Bericht Semesterprojekt

Von: Louis Feuillet

# Inhalt

[Inhalt 1](#_Toc187075172)

[1 Einleitung 2](#_Toc187075173)

[1.1 Problemstellung 2](#_Toc187075174)

[2 Datensatz 2](#_Toc187075175)

[2.1 Verkaufszahlen Bäckerei 2](#_Toc187075176)

[2.2 Google-Trends 2](#_Toc187075177)

[2.3 Zuckerpreis 2](#_Toc187075178)

[3 Explorative Datenanalyse 3](#_Toc187075179)

[4 Methodik 5](#_Toc187075180)

[4.1 Kausalität 5](#_Toc187075181)

[4.2 Modellauswahl 6](#_Toc187075182)

[4.2.1 SARIMA 6](#_Toc187075183)

[4.2.2 LSTM 8](#_Toc187075184)

[5 Handlungsempfehlung 9](#_Toc187075185)

[Anhang 9](#_Toc187075186)

[5.1 Literaturverzeichnis 10](#_Toc187075187)

# Einleitung

Kleinbetriebe wie Bäckereien Leben vom Tagesgeschäft. Aus diesem Grund muss die breite Kundschaft immer zufrieden gestellt werden. Dies kann nur sichergestellt werden, wenn alle Verkaufsprodukte immer zu Verfügung stehen. Allerdings besteht dann die Gefahr, dass die Qualität durch Überlagerung sich verschlechtert und die Verwertung sich verschlechtert.

## Problemstellung

Aus den oben genannten Gründen wird versucht, anhand von Verkaufsdaten einer Bäckerei, den zu erwarteter Verkauf zu prognostizieren.

Die untersuchten Verkaufsdaten werden hierbei nach ihren Eigenschaften untersucht. Zudem werden verschiedenen Modelle nach ihrer Genauigkeit und Komplexität geprüft.

Durch den Untersuch wird anschliessend eine Handlungsempfehlung umschrieben.

# Datensatz

## Verkaufszahlen Bäckerei

Der Datensatz "French Bakery Daily Sales" auf Kaggle enthält detaillierte Verkaufsdaten einer Bäckerei. Hauptmerkmale des Datensatzes ist der Zeitraum, welcher zwischen Januar 2021 und September 2022 liegt. Der Datensatz selbst umfasst für jeden Verkauf eines Artikels, das Verkaufsdatum, Zeitpunkt, Artikelnummer des verkauften Gebäcks sowie die Anzahl der Artikel. Des Weiteren ist jeder Verkaufsvorgang als Ticket unterteilt.

Der Datensatz wird als CSV-Datei abgerufen.

Wichtig ist hier, dass es keine örtliche Angabe über die Bäckerei gibt. Es ist also nicht möglich Ortspezifische Variabeln wie Wetter hinzuzufügen.

## Google-Trends

Zusätzlich wird, der Google-Trends-Daten zum Interesse an "Brot" (englisch "bread") mit den Verkaufsdaten der Bäckerei kombiniert. Dabei werden relative Suchwerte durch Skalierung pro Tag aufgelistet. Das Heist der Wert beträgt zwischen 0 100. So wird die Popularität des Begriffs " Brot " in einem bestimmten Zeitraum bestimmt.

Die Daten werden direkt als API abgerufen.

## Zuckerpreis

Über Yahoo Finance werden weltweite Zuckepreise abgerufen. Zur Verwendung einer zusätzlichen Variable, wird der Schlusskurs des Zucker-futures an diesem Tag verwendet.

Die Daten werden direkt als API abgerufen.

# Explorative Datenanalyse

Abbildung 1 zeigt die Verkaufszahlen der Top 10 Artikel. Auffallend ist hier, dass Verkaufswerte bei gewissen Artikeln die gleiche Saisonalität und Trend aufzeigen. Dies betrifft vor allem Artikel, welche einen täglichen Umsatz von mindesten 100 Euro haben. Artikel mit weniger Umsatz zeigen kaum noch Saisonalität und Trend.

