

Vorhersage von Hotlinenutzung mit sensiblen Anrufszenarien

**Analyse der Telefonhotline zu Gewalt gegen Frauen in Brasilien
'Ligue 180'**

Chiara Vogt Melian

8. Januar 2025

Abstract

Anmerkungen

Die wörtlichen Zitate auf Englisch wurden aus Gründen der Nachvollziehbarkeit mit dem Übersetzer deepl.com übersetzt.

Inhaltsverzeichnis

Abstract	II
Anmerkungen	III
Abbildungsverzeichnis	VI
Tabellenverzeichnis	VIII
Symbolverzeichnis	X
Abkürzungsverzeichnis	XI
1 Einleitung	1
1.1 Potenzial der Prediction Models für	1
1.2 Fragestellung	1
1.3 Gliederung der Arbeit	1
1.3.1 Forecasting der Anrufquote	1
1.3.2 Clusteranalyse der Nutzer:innengruppen	1
1.3.3 Raumcluster und Zusammenhänge	1
2 Grundlagen	2
2.1 Verwandte Arbeiten (Influential Research)	2
2.2 Erwartete Ergebnisse	2
3 Vorbereitung der Daten	3
3.1 Der 'Ligue 180'-Datensatz	3
3.2 Datenauswahl	3
3.2.1 Feature Vorbereitung	6
3.2.2 Feature Auswahl	6
3.3 Künstliche Intelligenz	6
4 Explorative Datenanalyse	10
4.1 Datenmenge	10
4.2 Verständnis für Daten	10
4.3 Forecasting der Anrufquote	10
4.3.1 Data Collection and Cleaning	11
4.3.2 Data Analysis	11
5 Vergleich der Zeitreihenvorhersagemodelle	13
5.1 SARIMA Modell	13
5.1.1 Anwendungsbereich	13
5.1.2 Bestandteile des Modells	13
5.1.3 Voraussetzungen des Modells	15
5.1.4 Auswahl der Parameter	15
5.1.5 Bewertung und Überarbeitung des Modells	16

5.2	Hyperparameteroptimization	16
6	Methodik	17
6.1	Defining Modelling Objective	17
6.2	Datenvorverarbeitung	17
6.2.1	Vorverarbeitungsschritte	17
6.3	Auswertungsmethoden	18
7	Results	19
7.1	Structure Learning	19
7.2	Auswertungskriterien	19
8	Discussion	20
8.1	Preparation of Data	20
8.2	Structure Learning	20
8.3	Auswertungskriterien	20
9	Limitations	21
10	Conclusion	22
	Bibliography	23
A	Appendix	I

Abbildungsverzeichnis

3.1	4
3.2	5
3.3	5

Tabellenverzeichnis

3.1	hash: d2d9664e8a2fb21d0a441753b3532b3a (Vorkommen: 69)	8
3.2	Should be a caption	8
3.3	Should be a caption	9
3.4	days_to_holiday for daily dataset from 2014 to 2023	9

1

¹Alle Tabellen wurden eigenständig erstellt, siehe *R*-Code

Symbolverzeichnis

Symbol	Bedeutung
H_0 :	Nullhypothese
H_1 :	Alternativhypothese
\mathbf{I}_n :	Einheitsmatrix der Dimension n
k	Anzahl der unabhängigen Variablen
$L(\cdot)$	Plausibilitätsfunktion bzw. Likelihood-Funktion
$\ell(\cdot)$	logarithmische Plausibilitätsfunktion bzw. log-Likelihood-Funktion
n :	Stichprobenumfang
p	Anzahl der Regressionsparameter
R^2 :	Bestimmtheitsmaß
\overline{R}^2 :	adjustiertes Bestimmtheitsmaß
\mathbf{X} :	Versuchsplanmatrix
$\beta_1, \beta_2, \dots, \beta_k$:	unbekannte Regressionsparameter
$G_\Phi(\vartheta)$	Gütefunktion

Abkürzungsverzeichnis

Kapitel 1

Einleitung

1.1 Potenzial der Prediction Models für

1.2 Fragestellung

Im nächsten Abschnitt wird erläutert, welche Schwerpunkte die Analyse des Datensatzes *Ligue 180* beinhaltet. In Form von Forschungsfragen und Hypothesen leiten diese inhaltlich durch die Arbeit.

