MDS501
Unit 3: Data Analysis Technique

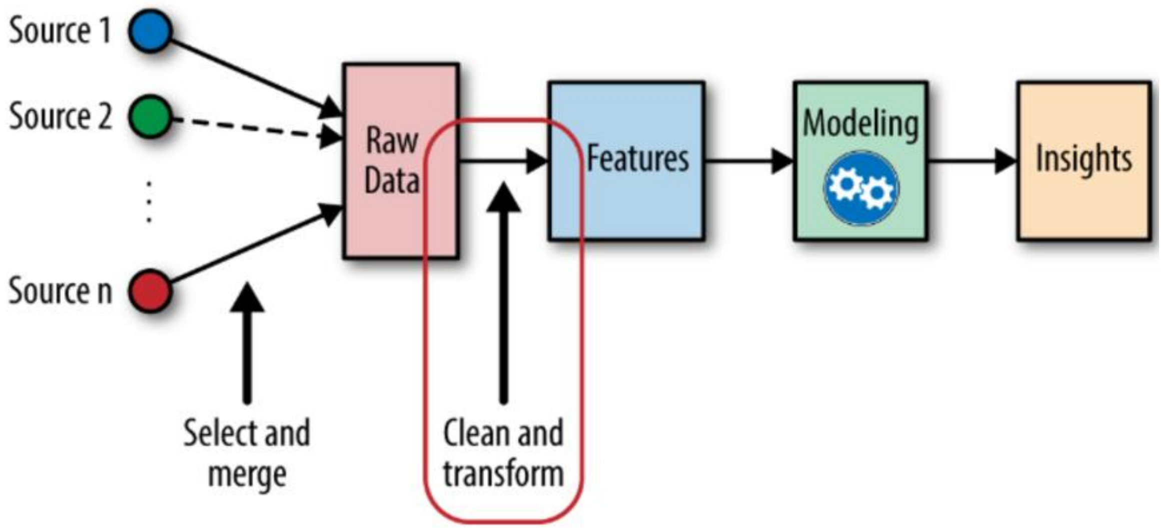
Dipesh Koirala

Outline

- Feature generation and feature selection algorithms: filters, wrappers, decision trees, random forests;
- Predictive data analysis: Introduction to predictive data analysis and its common applications.;
- Time series data analytics

Feature Engineering

- Is the process of *selecting, manipulating and transforming raw data into features*
- Because model performance largely rests on the quality of data used, feature engineering is a crucial preprocessing technique to improve the performance and efficiency of machine learning models.



- It **optimizes ML model performance** by transforming and selecting relevant features.

Feature Engineering

- Say, *you were setting up a gift shop* and your supplier dumps all the toys that you asked for in a room.

- Source:

https://www.analyticsvidhya.com/blog/2021/10/a-beginners-guide-to-feature-engineering-everything-you-need-to-know/

Feature Engineering

- **Includes:**
- **Feature Generation** : creating new features from raw data
- **Feature Transformation** : scaling, normalization, log transforms
- **Feature Selection** : choosing the most important features
- **Feature Extraction** : reducing dimensionality, e.g., PCA
- **Other cleaning steps** : handling missing values & outliers.

Feature Engineering

Why is Feature Engineering Important?

- Improves Model Performance
- Reduces Overfitting
- Enables Interpretability
- Handles Non-Linearity
- Optimizes Computational Efficiency

Feature Generation

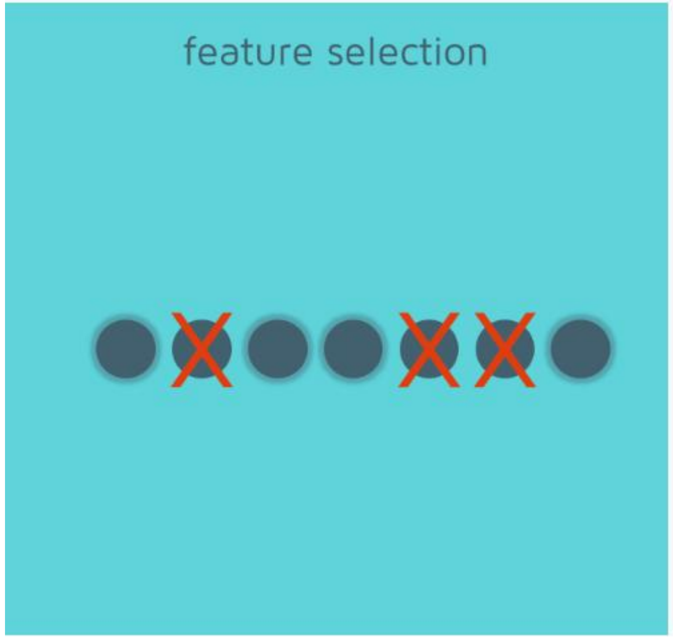
- Is the process of *creating new features from the existing* that better relate to the target.
- Creating a new feature from one or multiple features using derivation, multiplication or addition.