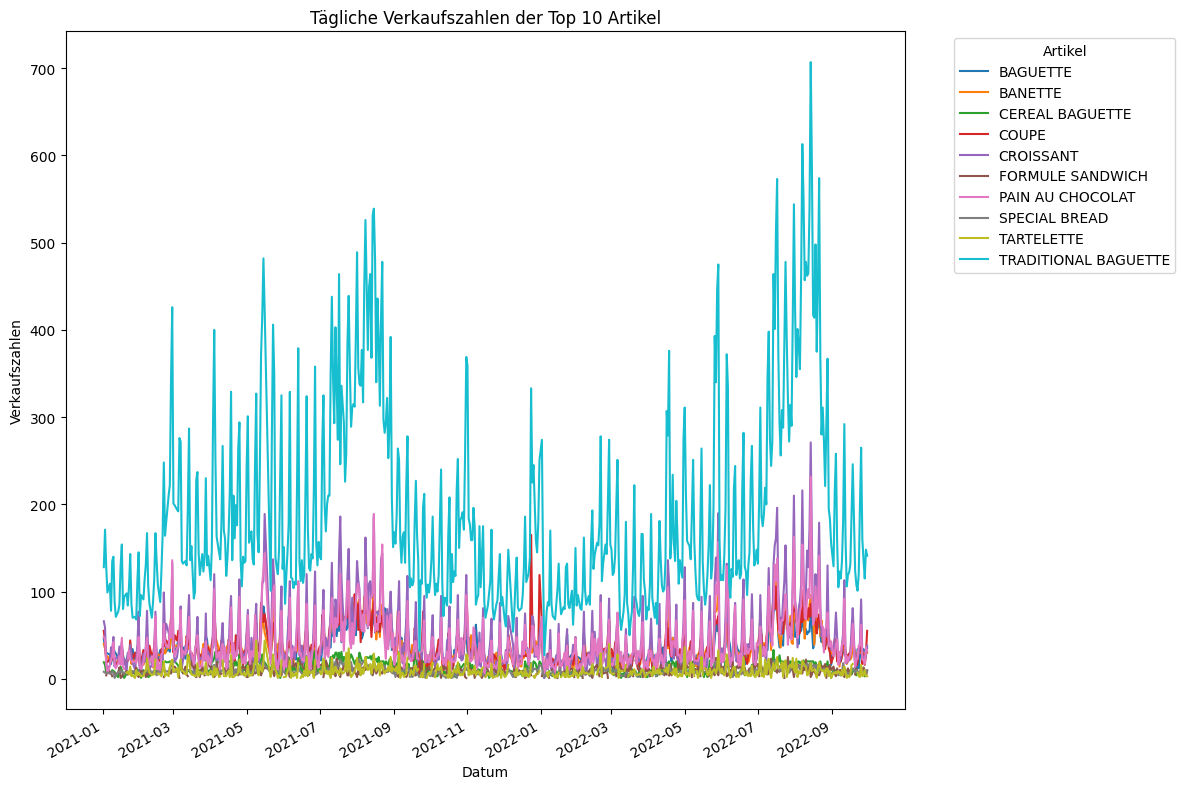


Abbildung 1:Tägliche Verkaufszahlen der Top 10 Artikel

Abbildung 2 stellt den Umsatz der Top 10 Artikel über den ganzen gemessenen Zeitraum dar. Dabei zeigt sich, dass das "Traditional Baguette" mit Abstand am besten verkauft wurde.

|  |  |
| --- | --- |
| Abbildung 2:Gesamtumsatz der Top 10 Artikel | Abbildung 3: Anzahl durchgeführte Verkäufe pro Tag in Stunden |

Abbildung 3 zeigt die Anzahl durchgeführte Verkäufe pro Tag in Stunden. Dabei wird ein tägliches Muster mio Arbeitsbeginn und Ende mit Mittagspause dargestellt.

