

Logistic Regression - Student Learning Guide

1. What is Feature Engineering?

Feature engineering is the process of transforming raw data into meaningful inputs for machine learning models.

It helps models learn better by creating cleaner, more informative, and better-scaled features.

2. Log Transformation

Log Transform:

- Used to reduce skewness in data (e.g., age or income).
- Makes highly skewed distributions more normal.
- Example: `df['age_log'] = np.log1p(df['age'])`

3. Binning

Binning:

- Groups continuous variables into categories.
- Useful for segmenting ages, income levels, etc.
- Example: `pd.cut(df['age'], bins=[0,10,20,30,...], labels=['0-10','10-20',...])`

4. One-Hot Encoding

One-Hot Encoding:

- Converts categorical variables into binary columns.
- Ensures models can process string-based categories.
- Example: `pd.get_dummies(df['country'])`

5. Feature Scaling

Feature Scaling:

- Standardization (Z-score): centers data around 0 with unit variance.
- Min-Max Scaling: rescales features to [0, 1] range.
- Important for distance-based or linear models.

Logistic Regression - Student Learning Guide

6. Multicollinearity and VIF

VIF (Variance Inflation Factor):

- Measures multicollinearity among numeric features.
- High VIF (> 10) means a feature is highly linearly correlated with others.
- Drop one of the correlated features to reduce redundancy.