Technische Hochschule Augsburg

Forschungsinstitut für Informationsmanagement

Fraunhofer FIT

Data Analytics

Session 7: Introduction to Micro Projects

Prof. Dr. Wolfgang Kratsch

Augsburg Technical University of Applied Sciences,
Faculty of Computer Science,
Professor of Applied AI in Digital Value Creation

FIM Research Center for Information Management

Branch Business & Information Systems Engineering of the Fraunhofer FIT

www.fim-rc.de
www.wirtschaftsinformatik.fraunhofer.de

DIGITAL SCHMIEDE BAYERN

GESTALTE DIE VERWALTUNG VON MORGEN

Bewerbungsschluss: 06.06.2024
Programmzeitraum: 05.08.2024 – 29.10.2024
Standort: München

Was wir bieten...

DIGITALISIERUNG MIT IMPACT

Arbeite an echten Herausforderungen der bayerischen Verwaltung und schaffe nachhaltigen Mehrwert für die Bürgerinnen und Bürger.

FACHLICHE & METHODISCHE WEITERBILDUNG

Lerne agiles Arbeiten und neuste digitale Innovationsmethoden kennen und wende diese direkt im Projektkontext an.

NETZWERKAUFBAU

Vernetze Dich mit anderen Fellows und treffe spannende Experten und Mentoren aus der öffentlichen Verwaltung und darüber hinaus.

ZUSÄTZLICHE BENEFITS

Neben einem finanziell vergüteten Stipendium während des Programms bekommst Du ein Zeugnis, das Deine Leistungen und Fähigkeiten hervorhebt.

Wen wir suchen...

- ✓ **Junge Digitaltalente** ab dem 4. Semester mit **betriebswirtschaftlichen, gestalterischen,** oder **technischen** Fähigkeiten und Interesse
- ✓ Begeisterung und Affinität für die menschzentrierte Entwicklung digitaler Innovationen
- ✓ Leidenschaft und Motivation unsere öffentliche Verwaltung zu verbessern
- ✓ Deutschkenntnisse (min. B1) und Bereitschaft vor Ort zu arbeiten

www.digitalschmiede.bayern

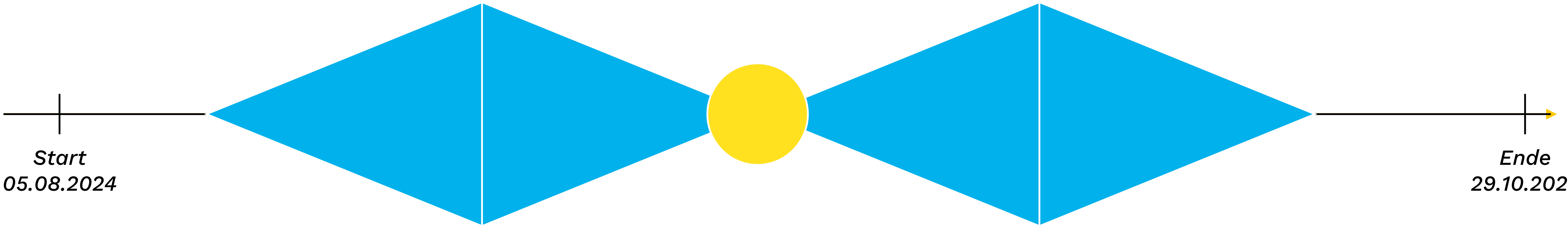
Ein Programm der: byte BAYERISCHE AGENTUR FÜR DIGITALES

in Kooperation mit: Bayerisches Staatsministerium für Digitales

organisiert von: mantro Fraunhofer FIT fim Forschungsinstitut für Informationsmanagement

