Data Mining for Business in Python

Data Mining for Business in Python 2021

1. Survival Analysis
2. Cox Proportional Hazard
3. CHAID
4. Gaussian Mixture Model
5. Dimension Reduction
6. Association Rule Learning
7. Random Forest
8. LIME
9. XGBoost and SHAP

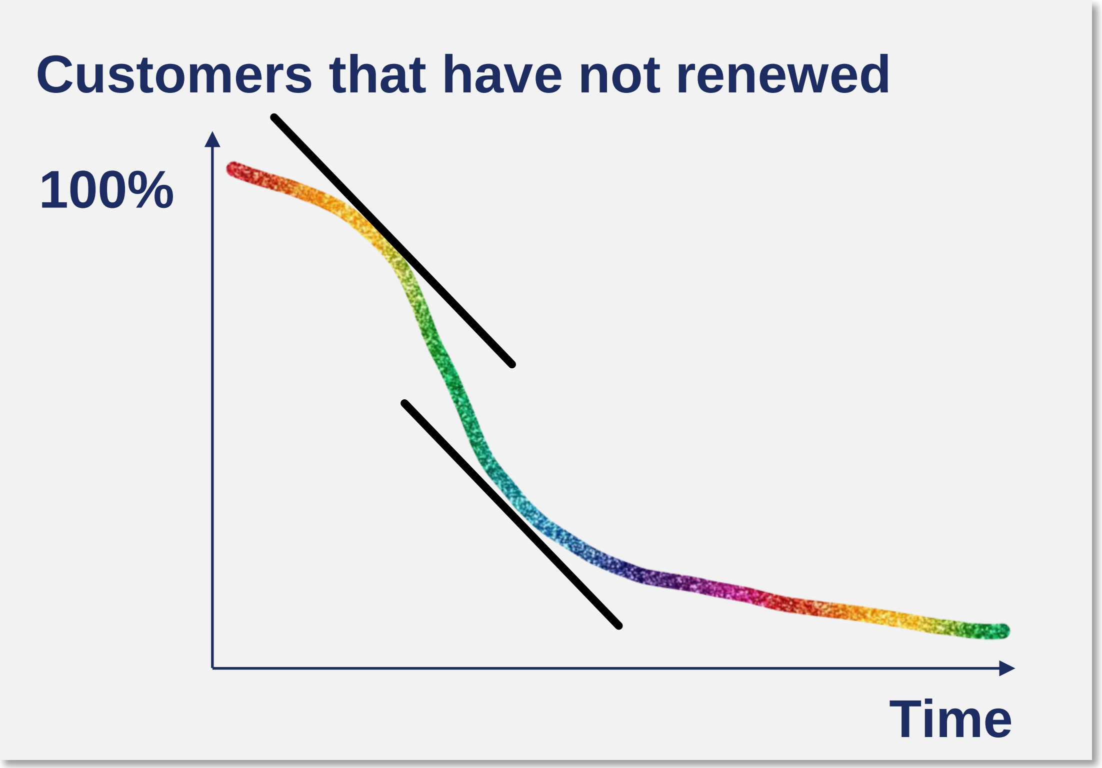
Survival Analysis

Introduction to Survival Analysis

Context

- Survival Analysis is very common for Subscription type businesses and very apt to study customer churn
- Imagine a customer decides to cancel their subscription. How long do you wait until you try to get that customer?
- 1 day
- 1 week
- 1 month

Visualization



Case Study Briefing

Lung Cancer

Case study[1]

Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group.

- Determine the survival curve through the Kaplan Meyer Estimator
- Understand differences between Males and Females

1: Author Terry M Therneau [aut, cre], Thomas Lumley [ctb, trl] (original S->R port and R maintainer until 2009), Atkinson Elizabeth [ctb], Crowson Cynthia [ctb]

Survival Analysis Step by Step

Prepare Dataset

↓

Perform Survival Curve

↓

Visualize and Interpret Results

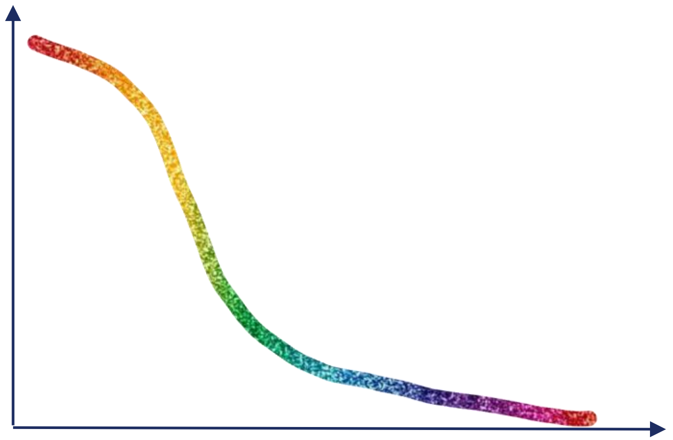
↓

Perform Log Rank Test (if the case requires it)

Kaplan-Meier Estimator

Explanation

- Non-parametric statistic used to estimate the survival function (probability of a person surviving) from the lifetime data.
- In medical research, it is often used to measure the fraction of patients living for a specific time after treatment or diagnosis.



Formula

$$S(t_i) = S(t_{i-1}) * (1 - \frac{d_i}{n_i})$$

Where:

$S(t_i)$ = probability of survival at time t

d_i = number of events at time t

n_i = number of survivors at time t

Censoring

Types

Description

Right Censoring:
The subject under observation is still alive. In this case, we can not have our timing when our event of interest (death) occurs.

Left Censoring:
The event cannot be observed for some reason. The event may also have started before recording.

Interval Censoring:
We only have data for a specific interval, so it is possible that the event of interest does not occur during that time.

Log Rank Test

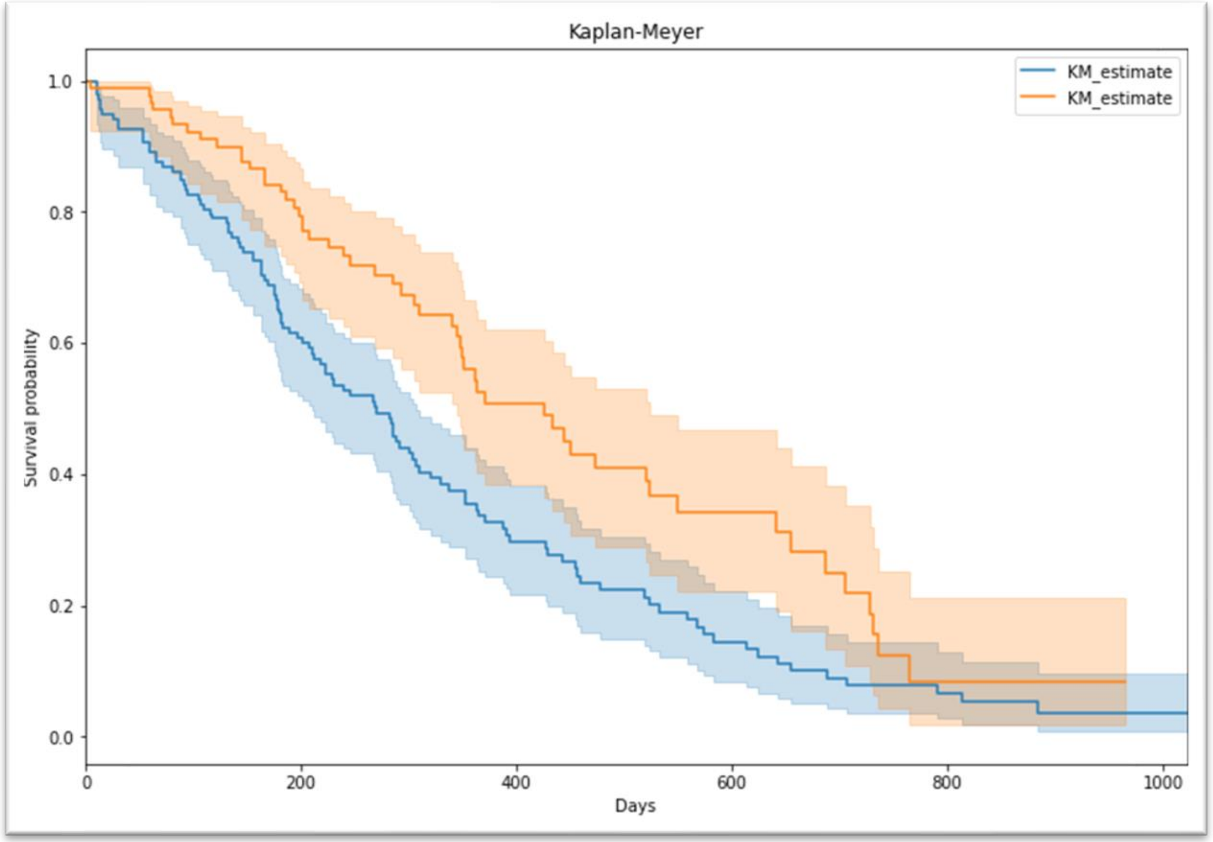
Context

Goal:
To test if there are statistical differences in the survival distribution of >= 2 groups

Null Hypothesis:
There is no difference between both groups

If p-value > 0.05:
There is no difference between both groups

Visualization



Survival Analysis extra Resources

Deep dives

Book
Survival Analysis
Stephen P. Jenkins

Challenge – Electronic components

Experiment

In 1988, an experiment was designed and implemented at one AT&T's factory.

The goal was to investigate alternatives in the "wave soldering" procedure for mounting electronic components to printed circuit boards.

The response is the number of visible solder skips.

Output

1. Transform Solder Variable into 1 and 0
2. Fit the Kaplan-Meyer estimator
3. Plot Survival curves
4. Do Logrank test for the Panel variable. Use `multivariate_logrank_test`
   Or be creative :D

Dataset source: Survival package from CRAN

Cox Proportional Hazard Regression

Cox Proportional Hazard Regression

Explanation

- Survival Analysis does not allow other predictors
- At best, you can split in groups of gender, age, etc… and perform a Log-rank test
- Thus, Cox Proportional Hazard regressions helps to determine the relationship between the survival time of a subject and one or more predictor variables
- $\exp(b_n)$ are called the Hazard Ratios (HR)

Formula

$$h(t) = h_o(t) * \exp(b_1 * x_1 + b_n * x_n)$$

Where:

$h_o(t)$ = baseline hazard

b_n = impact coefficients

x_n = covariates

Result interpretation:

HR > 1: increase
HR < 1: decrease
HR = 1: neutral

Cox Proportional Hazard Regression Step by Step

Prepare Dataset

↓

Cox Proportional Regression

↓

Visualize and Interpret Results

Case Study Briefing

Case study[1]

Lung Cancer

Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group.

1. Driver Analysis with Cox Proportional Hazard
2. Visualize Results

1: Author Terry M Therneau [aut, cre], Thomas Lumley [ctb, trl] (original S->R port and R maintainer until 2009), Atkinson Elizabeth [ctb], Crowson Cynthia [ctb]

Cox Proportional Hazard extra Resources

Deep dives

Time-dependent covariates in the cox proportional-hazards regression model
Lloyd D. Fisher and D. Y. Lin 1999

Cox Proportional-Hazards Regression for Survival Data in R
John Fox & Sanford Weisberg

Challenge – Veteran Lung Cancer A/B test

Experiment

Randomised trial of two treatment regimens for lung cancer.

