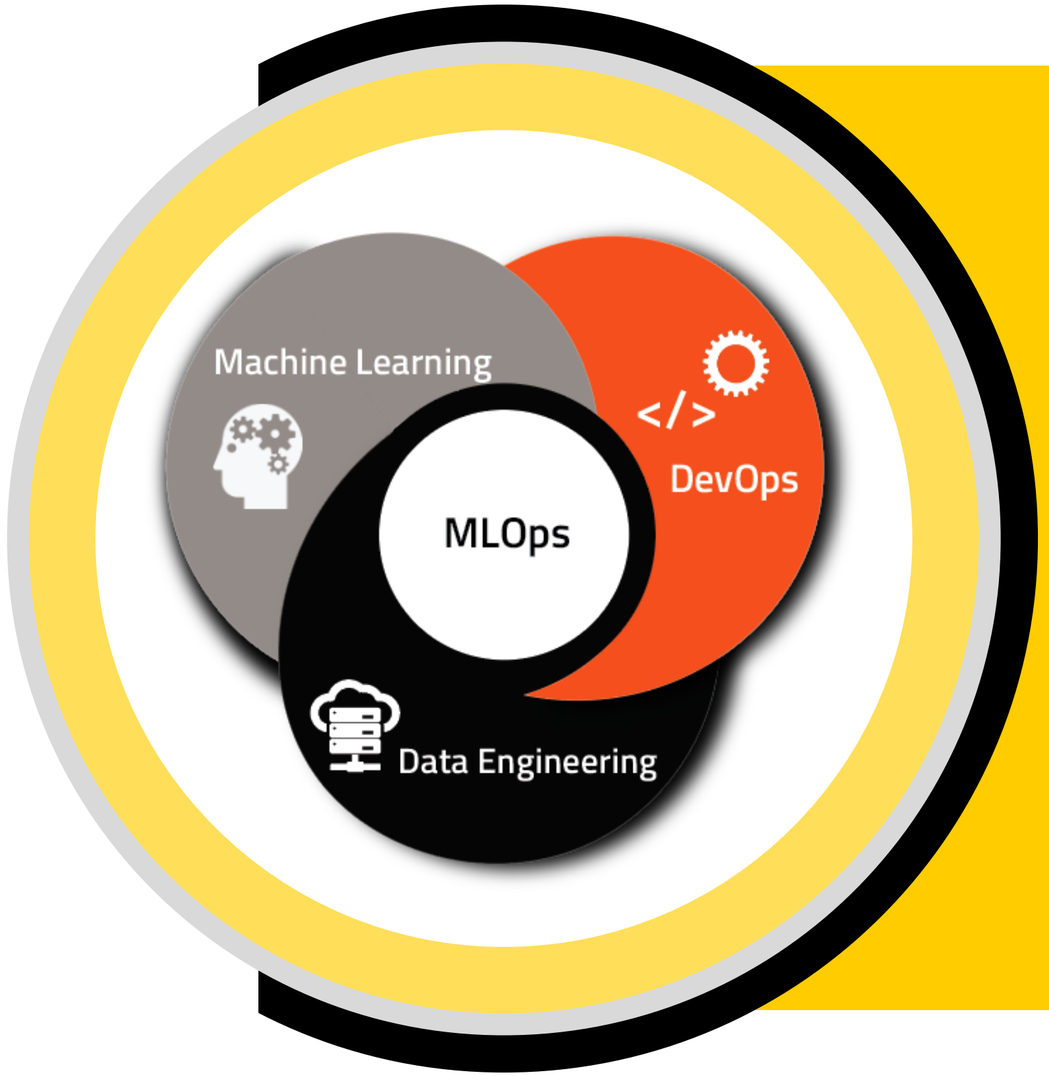
StatusNeo

statusneo.com



MLOps

Operationalizing Machine learning model to production

SESSION 3

StatusNeo

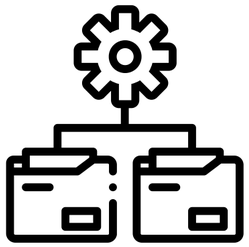
AGENDA

- Research Environment
  - ML Pipeline Overview
  - Feature Engineering
  - Feature Engineering techniques
  - Feature Selection
  - Training a Machine Learning Model
- Lab
  - Code Walk Through for House Price Prediction Dataset
  - Python Library version discussion
  - Learning how to use Pytest
  - Details About Tox

