
Feature Engineering & Feature Selection

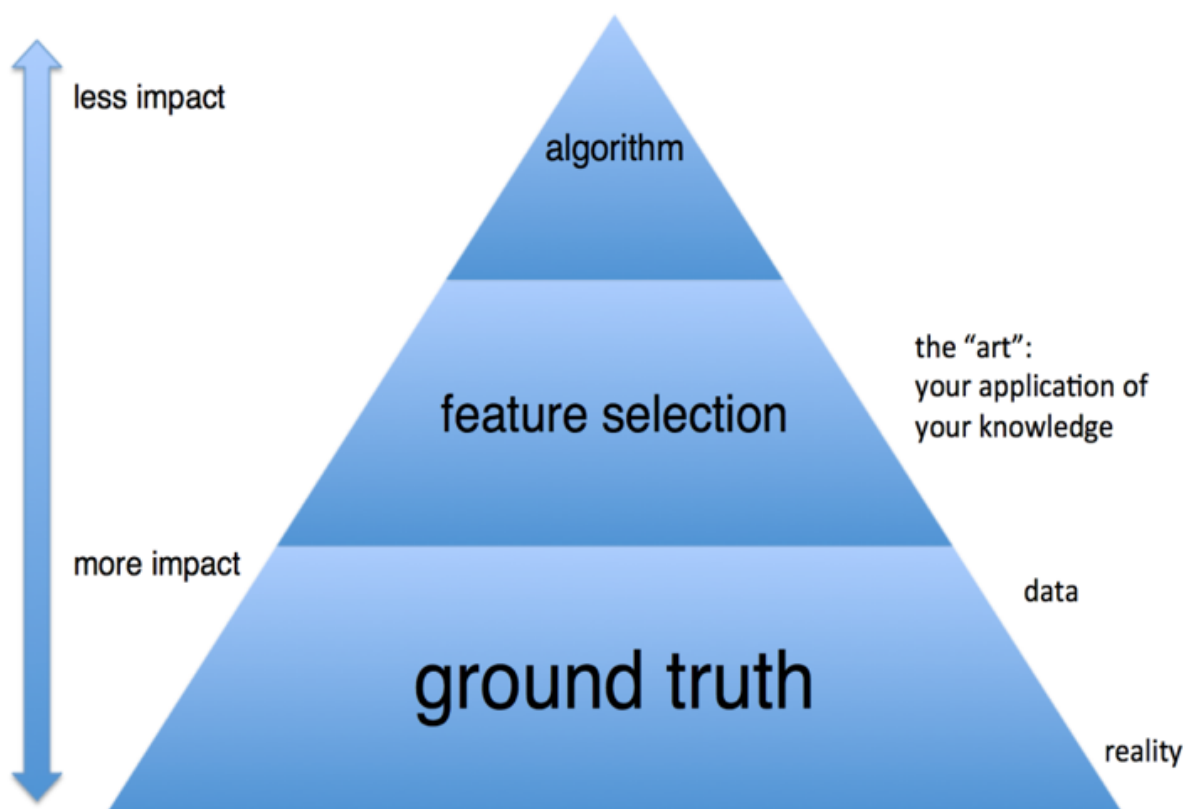
A comprehensive guide [pdf] [markdown] for **Feature Engineering** and **Feature Selection**, with implementations and examples in Python.

Motivation

Feature Engineering & Selection is the most essential part of building a useable machine learning project, even though hundreds of cutting-edge machine learning algorithms coming in these days like deep learning and transfer learning. Indeed, like what Prof Domingos, the author of 'The Master Algorithm' says:

“At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used.”

— Prof. Pedro Domingos



Data and feature has the most impact on a ML project and sets the limit of how well we can do, while models and algorithms are just approaching that limit. However, few materials could be found that

systematically introduce the art of feature engineering, and even fewer could explain the rationale behind. This repo is my personal notes from learning ML and serves as a reference for Feature Engineering & Selection.

Download

Download the PDF here:

- [https://github.com/Yimeng-Zhang/feature-engineering-and-feature-selection/blob/master/A%20Short%20Gu](https://github.com/Yimeng-Zhang/feature-engineering-and-feature-selection/blob/master/A%20Short%20Guide%20to%20Feature%20Engineering%20and%20Selection.pdf)

Same, but in markdown:

- <https://github.com/Yimeng-Zhang/feature-engineering-and-feature-selection/blob/master/A%20Short%20Gu>

PDF has a much readable format, while Markdown has auto-generated anchor link to navigate from outer source. GitHub sucks at displaying markdown with complex grammar, so I would suggest read the PDF or download the repo and read markdown with Typora.

What You'll Learn

Not only a collection of hands-on functions, but also explanation on **Why**, **How** and **When** to adopt **Which** techniques of feature engineering in data mining.

- the nature and risk of data problem we often encounter
- explanation of the various feature engineering & selection techniques
- rationale to use it
- pros & cons of each method
- code & example

Getting Started

This repo is mainly used as a reference for anyone who are doing feature engineering, and most of the modules are implemented through scikit-learn or its communities.

To run the demos or use the customized function, please download the ZIP file from the repo or just copy-paste any part of the code you find helpful. They should all be very easy to understand.

Required Dependencies:

- Python 3.5, 3.6 or 3.7
- numpy>=1.15

-
- pandas>=0.23
 - scipy>=1.1.0
 - scikit_learn>=0.20.1
 - seaborn>=0.9.0

Table of Contents and Code Examples

Below is a list of methods currently implemented in the repo.