Challenge[1]

1. Transform Cell Type into dummy variables.
   Use pd.get_dummies or drop the variable
2. Cox Proportional Hazard Regression
3. Plot CPH results

Dataset source: Survival package from CRAN

CHAID

Case Study Briefing

Labor Market Ethnic Discrimination

Case study[1]

Cross-section data about resume, call-back and employer information

4,870 fictitious resumes sent in response to employment advertisements in Chicago and Boston in 2001

The resumes contained information concerning the ethnicity of the applicant.

Bertrand, M. and Mullainathan, S. (2004).
Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. American
Economic Review, 94, 991–1013.

CHAID Step by Step

Variable selection

Transforming continuous variables into categorical

Do your first tree

Prune it for better interpretability

Factors influencing call-backs

✓ High ? Medium X Low

Employment gaps

Honorary degree	Yes	No
Yes	?	✓
No	X	?

Computer skills

College	Yes	No
Yes	✓	?
No	?	X

Resume quality

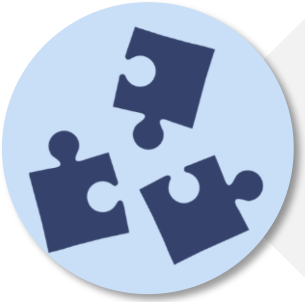
Work experience	Yes	No
High	✓	?
Low	?	X

Military experience

Special skills	Yes	No
Yes	✓	?
No	?	X

Complexity increases as you deep dive in your problem

Why?



Description

Problem depth:
Having more than 20, 50 or 100 drivers increases the complexity

Importance:
how do you know which driver actually matters most?

Relevance:
Some variables might be relevant in combination with some, but not all

One of the CHAID‘s benefits is that figures out which drivers are more important

Which?

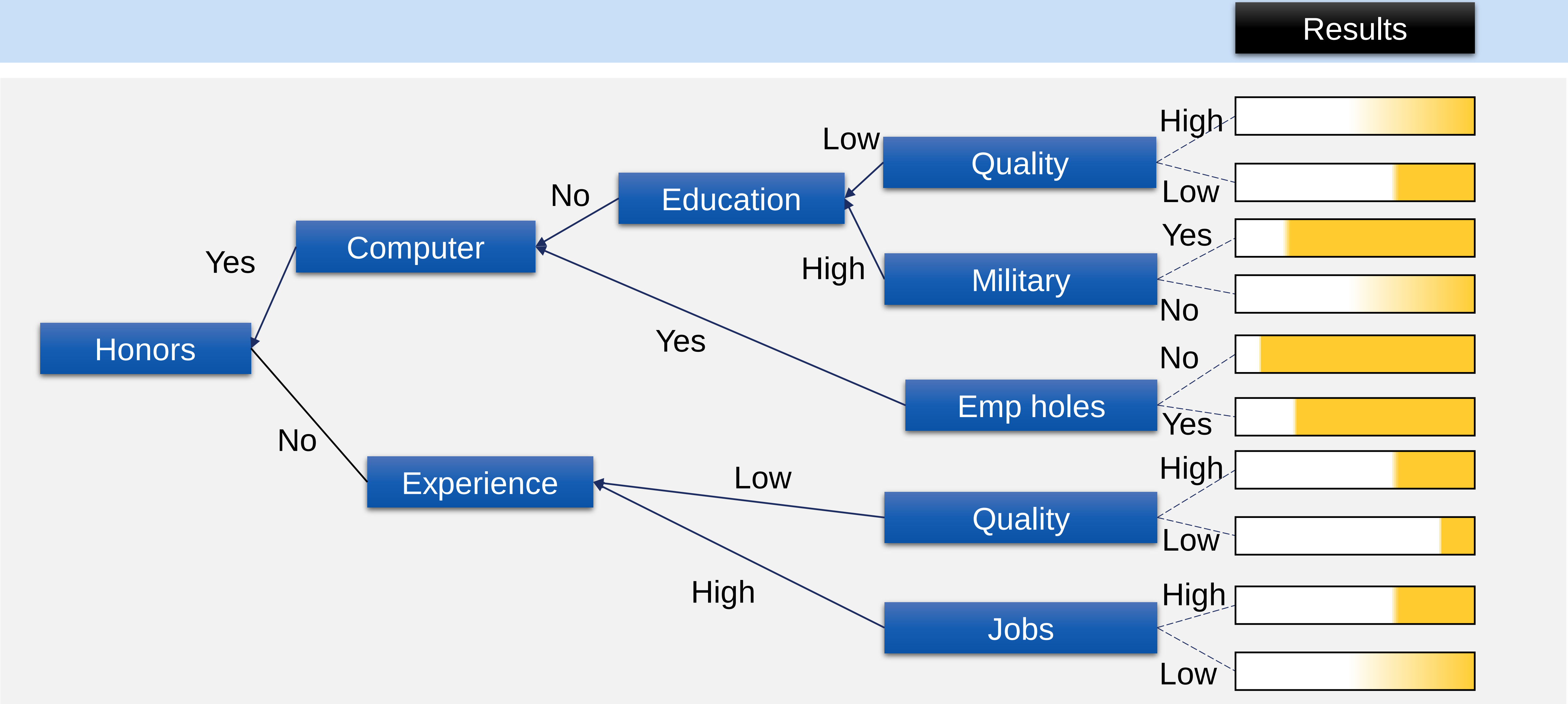
Description

Importance ranking:
CHAID figures out which drivers matter more, by doing significance tests

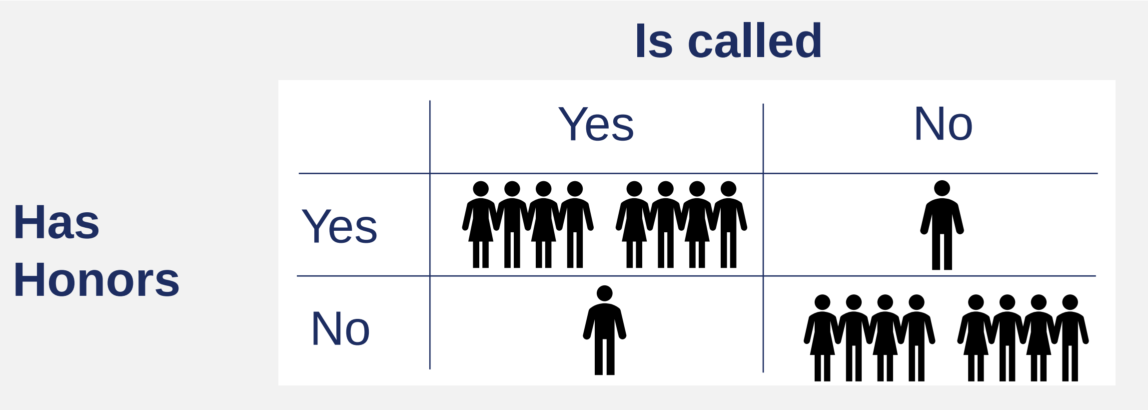
Segmented Driver Analysis:
CHAID will segment the population and perform driver analysis for each of them.

Interpretability:
CHAID provides easy to read graphs with customer segments

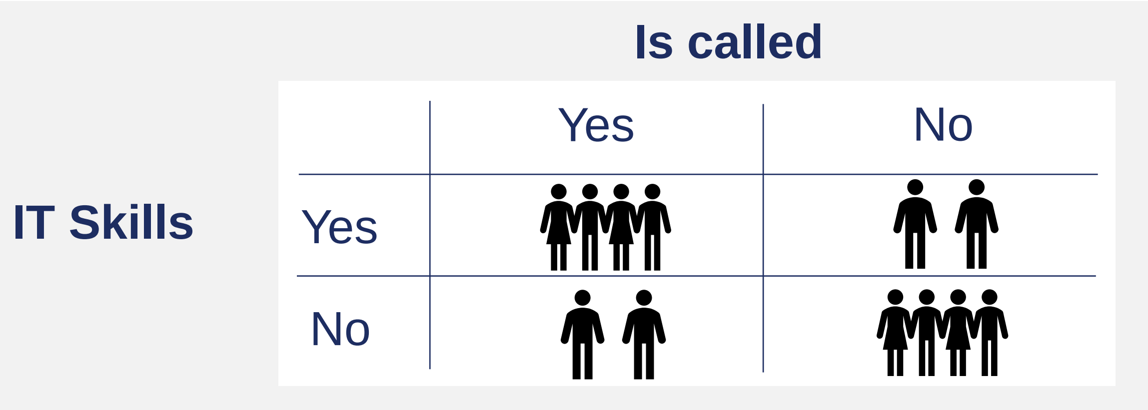
Let‘s see how it works visually



How CHAID processes



Of the people who have honors:



What does it do?

CHAID looks at all predictors and tries to find the one where the “yes” is most different from the “no”

How does it work?

CHAID performs a Chi-square test. It shows whether the frequencies of the categorical variables are different or not. Very similar to t-test, but it is a test of variance, and ideal for categorical variables.

And then?

After it finds the first segment split, tries to find the next where the “yes” differs most from the “no”

Last few things consider

Which?

Description

Tree size:
You can choose how many levels the tree will have

Bucket size:
You can choose a minimum threshold that you want your buckets to have

Continuous variables:
CHAID accepts only categorical variables

CHAID extra Resources

Deep dives

A CHAID Based Performance Prediction Model in Educational Data Mining
M. Ramaswami and R. Bhaskaran, 2010

Tree Structured Data Analysis: AID, CHAID and CART
Leland Wilkinson 1992

Challenge – Police Racial Bias

Challenge[1]

Description

You have been hired to understand to investigate Vehicle searches by the police, and if there is racial bias

1. Create a dataset with these 5 variables: problem, vehicleSearch, race, gender, policePrecinct
2. Transform string variables into dummy.
3. Get names of Dependent and Independent variables
4. Perform CHAID and visualize. Set max depth to 2

Dataset source: carStops package from CRAN

Clustering: Gaussian Mixture Model

Case Study Briefing – Country Segmentation

Case study[1]

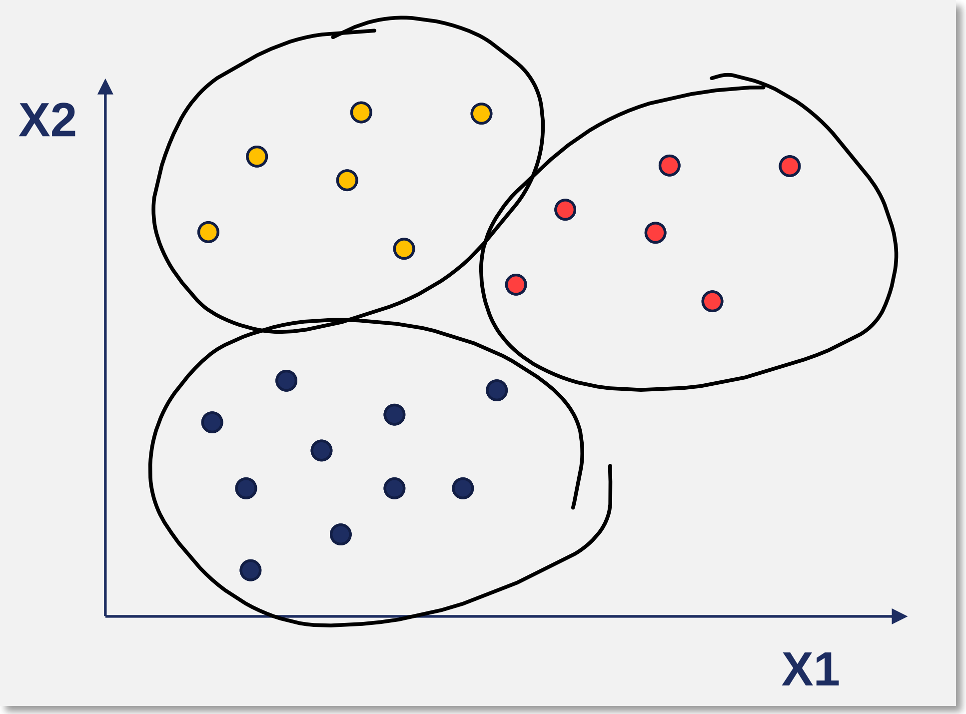
Socio-Economic Data

Data with country socio-economic data

1. Find optimal Number of cluster
2. Visualize optimal number of clusters
3. Create clusters
4. Interpret the clusters

What are clustering techniques?

