

Remember that professional DS's spend most of their time on the stages that lead to this point:  
 (1) Exploring the data; (2) Cleaning the data;  
 (3) Engineering new features.  
 This is because of our dear Boolean:  
 Better Data > Fancier Algorithms.  
 Everything was done aiming to boost our performance and quality on this last stage.

2. Plot the distributions:  
 Numerical  
 Categorical  
 Segmentation

Remember that your aim here is to gain intuition about the data. By the end of your Exploratory Analysis step, you'll have a pretty good understanding of the dataset, some notes for data cleaning, and possibly some ideas for feature engineering.

2. Tune Hyper parameters

3. Cross Validate

4. Select Wining model

1. Get a "feel" of the data with basic questions

Exploratory Analysis  
10 %

3. Study Correlations

1. Split Dataset (Train-Test)

Model Training  
15%

5

THANKS TO:  
ELITEDATASCIENCE



1



# JESUSPRZR Data Science workflow process (Modeling Oriented)

BETTER DATA > FANCIER ALGORITHMS

1. Remove Unwanted Observations

Depending of the quality of your data, your projects will live or die. Doing data cleaning properly can really save you from a ton of headaches down the road, so please don't rush this step.

Some of these are:  
 Lasso regression  
 Ridge regression  
 Elastic-Net  
 Random forest  
 Boosted tree

Algorithm Selection  
10%



4

2



Data Cleaning  
20%

2. Fix Structural Errors

Input = Output

The most effective algorithms typically offer: regularization, automatic feature selection, ability to express nonlinear relationships, and/or ensemble.

5. Remove Unused Features

4. Add Dummy Variables

3. Combine Sparse Classes

Feature Engineering  
25%

2. Create Interaction Features

1. Infuse Domain Knowledge



4. Handle Missing Data

3. Filter Unwanted Outliers

by understanding these concepts (regularization, ensembling, automatic feature selection, etc.) we also get to understand why some algorithms tend to perform better than others.

Remember that of the process of doing ML this is the part on which DS's spend more time. Feature engineering is about creating new input features from your existing ones. One highly predictive feature makes up for 10 duds.