

Introduction To Data Science

Eric@Udacity

Content

- Data cleaning and exploratory data analysis
- Feature engineering
- Model construction
- Model evaluation

1. Data cleaning and exploratory data analysis

- Understand data

Structured data: categorical, numerical

Unstructured data: text, image

- Data cleaning
- Exploratory data analysis
- Unbalanced data
- Supervised Binning method

2. Feature engineering

- Tf-idf
- PCA
- Lasso
- Feature selection

Remove low variance features

Univariate feature selection

Recursive feature elimination

Selection features through model

3. Modeling

- GLM-Generalized Linear Regression
- Naïve Bayes
- Tree Model
- Neural Network
- K-means, KNN
- Ensemble methods
- SVM*

4. Model evaluation

- Confusing matrix

	Real: positive	Real: negative
Predict: positive	True Positive	False Positive
Predict: Negative	False Negative	True Negative

- Regression & Classification Evaluation
- Bias & Variance
- Learning Curve