Visualization

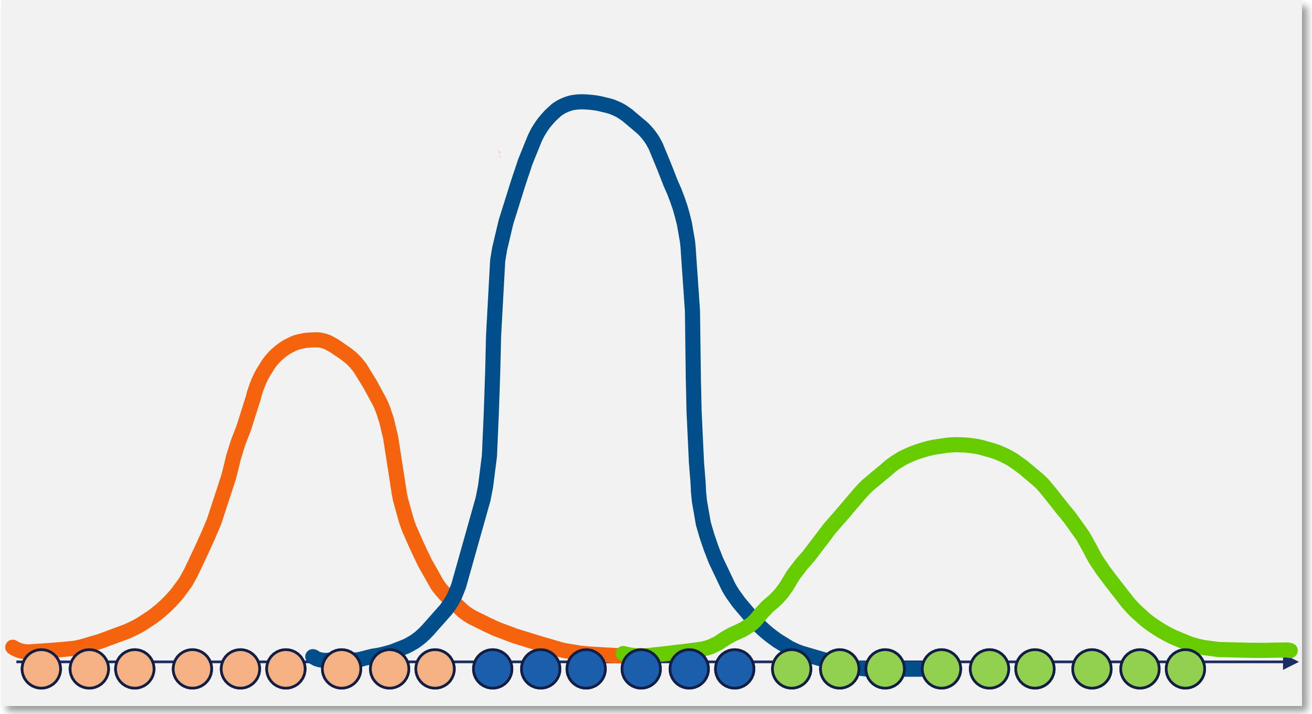


Key ideas

- Groups observations in terms of their characteristics
- Main task of exploratory data mining
- Clustering is an art rather than Science

Gaussian Mixture Model

Visualization



Key ideas

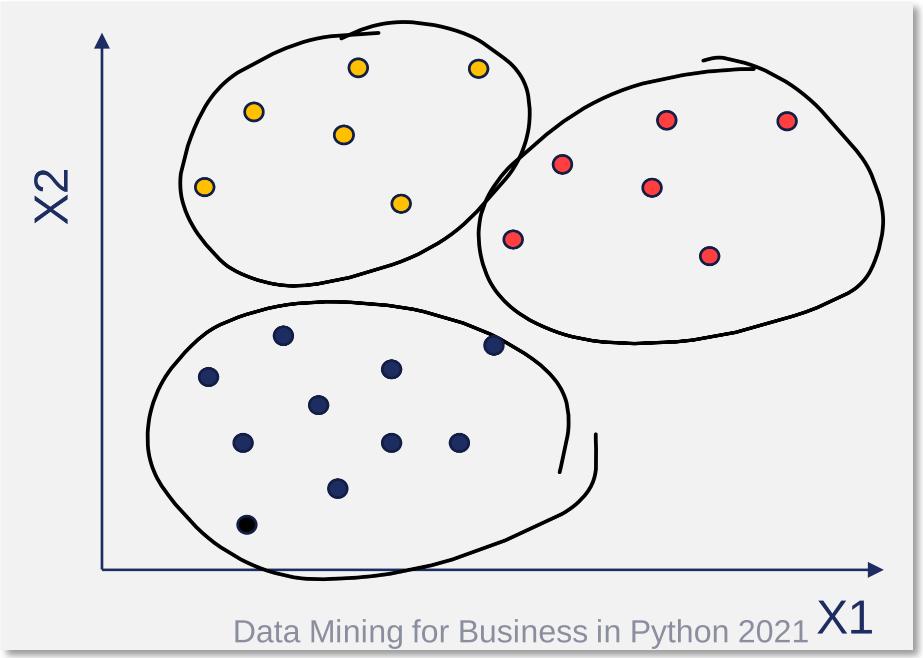
- Gaussian Mixture Model is a probabilistic method for clustering
- Better to use than traditional clustering algorithms, like Kmeans
- The probabilities allow to better evaluate edge cases

Gaussian Mixture Model vs. Kmeans

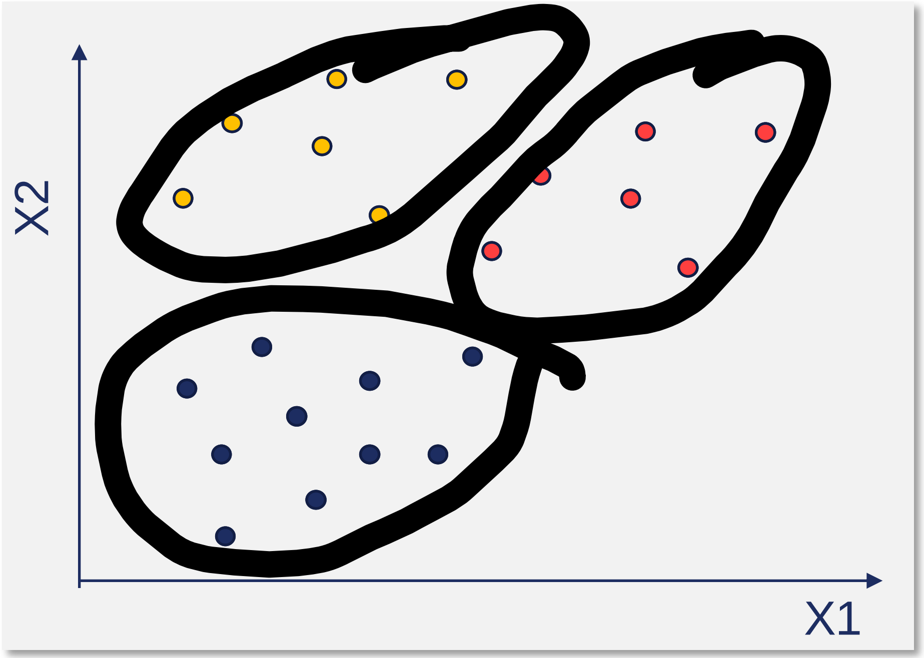
Key ideas

- No need to standardize data
- The cluster sizes do not have specific structures that might or might not apply.
- Faster to compute
- Poor at dealing low amount of data points

Kmeans



Gaussian Mixture Model

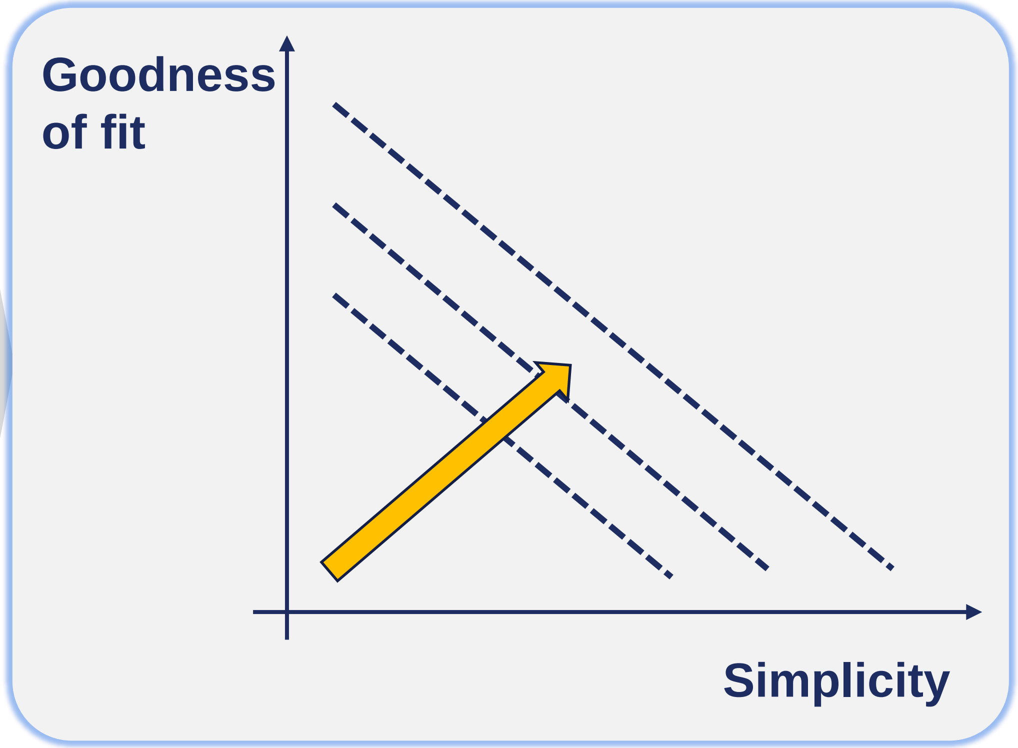


Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Key Ideas

- AIC and BIC helps us determining the optimal number of clusters
- AIC and BIC provide a means to select a model
- Trade-off between simplicity and goodness of fit
- Deal with overfitting and underfitting

