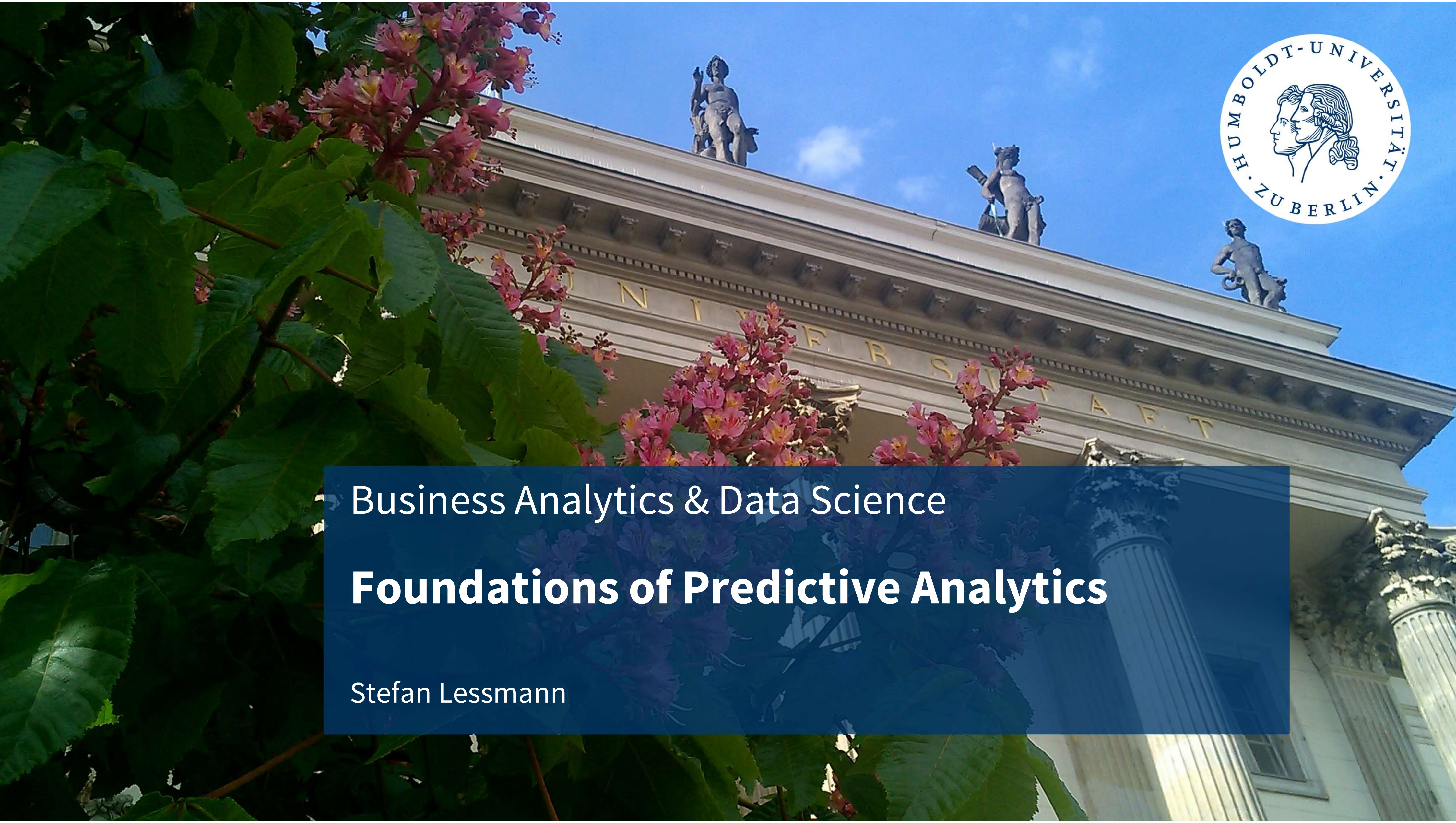


Business Analytics & Data Science

Foundations of Predictive Analytics

Stefan Lessmann

Agenda

Introduction
Labeled & unlabeled data, principles of prediction, regression & classification