1.3 Gliederung der Arbeit

1.3.1 Forecasting der Anrufquote

Die Arbeit wird drei Themenbereiche abdecken und die Analyse des Datensatzes dabei jeweils um eine Komponente erweitern. Der erste Teil der Analyse betrachtet die Menge der Anrufe, und hat zum Ziel vorherzusagen, wann wie viele Anrufe getätigt werden, um eine ausreichende Deckung der Hotline sicherzustellen. Dafür werden die Methoden der Time Series Analysis und des Forecasting genutzt.

Auf deutsch nennt sich der Bereich der Methode Entwicklungsprognose oder Zeitreihenanalyse und gehört zu dem Bereich der quantitativen Prognosetechniken [7, S. 11].

1.3.2 Clusteranalyse der Nutzer:innengruppen

1.3.3 Raumcluster und Zusammenhänge

Kapitel 2

Grundlagen

2.1 Verwandte Arbeiten (Influential Research)

2.2 Erwartete Ergebnisse

Kapitel 3

Vorbereitung der Daten

3.1 Der 'Ligue 180'-Datensatz

Die Daten sind aus einem durch die brasilianische Regierung zur Verfügung gestellten Datensatz des Ministeriums für Menschen- und Bürgerrecht¹. Der Datensatz beinhaltet die Meldungen von Gewalt gegen Frauen, die bei der Hotline 'Ligue 180' eingegangen sind. Er besteht aus 17 einzelnen Datensätzen und erstreckt sich über die Jahre 2014 bis zum ersten Halbjahr 2024². Insgesamt enthalten alle Datensätze zusammen 2.406.112 Einträge; ab 2020 hat der Datensatz die gleiche Struktur und ergibt 1.681.161 Einträge mit 62 Kategorien. Diese Menge an Daten ist selbst bei einer großen Anzahl ungültiger Werte ausreichend für die maschinelle Verarbeitung.

Durch das brasilianische Gesetz des Zugangs zu Informationen 12.527³ werden Leitlinien festgesetzt, die den Zugang zu staatlich erhobenen Daten ermöglichen. Dadurch kann auf den Datensatz 'Central de Atendimento à Mulher – Ligue 180' über das *Portal der offenen Daten* zugegriffen werden⁴.

3.2 Datenauswahl

Eine anfängliche Durchsicht der Einträge zeigt, dass nicht alle für eine Verarbeitung geeignet sind. Die Rohdaten haben eine Menge von 2.406.097 Einträgen, darin befinden sich ab 2020 Daten mit gleichem Key *hash*⁵ und gleichem Timestamp. Diese wurden augen-

¹Übersetzung: Ministério dos Direitos Humanos e da Cidadania, URL: <https://www.gov.br/mdh/pt-br>, zuletzt aufgerufen am: 30.09.2024, 15:01 Uhr

²Quelle der Daten: <https://www.gov.br/mdh/pt-br/aceso-a-informacao/dados-abertos/ligue180> und <https://dados.gov.br/dados/conjuntos-dados/central-de-atendimento-a-mulher--ligue-180> (aufgerufen am 02.10.2024, 07:08 Uhr).

³http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/112527.htm (aufgerufen am 02.10.2024, 08:41 Uhr)

⁴<https://dados.gov.br/dados/conjuntos-dados/central-de-atendimento-a-mulher--ligue-180> (aufgerufen am 02.10.2024, 08:48 Uhr)

⁵Für das erste Semester 2020 heißt der Key *hash_par_vitima_suspeito*.

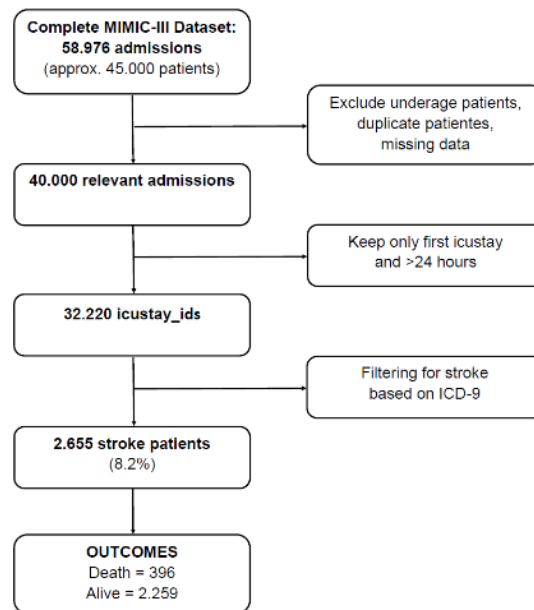


Figure 1: Overview of Filtering Steps

Abbildung 3.1:

scheinlich mehrmals aufgenommen, um mehrere Tatschwerpunkte festzuhalten. In Tabelle 3.1 kann beispielhaft nachvollzogen werden, welche Unterschiede in Werten mit dem gleichen *hash* zu finden sind. Nach der Eliminierung dieser Duplikate verbleiben 1.069.407 Datenpunkte.

