Overview

Tuesday, April 23, 2024 9:58 PM

Feature Engineering:

- This is the process of creating new features from the existing features in the dataset.
- The primary aim is to try and come up with new features that can help boost model performance.
- This is a completely experimental step and is limited only by the creativity of the programmer.

Data Pre-processing:

- The process of converting the dataset into a form suitable for training a model
- Almost all features require some sort of pre-processing before model training
- Ex: *scaling numeric variables, encoding categorical variables*

Libraries:

- Pandas
- Scikit-learn
- Tsfresh
- Feature Engine
- NLTK

Numeric Variables

Tuesday, April 23, 2024 10:22 PM

1. Extreme Values:

- **Ask:**
  - Are the extreme values genuine ?
  - Are the extreme values erroneous ?
- **How to Identify?**
  - Mean and Standard Deviation
  - IQR
  - Beyond / below percentiles
  - Isolation Forest
  - Plots
    - *Box Plot*
    - *KDE Plot + Rug Plot*
- **Remedial Steps:**
  - Delete extreme values
  - Cap / Floor extreme values
    - *Arbitrary values*
    - *Mean +/- (3 * Standard Deviation)*
    - *Q1 - 1.5 * 1QR, Q3 + 1.5 * IQR*
    - *5th and 95th percentiles*
  - Represent extreme values as missing (for later imputation)

2. Missing Values:

- **Ask:**

- Are values missing because they don't exist?
- Are values missing because they weren't recorded?

- **Remedial Steps:**

  - Delete missing values
  - Create Indicator columns
  - Impute missing values
    - Mean
    - Mode
    - Median
    - Constant Value
    - Algorithms (MICE, KNN)

3. Mathematical Operations:

- **Combining Features:**

  - Aggregations:

    - Sum
    - Max
    - Min
    - Mean

  - Relativity:

    - Ratios
    - Differences

  - Decision Trees

- □ Use subset of input features to train a decision tree
- □ Use predictions from the tree as new feature

- Polynomial Features

- **Functions:**

  - Log
  - Square Root
  - Reciprocal
  - Exponential

- **Transformations**:

  - Power Transformations
  - Box-Cox transformation
  - Yeo-Johnson transformation

- **Feature Discretization:**

  - Equal Width
  - Equal Frequency
  - Arbitrary Intervals
  - K-Means Clustering

4. Feature Scaling

- **Standardization**

- **Normalization**

- **Median and IQR**

Categorical Variables

Wednesday, April 24, 2024 5:07 AM

1. **Missing Values:**
    - **Most common value**
    - **Arbitrary Value**
2. **Manipulating Categories:**
    - **Group Rare Categories**
    - **Replace Related Categories with Meaningful Values**
    - **Convert to Binary Categories**
3. **Encoding Categories:**
    - **Ordinal Encoding**
    - **One-hot Encoding**
    - **Rare-label Encoding**
    - **Frequency Encoding**
    - **Target Mean Encoding**

Date and Time Variables

Wednesday, April 24, 2024 11:10 PM

1. Extracting Features from Dates:

- Day
- Month
- Year
- Quarter
- Weekday
- Day of the Year
- Day of the Week
- Year Half (from Quarter)

2. Extracting Features from Timestamps:

- Hour
- Minute
- Second

3. Creating new Features:

- Difference between 2 features time/date
- Represent "time lag" in minutes, seconds