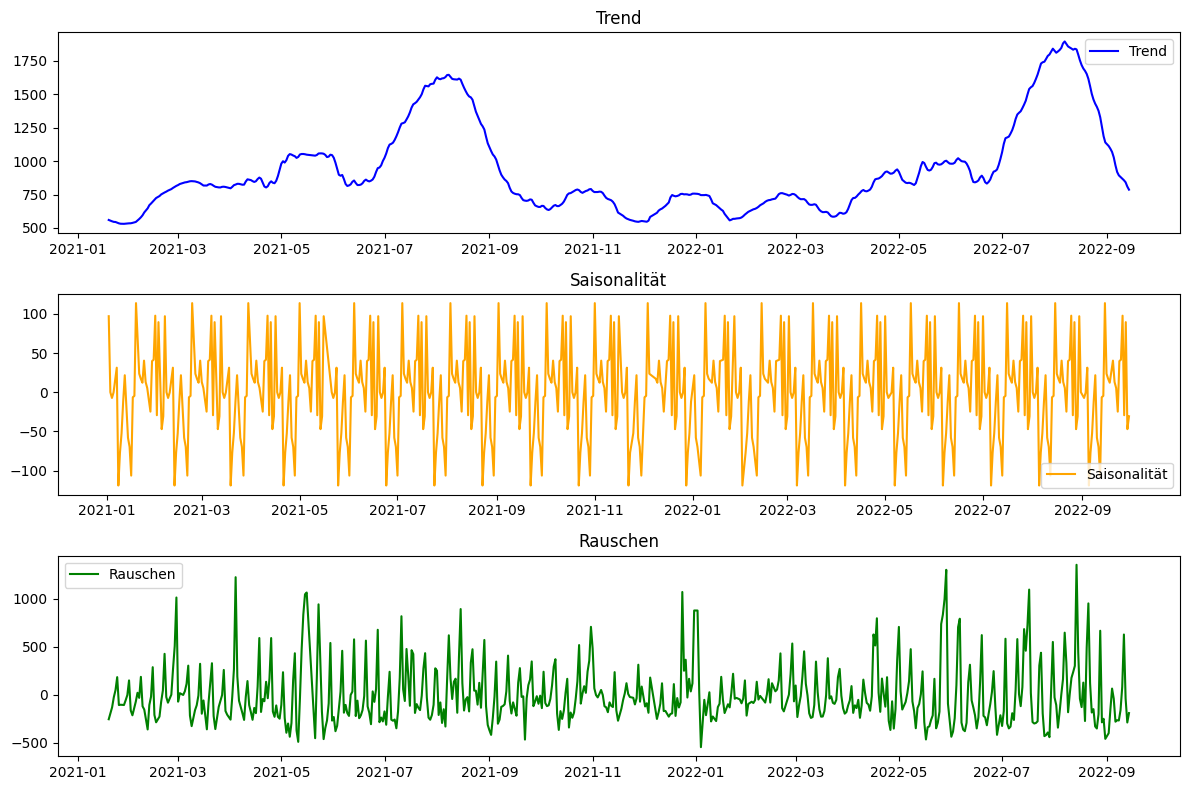


Abbildung 4:Trend, Saisonalität und Rauschen der Gesamtverkaufszahlen

Trend, Saisonalität und Rauschen der Gesamtverkaufszahlen sind in Abbildung 4 ersichtlich. Auffallend ist hier, dass der Trend klar den Herbst als Zeitpunkt aufzeigt, wo am meisten verkauft wurde.

|  |  |
| --- | --- |
| Die Saisionalität hingegen wiederholt sich immer periodisch 12 mal pro Jahr. Dies zeigt, dass eine Periode einen Monat lang ist. Beim Rauschen ist kein klares Muster ersichtlich.  Abbildung 5 stellt die des gleitenden Durchschnitts des Umsatzes dar. Das heisst, dass die Saisonalitäten weggerechnet wurden. | Abbildung 5:gleitender Durchschnitt der Verkaufszahlen |

# Methodik

## Kausalität

|  |  |
| --- | --- |
| Abbildung 6 zeigt Die Pearson-Korrelation zwischen den Variablen Gesamtverkauf, Google-Trend "bread" und dem Zuckerpreis.  Es besteht eine moderate positive Korrelation (0.39) zwischen Gesamtverkauf und Google-Trend "bread". Dies deutet darauf hin, dass ein steigendes Interesse an "bread" bei Google tendenziell mit höheren Einnahmen assoziiert ist. | Abbildung 6: Korrelationsmatrix genutzten Variabklen |

Es gibt eine sehr schwache positive Korrelation (0.08) zwischen Gesamtverkauf und Zucker Preis. Dies bedeutet, dass die Einnahmen kaum vom Zuckerpreis beeinflusst werden.