Time Prediction Analysis

Für die Vorhersage der Menge der Anrufe wird ein neuer Datensatz geschaffen, der beinhaltet, wie viele Anrufe pro Stunde oder pro Tag getätigt werden.

Für die stündliche Vorhersage werden die Datenpunkte gelöscht, die keinen Stundenwert enthalten. Das ergibt eine Menge von 40.926 Datenpunkten.

Die Anrufe pro Tag ergeben nach der Bereinigung 3.652 Datenpunkte.

Ursprüngliche Menge an Einträgen: 2406097

Einträge mit hash (2020-2023): 1739566

Anzahl an einzigartigen hash: 402876

Für den Datensatz mit Stunden: 1811098 (Aussortieren der Daten ohne Stundenwert)

Datensatz ohne Duplikate: 1069407

Datensatz Anrufe pro Stunde: 40926 (Auswahlkriterium ist die Spalte hash. Diese darf jeweils nur einmal vorhanden sein)

Datensatz Anrufe pro Tag: 3652 (Auswahlkriterium hash)

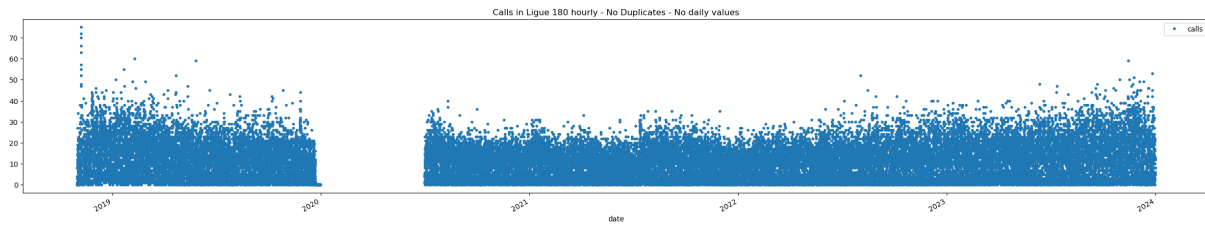


Abbildung 3.2:

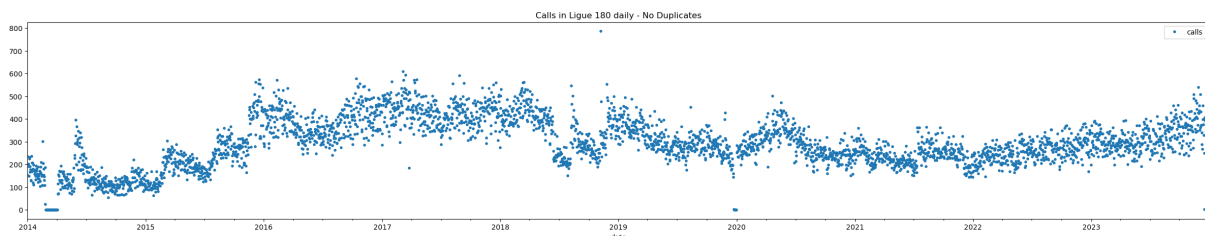


Abbildung 3.3:

Umgang mit Fehlenden Daten

Fehlende Daten in hourly_call_volume

Die fehlende Uhrzeit in der Zeitreihe mit stündlichen Daten ergibt insgesamt sechs Monate fehlender Daten in 2020.

