Learning Objectives

End to End Model Building

Benefits of Feature Selection

Boruta

DPhi

Problem Solving

- **Define Objective or understand the problem statement**
- Data Requirements
- Data Collection
- **Exploratory Data Analysis**
- **Data Pre-processing**
- **Build a model**
- **Evaluate**
- **Optimise**
- **Production**
- **Monitor**
- **You keep Optimising it every now and then**

Objective/Problem Statement

The goal of the model is to **chances of survival of a patient after 1 year of treatment**

DPhi

Data Requirements & Collection

We have the data!

<will talk about real-world case in at the end>

DPhi

Understanding the Data

- ID_Patient_Care_Situation: Care situation of a patient during treatment
- Diagnosed_Condition: The diagnosed condition of the patient
- ID_Patient: Patient identifier number
- Treatment_with_drugs: Class of drugs used during treatment
- Patient_Age: Age of the patient
- Patient_Body_Mass_Index: A calculated value based on the patient's weight, height, etc.
- Patient_Smoker: If the patient was a smoker or not
- Patient_Rural_Urban: If the patient stayed in Rural or Urban part of the country

Understanding the Data

- Previous_Condition: Condition of the patient before the start of the treatment (This variable is splitted into 8 columns - A, B, C, D, E, F, Z and Number_of_prev_cond. A, B, C, D, E, F and Z are the previous conditions of the patient.

- Survived_1_year: If the patient survived after one year (0 means did not survive; 1 means survived)

Hands-on

Hands-on

One hot encoding

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Used for nominal data without any order

DPhi

Too many Features - Problem & Solution

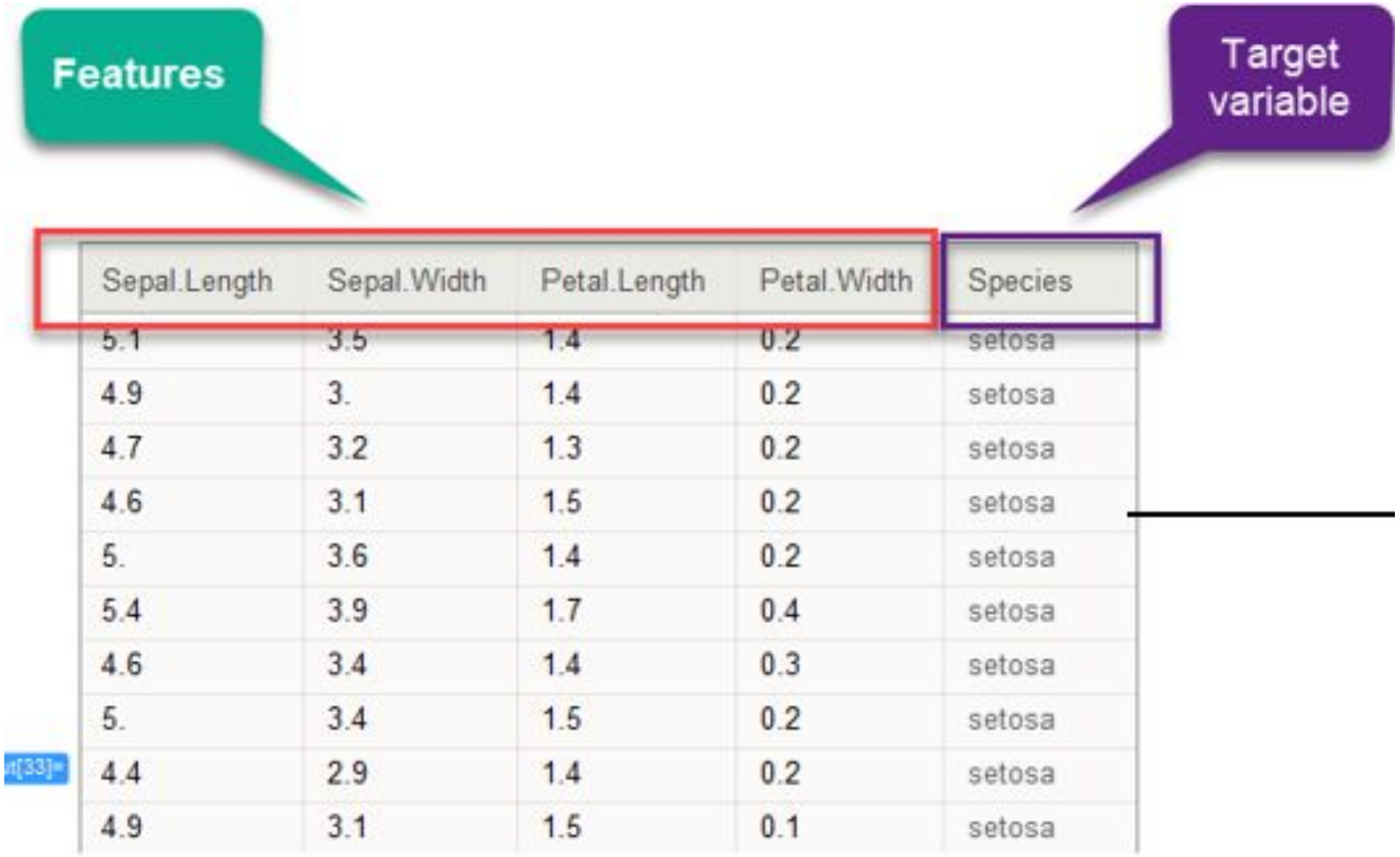
Problem:

Gone are the days when you had 5 variables to fit your linear regression: Modern datasets contain more variables/features to choose from. A dataset with 50 or more features -> more than 1 million observations.

Solution: Feature Selection & Feature importance

We will come to why/benefits very soon!

Iris Dataset



DPhi

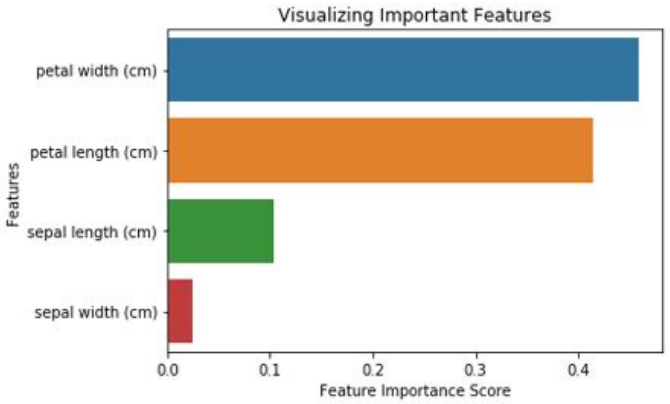
Feature Importance

Features

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3.	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.	3.6	1.4	0.2
5.4	3.9	1.7	0.4
4.6	3.4	1.4	0.3
5.	3.4	1.5	0.2
4.4	2.9	1.4	0.2
4.9	3.1	1.5	0.1