- **E.g.,** Numerical : height * weight => BMI

  Categorical : One-hot encoding, label encoding, target encoding

Feature Selection

- Is the process of selecting the most relevant features of a dataset to use when building and training a machine learning (ML) model.
- not all features are equally valuable or informative, and using irrelevant features can lead to poor model performance and longer training times.



Feature Selection

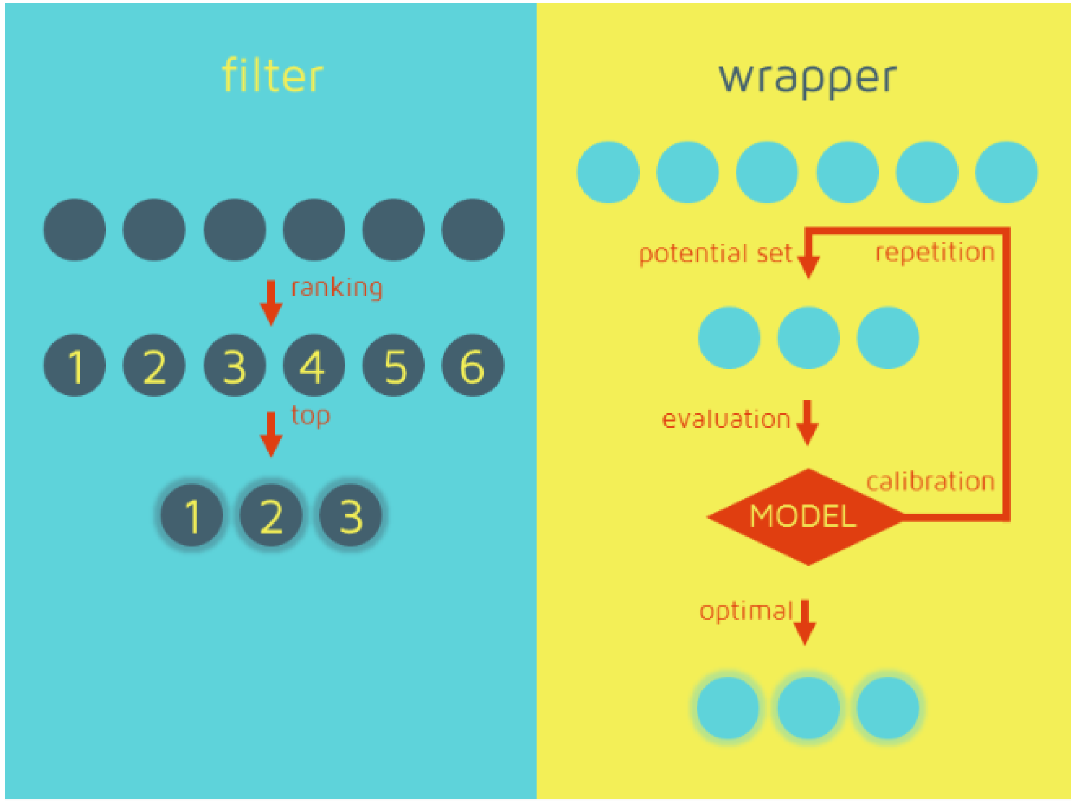
- Too many features may adversely affect the model performance.
- This is because as the number of features increases, it becomes more difficult for the model to learn mappings between features and target **(this is known as the curse of dimensionality).**

Feature Selection Techniques:

