Feature Engineering

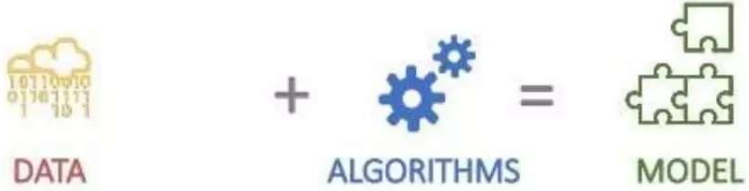
Pertemuan 6

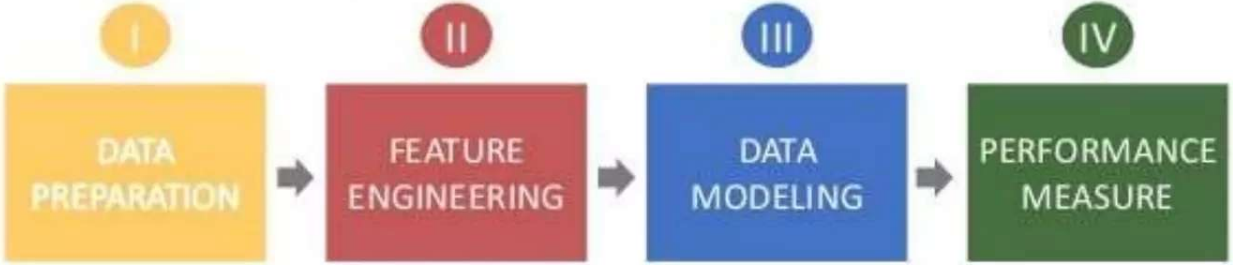
Review

Fadil Indra Sanjaya, S.Kom., M.Kom

Universitas Teknologi Yogyakarta

How does a Machine learning model work?





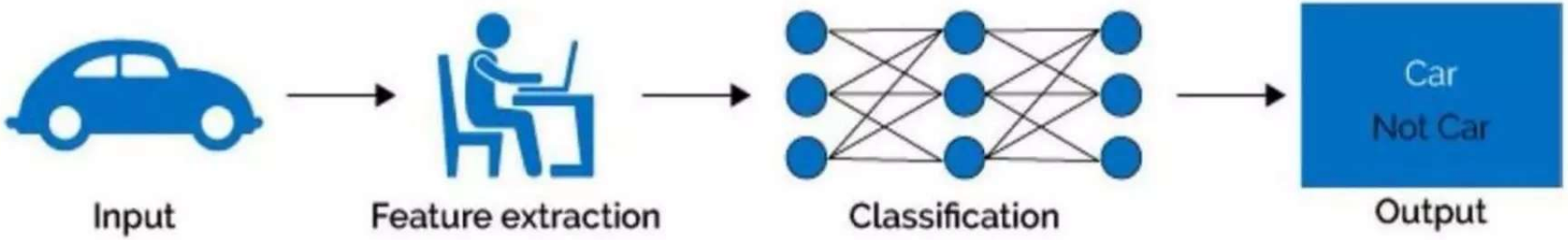
The Data

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302
5	SpiceJet	24/06/2019	Kolkata	Banglore	CCU → BLR	09:00	11:25	2h 25m	non-stop	No info	3873
6	Jet Airways	12/03/2019	Banglore	New Delhi	BLR → BOM → DEL	18:55	10:25 13 Mar	15h 30m	1 stop	In-flight meal not included	11087
7	Jet Airways	01/03/2019	Banglore	New Delhi	BLR → BOM → DEL	08:00	05:05 02 Mar	21h 5m	1 stop	No info	22270
8	Jet Airways	12/03/2019	Banglore	New Delhi	BLR → BOM → DEL	08:55	10:25 13 Mar	25h 30m	1 stop	In-flight meal not included	11087
9	Multiple carriers	27/05/2019	Delhi	Cochin	DEL → BOM → COK	11:25	19:15	7h 50m	1 stop	No info	8625
10	Air India	1/06/2019	Delhi	Cochin	DEL → BLR → COK	09:45	23:00	13h 15m	1 stop	No info	8907

Why do we need to refine the Datasets ?

