OTH AMBERG-WEIDEN
Fakultät Elektrotechnik, Medien und Informatik
Prof. Dr. Christian Bergler, Prof. Dr. Fabian Brunner
Wintersemester 2025/2026

Projektarbeit Data Analytics

Ausgabe: 15.12.2025 – 18:00:00 Deadline: 18.01.2026 – 23:59:59

Ziel der Projektarbeit

In dieser Projektarbeit sollen historische bis tagesaktuelle Finanzmarktdaten aus zwei umfangreichen Datenquellen – den NASDAQ Daily Stock Prices (1962–2025) sowie den Crypto Currencies Daily Prices (2010–2025) – systematisch analysiert werden. Ziel ist es insbesondere, die Kursentwicklung traditioneller Aktienmärkte und moderner Kryptowährungen vergleichend zu untersuchen, um ein fundiertes und umfassendes Verständnis der jeweiligen Marktmechanismen, Dynamiken, Einflussfaktoren und Zusammenhänge über mehrere Jahre hinweg zu gewinnen. Der Schwerpunkt der Untersuchung liegt dabei auf dem NASDAQ-Aktienmarkt, dessen Kursentwicklungen, strukturelle Merkmale und langfristige Dynamiken detailliert untersucht werden. Der Crypto Currencies Datensatz wird ergänzend herangezogen, um die Kurs-, Volatilitäts- und Renditestrukturen des NASDAQ in Relation zu digitalen Vermögenswerten zu analysieren. Dadurch soll untersucht werden, inwieweit sich Entwicklungen im NASDAQ-Markt von denen des Kryptomarktes unterscheiden, ähneln oder möglicherweise gegenseitig beeinflussen. Ziel der Projektarbeit ist es insbesondere, ein tiefgehendes Verständnis der kursbestimmenden Faktoren innerhalb des NASDAQ-Marktes zu gewinnen. Dazu gehört die Identifikation charakteristischer Muster, langfristiger Trends sowie marktphasenabhängiger Volatilitätsstrukturen, wie sie für technologieorientierte und wachstumsstarke Unternehmen im NASDAQ typisch sind. Für die Analyse wird ein einheitlicher Betrachtungszeitraum von 2000 bis 2025 herangezogen, wodurch sowohl die langfristigen Entwicklungen im NASDAQ-Markt abgedeckt als auch die ab 2010 einsetzenden Kursbewegungen der Kryptowährungen konsistent in die Untersuchung integriert werden können.

Modalitäten

- Die Bearbeitung der Projektarbeit erfolgt in der Programmiersprache Python.
- Als Ergebnis ist ein Jupyter-Notebook namens `nachname_vorname.ipynb` zu erstellen und elektronisch via Moodle abzugeben, das den vollständigen Programmcode und alle Analyseergebnisse (eingebettete Grafiken, Outputs der Zellen, erläuternder Text) enthält. Eine separate schriftliche Ausarbeitung ist nicht erforderlich. Die Ergebnisse zu den einzelnen Aufgaben und die abgeleiteten Erkenntnisse sollen jedoch ebenso wie das methodische Vorgehen in Form von Fließtext innerhalb des Notebooks ausführlich textuell dokumentiert werden. Darüber hinaus ist für den Vortrag ein Foliensatz zu erstellen, der als Teil der Dokumentation mit abzugeben ist. Datensätze, die als Ergebnisse einzelner Aufgaben resultieren, sind, sofern explizit gefordert, ebenfalls mit abzugeben.
- Verwenden Sie für die Bearbeitung der Projektarbeit vorzugsweise die **Python-Version 3.13.5** (oder neuer). Um die verwendete Umgebung eindeutig zu charakterisieren und die

Ergebnisse reproduzieren zu können, reichen Sie bitte zusätzlich zum Jupyter Notebook ist eine vollständige `requirements.txt`-Datei mit ein.

