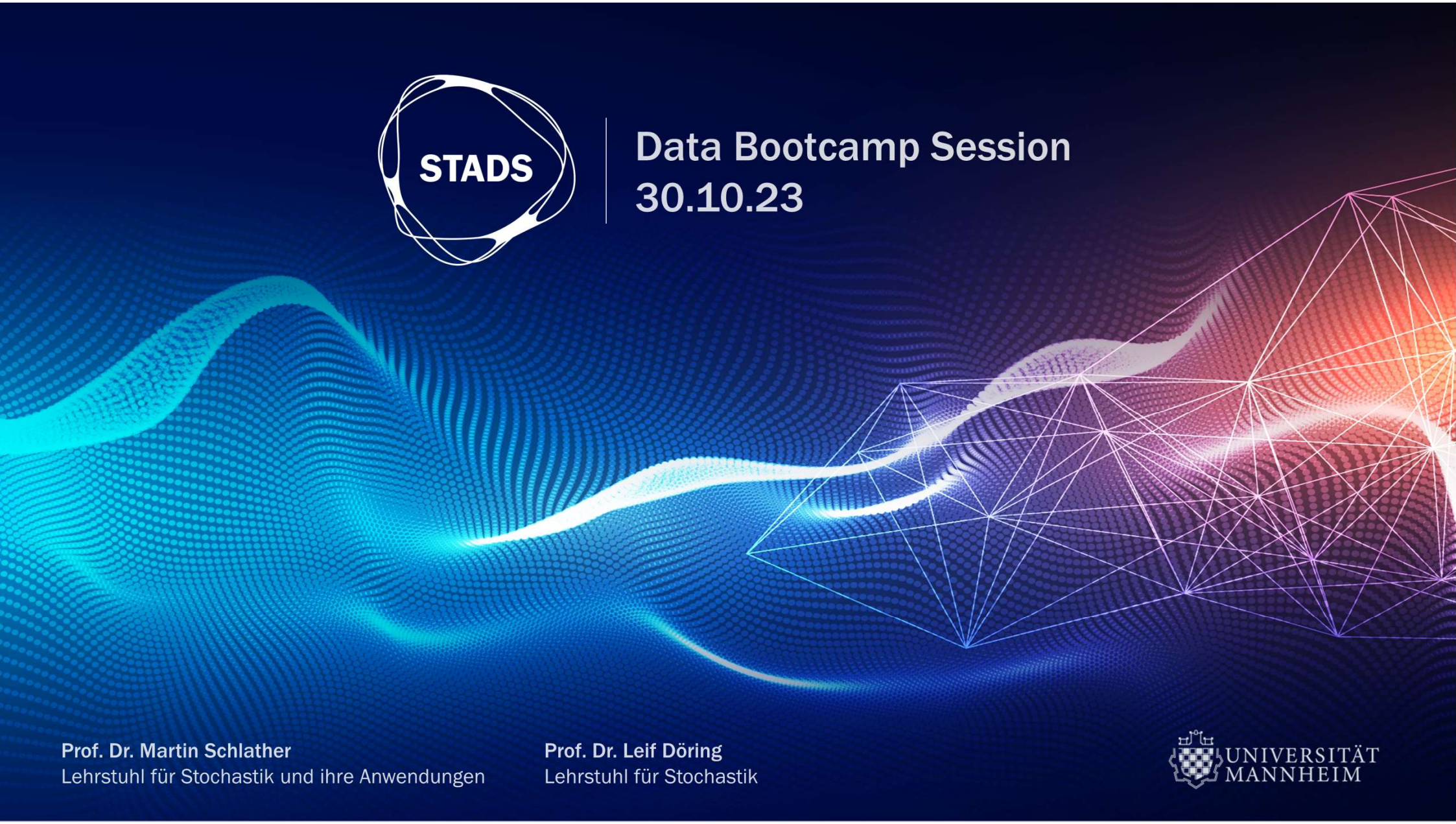

STADS
Data Bootcamp Session
30.10.23
Prof. Dr. Martin Schlather
Lehrstuhl für Stochastik und ihre Anwendungen
Prof. Dr. Leif Döring
Lehrstuhl für Stochastik
UNIVERSITÄT MANNHEIM


REGISTER NOW
STADS
Datathon
2023
02 & 03 NOV 2023
Solve a challenge together with your team and win great prizes! You have two days to do this, but much more awaits you.
STADS
Students' Association for Data Analytics & Statistics

RSM EBNER STOLZ

Challenge Beschreibung

Eine der Hauptaufgaben von Wirtschaftsprüfern besteht bspw. darin die Bewertung eines Bilanzpostens zum Geschäftsjahresende zu beurteilen. Die Höhe der zu bewertenden Bilanzpositionen hängt von den Geschäftsvorfällen auf den zugrundeliegenden Konten des Unternehmens ab. Sie werden in Form von Buchungen erfasst.

