Bähren, Tobias; Maasjosthusmann, Robin; Walter, Annika; Lehrbass, Frank

Research Report

Praktische Umsetzung von Business Analytics im Mediensektor: Predictive Analytics im Filmgeschäft

ifes Schriftenreihe, No. 21

Provided in Cooperation with:
ifes Institut für Empirie & Statistik, FOM Hochschule für Oekonomie & Management

Suggested Citation: Bähren, Tobias; Maasjosthusmann, Robin; Walter, Annika; Lehrbass, Frank (2020) : Praktische Umsetzung von Business Analytics im Mediensektor: Predictive Analytics im Filmgeschäft, ifes Schriftenreihe, No. 21, ISBN 978-3-89275-418-3, MA Akademie Verlags- und Druck-Gesellschaft mbH, Essen

ifes Schriftenreihe

Band 21

Bianca Krol (Hrsg.)

Praktische Umsetzung von Business Analytics im Mediensektor: Predictive Analytics im Filmgeschäft

~

Tobias Bähren, Robin Maasjosthusmann,
Annika Walter, Frank Lehrbass

FOM Hochschule
ifes
Institut für Empirie & Statistik
der FOM Hochschule
für Oekonomie & Management

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

MA Akademie Verlags- und Druck-Gesellschaft mbH
Leimkugelstraße 6, 45141 Essen
info@mav-verlag.de

Tobias Bähren, Robin Maasjosthusmann, Annika Walter, Frank Lehrbass

Praktische Umsetzung von Business Analytics im Mediensektor:
Predictive Analytics im Filmgeschäft

ifes Institut für Empirie & Statistik
der FOM Hochschule für Oekonomie & Management

ifes Schriftenreihe
Band 21, 2020

ISBN (Print) 978-3-89275-417-6
ISBN (eBook) 978-3-89275-418-3
ISSN (Print) 2191-3366
ISSN (eBook) 2569-5355

Inhaltsverzeichnis

Abkürzungsverzeichnis

API	Application Programming Interface
CPIAUCNS	Consumer Price Index for all Urban Consumers: All Items in U.S. City Average
CRISP-DM	Cross-Industry Standard Process for Data Mining
EDA	Explorative Datenanalyse
IDs	Identifikatoren
ISO	Internationale Organisation für Normung
URL	Uniform Resource Locator

Abbildungsverzeichnis

Tabellenverzeichnis

Einleitung

In der vorliegenden Studie wird der gesamte Business Analytics-Prozess nach Seiter (2017) an einem fiktiven Fallbeispiel der Filmproduktionsfirma CINEMAKE[1] bearbeitet.

CINEMAKE sei eine Produktionsfirma von Filmen, welche in der Regel im Kino ausgestrahlt werden. Der Erfolg dieser Filme wird durch die Einspielergebnisse an den Kinokassen, dem sogenannten „Box Office", gemessen. Das Portfolio der Filmproduktionen umfasst verschiedene Genres der Filmbranche, wie beispielsweise Science-Fiction, Fantasy, Action, Drama oder Komödie. Bei den bisher produzierten Filmen war allerdings noch kein Kassenschlager dabei und die großen Erfolge blieben somit aus. Soweit also die Fiktion einer Firma, um den Rahmen für Business Analytics zu schaffen.

In der Literatur zum Thema Vorhersage der Einspielergebnisse von Kinofilmen zeigt sich, dass es keinen typischen erfolgreichen Film gibt, da die Einspielergebnisse von Filmen generell nicht zu einem Durchschnitt konvergieren (vgl. de Vany, Walls, 1999, S. 285). Der Erfolg von einem Film hängt von vielen Faktoren ab und ist schwer zu prognostizieren. Eine Studie von 2013, bei welcher 435 Filme aus den Jahren 1965 bis 2010 untersucht wurden, legt allerdings nahe, dass die Hauptdarsteller und der Regisseur am ehesten den Erfolg eines Films bestimmen. Außerdem beeinflussen nach dieser Studie ein hohes Filmbudget und eine längere Laufzeit des Films das Einspielergebnis positiv (vgl. Kim, 2013, S. 1071).

Da es in letzter Zeit zu vermehrten Umsatzeinbußen in der Filmindustrie kam, kommen sogar in Hollywood immer mehr große Produktionsfirmen in die Bedrängnis Erfolge zu erzielen (vgl. Ahmed et al., 2019, S. 1). Die Filmproduzenten sind aus diesem Grund auf der Suche nach einer neuen Formel zur Vorhersage des Erfolgs von Filmen vor deren Produktion, da die bisher verwendeten Systeme nur eine geringe Genauigkeit aufweisen (vgl. Ahmed et al., 2019, S. 1). Beispielsweise entschied sich eines der bekanntesten Filmstudios in Hollywood, Warner Brothers, zuletzt für den Einsatz von künstlicher Intelligenz, um den Auswahlprozess der zu produzierenden Filme anhand von Erfolgsprognosen zu unterstützen (vgl. Steinitz, 2020, o. S.). In einer ähnlichen Situation befindet sich

[1] CINEMAKE stellt eine fiktive Filmproduktionsfirma exklusiv für diese Anwendung des Business Analytics-Prozesses dar. Es besteht kein Bezug zu tatsächlich bestehenden Produktionsfirmen unter diesem Namen.

ebenfalls die Produktionsfirma CINEMAKE, auf welche wir in den folgenden Kapiteln näher eingehen werden.

Der Aufbau dieser Arbeit orientiert sich an den vier Teilprozessen des Business Analytics-Prozesses von Seiter. Die vier Teilprozesse umfassen die Bereiche *Framing*, *Allocation*, *Analytics* und *Preparation* (vgl. Seiter, 2017, S. 33).

Zunächst wird im folgenden Kapitel *Framing* die Problemstellung des Unternehmens CINEMAKE genauer erläutert und daraus ein Analytics-Problem abgeleitet, welches durch Algorithmen zu lösen ist (vgl. Seiter, 2017, S. 23).

Darauf aufbauend werden im dritten Kapitel *Allocation* die benötigten Ressourcen zur Problemlösung dargestellt. Die Ressourcen stammen in der Regel aus den Bereichen Daten, Informationstechnologie und Personal (vgl. Seiter, 2017, S. 26). Dieser Abschnitt beinhaltet unter anderem die Erstellung des Datenkorpus, welcher für die späteren Analysen verwendet wird.

Die nächste Phase des Business Analytics-Prozesses und das vierte Kapitel dieser Arbeit beinhaltet den Bereich *Analytics*. Dieser Abschnitt wird in der Regel in enger Abstimmung mit der vorhergehenden Phase bearbeitet. Diese beiden Phasen, *Allocation* und *Analytics*, können aus diesem Grund häufiger durchlaufen werden (vgl. Seiter, 2017, S. 32). *Analytics* umfasst den eigentlichen Teil der Datenanalyse und die damit zusammenhängende Datenaufbereitung und Evaluation der Ergebnisse. Das Ziel ist dabei, das vorliegende Problem mit gewonnenen Evidenzen zu lösen (vgl. Seiter, 2017, S. 29). CINEMAKE entschied sich, in diesem Teilprozess nach dem *Cross-Industry Standard Process for Data Mining*, CRISP-DM, vorzugehen. Dieser Standardprozess wurde von Analysten aus unterschiedlichen Bereichen entwickelt und bietet einen Überblick über den Lebenszyklus eines Data Mining Projektes (vgl. Wirth, Hipp, 2000, S. 32).

Preparation ist der letzte Teilprozess des Business Analytics-Prozesses, welcher die Aufbereitung der Ergebnisse für deren optimalen Einsatz umfasst. Hierzu zählen die verständliche Übermittlung der Ergebnisse zur Anwendung dieser, die Erklärung der zugrundeliegenden Mechanismen sowie die Darstellung der Grenzen der Anwendung (vgl. Seiter, 2017, S. 31-32).

Framing

Zu Beginn des Business Analytics-Prozesses wird die aktuelle Situation der Produktionsfirma CINEMAKE sowie das vorliegende Problem genauer beschrieben. Anschließend wird mittels Operationalisierung die unternehmerische Problemstellung weiter ausgearbeitet, so dass das Analytics-Problem abgeleitet werden kann (vgl. Seiter, 2017, S.23). Hierzu wird bereits eine Idee zur Lösung des Problems dargestellt.

Betriebswirtschaftliche Problemstellung

Wie bereits erwähnt, blieben große finanzielle Erfolge bei den Filmproduktionen von CINEMAKE bisher aus. Die Auswahl der zu produzierenden Filme beginnt in der Regel damit, dass Filmvorschläge eingereicht oder Filmideen zur Produktion angeboten werden. Diese Vorschläge beziehungsweise Angebote werden anschließend bislang ohne feste Kriterien durch einen kleinen Kreis aus drei Personen der Managementebene priorisiert und letztendlich ausgewählt. Insgesamt wurden hiermit mehr Flops als Erfolge produziert. Das heißt die Einspielergebnisse blieben meist unter den Erwartungen und die gesamten Kosten, welche neben dem Filmbudget auch unter anderem das Werbebudget enthalten, konnten nicht ausgeglichen werden. Aus diesem Grund ist die momentane wirtschaftliche Situation des Unternehmens CINEMAKE kritisch.

Das Management von CINEMAKE identifizierte somit die betriebswirtschaftliche Problemstellung als mangelnde Erfolgsquote der eigenen Filmproduktionen aufgrund fehlender Nachvollziehbarkeit sowie Qualität des Auswahlprozesses. Damit CINEMAKE in der Zukunft weiter bestehen kann, dürfen die nächsten Produktionen keine Flops sein, so dass keine weiteren Verluste auftreten. Dies bedeutet, dass die Umsätze der produzierten Filme zumindest die Produktions- und Werbekosten abdecken müssen.

Die Geschäftsleitung von CINEMAKE wünscht als grundsätzliche Lösungsidee, dass zukünftig der Entscheidungsprozess der Filmproduktionen mit wissenschaftlichen Methoden, wie statistische Analysen, unterstützt wird und somit bessere sowie fundiertere Entscheidungen getroffen werden können. Letztendlich sollen nur noch diejenigen Filme produziert werden, welche mit hoher Wahrscheinlichkeit einen Erfolg aufweisen werden.

Analytics-Problem

Zunächst gilt zu klären, anhand welcher Eigenschaften Filmproduktionen der Vergangenheit als erfolgreich und somit sich lohnend eingestuft werden konnten. Basierend darauf, können entsprechende Klassen von Filmen gebildet werden, welche die Filme in Erfolge und Flops teilen.

Folglich lässt sich das Analytics-Problem aus der betriebswirtschaftlichen Problemstellung ableiten: *Auf Basis welcher Faktoren kann prognostiziert werden, ob ein Film ein Erfolg wird oder nicht?*

Anhand der Beantwortung dieser Frage könnten somit in Zukunft vorliegende Produktionsvorschläge bewertet werden, um letztendlich diejenigen Produktionen von Filmen mit den entsprechenden erfolgsversprechenden Faktoren zu priorisieren.

Das vorliegende Analytics-Problem kann zum einen durch deskriptive, zum anderen durch prädiktive Analysen gelöst werden. Zur deskriptiven Analyse zählt eine ausführliche explorative Datenanalyse, kurz EDA, von Filmdaten, um ein generelles Verständnis des bisher unbekannten Kundeninteresses, das heißt des Interesses von Kinobesuchern, zu erhalten und um ggf. erste wichtige Erfolgsfaktoren zu identifizieren. Um den Erfolg beziehungsweise Misserfolg eines Films prognostizieren und hierfür relevante Faktoren bestimmen zu können, bedarf es Analysen durch Klassifizierungs- oder Regressionsmodelle.

Allocation

Nach erfolgreicher Definition des Analytics-Problems und erster Darstellung einer Lösungsidee, werden in der zweiten Phase des Business Analytics-Prozesses die zur Lösung des Problems benötigten Mittel genauer erläutert. Zunächst wird die Erstellung des Datenkorpus beschrieben, welcher für die Analyse verwendet wird. Die Ressourcen Informationstechnologie und Personal werden im Anschluss betrachtet.

Erstellen des Datenkorpus

Da sich die Geschäftsführung dafür ausgesprochen hat, die Auswahl der Produktionen zukünftig auf Basis von statistischen Analysen durchzuführen, hat das Data Science Team sich entschieden, verschiedene Datenquellen zu analysieren. Durch diesen Vorgang sollen geeignete Quellen erschlossen werden, die die Basis für anschließende Analysen darstellen.

Zuerst wurde hierbei auf die firmeninterne Historie zurückgegriffen, sogenannte Primärdaten. Diese bieten den Vorteil, dass alle Datenpunkte bekannt und zugänglich sind (vgl. Seiter, 2017, S. 27). Allerdings beschränkt sich die Anzahl der Elemente (produzierte Filme), aufgrund der kurzen Unternehmensgeschichte, auf eine niedrige zweistellige Zahl. Aufgrund dieses Faktors und dem Fehlen von erfolgreichen Filmen in der Firmengeschichte, bietet diese keine ausreichende Grundlage für eine umfassende Analyse.

Daher wurde entschieden, auf externe Quellen, Sekundärdaten, zurückzugreifen. Da das Unternehmen nicht Teil eines Unternehmensverbundes ist, fällt der Datenaustausch innerhalb dessen als mögliche Datenquelle weg. In der Filmindustrie gibt es allerdings die Besonderheit, dass viele der wichtigsten Datenpunkte von allen Produktionen öffentlich zugänglich sind. Einige hiervon sind vollständig bekannt und faktisch korrekt. Unter diese fallen zum Beispiel die beteiligten Schauspieler oder die Produktionscrew. Andere Daten sind nur ungefähr bekannt oder werden annähernd geschätzt. Hier können das eingesetzte Budget (zumeist abgeschätzt) oder das Einspielergebnis (grundlegend ermittelbar über die Ticketverkäufe, aber nicht abschließend genau bestimmbar) genannt werden.

Diese Daten werden an verschiedenen Stellen im Internet zur Verfügung gestellt. Unterschieden werden kann hierbei zwischen kostenlosen und kostenpflichtigen Angeboten. Zu den bekanntesten kostenlosen Angeboten zählen

Seiten wie *www.imdb.com* oder *www.themoviedb.org*. Bei den kostenpflichtigen Angeboten sind *www.boxofficemojo.com* oder die *Pro-Version* von *www.imdb.com* zu nennen.

Nach einer Stichprobenanalyse konnte nicht festgestellt werden, dass die kostenpflichtigen Angebote mehr oder genauere Werte zur Verfügung stellen. Bei abweichenden Angaben ließ sich nicht ermitteln, welche Angaben besser sind.

Daher wurde unter dem Gesichtspunkt der Kostenoptimierung entschieden, eine kostenlose Alternative zu verwenden. Hier sticht *www.themoviedb.org* durch das ebenfalls kostenlos zugängliche Application Programming Interface, API, heraus, siehe *https://developers.themoviedb.org*. Da die Ansprache des API am Thema dieses Dokuments vorbeigeht, wird hier nicht weiter darauf eingegangen. Es wird aber der Ablauf der Informationsanreicherung beschrieben.

Zunächst musste eine Auswahl getroffen werden, welche Filme in den Datensatz einfließen sollen. Um in der Analyse auf aktuelle Trends und Entwicklungen stoßen zu können, wurde entschieden auf die letzten *10* Jahre, *2009* bis *2019*, zurückzublicken. Hierbei sollen aus jedem Jahr alle gelisteten Filme verwendet werden.

Um spezifische Informationen zu Filmen abfragen zu können, müssen ihre datenbankinternen Identifikatoren, IDs, bekannt sein. Diese wurden über die Abfrage der *10000* erfolgreichsten Filme pro Jahre ermittelt. Die Anzahl *10000* stellt hierbei das Maximum einer Anfrage an das API dar. Nach einer ersten Analyse der erhaltenen Daten wurde festgestellt, dass ein Großteil der Filme kein Budget oder Einnahmen hinterlegt hat. Pro Jahr beschränkt sich die Anzahl der Filme mit hinterlegten Daten auf 400-500. Da Filme ohne diese Informationen keinen Mehrwert für das Analyse-Projekt liefern, wurde entschieden, die Informationsanreicherung nur für Filme mit diesen Informationen durchzuführen. Dies stellt den ersten Schritt der Datenbereinigung dar, da sich dafür entschieden wurde, die Datensätze mit fehlenden Werten zu entfernen anstatt sie mit Methoden wie der Mittelwertberechnung oder das Setzen auf einen festen Wert zu füllen (vgl. Larose, 2015, S. 93ff).