- Die **Bearbeitung der Projektarbeit** ist in **Gruppen von maximal zwei Personen** zulässig. Im Fall einer Zweierabgabe genügt es, wenn ein Gruppenmitglied die Arbeit elektronisch einreicht. Im Kopf des Dokuments sind alle Gruppenmitglieder zu benennen.
- Die Abgabe der Dokumente hat **bis spätestens 18.01.2026 um 23:59:59 Uhr** über Moodle zu erfolgen.
- Eine **Präsentation** der Ergebnisse in Form eines ca. **20-minütigen Kurzvortrags pro Projektgruppe erfolgt am 03. Februar 2026 (DC 1.07, ab 08:30 bis ca. 15:00 Uhr)**. Ein zeitlicher Ablaufplan wird auf der Basis der gebildeten Einzel-/Zweiergruppen erstellt und rechtzeitig vorher bekannt gegeben.
- Die unten **beigefügte schriftliche Erklärung** (s. Anhang) ist von allen Gruppenmitgliedern **auszufüllen, zu scannen und mit den eingereichten Dokumenten hochzuladen**.

Anforderungen und Bewertungsgrundlagen

- Der Code ist lauffähig und erfüllt die in den Aufgaben gestellten Anforderungen.
- Der Code ist klar strukturiert, gut lesbar, nachvollziehbar und ausreichend kommentiert.
- Der Code ist elegant, effizient und verwendet, sofern verfügbar, bereits vorhandene Python-Funktionen zur Bearbeitung der gestellten Analyseaufgaben.
- Das eingereichte Jupyter-Notebook ist ansprechend und übersichtlich gestaltet. Verwenden Sie dazu die Strukturierungsmöglichkeiten, die die Markdown-Sprache bietet. Das Dokument soll mit einer Gliederung mit Verlinkung zu den Lösungen der einzelnen Aufgaben versehen werden und eine abschließende Zusammenfassung der Analyseergebnisse und der gewonnenen Erkenntnisse enthalten. Sofern Sie externe Quellen verwenden, geben Sie diese in einem Quellenverzeichnis an. Alle verwendeten Hilfsmittel und Quellen sind anzugeben.
- Die editorielle Qualität des Dokuments fließt in die Bewertung ein.
- Die in den einzelnen Teilaufgaben gewonnenen Erkenntnisse sind ausführlich visuell und textuell dokumentiert und in den Anwendungskontext eingeordnet. Beschreiben Sie nicht nur, **was** Sie beobachten, sondern gehen Sie auch auf mögliche Ursachen und weiterführende Zusammenhänge ein, ggf. unter Einbeziehung externer Quellen.
- Alle Ausführungen sind klar formuliert, nachvollziehbar und durch die Daten belegbar.
- Die erstellten Diagramme sind ansprechend und übersichtlich gestaltet und transportieren eine klare Botschaft. Sie sind insbesondere ausreichend beschriftet (z.B. Titel, Achsenbeschriftungen, Einheiten etc.).
- Die bei der Datenvorbereitung und -analyse durchgeführten Schritte sind fachlich und methodisch korrekt ausgeführt und hinreichend motiviert und dokumentiert worden. Beschreiben Sie nicht nur, **wie** Sie vorgehen, sondern auch **warum**.
- Die Ergebnisse werden im Rahmen eines Vortrags ansprechend und überzeugend präsentiert. Die Qualität der Folien, die zur Dokumentation gehören, fließt in die Bewertung ein.

Gegebene Daten

Die nachfolgend bereitgestellten Datensätze bilden die Grundlage für sämtliche Analyseaufgaben im Rahmen dieser Studienarbeit. Sie ermöglichen umfassende explorative Analysen, statistische Untersuchungen, sowie den Aufbau von Vorhersagemodellen. Ergänzende interne Variablen oder aggregierte Kennzahlen sind im Rahmen der Datenverarbeitung zu berechnen. Alle für die Durchführung der Projektarbeit relevanten Datensätze sind über den folgenden `myFiles`-Link abrufbar und stehen dort gesammelt zum Download zur Verfügung:

Datenkorpora Projektarbeit Data Analytics & Engineering (WS 25/26)

Die bereitgestellten **Datensätze** (NASDAQ, Crypto Currencies) **umfassen** alle verfügbaren **Kursinformationen** und liegen **bis einschließlich Freitag, den 12.12.2025**, in aktueller Form vor. Der insgesamte **Betrachtungszeitraum** soll die Jahre **2000 bis 2025** umfassen. Sofern im Rahmen der Analysen zusätzliche Daten oder weitere Informationsquellen erforderlich sind, wird deren Verwendung jeweils ausdrücklich und nachvollziehbar im Verlauf der Arbeit und Aufgabenbeschreibung spezifiziert. Die gegebenen Daten brauchen nicht wieder mit abgegeben werden. Bei der Korrektur wird davon ausgegangen, dass Sie sie in einem Unterverzeichnis namens `data` abgelegt haben, von wo sie eingelesen werden. Der Pfad zu diesem Verzeichnis soll als Variable im Kopf des Jupyter-Notebooks gesetzt und anschließend beim Einlesen verwendet werden.