Pseudo-visualization



Gaussian Mixture Model Step by Step

Prepare Dataset

Find Optimal Clusters

Perform Gaussian Mixture Model

Interpret results

Gaussian Mixture Model extra Resources

Deep dives

On the Number of Components in a Gaussian mixture model
Geoffrey J. McLachlan, Suren Rathnayake

The Infinite Gaussian Mixture Model
Carl Edward Rasmussen, 2000

Challenge – Wine Quality

Description

You are a wanna be Wine Connoisseur, trying to find the best wines for your parties using Data Mining

Challenge[1]

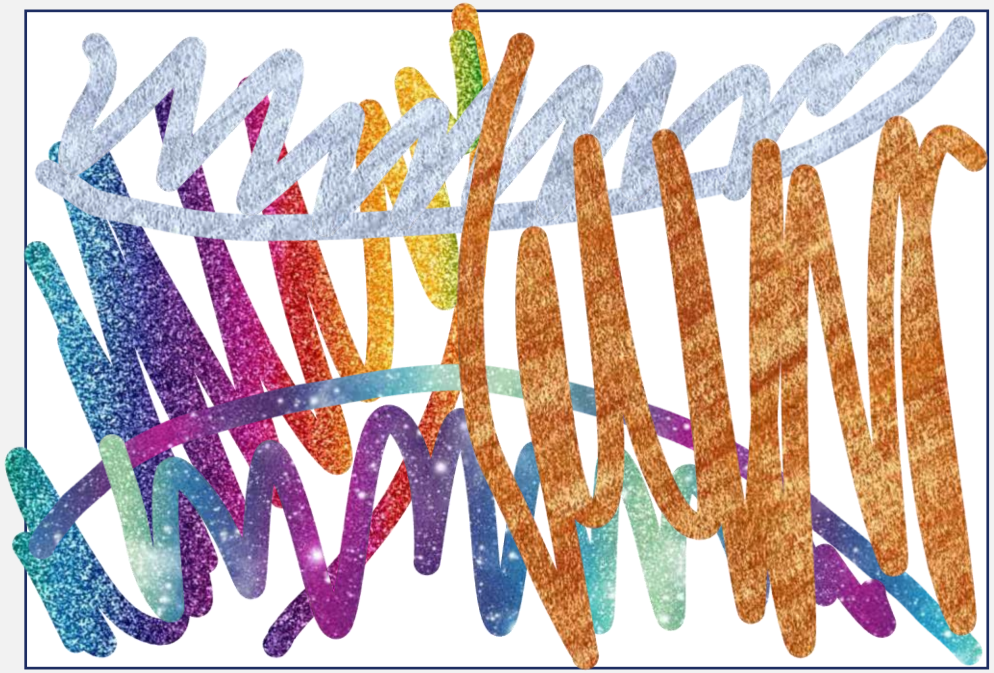
1. Determine the Optimal number of Clusters
2. Perform Gaussian Mixture Model
3. Interpret Results

Paulo Cortez,
University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009

Dimension Reduction

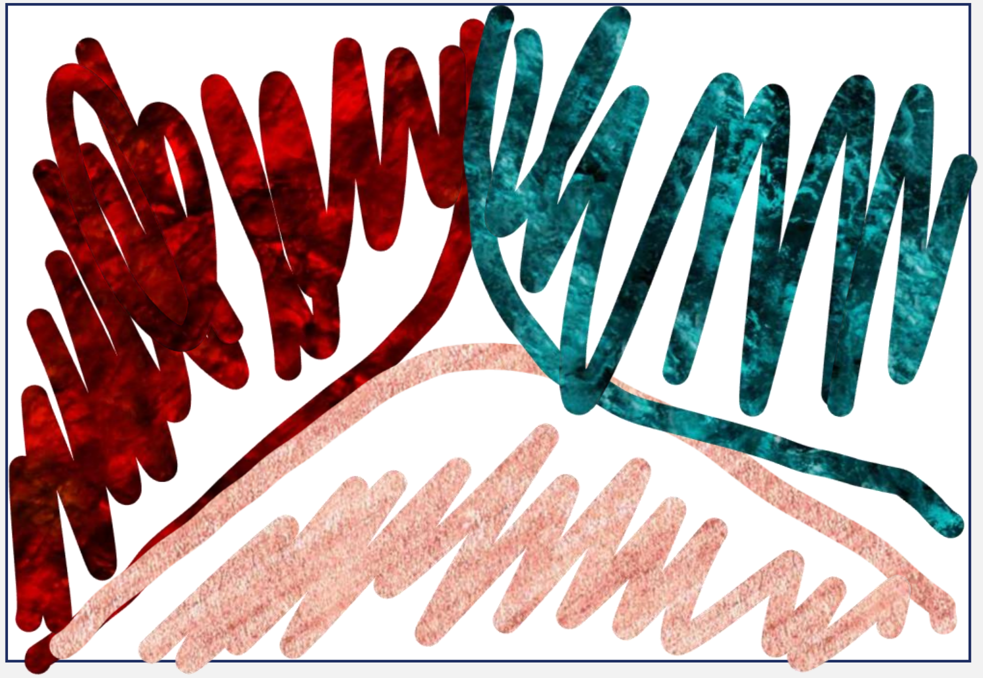
Dimension Reduction Goal

Data set with 4 independent variables



Variance

Components after Dimension Reduction



Variance

You have more information than you need

Problem statement

Dimension Reduction helps to solve

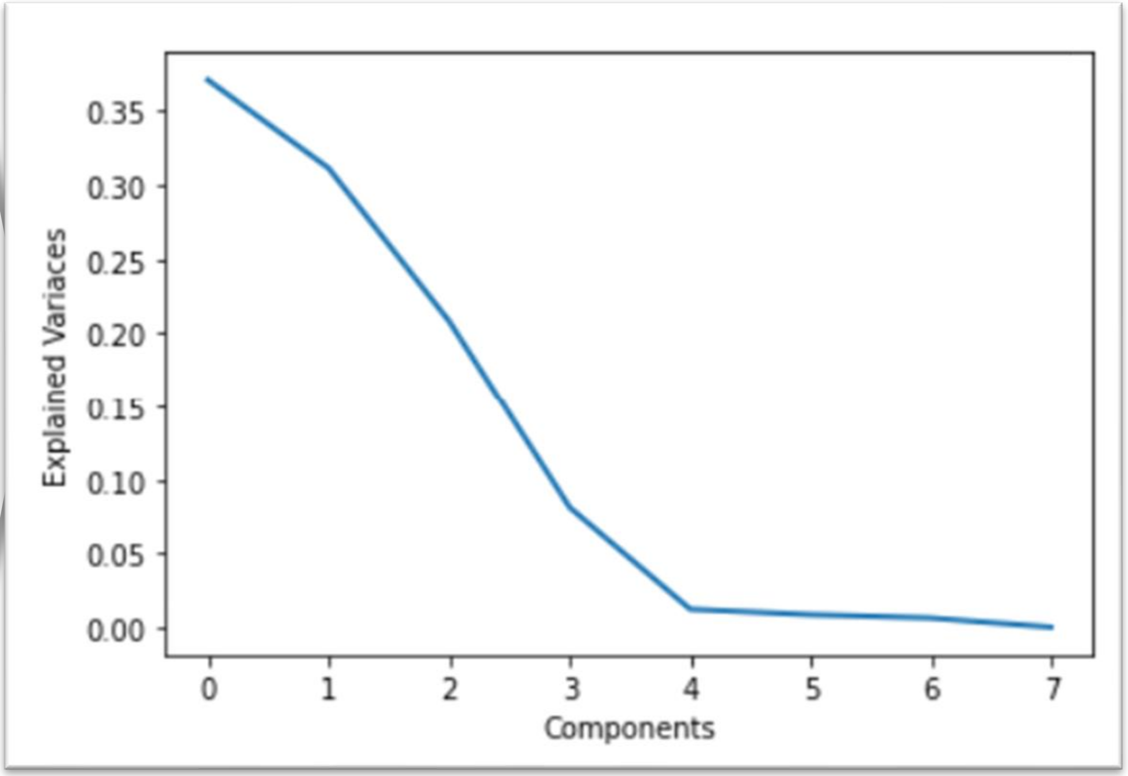
1. Multicollinearity issues
2. Computational issues of large number of predictors
3. Noisy models due to overfitting
4. Create new variables (called components)
5. Pre processing data for predictive models or forecasting

What is Principal Component Analysis?

Key Ideas

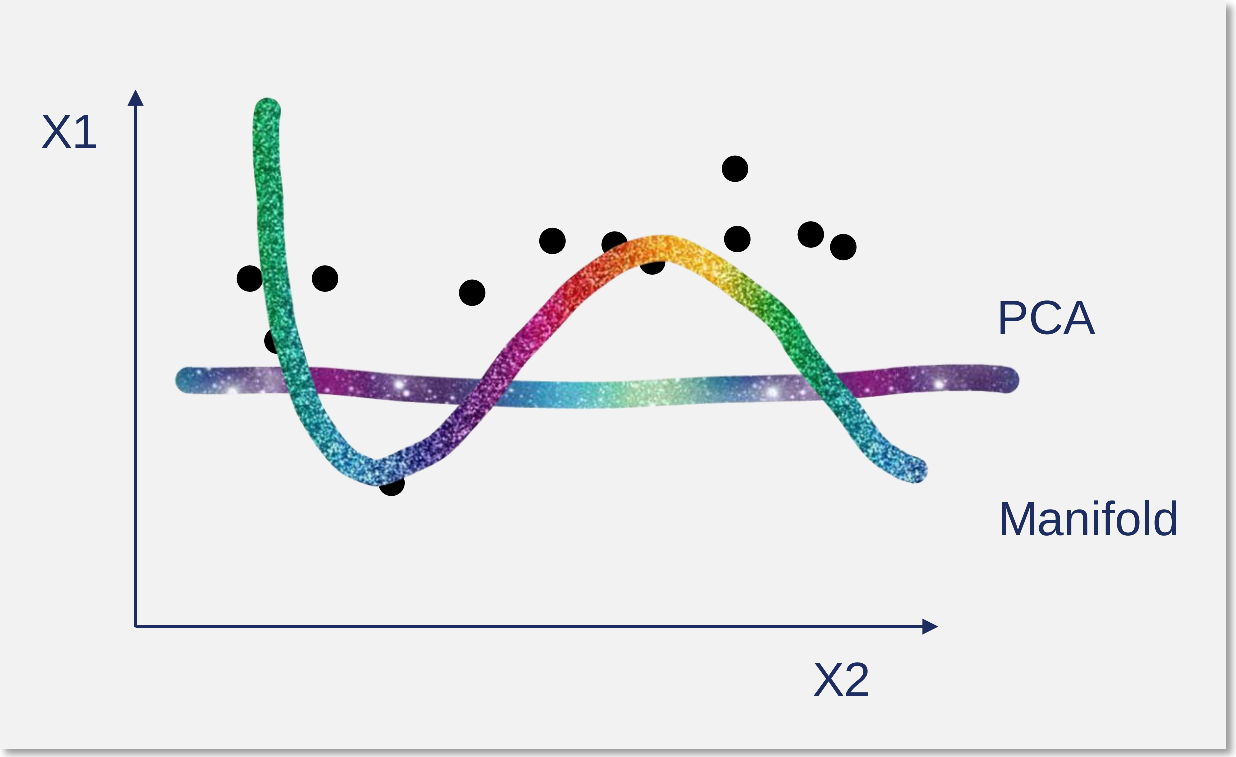
- An algorithm for Dimension Reduction
- Linearly Transforms variables into components
- Components can be determined by the percentage of variance explained
- Choosing Components is more of an art than a science