Nachdem nun eine Liste aller datenbankinternen IDs vorlag, konnten weitere Informationen abgefragt werden. Für die Feature-Liste wurden hierbei pro Film vier verschiedene Endpoints des API angesprochen. Der erste Endpoint liefert

primäre Informationen zum Film, der Zweite die auftretenden Schauspieler, der Dritte die Produktionscrew und der Vierte die Schlagwörter des Films.

Die so gesammelten Daten liegen in einer annähernd strukturierten Form in einer Datenmatrix vor. Die Zeilen entsprechen hierbei den einzelnen Filmproduktionen, die Spalten wiederum den Attributen der Filme beziehungsweise Features. Die einzelnen Features des Datensatzes werden im Abschnitt 4.2 genauer beleuchtet.

Informationstechnologie

Um die Problemstellung mit Hilfe von Daten lösen zu können, wird eine entsprechende IT-Ausstattung benötigt.

Da die Daten durch das Team der Data Scientisten über das Internet sowie durch die Ansprache des freizugänglichen API gesammelt werden und so für weitere Analysen verfügbar gemacht werden können, werden keine weiteren datenliefernden Komponenten, wie Sensordaten oder Daten aus operationalen Systemen, benötigt. Ergänzend ist anzumerken, dass diese Daten auch in Zukunft durch das Data Scientisten Team erstellt beziehungsweise gesammelt werden und somit auch dann keine zusätzlichen Komponenten notwendig sind.

Die Data Scientisten von CINEMAKE entschieden sich, die gewonnenen Daten in einem Analytics-Lab zur Verfügung zu stellen, welches dezentral im Unternehmen verortet ist. Hier werden die Daten ausschließlich mit dem Ziel der Problemlösung durch Business Analytics in entsprechend notwendiger Form bereitgestellt, das heißt für die folgenden Analysen des Teilprozesses *Analytics* (vgl. Seiter, 2017, S. 86). Außerdem wird das Versionsverwaltungsprogramm *Git* verwendet, um erstellten Programmcode zu verwalten und innerhalb des Analysten-Teams auszutauschen.

Zur Gewinnung von Evidenzen aus diesen Daten werden mehrere unterschiedliche Analytics-Plattformen von den Data Scientisten verwendet. Dies ermöglicht eine größere Auswahl der später anwendbaren Algorithmen (vgl. Seiter, 2017, S. 87). Hierzu zählen unter anderem die Analytics-Plattformen beziehungsweise Programmier/Skriptsprachen *Python*, *R* und *RapidMiner Studio*.

Personal

CINEMAKE verfügt, wie bereits erwähnt, über ein Team aus Data Scientisten, welches über umfangreiche Business Analytics-Kompetenzen verfügt und auf die nötigen Algorithmen spezialisiert ist. Dieses Team besteht aus drei Personen und einem wissenschaftlichen Berater (stundenweise zugeschaltet), die unterschiedliche Rollen übernehmen. Hierzu zählen die Datenbeschaffung und -aufbereitung zur Sammlung der Daten und Gewährleistung einer notwendigen Datenqualität, die Analyse der Daten durch die Auswahl geeigneter Algorithmen sowie die Interpretation und Visualisierung der gewonnenen Evidenzen, so dass die Ergebnisse in den betrieblichen Prozessen entsprechend genutzt werden können.

Als Fach-Experten steht das Management von CINEMAKE zur Verfügung, welches von den aktuellen Problemen direkt betroffen ist (vgl. Seiter, 2017, S. 28). Diese können die Data Scientisten bei betriebswirtschaftlichen Fragestellungen unterstützen.

Das Personal der IT-Abteilung des Unternehmens stellt den Data Scientisten die notwendigen Komponenten der IT-Architektur über den gesamten Prozess bereit.

Analytics: Vorgehen nach CRISP-DM

In diesem Abschnitt der eigentlichen Datenanalyse zur Lösung des Analytics-Problems wird nach dem branchenunabhängigen Standardprozess CRISP-DM vorgegangen. Dieser Prozess besteht aus sechs Phasen, welche ein Data Mining Projekt beschreiben. Hierzu zählen die Phasen *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation* und *Deployment*.

Wie von Wirth und Hipp beschrieben, ist der Ablauf beim Durchlaufen des CRISP-DM nicht unbedingt linear, sondern es kann zwischen den unterschiedlichen Phasen vor- und zurückgesprungen werden (vgl. Wirth, Hipp, 2000, S. 32). Im Rahmen dieses Projektes ist dies vor allem zwischen den Phasen *Data Understanding* und *Data Preparation* aufgetreten. Um in der Lage zu sein, alle Features in der Phase *Data Preparation* zu verarbeiten, musste zunächst ein grundlegendes Verständnis dieser erlangt werden. Nachdem dieses vorhanden war, konnten diese Features aufbereitet werden, um anschließend ein zweites Mal betrachtet zu werden und somit ein tieferes Verständnis zu erlangen.

Business Understanding Phase

In der ersten Phase des CRISP-DM sind zunächst die genauen Ziele des Projektes zu definieren (vgl. Larose, 2015, S. 72).

Da im Kapitel Framing bereits die betriebswirtschaftliche Problemstellung sowie das Analytics-Problem erarbeitet wurden, können die Ziele in diesem Abschnitt davon abgeleitet werden. Das Analytics-Problem besteht darin, dass die Faktoren, welche für den Erfolg eines Films verantwortlich sind, CINEMAKE bisher unbekannt sind und somit die produzierten Filme eine niedrige Erfolgsquote aufweisen.

Davon lässt sich das primäre Ziel des Projektes wie folgt ableiten: *Entwicklung eines Klassifizierungsmodells, das den Erfolg einer Filmproduktion prognostiziert und somit die Gewinnmaximierung unterstützt.*

Hierbei ist das sekundäre Ziel, ein besseres Kundenverständnis, das heißt der Kinobesucher, beziehungsweise ein besseres Verständnis der Erfolgsfaktoren durch eine umfangreiche EDA zu erhalten.

Somit handelt es sich bei den vorliegenden Zielen um ein Klassifizierungsproblem. Es gilt herauszufinden, welche zukünftigen Filme erfolgreich sein werden, basierend auf den zur Verfügung stehenden Daten.

Data Understanding Phase: Überblick über die Daten

In diesem Abschnitt soll ein erster Blick auf die Features des Datensatzes geworfen werden. Hierbei werden die vorhandenen Features erläutert sowie erste Analysen der Datentypen durchgeführt.

Der Datensatz hat zu Beginn der Datenaufbereitung 27 Features. In Tabelle 1 sind die Features aufgelistet und genauer erläutert. Als Basis für die Zielvariable der Analysen wurde das Feature *revenue* ausgewählt, welches das Einspielergebnis eines Films und somit den finanziellen Erfolg widerspiegelt.

Feature Name	Type	Beschreibung	Behalten[2]
adult	Flag	Gibt an, ob der Film ein R-Rating (Altersbeschränkung ab 17) hat	Ja
back-drop_path	Text	Pfad zu Hintergrundbild des Films auf *www.themoviedb.org*	Nein
budget	Number	Budget des Films in US-Dollar	Ja
genres	Object	Array gefüllt mit Objekten: Jedes Objekt stellt ein Genre dar, welches dem Film zugeordnet ist. Ein Objekt hat dabei die Eigenschaften *ID* und *Name*	Ja
homepage	Text	Uniform Resource Locator, URL, der Homepage des Films: Leer, wenn keine Homepage bekannt	Ja
movie_id	Object	ID des Films innerhalb der *www.themoviedb.org* Datenbank	Nein
imdb_id	Object	ID des Films innerhalb der *www.imdb.com* Datenbank (Querverweis).	Nein
original_language	Text	(Primäre) Sprache, in der der Film gedreht wurde	Ja

[2] Auf die ausgeschlossenen Features wird im Anschluss an die Tabelle genauer eingegangen

Feature Name	Type	Beschreibung	Behalten[2]
original_title	Text	Originaltitel des Films	Nein
overview	Text	Kurzbeschreibung des Films, entspricht Marketing-Text durch das Studio	Ja
popularity	Number	Gibt die Beliebtheit des Films anhand verschiedener Aktionen der User zu diesem Film auf *www.themoviedb.org* an	Nein
poster_path	Text	Pfad zum verwendeten Filmposter auf *www.themoviedb.org*	Nein
companies	Object	Array gefüllt mit Objekten: Jedes Objekt stellt eine Produktionsfirma, die am Film beteiligt war, dar. Ein Objekt hat dabei die Eigenschaften *name*, *ID*, *logo_path* und *origin_country*	Nein
countries	Object	Array gefüllt mit Objekten: Jedes Objekt stellt ein Produktionsland des Films dar. Ein Objekt hat dabei die Eigenschaften *iso_3166_1* (ISO-Ländercodeliste) und *name*	Ja
release_date	Date	Datum, an dem der Film veröffentlicht wurde (MM/DD/YYYY)	Ja
revenue	Number	Einnahmen des Films in US-Dollar	Ja
runtime	Number	Laufzeit des Films in Minuten	Ja
languages	Object	Array gefüllt mit Objekten: Jedes Objekt stellt eine Sprache dar, die im Film gesprochen wird. Jedes Objekt hat die Eigenschaften *iso_639_1* (ISO-Sprachencodes) und *name*	Ja

Feature Name	Type	Beschreibung	Behalten[2]
status	Text	Produktionsstatus des Films (z.B. *Rumored*, *in Production*, *Released*)	Nein
tagline	Text	Tagline des Films	Nein
title	Text	Englischer Titel des Films	Ja
video	Text	URL zu einem Trailer des Films	Nein
vote_average	Number	Durchschnittliche Benutzerbewertung auf *www.themoviedb.org*	Nein
vote_count	Number	Anzahl der abgegebenen Benutzerbewertungen	Nein
cast_list	Object	Array gefüllt mit Objekten: Jedes Objekt stellt einen Schauspieler dar, der an dem Film mitgewirkt hat. Jedes Objekt hat dabei die Eigenschaften *cast_id*, *character*, *credit_id*, *gender*, *id*, *name*, *order* und *profile_path*	Ja
crew_list	Object	Array gefüllt mit Objekten: Jedes Objekt stellt ein Produktionsteammitglied dar, das an dem Film mitgewirkt hat. Jedes Objekt hat dabei die Eigenschaften *credit_id, department, gender, id, job, name* und *profile_path*	Ja
keywords_list	Object	Array gefüllt mit Objekten: Jedes Objekt stellt ein Keyword dar, welches die Benutzer von *www.themoviedb.org* dem Film zugeordnet haben. Jedes Objekt hat dabei die Eigenschaften *id* und *name*	Ja

Tabelle 1: Überblick der ursprünglichen Features

Grundsätzlich lassen sich die ausgeschlossenen Features in vier Gruppen unterteilen.

Zunächst gibt es mit *status* ein obsoletes Feature. Da die abgefragten Filme alle einen bekannten Umsatz haben, wurden sie zwangsläufig veröffentlicht, haben also den Wert *released*. Damit haben alle Filme denselben Wert in *status*, wodurch dieses Feature keinen Mehrwert bei der Analyse bietet.

Als Zweites lassen sich die internen Features der Webseite *www.themoviedb.org* zusammenfassen. Diese sind *backdrop_path*, *movie_id*, *imdb_id*, *poster_path* und *video.* Alle geben Informationen zu Webseiten bezogenen Werten an und können daher nicht direkt mit dem Filmerfolg in Bezug gebracht werden. So kann es zum Beispiel sein, dass zwar kein Pfad unter *video* hinterlegt ist, aber der Film trotzdem einen Trailer hat.

Die dritte Gruppe umfasst alle Features, die erst nach Veröffentlichung des Films vorliegen und somit nicht für das Produktionsstudio bei der Auswahl von Filmprojekten verfügbar sind. Hierzu zählen *popularity*, *vote_average* und *vote_count*. Außerdem würden einige dieser Features indirekten Rückschluss auf das Einspielergebnis des Films geben. So kann davon ausgegangen werden, dass die Anzahl in *vote_count* mit der Anzahl der Personen, die den Film gesehen haben, korreliert. Die absolute Anzahl von Kinobesuchern korreliert wiederum stark mit dem Einspielergebnis des Films.

Zuletzt gibt es mit *original_title* und *tagline* zwei Features mit Text als Inhalt. *original_title* überschneidet sich mit dem Feature *title*, welches den englischen Titel des Films enthält. Da der englische Titel für das Analysten-Team leichter zu analysieren ist, wurde entschieden diesen zu verwenden. Die Tagline ist frei wählbar und lässt häufig ohne weitere Informationen nicht auf den Film oder seinen Inhalt schließen. So ist beispielsweise die Tagline „Every generation has a story" nicht unbedingt *Star Wars: The Force Awakens* zuzuordnen.

Data Preparation Phase

In diesem Abschnitt sollen die verbleibenden Features genauer betrachtet werden, um tiefere Erkenntnisse über ihren Inhalt und ihre Bedeutung im Rahmen der Klassifizierung zu erhalten. Außerdem soll der Datensatz für die anstehende Analyse weiter aufbereitet werden, indem weitere Features erzeugt oder beste-

hende Features entfernt werden. Für die EDA wird die Programmiersprache *Python* mit einigen Bibliotheken verwendet. Eine Übersicht der verwendeten Programme und Bibliotheken ist im Abschnitt 6.2 zu finden.

Nach dem Entfernen der im vorherigen Abschnitt genannten Features, sind noch 15 Features vorhanden. Von diesen 15 Features enthalten *sechs* Objekte mit weiteren Listen, *vier* Text, *drei* Nummern (zwei hiervon sind Dollarwerte), *ein* Feature Datumsangaben und *eins* stellt eine Flag (Boolean-Werte) dar.

Behandlung der fehlenden Werte

Zunächst soll geprüft werden, ob es fehlende Daten im Datensatz gibt. Betrachtet man Abbildung 1, ist zu erkennen, dass *vier* Features leere Einträge enthalten. Diese Features sind *homepage*, *overview runtime* und *tagline*. Hierbei wird *tagline* nicht weiter beachtet, weil dieses Feature, wie in Abschnitt *4.2* beschrieben, nicht in den Datensatz aufgenommen wird.

Betrachten wir das Feature *homepage*, ist zu erkennen, dass ein Film entweder eine URL hinterlegt hat oder das Feld leer gelassen wurde, wenn keine Homepage bekannt ist. Nach einer stichprobenartigen Untersuchung der URLs wurde entschieden, dass sie keine relevanten Informationen für die Analyse enthalten. Daher wird das Feature *homepage* in ein Flag-Feature umgewandelt, um darzustellen, ob eine Homepage existiert oder nicht. Hierbei werden die fehlenden Werte durch Nullen ersetzt.

Bei *overview* sind mit *17* nur relativ wenig Samples leer, daher könnten diese grundsätzlich aus dem Datensatz entfernt werden. Da die Anzahl der Samples allerdings im Allgemeinen schon relativ niedrig ist, soll dies vermieden werden. Da das Feature *overview* aus relativ langen Texten besteht, wurde entschieden die häufigsten Wörter zu extrahieren und diese mittels *one-hot-encoding* als eigene Features aufzunehmen. Hierbei können Samples ohne Werte in *overview* weiterverwendet werden, so dass keine Aktion zum Entfernen dieser notwendig ist.

Das Feature *runtime* wird, auf Basis der vorhandenen Business Intelligence, als wichtiger Faktor betrachtet. Daher sollen die fehlenden Werte möglichst gefüllt werden. Nach manueller Suche konnte die Laufzeit zu 22 Filmen gefunden werden. Die verbleibenden zwei Samples werden aus dem Datensatz entfernt.

```
[19]: nullseries = df.isnull().sum()
      nullseries[nullseries > 0]

[19]: homepage    1155
      overview      17
      runtime       24
      tagline      625
      dtype: int64
```

Abbildung 1: Features mit leeren Werten

Veröffentlichungsdatum

Das Feature *release_date* soll in der *Data Preparation* Phase als Grundlage für verschiedene Schritte verwendet werden, daher wird dieses als erstes verarbeitet. Zunächst ist das ursprüngliche Format *MM/DD/YYYY* nicht für alle Python-Bibliotheken, die verwendet werden sollen, geeignet. Daher wird dieses in das Format *YYYY-MM-DD* umgewandelt.

Anschließend werden sechs neue Features erstellt, die das Veröffentlichungsjahr (*release_date_year*), den -wochentag (*release_date_weekday*), den -monat (*release_date_month*), die -kalenderwoche (*release_date_weekofyear*), den -tag (*release_date_day*) und das -quartal (release_date_quarter) darstellen.

Number Features

Nun sollen die Features betrachtet werden, die Zahlenwerte enthalten. Hierbei handelt es sich um *revenue, budget* und *runtime*. Die beiden zuerst genannten stellen dabei Dollarwerte dar.