NASDAQ-Aktienkurse
Ausgangspunkt für die Analysen sind historische bis tagesaktuelle Kursdaten von Unternehmen, die am amerikanischen NASDAQ Stock Market gelistet sind, einer der größten und liquidesten Börsen weltweit, mit Sitz in New York City und dem höchsten Handelsvolumen in den USA. Für die Studienarbeit wird der Datensatz NASDAQ Daily Stock Prices verwendet. Die Daten werden öffentlich über die Plattform Kaggle gepflegt und bereitgestellt. Der Datensatz umfasst tägliche Handelsinformationen (Open, High, Low, Close) im Zeitraum von 1962 bis 2025 (tagesaktuell). Die Daten liegen in Form einzelner CSV-Dateien pro Unternehmen vor und ermöglichen sowohl kurz-, mittel- und langfristige Marktanalysen als auch detaillierte Betrachtungen einzelner Titel.

Meta-Information zu NASDAQ-Unternehmen
Ergänzend dazu steht ein Meta-Datensatz zur Verfügung, der Informationen zu einem Großteil der am NASDAQ gelisteten Aktien enthält. Dieser Datensatz wurde über das offizielle NASDAQ-Screener-Tool bereitgestellt und enthält unter anderem folgende Angaben: Symbol, Unternehmensname, Land, durchschnittliches Handelsvolumen, Sektor und Branche. Die Daten liegen in der bereits zur Verfügung gestellten Datei `nasdaq_screener.csv` vor und dienen als Grundlage für die Zuordnung einzelner Wertpapiere zu Branchen, Regionen oder Unternehmensstrukturen. Weiterhin wird eine Datei namens `nasdaq_company_addresses.csv` zur Verfügung gestellt, die als Ausgangspunkt für geobezogene Auswertungen Angaben zu den Adressen der Unternehmen enthält.

Kryptowährungen
Zusätzlich werden historische bis tagesaktuelle Kursdaten von Kryptowährungen bereitgestellt, die ebenfalls über Kaggle bezogen werden. Der Crypto Currencies Daily Prices (2010–2025) Datenkorpus enthält tägliche Handelsinformationen (Open, High, Low, Close) zahlreicher Kryptowährungen ab dem Jahr 2010 bis 2025 (tagesaktuell). Diese Daten dienen in der Studienarbeit primär dazu, die Kurs- und Volatilitätsentwicklungen des NASDAQ in Bezug zu digitalen Vermögenswerten zu setzen und Unterschiede oder Zusammenhänge zwischen beiden Märkten herauszuarbeiten.

Aufgaben

Aufgabe 1 (Datenvorbereitung)

Im Rahmen der Studienarbeit sollen die bereitgestellten Finanzmarktdaten – bestehend aus NASDAQ-Aktienkursen, ergänzenden Meta-Informationen sowie Kryptowährungsdaten – in einem ersten Schritt systematisch zusammengeführt, analysiert, bereinigt, aufbereitet und transformiert werden. Ziel es ist, vollständig bereinigte, homogene und reproduzierbare Datensätze für den Betrachtungszeitraum 2000-2025 zu erzeugen, welche als Grundlage für alle nachfolgenden Analysen dienen.

Die finalen Datengrundlagen sollen in den nachfolgenden `DataFrames` **abgespeichert** werden:

- `df_nasdaq_daily`, `df_nasdaq_weekly`
- `df_crypto_daily`, `df_crypto_weekly`
- `df_nasdaq_meta`

Alle final aufbereiteten Datengrundlagen (`DataFrames`) sind zusätzlich zur Notebook-Abgabe als `*.csv` Dateien mit **identischer Namensbezeichnung in Moodle hochzuladen**. Die Datensätze, die kursbezogene Informationen enthalten, sollen der Spalten-Struktur

`ticker, date, open, close`

folgen, wobei sich `open` bzw. `close` jeweils auf den Beginn und das Ende des Referenzzeitraums (Tag bzw. Woche) beziehen. Weiterhin soll pro Datensatz eine detaillierte Einschätzung über die insgesamte Datenqualität getroffen werden.