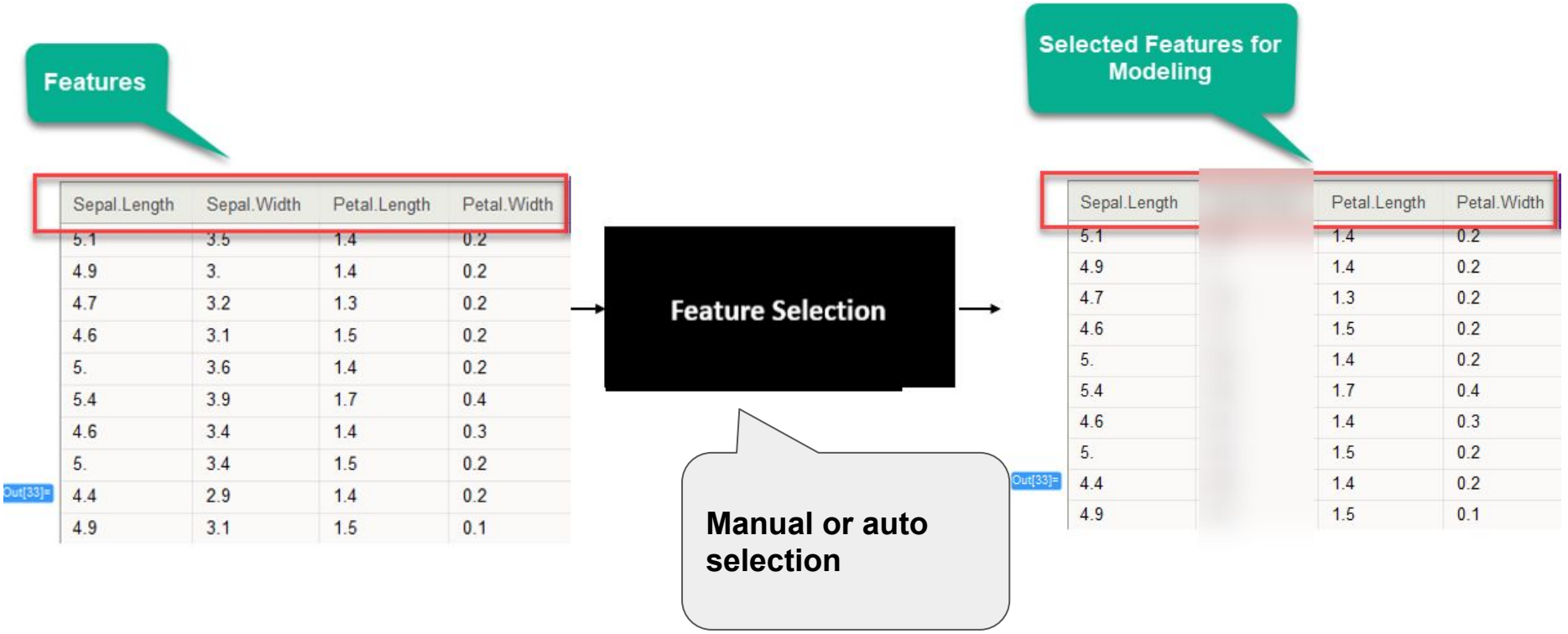
Out[33]=

Feature Importance



DPhi

Feature Selection



DPhi

Why Feature Selection?

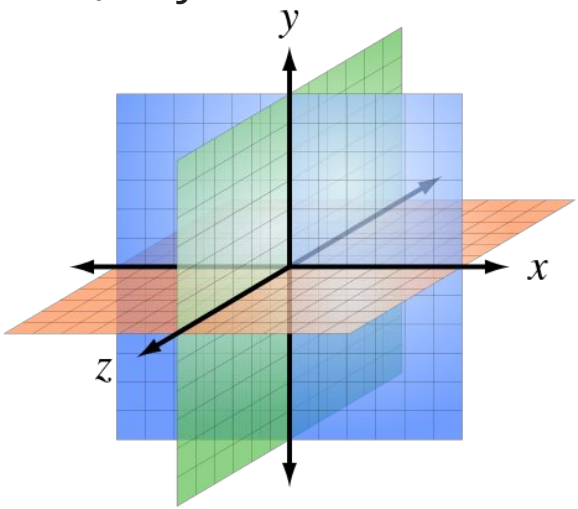
You already know a number of optimization methods by now and might think what's the need of reducing our data by feature selection if we can just optimize?

There's something known as "*The curse of dimensionality*".

In machine learning,

"dimensionality" = number of features (i.e. input variables) in your dataset.

When the number of features is very large relative to the number of observations(rows) in your dataset, certain algorithms struggle to train effective models. This is called the Curse of Dimensionality.



DPhi

Why Feature Selection?

