Data Mining Cup :

Prognose des Goldpreises

Namen :

- **Naceur Sayedi**
- **Abdullah Reslan**
- **Radhouen Abidi**

Gliederung

1. Projekt Zusammenfassung
2. Explorative Datenanalyse (EDA)
3. Modellierung
4. Modellauswahl & Begründung
5. Ergebnisse & Kennzahlen
6. Erkenntnisse & Fazit

Projekt Zusammenfassung

- Problemstellung : Gold Price Prediction für die nächsten 14 Tage
- Verwendete Methodik : CRISP-DM
- Verwendete Werkzeuge : Jupyter Notebooks, VS Code, Git und Gitlab
- Datenquelle : Yahoo Finance und FRED API's
- Einflussfaktoren :
    - USD_Index , EUR_USD, SP500, Oil_Price, Bitcoin_Price, VIX_Index
    - CPI, Unemployment_rate, GDP, Intrest_Rate, Zinsen, Gold-ETF-Bestände, M2-Geldmenge
    - Im Nootbook erstellte Zeit-Features
    - Monatliche und vierteljährliche Indikatoren wurden für eine konsistente Modelleingabe linear auf eine tägliche Frequenz interpoliert.

Explorative Datenanalyse (EDA)

- numerische fehlende Werte mit dem Mittelwert ausfüllen
- Keine Duplikate
- Alle Features mit einem Line-Plot plotten, um den Trend zu erkennen
- Fast alle Features zeigen einen wachsenden Trend
- Alle Features mit einem Histogramm plotten, um die Datenverteilung zu überprüfen.
- Man sieht sofort, dass der Mittelwert in einem älteren Bereich liegt
- Das heißt: alle neuen Werte liegen in einem Bereich, indem sie als von Modellen Ausreißer betrachtet werden
- Danach haben wir die Features Stationär gemacht, damit das mit ARIMAX funktioniert
- Prüfung mithilfe des ADF-Tests und Transformation mithilfe einer von uns erstellten Funktion

Explorative Datenanalyse (EDA)

- Danach haben wir die Korrelationen zwischen den Features geprüft
- Ergebnisse sind:
- Unemployment Rate hat eine Korrelation von -0.06 mit dem Gold Features ↠ entfernt
- EUR_USD hat eine Korrelation von -0.25 mit dem Gold Features ↠ entfernt
- VIX hat eine Korrelation von 0.22 mit dem Gold Features ↠ entfernt
- GPD ist sehr hoch mit sp500 korreliert , wobei der Wert 0.96 ist ↠ entfernt
- US_Intrest_Rate, M2_Euro, M2_US, CPI haben ein flaches Verhalten ↠ entfernt
- Bitcoin zeigt sehr schwankendes Verhalten, deswegen wurde es ebenfalls entfernt
- Das Training mit dem Bitcoin-Feature hat das Ergebnis nicht verbessert, sondern verschlechtert.

Modellierung - XGBoost

- Daten ab 01-01-2024
- Verschieden Datums wurden probiert, dieses Datum liefert das beste Ergebnis
- Grund dafür ist , dass die Verteilung der Daten gut ist und die neuen Preise nicht als Ausreißer betrachtet werden.
- Lag Features, Rolling Features und Zeit Features haben wir verwendet, trotzdem liefert das ein schlechtes Ergebnis, deswegen haben wir die entfernt
- Vorhersagen für die nächsten 14 Tage liefert nach Hyperparameter Tuning ein RMSE von : 47

Modellierung - XGBoost

Testen für 14 Tage mit verschiedener Möglichkeiten nach Hyperparameter Tuning

Datum	Nur Einflussfaktoren	Einflussfaktoren + Zeit-Features	Einflussfaktoren ohne irrelevante Features	Einflussfaktoren ohne irrelevante Features und Zeit-Features
Ab 2014	RMSE = 62,1	RMSE = 59,9	RMSE = 65,7	RMSE =84,7
Ab 2020	RMSE = 71	RMSE = 63	RMSE = 62	RMSE =77
Ab 2024	RMSE = 47	RMSE = 57	RMSE = 58,8	RMSE =55