Die methodische Herangehensweise ist grundsätzlich frei, jedoch sind bestimmte zentrale Aspekte der Datengrundlage sowie der damit verbundenen Schritte der Datenvorverarbeitung zwingend zu berücksichtigen, sinnvoll und angemessen zu adressieren. Die folgenden Hinweise dienen als Orientierungsrahmen und sollen im Kontext der eigenen datensatzspezifischen Analysen sorgfältig reflektiert, dokumentiert und fachgerecht umgesetzt werden.

Hinweise:

- Ordnungsgemäßes **Einlesen** der jeweiligen CSV-Dateien **inklusive einer fachgerechten Zusammenführung**, sofern dies analytisch erforderlich ist.
- Bereinigung des gesamten Datensatzes durch Entfernen aller Werte außerhalb des festgelegten **Untersuchungszeitraums 2000 bis 2025**.
- Sicherstellung einer konsistenten und **sachgemäßen Typisierung** aller Datenattribute.
- Durchführung weiterer geeigneter **Datenbereinigungsschritte** um strukturelle Anomalien (Sonderfälle) in den Daten zu erkennen und zu bereinigen. Dazu gehört u.a. die geeignete Berücksichtigung von **Aktiensplits**:
    a. **Forward-Split**: Der Aktienkurs wird proportional geteilt, sodass Anleger mehr Aktien zu einem niedrigeren Stückpreis erhalten. Der Gesamtwert der Position bleibt unverändert. Warum? Damit die Aktie günstiger und für mehr Anleger zugänglich wird. Beispiel: Aus 1 Aktie zu 400 € werden 4 Aktien zu je 100 €, der Gesamtwert bleibt 400 € (siehe NVIDIA – NVDA). Der Split-Faktor ist größer als 1.
    b. **Reverse-Split**: Mehrere bestehende Aktien werden zu einer Aktie zusammengelegt, wodurch der Stückpreis steigt. Auch hier bleibt der Gesamtwert für Anleger unverändert. Warum? Um einen zu niedrigen Kurs anzuheben und Börsenanforderungen zu erfüllen.

Beispiel: Aus 10 Aktien zu je 0,50 € wird 1 Aktie zu 5 €, der Gesamtwert bleibt 5 € (siehe E-Home Household Service Holdings Limited – EJH). Der Split-Faktor ist kleiner als 1.

Vergleichen Sie die **bereitgestellte NASDAQ-Split-Datei** `splits_2000_2025.csv` mit ihren täglichen Kursdaten, indem Sie für jede Aktie und jedes dort angegebene Split-Datum den zugehörigen Split-Faktor auslesen.

Prüfen Sie anschließend im NASDAQ-Datensatz, inwiefern zwischen dem letzten Handelstag vor dem Split und dem Tag des Splits (bzw. in diesem Zeitfenster) ein Kursverhältnis vorliegt, das dem angegebenen **Split-Faktor** entspricht ($\texttt{split-factor} = \frac{\texttt{close}_{t-1}}{\texttt{close}_t}$).

Ist dieses Verhältnis vorhanden, wurde der Split im Kursdatensatz noch nicht bereinigt. Ist kein entsprechender Sprung sichtbar, gilt der Split als bereits eingepreist und es ist keine weitere Anpassung notwendig. Führen Sie diesen Prozess automatisiert für alle im `splits_2000_2025.csv` enthaltenen Aktien und Split-Ereignisse durch und dokumentieren Sie kurz, wie viele Splits bereinigt werden mussten.

- Identifikation und sachgerechte **Behandlung potenziell fehlender Werte**, einschließlich ergänzender Datenbeschaffung **über externe Quellen** (Python-Modul `yfinance`). Ziel ist, den Anteil fehlender Kursdaten im Verhältnis zur Gesamtstichprobe zu minimieren.
- **Plausibilisierung der Datenkorrekturen** anhand visueller und statistischer Verfahren.
- **Transformation der Daten mittels Aggregationen** über die Zeit (`daily` vs. `weekly`).
- Konsolidierung und Aufbereitung aller erzeugten Datensätze zu einer einheitlichen, reproduzierbaren Form, einschließlich einer **detaillierten Bewertung der Datenqualität jedes einzelnen Datensatzes**.
- **Dokumentation** der getroffenen Entscheidungen sowie Begründung **des Vorgehens**.