Predictive Analytics Applications in Business
Use cases in business, return prediction case study, lessons learnt

Summary

Introduction

Labeled & unlabeled data, principles of prediction, regression & classification

Recap: Data Science Models and Algorithms

Modeling approaches

Predictive

Regression

- Linear Regression
- Neural network
- CART
- …

Classification

- Logistic regression
- Nearest neighbor
- Boosting
- …

Descriptive

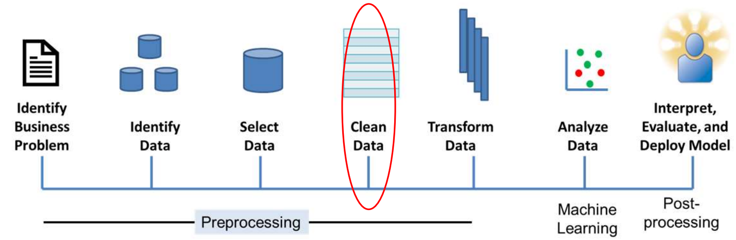
Segmentation

- Hierarchical
- K-means
- DBSCAN
- …

Rule/sequence mining

- Apriori-algorithm
- FP-growth
- PrefixSpan
- …

Recap: Structured Tabular Data



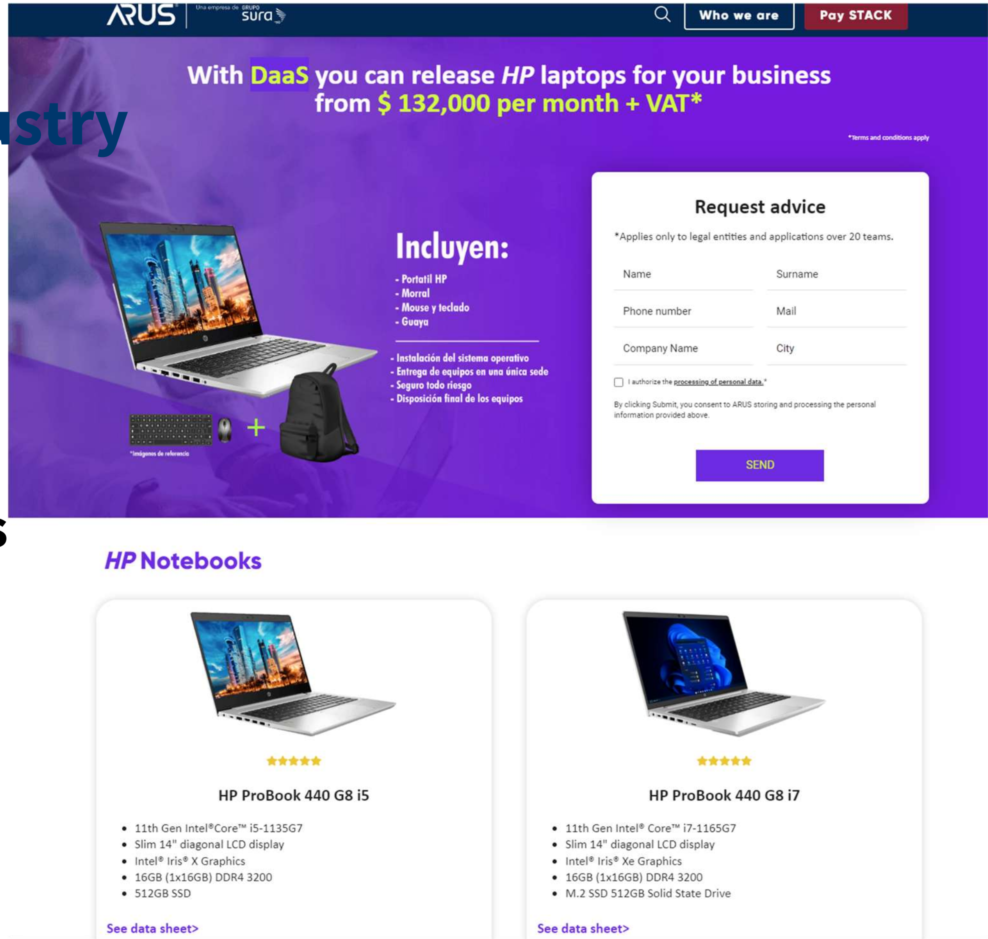
Cases / observations / examples

Age group	Gender	No. of orders	No. of returns	Avg. order volume	Total purchases	…
<18	M	3	1	7	€150	…
18-29	M	1	0	13	€75	…
<18	F	5	2	5	€33	…
30-50	M	2	0	2	€24	…
>50	F	1	0	25	€120	…
19-29	F	3	1	17	€41	…
>50	F	9	1	9	€284	…
18-29	M	2	2	14	€10	…
<18	F	1	0	11	€18	…

Variables/ characteristics / attributes/ features / predictors/ covariates

Recap: **Business Use Case II: Leasing Industry**

Service Provider for IT Leasing



- **Clients lease IT equipment for given period**
- **Provider receives monthly fee**
- **Client returns the item when contract expires**
- **Provides resales the used item in the second-hand market**
- **Business question:**

  how to price leasing contracts?

- **Data Science support**
  - Depreciation is the main unknown in the calculation of costs and, by extension prices
  - Forecast the resale price of used items in the second-hand market

Predictive Analytics (aka Supervised Machine Learning)

Estimate functional relationship between features and a target

- **Data includes features and a target variable**
  - Numerical target variable → regression
  - Discrete target variable → classification
- **Regression example: resale price forecasting in leasing**

Target, outcome, label, response (variable), dependent (variable)

Product	List price [$]	Age [month]	Industry	…	Observed resale price [$]