- Filters
- Wrappers
- Embedded

Feature Selection

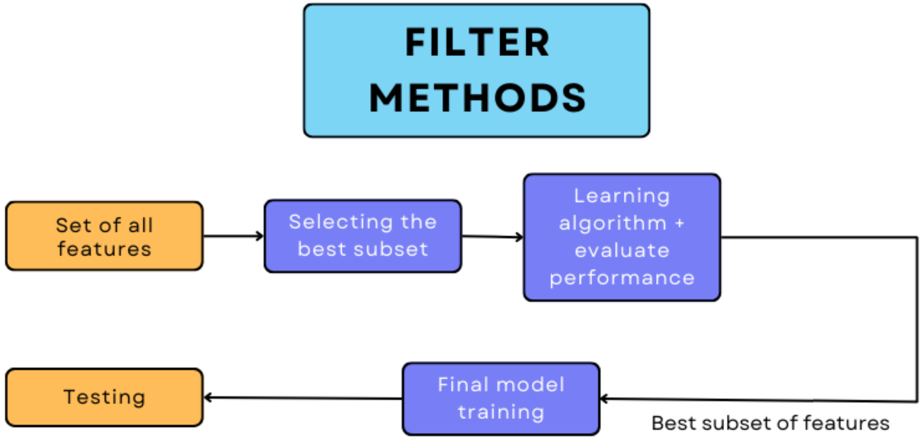
- Filter and Wrapper Methods



Feature Selection

- **Filter methods** are a type of feature selection method that works by selecting features based on some criteria prior to building the model.
- independently analyze each feature based *on **a pre-defined metric.***

- Variance thresholds
- Correlation
- Mutual Information
- Chi – Square test



Feature Selection

Filters Method:

- Variance threshold
- is the method to remove any features that have little to no variation in their values.
- This is because features with low variance do not contribute much information to a model.
- remove categorical features for which all or a majority of the values are the same.

hours_study	hours_TV	hours_sleep	height_cm	grade_level
1	4	10	155	8
2	3	10	151	8
3	4	8	160	8
3	3	8	160	8
3	2	6	156	8
4	3	6	150	8
3	2	8	164	8
4	2	8	151	8
5	1	10	158	8
5	1	10	152	8

Feature Selection

Filters Method:

- **Pearson's correlation coefficient**
- is useful for measuring the linear relationship between two numeric, continuous variables.

- When two features are highly correlated with one another, then keeping just one to be used in the model will be enough because otherwise they provide duplicate information.

Feature Selection

Filters Method:

- **Mutual Information**
- *Mutual information(MI)between two random variables is a non-negative value, which measures the dependency between the variables .*

The mutual information $I(X;Y)$ is calculated as:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right)$$

- If X an Y are independent, mutual information is 0.
- A high Mutual Information indicates feature provides a significant amount of information about the target, and it may be a crucial predictor in a machine learning model.

Feature Selection

Filters Method:

- **Chi-square Test**
- is a statistical test used to assess the relationship between two categorical variables.
- It is used in feature selection to analyze the relationship between a categorical feature and the target variable.

$$X^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$$

- H_0 : Two variables are independent (No association exists between feature and the target)
- H_1 : Two variables are dependent

Feature Selection

Wrappers Method

- Wrappers methods are a type of feature selection method that works by selecting features by training a machine learning model on different subsets.

  - Generate *possible subsets of features from the dataset.*
  - Train a machine learning model on each subset.
  - Evaluate the performance of each model using a specific criterion
  - Choose the subset of features that results in the best model performance

- Wrapper-based approaches can be likened to having a friend *who tries different dish and tells which are the great to eat.*

Feature Selection

Wrappers Method

- **Forward Selection:**
- Start with no features and *add one feature at a time.*
- Train the model with the current set of features and *evaluate its performance using a chosen metric* (e.g., accuracy)
- Stop when the addition of new features does not significantly improve the model's performance or when a maximum number of features is reached

- **Backward Elimination:**
- Start with *all features and remove one feature at a time*, eliminating the one that worsens the model the least.
- Typically uses statistical tests (like p-values in regression) to determine feature significance.

Feature Selection

Wrappers Method

	Pragnency	Glucose	Blod Pressure	Skin Thikness	Insulin	BMI	DFP	Age	Diabetes
0	1	85	66	29	0	26.6	0.351	31	0
1	8	183	64	0	0	23.3	0.672	32	1
2	1	89	66	23	94	28.1	0.167	21	0
3	0	137	40	35	168	43.1	2.288	33	1
4	5	116	74	0	0	25.6	0.201	30	0

Feature Selection

Wrappers Method

- **Recursive Feature Elimination**
- Repeatedly build the model and *remove the least important features*
- Uses model-based feature importance (e.g., coefficients in linear models for Evaluation)

