features can provide valuable insights. This may involve extracting the day of the week, month, quarter, or year. Time differences or durations can also be calculated. By calculating the time difference it is possible to store numeric data.

**Conclusion:-** Thus we have performed feature engineering on a data set. We performed various feature engineering techniques on a real world dataset and extracted features from the data.