Team Wetter – Prediction Vorgehen 19.11.2019

Understanding 
Data 
Preparation 
Modeling 
•CRISP 
Business 
U n ding 
Deployment 
Evaluation 
Cross Industry Standard Process for Data Mining 

1. Business understanding
   * What are project objectives?; How do you define “success” and how can you measure it?
     1. We want to develop a **prediction mode**l that predicts bike rental demand as a function of suitable features available in or derived from "*the datasets" -****The datasets refers to Bike and weather.***
     2. Using a cross validation set, we check on the failrate? Wie viel weiche ich im Durchschnitt ab? Siehe [Anhang 1](onenote:#Prediction%20-%20Vorgehen&section-id={B2DF27FC-0767-4FD8-A638-4BF02FFBA788}&page-id={25CFD70C-6FE7-44AD-B623-E8B1FE288635}&object-id={47412909-4D9D-4FB7-8010-9B32EBA996E7}&71&base-path=https://d.docs.live.net/376b2f829bf70940/Dokumente/Analytics%20and%20Applications/Group%20work.one)
     3. How do we define success? **What does the team want from us? What ist bike rental demand?**
   * Do we fully understand the domain we are operating in?
     1. We are in the bike rental domain.
2. Data understanding
   * Which analytics approach should be employed (regression, classification, etc.)?
     1. Research! Aufgabe bis nächste Woche!
   * For this approach, what are data requirements and how can data collection be organized?
     1. Über Stunden einmal Daten visualisieren
     2. Fahrrad und Wetter Daten. Primär Wetterdaten
     3. Wir teilen die gebrainstormten Daten nach Donnerstag auf und visualisieren diese bis Dienstag
        1. Masking / einlesen der Daten in data frame -> pandas und dann visualisieren
   * Descriptive statistics and visualization combined with business understanding facilitate data understanding
     1. Das ist das was die Gruppe von Alex und Johanna macht

**Schnittstelle zu geclusterten Fahrraddaten zu unserem Modell.**

1. Data Preperation - **genaue und saubere Aufteilung nächste Woche Dienstag**

* Data cleaning and preparation routines include, e.g.
  + Missing or invalid values elimination or imputation
  + Eliminating duplicate rows
  + Aligning formatting
  + Combining multiple data sources
  + Transforming and normalizing data (e.g. categorical to encoded features)
  + Engineering new features (e.g. via NLP, etc. )
* //„Arguably the most time-consuming step of the entire DS process is data cleaning and preparation„
* //Accelerate data preparation by automating common steps

1. Modeling
   1. Developing predictive or descriptive models
   2. Modeling is often a highly iterative process in which different features and models are tried

-> anlegen, trainieren und hyperparameter festlegen

1. Evaluation
   1. Assess the model’s quality and it‘s performance in the real world – How reliable is it?
   2. Use statistical tests and common test metrics (R2,RMSE, etc.) to compare model performance
   3. Ensure that the model properly addresses the business problem
   4. Refine model as needed
2. [Deployment] -> Outlook and Analysis

Anhang 1

Training set 
(e.g. 50%) 
Validation set 
(e.g. 20%) 
All data 
Test set 
(e.g. 30%) 

1. Divide data into training

set, holdout set, and test

set

2. Train algorithm on training

set (i.e., to learn

parameters), use holdout

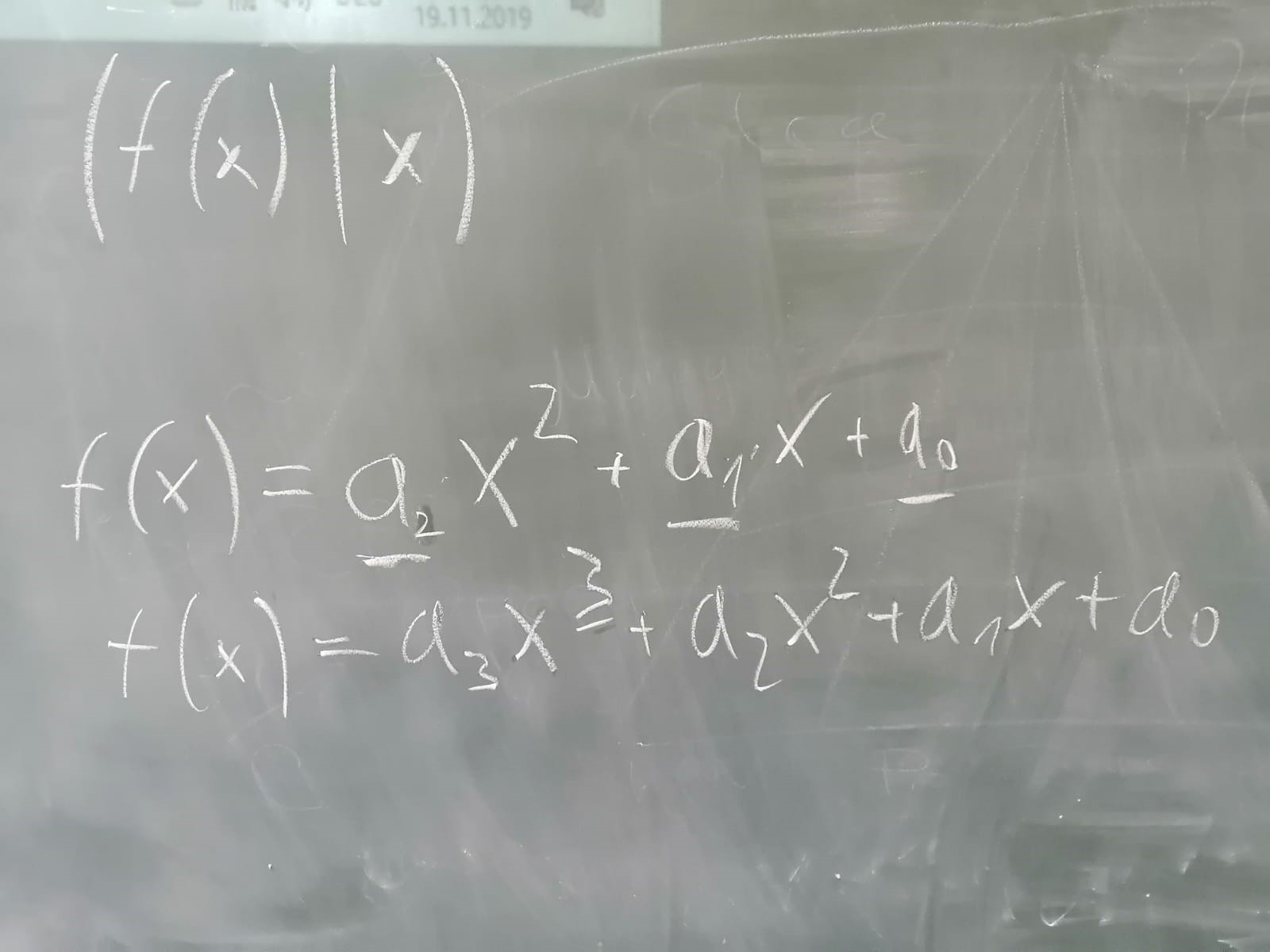
set to select

hyperparameters

3. (Optional) retrain system

on training + holdout

4. Evaluate performance



Diskussionspart: Der eigentliche Demand wird nicht gemessen. Es kann sein, dass zu einem Zeitpunkt keine Fahrräder mehr da sind. Dadurch steigen Nutzer auf andere Verkehrsmittel um. Es spiegelt also nicht den realen Bedarf wider.

Ein Indikator könnte jedoch sein, wenn alle Fahrräder an einem Standort weg sind.