1. Data Exploration

- 1.1 Variables
- 1.2 Variable Identification
 - Check Data Types [guide] [demo]
- 1.3 Univariate Analysis
 - Descriptive Analysis [guide] [demo]
 - Discrete Variable Barplot [guide] [demo]
 - Discrete Variable Countplot [guide] [demo]
 - Discrete Variable Boxplot [guide] [demo]
 - Continuous Variable Distplot [guide] [demo]
- 1.4 Bi-variate Analysis
 - Scatter Plot [guide] [demo]
 - Correlation Plot [guide] [demo]
 - Heat Map [guide] [demo]

2. Feature Cleaning

- 2.1 Missing Values
 - Missing Value Check [guide] [demo]

-
- Listwise Deletion [guide] [demo]
 - Mean/Median/Mode Imputation [guide] [demo]
 - End of distribution Imputation [guide] [demo]
 - Random Imputation [guide] [demo]
 - Arbitrary Value Imputation [guide] [demo]
 - Add a variable to denote NA [guide] [demo]
 - 2.2 Outliers
 - Detect by Arbitrary Boundary [guide] [demo]
 - Detect by Mean & Standard Deviation [guide] [demo]
 - Detect by IQR [guide] [demo]
 - Detect by MAD [guide] [demo]
 - Mean/Median/Mode Imputation [guide] [demo]
 - Discretization [guide] [demo]
 - Imputation with Arbitrary Value [guide] [demo]
 - Winsorization [guide] [demo]
 - Discard Outliers [guide] [demo]
 - 2.3 Rare Values
 - Mode Imputation [guide] [demo]
 - Grouping into One New Category [guide] [demo]
 - 2.4 High Cardinality
 - Grouping Labels with Business Understanding [guide]
-

-
- Grouping Labels with Rare Occurrence into One Category [guide] [demo]
 - Grouping Labels with Decision Tree [guide] [demo]

3. Feature Engineering - 3.1 Feature Scaling

- Normalization - Standardization [guide] [demo]
- Min-Max Scaling [guide] [demo]
- Robust Scaling [guide] [demo] - 3.2 Discretize
- Equal Width Binning [guide] [demo]
- Equal Frequency Binning [guide] [demo]
- K-means Binning [guide] [demo]
- Discretization by Decision Trees [guide] [demo]
- ChiMerge [guide] [demo] - 3.3 Feature Encoding
- One-hot Encoding [guide] [demo]
- Ordinal-Encoding [guide] [demo]
- Count/frequency Encoding [guide]
- Mean Encoding [guide] [demo]
- WOE Encoding [guide] [demo]
- Target Encoding [guide] [demo] - 3.4 Feature Transformation
- Logarithmic Transformation [guide] [demo]
- Reciprocal Transformation [guide] [demo]
- Square Root Transformation [guide] [demo]
- Exponential Transformation [guide] [demo]
- Box-cox Transformation [guide] [demo]
- Quantile Transformation [guide] [demo] - 3.5 Feature Generation
- Missing Data Derived [guide] [demo]
- Simple Stats [guide]
- Crossing [guide]
- Ratio & Proportion [guide]
- Cross Product [guide]
- Polynomial [guide] [demo]
- Feature Learning by Tree [guide] [demo]
- Feature Learning by Deep Network [guide]

4. Feature Selection

- 4.1 Filter Method
 - Variance [guide] [demo]

-
- Correlation [guide] [demo]
 - Chi-Square [guide] [demo]
 - Mutual Information Filter [guide] [demo]
 - Information Value (IV) [guide]
 - 4.2 Wrapper Method
 - Forward Selection [guide] [demo]
 - Backward Elimination [guide] [demo]
 - Exhaustive Feature Selection [guide] [demo]
 - Genetic Algorithm [guide]
 - 4.3 Embedded Method
 - Lasso (L1) [guide] [demo]
 - Random Forest Importance [guide] [demo]
 - Gradient Boosted Trees Importance [guide] [demo]
 - 4.4 Feature Shuffling
 - Random Shuffling [guide] [demo]
 - 4.5 Hybrid Method
 - Recursive Feature Selection [guide] [demo]
 - Recursive Feature Addition [guide] [demo]

Key Links and Resources

- Udemy's Feature Engineering online course

<https://www.udemy.com/feature-engineering-for-machine-learning/>

- Udemy's Feature Selection online course

<https://www.udemy.com/feature-selection-for-machine-learning>

- JMLR Special Issue on Variable and Feature Selection

<http://jmlr.org/papers/special/feature03.html>

- Data Analysis Using Regression and Multilevel/Hierarchical Models, Chapter 25: Missing data

<http://www.stat.columbia.edu/~gelman/arm/missing.pdf>

- Data mining and the impact of missing data

<http://core.ecu.edu/omgt/krosj/IMDSDataMining2003.pdf>

- PyOD: A Python Toolkit for Scalable Outlier Detection

<https://github.com/yzhao062/pyod>

- Weight of Evidence (WoE) Introductory Overview

<http://documentation.statsoft.com/StatisticaHelp.aspx?path=WeightofEvidence/WeightofEvidenceWoEIntroductory>

- About Feature Scaling and Normalization

http://sebastianraschka.com/Articles/2014_about_feature_scaling.html

- Feature Generation with RF, GBDT and Xgboost

https://blog.csdn.net/anshuai_aw1/article/details/82983997

- A review of feature selection methods with applications

<https://ieeexplore.ieee.org/iel7/7153596/7160221/07160458.pdf>