Fehlende Daten in daily_calls

Im Datensatz daily_calls verzeichnen 45 Tage keine Anrufe. Diese sind in einem Block vom 26. Januar 2014 bis 03. April 2014 (37 Tage) zu finden und in einem zweiten Block vom 24. Dezember 2019 bis 31. Dezember 2019 (8 Tage). Da die Nullwerte jeweils als Block vorzufinden sind, ist es wahrscheinlicher, dass fälschlicherweise keine Daten erhoben wurden bzw. in dem online verfügbaren Datensatz nicht vorhanden sind. Weniger wahrscheinlich ist, dass in diesen Zeiträumen keine Anzeigen stattfanden; insbesondere für den zweiten Block, der Weihnachten und Silvester umspannt. Vorherige Forschungsergebnisse haben gezeigt, dass gerade Familienfeiern zu einem Anstieg häuslicher Gewalt führen. Außerdem finden sich Zusammenfassungen mit Statistiken zu diesem Datensatz, hier werden Zahlen zu der allgemeinen Menge an Anrufen gezeigt, z.B. einem Anrufvolumen in 2014 von 1.348 pro Tag und 40.425 Gespräche pro Monat [6, S. 5]. Aus dem Bericht lässt sich keine konkrete Zahl von Anzeigen pro Monat oder pro Tag ableiten, jedoch zeigen die Zahlen, dass es nicht der Fall ist, dass in den Zeiträumen keine Anzeigen getätigt wurden [6, 4].

Deshalb ist es wahrscheinlich, dass es in diesen Zeiträumen Anrufe gab

Dezember 2019 Nullwerte für März 2019: Daten sind für den Monat vorhanden im Balanco 2019, warum diese Daten nicht in der Statistik sind ist nicht klar. Dezember ist auffällig weil es vom 24. bis 31. ist: einerseits Vermutung, dass da besonders viele Anrufe sind wegen Feiertage, Alkohol, Familienstreit aka nicht null. Deswegen kann ich mir nicht vorstellen, dass die zu hatten. Vielleicht haben sie aber keine Anzeigen aufgenommen,

- Suchstrategien, z.B. uninformierte Suche, heuristische Suche,
- Wissensrepräsentationen, z.B. Ontologien und Semantic Web,
- Machine Learning, z.B. neuronale Netze und Deep Learning in Kombination von supervised und unsupervised Learning und Reinforcement Learning,
- Natural Language Processing,
- Computervision,
- Robotik.

Bei den verschiedenen Ansätzen wird zwischen symbolischen und nicht-symbolischen unterschieden. Symbolische Ansätze bezeichnen Techniken und Ansätze des Maschinellen Lernens bei denen die Merkmale und Zusammenhänge von Mustern expliziert und repräsentiert sind. Das Verhalten dieser Techniken ist nachvollziehbar, kalkulierbar und interpretierbar. Zu diesen Techniken gehören Entscheidungsbäume, Ontologien, das Semantic Web und Wissensgraphen [2, S. 9]. Nicht-symbolische Ansätze speichern Wissen implizit. Dieses Wissen muss anhand großer Datensätze trainiert werden. Die Ergebnisse der Wissensrepräsentationen sind deshalb unvorhersehbar und manchmal auch nicht nachvollziehbar [2, S. 9]

Tabelle 3.1: hash: d2d9664e8a2fb21d0a441753b3532b3a (Vorkommen: 69)