- **Exhaustive Feature Selection**
- This is a brute-force evaluation of each feature subset.
- This means it tries *every possible combination of the variables and returns the best-performing subset.*

Feature Selection

Decision Trees

- Embedded Method
- Computes gini impurity or IG for classification and Variance Reduction for regression to select splits.
- Decision Trees can be used to compute feature importance that computes how much feature contributes to reducing impurity

```
dmodel = DecisionTreeClassifier()
dmodel.fit(X,y)

importance = dmodel.feature_importances_
print(importance)
```

```
[0.05465348 0.32419382 0.11036545 0.01257586 0.04193103 0.21131163
 0.13368582 0.11128291]
```

Feature Selection

Random Forests

- Embedded Methods
- Feature importance is averaged over all trees.

Predictive Data Analysis

- Predictive analytics is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.

- Financial markets (stock prices)
- Demand forecasting
- Risk assessment
- Weather forecasting etc.

Predictive Data Analysis

- Predictive analytics models are designed to assess historical data, discover patterns, observe trends, and use that information to predict future trends.

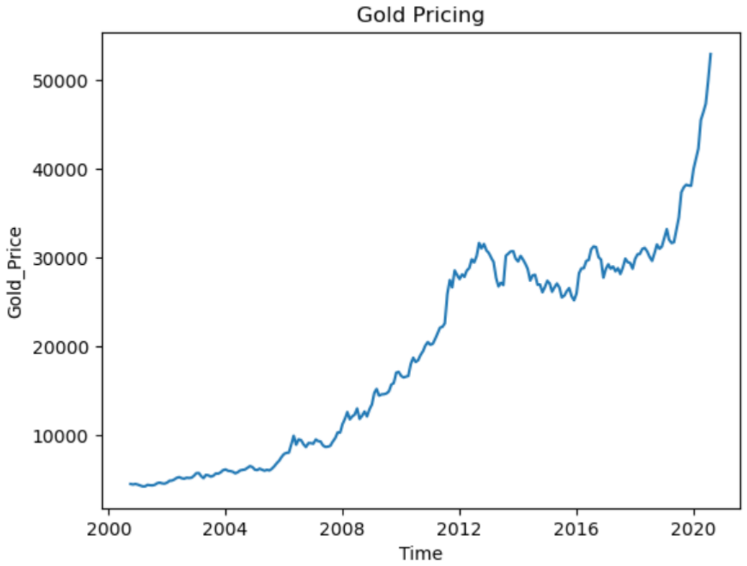
Types of predictive modeling:

- Classification Models
- Regression Models
- Clustering Models
- Time Series Models

- **Note:** Regression, Classification and Clustering Models are already covered.

Time Series Data Analytics

- Time series data is a sequence of data points collected or recorded at specific time intervals.
- E.g., daily stock prices, monthly sales figures, weather data.



- Time Series Analytics involves analyzing and forecasting data points collected or recorded sequentially over times.

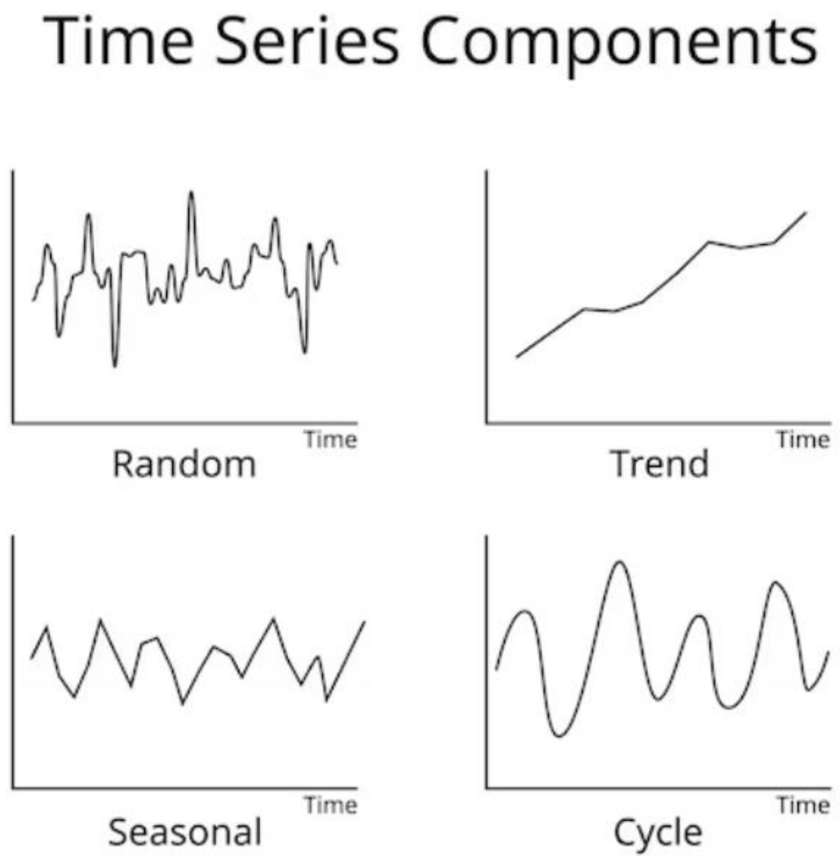
- Financial markets (stock prices)
- Weather forecasting
- Sales predictions
- Economic indicators

