Einschränkungen und Datenqualität der verwendeten Datensätze

Die verwendeten Daten von öffentlich zugänglichen Quellen - Hauptquelle Kaggle. Dabei ist zu beachten, dass die Kaggle Daten nicht in jedem Fall den hohen Qualitätsansprüchen professioneller Finanzmarktdaten genügen. Ein zentrales Beispiel betrifft den zugrundeliegenden Goldpreis. In den Kaggle-Daten ist nicht exakt nachvollziehbar, wie "Gold" definiert ist. Häufig handelt es sich dabei nicht um den Spotpreis von physischem Gold, sondern um einen ETF, der den Goldpreis abbilden soll. Für saubere Vorhersagemodelle wäre es sinnvoll, diesen ETF klar zu identifizieren und seine Entwicklung mit verfügbaren Gold-ETFs aus verlässlichen Quellen wie Yahoo Finance (z.B. GLD, IAU, etc.) abzugleichen.

Ähnliches gilt für berechnete Indikatoren wie "Trend"-Variablen. Hier ist die exakte Berechnungsgrundlage unklar. Für ein robustes Modellverständnis wäre es notwendig, die verwendeten Methoden zur Trenderkennung transparent zu dokumentieren oder selbst neu zu berechnen.

Warum log-Transformation von Handelsvolumina sinnvoll ist:

Handelsvolumina unterliegen häufig starken Ausreißern und besitzen eine rechtsschiefe Verteilung. Durch die Anwendung des natürlichen Logarithmus (log(x) bei uns reicht weil in unserem Datensatz gibt es keine Null-Werte, log(x + 1) zur Vermeidung von Problemen bei Nullen nicht nötig) werden die Daten symmetrischer und Extrema abgeflacht. Dies verbessert die Stabilität und Interpretierbarkeit von Regressions- und Machine-Learning-Modellen.

Warum Prozentuale Veränderungen von Preisen verwendet werden:

Absolute Preisniveaus sind oft wenig vergleichbar, insbesondere über verschiedene Vermögensklassen hinweg. Prozentuale Preisveränderungen (Returns) standardisieren die Bewegungen auf eine einheitliche Skala und machen Zusammenhänge sowie Abhängigkeiten zwischen Variablen modellierbar. Zudem folgen Renditen häufig besser den Annahmen klassischer Regressionsmodelle (z.B. Stationarität) als Preisniveaus.

Warum nur Adjusted Close bzw. Close verwendet wird:

Die Preise "Open", "High", "Low", "Close" sowie "Adjusted Close" eines Handelstages sind meist stark miteinander korreliert. Um Multikollinearität in Modellen zu vermeiden, wird üblicherweise nur eine dieser Preisvariablen verwendet. "Adjusted Close" berücksichtigt Dividendenzahlungen und Aktiensplits und ist daher die bevorzugte Wahl. Falls "Adjusted Close" nicht verfügbar ist, wird ersatzweise der "Close"-Preis verwendet.

Ist unser Dataset vollständig?

CODE – DATEN-CHECK:
Wir haben das Paket pandas_market_calendars verwendet und den NYSE-Kalender herangezogen, um zu überprüfen, ob alle Handelstage enthalten sind. Dabei haben wir festgestellt, dass 53 NYSE-Handelstage fehlen:

Fehlende NYSE-Handelstage (53):
2012-01-26, 2012-08-15, 2012-10-02, 2012-10-08,
2012-11-12, 2013-04-16, 2013-04-17, 2013-04-18,

2013-04-19, 2013-04-22, 2013-04-23, 2013-04-24,
2013-04-25, 2013-04-26, 2013-04-29, 2013-04-30,
2013-06-26, 2013-06-27, 2013-06-28, 2013-07-31,
2013-08-15, 2013-10-02, 2013-10-31, 2013-12-16,
2013-12-17, 2013-12-18, 2013-12-19, 2013-12-20,
2013-12-23, 2013-12-24, 2013-12-26, 2013-12-27,
2013-12-30, 2013-12-31, 2014-01-31, 2014-02-28,
2014-04-24, 2014-04-30, 2014-08-15, 2014-10-02,
2014-10-15, 2015-01-26, 2015-10-02, 2016-01-26,
2016-08-15, 2016-10-31, 2016-11-11, 2017-01-26,
2017-08-15, 2017-10-02, 2018-01-26, 2018-08-15,
2018-10-02.

Hintergrund zu den fehlenden Tagen:
An diesen Tagen (u.a) fehlen Daten in unserem Dataset:

- 26. Januar: Indischer Feiertag (Republic Day), beeinflusst Gold- und Öl-Märkte; die NYSE ist jedoch regulär geöffnet.
- 15. August: Indischer Unabhängigkeitstag, geringere Liquidität auf den Rohstoffmärkten.
- 2. Oktober: Gandhi Jayanti (Indien), ebenfalls Einfluss auf globale Rohstoffmärkte.

Diese Tage haben direkten Einfluss auf den Gold- und teilweise auch auf den Öl-Markt durch globale Handelsverflechtungen – auch wenn die NYSE offiziell geöffnet ist.

Die Liste der fehlenden Tage stimmt weitgehend mit bekannten indischen Feiertagen überein, die häufig mit Handelsunterbrechungen an der MCX (Multi Commodity Exchange) einhergehen.

Was genau werden wir vorhersagen?

Wir werden die Tagesrendite von Gold vorhersagen. Die Tagesrendite haben wir folgendermaßen definiert:

$$R_{t+1} = \left(AjGoldPric\,e_{t+1} - AjGoldPrice\,t_t \right) / AjGoldPric\,e_t$$

Doch was ist AjGoldPrice?

CODE – DATEN-CHECK:
Wir haben die Preisreihe von Kaggle mit dem GLD-ETF verglichen und festgestellt, dass sie identisch ist.

GLD ist ein börsengehandelter Fonds (Exchange Traded Fund, ETF), der den Preis von physischem Gold möglichst genau abbilden soll.
Der Fonds hält physisches Gold in Tresoren (meist in London) und bildet den Goldpreis ab, indem er die Wertentwicklung von Gold möglichst 1:1 nachbildet, abzüglich Verwaltungsgebühren.

GLD ist ein sehr guter Indikator für den Goldpreis, es gibt jedoch minimale Abweichungen zum Spotpreis von Gold, unter anderem durch Verwaltungsgebühren sowie Angebot und Nachfrage am ETF-Markt.