'Profissão_do_suspeito'	'CONSELHEIRO TUTELAR'
'violacao'	'INTEGRIDADE>PATRIMONIAL>INDIVIDUAL' 'INTEGRIDADE>FÍSICA>MAUS TRATOS' 'INTEGRIDADE>PSÍQUICA>AMEAÇA ou COAÇÃO' 'INTEGRIDADE>PSÍQUICA>CONSTRANGIMENTO'
'Motivação'	'PARA FINS DE EXPLORAÇÃO DO TRABALHO.COMÉRCIO/ INDÚSTRIA' 'EM RAZÃO DE CONDIÇÕES FÍSICAS, SENSORIAIS, INTELECTUAIS OU M' 'POR CONFLITO AGRÁRIO.QUILOMBOLAS' ... 'PARA FINS DE EXPLORAÇÃO SEXUAL' 'EM RAZÃO DA RELAÇÃO DE EN' 'POR CONFLITO AGRÁRIO.PESCA' 'COM HUMILHAÇÃO' 'EM RAZÃO DA PROFISSÃO' ... 'DA COABITAÇÃO/ CONVIVÊNCIA FAMILIAR/ RELAÇÃO AFETIVA' 'FOI PRATICADO POR DUAS OU MAIS PESSOAS' 'PARA FINS DE EXPLORAÇÃO DO TRABALHO.DOMÉSTICO' 'PARA FINS DE REMOÇÃO DE ÓRGÃOS/ TRÁFICO DE ÓRGÃOS' 'EM PÚBLICO OU POR MEIO QUE FACILITE A DIVULGAÇÃO/ NO ÂMBITO' 'POR CONDUTAS EXCESSIVAS/ DESNECESSÁRIAS/ DESACONSELHADAS' 'NA FORMA DE AUXÍLIO/INSTIGAÇÃO/INDUZIMENTO/INCITAÇÃO' 'PARA FINS DE EXPLORAÇÃO DO TRABALHO.INFORMAL' 'FALTA DE ACESSIBILIDADE.NOS MEIOS DE TRANSPORTE' 'EM DESCUMPRIMENTO DE MEDIDA PROTETIVA' 'EM RAZÃO DO SEXO' 'RESULTANDO EM LESÃO SEGUIDA DE MORTE' 'RESULTANDO EM LESÃO' 'NA FORMA CULPOSA' 'POR VIOLÊNCIA INSTITUCIONAL' 'COM FINS CO' 'RESULTANDO EM LESÃO GRAVE' 'DO AGRESSOR POSSUIR INFLUÊNCIA JUNTO ÀS AUTORIDADES LOCAIS' 'NA RELAÇÃO FAMILIAR' 'EM RAZÃO DE CONFLITO DE IDEIAS' 'EM RAZÃO DA RELIGIÃO' 'EM RAZÃO DE RAÇA/COR' 'PARA FINS DE EXPLORAÇÃO DO TRABALHO.OUTROS' 'PARA OBTENÇÃO DE BENEFÍCIO FINANCEIRO/ GANÂNCIA' 'EM RAZÃO' 'POR CRIME AMBIENTAL.COM FINS DE EXTRATIVISMO.MINERAL' 'POR CRIME AMBIENTAL.PESCA' 'RESULTANDO EM UMA DEFICIÊNCIA EM RAZÃO DA VIOLÊNCIA' 'EM RAZÃO DE ORIENTAÇÃO SEXUAL/ IDEOLOGIA DE GÊNERO' 'POR CRIME AMBIENTAL.DE CAÇA' 'PARA FINS DE EXPLORAÇÃO DO TI' 'COM RESULTADO MORTE' 'EM RAZÃO DA ORIGEM' 'POR MOTIVO VIL, TORPE, INSIDIOSO, CRUEL, À TRAIÇÃO, OU POR DIN' 'EM RAZÃO DE SER MULHER' 'COM VÍTIMA EM SITUAÇÃO DE RUA' 'FALTA DE ACESSIBILIDADE.NO ESPAÇO EDIFICADO' 'PARA FINS DE ATIVIDADE ILÍCITA' 'RESULTANDO EM LESÃO GRAVÍSSIM' 'POR CRIME AMBIENTAL.COM FINS DE EXTRATIVISMO.VEGETAL' 'NA F' 'POR CONFLITO AGRÁRIO.DE COMUNIDADES TRADICIONAIS' 'POR CRIME AMBIENTAL.PARA EXPANSÃO AGROPECUÁRIA' 'FALTA DE ACESSIBILIDADE.NA COMUNICAÇÃO' 'EM RAZÃO DE SER COMUNICADOR SOCIAL' 'EM RAZÃO DE QUAISQUER FORMAS DE DISCRIMINAÇÃO' 'VALENDO-SE DA HOSPITALIDADE' 'POR CONFLITO AGRÁRIO.INDÍGENA' 'FALTA DE ACESSIBILIDADE.NOS SISTEMAS DE COMUNICAÇÃO OU DE T' 'POR CONFLITO AGRÁRIO DE CAÇA' 'PARA FINS DE ADOÇÃO'

Tabelle 3.2: Should be a caption

Tabelle 3.3: Should be a caption

Tabelle 3.4: days_to_holiday for daily dataset from 2014 to 2023

Kapitel 4

Explorative Datenanalyse

Im folgenden Kapitel wird Anhand der Methoden in [8] erforscht, in welchem Zusammenhang die Daten miteinander stehen.

4.1 Datenmenge

Erste Fragen zu dem Datensatz aus [8, S.77f.]. Die kleinste Einheit im Datensatz, also der unique identifier: hash Menge der Daten: Spalten x Reihe Key categorical values and the frequencies of each value? Verteilung kontinuierlicher Variablen Zusammenhänge zwischen den einzelnen Variablen Variablen mit Werten außerhalb des Erwartungsbereichs und fehlende Werte

4.2 Verständnis für Daten

”When working with time series, it is critical that you learn more about the data you are working with and how it relates to the problem you are attempting to solve. For example, when working with manufacturing or sales data, you cannot assume that an organization’s working day is Monday to Friday or whether it uses the standard calendar year or fiscal year. You should also consider understanding any holiday schedule, annual shutdowns, and other matters related to the business operation.” Atwan.2022 *Time Series Analysis with Python Cookbook*.