Dollar Features

In diesem Abschnitt sollen die beiden Features *budget und revenue*, als die einzigen Features mit Dollarwerten, genauer betrachtet werden. r*evenue* stellt hierbei die Grundlage für die spätere Zielvariable der Analysen dar.

Da die Filme über einen Zeitraum von zehn Jahren veröffentlicht wurden, soll zunächst betrachtet werden, welchen Einfluss die Inflation in diesem Zeitraum hatte. Hierfür wurde auf den *Consumer Price Index for all Urban Consumers: All*

Items in U.S. City Average, *CPIAUCNS*, zurückgegriffen, um die Inflationsraten für die Monate der Veröffentlichung des jeweiligen Films zu berechnen (vgl. U.S. Bureau of Labor Statistics, 2019, o. S.). Der *CPIAUCNS* enthält den Preis für einen fest definierten Warenkorb für jeden Monat, so dass man die Preisentwicklung über die Jahre nachvollziehen kann. Im Rahmen dieser Analyse wurde der Wert des Monats Oktober 2019 als 1,0 definiert und somit berechnet, mit welchem Faktor die Dollarwerte der älteren Filme multipliziert werden müssen, um die Inflation auszugleichen.

Für die Multiplikation wird das Feature *release_date* betrachtet, um den passenden Eintrag aus der *CPIAUCNS*-Datenbank auszuwählen. Betrachtet man zum Beispiel den Film *Avatar,* der im April 2009 veröffentlicht wurde, müssen die Dollar Features *budget* und *revenue* zum Inflationsausgleich mit dem Faktor *1,2068* multipliziert werden. Für *budget* wird dieselbe Inflationsrate wie für *revenue* verwendet, da sich der Zeitraum der Ausgaben nicht genau bestimmen lässt.

Abbildung 2: Dollar Features mit und ohne Inflationsbereinigung

In Abbildung 2 sind jeweils die Verläufe der Einnahmen (*revenue*) und des Budgets (*budget*) aller Jahre des betrachteten Zeitraums mit und ohne Inflationsbereinigung dargestellt. Es ist zu erkennen, dass der Unterschied vor allem in

den ersten Jahren deutlich ist. Für das Jahr 2009 ergibt sich zum Beispiel ein Unterschied von *4.58* Milliarden Dollar. Bei 249 Filmen aus diesem Jahr bedeutet dies, dass jeder dieser Filme im Jahr 2019 durchschnittlich *18.3* Millionen Dollar mehr eingespielt hätte. Um diesen Fakt bei dem Modell zu berücksichtigen, wurde entschieden, von nun an die inflationsbereinigten Dollarwerte zu verwenden.

Betrachtet man die Verteilung der Dollar-Features, ist erkennbar das beide stark rechtsschief sind (siehe Abbildung 3 links). Um eine symmetrischere Verteilung zu erhalten, wurde entschieden, die Werte zu logarithmieren. Hierfür wurde die *Numpy* Funktion *log1p* verwendet, die den natürlichen Algorithmus von 1 plus den Wert des übergebenen Elements zurückgibt (vgl. The SciPy community, 2019, o. S.). Die hierdurch entstehende Verteilung ist zwar immer noch nicht symmetrisch (nun linksschief), aber deutlich näher hieran als zuvor (siehe Abbildung 3 rechts). Daher wurde entschieden, von nun an mit den logarithmierten Werten der Dollar-Features weiterzuarbeiten.

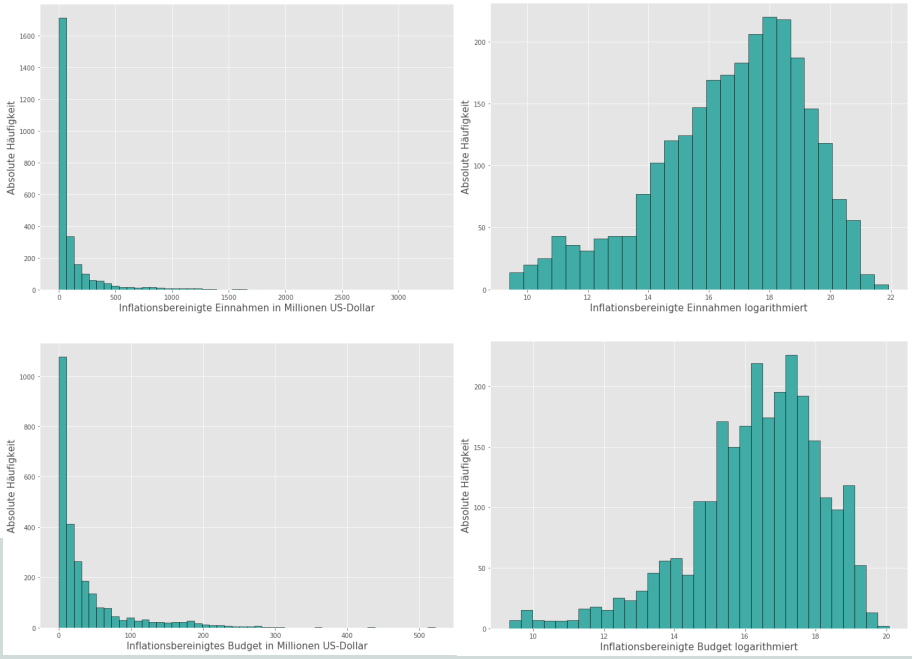

Abbildung 3: Verteilung der Dollar Features (Dollar vs. logarithmiert)

Runtime

Die Laufzeit der Filme wird in Minuten angegeben. Wie in Abschnitt *4.3.1 Behandlung der fehlenden Werte* beschrieben, konnten fast alle fehlenden Werte ergänzt werden. Betrachtet man den Boxplot in Abbildung 4, ist ersichtlich, dass es sowohl am unteren als auch am oberen Ende Ausreißer gibt.

Besonders extrem ist hier der Ausreißer mit *338* Minuten Laufzeit. Dieser ist um *131* Minute länger, als der nächste Ausreißer und liegt somit mehr als die durchschnittliche Länge aller Filme über diesem. Der Name des Films lautet *Carlos* und die Laufzeit konnte (in der Langfassung) bestätigt werden, wobei andere Quellen hier zwischen *331* und *338* Minuten schwanken.

Die Ausreißer am unteren Ende sind die Filme *The Vegetable* (*13*); *La Donna* (*14*); *Run, Hide, Fight* (*15*); *Barbet: L'Homme de la situation* (*20*) und *Blood Quantum* (*22*). Jeder dieser Filme ist ein Kurzfilm, womit die Kürze der Laufzeit erklärbar ist.

Da alle Ausreißer plausibel sind, wurden sie im Datensatz belassen und keine weitere Verarbeitung am Feature *runtime* durchgeführt.

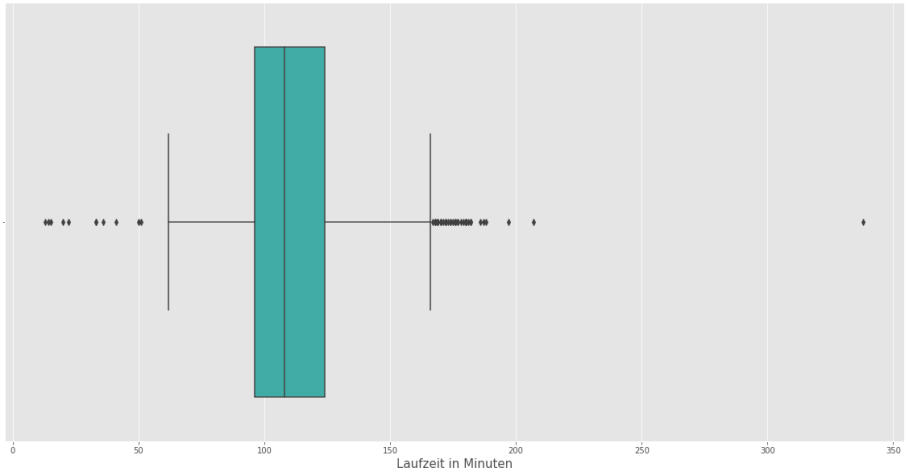

Abbildung 4: Boxplot Laufzeit der Filme

Listen Features

Da innerhalb einiger Features Objekte mit weiteren Eigenschaften enthalten sind, sollen diese im nächsten Schritt genauer betrachtet werden. Die zu betrachtenden Features sind hierbei *genres*, *countries*, *languages*, *cast_list*, *crew_list* und *keywords_list*.

Genres

genres enthält für jeden Film bis zu n-Einträge. Um ein Gefühl über die Verteilung der Genres-Anzahl zu erhalten, wurde eine Übersicht erstellt, wie viele Filme jeweils wie viele Genres zugeordnet haben. Dies ist in Tabelle 2 dargestellt.

Genres	2	3	1	4	5	6	0	7
Filme	874	845	494	309	80	16	15	1

Tabelle 2: Anzahl Filme mit n-Genres

Es ist klar zu erkennen, dass die meisten Filme bis zu vier Genres haben. Ausreißer mit mehr als vier Genres sind eher selten. Wie zu erkennen ist, gibt es 15 Samples ohne Genre. Da der Listenkörper für eine leere Liste in der jeweiligen Zelle allerdings vorhanden ist, sind diese nicht in der zuvor durchgeführten Prüfung auf leere Werte aufgetaucht.

Für ein besseres Verständnis der Filmdaten werden die häufigsten Genres ermittelt. Hierfür wurde die absolute Häufigkeit jedes Genres festgestellt, welche in Abbildung 5 mittels Wordcloud dargestellt werden.

Um die Liste im Feature *genres* aufzulösen wurde sich dazu entschieden auf *genres one-hot-encoding* anzuwenden, wobei nur die 15 häufigsten Genres als eigenes Feature übernommen werden sollen. Die neuen Feature-Bezeichnungen enthalten hierbei weiterhin die Namen der Genres, um eine einfache Interpretierbarkeit zu ermöglichen (z.B. genre_comedy). Die 15 häufigsten Genres lauten: *Drama* (1233), *Comedy* (912), *Action* (731), *Thriller* (677), *Romance* (434), *Adventure* (432), *Crime* (347), *Science Fiction* (273), *Family* (262), *Horror* (259), *Fantasy* (252), *Mystery* (197), *Animation* (168), *History* (116) und *War* (81).

Abbildung 5: Wordcloud Genres

Countries und languages

countries und *languages* wurden zusammen betrachtet, da erwartet wurde, dass diese Features ähnliche Informationen liefern. Filme aus demselben Land werden zumeist in den gleichen Sprachen gedreht. Da die Sprachen die beteiligten Länder widerspiegeln und zusätzlich in Ländern mit mehreren Sprachen diese getrennt darstellen können, wurde entschieden weiter mit dem Feature *languages* zu arbeiten und *countries* zu entfernen.

Sprachen	1	2	3	4	5	6	8	9	10
Filme	1976	407	148	62	25	10	3	2	1

Tabelle 3: Anzahl Filme mit n-Sprachen

Zunächst wurde ermittelt, wie häufig n-Sprachen in Filmen vorkommen. Das Ergebnis ist in Tabelle 3 dargestellt. Es ist klar zu erkennen, dass ein Großteil der Filme nur eine Sprache enthält, mehr als 2 Sprachen sind als Ausnahme anzusehen. Anzumerken ist, dass hierbei zunächst für 17 Filme null Sprachen hinterlegt waren. Auch dies ist auf leere Listen in dem Feature zurückzuführen, die verhindert haben, dass dies bei der ursprünglichen Prüfung auf leere Werte aufgefallen ist. Für die 17 Filme wurde manuell nach den gesprochenen Sprachen gesucht und für alle wurden diese gefunden und ergänzt. Das obere Ende, neun oder zehn Sprachen, wurde ebenfalls geprüft. Die Titel der Filme lauten *2012*, *Pina* und *Son of Saul*. Für alle konnte die angegebene Zahl verifiziert (*Pina* und *Son of Saul*) oder zumindest als realistisch (*2012*, viele Szenen mit unterschiedlichen Sprachen im Hintergrund) eingestuft werden.

Um zu erkennen, ob gewisse Sprachen einen großen Einfluss auf den Erfolg des Films haben, wurde außerdem ermittelt, welche Sprache in wie vielen Filmen vorkommt. Wenig überraschend ist hier *Englisch* mit *1982* Filmen die häufigste Sprache. Der Abstand zur zweithäufigsten Sprache (*Spanisch*, *199* Filme) ist allerdings überraschend groß.

Da sowohl die Anzahl der vorhandenen Sprachen als auch die Häufigkeit der Sprachen, so eindeutige Schwerpunkte haben, wurde dagegen entschieden, diese als Features in den Datensatz aufzunehmen. Ein weiterer Grund hierfür ist, dass mit dem Feature *original_language* ein weiteres Feature Bezug auf die gesprochene Sprache hat. Hierbei enthält *original_language* zu 98,33 Prozent eine der Sprachen aus *languages*, so dass hier eine große Korrelation vorliegen würde.

Cast

Da die „Starpower" von Filmen häufig mit Erfolg in Verbindung gebracht wird (vgl. Kim, 2013, S. 1071), wird das Feature *cast_list* von den Fachexperten als sehr wichtig für die Analyse angesehen. Im Datensatz sind *44497* verschiedene Schauspieler vertreten. Die Geschlechterverteilung der gespielten Rollen beträgt *43,46* Prozent männlich, *24* Prozent weiblich und *32,54* Prozent unbekannt. Betrachtet man dies im Kontext der weltweiten Geschlechterverteilung, wo im betrachteten Zeitraum auf 100 Männer 101,7 Frauen kommen, ist die deutliche Überhand der männlichen Rollen verwunderlich (vgl. United Nations, 2019, Spalten T,U,V). Aufgrund der großen Anzahl von unbestimmten Geschlechtern

wurde allerdings entschieden, das Geschlecht der Rolle nicht als Feature aufzunehmen.

Als nächstes wurde die Anzahl der Schauspieler pro Film betrachtet. In Abbildung 6 sind die *50* häufigsten Besatzungsgrößen der Filme dargestellt. Es ist zu erkennen, dass der Schwerpunkt bei *10* bis *20* Besatzungsmitgliedern liegt, es allerdings auch deutlich größere oder kleinere Besatzungen gibt. Um diesen Faktor mit in die Analyse aufnehmen zu können, wurde das neue Feature *num_cast* erstellt, welches die Anzahl an gelisteten Schauspielern des Filmes enthält.

Da die Anzahl der beteiligten Schauspieler zu groß ist, um alle mittels *one-hot-encoding* als eigene Features aufzunehmen, wurde entschieden nur eine Auswahl aufzunehmen. Hierbei wurde zugunsten eines zweiteiligen Prozesses entschieden. Zunächst wurde darauf geprüft, in wie vielen Filmen ein Schauspieler beteiligt ist. Dies basiert auf der Annahme, dass Schauspieler mit wenig Filmen keinen großen Einfluss auf die Vorhersage haben. In Abbildung 7 ist dargestellt, wie viele Schauspieler in jeweils n-Filmen beteiligt waren. Die extreme Mehrheit ist nur bei einem Film vertreten. Es wurde entschieden, nur Schauspieler mit mehr als fünf Auftritten weiter zu betrachten. Das entsprechende Diagramm ist als Abbildung 28 im Anhang zu finden.

Als nächster Schritt wurde der Einfluss eines Schauspielers auf die Einnahmen der Filme betrachtet. Als Einnahmen der Filme wurden hierbei die logarithmierten und inflationsbereinigten Werte des Features *revenue* verwendet (siehe Abschnitt *4.3.3.1 Dollar Features*). Für jeden Schauspieler wurde der Median der Einnahmen der Filme mit und ohne ihn berechnet. Ein Schauspieler wurde als Feature in den Datensatz mit aufgenommen, wenn der Median in den Filmen, in denen er beteiligt war, um 15 Prozent höher oder niedriger als in den Filmen ohne ihn war. Hierdurch wurden insgesamt *156* schauspielerspezifische Features erstellt (z.B. cast_name_chris_evans).