Die Abbildung 7 und Abbildung 8 zeigen die Ergebnisse des Granger-Kausalitätstests für die Variablen Zuckerpreis und Google-Trend "bread" in Bezug auf die Zielvariable Gesamtverkauf.

Alle Werte (für Lags 1–4) liegt der Zuckerpreisunterhalb der roten Signifikanzlinie (5%-Signifikanzniveau: F = 3.84). Der Zuckerpreis hat bei allen getesteten Verzögerungen keine signifikante Granger-Kausalität auf die Gesamtverkauf.

Für alle getesteten Lags bei Google-Trend "bread" liegen die Werte deutlich über der roten Signifikanzlinie, besonders bei Lag 1.

|  |  |
| --- | --- |
| Abbildung 7: Granger Kausalität von Gesamtverkauf und den Zuckerpreis | Abbildung 8: Granger Kausalität von Gesamtverkauf und Google-Trend "Brot" |

## Modellauswahl

### SARIMA

Der Datensatz mit dem Gesamtverkauf wurde in Trainings- (80% )und Testdaten (20%) aufgeteilt:

Der Parameter p gibt die Anzahl der vorherigen Zeitpunkte an, die in das Modell aufgenommen werden, um den aktuellen Wert vorherzusagen. Aus dem ACF-Plot ist zu sehen, dass die ersten Lags eine starke Autokorrelation zeigen. Daher wurde p=2 gewählt, um diese kurzfristige Abhängigkeit zu modellieren.

Der Parameter d gibt an, wie oft die Zeitreihe differenziert werden muss, um stationär zu werden. Stationarität ist eine Voraussetzung für ARIMA-Modelle. Der Dickey-Fuller-Test zeigt, dass die Zeitreihe nicht stationär ist. Eine einfache Differenzierung d=1 wurde angewandt, um dies zu korrigieren.

Der Parameter q gibt die Anzahl der vorherigen Fehlerterme an, die im Modell berücksichtigt werden, um den aktuellen Wert zu glätten. Aus dem ACF-Plot und der Analyse der Residuen zeigt ein kleiner q-Wert eine Verbesserung der Modellgüte. Die erste Differenzierung q=1 führt dazu, dass eine einfache glättende Komponente ausreicht.

Der Parameter P gibt die Anzahl der saisonalen Lags an, die für die Vorhersage des aktuellen Werts verwendet werden. Basierend auf dem saisonalen Muster oder der erwarteten Wiederholungsstruktur (falls vorhanden) wurde P=2 gewählt, um eine zweifache saisonale Abhängigkeit zu modellieren.

Der Parameter D gibt an, wie oft saisonale Differenzierung durchgeführt werden muss, um die saisonale Stationarität herzustellen. Saisonale Differenzierung D=1 wurde verwendet, um Trends zu entfernen, die in saisonalen Perioden wiederkehren können.

Der Parameter Q gibt die Anzahl der saisonalen Fehlerterme an, die im Modell berücksichtigt werden. Basierend auf saisonalen Schwankungen in den Daten ist die Berücksichtigung von zwei vorhergehenden Fehlertermen Q=1 nützlich, um saisonale Muster zu glätten.

Der Parameter s gibt die Länge der saisonalen Periode an. Die Daten werden monatlich aggregiert, und s=7 modelliert die Wiederholungsmuster innerhalb einer Woche.

Die SARIMA-Vorhersagen sind in Abbildung 9 mit Vergleich Echtwerten dargestellt. Der MAE ist dabei 470. Die SARIMA Voraussagen sind deutlich geglättet und erfassen die hohen Schwankungen der Testdaten nicht. Das Modell hat Schwierigkeiten, kurzfristige Variationen und Spitzen (z. B. hohe Umsätze) zu modellieren. Das Modell liefert einen stabilen Mittelwert, aber keine gute Abbildung der Variabilität.

