FH University of Applied Sciences TECHNIKUM WIEN

Proposal

Degree Program Informatik/Computer Science

Ein hybrider Ansatz zur Aktienkursprognose: Kombination von Machine Learning und Google Trends für NVIDIA, Google und Microsoft

By: Nathaniel Ace Panganiban
Student number: 2210257040

Advisor: Christian Brandstätter

Wien, 18.02.2025

Inhaltsverzeichnis

1. Einführung

1.1 Stand der Technik

Die Vorhersage von Aktienkursen stellt eine zentrale Herausforderung in der Finanzwelt dar. Seit Jahrzehnten werden verschiedene statistische und maschinelle Lernverfahren zur Prognose von Marktbewegungen eingesetzt. Zu den traditionellen Methoden gehören autoregressive Modelle wie ARIMA (AutoRegressive Integrated Moving Average) und GARCH (Generalized Autoregressive Conditional Heteroskedasticity) [1]. Diese Methoden sind besonders wertvoll, wenn es um die Modellierung von Volatilitäten und kurzfristigen Marktbewegungen geht. Dennoch haben sie Einschränkungen, insbesondere in der Erkennung nichtlinearer Zusammenhänge und der Berücksichtigung externer Einflussfaktoren.

Mit der zunehmenden Rechenleistung und Verfügbarkeit großer Datenmengen haben sich Machine Learning-Techniken etabliert, die in der Lage sind, komplexe Muster in Finanzzeitreihen zu erkennen. Modelle wie Random Forests oder Gradient Boosting Machines (XGBoost) werden häufig eingesetzt, um Kursbewegungen anhand historischer Daten vorherzusagen [2]. Noch leistungsfähiger sind Deep Learning-Modelle wie Long Short-Term Memory (LSTM) Netzwerke, die besonders gut für sequenzielle Daten geeignet sind und langfristige Abhängigkeiten zwischen Variablen erfassen können [3].

Neben klassischen Kursdaten werden zunehmend alternative Datenquellen zur Verbesserung der Vorhersagequalität genutzt. Google Trends, ein öffentlich zugänglicher Dienst von Google, erlaubt die Analyse des Suchverhaltens von Nutzern weltweit. Untersuchungen zeigen, dass bestimmte Suchbegriffe stark mit Marktbewegungen korrelieren können, da sie das Investoreninteresse widerspiegeln [4]. Trotz dieses Potenzials wurde Google Trends bisher nur selten systematisch in Machine Learning-Modelle zur Aktienkursprognose integriert.

1.2 Warum die gewählten Modelle für das Hybrid-Modell

Das vorgeschlagene Hybridmodell kombiniert verschiedene Modellierungsansätze, um die Stärken jedes einzelnen Verfahrens optimal zu nutzen.

- **GARCH-Modell**: Es ist speziell darauf ausgerichtet, Volatilität zu erfassen und Marktunsicherheiten zu modellieren. Volatilitätsvorhersagen sind essenziell für die Risikobewertung und können wertvolle Informationen über bevorstehende Kursbewegungen liefern [1].
- **LSTM oder XGBoost**: Diese Modelle werden zur Identifikation nichtlinearer Muster und langfristiger Abhängigkeiten eingesetzt. Während LSTM besonders für

sequenzielle Zeitreihenanalysen geeignet ist, bietet XGBoost hohe Effizienz, Skalierbarkeit und Interpretierbarkeit durch Feature-Wichtungen [2], [3].

- **Google Trends-Daten**: Die Integration von Google Trends als externer Indikator für Marktstimmung und Investoreninteresse ermöglicht eine dynamische Anpassung der Prognosemodelle an aktuelle Entwicklungen [4].

Die Kombination dieser drei Komponenten erlaubt es, sowohl historische Marktbewegungen als auch aktuelle Marktsignale in die Prognose mit einzubeziehen. Dadurch kann eine bessere Vorhersagegenauigkeit erreicht werden als mit rein datengetriebenen oder statistischen Modellen allein [8].

1.3 Motivation

Klassische Aktienkursprognosemodelle basieren hauptsächlich auf historischen Preisdaten und technischen Indikatoren. Dies führt jedoch zu mehreren Problemen:

1. **Fehlende externe Einflussfaktoren**: Traditionelle Modelle berücksichtigen keine externen Datenquellen, die das Investorenverhalten oder Markttrends widerspiegeln könnten.
2. **Begrenzte Mustererkennung**: Lineare Modelle sind oft nicht in der Lage, komplexe nichtlineare Zusammenhänge und Marktreaktionen zu erfassen.
3. **Mangelnde Interpretierbarkeit**: Viele Machine Learning-Modelle liefern präzise Vorhersagen, aber ihre Entscheidungsprozesse bleiben oft intransparent.