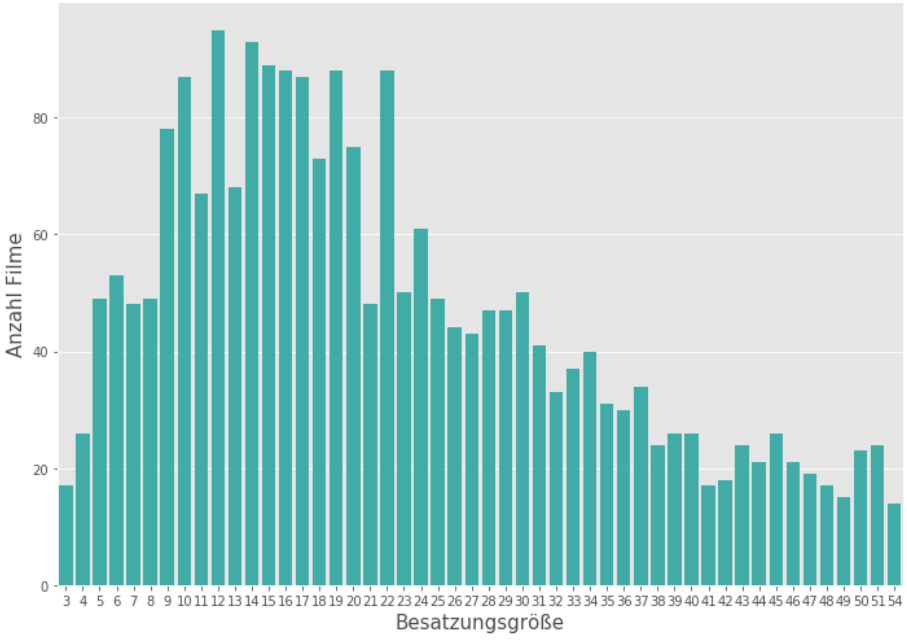

Abbildung 6: Verteilung der Besatzungsgröße der Filme

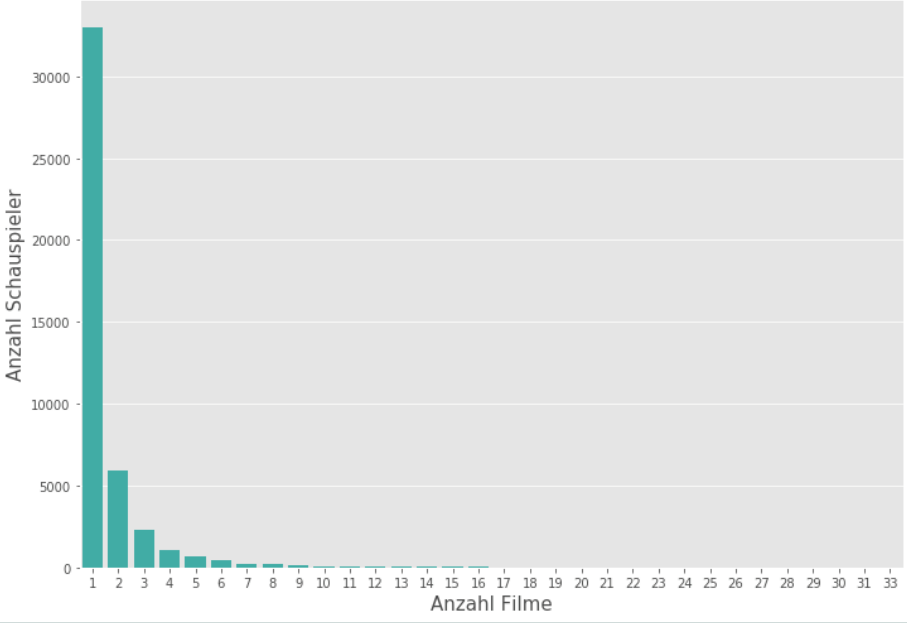

Abbildung 7: Anzahl Schauspieler mit n-Filmen

Crew

Das Vorgehen für das Feature *crew* ist in einigen Teilen identisch zu dem des Features *cast*. Zunächst wurde auch hier erst dargestellt, wie viele Filme ein Produktionsteam von der Größe n haben. In Abbildung 8: Größe des Produktionsteams pro Film ist hierbei zu erkennen, dass es keinen so klaren Schwerpunkt gibt wie bei den Schauspielern, aber ein Großteil der Filme unter 25 Personen gelistet hat. Dies und die hohe Anzahl an Filmen, die nur einstellige Anzahlen an Personen haben, lässt darauf schließen, dass bei vielen Filmen nicht alle Personen mit aufgelistet wurden. Hierdurch kann es zwar der Fall sein, dass entscheidende Personen nicht gelistet wurden und daher in der Analyse als wichtiges Feature fehlen; es wird aber davon ausgegangen, dass einflussreiche Personen eher hier gelistet werden, als weniger einflussreiche. Daher wird davon ausgegangen, dass Personen, die im originalen Datensatz nicht vorhanden sind, auch nicht durch das Auswahlverfahren ausgewählt worden wären, um in den Datensatz aufgenommen zu werden. Die Größe des Produktionsteams wird als Feature *num_crew* mit in den Datensatz aufgenommen.

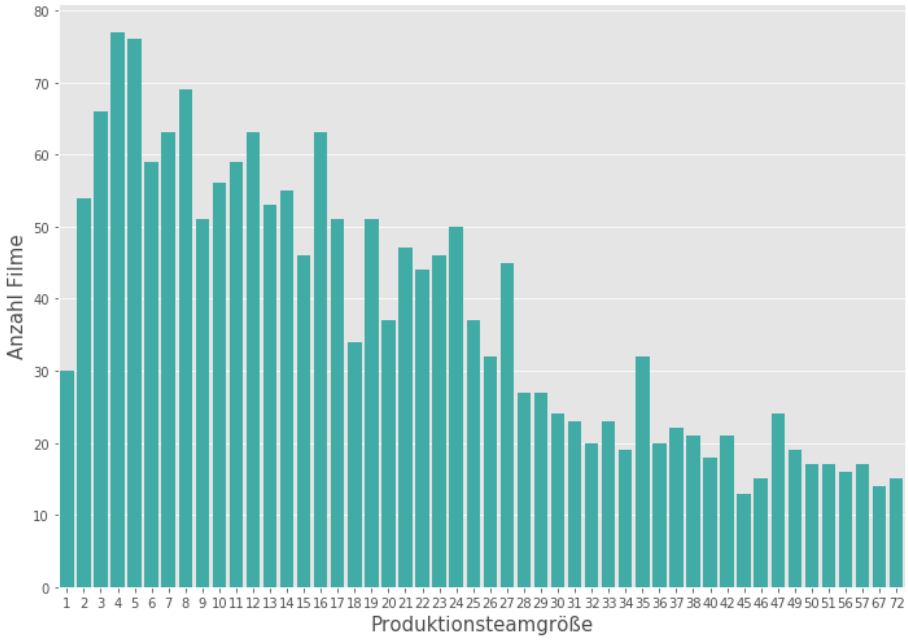

Abbildung 8: Größe des Produktionsteams pro Film

Bei den Produktionsteams sind *63322* verschiedene Personen aufgeführt. Da auch diese Zahl zu hoch ist, um *one-hot-encoding* auf sie anzuwenden, wurde das gleiche Vorgehen wie bei den Schauspielern gewählt. In Abbildung 9 ist zu erkennen, dass hierbei noch mehr Personen nur in wenigen Filmen beteiligt sind. Daher wurde die Grenze hier bei mehr als *10* Beteiligungen gesetzt. Beim Filtern anhand des Medians bei den Einnahmen wurde der Grenzwert auf 17,5 Prozent höheren oder niedrigeren Median gesetzt. Am Ende des Filterprozesses wurden *83* neue Features erstellt (z.B. crew_name_david_farmer).

Das Feature *crew* enthält weitere interessante Eigenschaften, so sind noch die ausgeführten Berufe (*job*) und die Abteilungen (d*epartment)* der jeweiligen Personen aufgeführt. Es werden insgesamt *760* Berufe gelistet. Es wurde entschieden, die *15* Häufigsten mittels *one-hot-encoding* als Features aufzunehmen. Die *15* häufigsten Berufe lauten: *Producer* (6761), *Executive Producer* (5000), *Animation* (3751), *Director* (2827), *Editor* (2684), *Casting* (2628), *Screenplay* (2483), *Art Direction* (2474), *Visual Effects Supervisor* (2138), *Director of Photography* (2058), *Product Design* (1884), *Original Music Composer* (1863), *Writer* (1821), *Makeup Artist* (1795) und *Costume Design* (1760).

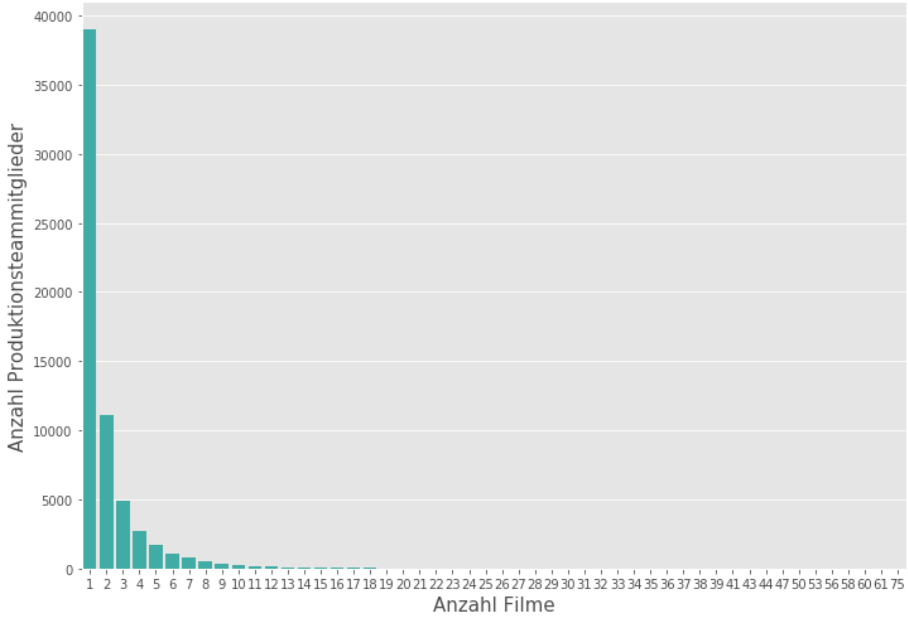

Abbildung 9: Anzahl der Produktionsteammitglieder mit n-Filmen

Bei den Abteilungen werden *12* Verschiedene gelistet, wobei die Abteilung *Actor* mit dem Stunt Double Giorgio Antoni nur ein Mitglied hat. Daher wird diese Abteilung nicht weiter betrachtet. Für die anderen Abteilungen (*Art*, *Camera*, *Costume & Makeup*, *Crew*, *Directing*, *Editing*, *Lighting*, *Production*, *Sound*, *Visual Effects* und *Writing*) wurde jeweils ein Feature angelegt, das die Anzahl an Personen, die an dem Film in der jeweiligen Abteilung gearbeitet haben, enthält. (z.B. *department_num_art*).

Keywords

Das letzte Feature, das eine Liste enthält, ist *keywords*. Dieses stellt insoweit eine Besonderheit dar, dass dieses manuell durch die Nutzer von *www.themoviedb.org* vergeben wurde. Daher liegen diese nicht automatisch bei dem Bewerten von vorliegenden Film-Ideen vor. Es wird allerdings davon ausgegangen, dass diese während des Bewertungsprozesses, firmenintern anhand des vorliegenden Drehbuchs, erstellt werden können.

In Abbildung 10 ist die Verteilung der Anzahl von Keywords dargestellt. Es ist zu sehen, dass viele Filme (*318*) keine Keywords haben. Dies ist nicht ideal für die Analyse, aber lässt sich durch die Notwendigkeit der manuellen Erstellung durch die Nutzer erklären. Mit Blick auf den hierfür notwendigen Zeitaufwand, wurde dagegen entschieden, die Keywords für die *318* durch das Analysten-Team erstellen zu lassen. Es wurde außerdem entschieden, die 30 häufigsten Keywords mittels *one-hot-encoding* in den Datensatz aufzunehmen (z.B. *keyword_during-creditsstinger*). Um einen besseren Einblick in die Filmdaten zu erhalten, wurde außerdem eine Wordcloud der 100 häufigsten Keywords erstellt, welche in Abbildung 23 im Abschnitt 4.4.3.2 zu finden ist.

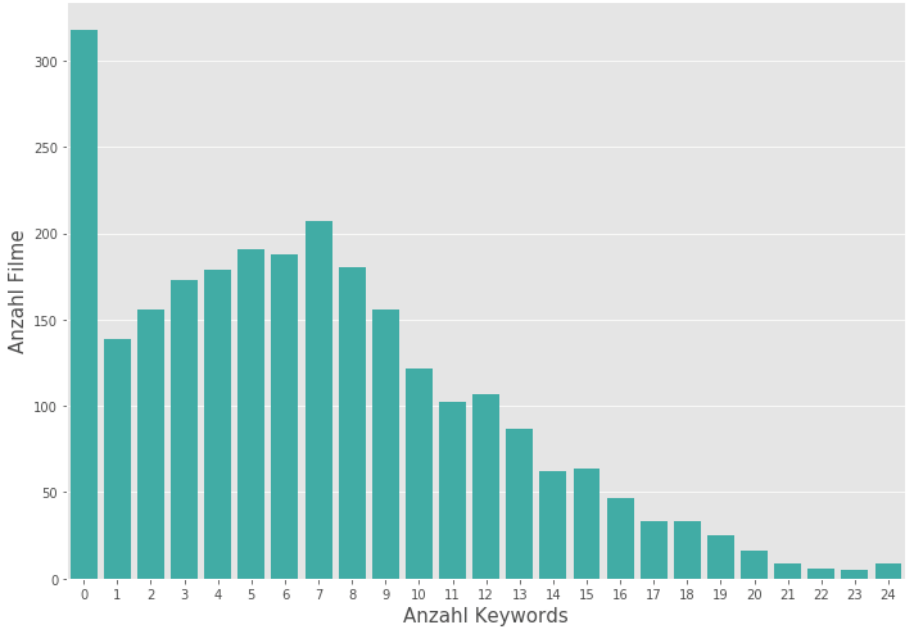

Abbildung 10: Anzahl der Filme mit n-Keywords

Text Features

Features mit Text als Inhalt stellen eine Besonderheit in dem Datensatz dar. Während es zwar eine Vielzahl von Methoden gibt Text-Features aufzubereiten (z.B. Füllwortentfernung oder Stemming/ Lemmatization) und weiterzuverarbeiten (z.B. N-Grams oder one-hot-encoding) führt dies häufig zu dem Hinzufügen von sehr vielen neuen Features. Um die Anzahl der Features des Datensatzes in einem sinnvollen Rahmen zu halten, wurde daher versucht diese Features vorab zu analysieren und nur begrenzt mit in den Datensatz aufzunehmen.

Title

In *title* ist der englische Name des Films hinterlegt. Der Titel des Films kann durch das Filmstudio grundsätzlich frei gewählt werden. Er muss nicht unbedingt einen direkten Hinweis auf den Inhalt des Films geben und repräsentiert oft einen Eigennamen (z.B. *Minions*). Um einen besseren Einblick in den Aufbau der Filmtitel in dem Datensatz zu erhalten, wurde die Wordcloud in Abbildung 11 erstellt. In ihr werden die am häufigsten verwendeten Worte in den Titeln

der Filme dargestellt. Wie zu erkennen, sind unspezifische Worte wie *Man*, *Movie*, *Day* oder *Life* dominant. Interessant ist, wie häufig das Wort *Part* vorkommt, was darauf schließen lässt, dass viele Fortsetzungen beziehungsweise Reihen in dem Datensatz vorhanden sind. Spezifische Begriffe wie *3D* oder *Monster* werden ebenfalls durch die verteilten Keywords (siehe *4.3.4.5 Keywords*) abgedeckt. Daher wurde entschieden, das Feature *title* aus dem Datensatz zu entfernen und keine neuen Features aus ihm abzuleiten.