Über das Geschäftsjahr hinweg werden bspw. Vermögensgegenstände wie Forderungen auf einem Konto erfasst und bei Ausgleich wieder ausgebucht. In vielen Fällen gibt es ein so genanntes Nebenbuch, das über die Zusammensetzung eines Kontos Auskunft geben kann, doch dies trifft nicht auf alle Konten zu.

Mitten im größten Stress braucht der Prüfungsleiter diese Information für ein Konto, doch die Buchhaltung ist mit anderen Themen schon völlig überfordert. Könnt Ihr mit eurem analytischen Geschick und algorithmischem Denken dieses numerische Rätsel knacken?

.msg

Challenge Description

The Challenge is about predicting Price shocks. A price shock is a sudden surge in material prices, which can lead to rising production costs and even to an uneconomical production.

In order to be able to predict these shocks, a model was trained using key factors that have caused sudden increases in prices in the past. The model can predict future prices and identify possible shocks based on currently occurring key factors. To provide these current factors to the model as features, they must be extracted from current news articles.

This is where the challenge begins.

STADS

Challenge Description

McCar, a fictitious used car dealer from Mannheim, always offers his customers the best prices for their new dream car.

Due to constant growth in recent years, McCar unfortunately no longer keeps up with the pricing. Every day, the optimal sales price has to be found for dozens of vehicles.

McCar has a data set with the characteristics and sales prices of the last years at its disposal - can you support them and develop a model that predicts the optimal sales price for every vehicle in the future?

Agenda

Organizational information

- Data Bootcamp

STADS

Data Bootcamp

1. What to expect as challenges 
2. How to deal with the Problems 
3. Tips and Tricks regarding the Presentation 

Different perspectives to classify a solution as successful

Usually, the participants face some common problems, regardless of the specific challenge they have to solve. On the other hand, the companies have certain requirements that they would like to see met.

Time for Problem Processing: 15 h

What is taken into account for successful challenge completion?

- Completeness in the sense of the challenge
- Implementability
- Idea
- Creativity
- Overview of the code
- Overall impression

What are typical problems to face when solving the challenge?

- Understanding the Challenge and the Result that the Company wants
- Data Cleaning and Preprocessing
- Feature Engineering
- Model Optimization
- Explainability to the Jury

Data Bootcamp

1. What to expect as challenges 
2. How to deal with the Problems 
3. Tips and Tricks regarding the Presentation 

HOW TO DEAL WITH THE PROBLEMS

Rule No. 1: Company is KING

The jury of the first round consists of employees of the company who have a personal relationship with the challenge they bring. Therefore, always respect and value their comments and opinions. Read between the lines to get the best result for both sides.

Understand the Challenge!

- Task
- Tools
- Techniques
- Data
- Metrics

What does the Company want?

- Target user
- Transformation
- Scalability
- Adaptability
- Integration
- Cost-Efficiency
- Competitive Advantage

How are we going to explain the solution to the Jury?

- Benefits and Impact
- Addressing Concerns and Questions
- UX

HOW TO DEAL WITH THE PROBLEMS

Data Cleaning and Preprocessing - Strategies

Handling Missing Values

1. Imputation: Fill missing values with the mean, median, mode, or a custom value based on the context.
2. Deletion: Remove rows or columns with missing values if they are insignificant or cannot be imputed.
3. Advanced Imputation Techniques: Use more complex methods like K-nearest neighbors (KNN) imputation or regression imputation.

Outlier Detection and Treatment

1. Statistical Methods: Identify outliers using measures like Z-scores or IQR (interquartile range) and decide whether to remove or transform them.
2. Winsorizing: Cap extreme values by replacing them with the nearest non-outlying value.
3. Smoothing Techniques: Apply moving averages or other smoothing methods to reduce noise in time series or signal data.