Time Series Data Analytics

Components of Time Series Data

- Time series data is generally comprised of different components that characterize the patterns and behavior of the data over time.

- Trends
- Seasonality
- Cycles
- Noise



Time Series Data Analytics

Components of Time Series Data

- **Trends:**
- Long-term increases, decreases, or stationary movement
- Trends indicate the long-term movement in the data and can reveal overall growth or decline. For e.g., e-commerce sales may show an upward trend over the last five years.

- **Seasonality:**
- refers to predictable patterns that recur regularly, like yearly retail spikes during the holiday season.
- Seasonal components exhibit fluctuations fixed in timing, direction, and magnitude. For instance, electricity usage may surge every summer as people turn on their air conditioners.

Time Series Data Analytics

Components of Time Series Data

- **Cycles:**
- demonstrate fluctuations that do not have a fixed period, such as economic expansions and recessions.
- Business cycles that oscillate between growth and decline are an example.

- **Noise:**
- noise encompasses the residual variability in the data that the other components cannot explain.

- Source: https://www.sigmacomputing.com/blog/what-is-time-series-analysis

Time Series Data Analytics

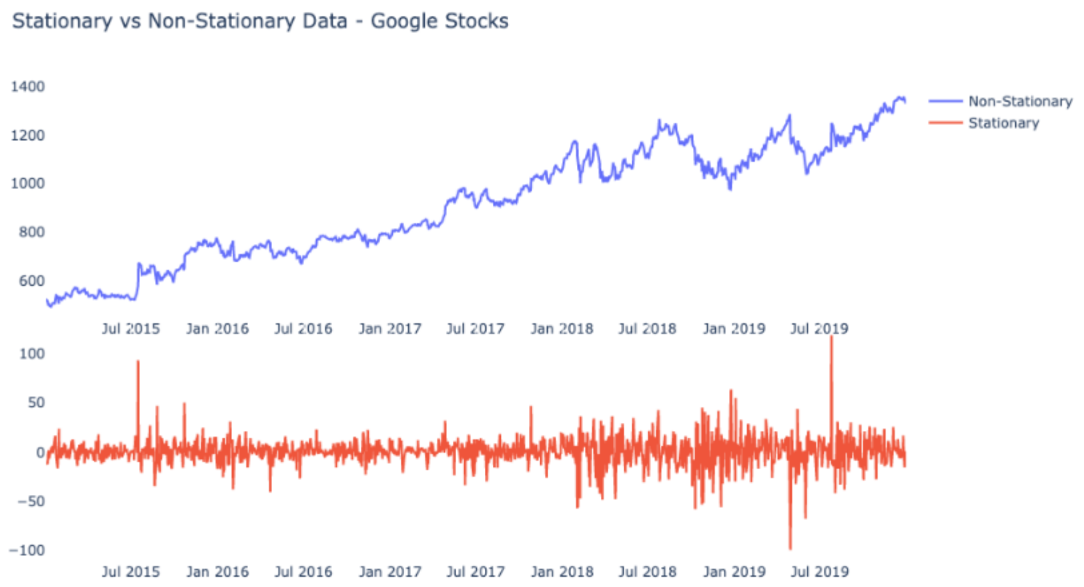
Models

- AR
- MA
- ARIMA, SARIMA, ARIMAX
- Exponential Smoothing
- Prophet
- LSTM, Transformers

Time Series Data Analytics

Stationarity

- Constant Mean (No trend)
- Constant Variance
- Constant Autocorrelation (Seasonality/patterns should not change)



End of Unit 3

Dipesh Koirala