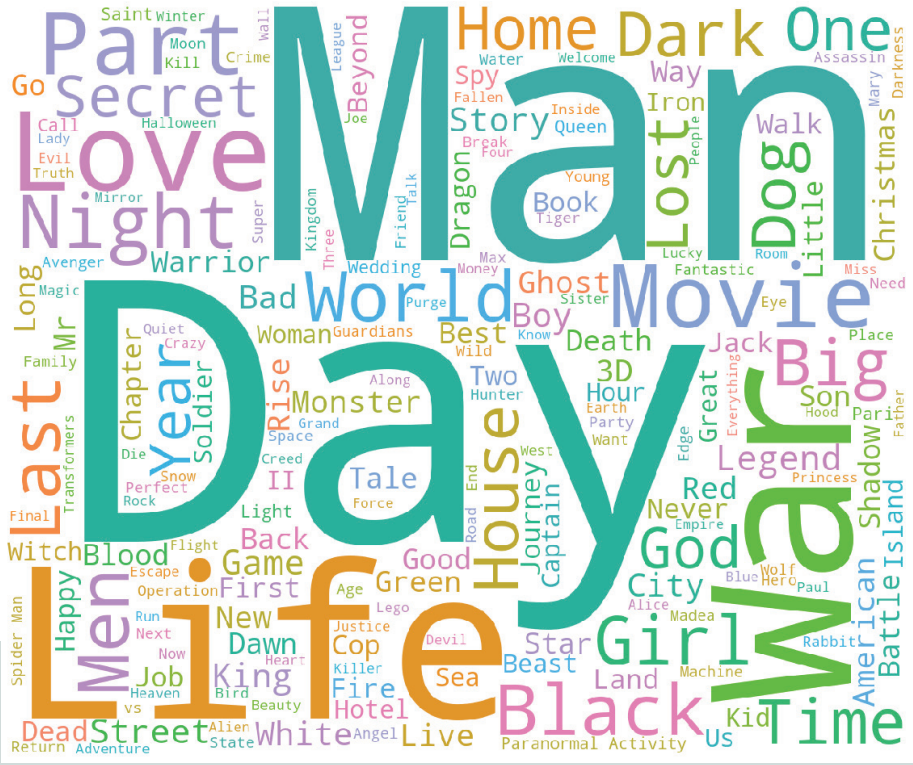

Abbildung 11: Wordcloud Worthäufigkeit in den Titeln

Overview

Das Feature *overview* gibt die vom Filmstudio veröffentlichte Beschreibung des Films an. Ihr Ziel ist es, die Zuschauer für den Film zu interessieren, so dass sie sich ihn im Kino anschauen. Da den Zuschauern durch die Beschreibung nicht bereits zu viel Handlung verraten werden soll, wird jene eher vage gehalten.

In unserem Datensatz ist *overview* durchschnittlich *280,5* Zeichen lang und somit nur wenig länger als die durchschnittliche Länge der Keyword-Listen mit *261,8* Zeichen (siehe Abschnitt 4.3.4.5 *Keywords*). Bei diesen Werten ist zu beachten, dass in dem Listen Feature noch zusätzliche Zeichen durch die Python-Syntax enthalten sind (z.B. geschweifte Klammern), in den Beschreibungen aber auch Satzzeichen, welche in den Listen nicht auftreten. Das Analysten Team geht nach einer Stichprobe davon aus, dass sich die Anzahl dieser zusätzlichen Zeichen ungefähr ausgleicht. Zu beachten ist zusätzlich, dass *overview* (*17*) deutlich weniger Samples ohne Inhalt enthält als *keywords* (*318*). Da leere Zeilen den Durchschnitt deutlich senken, müssen die anderen Filme eine entsprechend höhere Anzahl an Zeichen enthalten, um diesen hohen Durchschnitt zu halten.

In Abbildung 12 ist die Wordcloud über *overview* dargestellt. Es ist ersichtlich, dass hier wie bei *title* vor allem unspezifische Wörter, die zu einer großen Anzahl von Filmen passen vorkommen. So sind die Wörter *one* und *life* zum Beispiel in *18,1* Prozent (*476*) beziehungsweise *17,5* Prozent (*461*) der Beschreibungen enthalten. Daher wurde auch hier entschieden, diese nicht als zusätzliche Features aufzunehmen.

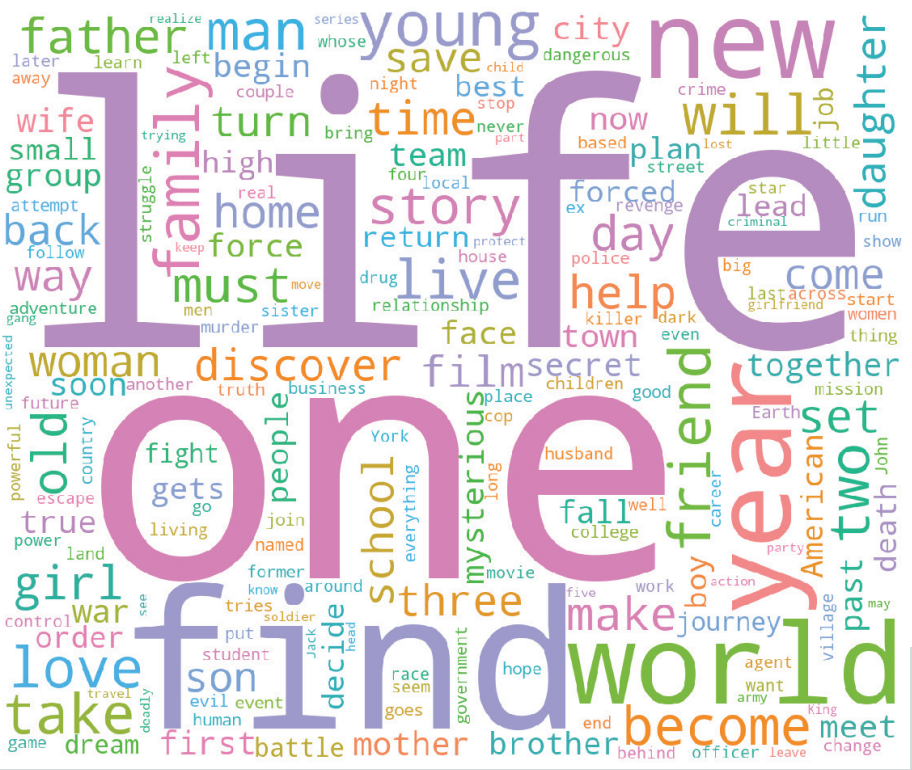

Abbildung 12: Wordcloud Overview

Original_language

Wie in Abschnitt *4.3.4.2 Countries und languages* beschrieben, wird das Feature *original_language* als einziges Feature verwendet, um die Lokalität des Films widerzuspiegeln. Betrachtet man die Verteilung der Sprachen in Tabelle 4, wird klar, dass eine große Mehrheit der Filme *Englisch* (*en*) als Sprache hat. Die Verteilung ist hierbei so eindeutig, dass zunächst überlegt wurde auch das Feature *original_language* aus dem Datensatz zu entfernen. Allerdings zeigt die Verteilung der Einnahmen auf die verschiedenen Sprachen in Abbildung 13, dass es klare Unterschiede zwischen ihnen gibt. So hat *Chinesisch* (*zh*) mit *18,31* den höchsten Median aller Sprachen und liegt somit noch vor Englisch (*17,76*). Den niedrigsten Median der in Abbildung 13 dargestellten Sprachen hat *ml* (*Malayalam*), eine Sprache aus Indien. Den niedrigsten Median aller Originalsprachen im Datensatz hat *rumänisch* (*ro*) mit *11,59.* Es wurde entschieden, die 10

häufigsten Sprachen mittels *one-hot-encoding* als eigene Features mit aufzunehmen.

Sprache	en	hi	ru	ml	fr	ta	es	ko	zh
Filme	1890	153	48	70	54	50	49	41	31

Tabelle 4: Übersicht der 10 häufigsten Originalsprachen

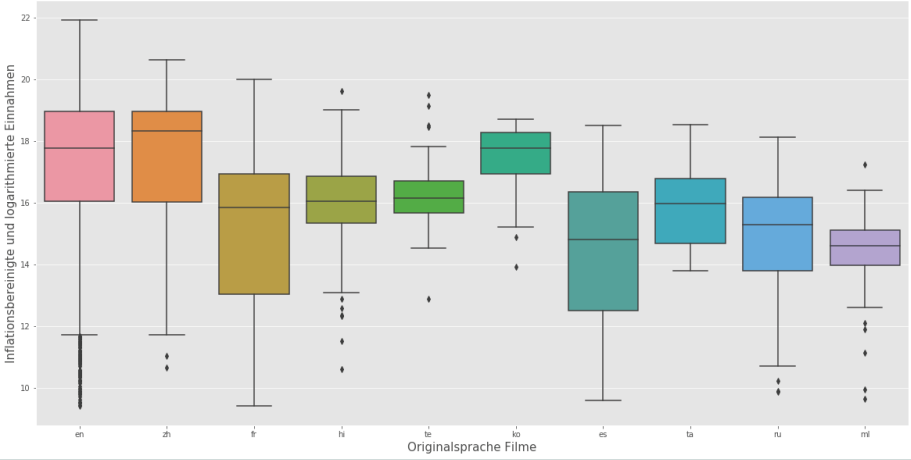

Abbildung 13: Boxplots Einnahmen zu Originalsprachen

Homepage

Wie in Abschnitt *4.3.1 Behandlung der fehlenden Werte* bereits beschrieben, wird der Text in diesem Feature in die Boolean-Werte *1* und *0* umgewandelt. Alle Filme, die in diesem Feature Text enthalten (URL zur Filmwebseite), erhalten *1* als Wert und alle anderen *0*. Hierdurch wird dieses Feature ebenfalls zu einem *Flag*--Feature, dieses trägt den Namen *has_homepage*. Insgesamt haben *1479* Filme eine Webseite hinterlegt und *1151* Filme nicht. Wie in Abbildung 14 zu sehen, haben Filme mit einer Webseite im allgemeinen höhere Einnahmen als Filme ohne. Das Feature *homepage* wird aus dem Datensatz entfernt.

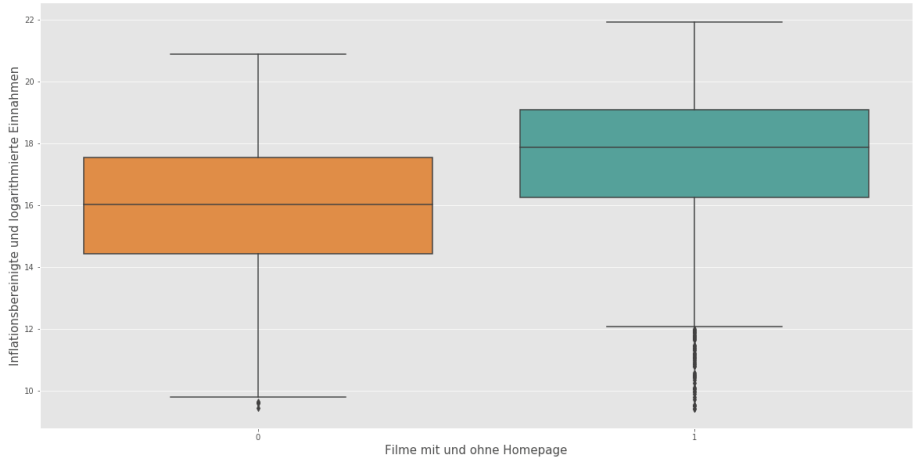

Abbildung 14: Boxplot der Einnahmen von Filmen mit/ ohne Homepage

```
False    2633
Name: adult, dtype: int64
```

Abbildung 15: Ausprägungen Adult

Flag Features

Das einzige Feature, welches direkt in die *Flag Feature*-Kategorie gezählt werden kann ist *adult*, welches angibt ob ein Film ein R-Rating als Altersfreigabe hat oder nicht. Betrachtet man die Ausprägungen dieses Features (Abbildung 15) wird ersichtlich, dass alle Filme in dem vorhandenen Datensatz kein R-Rating haben (*False* als Wert). Da ein Feature, mit nur einer Ausprägung, keinen Mehrwert in der Klassifizierung bietet, wird das Feature adult aus dem Datensatz entfernt (vgl. Larose, 2015, S.120).

Data Understanding Phase: Analyse der Daten

Nachdem ein erster Überblick erhalten sowie eine umfangreiche Aufbereitung der vorliegenden Daten in den zwei vorhergehenden Abschnitten 4.2 und 4.3 durchgeführt werden konnte, werden nun die Filmdaten genauer analysiert. Der Fokus liegt hierbei auf dem in Abschnitt 4.1 definierten sekundären Ziel, das

heißt auf der Untersuchung des Kundeninteresses und der wichtigsten Erfolgsfaktoren.

Zusammengefasst liegt nach Abschluss der *Data Preparation* Phase dem Analysten-Team ein Datensatz mit *336* Spalten, das heißt Features, und *2630* Zeilen, welche den Filmen entsprechen, vor. Die Features unterteilen sich unter anderem in *vier* Dollar Features, welche sich in Einnahmen und Budget splitten lassen mit den jeweils inflationsbereinigten beziehungsweise logarithmierten Werten, *ein* Number-Feature bezüglich der Filmdauer, *sechs* Features in Bezug auf das Veröffentlichungsdatum, *15* Features der häufigsten Genres plus *ein* Feature mit der Anzahl der Genres, *156* Features der relevantesten Schauspieler sowie *83* der relevantesten Produktionsteammitglieder plus jeweils *ein* Feature mit der Anzahl der Schauspieler beziehungsweise der Größe des Produktionsteams, *15* Features der häufigsten Jobs und *11* Features der häufigsten Abteilungen, welche die Anzahl der Personen in den jeweiligen Abteilungen und Jobs der jeweiligen Filme enthalten, und in *30* Features der am häufigsten verwendeten Keywords plus *ein* Feature, welches die Anzahl der Keywords je Film enthält. Außerdem wurden *10* Text-Features der häufigsten Originalsprachen erstellt und *ein* Flag-Feature in Bezug auf die Existenz einer Homepage. Das Text-Feature, welches die Titel der Filme enthält, wurde zur späteren Modellbildung aus dem Datensatz entfernt, da hier überwiegend allgemeine Begriffe enthalten sind, wird aber zur Untersuchung des Kundeninteresses weiter betrachtet.

Analyse der Number Features

Zunächst werden die Number Features näher betrachtet, hierzu zählen auch die Zielvariablen *revenue_inf* und *log_revenue_inf*. Das Dollar-Feature *budget_inf* wird insbesondere in Bezug auf die eingespielten Einnahmen genauer analysiert. Anschließend wird untersucht, inwiefern die Filmlänge einen Einfluss auf den Erfolg eines Films hat.

Verteilung Revenue

Um ein Gefühl für die Zielvariable *revenue_inf* zu erhalten, wird zu Beginn deren Verteilung betrachtet. Das höchste Einspielergebnis eines Films liegt inflationsbereinigt bei *3,3* Milliarden Dollar, wobei es sich um den Film *Avatar* aus dem Jahr 2009 handelt. Der zweite Platz in Bezug auf die Einnahmen belegt *Avengers: Endgame* von 2019 mit *2,8* Milliarden Dollar, der dritte Platz wird von *Star Wars: The Force Awakens* von 2015 mit *2,3* Milliarden Dollar belegt. Bei

diesen Filmen handelt es sich um Ausreißer in Bezug auf die Variable *revenue_inf*, was man im dargestellten Boxplot in Abbildung 16 gut erkennen kann. Diese Werte wurden entsprechend überprüft und als korrekt eingeordnet. Im Gegensatz hierzu liegt das Minimum der Einnahmen bei *12.228,8* Dollar. Im untenstehenden Boxplot kann man außerdem erneut erkennen, dass die Variable *revenue_inf* rechtsschief verteilt ist, da der Median, welcher einen Wert von *26,8* Millionen Dollar aufweist, unter den durchschnittlichen Einnahmen pro Film von *116,9* Millionen Dollar liegt.

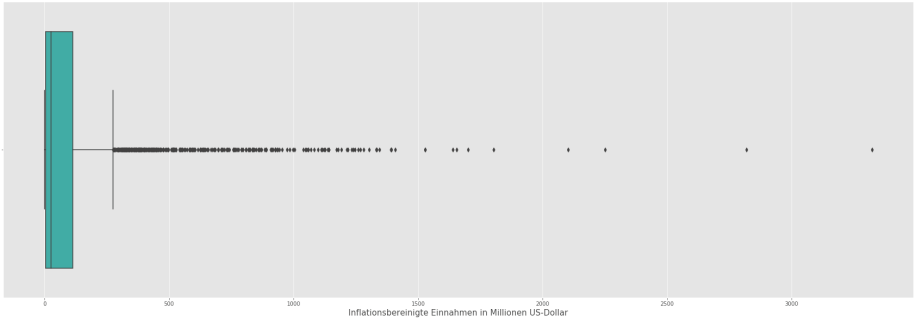

Abbildung 16: Boxplot der inflationsbereinigten Einnahmen in Millionen Dollar

Verteilung Budget

Betrachtet man nun die Verteilung der inflationsbereinigten Variable *budget_inf*, stellt man fest, dass auch hier Ausreißer existieren. Die höchsten Produktionskosten der vorliegenden Filmdaten liegen inflationsbereinigt bei *1,1* Milliarden Dollar, welche dem Film *Hello: A Portrait of Leslie Phillips* zugeordnet werden können. Dieser Wert weist einen hohen Abstand zu den nächst höheren Budget-Werten auf. Außerdem konnte diese Information durch Recherche in unterschiedlichen Quellen zu den bisher teuersten Filmen nicht verifiziert werden. Aus diesen Gründen wurde dieser Ausreißer als fehlerhaft deklariert und aus dem Datensatz entfernt. Die nächst höchsten Produktionskosten liegen bei *521,6* Millionen Dollar des Films *Justice League* und bei *432,8* Millionen Dollar, was den Kosten des Films *Pirates of the Caribbean: On Stranger Tides* entspricht. In Abbildung 17 wird der Boxplot der Verteilung der Variable *budget_inf* dargestellt, welcher eine Ähnlichkeit zur Verteilung der Einnahmen aufweist. Es ist ebenfalls zu erkennen, dass der Median unter dem Mittelwert liegt und somit

die rechtsschiefe Verteilung zu erkennen ist. Diese Werte liegen bei *15,1* Millionen Dollar und bei *37,2* Millionen Dollar.