Dell XPS 15’	2,500	36	Mining	…	**347**
Dell XPS 15’	2,500	24	Health	…	**416**
Dell XPS 17’	3,000	36	Manufacturing	…	**538**
HP Envy 17’	1,300	24	Office	…	**121**
HP EliteBook 850	1,900	36	Manufacturing	…	**172**
Lenovo Yoga 11’	799	12	Office	…	**88**
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Predictive Analytics (aka Supervised Machine Learning)

Estimate functional relationship between features and a target

- **Data includes features and a target variable**
  - Numerical target variable → regression
  - Discrete target variable → classification
- **Classification example: credit risk modelling**

Target, outcome, label, response (variable), dependent (variable)

Bureau score	Collateral	Debt/Income	Years at address	Age	…	Default (e.g., 90 days late)
650	Yes	20%	2	<21	…	**No**
280	No	43%	0	21-29	…	**Yes**
750	Yes	27%	8	30-39	…	**No**
600	Yes	18%	4	40-50	…	**No**
575	No	33%	12	>50	…	**No**
715	Yes	24%	1	21-29	…	**No**
580	No	18%	6	40-50	…	**Yes**

Formalization of Supervised Machine Learning

Resale price forecasting example

- **We aim at forecasting resale prices (our target variable) denoted by Y**
- **We assume that resale prices Y depend on features X**
  - We do not know how exactly resale prices depend on feature values
  - But we have access to historical data $\mathcal{D} = \{Y_i, \boldsymbol{X}_i\}_{i=1}^{n}$ that exemplifies the relationship
- **At decision time (e.g., when forecast is needed), we can observe X but not Y**
- **We use algorithms to learn a model f that maps from features to target $f(\boldsymbol{X}) \longrightarrow Y$**

Product	List price [$]	Age [month]	Industry	…
…	…	…	…	…

$$\boldsymbol{X} = (X_1, X_2, \dots, X_m) \in \mathbb{R}^m$$

$$\boldsymbol{X} \xrightarrow{f(X)} Y$$

Observed resale price [$]
…

$$Y \in \mathbb{R}$$

Two-Stage Paradigm

Characteristic of supervised (and other forms of) ML

Stage 1: Model Training

Data-driven development of a predictive model using labelled data $\mathcal{D} = \{Y_i, \boldsymbol{X}_i\}_{i=1}^{n}$