4.3 Forecasting der Anrufquote

Anhand des Methodenbuchs *Time Series Analysis and Forecasting* der Autoren MONTGOMERY, JENNINGS, KULAHCI 2015 [5] werden im folgenden die Schritte beschrieben, die für eine Analyse im Bereich Forecasting benötigt werden.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t \quad (5.1)$$

Darin ist y_t die Zeitreihe, c ist der Durchschnitt der Veränderung von aufeinanderfolgenden Werten und ε_t bezeichnet das weiße Rauschen [3, Kapitel 8.1, 8.3].

Autoregressive Modelle verlangen stationäre Daten.

Moving Average Modell

In diesem Modell werden nicht die vergangenen Werten für die Vorhersage genutzt, sondern die vergangenen Vorhersagefehler mithilfe des gleitenden Durchschnitts (Moving Average) [3, Kapitel 8.4]. Das Modell **MA(q) model** wird mit folgender Formel dargestellt:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (5.2)$$

Non-Seasonal ARIMA Modell

Das nicht saisonale ARIMA (AutoRegressive Integrated Moving Average) Modell ist die Kombination des AR und des MA Modells. Integration stellt hier das Gegenteil von Differenzierung dar [3, Kapitel 8.5]. Das ARIMA Modell wird mit folgender Formel beschrieben:

$$y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (5.3)$$

Das **ARIMA(p, d, q) model** hat die Parameter

- p = Ordnung des autoregressiven Teils;
- d = Grad der ersten Differenzenbildung;
- q = Ordnung des Teils des gleitenden Durchschnitts.

Seasonal ARIMA Modell

Es ist möglich dem ARIMA Modell eine saisonale Komponente hinzuzufügen. Die saisonale Komponente wird durch einen weiteren Term in Großbuchstaben beschrieben. Das Modell wird beschrieben mit:

$$ARIMA(p, d, q)(P, D, Q)_m \quad (5.4)$$

Der Parameter m bezeichnet die Anzahl der Beobachtungen pro Jahr. Der Autor [3, Kapitel 8.9] benennt das Beispiel von $ARIMA(1,1,1)(1,1,1)_4$, das die Formel

$$(1 - \phi_1 B) (1 - \Phi_1 B^4) (1 - B) (1 - B^4) y_t = (1 + \theta_1 B) (1 + \Theta_1 B^4) \varepsilon_t \quad (5.5)$$

herbeiführt. Daran kann erkannt werden, dass die neu hinzugefügten saisonalen Terme mit den bestehenden Termen multipliziert werden.

5.1.3 Voraussetzungen des Modells

Damit ein Datensatz in ein SARIMA Modell verarbeitet werden kann, muss dieser das Kriterium der Stationarität erfüllen.

Stationarität

Stationarität bezeichnet die Eigenschaft, dass die Eigenschaften einer Zeitreihe, wie Mittelwert und Varianz über die Zeit konstant bleiben [7, S. 26]. Folgende drei Eigenschaften müssen erfüllt sein, damit ein Prozess als stationär (oder schwach stationär) bezeichnet werden kann:

1. $\mu(t) =: \mu$ (der Mittelwert) hängt nicht von der Zeit t ab.
2. $\sigma^2(t) =: \sigma^2$ (die Varianz) hängt nicht von der Zeit t ab.
3. $\gamma(t_1, t_2) =: \gamma(t_2 - t_1)$ (die Autokovarianz) hängt nur von der Zeitdifferenz $t_2 - t_1$ ab [7, S. 26].

Definition: Erfüllt ein stochastischer Prozess 2. Ordnung die Eigenschaften (1) bis (3), so heißt er schwach stationär [7, S. 26].

Der Mittelwert μ der Datenreihe verändert sich nicht, somit hat die Datenreihe keine Trends, also über die Zeit fallende oder steigende Werte.

Die Varianz σ^2 bezeichnet die Entfernung der Werte um den Mittelwert (Streuung der Werte); auch diese bleibt konstant.

Die Autokovarianz $\gamma(t_1, t_2)$ beschreibt den Zusammenhang zwischen zwei Werten in der Zeitreihe. Bei einer stationären Zeitreihe darf dieser Zusammenhang nur durch die Zeitdifferenz bestimmt werden [7, S. 26]. Eine stationäre Zeitreihe sollte also zu jedem Zeitpunkt ähnlich aussehen. Jedoch kann eine solche Zeitreihe auch zyklische Muster haben, die aber keinem Trend folgen [3, Kapitel 8.1].