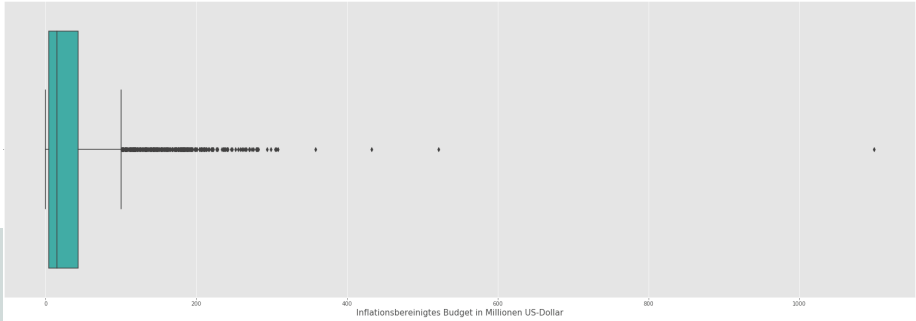

Abbildung 17: Boxplot des inflationsbereinigten Budgets in Millionen Dollar

Zusammenhang Revenue – Budget

Wie zu Beginn des Dokuments erwähnt, werden gemäß einer Studie die Einspielergebnisse durch das eingesetzte Filmbudget positiv beeinflusst (vgl. Kim, 2013, S. 1071). Eine Erklärung hierfür ist, dass die Qualität der Filme durch verschiedene Faktoren steigt, wenn mehr Geld in die Produktion gesteckt wird. Dies bedeutet beispielsweise, dass bekanntere und somit teurere Schauspieler gebucht werden können oder mehr Budget in aufwendige Spezialeffekte oder Kostüme investiert werden kann. Daraus ergibt sich wiederum, dass der Film eine breitere Masse anspricht und überzeugt, sich diesen Film anzuschauen. Dieses Verhalten spiegelt sich auch in den vorliegenden Daten wider.

In Abbildung 18 kann man deutlich einen linearen Zusammenhang der beiden inflationsbereinigten und logarithmierten Variablen *inf_rev_log* und *inf_bud_log* erkennen. Der berechnete Korrelationskoeffizient der beiden Features liegt bei *0,74*, was ebenfalls die Vermutung eines positiven Zusammenhangs unterstreicht.

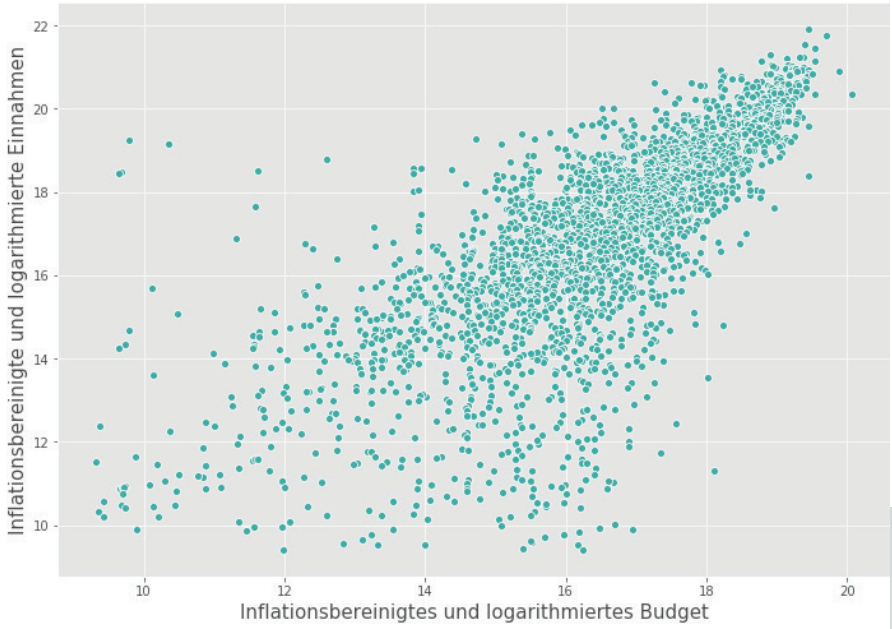

Abbildung 18: Zusammenhang Einnahmen und Budget

Zusammenhang Revenue – Runtime

Wie bereits im Abschnitt 4.3.3.2 erwähnt und in Abbildung 4 erkannt, gibt es einige Ausreißer in der Variablen *runtime*, welche vollständig erklärt werden konnten. Im dargestellten Boxplot ist außerdem zu erkennen, dass ein Großteil der Filme eine Laufzeit zwischen *96* und *124* Minuten hat. Der Median und Mittelwert liegen beide bei ca. *110* Minuten, was auch auf Basis eigener Erfahrungen eine übliche Spielzeit ist.

Mit Blick auf Abbildung 19 gibt es keinen ersichtlichen linearen Zusammenhang zwischen der Laufzeit eines Films und dessen Einspielergebnisse. Es ist allerdings tendenziell zu erkennen, dass Filme mit einer Laufzeit unter *90* Minuten eher weniger Einnahmen generieren. Dies kann wiederum daran liegen, dass Filme mit einer längeren Laufzeit mehr Budget benötigen und somit, aufgrund des linearen Zusammenhangs, höhere Ergebnisse einspielen.

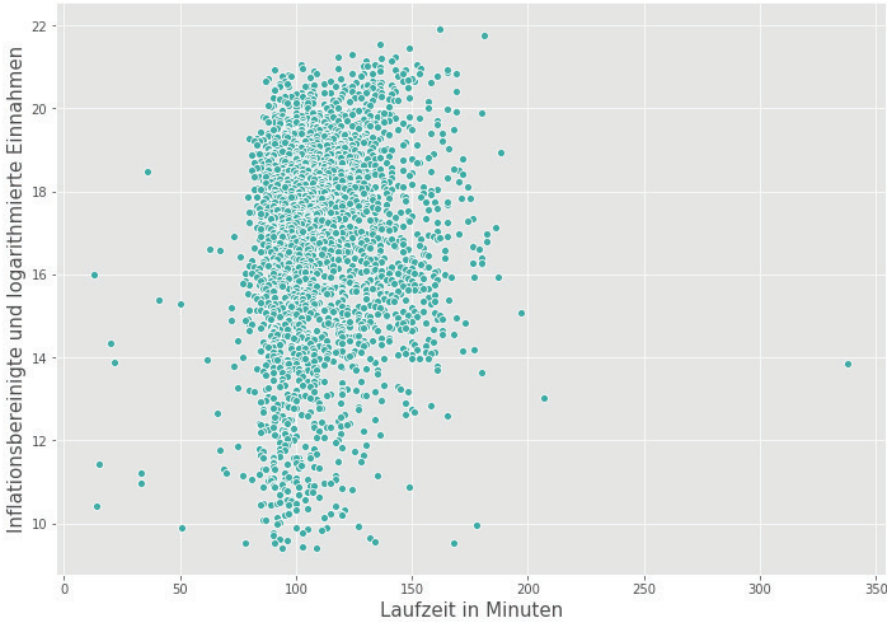

Abbildung 19: Zusammenhang Einnahmen und Laufzeit

Analyse des Veröffentlichungsdatums

In diesem Abschnitt werden die Features des Veröffentlichungsdatums in Bezug auf die Einspielergebnisse untersucht.

Da die Variable *revenue* bezüglich Inflation für weitere Analysen bereinigt wurde, wurden die dadurch begründeten Schwankungen entfernt. Des Weiteren kann in Abbildung 2 erkannt werden, dass das Jahr der Filmveröffentlichung keine besondere Rolle spielt, da die Einnahmen über den betrachteten Zeitraum relativ konstant blieben. Eine Ausnahme liegt im Jahr 2016 vor, in welchem insgesamt die höchsten Einnahmen generiert werden konnten. Da das gewünschte Modell auf zukünftige Filmproduktionen angewendet werden soll und somit der Zusammenhang der Einnahmen und der vergangenen Jahre hinfällig ist, wird dieses Feature nicht näher betrachtet.

Schaut man sich allerdings die Verteilung der Einnahmen in Bezug auf die Veröffentlichungsmonate in Abbildung 20 an, sind Schwankungen erkennbar. Die Schwankungen entsprechen in etwa einer Welle mit zwei erkennbaren Tiefen.

Zum einen zeichnet sich ein Sommerloch in den Monaten August und September ab, welches sich vermutlich durch üblicherweise gutes Wetter und die Ferienzeit erklären lässt. Zum anderen werden im Januar verhältnismäßig wenig Einnahmen generiert, was gegebenenfalls damit zusammenhängen kann, dass die „Kinosaison" nach der Sommerpause ab Oktober startet und viele Produktionsstudios ihre „Kassenschlager" im November oder Dezember veröffentlichen. Vielversprechende Monate der Veröffentlichung sind hingegen April und Dezember, welche die höchsten Einnahmen aufweisen.

Bei der Betrachtung der Wochentage der Filmveröffentlichung in Bezug auf die eingespielten Ergebnisse, ist kein relevanter Einfluss zu erkennen. Allerdings werden donnerstags die meisten Filme veröffentlicht, was auch in Deutschland in der Regel der Fall ist.

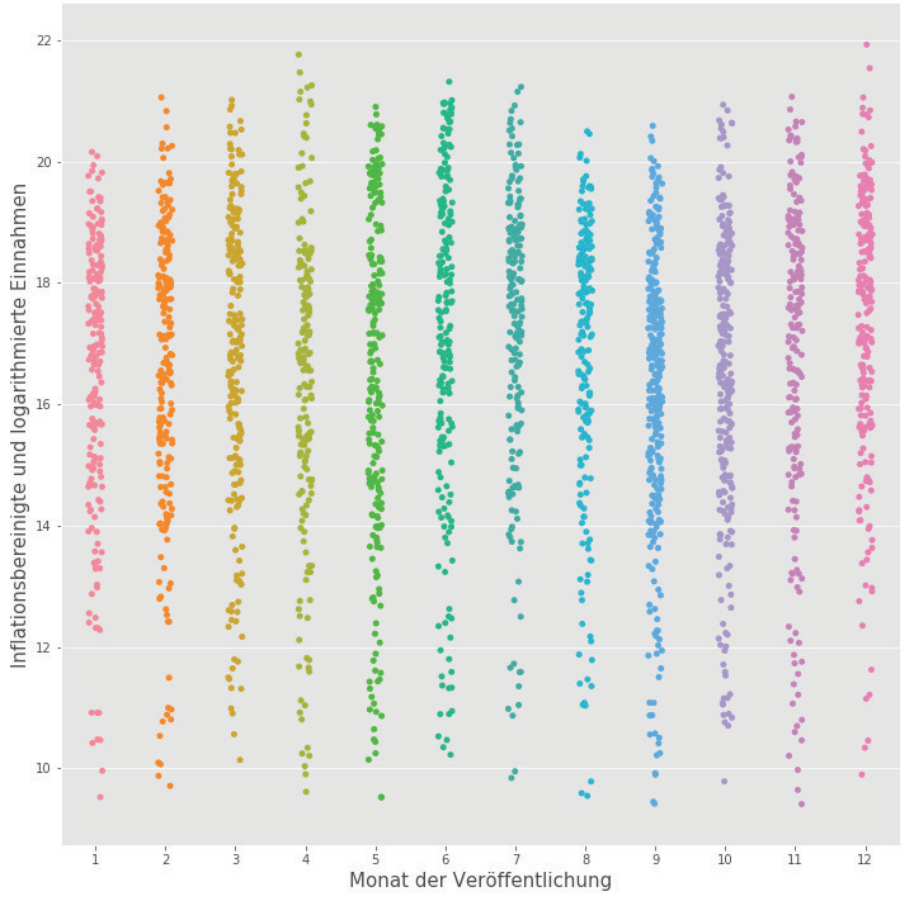

Abbildung 20: Zusammenhang Einnahmen und Monat der Veröffentlichung

Analyse der Listen Features

Nun werden die aufbereiteten Listen Features näher betrachtet. Zunächst wird der Zusammenhang der Filmgenres zur Zielvariablen untersucht. Im Anschluss wird auf die Keywords der Filme näher eingegangen. Die Analyse der Features der relevantesten Schauspieler und Produktionsteammitglieder lassen wir an dieser Stelle außen vor, da diese bereits in der vorherigen Phase *Data Preparation* umfangreich untersucht wurden.

Zusammenhang Revenue – Genres

Wie bereits in Abschnitt 4.3.4.1 erkannt, werden den meisten Filmen ein bis vier Genres zugeordnet. *Drama*, *Comedy* und *Action* zählen zu den häufigst verwendeten Filmgenres. In Bezug auf die Einspielergebnisse sieht es ähnlich aus. Wie man in Abbildung 21 erkennen kann, werden die meisten Einnahmen generiert, wenn den Filmen drei beziehungsweise vier Genres zugeordnet wurden. Eine Erklärung hierfür ist, dass durch die Zuordnung mehrerer Genres der Geschmack einer breiteren Masse getroffen wird. Eine Hinzunahme weiterer Genres lässt die Einnahmen allerdings wieder sinken, da es vermutlich oft nicht oder nur schwer möglich ist, einen Film in mehr Richtungen auszuprägen. Falls es doch umgesetzt wird, liegt nahe, dass dadurch die Qualität des Films sinkt und es somit zu geringeren Einnahmen kommt.

Betrachtet man die positiven Einflüsse der einzelnen Genres, wird klar, dass diese nicht unbedingt mit der Häufigkeit jener übereinstimmen müssen. Die Genres *Drama* und *Comedy*, welche am häufigsten in den Daten auftreten, beeinflussen nicht beziehungsweise sogar negativ die inflationsbereinigten, logarithmierten Einnahmen (siehe Abbildung 22). Das Genre *Action* wird wiederum häufig verwendet und beeinflusst gleichzeitig auch die Einnahmen positiv. Weitere Genres mit positivem Zusammenhang mit den Einspielergebnissen sind Adventure, *Science-Fiction*, *Family*, *Fantasy* und *Animation*.