Training data incl. Y

i	Y	X_1	X_2	…	X_m
1	…	…	…	…	…
2	…	…	…	…	…
…	…	…	…	…	…
n	…	…	…	…	…

Learning Algorithm

Model

Stage 2: Model Testing & Use

Application of trained model to novel data yields output (e.g., forecasts)

New data w/o Y

i	X_1	X_2	…	X_m
$n+1$	…	…	…	…
$n+2$	…	…	…	…
…	…	…	…	…
N	…	…	…	…

Forecasts of Y

i	$\hat{Y}$
$n+1$	…
$n+2$	…
…	…
N	…

Two-Stage Paradigm
Linear regression example

- **Model specification**
  - Continuous target variable
  - Linear, additive relationship
  - Random variation

$$Y = b + \boldsymbol{w}\boldsymbol{X} + \epsilon$$

- **Model estimation**
  - Determine free parameter $\boldsymbol{w}$
  - Find $\widehat{\boldsymbol{w}}$ that maximizes model fit
  - Objective: minimize least-squares loss

$$\widehat{\boldsymbol{w}} \leftarrow \underset{\boldsymbol{w},b}{\operatorname{argmin}} \left(y_i - (b + \boldsymbol{w}\boldsymbol{X}_i)\right)^2, \quad i = 1, \dots, n$$

- **Model**
  - Estimated coefficients
  - Facilitates forecasting

$$\hat{Y} = \hat{b} + \widehat{\boldsymbol{w}}\boldsymbol{X}$$

Training data incl. Y

i	Y	X_1	X_2	…	X_m
1	…	…	…	…	…
2	…	…	…	…	…
…	…	…	…	…	…
n	…	…	…	…	…

Learning Algorithm

Model $\widehat{\boldsymbol{w}}$

New data w/o Y

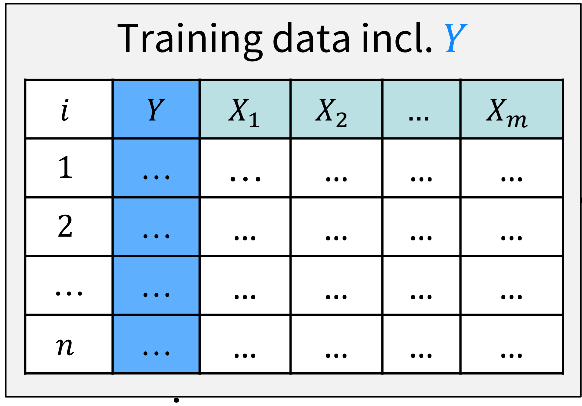
i	X_1	X_2	…	X_m
$n+1$	…	…	…	…
$n+2$	…	…	…	…