5.1.4 Auswahl der Parameter

Die Parameter p, d, q und P, D, Q sowie m müssen ausgewählt werden, um das beste SARIMA Modell zu erstellen.

Autocorrelation Function ACF und Partial Autocorrelation Function PACF

Für die Bestimmung der Parameter sind zwei Funktionen hilfreich: die Autocorrelation Function (ACF) und Partial Autocorrelation Function (PACF). Aus einer Zeitreihe ist es meist nicht möglich die Autokorrelation einer Zeitreihe herauszulesen, mithilfe der ACF und PACF-Plots kann das jedoch gelingen [3, Kapitel 8.5].

In einem ACF-Plot wird die Autokorrelation dargestellt, indem die Beziehung von y_t zu y_{t-k} entlang der Reihe von k gezeigt wird. Damit die

5.1.5 Bewertung und Überarbeitung des Modells

5.2 Hyperparameteroptimization

Fast alle Vorhersagemethoden braucht von außen festgelegte Parameter, um geeignete Vorhersagen zu modellieren. Diese festgelegten Parameter sind Hyperparameter. Um die zu finden, können verschiedene Methoden der Hyperparameteroptimierung angewendet werden. Klassischerweise wird dafür die Gittersuche (Gridsearch), die zufällige Suche (Random Search), die bayesische Optimierung oder durch Expertise und heuristische Annäherung händisch ausgewählte Parameter.

Heuristische Annäherung an passende Parameter

Optuna

Kapitel 6

Methodik

6.1 Defining Modelling Objective

6.2 Datenvorverarbeitung

In dem Leitfaden *Data Preparation for Machine Learning* beschreibt [1] die Datenvorverarbeitung als elementaren Bestandteil der Analyse. Er argumentiert, dass die Verarbeitungsalgorithmen seit Jahren bekannt sind, jedoch die evaluierten Datensätze die Innovation mitbringen, weswegen diese besonders gut verarbeitbar sein sollten [1, 14]. Demnach ist das, was die Qualität der Studien ausmacht, nicht der ausgewählte Algorithmus, sondern die Verarbeitung der Daten von Rohdaten zu Daten für die maschinelle Verarbeitung [1, 9]. Dabei muss herausgefunden werden, wie die zugrunde liegende Struktur des Datensatz bestmöglich herausgearbeitet werden kann [1, 8]. Das ist abhängig davon, welcher Algorithmus verwendet werden soll, um das Modell zu erstellen [1, 12].

6.2.1 Vorverarbeitungsschritte

Je nach Algorithmus können irrelevante oder korrelierende Daten die Vorhersage des Modells verschlechtern. Die Wahl der Vorverarbeitung hängt zudem auch von dem vorliegenden Datensatz ab. Im Prozess des Feature Engineerings werden entscheidende Features, die das Design des Modells mitbestimmen herausgearbeitet [1, 12].

Bei der Wahl der Vorverarbeitungsschritte können zwei Es gibt den Ansatz durch das Modellieren herauszufinden, welche Schritte in der Datenvorverarbeitung notwendig sind. So können die Zusammenhänge der Daten beim Erstellen des Modells herausgearbeitet werden [1, 13]. Auf der anderen Seite gibt es die Möglichkeit, sich für jede Datenzeile zu überlegen, wie diese Aussehen müsste, um bestmöglich deren Charakter herauszuarbeiten [1, 13]. Der Autor plädiert dafür, eine Balance zwischen den beiden Ansätzen zu finden, da beide Ansätze mächtig seien und zu einer Ausarbeitung der zugrunde liegenden Struktur beitragen [1, 13].

Entscheidend ist laut [1, 25ff.] ebenso die Reihenfolge der Vorverarbeitungsschritte. Der Autor benennt folgende Schritte als gängige Praxis:

- Prepare Dataset
- Split Data
- Evaluate Model.

Diese Reihenfolge sieht er kritisch, da sie zu data leakage¹ führen kann und somit das Modell beeinflussen kann [1, 26]. Ein Beispiel für data leakage findet sich in der Normalisierung einer Variable in der Skala von 0 bis 1. Dafür muss Minimum und Maximum aller Variablen identifiziert werden und somit erhält das Model Informationen über die globale Verteilung und somit über die Testdaten. Bei vielen Methoden, um Daten zu normalisieren und leere Felder zu füllen greift dieser Mechanismus.