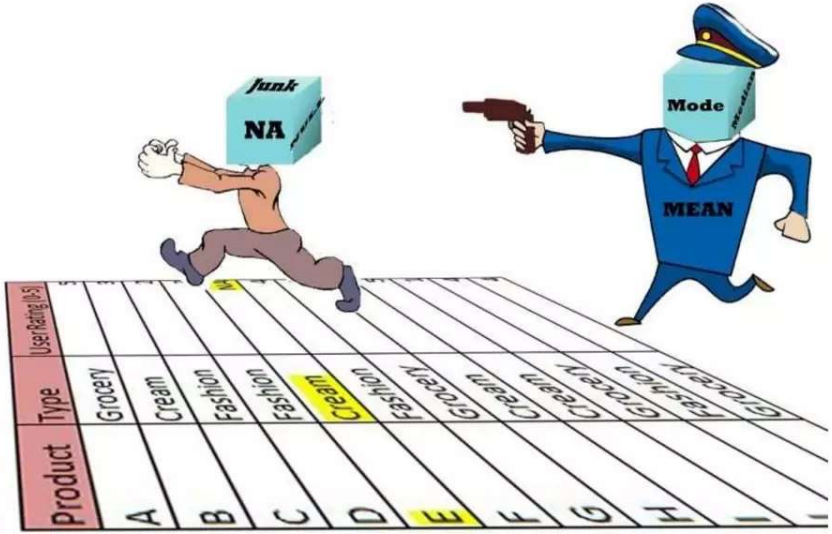
What is Feature Engineering ?

The process of extracting features from raw data is called Feature Engineering

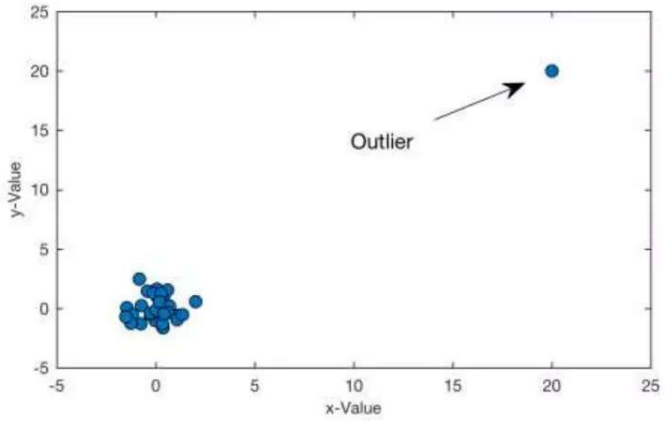


Methods for Engineering the Features

1. Imputation



2. Coping with Outliers

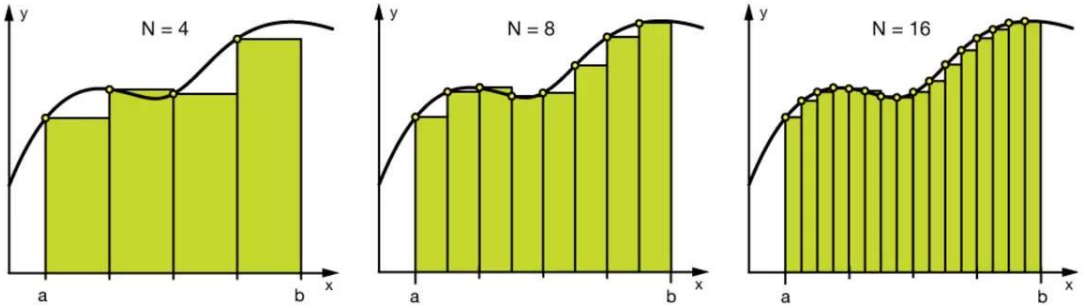


- Outliers can cause a real statistical trouble if we are concerned about the majority of the data
- They are sometimes helpful or rather most essential element of our model
  eg. Anomaly detection
- In majority of the cases we have to get rid of the outliers
- This could be done by eliminating the data points whose standard deviation is relatively high

3. Binning

- Binning is helpful is club together the values which are in a similar range
- Helps in converting discrete feature values to a categorical feature values

Age Groups	Category
<12	Children
13-19	Teen agers
20-59	Adults
>60	Elderly



4. Encoding Techniques

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

5. Feature Split

- Sometimes a feature has some information which could not be used as it is.
- Feature splitting is used when the information could be split directly into two or more features without advanced engineering

```
#String extraction example
data.title.head()
0                       Toy Story (1995)
1                         Jumanji (1995)
2               Grumpier Old Men (1995)
3              Waiting to Exhale (1995)
4    Father of the Bride Part II (1995)

data.title.str.split("(", n=1, expand=True)[1].str.split(")", n=1,
expand=True)[0]
0    1995
1    1995
2    1995
3    1995
4    1995
```

6. Scaling

- From Machine Learning point of view, all numeric features should be in a similar range
- Otherwise it gets difficult for the algorithm to fit those features.
- To avoid this issue an engineering technique called Scaling is used
- There are many mathematical methods for scaling but the most popular one is 'Normalization'

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

City	Population	Avg Age
A	54000	51.7
B	130000	45.9
C	78000	57.1
D	60000	48.6
E	92000	53.4

7. Date Extraction

- One of the most essential problem with data having dates mentioned
- These dates cannot be understood by the algorithm in their present forms
- Relevant information can be extracted from these dates according to the requirement and format of the date (This could be tricky)

  Eg. 1. Number of days between start and end date

  2. Extracting Day, Month and Year in different columns, etc.

March,15 2012

15 March 2012

March 15th ,2012

15/03/12 (Br)

03/15/12 (Am)

Who is a good Data Scientist ?

