EMBA Program
MB-511

Data Science for Managerial Decisions (MB 511)

Instructor
Anant Prakash Awasthi

Data Science for Managerial Decisions (MB 511)

EMBA Program
MB-511

Program Overview

- Introduction to Data Science
- Information Technology An Overview
- Applications of Data Science in various fields
- MIS and Control Systems
- Data Collection and Data Pre-Processing
- Building Information Systems
- Support Systems for Management Decisions

Data Collection and Data Pre-Processing

EMBA Program
MB-511

- Introduction to Data Collection
- Methods of Data Collection in Management
- Designing Data Collection Instruments
- Sampling Techniques
- Data Collection Planning and Management
- Understanding Data Pre-Processing
- Data Cleaning Techniques
- Quality Assurance in Data Pre-Processing

Data Collection and Data Pre-Processing

Data Cleaning Techniques

EMBA Program
MB-511

Data cleaning is a crucial step in the data preprocessing pipeline, ensuring that datasets are accurate, consistent, and ready for analysis.

- Handling missing values
- Handling duplicate data
- Data transformation
- Outlier detection and removal
- Data formatting
- Handling inconsistent data
- Handling categorical data
- Feature engineering
- Data anonymization and masking
- Data validation

Data Collection and Data Pre-Processing

Handling missing values

EMBA Program
MB-511

Statistically, missing values refer to the absence of data points or observations in a dataset. Handling missing values is essential because they can affect the statistical analysis and modeling process, leading to biased results or reduced accuracy.

Handling Missing Values:

Deletion: Remove rows or columns with missing values.

Imputation: Fill missing values with a statistical measure like mean, median, or mode.

Prediction: Predict missing values based on other features using algorithms like K-nearest neighbors or regression.

Data Collection and Data Pre-Processing

Managing Outliers

EMBA Program
MB-511

Outliers are data points that deviate significantly from the rest of the dataset, often due to errors in data collection, measurement variability, or genuinely rare phenomena. Detecting and treating outliers is essential in statistical analysis and machine learning to prevent them from unduly influencing results.

Outlier Detection:

Outlier detection involves identifying data points that are unusually distant from the majority of the dataset. Various statistical methods and visualization techniques can be used for outlier detection, including Z-Score, Interquartile Range (IQR), Mahalanobis Distance, box plots, and scatter plots. These methods help highlight observations that fall outside expected ranges or patterns.

Data Collection and Data Pre-Processing

Managing Outliers

EMBA Program
MB-511

Once outliers are identified, several approaches can be used for treatment:

- Deletion: Outliers can be removed from the dataset. However, this approach should be used cautiously as it may result in loss of valuable information, especially in smaller datasets.
- Winsorization: This method involves replacing extreme values with less extreme ones. Winsorization caps the outliers at a specified percentile (e.g., replacing values above the 99th percentile with the value at the 99th percentile).
- Transformation: Data transformation techniques such as logarithmic, square root, or inverse transformations can be applied to mitigate the impact of outliers. These transformations can make the data more normally distributed and reduce the influence of extreme values.
- Modeling Robustness: Using statistical methods or machine learning algorithms that are less sensitive to outliers can also be an effective approach. Robust methods include robust regression techniques or non-parametric statistical tests.

Data Collection and Data Pre-Processing

Handling Categorical Data

EMBA Program
MB-511

Categorical variable encoding is necessary in machine learning because many algorithms require numerical input data. Categorical variables, which represent qualitative data with distinct categories or groups, need to be converted into numerical format for the algorithms to process them effectively.

One-Hot Encoding:

- Represents each category as a binary vector where each category is transformed into a binary column.
- Widely used for categorical variables with low cardinality (few unique categories).

Label Encoding:

- Assigns a unique integer label to each category.
- Suitable for ordinal categorical variables where there is an inherent order among the categories.

Ordinal Encoding:

- Similar to label encoding but explicitly defines the order of the categories.
- Assigns integer labels based on the predefined order of the categories.