StatusNeo

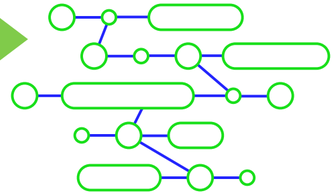
ML Pipeline Overview

Data Gathering

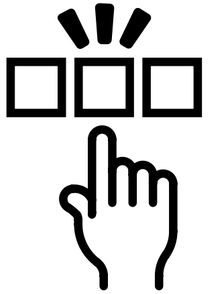


Data Analysis

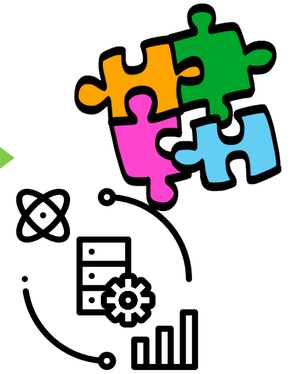
Feature Engineering



Feature Selection



ML Model Building



Model Deployment

StatusNeo

Feature Engineering

Missing Data

Missing values within a variable
Value does not exist, Data collected from survey, or data is calculated from other two variable.

Labels

Strings in categorical variable.
Certainly not with python in scikit-learrn

1. cardinality: High number of labels
2. Rare Labels: infrequent Categories
3. Categories: strings

Distribution

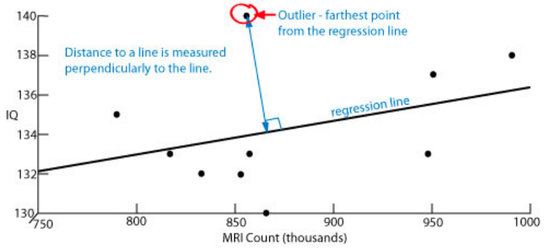
Normal vs skewed

Outliers

Unsual or unexpected values

Outliers

Linear Models



Adaboost

Tremendous weight

Bad Generalization

Feature Magnitude Scale

Machine Learning Models sensitive to feature scale:

- Linear & Logistic Regression
- Neural Networks
- Support Vector Machines
- KNN
- K-means clustering
- Linear Discriminant Analysis
- Principal Component Analysis

Tree based ML Models insensitive to feature side

- Classification and Regression Trees
- Random Forests
- Gradient Boosted Trees

For example in Linear Model :
Variable:
Area: sq kms
No of rooms:
Variable Area will be predominate the no of rooms variable.

StatusNeo

Feature Engineering

- Transform Variables
- Feature Extraction
- Create New Features

StatusNeo

Missing Data Imputations Technique

Numerical Variable

- Mean/ Median Imputation
- Arbitrary Value Imputation
- End of tail imputation

Categorical Variables

- Frequent Category Imputation
- Adding a "missing" category

Both

- Complete case Analysis
- Adding a "missing" indicator
- Random sample Imputation