Forecasts of Y

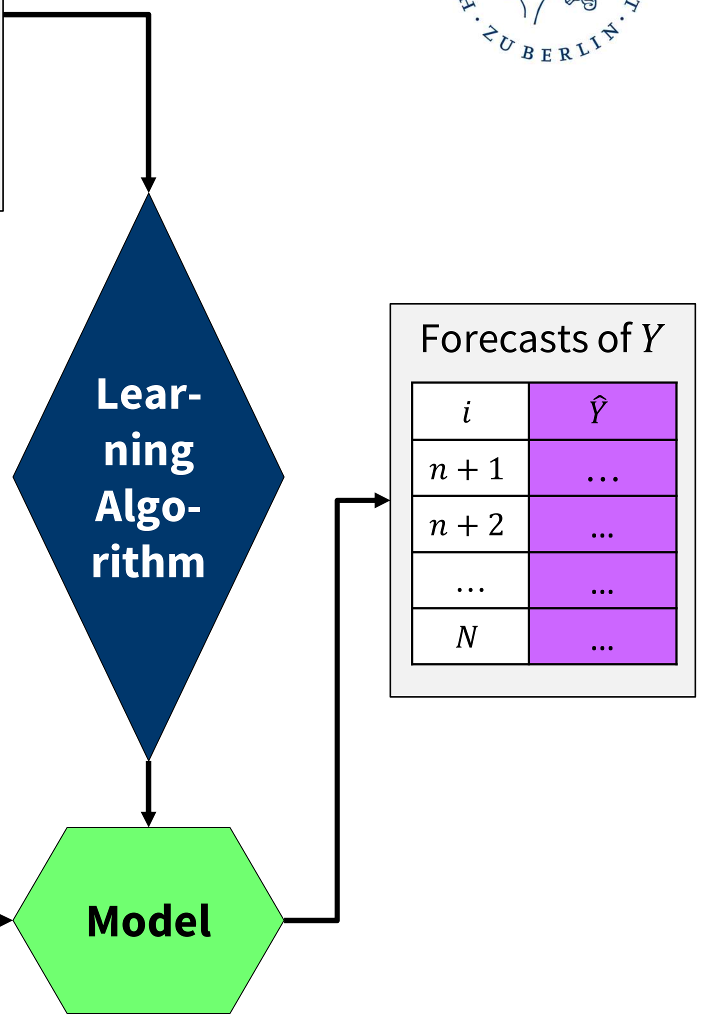
i	$\hat{Y}$
$n+1$	…
$n+2$	…
…	…
N	…

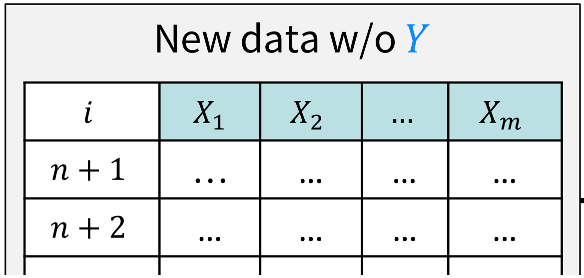
Two-Stage Paradigm

Supervised ML in general



- **Learning algorithm**
  - (Semi-)Parametric approaches mimic linear regression
  - Nonparametric approaches make no assumptions about DGP*
- **Model training**
  - Empirical risk minimization: maximize model fit on training data
  - Structural risk minimization: balance model fit vs. complexity
  - Minimize a loss function
- **Model**
  - Form varies across algorithms
  - Function with estimated parameters
  - Decision rules or tree-structure





*DGP: Data generating process

Algorithms for Supervised Learning (Selection)

A subjective view and a bit of guidance

Supervised learning algorithms

Tree- and prototype-based algorithms (non-parametric)

- CART
- CHAID
- C4.5

Individual trees

- Bagging / Random Forest
- Gradient Boosting / XGB

Ensembles (many trees)

- Nearest neighbors
- …

Regression-type algorithms (semi-/parametric)

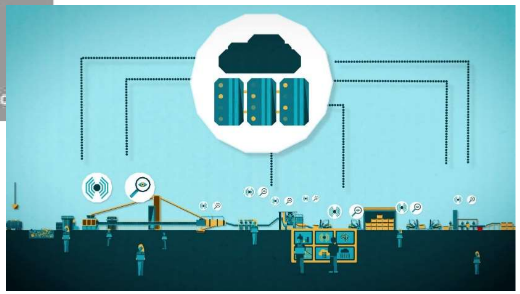
- Linear regression
- Generalized linear models (GLM)
- Generalized additive models (GAM)
- Artificial neural networks (ANN)
- Support vector machines (SVM
- …

Predictive Analytics Applications in Business

Use cases in business, return prediction case study, lessons learnt

Use Cases for Prediction Models in Business (Selection!)