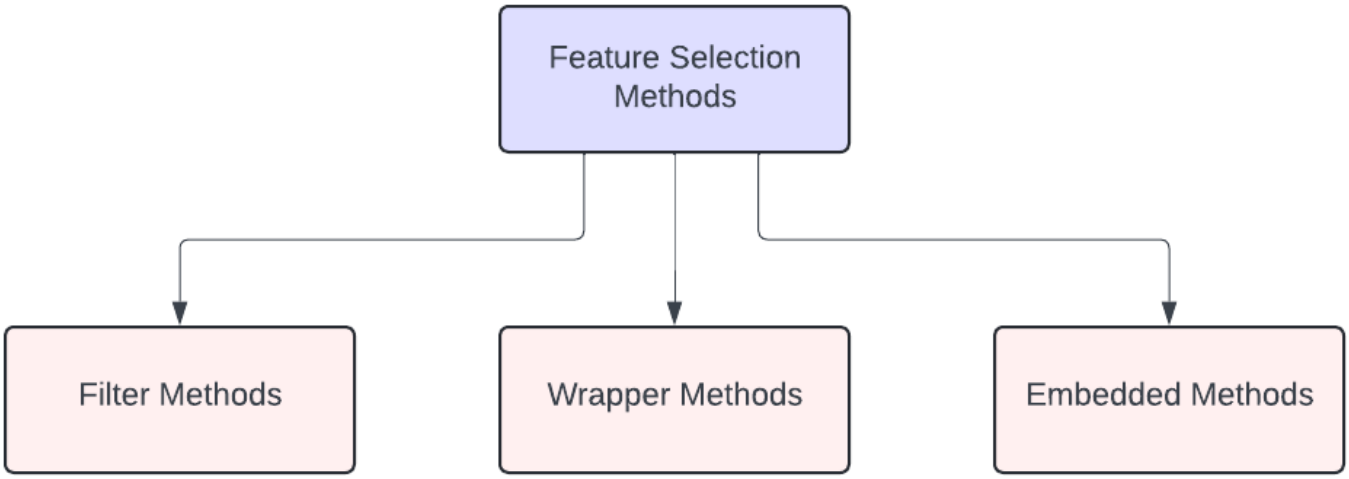
Text Variables

Thursday, April 25, 2024 4:57 AM

1. **Aggregations:**
    - Frequency of characters
    - Frequency of words
    - Frequency of unique words
2. **Ratios:**
    - Take ratios from above created features

Feature Selection

Wednesday, May 1, 2024 12:04 AM



1. Filter Methods:

- Involves ranking each variable in the dataset and then selection
- Variables are ranked based on some statistical metric calculated
  - *Chi-square test statistic*
  - *F-test statistic*
  - *Correlation Coefficient*
  - *Mutual Information*
- These techniques don't involve any Machine Learning flavours
- These techniques are fast and easily scalable

2. Wrapper Methods:

- These techniques involve generating various subsets of the dataset
- Each subset is evaluated against a Machine Learning model
  - Involves training multiple models

- Trains one model for each subset

- The variables present in the best performing subset are the selected features

- Algorithms:

  - *Exhaustive Search*
  - *Forward Elimination*
  - *Backward Elimination*

3. Embedded Methods:

- The selection mechanism used, is built-in into the Machine Learning model as part of its training phase

  - *Linear Regression*
  - *LASSO Regression*
  - *Logistic Regression*
  - *Linear Models*
  - *Decision Trees*
  - *Random Forest*
  - *Tree-based Models*

- These techniques involve training only a ***single model***

- Features are selected by ranking based on estimated coefficients or feature importances produced by the trained model

Major Components

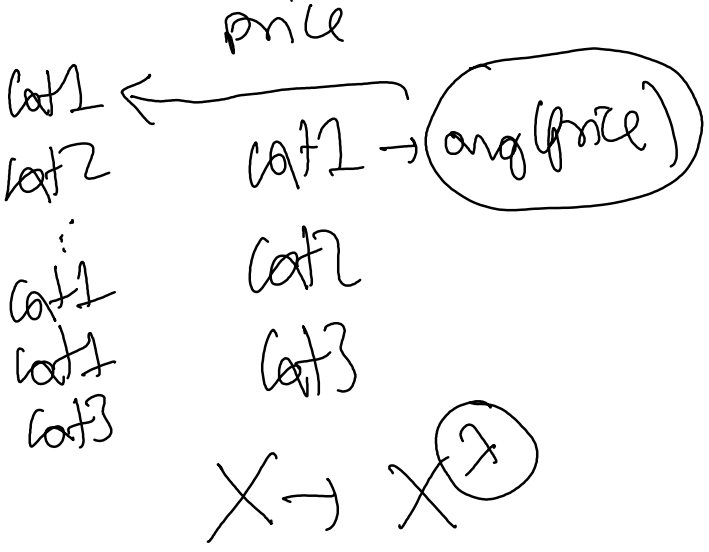
Thursday, April 25, 2024 10:05 PM

- **Built-in Transformers**
- **Custom Transformers**
  - *Python Class*
  - *Python Function*
- **Function Transformer**
- **Feature Union**
- **Pipeline**
- **Column Transformer**