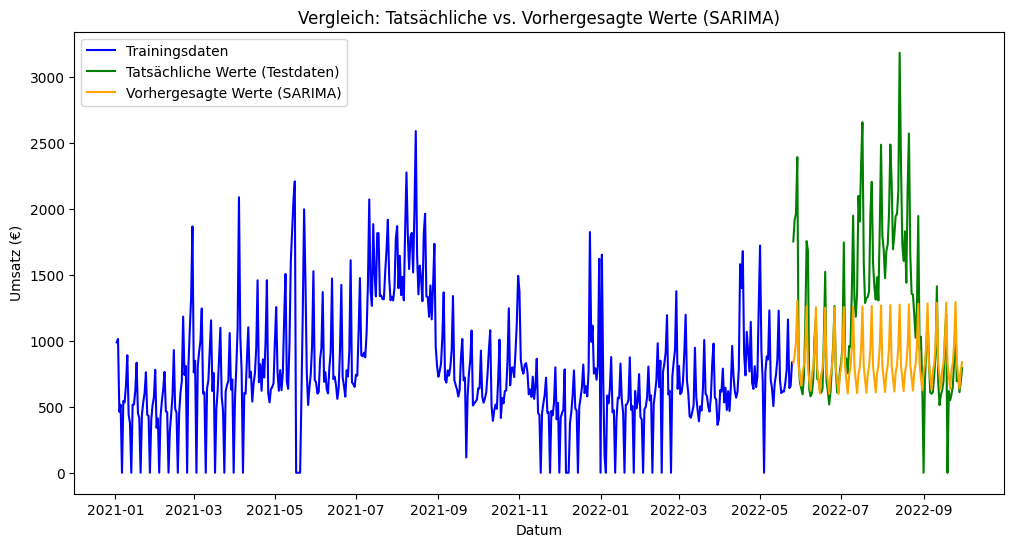


Abbildung 9:Vergleich SARIMA Voraussage mit tatsächlichen Werten

Abbildung 10 zeigt das Residuum, also die Differenz zwischen den tatsächlichen und den vorhergesagten Werten. Das Residuum zeigt klare Muster und keine zufällige Verteilung. Besonders in den Perioden mit hohen Umsatzspitzen (z. B. August 2022) sind die Residuen stark positiv. Das Modell unterschätzt Spitzenwerte und überschätzt einige niedrigere Werte, was auf eine fehlende Modellanpassung hinweist.

Abbildung 11zeigt den ACF-Plot also die Stärke der Autokorrelation der Residuen für verschiedene Lags. Auch hier gibt es signifikante Autokorrelationen bei den ersten Lags (1–10), was zeigt, dass die Residuen nicht zufällig sind. Die Autokorrelation zeigt, dass das Modell die Struktur der Zeitreihe nicht vollständig modelliert hat.

|  |  |
| --- | --- |
| Abbildung 10: Residium zwischen den tatsächlichen und den vorhergesagten Werten (Fehler des Modells). | Abbildung 11: Stärke der Autokorrelation der Residuen für verschiedene Lags |

### LSTM

Da die Resultate bei SARIMA nicht wünschenswert waren wurde entschieden, ein LSTM anzuwenden. Anders als bei SARIMA wurde zusätzlich jeder Verkaufsartikel einzeln als Feature betrachtet.

Der Datensatz mit dem Gesamtverkauf wurde in Trainings- (80%) und Testdaten (20%) aufgeteilt: Die Daten wurden zuerst skaliert, sodass alle Werte in einem Bereich von 0 bis 1 liegen. Das Modell hat dabei folgende Schichten:

* Schicht 1: Die erste LSTM-Schicht hat 50 Neuronen und gibt eine Sequenz aus, da noch eine zweite LSTM-Schicht folgt.
* Schicht 2 (Dropout): 20% der Neuronen der ersten Schicht werden zufällig deaktiviert, um Überanpassung zu vermeiden.
* Schicht 3: Eine zweite LSTM-Schicht mit 50 Neuronen gibt nur den letzten Wert der Sequenz aus, um die Vorhersage zu ermöglichen.
* Schicht 4 (Dropout): Erneut werden 20% der Neuronen deaktiviert, um Überanpassung weiter zu reduzieren.
* Schicht 5 (Dense): Eine finale Schicht mit nur einem Neuron gibt den vorhergesagten Umsatzwert aus.

Das Modell wird über 50 Epochen trainiert, wobei jeweils 32 Datenpunkte in einem Durchgang verarbeitet werden. Während des Trainings wird das Modell auf den Testdaten validiert.