Visualization



PCA vs Manifold

Visualization



Key ideas

- There are inherent curves in the relationship among the data that have information
- Methods like PCA cannot absorb that information because of their linearity
- No need to standardize data
- Con: Manifold is less interpretable than PCA
- Con : No good quantitative way of determining components.
- There are several algorithms for Manifold. We will use t-SNE

Pros and Cons t-SNE (t-Distributed Stochastic Neighbor Embedding)

+

−

Excellent in high dimensional datasets 1

Focuses on preserving local structures 2

Easy implementation 3

1 Very Computationally intensive

Dimension Reduction extra Resources

Deep dives

Principal component analysis
Herve Abdi ´ and Lynne J. Williams

What is principal component analysis?
Markus Ringnér

Algorithms for manifold learning
Lawrence Cayton

Large-Scale Manifold Learning
Ameet Talwalkar, Courant Sanjiv Kumar, and Henry Rowley

Challenge - Abalone

Challenge[1]

Description

The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope - a time-consuming task. Other measurements, which are easier to obtain, are used to predict the age.

1. Transform gender variable and remove rings variable
2. Perform Correlation Matrix and Standardize data
3. Find Optimal Number of Clusters
4. Perform PCA and interpret components
5. Perform t-SNE and visualize results

Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994)
"The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait",
Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)

Association Rule Learning – Apriori

Case Study Briefing - Groceries

Transaction Data

Data customer grocery shopping purchases

Case study[1]

1. We have a file with almost 10k transactions
2. We need to find patterns in our data to maximize baskets
3. Perform Association Rule Learning

Michael Hahsler, Kurt Hornik, and Thomas Reutterer (2006)
Implications of probabilistic data modeling for mining association rules.
In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nuernberger, and W. Gaul, editors, From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization, pages 598–605. Springer-Verlag.

Association Rule Learning Step by Step

Prepare Dataset

↓

Define Support

↓

Define Confidence

↓

Execute Association Rules Learning

↓

Visualize Results

The output of Association Rule Learning Algorithm

If…		Then…
Game of Thrones	→	Lord of the rings
Burger	→	Fries
Jay-Z	→	Kanye

Key Ideas

The output is an **If…then…** type of analysis

Association Rule Learning is a very simple recommender system

Concepts you need to know - Support

Methodological background

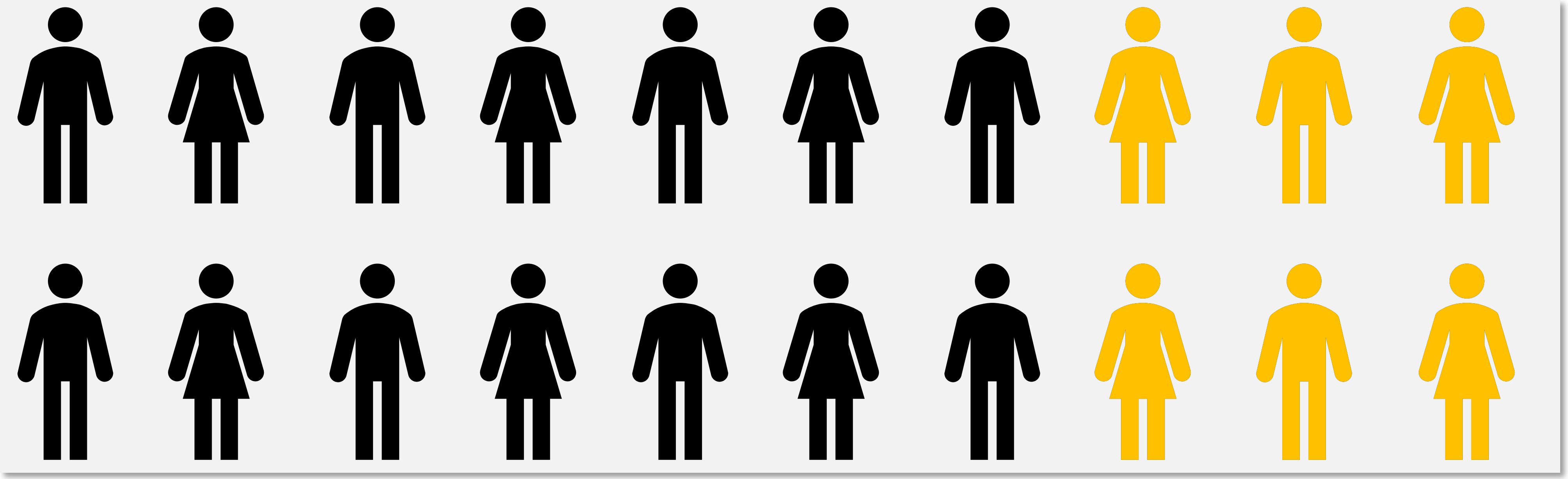
$$Support\ (Burgers) = \frac{\#\ Transactions\ with\ Burger}{Total\ Transations}$$

To consider

- It does not matter if Burgers happen more than once per transaction
- Support indicates the Relevance of the item

Burger Support Visualization

Visualization



Data

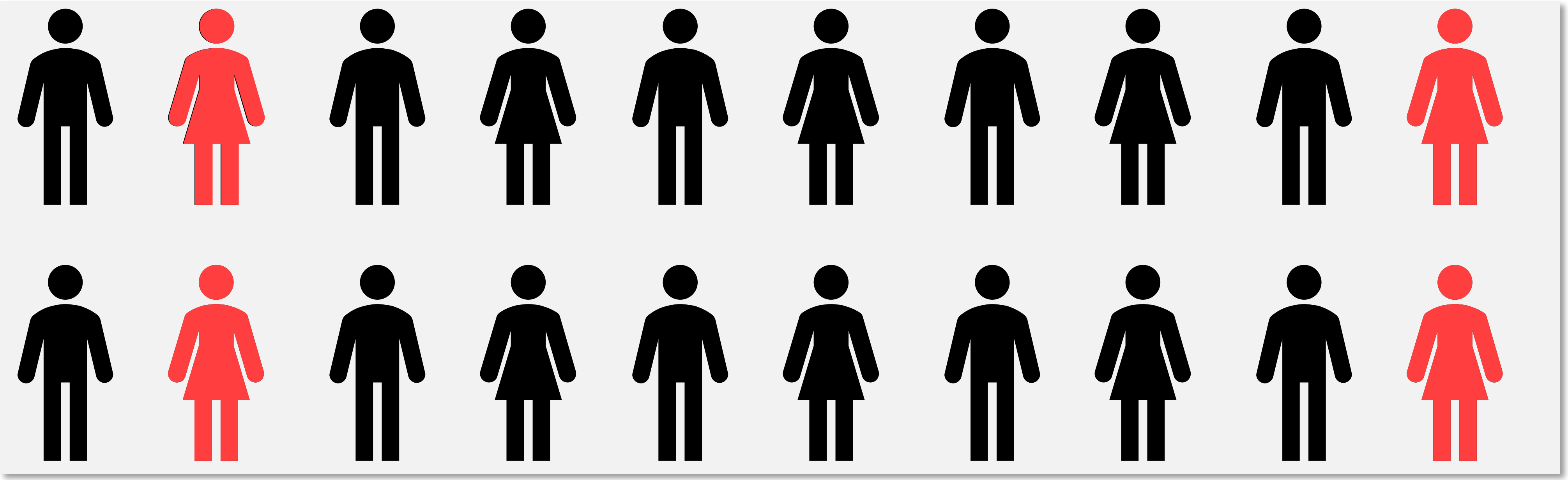
- Population is 20
- 6 people like burgers

Formula

$$Support\ (Burgers) = \frac{6}{20} = 30\%$$

Mayo Support Visualization

Visualization



Data

- Population is 20
- 4 people like Mayo

Formula

$$Support\ (Mayo) = \frac{4}{20} = 20\%$$

Concepts you need to know - Confidence

Methodological background

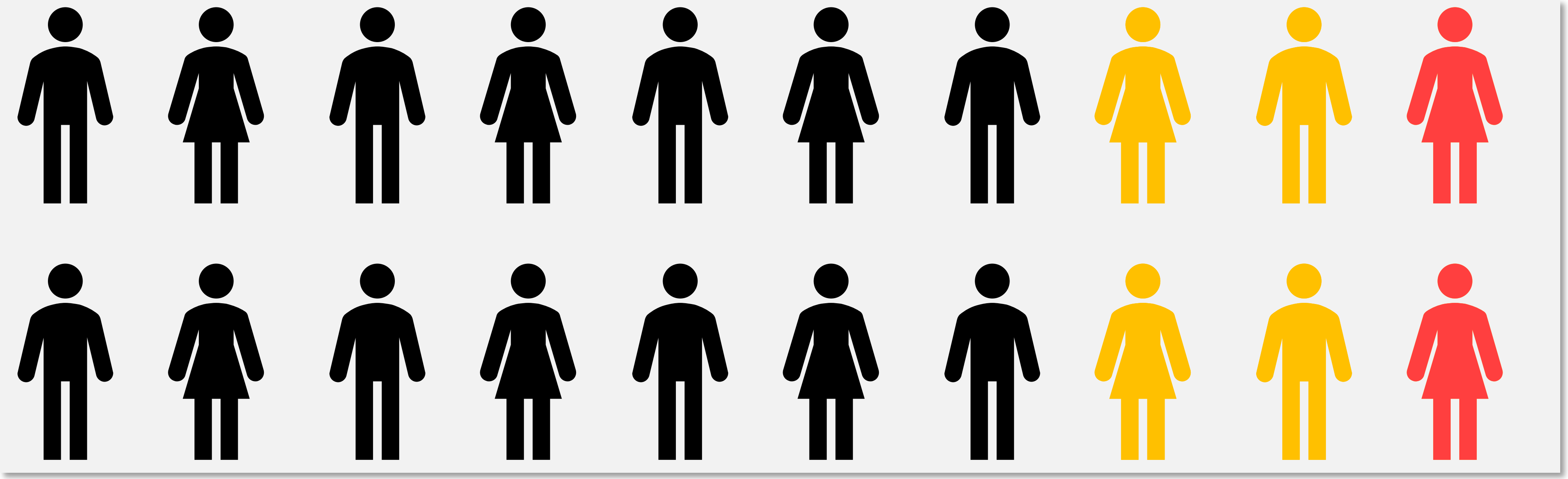
$$Confidence\ (Mayo|burgers) = \frac{\#\ Transactions\ with\ Burger\ \&\ Mayo}{Total\ Burger\ Transactions}$$

To consider

- It does not matter if Burgers or Mayo happen more than once per transaction
- Confidence indicates the strength of the relationship

Confidence Visualization

Visualization



Data

- Population is 20
- 6 people like burgers
- Of the 6, 2 like Mayo

Formula

$$Confidence\ (Mayo|burgers) = \frac{2}{6} = 33\%$$

Concepts you need to know - Lift

Methodological background

$$Lift\ (Mayo|burgers) = \frac{Confidence\ (Mayo|burgers)}{Support\ (Mayo)}$$

Key idea

- Lift measures the likelihood of buying Mayo and Burgers together vs. Just buying Mayo
- Lift bigger than 1 means increased likelihood to buy

Apriori is an Association Rule Learning Algorithm

What is it?

