



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Data wrangling and Feature Engineering

Feature Engineering

Dr. Ramakrishna Dantu

Associate Professor, BITS Pilani

Introduction to Data Science



Disclaimer and Acknowledgement



Disclaimer

- The content for these slides has been obtained from books and various other source on the Internet
- I here by acknowledge all the contributors for their material and inputs.
- I have provided source information wherever necessary
- I have added and modified the content to suit the requirements of the course

Introduction to Data Science



Data wrangling and Feature Engineering – Part-1

- Data cleaning
- Data Aggregation, Sampling,
- Handling Numeric Data
 - Discretization, Binarization
 - Normalization
 - Data Smoothing
- Dealing with textual Data
- Managing Categorical Attributes
 - Transforming Categorical to Numerical Values
 - Encoding techniques
- Feature Engineering
 - Feature Extraction (Dimensionality Reduction)
 - Feature Construction
 - Feature Subset selection
 - Filter methods
 - Wrapper methods
 - Embedded methods
 - Feature Learning
- Case Study involving FE tasks





Feature Engineering

Feature Engineering



Motivation for Feature Engineering

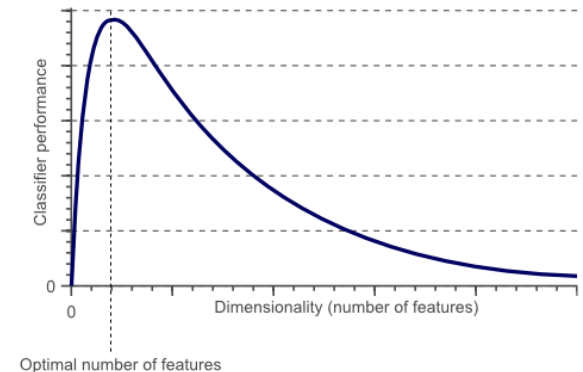
- Real Estate price prediction
 - Problem Definition:
 - Given the characteristics in terms of landmark, facilities, complaints etc. of property predict the sale price.
 - Data Set Description:
 - The dataset has 108 Features and 6664 instances. The dependent variable is the sale price.
- Predicting returning customers
 - Given a data set of customer purchasing data gathered over a year, predict which customers will return to make a purchase.
 - The set contains around 350 million rows and 11 features
- Predicting whether a patient gets cancer or not
 - Given a data set of cancer patients and the tests they have undergone to predict whether a patient will have cancer or not
 - The set contains around 100 million rows and 35 features

Feature Engineering



Motivation for Feature Engineering

- Curse of Dimensionality
 - As the number of features increases, performance of ML algorithm deteriorates
 - As the number of variables increases, the required number of samples (to achieve the same accuracy) grows exponentially
- Hughes Phenomenon
 - *"given fixed number of data points, performance of a regressor or a classifier first increases but later decreases as the number of dimensions of the data increases"*
- Reasons for this phenomenon
 - Redundant Features – Carry same information in some other form (no value addition)
 - Correlation between features – the presence of one feature influence the other
 - Irrelevant Features - those that are simply unnecessary
- So, one motivation for feature engineering is dimensionality reduction
- The other motivation is to improve a learning algorithm's performance



Feature Engineering



So, what is Feature Engineering (FE)?

- We all know that in any system, if garbage in, then garbage out
- Our model will only be capable of learning properly if the training data contains enough relevant features and not too many irrelevant ones
- Success of a ML project depends, in part, on coming up with a good set of features to train on
- FE is the process of selecting and extracting useful, predictive signals from data.
- FE involves the following steps:
 - Feature selection (selecting the most useful features to train on among existing features)
 - Feature extraction (combining existing features to produce a more useful one)
 - Creating new features by gathering new data
- The goal is to create a set of features that best represent the information contained in the data
 - This produces a simpler model that generalizes well to future observations.

Feature Engineering



Feature

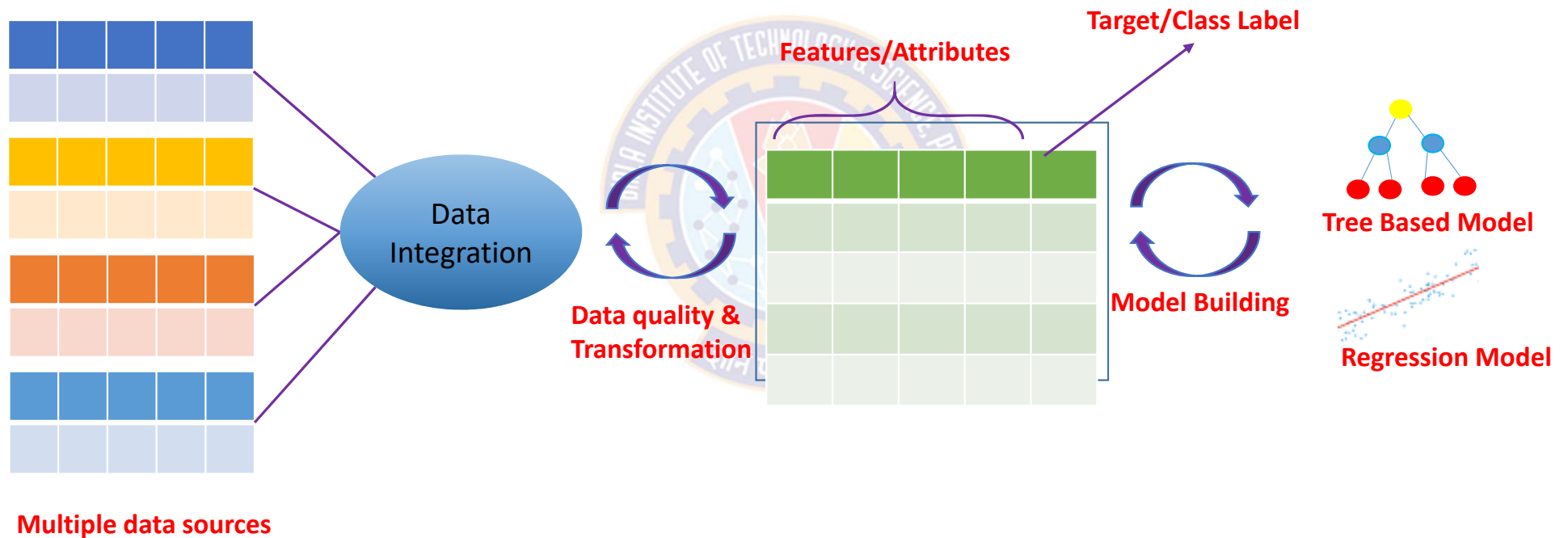
- Feature is a property (attribute) of an object (Entity) under study
- Features are the building blocks of datasets