DIGITAL SCHMIEDE BAYERN

Programmphasen im Überblick



Start
05.08.2024

Ende
29.10.2024

On-boarding	Verstehen	Verknüpfen	Gestalten	Umsetzen	Off-boarding
1 Woche	4 Wochen	2 Wochen	3 Wochen	3 Wochen	1 Woche
Teambuildung & Einführung	*Entwicklung eines Problemverständnis*	*Definition einer Problemstellung*	*Exploration von Lösungsräumen*	*Verproben eines Lösungsansatzes*	*Feedback & Übergabe*

www.digitalschmiede.bayern

DIGITAL SCHMIEDE BAYERN

Bewerbungsschluss: **06.06.2024**

Infoveranstaltungen:

- 14.05.2024, 18 Uhr (virtuell), Teilnahme über Zoom
- **22.05.2024, 18 Uhr (vor Ort), THA, Raum M 101**
- 27.05.2024, 18 Uhr (virtuell), Teilnahme über Zoom
- 03.06.2024, 18 Uhr (virtuell), Teilnahme über Zoom

Wöchentliche Drop-in-Termine für Deine Fragen:

- Jeden Dienstag vom 07.05. bis zum 04.06.2024 von 17:00 bis 17:30 Uhr (virtuell), Teilnahme über Zoom

JETZT BEWERBEN!

www.digitalschmiede.bayern

Ein Programm der: byte BAYERISCHE AGENTUR FÜR DIGITALES

in Kooperation mit: Bayerisches Staatsministerium für Digitales

organisiert von: mantro Fraunhofer FIT fim Forschungsinstitut für Informationsmanagement

Content of today’s lecture and tutorial

		02:00 p. m. - 03:30 p. m. (Room: W315)	03:40 p. m. - 05:10 p. m. (Room: W315)
1	27.03.2024	Introduction to Data Analytics	Python Setup & Getting Started
2	03.04.2024	Introduction to Python	Python Basics
3	10.04.2024	Data Engineering & Management	Data Engineering with Python
4	17.04.2024	Modeling I	Modeling with Python I
5	24.04.2024	Modeling II	Modeling with Python II
6	15.05.2024	Storytelling with Data (Guest Lecture)	Storytelling with Data
7	22.05.2024	**Introduction to Micro Projects** Evaluation	Evaluation
8	23.05.-11.06.2024	Micro Project Phase	
9	12.06.2024	Micro Project Presentations	
10	26.06.2024	Advanced Use Cases & Frontier Topics	Q&A Session
11	tba	Exam (60 min, written test, no accompanying materials permitted)	

Micro Projects

- Projects will be carried out in groups of 4 to 5 students per topic
- You can submit your preferences by ranking the topics on Moodle by tomorrow (May 23rd, 11:59 pm). We will optimize the overall matching of preferences and communicate the final assignment by Friday (May 24th).
- Presentation:
  - June 12th in our regular slot (2:00 pm – 5:10 pm, Room W315)
  - 20 minutes presentation, 10 minutes discussion
- Grading:
  - You can earn up to 15 bonus points on the exam.
  - Bonus points are only valid for the upcoming exam for summer term 2024.
  - Although you cannot get a better grade than 1.0 using bonus points, you do not need to pass the exam with >= 4.0 to benefit from bonus points.
  - Evaluation criteria (non-exclusive): project outcome (quality and creativity), learning path, individual contribution to analysis, data visualization, presentation (everybody must have an active part in the presentation), and discussion.
  - Participation in the micro projects is voluntarily. In case you do not submit your preferences by tomorrow, we assume that you decline to participate.

Topic 1: Predictive Process Monitoring

Description

Your team has been tasked with helping the Swedish car manufacturer Volvo improve their incident management process within their IT department. The goal is to predict the service level (first, second, or third) that will ultimately resolve an incident. This involves analyzing historical event data from the incident management system to develop a predictive model.

Tasks

- Describe and clean the dataset, including visualization of key features.
- Choose and train a classification model to predict the service level.
- Evaluate the model's performance on unseen data.
- Present the findings and steps taken to achieve the goals.