Modellierung - Prophet

- Die Parameter wurden mit der Prophet.diagnostics-Bibliothek eingestellt
- Prophet.diagnostics-Bibliothek bietet eine „Time-Series-Cross-Validation" und „Performance Metrics" an ,um die Ergebnisse der Cross-Validation zu bewerten (RMSE,MAPE,..)
- Dask wurde benutzt, um die Cross-Validation-Schleifen zu beschleunigen
- Das Modell wurde nur mit dem Goldpreis , mit relevanten Faktoren und dann mit Lag Features trainiert und evaluiert

Modellierung - Prophet

Changepoint_prior_scale = [0.001 bis 0.5]

Seasonality_prior_scale =[0.01 bis 10]

Wichtigste Experimente: (09.06.2025 – 22.06.2025)

Experiment	Beschreibung	Parameters	RMSE
Base-Modell	Nur Gold Price	Changepoint_prior_scale = 0.3 Seasonality_prior_scale = 7	175.68
Base-Modell + Features	Gold Price + Regressors	Changepoint_prior_scale = 0.1 Seasonality_prior_scale = 1	172.55
+ Lag Features	(Lag_7, Lag_14)	Changepoint_prior_scale = 0.001 Seasonality_prior_scale = 1	265.22

Modellierung - ARIMAX

Feature Engineering:

- Lag-basierte Features: Lag_1, Lag_7, Lag_14
- Differenzierung & Transformation zur Erreichung von Stationarität
- Reduktion der Merkmale basierend auf Korrelationsmatrix

  ↠ z. B. entfernt: Bitcoin, CPI, Zinsen, EUR/USD

Experiment-Setup:

- Zeitreihen-Cross-Validation mit TimeSeriesSplit
- Hyperparameteroptimierung mit Optuna
- Experiment-Tracking mit MLflow

Modellierung - ARIMAX

Wichtigste Experimente: (09.06.2025 – 22.06.2025)

Experiment	Beschreibung	Parameters	RMSE
Base-Modell	Nur Gold Price	{'p': 2, 'd': 0, 'q': 2}	63.70
Base-Modell + Features	Gold Price + Alle Merkmale	{'p': 2, 'd': 1, 'q': 1}	53.02
+ Lag Features	(Lag_1, Lag_7, Lag_14)	{'p': 3, 'd': 0, 'q': 1}	47.81
Feature Reduction	Nur korrelierte Behalten, mit lag Features	{'p': 2, 'd': 1, 'q': 1}	48.76
Feature Reduction	Nur korrelierte Behalten, ohne lag Features	{'p': 0, 'd': 0, 'q': 3}	46.14

Modellauswahl & Begründung

Prognosezeitraum: 09.06.2025 – 22.06.2025

Modell	RMSE
ARIMAX	46.14
XGBoost	47
Prophet	172.55

Ergebnisse & Kennzahlen

Herausforderung:

- Exogene Daten sind für zukünftige Zeitpunkte oft nicht verfügbar, was reale Vorhersagen erschwert.