Sequence of Steps

Tuesday, April 30, 2024 2:18 AM

1. **Import Libraries**
2. **Display Settings**
    - Pandas --> *Display all columns*
    - Warnings --> *Ignore warnings*
    - Scikit-learn --> *Transform Output*
3. **Read Training Data**
4. **Transformation Operations (column-wise)**
    - **<u>Airline:</u>**
        - Imputation
        - Group Rare Labels
        - One-hot Encoding
    - **<u>Date of Journey:</u>**
        - Date-time features
        - Min Max Scaling
    - **<u>Source / Destination:</u>**
        - Group Rare Labels
        - Mean Encoding
        - Power Transformer
        - Is North City
    - **<u>Departure / Arrival time:</u>**



- Date-time features
- Min Max Scaling

} time pipe1

- Part of Day
- Count Encoding
- Min Max Scaling

} pipe2

→

- **Duration:**

  - Capping by Quantiles (*Winsorizer*)
  - Imputation

    - Duration categories
    - Ordinal Encoding (*ordinal column; specify categories*)

    - RBF Percentiles Similarity (RBF Kernel)
    - Power Transformer

    Compute the rbf (gaussian) kernel between X and Y.

    K(x, y) = exp(-gamma ||x-y||^2)

    for each pair of rows x in X and y in Y.

    - Over arbitrary minutes

    - Standard Scaling

- **Total Stops:**

  - Imputation
  - Is Direct Flight

- **Additional Info:**

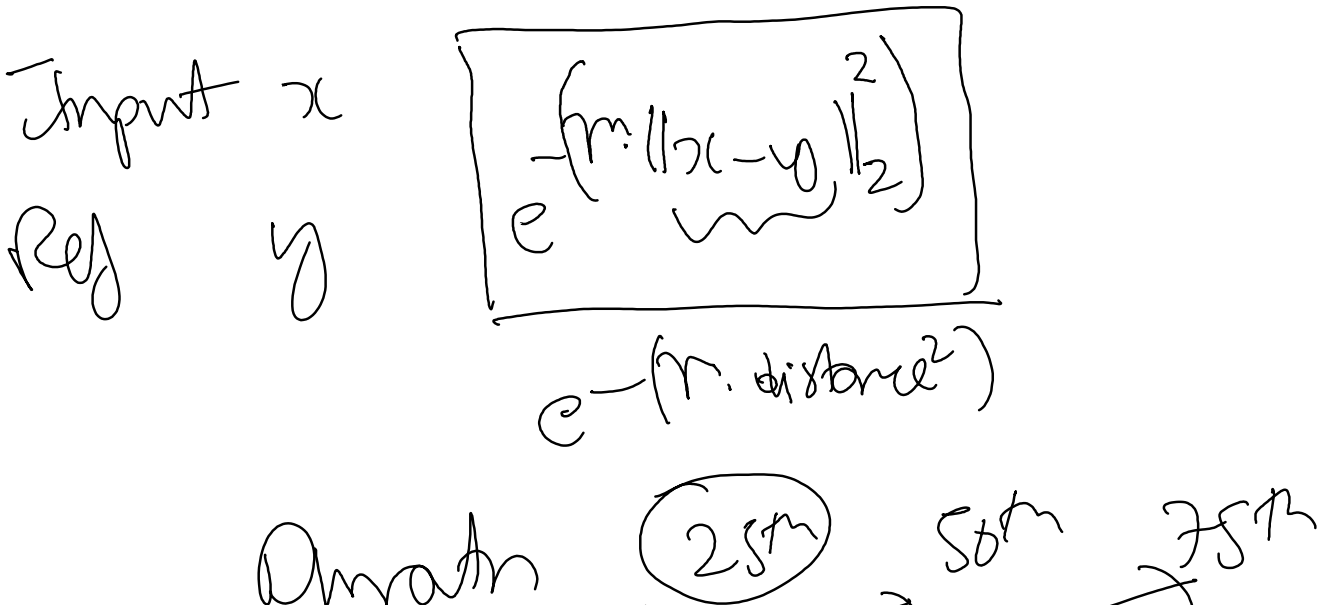
- Imputation
    - □ Group Rare Labels
    - □ One-hot Encoding
    - □ Have Info