Dataset

The dataset consists of 6571 records detailing the incident management process. Each record includes event data generated by up to eleven process activities and the outcome of each process instance, specifically identifying which service level ultimately resolved the incident.

Challenges

- Data Preparation: Develop a suitable initial dataset from the raw event data.
- Feature Selection: Identify relevant features within the event data that are most predictive of the service level.

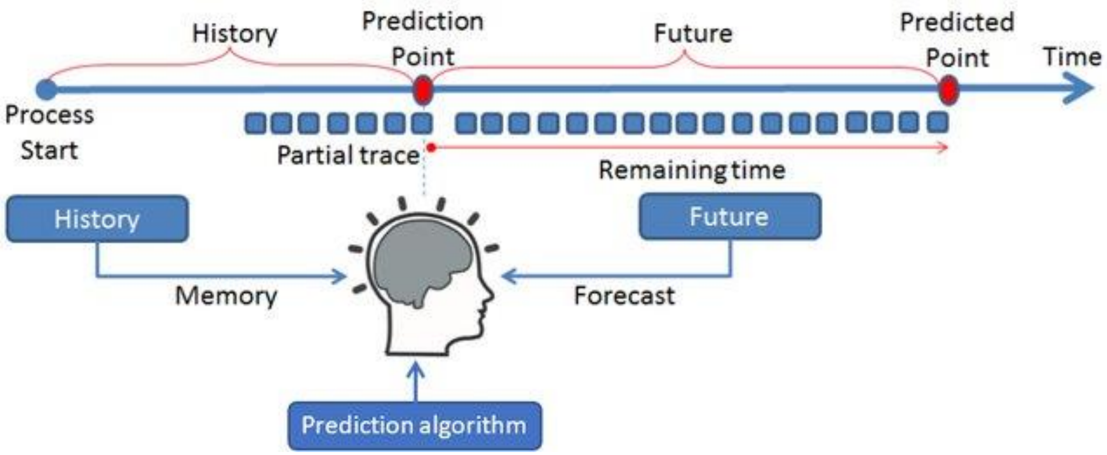


Image Source: Verenich, I., Dumas, M., Rosa, M. L., Maggi, F. M., & Teinemaa, I. (2019). Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. ACM Transactions on Intelligent Systems and Technology (TIST), 10(4), 1-34.

Topic 2: Building Analytics – Classifying Building Functions

Description

Your team has been tasked with predicting the primary functions of buildings using structured data provided by the city of Hamburg. The goal is to develop a classification model that accurately identifies building functions. This involves analyzing the provided data and overcoming challenges related to feature engineering and data quality.

Tasks

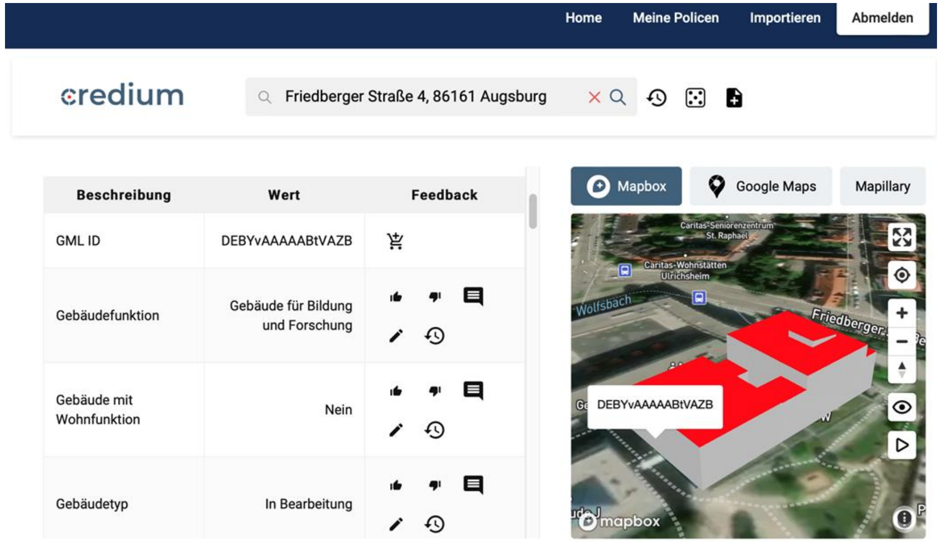
- Describe and clean the dataset, including visualization of key features.
- Choose and train a classification model to predict building functions.
- Evaluate the model's performance on the dataset.
- Present the findings and steps taken to achieve the goals.