Data Collection and Data Pre-Processing

Feature engineering

EMBA Program
MB-511

Feature engineering is the process of creating new features or transforming existing ones to improve the performance of machine learning models. It involves selecting, extracting, and modifying features from raw data to make them more informative and suitable for the specific task at hand.

Importance of Feature Engineering:

- Good features are essential for building accurate and robust machine learning models.
- Feature engineering can help models better capture the underlying patterns and relationships in the data, leading to improved performance.

Data Collection and Data Pre-Processing

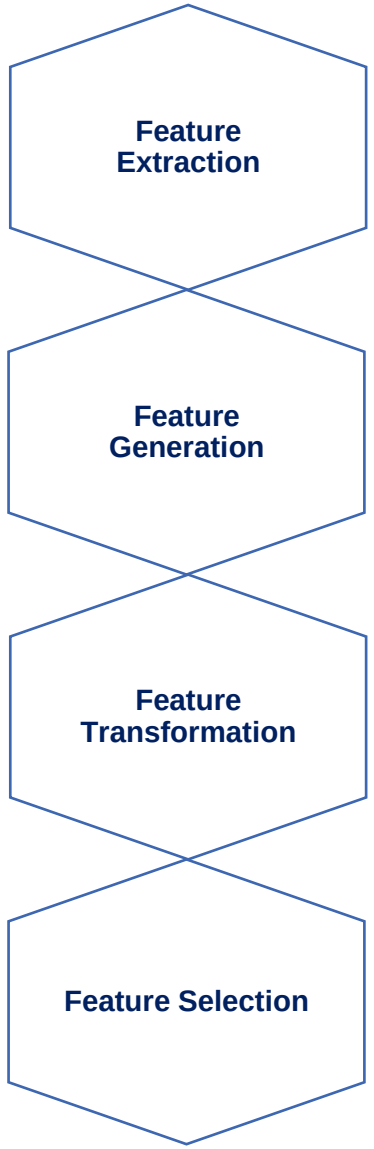
Feature engineering – Types

Feature Extraction: Creating new features from existing ones to capture additional information or patterns. Techniques include principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and autoencoders.

Feature Generation: Creating new features based on domain knowledge, intuition, or insights gained from exploratory data analysis (EDA). This could involve creating interaction terms, polynomial features, or aggregating information from multiple variables.

Feature Transformation: Modifying the existing features to make them more suitable for the model. This includes scaling, normalization, binning, and handling skewness or outliers.

Feature Selection: Choosing the most relevant features from the dataset to reduce dimensionality and computational complexity.



Data Collection and Data Pre-Processing

Feature engineering – Common Techniques

- Handling Categorical Variables: Encoding categorical variables into numerical format using techniques like one-hot encoding, label encoding, or target encoding.
- Handling Missing Values: Imputing missing values using mean, median, mode, or more sophisticated techniques such as predictive modeling or multiple imputation.
- Scaling and Normalization: Scaling numerical features to a similar range or normalizing them to have a mean of 0 and a standard deviation of 1 to improve model convergence and performance.
- Creating Interaction Terms: Multiplying or combining two or more features to capture interactions or nonlinear relationships that may be important for the model.
- Dimensionality Reduction: Using techniques like PCA or feature selection algorithms to reduce the number of features while preserving as much relevant information as possible.

EMBA Program
MB-511

Data Collection and Data Pre-Processing

Quality Assurance in Data Pre-Processing

EMBA Program
MB-511

Ensuring the accuracy and reliability of data in an information system is crucial for making informed decisions, maintaining trust in the system, and avoiding costly errors. Ensuring the accuracy and reliability of data in an information system is of paramount importance for several reasons.

- Improved Data Quality
- Reduced Errors and Bias
- Enhanced Data Consistency
- Increased Trust and Confidence
- Cost Savings
- Compliance with Regulations
- Streamlined Data Analysis
- Facilitated Decision Making

EMBA Program
MB-511

Case Study

- **Missing Value imputation**
- **Outlier Treatment**
- **Categorical Encoding Methods**

EMBA Program
MB-511

Have a question?

Feel Free to Reach out at

- +91-88846-92929 (WhatsApp)
- anant.awasthi@outlook.com (E-Mail)