Key characteristics

1. Mines frequent itemsets for Boolean Association Rules
2. Works by finding items that have occurred a minimum number of times (Support)
3. And the corresponding itemsets that pass a certain cut-off (confidence

Limitations

1. Slow in processing Itemsets
2. Only allows Boolean values

Association Rule Learning extra Resources

Deep dives

Online Association Rule Mining
Christian Hidber

Association Rule Mining: A Survey
Qiankun Zhao Nanyang and Sourav S. Bhowmick

Algorithms for Association Rule Mining – A General Survey and Comparison
Jochen Hipp, Ulrich Guntz, and Gholamreza Nakhaeizadeh

Challenge - NYC restaurants cuisine, borough and sanitary grade

Challenge[1]

Description

You have a dataset with NYC restaurants, their boroughs and sanitaty grade

1. Create a list with the transactions
2. Encode the transaction list into a Dataframe
3. Perform Association Rules Learning. Play around with support and confidence
4. Visualize the results

Random Forest

You were hired to figure out which the main drivers of customers that sign up to a savings account in a bank

Problem Relevance

Description

Customer churn:
Calling a customer who cannot sign up can lead for he/she to unsubscribe

Opportunity cost:
Sending to wrong product for the customer to sign up can create a loss in the case the customer would be interesting to sign up for another

Relevance:
Sending constinuously information that the customer is not interested can potentially lead for lower open rate willingness in the future

[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Random Forest Step by Step

Prepare Dataset

↓

Split into training and test set

↓

Perform Random Forest

↓

Predict using the Random Forest

↓

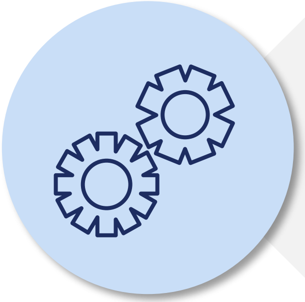
Model Assessment

↓

Execute Driver Importance

Random Forest is an Ensemble Learning Algorithm

What is it?

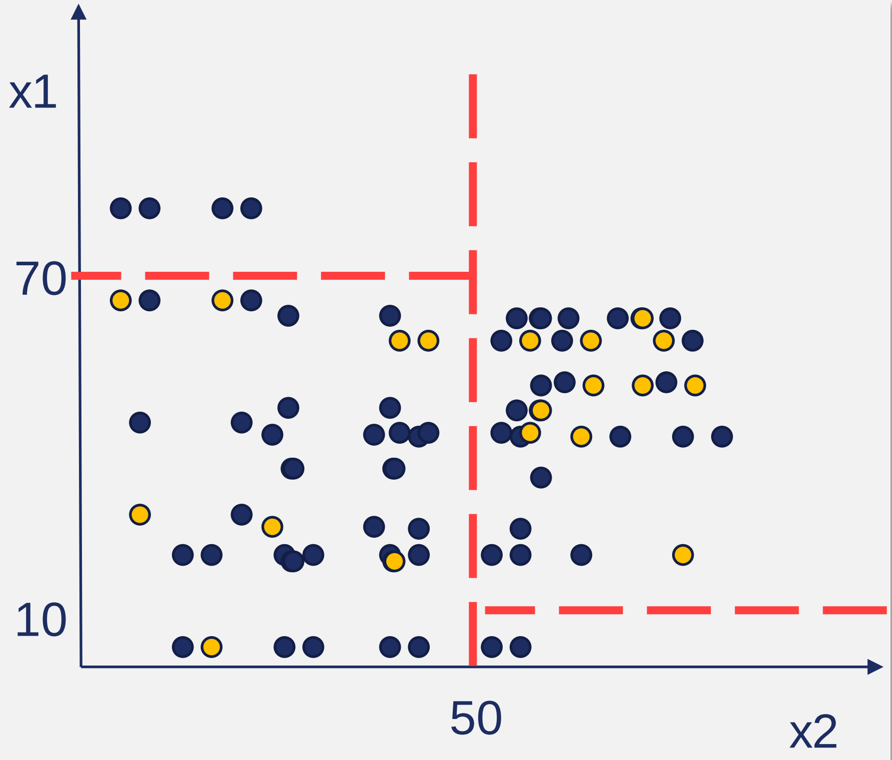


Description

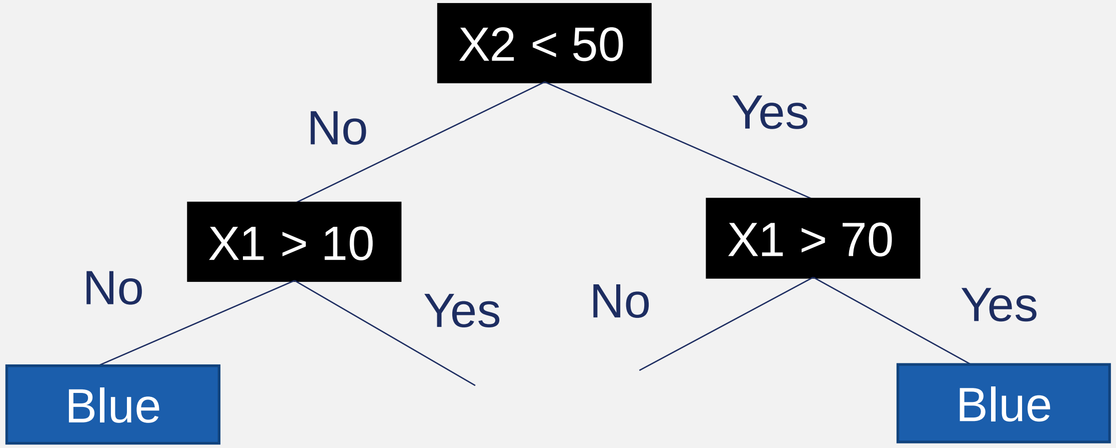
1. Ensemble Learning is when you have a plurality of models predicting your output
2. In simple words, ensemble is an average of Models
3. A Random Forest is a combination of decision trees

How do Decision trees work?

Visualization



Decision tree



Key Ideas:

- A split or leaf is done taken a maximum entropy logic
  - Where would it yield more information
- The prediction would be done based on the relative frequency

Random Forest is an Ensemble Learning Algorithm

What is it?

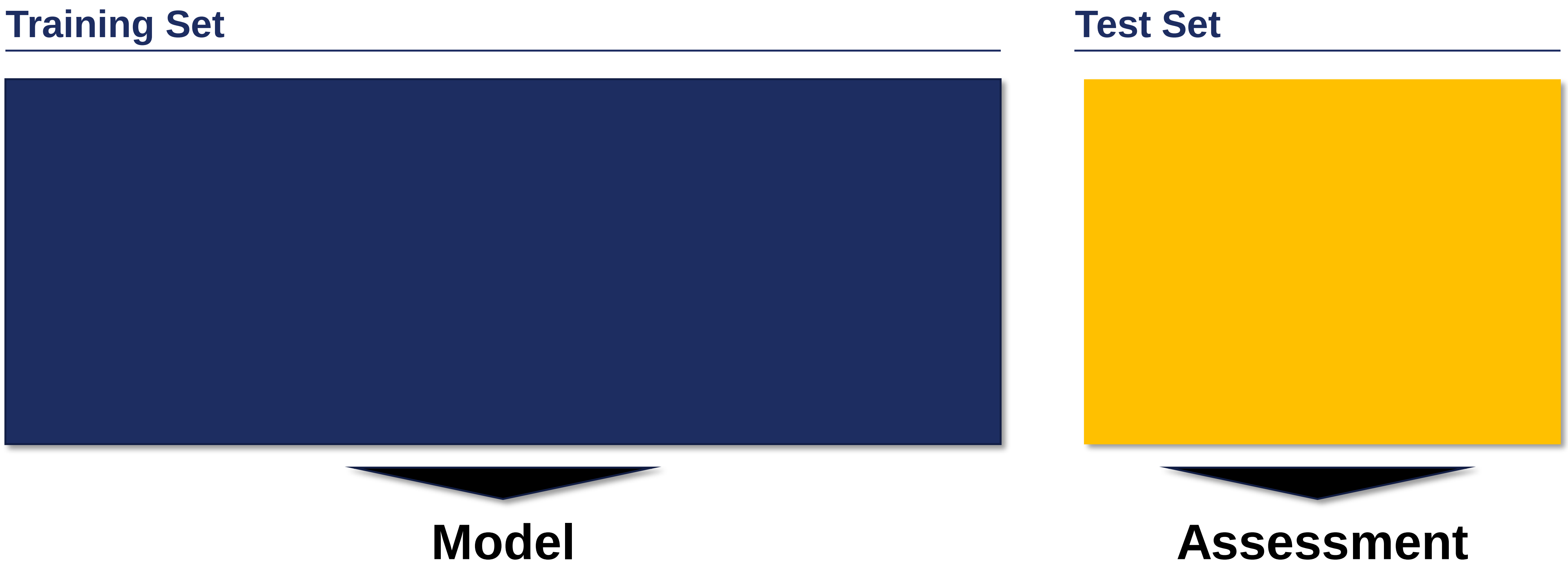
Description

1. Ensemble Learning is when you have a plurality of models predicting your output
2. In simple words, ensemble is an average of Models
3. A Random Forest is a combination of decision trees
4. Can be used for Regression and Classification problems
5. Random Forests have a tendency to overfit

Let‘s imagine this is our full data set

Description

Splitting between training and test enables an unbiased model assessment



The Confusion Matrix allows to access the results of a classifier

Confusion Matrix

		Truth	
		False	True
Predicted	False	True negative	False Negative
	True	False Positive	True positive

Accuracy

- Accuracy = (True positive + True negative) / All
- Used when we have balanced dataset

Sensitivity or Recall

- True positive / (true positive + false negative)
- Used when skewed towards False values