Aus diesem Grund plädiert [1] für diese Schritte, um die Daten vorzuverarbeiten:

- Split Data
- Fit data preparation on Training Dataset
- Apply Data Preparation to Train and Test Datasets
- Evaluate Models.

Nicht nur die Vorbereitung der Daten sollte nur mit dem Trainingsdatensatz erfolgen; dieser Ansatz sollte die gesamte Modellierung umfassen, wie Feature Selection, Feature Engineering und Dimensionality Reduction [1, 27].

Diese Vorannahmen führen dazu, dass der erste Schritt der Vorverarbeitung die Einteilung in Test- und Trainingsdaten ist.

Unterteilung der Daten in Trainings und Testdaten

Einladen der Daten als Panda-Dataframe

Verbinden der Jahre 2021 bis 2023

Nutzen der Methode train_test_split von sklearn.model_selection

Grundlagen der Datenbereinigung

6.3 Auswertungsmethoden

¹Data leakage bedeutet, dass das Model Informationen erhält, die zu einem Vorteil für eine bessere Modellierung werden. Das können Testdaten sein, die den Trainingsdaten zur Verfügung stehen oder wenn Informationen aus der Zukunft der Vergagenheit zur Verfügung stehen [9, 93].

Kapitel 7

Results

7.1 Structure Learning

7.2 Auswertungskriterien

Kapitel 8

Discussion

8.1 Preparation of Data

8.2 Structure Learning

8.3 Auswertungskriterien

Kapitel 9

Limitations

Kapitel 10

Conclusion

Bibliography

Literaturverzeichnis

- [1] BROWNLEE, J. : *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python* <https://books.google.com/books?hl=en&lr=&id=uAPuDwAAQBAJ&oi=fnd&pg=PP1&dq=+Machine+Learning+Data+Preparation&ots=Cl6LscgNoX&sig=U4FadKC2XWMUxnpqRg6oAcpFjQg>
- [2] EGE, B. : Einblick in die Welt der künstlichen Intelligenz. Version: 2021. http://dx.doi.org/10.1007/978-3-658-31938-0_{_}1. In: EGE, B. (Hrsg.) ; PASCHKE, A. (Hrsg.): *Semantische Datenintelligenz im Einsatz*. Wiesbaden and Heidelberg : Springer Vieweg, 2021. – DOI 10.1007/978-3-658-31938-0_1. – ISBN 978-3-658-31937-3, S. 1–17
- [3] HYNDMAN, R. J. ; ATHANASOPOULOS, G. : *Forecasting: Principles and practice*. 2nd edition. Lexington, Ky. : Otexts online open-access textbook, May 2018. – ISBN 978-0987507112
- [4] MINISTÉRIO DA MULHER, DA FAMÍLIA E DOS DIREITOS HUMANOS: *Balanço 2019: Ligue 180 - Central de Atendimento à Mulher*. Brasília/DF, Brasil,
- [5] MONTGOMERY, D. C. ; JENNINGS, C. L. ; KULAHCI, M. : *Introduction to time series analysis and forecasting*. Second edition. Hoboken, New Jersey : Wiley, 2015 (Wiley series in probability and statistics). – ISBN 978-1-118-74511-3
- [6] SECRETARIA DE POLÍTICAS PARA AS MULHERES: *Balanço 2014: Ligue 180 - Central de Atendimento à Mulher*. Brasília/DF, Brasil,
- [7] VOGEL, J. : *Prognose von Zeitreihen: Eine Einführung für Wirtschaftswissenschaftler*. Wiesbaden : Springer Gabler, 2015. – ISBN 978-3-658-06836-3
- [8] WALKER, M. : *Python data cleaning cookbook: Prepare your data for analysis with pandas, NumPy, Matplotlib, scikit-learn and OpenAI*. Second edition. Packt Publishing Ltd (Expert insight). <https://learning.oreilly.com/library/view/-/9781803239873/?ar>. – ISBN 978-1-80323-987-3
- [9] ZHENG, A. ; CASARI, A. : *Feature engineering for machine learning: Principles and techniques for data scientists*. Beijing and Boston and Farnham and Sebastopol and

Tokyo and Beijing and Boston and Farnham and Sebastopol and Tokyo : O'Reilly,
2018. – ISBN 1491953241

Anhang A

Appendix

Erklärung der Urheberschaft

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form in keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Ort, Datum

Unterschrift