Vorhersage von Hotlinenutzung mit sensiblen Anrufszenarien

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

Chiara Vogt Melian

20. November 2024

Abstract

Anmerkungen

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

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

[1]

[1] Alle 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
$\beta_1, \beta_2, \ldots, \beta_k$:	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.

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 Bürgerrecht[1]. 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 2024[2]. 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 12.527[3] 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 werden[4].

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-

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

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

[3] `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.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)

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

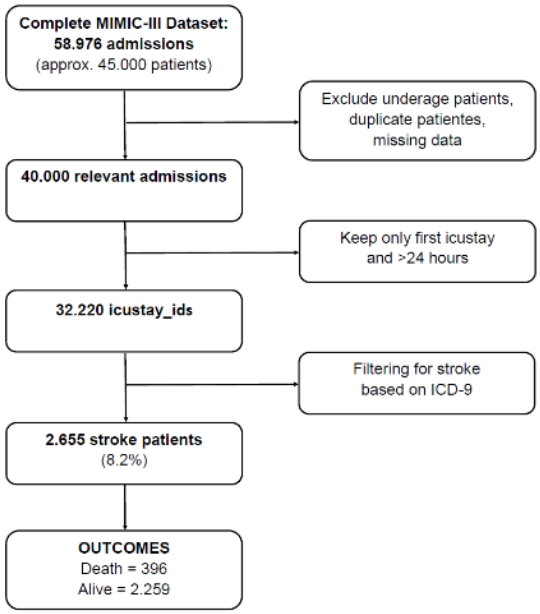

Figure 1: Overview of Filtering Steps

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)

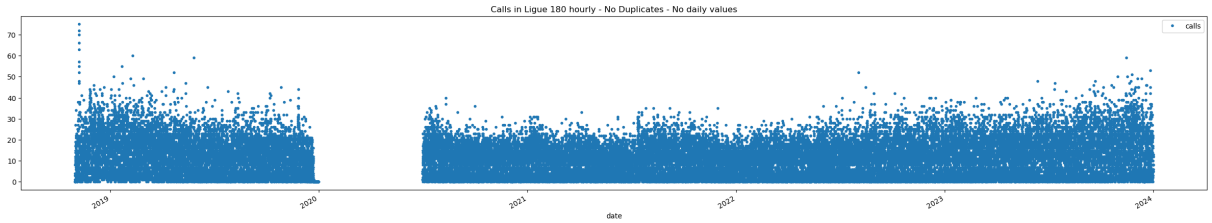

Abbildung 3.2:

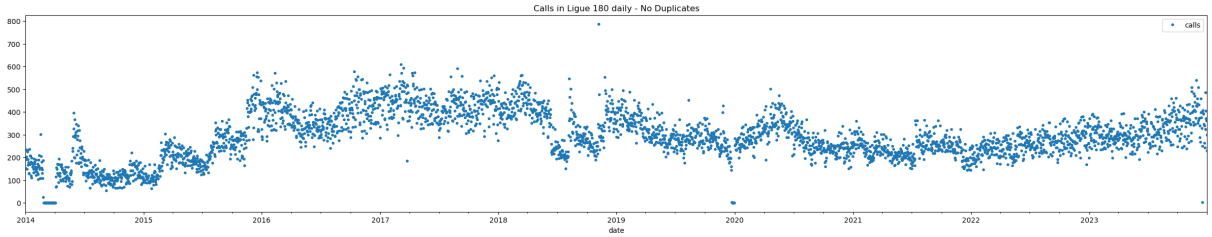

Abbildung 3.3:

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,

aber dann funktionieren auch die Schutzmaßnahmen nicht.

Clusteranalyse von Nutzungsgruppen

3.2.1 Feature Vorbereitung

Der Datensatz enthält 62 verschiedene Kategorien.
- Zusammenführen der Datensätze, inwieweit möglich?
- Tabelle mit verschiedenen Kategorien
- Tabelle mit Menge der Values
- Zusammenführung von Values - Übersicht
- Zusammenhangsgrafik der Kategorien (die vom Tablet)
- Anhang mit Kategorien und Werten

H1

- Daten die zwischen leeren Tagen Monaten liegen bei df_hourly?
- hinzufügen der Ferientage: Ferientage eingetragen, mit Differenz: -1 bedeutet einen Tag nach Feiertag und +1 einen Tag vor Feiertag - Ergebnisse sind 3.2.1 - Das gleiche für df_daily gemacht

3.2.2 Feature Auswahl

H1

-> Nullwerte für df_hourly schon gezeigt

- Daraus Schlussfolgerung welche Daten gewählt werden
- Tabelle mit Nullwerten
-

3.3 Künstliche Intelligenz

Die Künstliche Intelligenz ist das mit vielen verschiedenen Hoffnungen verbundene zukunftsweisendste Forschungfeld der Gegenwart. Die großen Hoffnungen und Erwartungen begründen sich auch in den vielfältigen Ansätzen und Techniken, die in den Bereich der künstlichen Intelligenz fallen [3, S. 2]. So gehört zum Bereich künstliche Intelligenz laut [3, S. 2]:

- die Logik, z.B. Aussagenlogik, Prädikatenlogik, Entscheidbarkeit,
- unsicheres Wissen und Schlussfolgerungen, z.B. Bayessche Netze, Fuzzy-Logik,

- 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 [3, 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 [3, 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 ÂMBIT '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 LESÃ '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 RAZÃ '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 GRAVÍSSI '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

2020-01-31	31
2020-02-29	29
2020-03-31	31
2020-04-30	30
2020-05-31	31
2020-06-30	30

Tabelle 3.3: Should be a caption

1	90
2	90
0	98
-1	88
3	88
...	
-56	10
-57	10
-58	10
-59	10
-60	10

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 [7] erforscht, in welchem Zusammenhang die Daten miteinander stehen.