Categorical Encoding Techniques

Numerical Variable

- One hot encoding
- Count/ frequency encoding
- Ordinal / label encoding

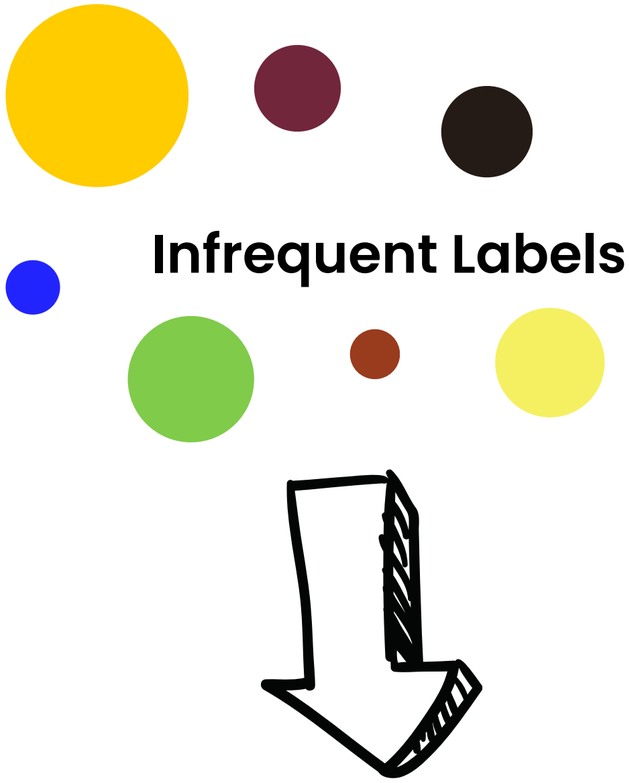
Categorical Variables

- Ordered Label Encoding
- Mean encoding
- Weight of Evidence

Both

- Binary Encoding
- Feature hashing
- Others

StatusNeo

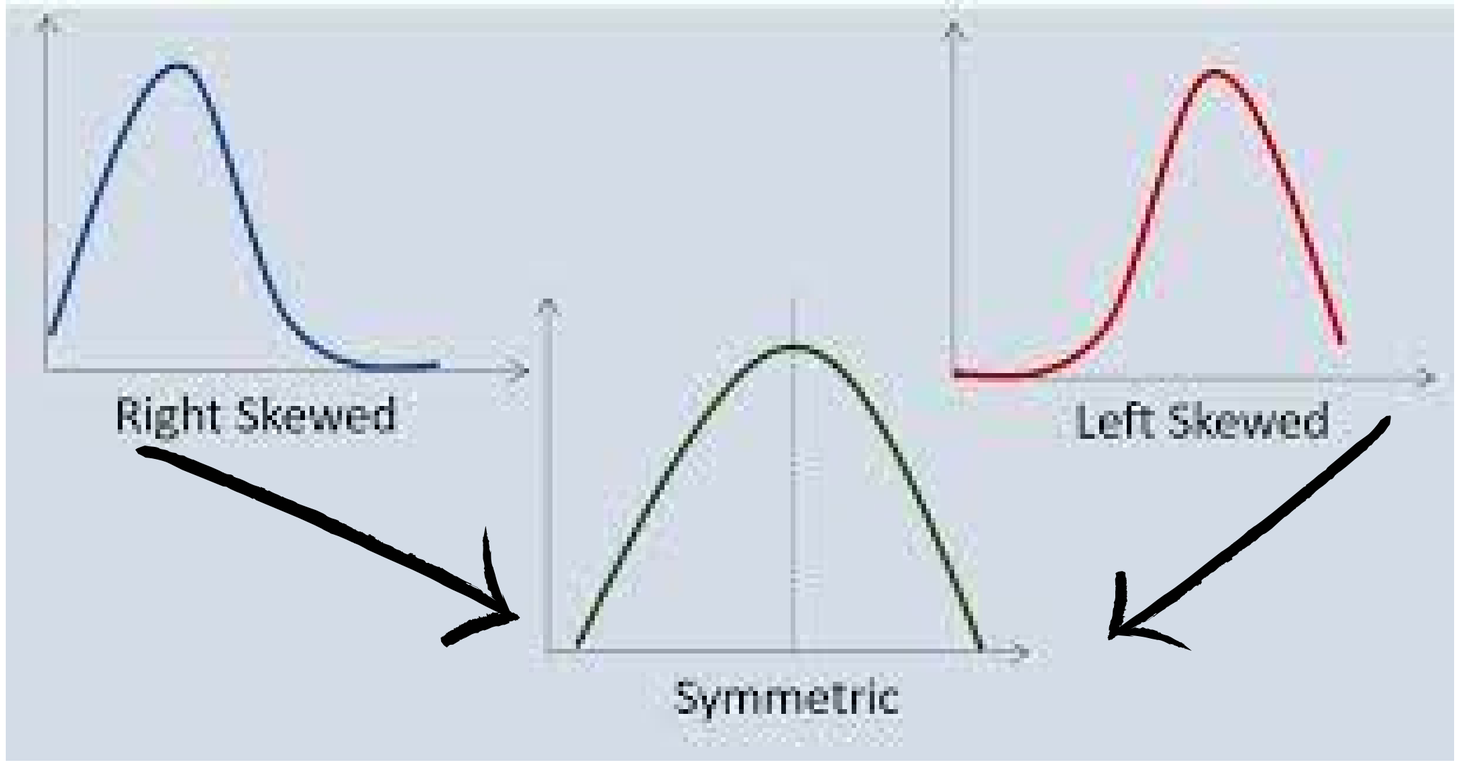


Encoding Techniques:

Rare Label

- ✓ One Hot Encoding of frequent categories
- ✓ Grouping of rare categories

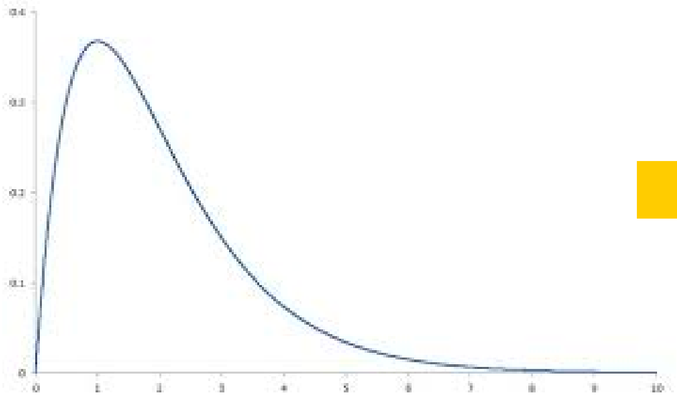
Transform Variable Distributions

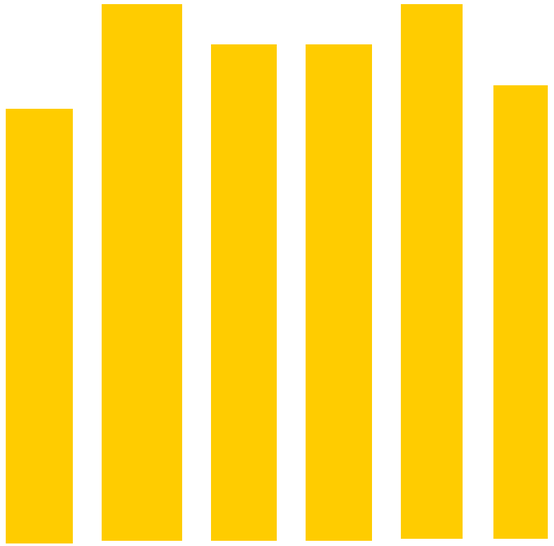


Variable transformation

- Logarithmic
- Exponential
- Reciprocal
- Box Cox
- Yeo-Johnson

Discretisation





Variable transformation

StatusNeo

Different ways of Feature Selection

Here are some ways of selecting the best features out of all the features to increase the model performance as the irrelevant features decrease the model performance of the machine learning or deep learning model.

Filter Methods:
Select features based on statistical measures such as correlation or chi-squared test.For example- Correlation-based Feature Selection, chi2 test, SelectKBest, and ANOVA F-value.

Wrapper Methods:
Select features by evaluating their combinations using a predictive model.For example- Recursive Feature Elimination, Backward Feature Elimination, Forward Feature Selection

Embedded Methods:
Select features by learning their importance during model training.For example- Lasso Regression, Ridge Regression, and Random Forest.

Hybrid Methods:
Combine the strengths of filter and wrapper methods. For Example- SelectFromModel

Dimensionality Reduction Techniques:
Reduce the dimensionality of the dataset and select the most important features. For Example- pca, lda, ica