Specifiticy or False Positive Rate

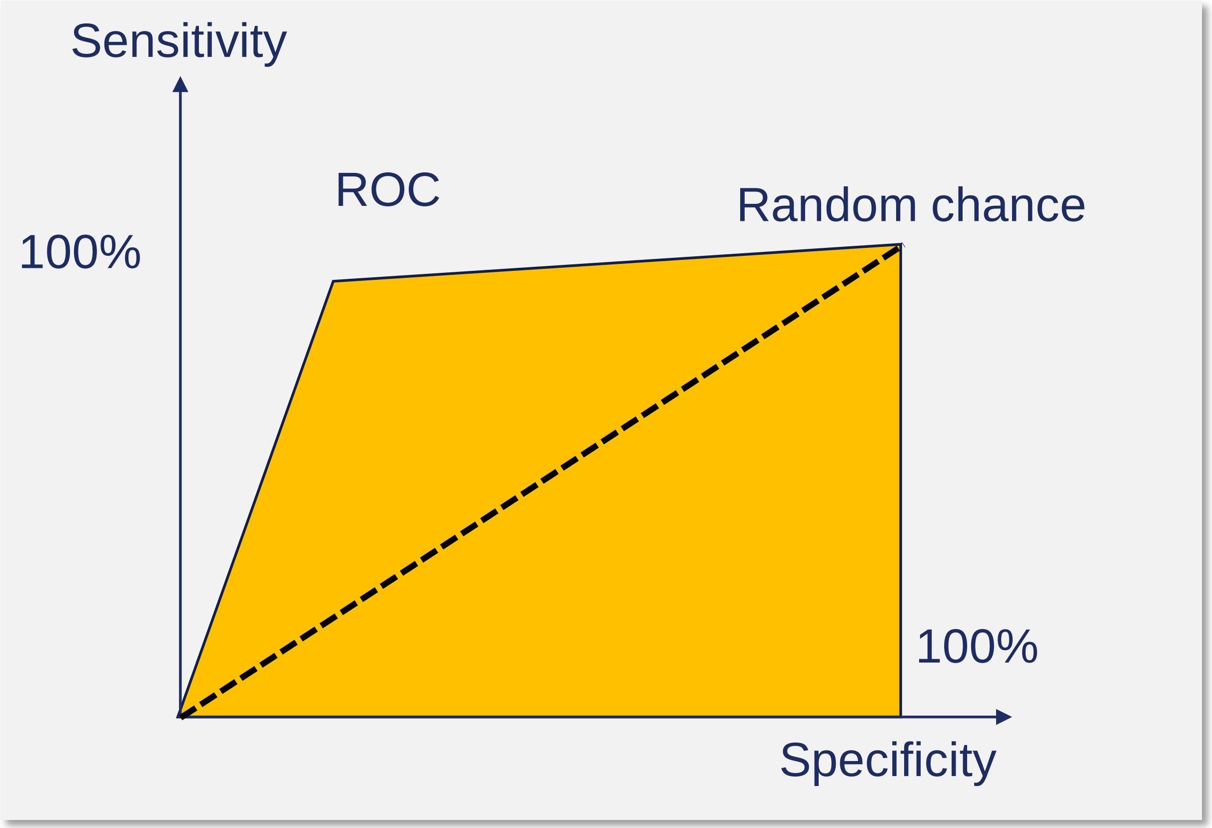
- True negative / (true negative + false positive)
- Used when skewed towards True values

Precision

- True Positive / (true positive + false positive)
- Used when skewed towards False values

Area under the ROC curve (AUC)

Visualization



Key ideas

- AUC is a performance measure for classification problems
- It tells us how well the model is able to distinguish between positives and negatives

The F1 score should be used when we have an unbalanced dataset

Confusion Matrix

		Truth	
		False	True
Predicted	False	True negative	False Negative
	True	False Positive	True positive

Sensitivity or Recall

- Accuracy = (True positive + True negative) / All
- Used when we have balanced dataset

Precision

- True Positive / (true positive + false positive)
- Used when we are skewed towards True values

F-score

- 2 * (precision * recall) / (precision + recall)
- Used for unbalanced dataset

Random Forest extra Resources

Deep dives

How Many Trees in a Random Forest?
Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranauskas

Random forest classifier for remote sensing classification
M. Pal

Real-Time Human Pose Recognition in Parts from Single Depth Images
Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake

A Random Forest Guided Tour
Gérard Biau and Erwan Scornet

Random Forest Challenge – Extramarital affairs

Challenge[1]

A Theory of Extramarital Affairs

Key characteristics of cheaters

1. Isolate X and Y
2. Transform Y into binary format
3. Create a dummy variable out of the occupation variable
4. Transform X string variables into dummies
5. Perform Random Forest
6. Create Importance drivers

LIME

Interpreting Advanced Machine Learning Models

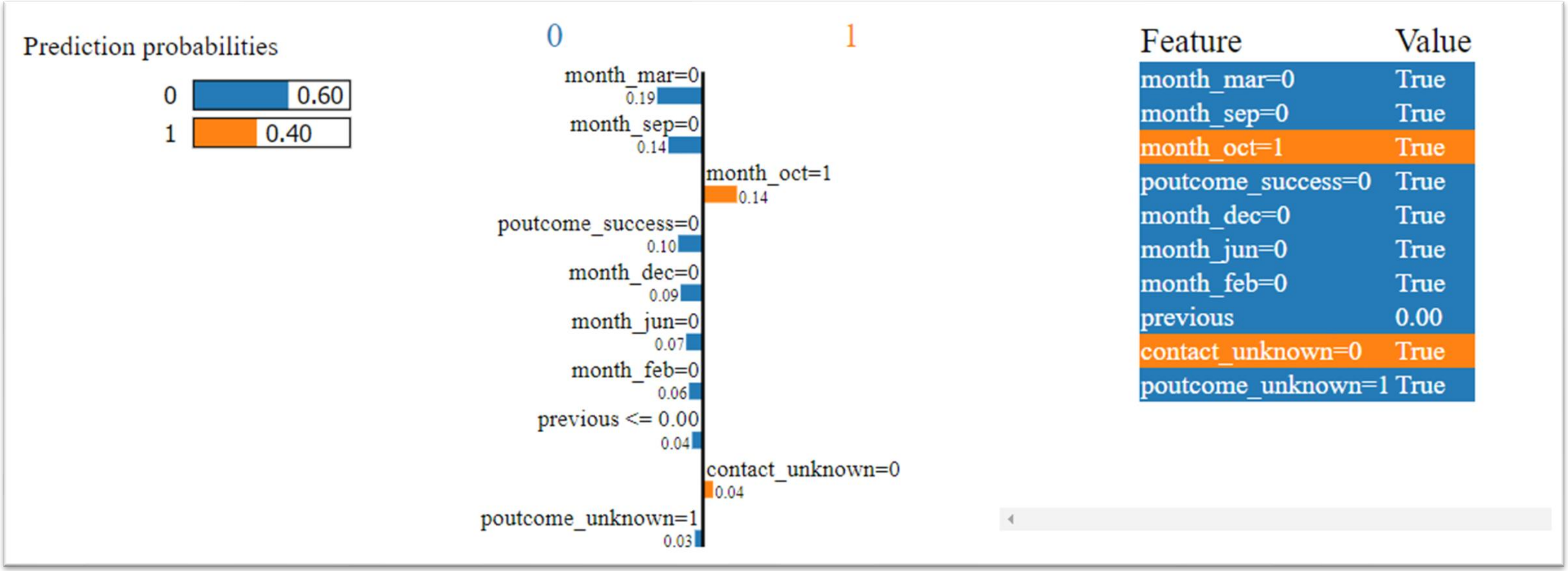
Problem Statement

- How do we explain Advanced Machine Learning models?
- How can we trust something that does explain itself?
- From a Data Mining perspective, it feels like a great loss to not be able to take advantage of the Data Science newer algorithms

Introducing Lime

- Local interpretable model-agnostic explanations -> works with most models
- LIME is the application of surrogate models
- Surrogate models are trained to approximate the predictions of the underlying black box model
- LIME is best applied to Classification problems!
- LIME focus is on explaining individual predictions

LIME explanation example



LIME extra Resources

Deep dives

Model-Agnostic Interpretability of Machine Learning
Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin

Statistical stability indices for LIME: obtaining reliable explanations for Machine Learning models
Giorgio Visania,b, Enrico Baglib , Federico Chesania , Alessandro Poluzzib and Davide Capuzzo

Challenge – Understanding Remote Work predictions

Challenge[1]

Stackoverflow dataset

Worker‘s characteristics, and job related queries

1. Install LIME
2. Transform string variables
3. Isolate X and Y
4. Perform Random Forest
5. Prepare LIME explainer
6. Use LIME to explain a couple of instances

SHAP

Case Study Briefing – Car prices

Case study[1]

Pricing a car

List of cars, their price, and characteristics

1. Build a XGBoost model to measure accuracy

2. Use SHAP to get insights

XGBoost and SHAP step by step

Prepare dataset, isolate X and Y

Split into Training and Test Set, and create Matrices

Set Parameters

Run XGBoost

Assess Model

Implement SHAP

XGBoost is a state-of-art Machine Learning Algorithm

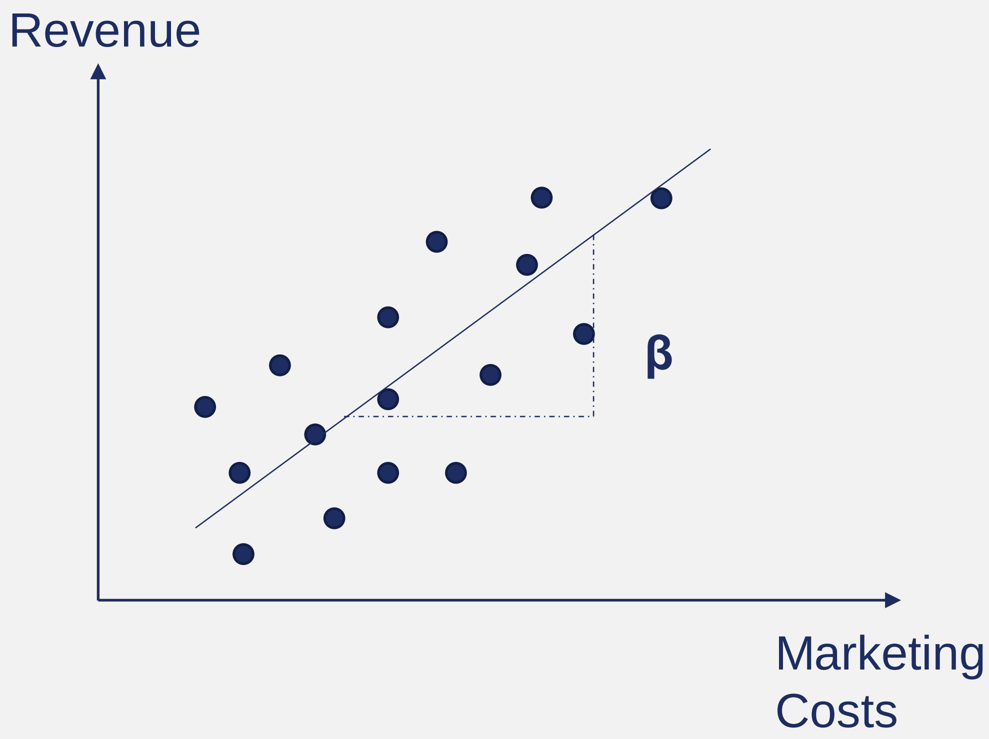
What is it?

