Vinoothna Bavireddy

Feature Engineering

A feature in machine learning is defined as a column or an attribute of the data. Feature engineering is the process by which knowledge of data is used to construct explanatory variables, features, that can be used to train a predictive model(Rencberoglu,2019). Feature engineering primarily focuses on extracting important attributes in the raw data and engineering them into formats that will significantly improve the performance of machine learning models.

Unprocessed data come with issues like missing values and information redundancy. These problems can be solved by data cleaning, normalization and feature selection.

Engineering Features

- **Handling missing values:**
    1) For Categorical data
        a) Mode filling: Fill missing values with the most popular/frequent/modal class.
        b) Temporal filling (forward or backward fill): Fill missing values with the preceding value (top-down) or with the succeeding value (bottom-up).
        c) Encoding and fill: In this method, you can encode the values using different strategies, and then fill with either the mean, mode, or the median.
    2) For Numerical Data
        a) Filling with mean, mode, or median.
        b) Temporal filling (backward or forward filling).
        c) Use machine learning models: Train a machine learning model to learn the most appropriate fill values.
- **Normalization of Data:**

1) StandardScaler: Standardizing the attributes by subtracting the mean and scaling to the unit variance.
2) RobustScaler: Using statistical methods to handle outliers.
3) MinMaxScaler: Normalizing attributes by putting them in a certain range.

- **Splitting Features:**
  This involves binning data into attributes that can help the algorithm predict patterns. For example: Breaking sales data into quarters (Q1,Q2,Q3,Q4) to measure performance or identify seasons with high/ low sales.
- **Data Enrichment**
  Data enrichment is nothing but adding data to our existing dataset from external data sources that can help increase the features in our dataset and improve the predictions. For example using census data to healthcare data to understand how social determinants of health impact hospital readmission rates.
- **Feature Transformation**

This involves aggregating or grouping data to generate new features/ attributes.This will help identify trends in the model and how two features can be combined to provide new insights.
For example: horsepower and price of a car.

Attribute Selection
To understand which features are to be selected we need to first understand what the dependent (the feature being predicted) and independent variables in the dataset.
Methods to select attributes include:

1) Correlation
   The variables that are to be selected can be identified by calculating the correlation coefficient. By understanding the type of correlation (positive/negative/ no) the different attributes that can affect our dependent variable can be selected.
2) Near Zero Variance
   Some attributes have constant and do not affect our variable of interest. Such columns can be eliminated from our model using certain functions. For example: nzv() in R.

Some other techniques include **Principal component analysis (PCA) and Linear discriminant analysis (LDA)** to maximize the variance in the dataset.

Conclusion:

“GARBAGE IN GARBAGE OUT” - Feature engineering will help clean and process data to produce meaningful insights.

References:

1. Hannah Patrick. (2019, February 19). The importance of feature engineering and selection. Rittman Mead. https://www.rittmanmead.com/blog/2019/02/the-importance-of-feature-engineering-and-selection/
2. Odegua, R. (2020, May 9). A practical guide to feature engineering in Python. Medium. https://heartbeat.fritz.ai/a-practical-guide-to-feature-engineering-in-python-8326e40747c8
3. Rençberoğlu, E. (2019, April 1). Fundamental techniques of feature engineering for machine learning. Medium. https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114