Handling Duplicates

1. Identify and remove duplicate rows or records from the dataset.

   ! Often more complicated than first thought !

HOW TO DEAL WITH THE PROBLEMS

Data Cleaning and Preprocessing - Strategies

Handling Imbalanced Classes

1. Apply techniques like resampling (e.g., oversampling, undersampling) or using algorithms that handle imbalanced data.
2. Identify and remove highly correlated features to avoid redundancy in the dataset.

Standardizing Data

1. Scale numerical features to have zero mean and unit variance (z-score normalization) to ensure consistent units and magnitudes.
2. Apply transformations (e.g., logarithmic or power transformations) to reduce skewness in data distributions.
3. Standardizing Units: Convert different units to a common standard (e.g., converting miles to kilometers).
4. Correcting Typos and Misspellings: Use spell-checking tools or algorithms to identify and correct errors in text data.

Dealing with Categorical Variables

1. One-Hot Encoding: Convert categorical variables into binary (0/1) columns for each category.
2. Label Encoding: Assign unique numerical labels to each category.

HOW TO DEAL WITH THE PROBLEMS

Data Cleaning and Preprocessing - Strategies

Text Data Cleaning

1. Removing Punctuation and Special Characters: Strip non-alphanumeric characters from text.
2. Lowercasing: Convert all text to lowercase for consistent processing.
3. Stopword Removal: Eliminate common words (e.g., "the", "and") that may not carry significant meaning.

Data Validation and Constraints

1. Apply business rules or constraints to validate data entries (e.g., age should be a positive integer).

Visual Inspection and EDA

1. Use visualization techniques to identify anomalies, patterns, or inconsistencies in the data.

HOW TO DEAL WITH THE PROBLEMS

Feature Engineering in General

The real work and concept creating, that Feature Engineering covers, always follows after having a deep understanding of the data and the problem. Starting first with this point is mostly ineffective and costs a LOT of time.

... important for implementing Linear Regression, Decision Trees, Random Forests

Strategy for Creating Informative Features

- Extracting relevant attributes
- Performing aggregation for evaluating the importance of the features
- Generating new variables („age“*“income“, squaring variables)

ChatGPT
Python: skit-learn, tensorflow, numpy, pandas, matplotlib, ...

Dimensionality Reduction Techniques

Principal Component Analysis
How can we retain as much of the original variance?

Linear Discriminant Analysis
Which linear combination of features separates different classes best?

HOW TO DEAL WITH THE PROBLEMS

Feature Engineering Techniques

Recursive Feature Elimination (RFE): recursively removes less important features

Mutual Information: measures dependency between two variables, relevance with respect to target variable

SelectKBest, SelectPersentile: select top k features based on scoring function

LASSO (L1 Regularization): add penalty term based on loss and sum of coefficients

Tree-Based Methods (e.g., Random Forest Feature Importance): evaluating on how much feature contributes to reducing impurity

Sequential Feature Selection: iteratively adding or removing features based on their impact on model performance

Embedded Methods: Integrate feature selection directly into the model training process

PCA & LDA

HOW TO DEAL WITH THE PROBLEMS

Model Selection matching the Task

Classification
... for categorizing data into distinct classes
e.g., spam detection, image recognition
Algorithms: Logistic Regression, Decision Tree, Random Forest, SVM, NN

Regression
... when predicting continuous values
e.g., house prices, temperature
Algorithms: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, NN