Description

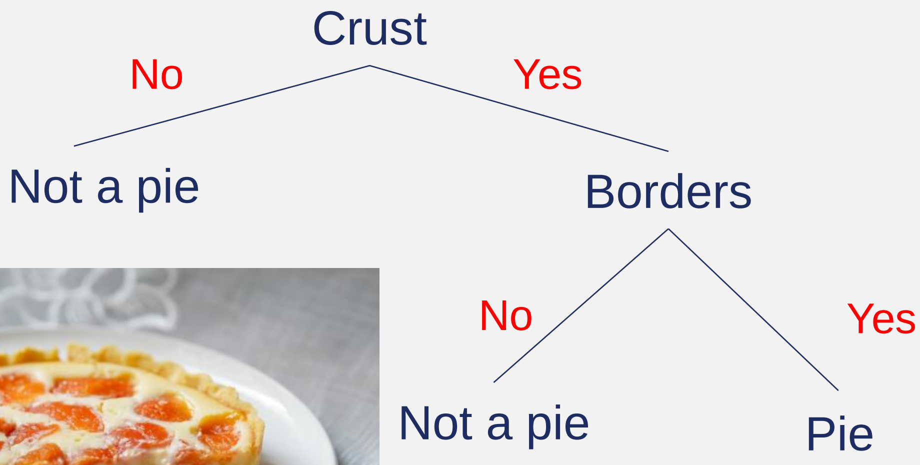
1. Stands for Extreme Gradient Boosting
2. Can be contructed with a tree based algorithm or linear (worse results)
3. It is an emsemble algorithm
4. Each new model is built upon the precedent one -> continuous improvement
5. Can be used for both Regression and Classification

Linear vs Decision Trees

Linear Approach



Decision Tree



XGBoost gives different weights depending on how difficult it is to predict

First Tree

Outcome		Predictor	Weight
✓ 1	←	X	25%
✓ 0	←	X	25%
✗ 0	←	X	25%
✗ 1	←	X	25%

Second Tree

Outcome		Predictor	Weight
✗ 1	←	X	20%
✓ 0	←	X	20%
✗ 0	←	X	30%
✓ 1	←	X	30%

Third Tree

Outcome		Predictor	Weight
✗ 1	←	X	23%
✓ 0	←	X	15%
✓ 0	←	X	35%
✓ 1	←	X	27%

XGBoost looks at parts of the observations at a time

First Tree

	Outcome		Predictor	Weight
✓	1	←	X1	25%
✓	0	←	X2	25%
✗	1	←	X4	25%

Second Tree

	Outcome		Predictor	Weight
✗	1	←	X1	20%
✓	0	←	X2	20%
✗	0	←	X3	30%

Third Tree

	Outcome		Predictor	Weight
✗	1	←	X1	23%
✓	0	←	X3	35%
✓	1	←	X4	27%

Key Idea

XGBoost only looks at a fraction of the observation at the time

Observations that are more difficult to predict are given a bigger weight

The logic is similar for Regression-based tasks

First Tree

Error	Outcome		Predictor	Weight
- 5	15	←	X1	33%
2	22	←	X2	33%
4	34	←	X4	33%

Second tree

Error	Outcome		Predictor	Weight
- 1	19	←	X1	40%
-1	25	←	X2	30%
3	35	←	X4	35%

XGBoost also gives different weights to different predictors

First Tree

Error	Outcome	X1	X2	X3	Weight
-5	15	50%	50%		33%
2	22				33%
4	34				33%

Second Tree

Error	Outcome	X1	X2	X3	Weight
-1	19	50%		50%	40%
-1	25				30%
3	35				35%

Third Tree

Error	Outcome	X1	X2	X3	Weight
1	21		40%	60%	35%
0	24				30%
2	36				40%

Key Idea

Predictors also have different weights if they yield different model results

XGBoost quirks

Description

Which?

NA:
Unlike other regression models, XGBoost treats NA‘s as information

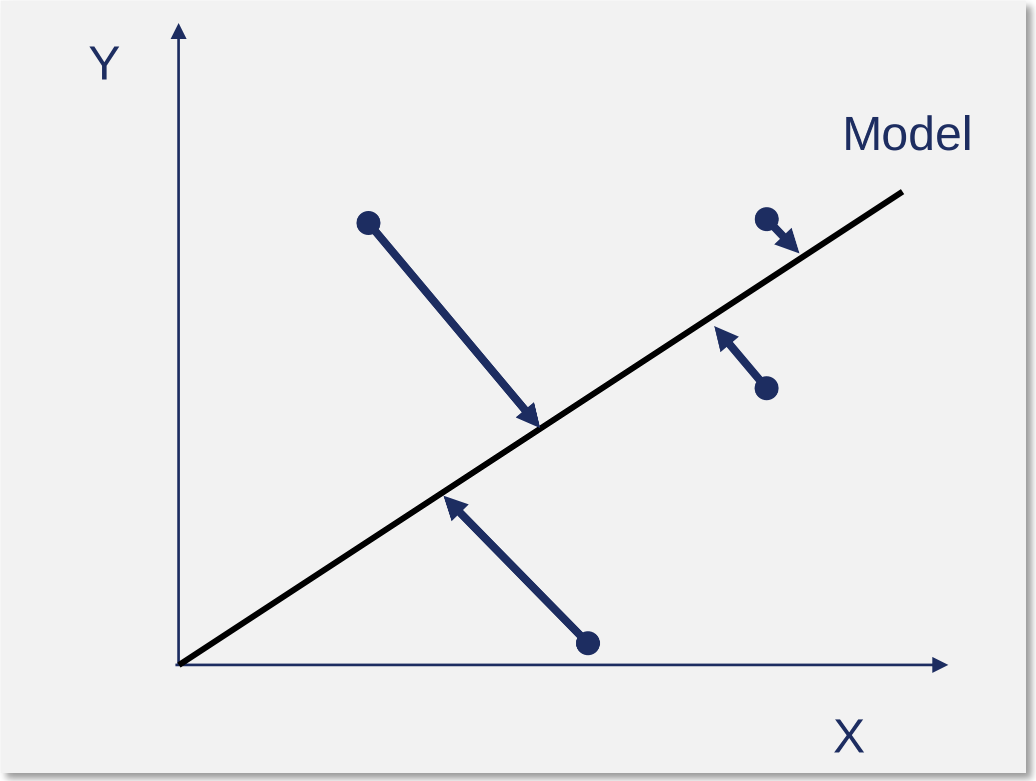
Non-linearity:
XGBoost is excellent dealing with non-linearity relationship between the dependent and the independent variables.

XGBoost has 7 main tuning parameters

Parameter	Description
Minimum Child weight	Relates to the sum of the weights of each observation. Low values can mean that maybe not a lot of observations are in the round
ETA	Learning Rate. How fast do you want the model to learn?
Max depth	How big should the tree be? Bigger trees go into more detail
Gamma	How fast should the tree be split?
Subsample	Share of observations in each tree?
Colsample by tree	How much of the tree should be analysed per round?
Number of rounds	How many times do we want the analysis to be run?

Mean Absolute Error (MAE) vs Root Squared Mean Error (RSME)

Visualization



Key ideas

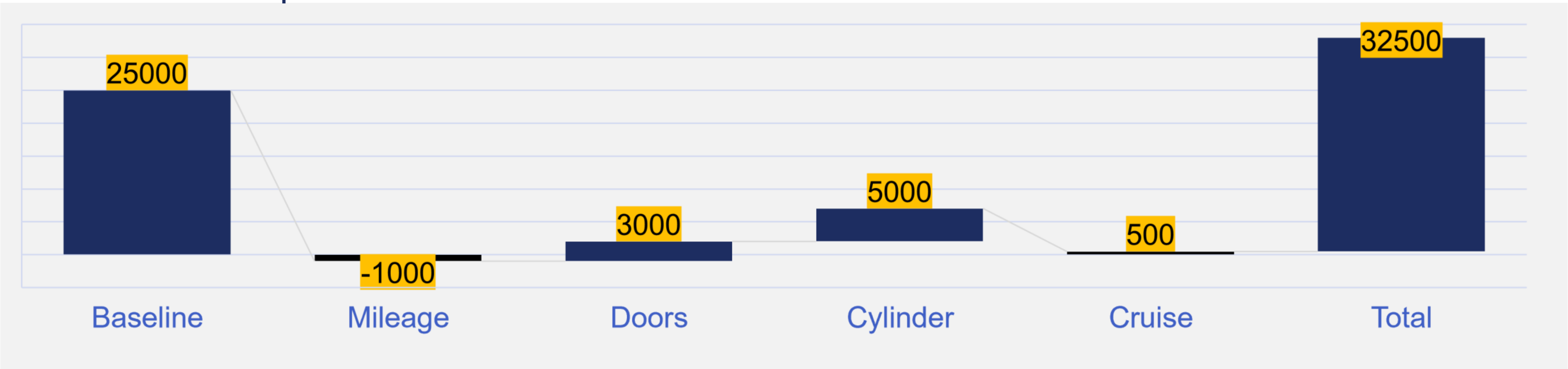
- MAE and RSME are performance indicators for Regression models with continuous dependent variables

$$MAE = \frac{\sum |y - \hat{y}|}{n} \qquad RSME = \sqrt{\frac{\sum (\hat{y} - y)^2}{n}}$$

- RSME is quite useful for models with extremes / outliers
- MAE is more interpretable.

Introduction to SHAP

- SHapley Additive exPlanations were introducted by Lundberg and Lee (2016)
- SHAP aims to explain each instance by computing the marginal contribution of each feature to the prediction



- SHAP computes each value using coalitional game theory

There are 3 main areas of insights

Global Interpretability

- The SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable

Local Interpretability

- Each observation gets its own set of SHAP values
- We can explain why a case receives its prediction and the contributions of the predictors

Dependency Plots

- Shows the relations between an independent variable and the output
- Also shows how the predictor interactor with its closest independent variable

XGBoost and SHAP extra Resources

Deep dives

XGBoost: A Scalable Tree Boosting System
Tianqi Chen and Carlos Guestrin

A Unified Approach to Interpreting Model Predictions
Scott M. Lundberg and Su-In Lee

Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis
Amir Bahador Parsaa, Ali Movahedia, Homa Taghipoura, Sybil Derribleb, and Abolfazl (Kouros) Mohammadian

Challenge – Understanding house price drivers

Dataset with house characteristics and prices

Challenge[1]

1. Install SHAP and import libraries
2. Transform string variables
3. Isolate X and Y, and generate XGBoost matrix
4. Set parameters and run XGBoost
5. Local interpretability
6. Dependency plots
7. Global interpretability