4.1 Datenmenge

Erste Fragen zu dem Datensatz aus [7, 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.äus 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.

4.3.1 Data Collection and Cleaning

Zu Beginn ist es wichtig die vorhandenen verfügbaren Daten auf ihre Nützlichkeit zu prüfen. Es können Werte fehlen, Daten verfälscht sein oder jegliches anderes Datenproblem auftreten. Hier muss festgehalten werden, welche Daten weshalb genutzt werden [5, 14]. Außerdem sollten die Daten bestimmten Gütekriterien entsprechen, laut [5, 17] sind diese *„accuracy, timeliness, completeness, reprensatativeness, and consistency*[1]*."* Im folgenden Abschnitt wird geprüft, ob die Daten diese Kriterien erfüllen.

Accuracy

Die Genauigkeit misst die Nähe der abgebildeten Daten zu den Daten in der Realität.

Timeliness

Die Daten sollten bei einem Forecasting in die Zukunft so aktuell wie möglich gehalten werden, um eine genaue Vorhersage sicherzustellen.

Completeness

Die Vollständigkeit verlangt vom Datensatz, dass es keine Ausreißer und fehlenden Werte gibt.

Representativeness

Die vorhandenen Daten sollten für die Fragestellung geeignet sein, beispielsweise ist es nicht sinnvoll, die Daten der Hotline zu verwenden, um im allgemeinen Vorherzusagen, wann Gewalt gegen Frauen in Brasilien geschieht.

Consistency

Dieses Gütekriterium verlangt Konsistenz innerhalb der Zeitreihe, nicht nur inhaltlich sondern auch Formate und die Bedeutung von Kategorien und die Struktur des Datensatzes.

4.3.2 Data Analysis

Im Vorbereitungsschritt der Datenanalyse werden ist es wichtig die Zeitreihe auch visuell aufzubereiten. Dadurch können saisonale Muster oder Trends erkannt werden[5, 14f.]. Im Fall der vorliegenden Daten ergeben sich starke Differenzen in den Aufzeichnungen zwischen den Jahren 2014 und 2023.

[1] Übersetzung: *„Genauigkeit, Aktualität, Vollständigkeit, Repräsentativität und Konsistenz."*

Ein weiterer Teil der vorbereitenden Analyse sind die Berechnung der statistischen Kennzahlen, wie Sample Mean, Standardabweichung, Perzentile und Autokorrelationen.
Ebenfalls ist es sinnvoll potenzielle Ausreißer festzuhalten, um sie später auf ihre Passbarkeit zu prüfen. Diese Vorbereitungen sind wichtig um ein Gefühl für die Daten zu bekommen und die Ergebnisse besser einordnen zu können [5, 15].

4.3.3 Modellvergleich

Kapitel 5

Methodik

5.1 Defining Modelling Objective

5.2 Datenvorverarbeitung

In dem Leitfaden *Data Preparartion for Machine Learning* beschreibt [2] die Datenvorbereitung 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 [2, 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 [2, 9]. Dabei muss herausgefunden werden, wie die zugrunde liegende Struktur des Datensatz bestmöglich herausgearbeitet werden kann [2, 8]. Das ist abhängig davon, welcher Algorithmus verwendet werden soll, um das Modell zu erstellen [2, 12].

5.2.1 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 Models mitbestimmen herausgearbeitet [2, 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 [2, 13]. Auf der anderen Seite gibt es die Möglichkeit, sich für jede Datenzeile zu überlegen, wie diese Aussehen müsste, um bestmöglich deren Charakter herauszuarbeiten [2, 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 beitrügen [2, 13].

Entscheidend ist laut [2, 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 leakage[1] führen kann und somit das Modell beeinflussen kann [2, 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 [2] 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 [2, 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

5.3 Auswertungsmethoden

[1] Data 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 [8, 93].

Kapitel 6

Results

6.1 Structure Learning

6.2 Auswertungskriterien

Kapitel 7

Discussion

7.1 Preparation of Data

7.2 Structure Learning

7.3 Auswertungskriterien

Kapitel 8

Limitations

Kapitel 9

Conclusion

Bibliography

Literaturverzeichnis

[1] Atwan, T. A.: *TIME SERIES ANALYSIS WITH PYTHON COOKBOOK: Practical recipes for exploratory data analysis, data preparation, forecasting, and model evaluation.* PACKT PUBLISHING LIMITED https://learning.oreilly.com/library/view/-/9781801075541/?ar. – ISBN 978–1801075541

[2] 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

[3] 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

[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] 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

[8] 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