Aufgabe 2 (Datenanalyse)

Als **Datenausgangslage** sind die **zuvor abgespeicherten** und aufbereiteten `*.csv` **neu einzulesen**. Die daraus erzeugten `DataFrames` müssen in den folgenden Variablen abgelegt werden:

- `df_nasdaq_daily_pp`, `df_nasdaq_weekly_pp`
- `df_crypto_daily_pp`, `df_crypto_weekly_pp`
- `df_nasdaq_meta_pp`

Diese bilden die Basis für alle weiteren Verarbeitungsschritte.

Im Rahmen der anschließenden **explorativen und weiterführenden Datenanalyse** sollen grundlegende statistische Eigenschaften, marktphasenabhängige Dynamiken, Trends und strukturelle Muster des NASDAQ/Krypto-Markts untersucht werden.

Für die Untersuchung und Beantwortung der nachfolgend genannten Fragestellungen ist die konkrete Ausgestaltung der Analysen weitgehend frei und unterliegt keinen detaillierten methodischen Vorgaben. Die erzielten Ergebnisse sind jedoch mittels geeigneter Visualisierungen nachvollziehbar aufzubereiten, fachlich fundiert zu interpretieren und im Hinblick auf relevante kontextuelle Zusammenhänge zu analysieren.

Analytische Fragestellungen:

- Wie hat sich das gesamte **NASDAQ-Aktienportfolio historisch** über den Gesamtzeitraum (2000-2025) **verändert** (Anzahl gelisteter Aktien, Dynamik der Branchenstruktur und mittlere "Lebensdauer" von Aktien im NASDAQ)?

- Was waren bis heute die **"goldenen Jahre" des NASDAQ und des Kryptomarktes**? Untersuchen sie hierfür folgende Teilaspekte:

  a. **Analyse des jährlichen (prozentualen) Anteils an Aktien**, im NASDAQ (2000-2025) und Kryptomarkt (2010-2025), die Kursgewinne bzw. Kursverluste zu verzeichnen haben (z.B. 500 von insgesamt 1.000 Aktien sind im Jahr 2020 gefallen).
  b. **Analyse der tatsächlichen jährlichen Wertveränderungen** (aggregierte Gewinne bzw. Verluste) im NASDAQ und Kryptomarkt, um zu untersuchen, inwieweit der prozentuale Anteil fallender oder steigender Aktien mit der tatsächlichen Marktintensität korrespondiert. Ein hoher Anteil an Verlustwerten muss nicht zwangsläufig mit den größten absoluten oder relativen Gesamteinbußen einhergehen.
  c. Lassen sich darüber hinaus **erste Parallelen oder strukturelle Ähnlichkeiten** zum Gewinn- und Verlustverhalten zwischen NASDAQ vs. Krypto identifizieren?

- Wie gestaltet sich die **statistische Verteilung der Schlusskurse** (`close`) über alle NASDAQ-Unternehmen im gesamten Betrachtungszeitraum 2000-2025? **Visualisieren** Sie zunächst die Verteilung der Daten in geeigneter Form und **prüfen** Sie im Anschluss, inwieweit die Annahme **einer Normalverteilung für logarithmierte Schlusskurse** erfüllt ist, indem Sie den `Shapiro-Wilk` Hypothesentest durchführen.
- Führen Sie eine ausführliche **Korrelationsanalyse zwischen** den beiden berühmtesten Kryptowährungen – **Bitcoin (BTC) und Ethereum (ETH)** – zusammen mit Aktien aus dem **NASDAQ** durch? Unterstützen sie das Ganze visuell durch ein geeignetes **interaktives Diagramm** zwischen ausgewählten Aktien eines definierten Zeitraums und den beiden genannten Kryptowährungen.
- Berechnen Sie für jeden betrachteten NASDAQ-Titel sowie jede betrachtete Kryptowährung die **Gesamtrendite** im Zeitraum **2020-2025**, unter der Annahme, dass im Rahmen eines Sparplans **monatlich konstant 100 €** in jeden einzelnen Wert investiert wurde. Verwenden Sie hierzu die vorgegebene formale Berechnungsvorschrift zur Ermittlung der Gesamtrendite.