Durch die Integration von Google Trends als exogene Variable wird untersucht, ob das Suchverhalten der Nutzer als Frühindikator für Kursbewegungen genutzt werden kann [4]. Die Hypothese ist, dass steigende Suchvolumina für Begriffe wie „NVIDIA stock“, „Google stock“ und „Microsoft stock“ eine Korrelation zu positiven Kursentwicklungen aufweisen, während Suchbegriffe wie „sell NVIDIA“ oder „market crash“ mit negativen Entwicklungen in Verbindung stehen [9].

Zusätzlich wird Explainable AI (XAI) in Form von SHAP (Shapley Additive Explanations) verwendet, um zu verstehen, welche Faktoren die Modellvorhersagen am stärksten beeinflussen. Dies ermöglicht eine transparente Bewertung der Bedeutung einzelner Merkmale, insbesondere im Kontext der Google Trends-Daten [5].

2. Forschungsfragen

2.1 Ziele der Arbeit

- Entwicklung eines **Hybridmodells**, das traditionelle Finanzmodelle mit modernen Machine Learning-Techniken und Google Trends-Daten kombiniert.
- Untersuchung, ob alternative Datenquellen die **Genauigkeit von Aktienkursprognosen verbessern** können.
- Analyse der **Interpretierbarkeit** des Modells mithilfe von Explainable AI (XAI) durch SHAP-Werte.

2.2 Forschungsfragen

1. Wie beeinflusst die Wahl des Machine Learning-Modells (LSTM vs. XGBoost) die Vorhersagegenauigkeit des Hybridmodells?
2. Kann ein Hybridmodell aus GARCH + LSTM/XGBoost die Vorhersagegenauigkeit der NVIDIA, Google und Microsoft Aktien verbessern?
3. Wie wichtig sind historische Finanzdaten im Vergleich zu alternativen Datenquellen für ein hybrides Prognosemodell?
4. Wie stark korrelieren Google Trends-Daten mit den Aktienkursen von NVIDIA, Google und Microsoft?

3. Methoden

3.1 Datenbasis

Die Analyse basiert auf historischen Aktienkursdaten der Unternehmen NVIDIA, Google und Microsoft, bezogen von Finanzplattformen wie Yahoo Finance. Der Datensatz erstreckt sich über einen Zeitraum von 2015 bis 2024, um sowohl langfristige als auch kurzfristige Marktbewegungen zu erfassen. Die Daten umfassen:

- Eröffnungs-, Höchst-, Tief- und Schlusskurse
- Handelsvolumen
- Zusätzliche technische Indikatoren (z. B. gleitende Durchschnitte, Relative Strength Index (RSI))

Zusätzlich werden alternative Datenquellen genutzt:

- Google Trends-Daten, gesammelt über die pytrends-Bibliothek, um das Suchinteresse für relevante Begriffe wie „NVIDIA stock“, „buy Microsoft“ oder „sell Google“ zu messen.

3.2 Modellimplementierung

Das Modellierungsverfahren umfasst mehrere Schritte:

1. **Datenaufbereitung**: Normalisierung der Finanz- und Suchtrends-Daten, Entfernung von Ausreißern und Aggregation von Zeitreihenmerkmalen.
2. **Feature Engineering**: Erstellung neuer Variablen wie gleitende Durchschnitte, Volatilitätsmaße aus GARCH und Ableitung von Stimmungsindikatoren.
3. **Training und Optimierung**: Vergleich und Feinabstimmung der Hyperparameter für LSTM und XGBoost.
4. **Ensemble-Techniken**: Kombination der Modelle zur Verbesserung der Vorhersagegenauigkeit.

3.3 Evaluierungsmethoden

Die Evaluierung des Modells erfolgt anhand folgender Methoden:

- **Fehlermetriken**: Die Prognosegenauigkeit wird mittels Mean Squared Error (MSE) und Root Mean Squared Error (RMSE) gemessen. Diese Metriken quantifizieren die Abweichung zwischen den vorhergesagten und den tatsächlichen Kursen.
    - **Betrachtungszeitraum für RMSE und MSE:** Um realistische Bewertungen vorzunehmen, werden die Fehlerkennzahlen für unterschiedliche Zeitintervalle berechnet, darunter tägliche, wöchentliche und monatliche Vorhersagen. Dadurch kann analysiert werden, wie sich die Modellgenauigkeit über verschiedene Prognosehorizonte hinweg verändert.
    - **Train-Test-Split für Zeitreihen:** Da es sich um eine Zeitreihenanalyse handelt, wird kein zufälliger Split genutzt. Stattdessen wird eine gleitende Zeitfensterstrategie verwendet (TimeSeriesSplit in Scikit-Learn), um sicherzustellen, dass die Modelle nicht auf zukünftige Daten trainiert werden [7].
- **Finanzmetriken**: Neben den statistischen Fehlermetriken wird das Sharpe Ratio genutzt, um die Praktikabilität der Vorhersagen für Handelsstrategien zu bewerten. Das Sharpe Ratio misst das Risiko-Ertrags-Verhältnis der Modellprognosen und gibt an, wie profitabel die Vorhersagen im Vergleich zum eingegangenen Risiko sind [6].
- **Explainable AI (XAI) mit SHAP**: SHAP (Shapley Additive Explanations) wird zur Interpretierbarkeit des Modells eingesetzt. SHAP-Werte zeigen auf, welche Features den größten Einfluss auf die Modellprognosen haben. Dies ist insbesondere für