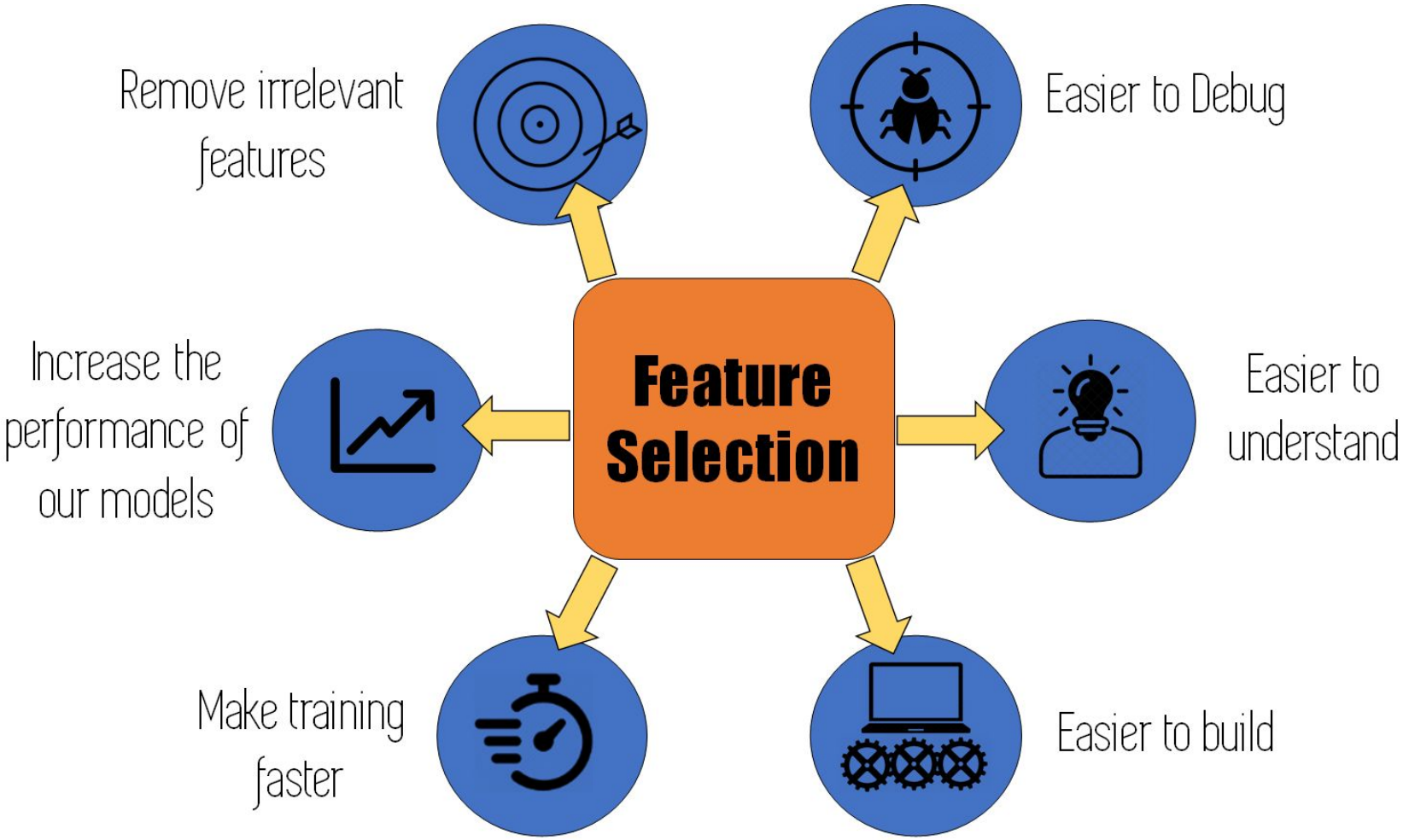


Image source: TowardsDataScience

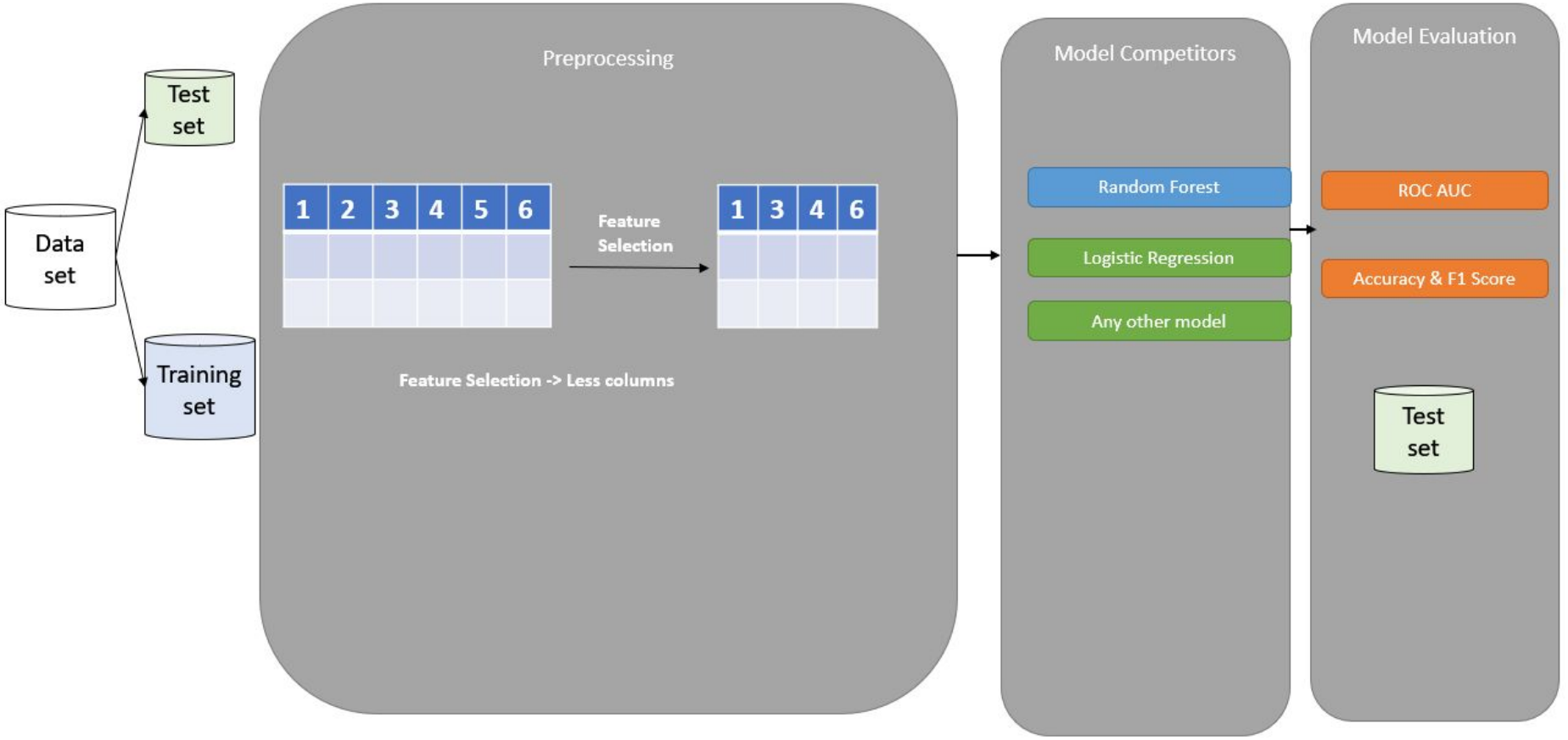
Benefits of performing Feature Selection

You might've gotten an idea of why feature selection is required by now.

Feature Selection helps us with the following:

- **Reduces Overfitting:** Less redundant data means less opportunity to make decisions based on noise(irrelevant data).
- **Improves Model Performance:** Less misleading data means our model's performance improves.
- **Reduces Training Time:** Less data means that algorithms train faster.

Setting up Experiments



DPhi

Boruta - Feature Selection Technique

Boruta follows an all-relevant feature selection method where it captures all features which are in some way or other relevant to the target variable.

In contrast, most of the traditional feature selection algorithms follow a minimal optimal method where they rely on a small subset of features which yields a minimal error on a chosen classifier.