Credit Scoring


Predictive Maintenance

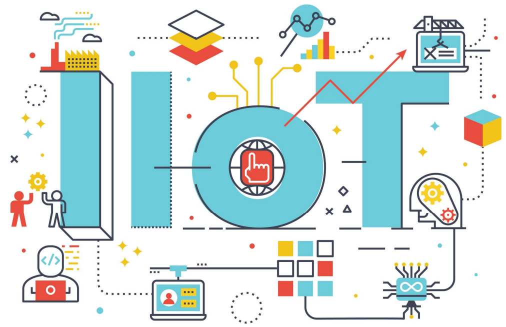

eCommerce Analytics

Anti Money Laundering


Social Media Analytics


Financial Forecasting

Fraud Detection


Internet of Things

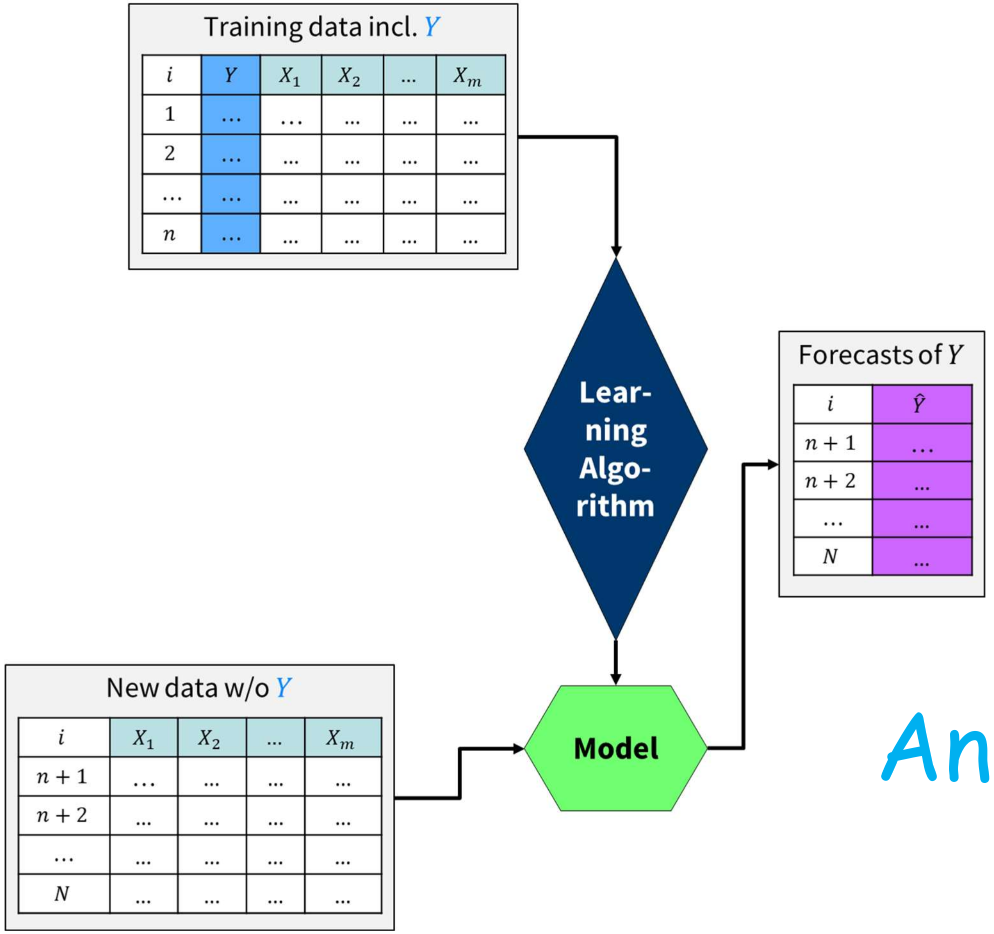
Case Study: Product Return Management in E-Commerce

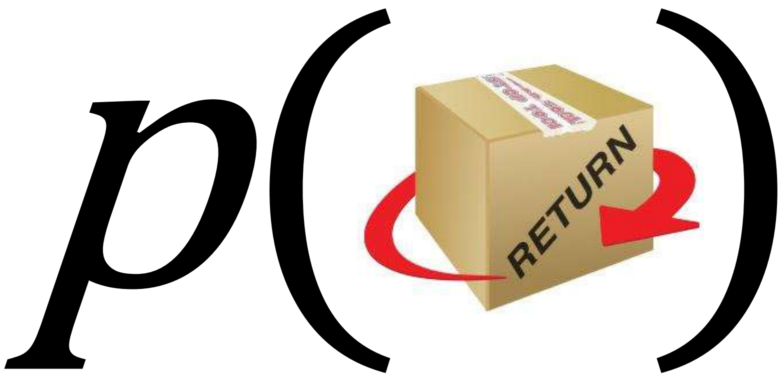
- Many online customers return items
- Costs to handle returns hurt e-tailors
- How can (predictive) analytics help?

https://www.invespcro.com/blog/ecommerce-product-return-rate-statistics/

Case Study: Product Return Management in E-Commerce

Analytical models can estimate a shopper's likelihood to return an item





And how is that useful ?