Building Area	Common Area	Type of Flooring	DistanceFrom BusDepot	Sale Price per square feet
11345	350	Marble	16503.22	6,715
2000	1334	Vitrified Tiles	16321.19	3,230
2544	924	Wood Vitrified Tiles	15619.92	6,588

Feature Engineering



Machine Learning Flow





Feature Creation

Feature Creation



Overview

- Involves creating new attributes that can capture important information in a data set much more efficiently than the original attributes
- Three general techniques:
 - Feature Extraction
 - Often used in data where there are many variables but small samples of data.
 - Example: extracting edges from images
 - Example: analysis of written texts where we want to extract opinions
 - Mapping Data to New Space (Mathematical transformation)
 - Extensively used in image processing
 - Example: Fourier and wavelet analysis
 - Feature Construction
 - Example: dividing mass by volume to get density

Feature Creation



Feature Extraction

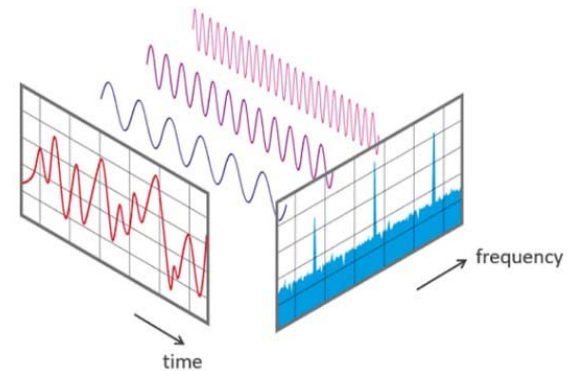
- Involves creation of new set of attributes from the original raw data
- Example: Bag of Words
 - Bag-of-Words widely used in natural language processing
 - Here, words (or features) are extracted from a sentence, document, website, etc.
 - These features are classified into frequency of use
 - So in this whole process, feature extraction is one of the most important parts.
- Example: Image Processing - Classifying the images containing face
 - Features can be extracted from the content of an image, such as
 - Color, Texture, Shape, Position, Dominant edges of image items
 - Set of animal photographs which contains face
 - Raw data is set of pixels
 - Higher level features are extracted out of it like presence or absence faces or other body parts
 - Many techniques, including feature extraction as well as algorithms, are applied to detect features such as shapes, edges, or motion in a digital image or video to process them

Feature Creation



Mapping Data to New Space

- The Fourier Transform is an important image processing tool which is used to decompose an image into its sine and cosine components
- The output of the transformation represents the image in the Fourier or frequency domain, while the input image is the spatial domain equivalent
- In the Fourier domain image, each point represents a particular frequency contained in the spatial domain image.
- The Fourier Transform is used in a wide range of applications, such as reducing blurring and noise, image filtering, image reconstruction and image compression (such as JPEG format)



Feature Creation



Feature Construction

- Create dummy features: Often used to convert categorical variable into numerical variables




Customer_ID	Gender	Paymet_Method	Online Banking	Credit Card	Debit Card
C001	FEMALE	Online Banking	1	0	0
C002	MALE	Online Banking	1	0	0
C003	FEMALE	Credit Card	0	1	0
C004	MALE	Debit Card	0	0	1

Feature Creation



Feature Construction

- Creating Derived Features
 - Involves creating a new feature using data from existing features
 - Calculating session duration



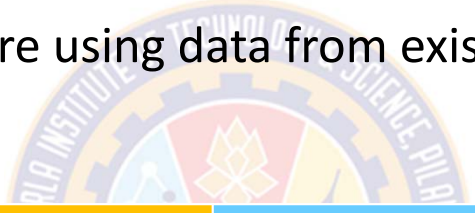
Customer_ID	Gender	Session_Begin	Session_End	Session_Duration
C001	FEMALE	15-06-2019 10.30	15-06-2019 11.15	45
C002	MALE	13-06-2019 8.00	13-06-2019 8.03	3
C003	FEMALE	2-06-2019 16.25	2-06-2019 18.35	126
C004	MALE	1-06-2019 11.20	1-06-2019 1.00	100

Feature Creation



Feature Construction

- Creating Derived Features
 - Involves creating a new feature using data from existing features
 - Calculating price per sqft



Area	Price (Rs)	Price/Sft (Rs)
1800	81,00,000	4500
2000	78,00,000	3900
1550	65,10,000	4200
2400	1,15,20,000	4800
3500	1,22,50,000	3500
2800	1,45,60,000	5200