Dataset

The dataset consists of 399,743 records and 71 attributes detailing various characteristics of buildings in Hamburg. The dataset includes building geometries and attributes such as building area, type, function, height, and other relevant features.

Challenges

- Data Quality: Managing inconsistent or missing data across different buildings.
- Feature Engineering: Identifying and transforming the most relevant features for accurate classification.



Beschreibung	Wert	Feedback
GML ID	DEBYvAAAAABtVAZB	
Gebäudefunktion	Gebäude für Bildung und Forschung	
Gebäude mit Wohnfunktion	Nein	
Gebäudetyp	In Bearbeitung	

Topic 3: Energy Analytics – FIM Smart Living Lab

Description

Your team has been tasked with forecasting energy production and usage in a smart building equipped with a rooftop PV system, electric vehicle charging infrastructure, and a battery storage system. The goal is to analyze time series data and compare different forecasting models to determine which data granularity (minute-based or quarter-hour-based) is better suited for forecasting energy metrics.

Dataset

- Time series with measurements taken every 60 seconds, containing 6,737,452 records from January 10, 2023, to April 1, 2023.
- Time series with measurements taken every 15 minutes, containing 531,547 records from January 1, 2023, to April 1, 2023.
- Description of the components in the energy management system, containing 77 records.

Challenges

- Data Preparation: Handling the large volume of time series data and dealing with missing values and irregularities.
- Identifying suitable methods for time series analysis.

Tasks

- Describe and clean the dataset.
- Choose and train a time series forecasting model.
- Compare model performance using minute-based and quarter-hour-based data.
- Present the findings and steps taken to achieve the goals.

Topic 4: EA Sports (FIFA)

Description

Your team has been tasked with validating whether FIFA player profiles can accurately predict real-life performance. The objective is to compare FIFA stats with actual performance metrics to determine if FIFA can serve as a digital twin for real-life football players. This involves analyzing the provided dataset and developing a predictive model.

Tasks

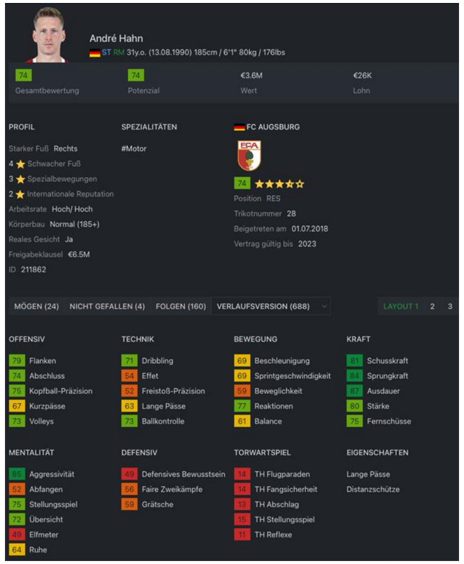
- Describe and clean the dataset, including visualization of key features.
- Train a model to predict real-life performance based on FIFA stats.
- Evaluate the model's performance and analyze which FIFA features are most predictive.
- Present the findings and steps taken to achieve the goals.

Dataset

The dataset consists of 2,137 records and 33 attributes detailing FIFA player statistics and corresponding real-life performance metrics. It includes sub-tables covering overall statistics and detailed stats for defenders, midfielders, and offensive players.