Dataset

Input features

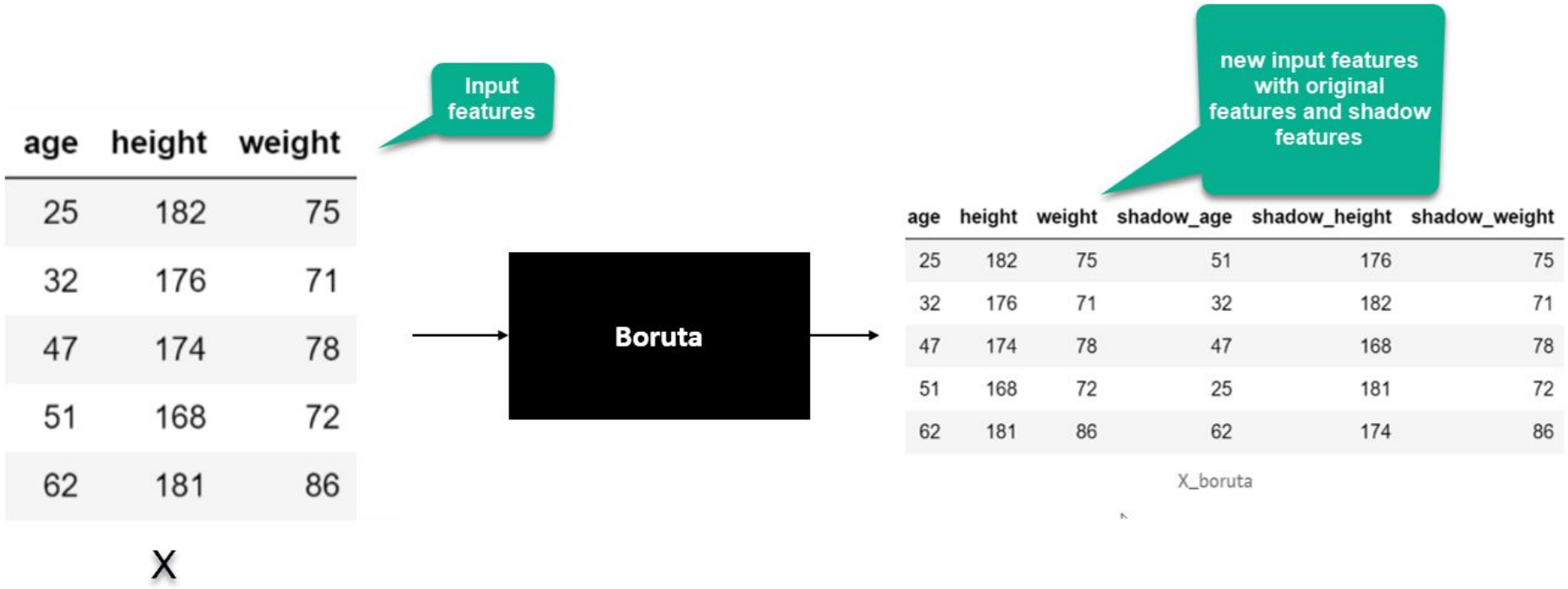
target variable

	age	height	weight
0	25	182	75
1	32	176	71
2	47	174	78
3	51	168	72
4	62	181	86

	income
0	20
1	32
2	45
3	55
4	61

X and y

Boruta Intuition



age	height	weight
25	182	75
32	176	71
47	174	78
51	168	72
62	181	86

age	height	weight	shadow_age	shadow_height	shadow_weight
25	182	75	51	176	75
32	176	71	32	182	71
47	174	78	47	168	78
51	168	72	25	181	72
62	181	86	62	174	86

X_boruta

DPhi

Boruta Intuition

New Dataset with Shadow Features

age	height	weight	shadow_age	shadow_height	shadow_weight
25	182	75	51	176	75
32	176	71	32	182	71
47	174	78	47	168	78
51	168	72	25	181	72
62	181	86	62	174	86

X_boruta

Boruta FS & Importance

Feature importance comparison of original features and shadow features

	age	height	weight	shadow_age	shadow_height	shadow_weight
feature importance %	39	19	8	11	14	9
hits	1	1	0	-	-	-

Outcome of one run

Boruta + Random forest classifier

DPhi

What makes it different?

- In essence, Boruta is trying to validate the importance of the feature by comparing with random shuffled copies (shadow features), which increases the robustness. This is done by simply comparing the number of times a feature did better with the shadow features using a z-score.

- **Comparison with standard random forest algorithm:** While fitting a random forest model on a data set, you can recursively get rid of features in each iteration which didn't perform well in the process. This will eventually lead to a minimal optimal subset of features as the method minimizes the error of random forest model. This happens by selecting an over-pruned version of the input data set, which in turn, throws away some relevant features. On the other hand, boruta find all features which are either strongly or weakly relevant to the decision variable. So this makes it very unique

Boruta References

- The following article explains it intuitively and it was implemented from scratch: https://towardsdatascience.com/boruta-explained-the-way-i-wish-someone-explained-it-to-me-4489d70e154a

- The below article is in R programming - but still it gives a nice comprehensive view about Boruta with a dataset: https://www.datacamp.com/community/tutorials/feature-selection-R-boruta

- https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/

Slide Download Link

You can download the slides here:
https://docs.google.com/presentation/d/1591auCusOl5Qs2GQ5OQFEG3Jdd44nCMnj8esuLe5dKQ/edit?usp=sharing

DPhi

That’s it for this unit. Thank you!

Feel free to post any queries on Discuss.

DPhi