Finanzmärkte von Bedeutung, da es Investoren hilft, die zugrunde liegenden Entscheidungsprozesse des Modells zu verstehen [5]. Konkret ermöglicht SHAP:

- **Ermittlung der wichtigsten Einflussfaktoren** für die Aktienkursprognose.
- **Analyse der Wechselwirkungen zwischen Features**, z. B. ob steigende Google Trends-Suchvolumina einen positiven oder negativen Einfluss auf die Aktienkurse haben.
- **Visuelle Darstellungen der Feature-Importance**, um Transparenz in den Modellentscheidungen zu gewährleisten [5].

Durch diese umfassenden Evaluierungsmethoden wird sichergestellt, dass das Hybridmodell nicht nur genaue Vorhersagen liefert, sondern auch für reale Investitionsentscheidungen nachvollziehbar und praktisch einsetzbar ist.

4. Erwartete Ergebnisse

- **Nachweis einer signifikanten Korrelation** zwischen Google Trends-Daten und den Aktienkursbewegungen von NVIDIA, Google und Microsoft.
- **Erhöhung der Vorhersagegenauigkeit** durch die Kombination von GARCH mit LSTM/XGBoost im Vergleich zu bestehenden traditionellen Modellen wie ARIMA oder klassischen linearen Regressionsmodellen.
- **Identifikation der einflussreichsten Google Trends-Suchbegriffe**, die mit positiven oder negativen Kursbewegungen assoziiert sind.
- **Erstellung eines robusten und skalierbaren Modells**, das Prognosen für die drei Unternehmen generieren kann.
- **Erhöhung der Interpretierbarkeit durch SHAP**, um eine transparente und nachvollziehbare Entscheidungsfindung zu ermöglichen.
- **Potenzielle Anwendung des Modells für algorithmische Handelsstrategien**, um Marktbewegungen frühzeitig zu erkennen und darauf zu reagieren.

5. Referenzen

[1] **Tsay, R. S.** *Analysis of Financial Time Series*. Hoboken, NJ, USA: John Wiley & Sons, 2010. [Online]. DOI: 10.1002/9780470644560.

[2] **T. Chen und C. Guestrin**, *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, S. 785-794. [Online]. DOI: 10.1145/2939672.2939785.

[3] **K. Kakade, I. Jain und A. K. Mishra**, Value-at-Risk forecasting: A hybrid ensemble learning GARCH-LSTM based approach. 2022. [Online]. DOI: 10.1016/j.resourpol.2022.102903.

[4] **T. Preis, H. S. Moat und H. E. Stanley**, *Quantifying trading behavior in financial markets using Google Trends*. *Scientific Reports*, Bd. 3, Nr. 1684, 2013. [Online]. DOI: 10.1038/srep01684.

[5] **S. M. Lundberg und S. I. Lee**, *A unified approach to interpreting model predictions*. In *Advances in Neural Information Processing Systems*, Bd. 30, 2017. [Online]. DOI: 10.48550/arXiv.1705.07874.

[6] **W. F. Sharpe**, *Mutual Fund Performance*. The Journal of Business, Bd. 39, Nr. 1, S. 119-138, 1966. [Online]. DOI: 10.1086/294846.

[7] **Scikit-Learn**, "Time Series Split," Scikit-Learn Documentation, 2024. [Online]. Verfügbar: https://scikit-learn.org/stable/modules/cross_validation.html#time-series-split. Abgerufen: 9. Februar 2025.

[8] **K. Kakade, A. K. Mishra, K. Ghate und S. Gupta**, Forecasting Commodity Market Returns Volatility: A Hybrid Ensemble Learning GARCH-LSTM based Approach. 2022. [Online]. DOI: 10.1002/isaf.1515.

[9] **M. Seo, S. Lee und G. Kim**, Forecasting the Volatility of Stock Market Index Using the Hybrid Models with Google Domestic Trends. 2019. [Online]. DOI: 10.1142/S0219477519500068.