Case Study: Product Return Management in E-Commerce

Return predictions support decision-making

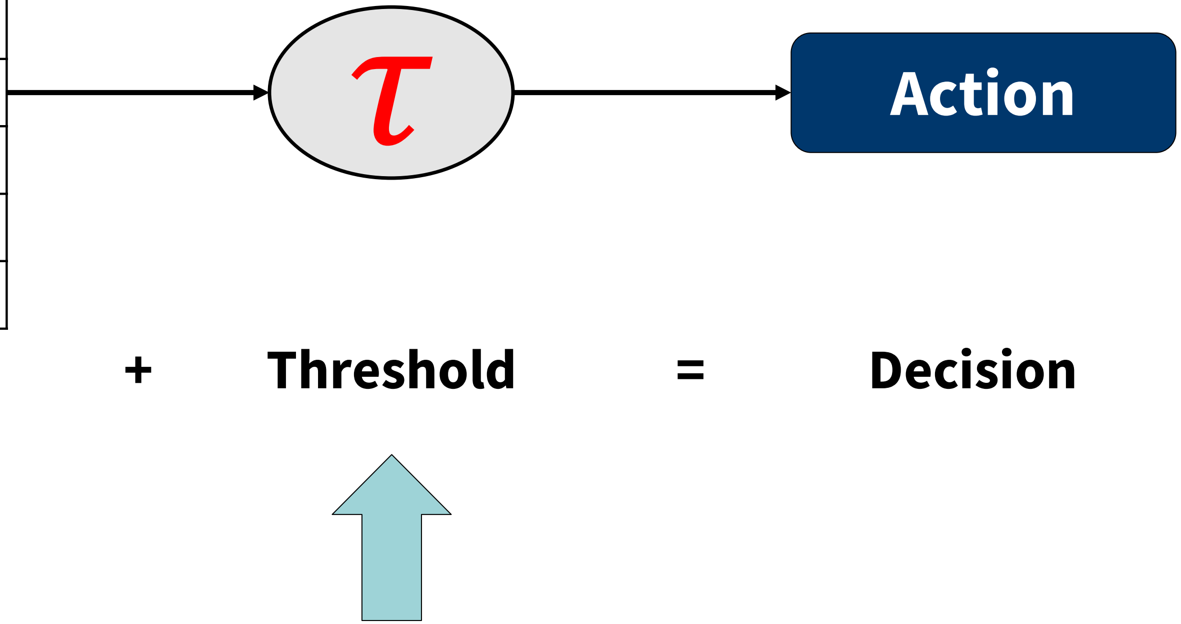
- **Use cases of model-based return probabilities**
  - Discourage buying items with high return probability
  - Recommend other items
  - Change the set of payment methods offered
  - Alter shipping costs

Case Study: Product Return Management in E-Commerce

Note that the prediction model does not make a decision

Product return management example revisited

Observation	*p*(return|features)
N+1	0.65
N+2	0.43
N+3	0.20
N+4	0.87
…	…
N+M	0.72



Prediction + **Threshold** = **Decision**

Domain knowledge, business rules, management strategy, cost-benefit considerations, …

Summary

Summary

Learning goals

- Data for supervised learning
- Regression versus classification
- Principle of predictive modeling (PM)

Findings

- PM requires past data with labels / target variable
- Regression involves predicting a numeric target
- Classification involves predicting a discrete target
- Two-step approach: model training and testing
- A prediction models maps from known feature values to unknown labels

What next

- How to prepare data for analysis
- Preprocessing pipeline & techniques

Thank you for your attention!

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel. +49.30.2093.5742
Fax. +49.30.2093.5741

stefan.lessmann@hu-berlin.de
http://bit.ly/hu-wi

www.hu-berlin.de