How does Feature Engineering differentiates between a good Data Scientist and a bad Data Scientist ?

Pemilihan Fitur (Feature Selection)
Mengidentifikasi fitur-fitur paling relevan untuk digunakan dalam model.

- **Metode Statistik**:
  - **ANOVA**: Untuk fitur kategorikal dan target kontinu.
  - **Chi-square test**: Untuk fitur dan target kategorikal.
  - **Correlation matrix**: Untuk menemukan hubungan antara fitur kontinu.
- **Metode Pemilihan Berdasarkan Model**:
  - **LASSO (L1 Regularization)**: Memilih fitur dengan koefisien terbesar.
  - **Feature Importance dari Tree-based Models**: Seperti Random Forest atau Gradient Boosting.
- **Metode Dimensional Reduction**:
  - **Principal Component Analysis (PCA)**: Mengurangi dimensi dengan mempertahankan informasi maksimum.
  - **t-SNE** atau **UMAP**: Untuk visualisasi dan clustering.

Transformasi Fitur (Feature Transformation)
Mengubah fitur agar lebih sesuai untuk model.

•**Skalabilitas Data**:

- **Standardization**: Mengubah data agar memiliki mean = 0 dan standar deviasi = 1.
- **Normalization**: Mengubah data agar berada dalam rentang tertentu (contoh: 0-1).

•**Transformasi Non-linear**:

- **Log Transformation**: Untuk menangani distribusi data yang sangat skewed.
- **Square root** atau **Exponential Transformation**: Untuk mengurangi atau meningkatkan skala variabel.

•**One-Hot Encoding**: Untuk mengonversi variabel kategorikal menjadi variabel dummy biner.

•**Ordinal Encoding**: Untuk kategori dengan urutan alami (contoh: tingkat pendidikan).

•**Binarization**: Mengubah fitur kontinu menjadi biner berdasarkan threshold.

Pembuatan Fitur (Feature Creation)
Menciptakan fitur baru dari data yang ada.

- **Arithmetic Transformations**:
  - Penjumlahan, pengurangan, perkalian, atau pembagian antar fitur.
- **Agregasi**:
  - Menggunakan **mean**, **sum**, **count**, **min**, atau **max** pada data grup.
- **Feature Interaction**:
  - Membuat fitur baru dengan mengalikan atau membagi dua fitur.
- **Time-based Features**:
  - Menyusun fitur seperti **year**, **month**, **day of week**, atau **season** dari data waktu.
- **Text Features**:
  - **TF-IDF**, **Word embeddings**, atau **n-grams** untuk fitur berbasis teks.

Pengisian Data Hilang (Handling Missing Values)

- **Mean/Median/Mode Imputation**: Mengisi nilai hilang dengan nilai rata-rata, median, atau modus.
- **Forward Fill/Backward Fill**: Untuk data time series.
- **Model-Based Imputation**: Menggunakan model machine learning untuk memprediksi nilai hilang.

Teknik Peningkatan Fitur (Feature Augmentation)
Menambahkan informasi eksternal atau konteks ke dalam data.

- **External Data Integration**:
  - Menambahkan informasi dari sumber lain (contoh: data cuaca, data pasar).
- **Domain-Specific Features**:
  - Fitur yang dirancang berdasarkan pengetahuan domain (contoh: volatilitas harga dalam data keuangan).

Penanganan Outlier

•**Winsorization**: Mengubah nilai ekstrem menjadi nilai batas (threshold).

•**Clipping**: Memotong nilai yang melampaui batas tertentu.

•**Transformation**: Seperti log atau square root untuk mereduksi dampak outlier.

Teknik Encoding Lanjutan untuk Data Kategorikal

- **Frequency Encoding**: Menggantikan kategori dengan frekuensi kemunculannya.
- **Target Encoding**: Mengganti kategori dengan rata-rata target variabel.
- **Leave-One-Out Encoding**: Target encoding yang mengecualikan data saat ini.

Reduksi Dimensi

Menghapus fitur yang redundant atau tidak penting.

•**Variance Threshold**: Menghapus fitur dengan varians rendah.

•**Feature Clustering**: Menggabungkan fitur yang saling berkorelasi tinggi.


thank you