  Ermitteln und benennen Sie anschließend das **lukrativste Aktien-Krypto-Triplet**, bestehend aus **zwei NASDAQ-Tickern und einer Kryptowährung**, das im Zeitraum **2020-2025** die höchste Gesamtrendite erzielt hat (Stichtag 31.12.2025), wobei angenommen werden soll, dass jeder der drei Werte mit einem Sparplan in Höhe von 100€ monatlich (vgl. oben) bespart wurde.

Gesamtrendite	$= \dfrac{\text{Portfoliowert} - \text{Investiertes Kapital}}{\text{Investiertes Kapital}}$	prozentuale Wertentwicklung
Portfoliowert	$= \sum_{i=1}^{N} (\text{Anteile}_i \cdot \text{aktueller Preis}_i)$	$N =$ Anzahl der Portfoliowerte
Anzahl Anteile	$= \dfrac{\text{Investiertes Kapital}}{\text{Preis am Kaufdatum}}$	monatlich gekaufte Stückzahl
Investiertes Kapital	$= m \cdot$ Sparrate pro Monat	$m =$ Anzahl Monate

  Erstellen Sie weiterhin eine **interaktive Visualisierung** mit `Plotly`, in der die zeitliche Wertentwicklung eines Sparplans mit konstanter Sparrate von monatlich 100€ für einen auszuwählenden Titel über einen vom Nutzer frei wählbaren Zeitraum (von... bis...) dargestellt wird. Transaktionskosten können vernachlässigt werden. In einem gemeinsamen Diagramm sind dabei sowohl das **kumuliert investierte Kapital** als auch der **Portfoliowert**

darzustellen. Der darzustellende Titel soll aus einer Dropdown-Liste ausgewählt werden können, wobei die beiden Kryptowährungen Bitcoin und Ethereum sowie die Aktien aus dem NASDAQ-100-Index (Zusammensetzung zum Stand Dezember 2025) auswählbar sein sollen.
Untersuchen Sie mit Hilfe dieses Diagramms die zeitliche Entwicklung eines Sparplans für die Krypto-Werte sowie für ausgewählte Aktien.

- Erstellen Sie eine **GEO-Visualisierung** mittels `folium`, um mögliche **regionale Cluster, geographische Muster und räumliche Zusammenhänge der NASDAQ-Unternehmen** zu identifizieren und diese zu analysieren. Die Standortinformationen der Unternehmen können der beigefügten Datei `nasdaq_company_addresses.csv` entnommen werden. Es brauchen nur Unternehmen berücksichtigt werden, die in den USA ansässig sind.
- Führen Sie eine frei gewählte, interessensgeleitete Analyse des NASDAQ- und/oder Kryptomarkts durch. Entwickeln Sie hierzu eine eigene Fragestellung und wenden Sie geeignete analytische Methoden nach eigenem Ermessen an.

Aufgabe 3 (Modellbildung)

Abschließend soll die zuvor aufbereitete **NASDAQ-Datenbasis** genutzt werden, um ein **lineares Regressionsmodel**l unter Verwendung der `Python` Bibliothek `Scikit-learn` zu entwickeln, das der **Prognose zukünftiger Schlusskurse** ($\texttt{close}_{t+1}$) der **NVIDIA-Aktie** dient. Ziel ist es, sowohl die Modellgüte im Testdatensatz zu bewerten als auch eine echte, fortlaufende, auf historischen Daten basierende, Prognose zukünftiger Schlusskurse zu erzeugen. Dabei ist ein **zweistufiges Evaluationsverfahren** anzuwenden, das zunächst die Leistungsfähigkeit des Modells auf zuvor nicht gesehenen Testdaten beurteilt und anschließend die Fähigkeit des Modells untersucht, zukünftige Werte außerhalb des Trainings- und Testzeitraums vorherzusagen.