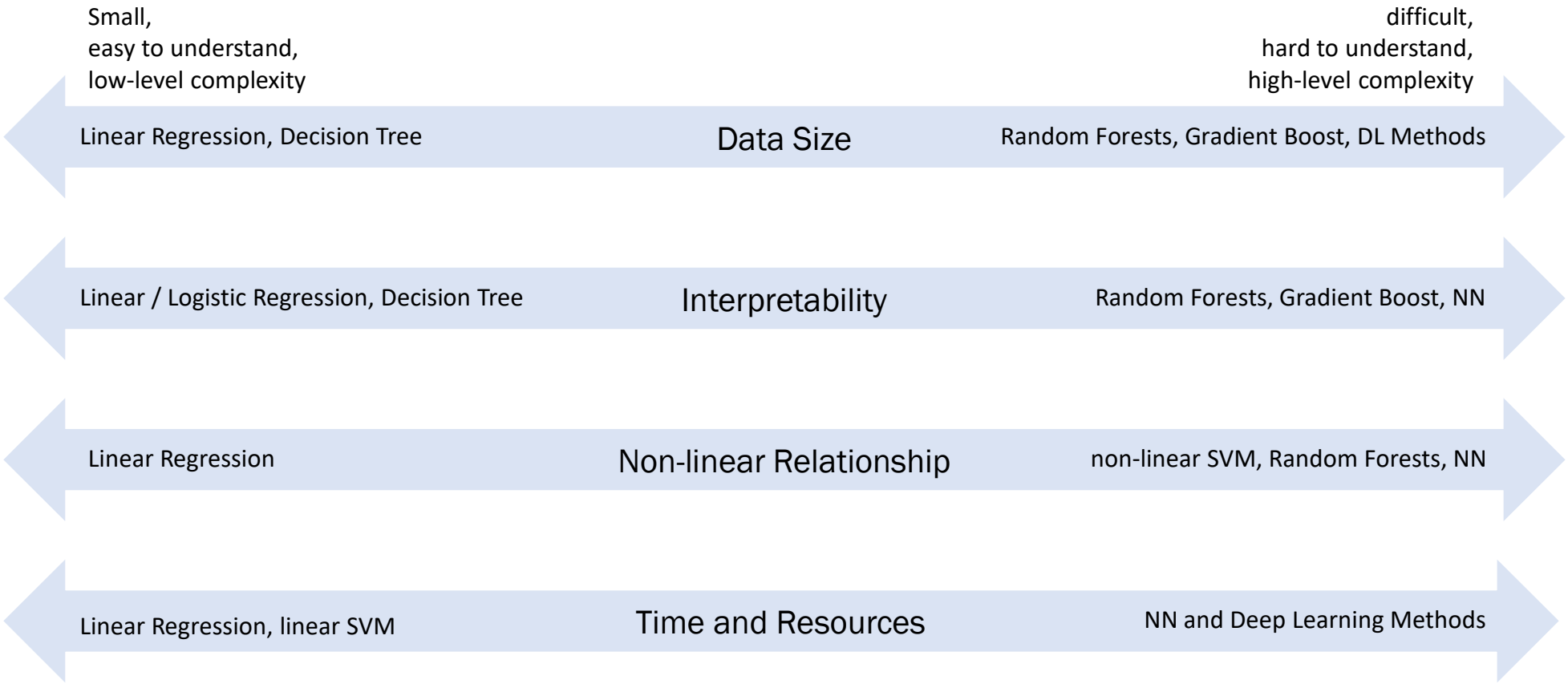
Clustering
... for grouping similar data points
e.g., customer segmentation
Algorithms: K-Means, DBSCAN, hierarchical clustering

Anomaly Detection
... to identify unusual patterns and outliers
e.g., fraud detection
Algorithms: Isolation Forest, One-Class SVM, Autoencoders

Recommender Systems
... to suggest items or content to users
e.g., movie recommendations
Algorithms: collaborative filtering, matrix factorization

HOW TO DEAL WITH THE PROBLEMS

Variables influencing Model Optimization



Data Bootcamp

1. What to expect as challenges 
2. How to deal with the Problems 
3. Tips and Tricks regarding the Presentation 

Timeline of the Evaluation of the Challenge Pitches

Try to present your solution in a way that makes the most of the short presentation time.

What to learn from others

It often happens that teams are stressed in the last minutes of the problem processing time. They forget that the presentation is the only time when the companies get to know their work and that during the processing time they did not have the opportunity to get the full insight into the development process of the team's solution.

Insiders of the last years presentations:

- Creative presentation of the problem (aka role play)
- User Experience (super fancy User Interface) vs. Quality of the code (life-run-through)

Plan enough time to prepare your presentation! Imagine that you want to "sell" your solution!

Companies want to see...

... a solution they can understand.

... scalability and future considerations / further development opportunities.
... uniqueness in the approach and / or the technical viewpoint from which you approached the problem.
... soft skills in answering questions regarding the team structure / task distribution / general presentation skills.

Thank you for your attention!

See you at the Datathon!

STADS