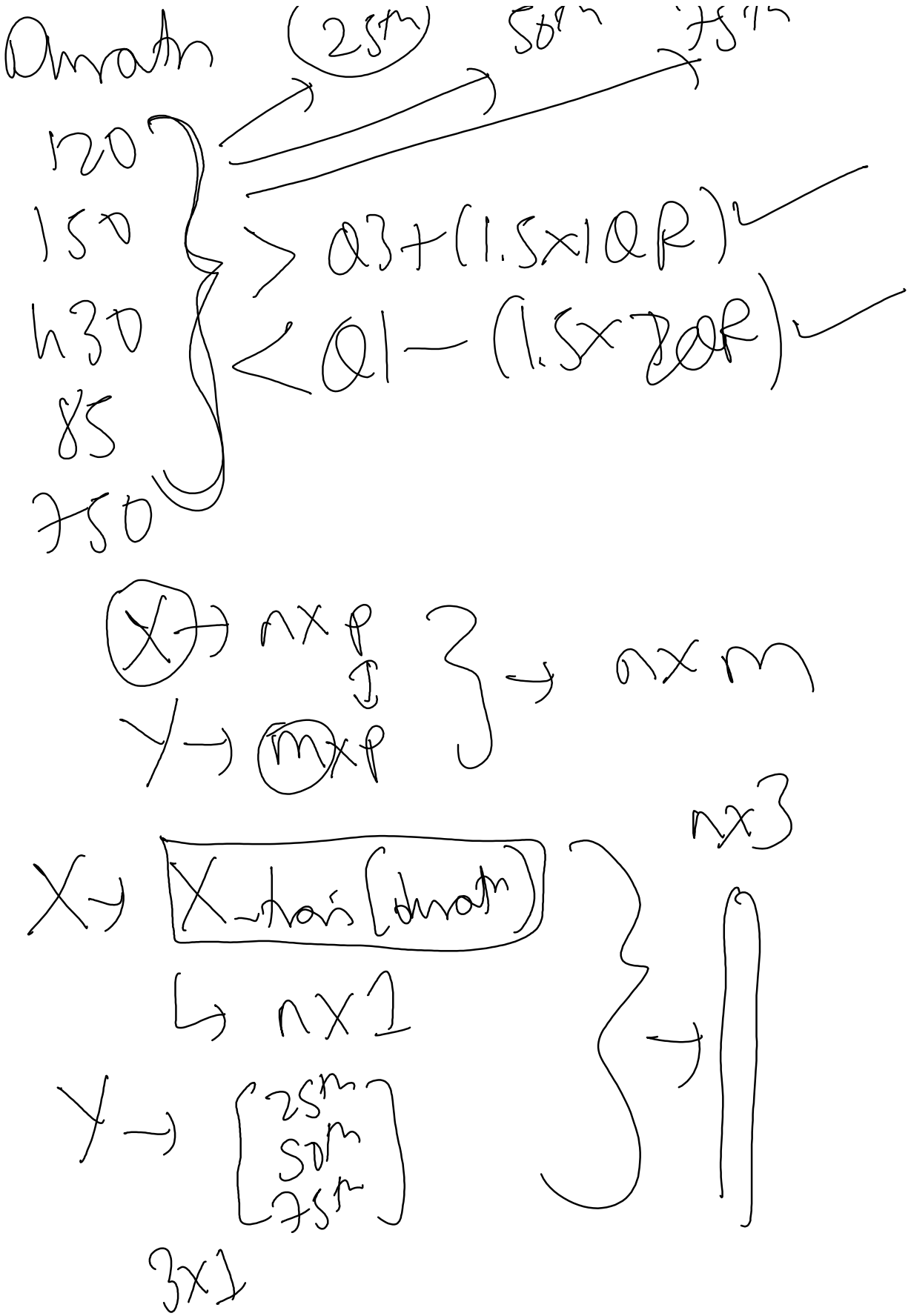
5. Feature Selection

- Selection by performance of each Individual Feature

6. Putting it all Together

7. Data Preprocessing




Qmath
120
150
430
85
750
25m
50m
75m
> Q3+(1.5x IQR)
< Q1-(1.5x IQR)
X → nxp
Y → mxp
nxm
X → X_train(Qmath)
nx1
Y → [25m, 50m, 75m]
3x1
nx3