Data Exploration and Preparation

Presented By

Aditya Joshi

Assistant Professor

Graphic Era Deemed to be University

Steps of Data Exploration and Preparation

Steps involved to understand, clean and prepare your data for building your predictive model.

1. Variable Identification
2. Univariate analysis
3. Bivariate analysis
4. Missing values treatment
5. Outliers treatment
6. Variable transformation
7. Variable creation

Feature Engineering

Exploratory analysis, Missing value treatment, identifying outliers and correct data inconsistencies are all the part of process of data preparation and exploration.

Variable identification:

Type of variable

Datatype: char, numeric

Variable category: Quantitative, Qualitative

Identifying Predictor(**input**):rain, Previous exam marks, income and Target(**output**): Play Cricket, Eligible for next sem, Loan Approved.

Univariate Analysis: explore variable one by one either it is qualitative or quantitative

Bivariate Analysis: Relation between two variables.

Missing Value treatment: Missing values in Training dataset can reduce the power /fit of a model can lead to biased model. It can do wrong prediction.

It can be handle by Manual Entry, mean, mode, median, most probable value, constant, delete.

Outliers: Observation that appears far away and diverge from an overall pattern in a sample.

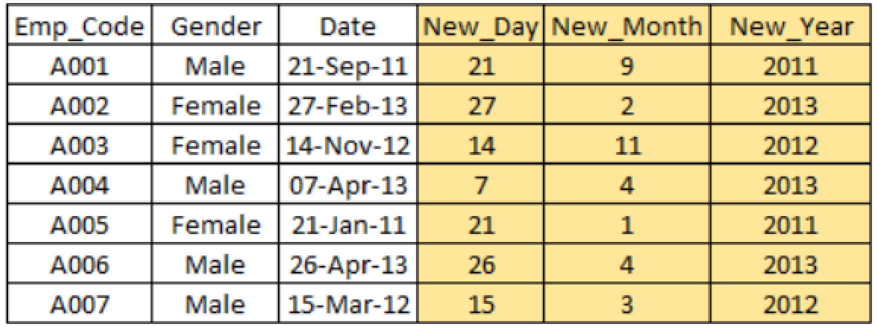
A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data. For example in the scores 25,29,3,32,85,33,27,28 both 3 and 85 are "outliers".

Someone like Elon Musk who has a net worth in the billions of dollars would be considered an outlier in terms of annual income.

Feature Engineering: it’s a science of extracting more information from existing data and not adding new data, we can make data more useful with existing data.

In data modelling, transformation refers to the replacement of a variable by a function. For instance, replacing a variable x by the square / cube root or logarithm x is a transformation.

Feature / Variable creation is a process to generate a new variables / features based on existing variable(s). For example, say, we have date(dd-mm-yy) as an input variable in a data set. We can generate new variables like day, month, year, week, weekday that may have better relationship with target variable. This step is used to highlight the hidden relationship in a variable



Emp_Code	Gender	Date	New_Day	New_Month	New_Year
A001	Male	21-Sep-11	21	9	2011
A002	Female	27-Feb-13	27	2	2013
A003	Female	14-Nov-12	14	11	2012
A004	Male	07-Apr-13	7	4	2013
A005	Female	21-Jan-11	21	1	2011
A006	Male	26-Apr-13	26	4	2013
A007	Male	15-Mar-12	15	3	2012