Im Rahmen der Modellbildung sind insbesondere die folgenden methodischen Anforderungen zu berücksichtigen und fachgerecht umzusetzen:

- **Feature Engineering**: Ableitung geeigneter Merkmale (existierende Features und daraus abgeleitete Features) auf Basis der historischen Zeitreihendaten, beginnend von 2015 bis 2025, um den $\texttt{close}_{t+1}$ Preis des nächsten Tages `t+1` vorherzusagen
- **Durchführung eines zeitgerechten Datensplits** in Training (2015-2023), Validierung (2024) und Test (2025). Auf Basis der Trainings- und Validierungsdaten wird anschließend ein lineares Regressionsmodell trainiert und optimiert, das im Anschluss mithilfe des Test-Splits zur Vorhersage zukünftiger Schlusskurse evaluiert wird. Die Validierungsdaten sollen insbesondere für das Feature Engineering herangezogen werden.
- In einem ersten Evaluationsschritt soll die **Güte des Modells** beurteilt werden, indem Sie den **mittleren relativen Fehler** (mean absolute percentage error, verfügbar in `sklearn.metrics`) berechnen und auswerten. Erstellen sie zudem eine geeignete **Visualisierung**, welche die **Modellvorhersagen den tatsächlichen Werten des ungesehenen** `test sets` gegenüberstellt. Welcher **funktionale Zusammenhang** liegt dem linearen Regressionsmodell zugrunde?
- In einem zweiten Evaluationsschritt soll eine echte fortlaufende Prognose zukünftiger Schlusskurse auf Basis der gesamten historischen Daten ($\texttt{close}_{t+1}$ prediction) sowie ein grafischer Vergleich der prognostizierten Werte mit den tatsächlichen Kursverläufen durchgeführt werden. Hier sollen zwei Ansätze realisiert werden:

    a. **Rückblickende Vorhersage (*ex-post*)**
    Vorhersage des NVIDIA-`close`-Kurses für den **2. Februar 2026** auf Basis der tatsächlich vorliegenden Kursdaten bis einschließlich **1. Februar 2026**. Diese Prognose kann

erst nach Veröffentlichung des Schlusskurses vom 1. Februar durchgeführt werden und erlaubt einen direkten Vergleich der Modellvorhersage mit dem real beobachteten Schlusskurs am 2. Februar (Best-Case-Evaluierung unter idealen Bedingungen). Für den Zeitraum 15. Dezember 2025 bis einschließlich 1. Februar sind die fehlenden Kursdaten mittels `yfinance` zu ergänzen.

b. **Zukunftsgerichtete und fortlaufende Prognose (*autoregressiv*)**
Vorhersage des NVIDIA-`close`-Kurses für den **2. Februar 2026** ausschließlich auf Basis der Datenlage am **18. Januar 2026** (Abgabedatum). Ab diesem Stichtag wird für jeden Folgetag jeweils der nächste Schlusskurs prognostiziert. Jede erzeugte Vorhersage ersetzt dabei den unbekannten tatsächlichen Kurswert und dient als Grundlage für die nächste Modellvorhersage. Dies führt zu einer schrittweisen Fehlerfortpflanzung und simuliert eine realistische Zukunftsprognose ohne Zugriff auf zukünftige Kursdaten.

- Betrachten sie mögliche Grenzen des eingesetzten linearen Regressionsmodells kritisch und beschreiben sie gegebenfalls alternative, für Zeitreihen oder Kursdaten geeignete, Modellierungsansätze.

Aufgabe 4 (Zusammenfassung)

Fassen Sie die **wesentlichen Erkenntnisse** Ihrer Studienarbeit in einer klar strukturierten, gut nachvollziehbaren und detaillierten Gesamtschau zusammen und schließen Sie diese mit einem **prägnanten übergeordneten Fazit** ab.

Anlage zur Projektarbeit Data Analytics

Wintersemester 2025/2026
Prof. Dr. Christian Bergler, Prof. Dr. Fabian Brunner

Füllen Sie die nachfolgende Erklärung entweder gemeinsam oder pro Gruppenmitglied aus und laden Sie eine gescannte Version mit Ihrer Einreichung auf Moodle hoch.

Name, Vorname Gruppenmitglied 1:

Matrikelnummer Gruppenmitglied 1:

Name, Vorname Gruppenmitglied 2:

Matrikelnummer Gruppenmitglied 2:

Erklärung

Hiermit wird erklärt, dass die eingereichte Projektarbeit ausschließlich von den o.g. Personen erstellt wurde. Alle verwendeten Hilfsmittel und Quellen sind in der Arbeit angegeben worden.

Ort, Datum

Unterschrift(en)