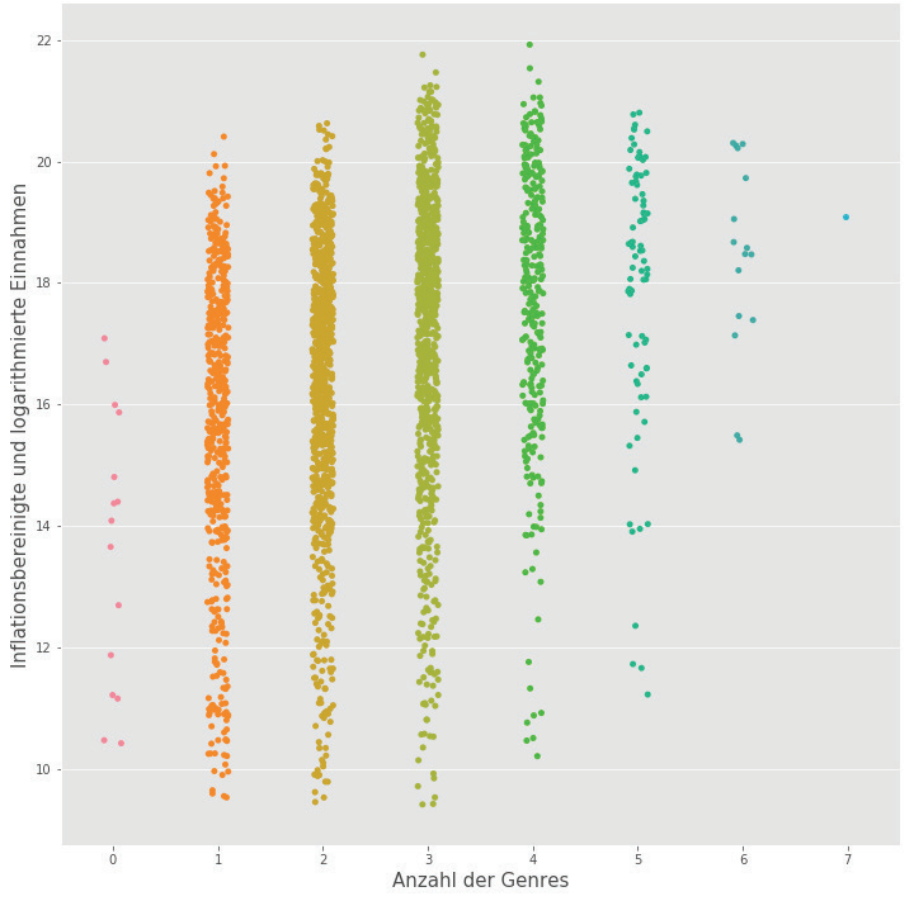

Abbildung 21: Zusammenhang Einnahmen und Anzahl der Genres

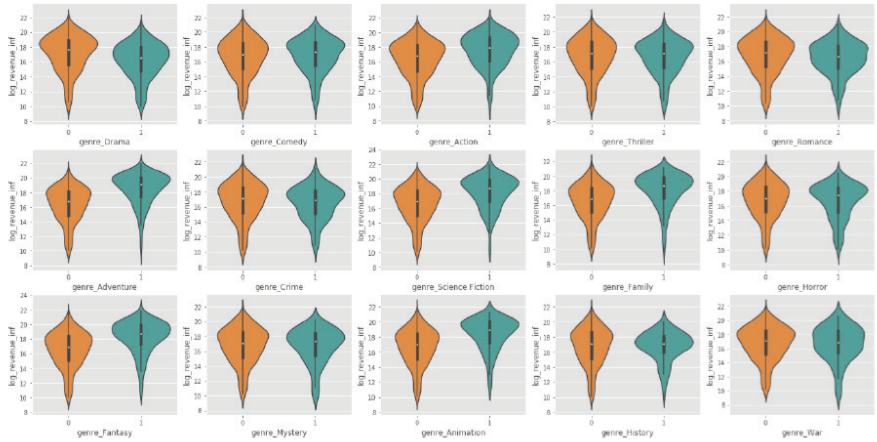

Abbildung 22: Violine-Plot Einnahmen und Genres

Zusammenhang Revenue – Keywords

Um einen besseren Einblick in die einzelnen Keywords der Filme zu erhalten, wurde eine Wordcloud der 100 häufigsten Keywords erstellt, welche in Abbildung 23 zu finden ist. Wie man erkennen kann, zählen zu den häufigsten Keywords Angaben wie *duringcreditsstinger*, *based on novel or book* und *sequel*.

Unter *duringcreditsstinger* kann verstanden werden, dass diese Filme sogenannte Mid-Credit-Szenen verwenden. Darunter fallen Szenen, welche während dem Abspann gezeigt werden, wie beispielsweise Outtakes. Die Häufigkeit des Keywords *based on novel or book* wird durch weitere ähnliche Keywords unterstützt, wie zum Beispiel *biography*, *based on a true story* oder *based on young adult novel*.

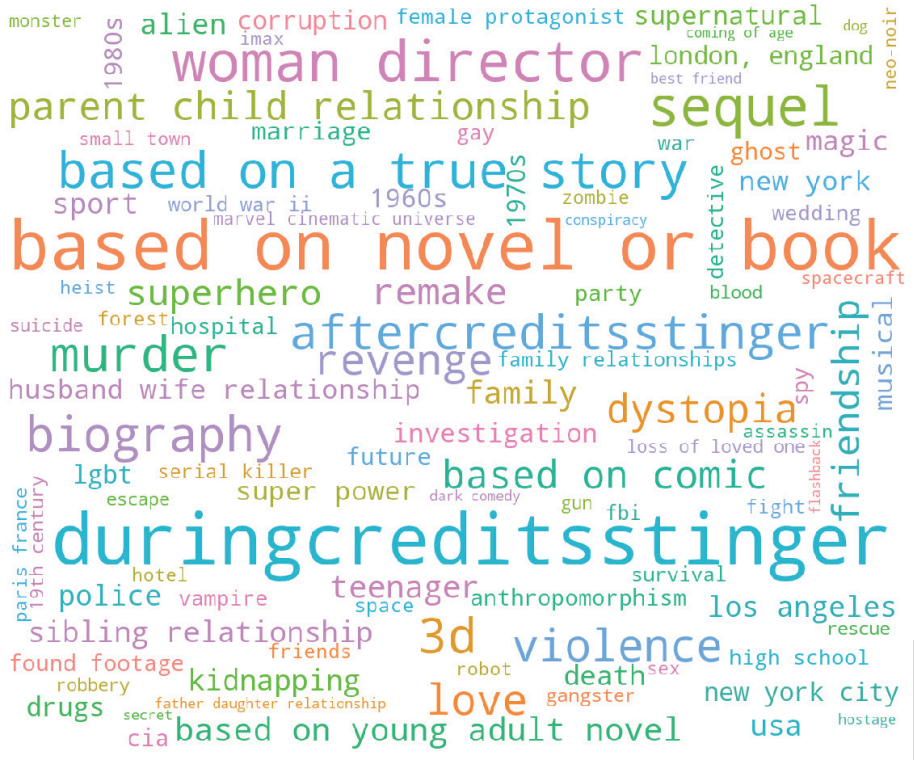

Abbildung 23: Wordcloud Keywords

In Abbildung 24 werden die Zusammenhänge der einzelnen Keywords mit den inflationsbereinigten, logarithmierten Einnahmen dargestellt. Im Gegensatz zu den Genres, lässt die Häufigkeit der Keywords mehr auf deren positiven Einfluss schließen. Zu den erfolgversprechendsten Keywords zählen die folgenden: *duringcreditsstinger*, *sequel*, *3d*, *aftercreditsstinger*, *dystopia*, *based on comic*, *superhero*, *based on young adult novel*, *alien* und *magic*.

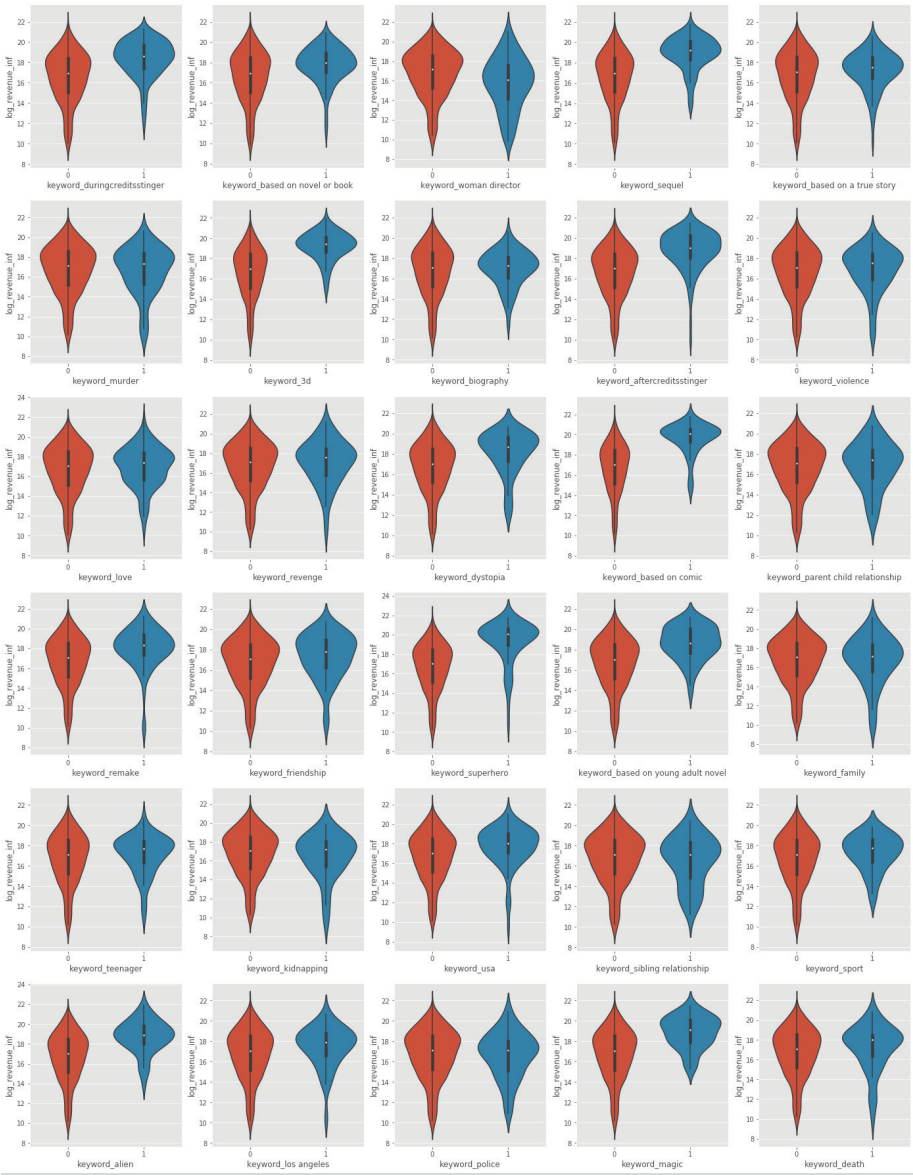

Abbildung 24: Violine-Plot Einnahmen und Keywords

Es ist zu erkennen, auch mit Blick auf die einflussreichsten Genres, dass Themen in Bezug auf Superhelden, Comics beziehungsweise Animation und Science-Fiction sehr erfolgsversprechend sind.

Das Keyword *3d* lässt darauf schließen, dass für die Produktion ein entsprechend hohes Budget benötigt wurde und es somit, aufgrund des linearen Zusammenhangs, zu höheren Einnahmen kommt. Außerdem sind die Preise für 3D-Filme in den Kinos in der Regel höher, wodurch ebenfalls mehr Einnahmen generiert werden.

Betrachtet man das Keyword *sequel* fällt auf, dass dieses eine hohe Häufigkeit sowie einen hohen Einfluss auf die Einnahmen aufweist. Analog hierzu, wie bereits in Abbildung 11 in Abschnitt 4.3.5.1 gesehen, kommt ebenfalls der Begriff *Part* im Filmtitel verhältnismäßig oft vor. Dies lässt darauf schließen, dass Filme, bei welchen es sich um Mehrteiler oder Fortsetzungen handelt, sich einer relativ hohen Beliebtheit erfreuen und somit Erfolg versprechen.

Modeling Phase

Im folgenden Abschnitt wird beschrieben, wie die Klassen zur Entscheidung, ob ein Film als Erfolg angesehen wird oder nicht, gebildet wurden. Anschließend werden unterschiedliche Modelle auf diese Klassen trainiert und miteinander verglichen, um hier ein bestmögliches Ergebnis zu erzielen.

Als Benchmark gilt ein einfaches Modell, dass im Anschluss durch das komplexere Modell geschlagen werden soll. Dafür wird einerseits die automatische Modellerstellung von *RapidMiner* verwendet, um einen ersten Überblick über die Leistungsfähigkeit der einzelnen Modelle auf dem Datensatz zu erhalten. Anschließend werden die Ergebnisse bewertet und darauf aufbauend manuell ein Modell erstellt sowie durch Feature-Selektion optimiert.

Im Anschluss werden die verschiedenen Modelle auf ihre Wirtschaftlichkeit geprüft und miteinander verglichen.

Klassifikation und Vorbereitung

Da in Zukunft möglichst wenig erfolglose Filme produziert werden sollen, werden dementsprechend die Filme binär in erfolgreiche und erfolglose Filme unterteilt. Um diese Teilung zu ermöglichen, wird ermittelt, wie das Verhältnis zwischen Umsatz und Budget ist. Dafür wird der Umsatz durch das Budget geteilt. Für die Berechnung werden die inflationsbereinigten Werte verwendet.

Das Budget enthält nicht alle Kosten, die anfallen, um einen Film zu veröffentlichen. Lediglich die Produktionskosten sind hierin enthalten. Alle weiteren Ausgaben, wie Marketing und Werbung fallen zusätzlich an. Zusätzlich fließen nicht alle Einnahmen vollständig an das Produktionsstudio zurück, sondern werden zu einem Teil auch von weiteren beteiligten Unternehmen, wie zum Beispiel den Kinoketten, einbehalten. Dementsprechend gilt ein Film, falls dieser einen Faktor von eins hat, noch nicht als erfolgreich, da dadurch noch nicht alle tatsächlichen Kosten abgedeckt sind.

Wie viel höher die Einnahmen eines Films im Vergleich zu dessen Produktionskosten sein müssen, hängt von der Höhe des Budgets und der Art des Films ab. Nach eigenen Erfahrungen definiert CINEMAKE einen Film als erfolgreich, wenn dieser mindestens das Zweifache oder mehr seines Budgets eingespielt hat. Wendet man dieses Kriterium auf den vorliegenden Datensatz an, werden *1330* Filme als Erfolg und *1300* Filme als Misserfolg klassifiziert. Erfolge werden in dem neu angelegten Feature *success* als *1* codiert und Misserfolge als *0*. Aus ökonometrischer Sicht und mit Blick auf maschinelles Lernen ist diese fast hälftige Aufteilung vorteilhaft. Probleme mit rare events wie man sie etwa vom Logit Modell kennt, können somit vermieden werden.

Als weitere Vorbereitung werden einige Features aus dem Datensatz entfernt. Hierunter fallen alle Features, welche eine Form der Zielvariablen darstellen, wie *revenue*, *revenue_inf* und *log_revenue_inf*. Zusätzlich wird von allen Features, welche das Budget widerspiegeln, wie im Abschnitt 4.3.3.1 *Dollar Features* erläutert, nur das logarithmierte und inflationsbereinigte Budget (*log_budget_inf*) verwendet. Somit wird das *budget* und *budget_inf* aus dem Datensatz entfernt. Mögliche Endogenitätsprobleme werden somit mitigiert.

Der Datensatz wird wie üblich aufgeteilt in einen Trainings- und Testdatensatz, um die unterschiedlichen Modelle trainieren und anschließend testen und vergleichen zu können. Die verwendeten numerischen Methoden benötigen eine grosse Stichprobe. Deshalb werden *90%* der vorliegenden Filmdaten für den Trainingsdatensatz verwendet und die restlichen *10%* für den Testdatensatz.

Die Trennung der Daten wird über einen Prozess in *RapidMiner* durchgeführt. Dafür wird ein Operator zum Teilen der Daten verwendet. Es muss definiert werden, wie viele *Subsets* es geben soll, sowie deren Größe. Das verwendete Stichprobenverfahren ist ein *stratified sampling* verfahren, dass sicherstellt, dass ein zufällig erstelltes *Subset* dieselbe Verteilung hat, wie der ursprüngliche Datensatz. Dies eignet sich vor allem für ähnlich große binäre Klassen, wie es in

dem oben beschriebenen Filmdatensatz der Fall ist. Für die Gewährleistung einer Reproduzierbarkeit wurde ein *local random Seed* von *2000* verwendet.

Einfaches Modell

Das einfache Modell ist so charakterisiert, dass es immer den am häufigsten vorkommenden Fall vorhersagt. Bei binären Problemen mit einer annähernd hälftigen Aufteilung wie hier ist somit sichergestellt, dass immer mindestens eine Trefferquote von 50 Prozent vorliegt. Alle weiteren Modelle werden gegen dieses Modell gestellt (Benchmarking), um zu erkennen, in wie weit der zusätzliche Modellierungsaufwand eine Verbesserung mit sich bringt.

Das einfache Modell wird anhand des Testdatensatzes bestimmt, da alle weiteren Modelle ebenfalls mit diesem *Subset* getestet werden. Der Testdatensatz hat eine Verteilung von *133* erfolgreichen und *130* erfolglosen Filmen, weswegen das einfache Modell immer einen Erfolg vorhersagen wird, wie in Tabelle 5 zu sehen ist.

	True 0	True 1	Precision
Prediction 0	0	0	0%
Prediction 1	130	133	50,57%
Genauigkeit			50,57%

Tabelle 5: Konfusionsmatrix Einfaches Modell

Mit dieser Vorhersage würde das Modell in *50,57* Prozent der Szenarien einen Film korrekt als Erfolg vorhersagen, was ebenfalls der Genauigkeit, synonym Accuracy, des Modells entspricht.

Ob sich der Einsatz des einfachen, synonym naiven, Modells finanziell lohnt, wird deutlich, wenn wir die Kosten der erfolglosen Filme den Einnahmen der erfolgreichen Filme gegenüberstellen wie es in Kapitel 4.6 *Evaluation Phase* getan wird.