Abbildung 12 zeigt eine deutliche Verbesserung gegenüber einfacheren Modellen wie SARIMA, die in früheren Ergebnissen einen höheren Fehler hatten. Der MAE beträgt 188.52. Bei einem maximalen Umsatz von über 3000 ist dieser Fehler niedrig und deutet darauf hin, dass das Modell die Umsätze gut vorhersagen kann.

Zusätzlich zeigt Abbildung 13, dass die meisten Lags liegen innerhalb des Konfidenzintervalls. Dies bedeutet, dass es keine signifikante Autokorrelation für diese Lags gibt.

|  |  |
| --- | --- |
| Abbildung 12: Vergleich von LSTM vorausgesagte Werte mit Echtwerten | Abbildung 13: Autokorrelationsplot (ACF) des LSTM-Modells |

# Handlungsempfehlung

Die Verwendung eines LSTM-Modells zur Umsatzprognose wird klar empfohlen, da es eine deutlich bessere Leistung im Vergleich zu SARIMA gezeigt hat. Der Mean Absolute Error (MAE) des LSTM-Modells beträgt 188.52, was im Vergleich zum MAE von 470 bei SARIMA eine deutliche Verbesserung darstellt. Das LSTM-Modell erfasst die Variabilität und Spitzenwerte besser als SARIMA, dass lediglich einen geglätteten Mittelwert liefert.

Artikelmerkmale (individuelle Verkaufszahlen pro Artikel) haben im LSTM-Modell bereits einen positiven Einfluss gezeigt. Die Aufnahme weiterer externer Variablen, wie z. B. Wetterdaten (falls Ortschaft bekannt), könnte Prognose weiter verbessern. Die Korrelation zwischen Google-Trends "Brot" und den Gesamtverkäufen (0.39) zeigt, dass ein steigendes Interesse an diesem Begriff ein guter Indikator für höhere Umsätze ist. Die sehr schwache Korrelation zwischen dem Zuckerpreis und den Gesamtverkäufen (0.08) deutet darauf hin, dass diese Variable aktuell keinen nennenswerten Einfluss auf die Verkäufe hat. Sie kann daher in zukünftigen Modellen vernachlässigt werden.

Die Analyse zeigt, dass die Verkäufe stark saisonal geprägt sind, mit monatlich wiederkehrenden Mustern. Dies bietet die Möglichkeit Produktions- und Lagerplanung besser an saisonale Spitzen und Täler anzupassen. Marketingkampagnen können gezielt in Monaten mit erwarteten Umsatzspitzen zu angepasst werden.

* Durchlesen
* Kontrolle
* Referenzen einfügen
* Kontroll + A alles aktualisieren

# Anhang

Der Vollständige Code wird dem Bericht beigefügt.

[Abbildung 1:Tägliche Verkaufszahlen der Top 10 Artikel 3](#_Toc187075188)

[Abbildung 2:Gesamtumsatz der Top 10 Artikel 3](#_Toc187075189)

[Abbildung 3:Anzahl durchgeführte Verkäufe pro Tag in Stunden 3](#_Toc187075190)

[Abbildung 4:Trend, Saisonalität und Rauschen der Gesamtverkaufszahlen 4](#_Toc187075191)

[Abbildung 5:gleitender Durchschnitt der Verkaufszahlen 4](#_Toc187075192)

[Abbildung 6: Korrelationsmatrix genutzten Variabklen 5](#_Toc187075193)

[Abbildung 7: Granger Kausalität von Gesamtverkauf und den Zuckerpreis 6](#_Toc187075194)

[Abbildung 8: Granger Kausalität von Gesamtverkauf und Google-Trend "Brot" 6](#_Toc187075195)

[Abbildung 9:Vergleich SARIMA Voraussage mit tatsächlichen Werten 7](#_Toc187075196)

[Abbildung 10: Residium zwischen den tatsächlichen und den vorhergesagten Werten (Fehler des Modells). 8](#_Toc187075197)

[Abbildung 11: Stärke der Autokorrelation der Residuen für verschiedene Lags 8](#_Toc187075198)

[Abbildung 12: Vergleich von LSTM vorausgesagte Werte mit Echtwerten 9](#_Toc187075199)

[Abbildung 13: Autokorrelationsplot (ACF) des LSTM Modells 9](#_Toc187075200)

## Literaturverzeichnis

**Im aktuellen Dokument sind keine Quellen vorhanden.**