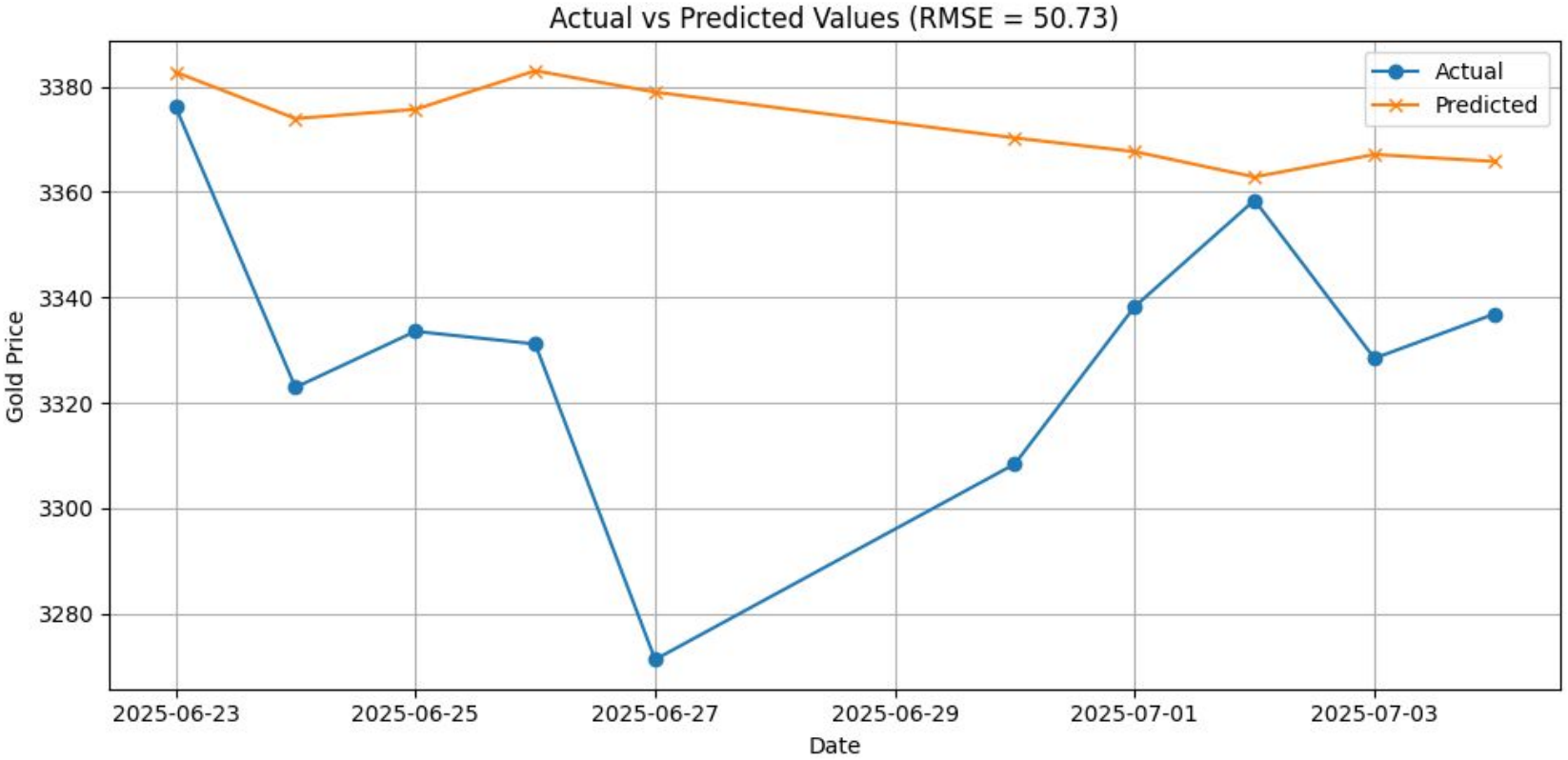
Lösung:

Da exogene Daten für zukünftige Zeitpunkte fehlen, haben wir sie durch synthetische Daten ersetzt:

- Zukünftige Werte wurden mit Hilfe des durchschnittlichen Trends und zufälligem Rauschen simuliert.
- Dadurch konnten wir realistische Forecasts trotz fehlender externer Variablen erstellen.

Ergebnis: Realistische Vorhersage mit akzeptabler Genauigkeit (RMSE: 50.73)

Ergebnisse & Kennzahlen



Erkenntnisse & Fazit

- Effektive Teamarbeit, Modellierung mit verschiedenen Algorithmen und Nutzung von EDA zur Datenanalyse.
- Einsatz von Optuna, Gridsearch , Prophet Diagnostics Bibliothek ,MLflow und GitLab zur Modelloptimierung und -verfolgung.

Danke für Ihre Aufmerksamkeit.

Haben Sie Fragen ?