Challenges

- Data Cleaning: Addressing missing or inconsistent data in the FIFA and real-life performance metrics.
- Feature Selection: Identifying FIFA stats that strongly correlate with real-life performance.



André Hahn
ST RM 31y.o. (13.08.1990) 185cm / 6'1" 80kg / 176lbs

74 Gesamtbewertung
74 Potenzial
€3.6M Wert
€26K Lohn

PROFIL
Starker Fuß Rechts
4 ★ Schwacher Fuß
3 ★ Spezialbewegungen
2 ★ Internationale Reputation
Arbeitsrate Hoch/ Hoch
Körperbau Normal (185+)
Reales Gesicht Ja
Freigabeklausel €6.5M
ID 211862

SPEZIALITÄTEN
#Motor

FC AUGSBURG
74 ★★★★☆
Position RES
Trikotnummer 28
Beigetreten am 01.07.2018
Vertrag gültig bis 2023

OFFENSIV
79 Flanken
74 Abschluss
75 Kopfball-Präzision
67 Kurzpässe
73 Volleys

TECHNIK
71 Dribbling
54 Effet
52 Freistoß-Präzision
63 Lange Pässe
73 Ballkontrolle

BEWEGUNG
69 Beschleunigung
69 Sprintgeschwindigkeit
59 Beweglichkeit
77 Reaktionen
61 Balance

KRAFT
81 Schusskraft
84 Sprungkraft
87 Ausdauer
80 Stärke
75 Fernschüsse

MENTALITÄT
85 Aggressivität
52 Abfangen
75 Stellungsspiel
72 Übersicht
49 Elfmeter
64 Ruhe

DEFENSIV
49 Defensives Bewusstsein
56 Faire Zweikämpfe
59 Grätsche

TORWARTSPIEL
14 TH Flugparaden
14 TH Fangsicherheit
13 TH Abschlag
15 TH Stellungsspiel
11 TH Reflexe

EIGENSCHAFTEN
Lange Pässe
Distanzschütze

EA SPORTS FIFA

Topic 5: Overtourism

Description

Your team has been tasked with forecasting visitor numbers at Scharbeutz beach to manage overcrowding effectively. The objective is to predict the days and hours with high visitor loads. This involves identifying and integrating relevant external data sources, such as weather data and local events, to create a robust predictive model.

Tasks

- Identify and additional data sources relevant to visitor load.
- Match these data sources with the provided measurements.
- Choose and train a forecasting model to predict visitor numbers.
- Present the findings and steps taken to achieve the goals.

Dataset

The dataset consists of 3,240 records with two columns: datetime and occupancy. It contains visitor load data recorded every four hours at Scharbeutz beach, capturing the number of visitors at different times of the day over various dates.

Challenges

- Data Integration: Identifying and integrating relevant external data sources, such as weather conditions and local events.
- Creating and transforming features to improve model accuracy.
- Identifying suitable methods for time series analysis.

Topic 6: Public Transport

Description

Your team has been tasked with forecasting arrival and departure delays across the public transportation network for a municipal transportation authority. The goal is to develop a predictive model that can accurately forecast the delay duration in minutes for both arrivals and departures at each station.

Dataset

The dataset consists of 156,696 records across 13 columns. It includes timestamps for planned and actual arrivals and departures, station identifiers, and recorded delay times at various stops within the network. Additionally, it records the arrival and departure delays in minutes.

Challenges

- Feature Engineering: Identifying influential features that affect delay times and creating new features to improve model predictions.
- Model Selection: Choosing appropriate models for time series forecasting.

Tasks

- Describe and clean the dataset, including visualization of key features.
- Choose and train a classification model to predict the duration of delays for both arrivals and departures.
- Evaluate the model's performance on unseen data.
- Present the findings and steps taken to achieve the goals.

VRR