Auto Modell

RapidMiner bietet viele Möglichkeiten und Operatoren an, um Daten zu bearbeiten und zu analysieren. Zusätzlich wird die Möglichkeit geboten, automatisch

Modelle zu trainieren. Hierfür und für andere Analysen müssen die Daten in *RapidMiner* eingelesen werden. Im Anschluss muss definiert werden, welches Feature die Zielvariable sein soll und welches die positiv belegte Klassifizierung ist. Zusätzlich wird die Möglichkeit geboten Strafterme festzulegen, wenn eine Vorhersage fehlerhaft ist. Im Schritt *Select Inputs* können die Features für die Modelle ausgewählt werden, wobei *RapidMiner* durch eine Ampelkodierung versucht relevante Features visuell hervorzuheben. Hierbei werden die Daten auf Korrelationen, fehlende Werte, zu einseitig oder zu verschieden befüllte Features geprüft. Dies ist für den verwendeten Datensatz nur bedingt notwendig, da bereits zuvor fehlende Werte und offensichtliche Redundanzen, wie Budget und inflationsbereinigtes Budget, bereinigt wurden. Aufgrund dieser Vorverarbeitung werden alle Features zum Trainieren des Modells verwendet.

Abschließend können die bevorzugten Modelle mit einigen Einstellungsmöglichkeiten ausgewählt werden. Zusätzlich werden weitere Vorverarbeitungsmöglichkeiten, wie eine weitere Feature-Selektion angeboten. Für den verwendeten Datensatz werden alle Modelle trainiert, um einen ersten Eindruck zu bekommen, welche Modelle vielversprechend sein könnten und zu prüfen, ob ein solches einfaches *Auto Modell* bereits besser sein kann als das einfache Modell.

In Abbildung 25 wird die Genauigkeit der einzelnen Modelle visuell dargestellt.

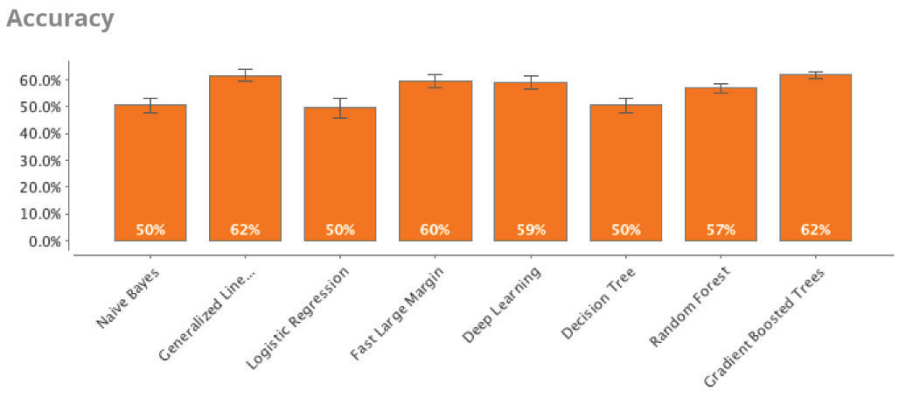

Abbildung 25: Genauigkeit Auto Modell

Die Genauigkeit beschreibt das Verhältnis der korrekten Vorhersagen zu der gesamten Anzahl an Beispielen. Hierfür wird der Datensatz von *RapidMiner* automatisch in einen Test- und einen Trainingsdatensatz geteilt. Das Modell *Naive Bayes* ähnelt dem einfachen Modell und sagt immer einen Erfolg voraus. Die

besten Modelle sind das *Generalized Linear Model* und die *Gradient Boosted Trees*. Diese beiden Modelle erzielen eine Genauigkeit von *62* Prozent.

Die Präzision der verschiedenen Modelle, die in Abbildung 26 abgebildet wird, beschreibt in Prozent, wie viele der als wahr vorhergesagten Werte wirklich wahr sind. Dies beschreibt die Genauigkeit der Vorhersage, ob ein Film erfolgreich sein wird. Dieser Wert sollte so hoch wie möglich sein und auch hierbei ist das *Generalized Linear Modell* mit *63* Prozent eins der erfolgreichsten.

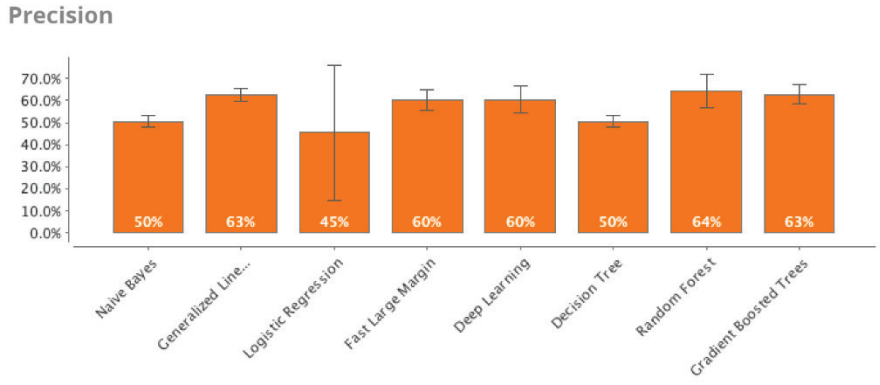

Abbildung 26: Präzision Auto Modell

Logistische Regression

Bei der Auswertung der Modelle des Automodel-Modus von *RapidMiner* im vorherigen Kapitel, ist zu erkennen, dass das *Generalized Linear Modell* deutlich besser abschneidet als die *Logistic Regression*. Dies ist verwunderlich, da die *Logistic Regression* eine auf binäre Probleme spezifizierte Version des *Generalized Linear Modells* darstellt. Aus diesem Grund sollten die beiden Modelle bei dem vorliegenden binären Problem ähnlich abschneiden, beziehungsweise das *Logistic Regression* Modell besser als gezeigt performen. Aufgrund des großen Unterschieds hat sich das Analytics Team dazu entschieden, das *Logistic Regression* Modell, synonym Logit Modell, manuell zu erstellen, um mögliche Fehler im Auto Modell auszuschließen und sich in Richtung mehr Nachvollziehbarkeit zu bewegen (vgl. Menard, 2002, S. V).

In *RapidMiner* wird der Algorithmus für eine logistische Regression verwendet, der durch die Firma *H20.ai* realisiert ist (vgl. RapidMiner, 2019, o. S.). Es ist eine

vereinfachte Form der in *RapidMiner* verwendeten „Generalized Linear Regression", bei der die Einstellungen so vorgenommen sind, dass eine logistische Regression verwendet wird.

Mit dem Ausführen und dem Start dieses Modells, wir ein einzelner H2O Cluster erzeugt, auf dem der Algorithmus ausgeführt wird.

Zum Trainieren des Modells, werden die Trainingsdaten in den Operator geladen. Zur Auflösung wird ein *solver* mit der Bezeichnung „L_BFGS" verwendet, der für Datensätze mit vielen Features geeignet ist (vgl. RapidMiner, 2019, o. S.).

Da der Algorithmus auf mehreren Threads laufen kann, ist die Reproduzierbarkeit nicht ohne weiteres gewährleistet, weswegen die Anzahl Threads auf maximal vier beschränkt wird. Um *Lambda* und *Alpha* für die Regularisierung festzulegen, wird Lambda auf *0,5* gesetzt. Mit *Alpha* wird die genaue Form des Strafterms gesteuert. In dem hier verwendeten Ansatz ist dieser Wert auf *0,05* gesetzt.

Numerische Werte werden für die Logistische Regression normalisiert, ein konstanter Wert wird ermittelt und kollineare Spalten werden entfernt. Da wir keine fehlenden Werte im Datensatz haben, ist es nicht relevant, wie mit diesen umgegangen wird und die maximale Anzahl der Wiederholung wird auf *100* gesetzt.

Mit den oben genannten Werten wird anschließend ein Logit-Modell trainiert, dass im Anschluss auf den Testdatensatz angewendet wird. Abschließend soll die Performance über einen entsprechenden Operator ausgegeben werden, der bestimmt, welche Informationen als Ergebnis angezeigt werden. Des Weiteren werden die Einflüsse der einzelnen Attribute auf das Ergebnis ausgegeben. Bei der visuellen Betrachtung der verschiedenen Einflussfaktoren in Abbildung 27 fällt auf, dass nicht alle Features einen Einfluss haben.

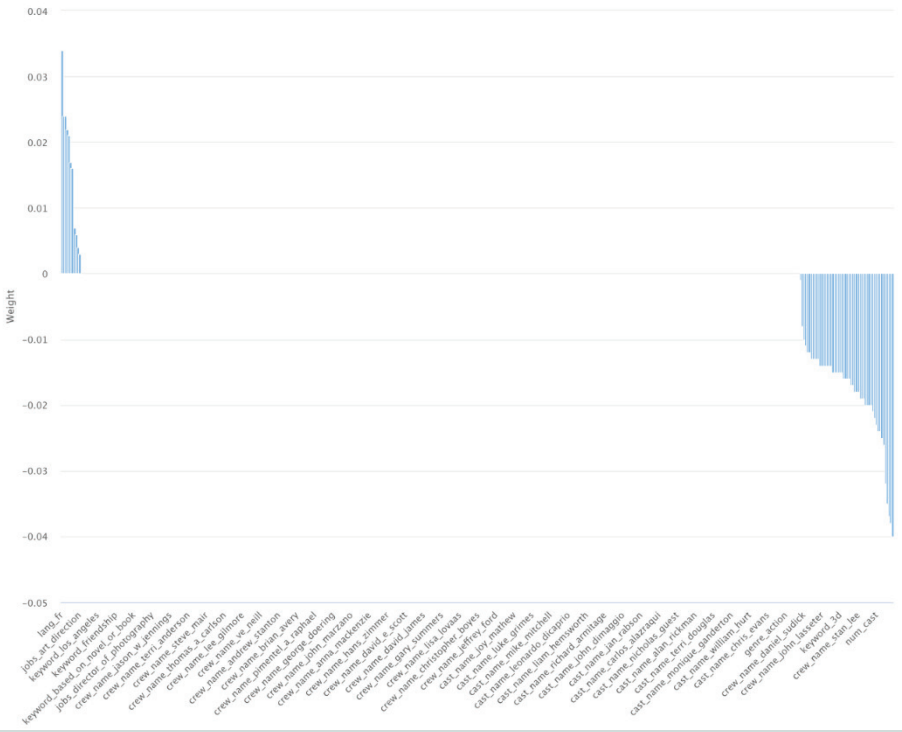

Abbildung 27: Logit-Modell Einflussfaktoren

Durch die Regularisierung via Lambda und Alpha werden von den ursprünglich *460* Features nur *13,7* Prozent (*63*) verwendet.

Die größte Gruppe von Features, die einen Einfluss haben, sind die Crewmitglieder, mit 23 Stück und der Anzahl der der Crewmitglieder. Anschließend kommt die Anzahl an Cast Mitgliedern, mit zehn Attributen und der Größe des Casts. Dann kommen die Keywords, sowie die Abteilungen mit jeweils sieben Werten und den Abschluss bilden die Sprachen mit vier sowie die verschiedenen Berufe mit drei Features. Zusätzlich geben mehrere Features die Personenanzahlen der Abteilungen oder Berufe an.

Die Qualität der Vorhersage des trainierten Logit-Modells wird in der unten abgebildeten Tabelle 6 dargestellt.

	True 0	True 1	Precision
Prediction 0	96	52	64,86%
Prediction 1	34	81	70,43%
Genauigkeit			67,30%

Tabelle 6: Konfusionsmatrix Logit-Modell I

Wie zu erkennen ist, sagt das Modell voraus, dass *115* Filme Erfolg und *148* Filme keinen Erfolg haben werden. Die tatsächlichen Werte für erfolgreiche Filme liegen bei *133* und für erfolglose Filme bei *130*.

Da *96* der Filme korrekt als erfolglos bestimmt werden, liegt die *Precision* hier bei *64,86* Prozent. Die Vorhersage von *81* erfolgreichen Filmen, die tatsächlich erfolgreich sind, sorgt für eine *Precision* von *70,43* Prozent wodurch das einfache Modell deutlich mit fast *20* Prozentpunkten geschlagen wird.

Da RapidMiner nicht in der Lage ist, die t- beziehungsweise p-Werte für das Logit-Modell auszugeben, wird der Datensatz auf die 63 nach Regularisierung relevanten Features gekürzt und in *RStudio* geladen. Hier werden diese Features mit einer logistischen Regression (via Funktionen glm & logistf, jedoch ohne Firth Bias, weil bereits regularisiert wurde) hinsichtlich ihrer Erklärungskraft geprüft. Die Ergebnisse befinden sich in Abbildung 29 im Anhang.

Als Ergebnis der logistischen Regression in *RStudio* ergibt sich, dass von den *63* Features *36* einen p-Wert zwischen *0,1* und *1* haben. Diese Features haben alle einen zu hohen p-Wert und gelten als nicht relevant und müssen deswegen entfernt werden. Von den übrigen *27* Features liegen fünf bei einem p-Wert von *0,05*, neun bei einem von *0,01*, nur drei bei einem Wert von *0,001* und zehn bei einem kleineren Wert. Bei diesen zehn Werten handelt es sich um *runtime, genre_comedy, genre_horror, department_num_crew, num_crew, num_keywords, keyword_sequel*[3]*, lang_fr, lang_ta* und *lang_ko*. Da in Erfahrung gebracht werden soll, ob ein Feature einen Einfluss hat, wird die Gegenhypothese geprüft, dass ein Feature keinen Einfluss hat. Je kleiner in diesem Fall der p-Wert ist, desto eher kann die Gegenhypothese verworfen und somit die

[3] Bei dem Feature *keyword_sequel* wurde darüber nachgedacht, dass ein Endogenitätsproblem vorliegen könnte, da in der Regel nur erfolgreiche Filme Fortsetzungen erhalten.

eigentliche Hypothese, dass das entsprechende Feature einen Einfluss hat, angenommen werden.

Um das Modell zu optimieren und statistisch nachvollziehbar zu bleiben, werden alle Features mit einem p-Wert unter *0,1* in einem neuen Datensatz hinzugefügt. Dieser Datensatz wird in *RapidMiner* wie der vorherige verarbeitet und erzielt das Ergebnis in der Tabelle 7.

	True 0	True 1	Precision
Prediction 0	78	43	64,46%
Prediction 1	52	90	63,38%
Genauigkeit			63,88%

Tabelle 7: Konfusionsmatrix Logit-Modell II

Mit diesem Modell wird eine Vorhersagegenauigkeit für erfolglose Filme von *64,46* Prozent und für erfolgreiche Filme von *63,38* Prozent erzielt. Somit liegt die allgemeine Genauigkeit bei *63,88* Prozent und schlägt das einfache Modell deutlich. Das Modell hat zwar eine geringere Genauigkeit als das zuvor beschriebene Logit-Modell I, ist aber statistisch nachvollziehbar, da die p-Werte ausschließlich diese *27* Features als relevant einstufen. Es ist unser Favorit und wird weiter unten kommerziell interpretiert.

Zusätzlich wird das GLM, welches in *RapidMiner* automatisch erstellt wurde, in der Tabelle 8 betrachtet.

	True 0	True 1	Precision
Prediction 0	241	156	60,71%
Prediction 1	132	222	62,71%
Genauigkeit			61,70%

Tabelle 8: Konfusionsmatrix GLM

Da der Testdatensatz von *RapidMiner* erstellt wird, weicht die gesamte Anzahl der Testdaten von den zuvor verwendeten Datensätzen ab. Hier wird das Model auf *751* Filme getestet, bei denen *373* erfolglos und *378* erfolgreich sind. Die Genauigkeit der Vorhersage eines Films liegt bei *61,7* Prozent und ist somit

ebenfalls um knapp elf Prozentpunkte besser als das einfache Modell aber dennoch um fast zwei Prozentpunkte schlechter als das von uns favorisierte Modell mit lediglich 27 Features.

Evaluation Phase

In der Evaluationsphase werden die erzielten Ergebnisse noch einmal zusammengefasst und die Modelle werden auf den durchschnittlichen Gewinn und somit auf die Wirtschaftlichkeit geprüft.

Aufgrund der Ausreißer wird mit dem Median des Budgets, welches für die Berechnungen aus den oben genannten Gründen verdoppelt wird, und des Umsatzes gerechnet. Des Weiteren werden die Berechnungen für beide Klassen getrennt voneinander durchgeführt, da andernfalls auch erfolglose Filme einen Gewinn erzielen würden. Filme, die der Kategorie *0* zugeordnet werden, machen einen Verlust von *19.462.733,86* Dollar im Median. Filme der Kategorie *1* erzielen im Median einen Gewinn von *51.998.981,50* Dollar. Das einfache Modell, welches eine Genauigkeit von *50,57%* hat, in dem es jeden Film als erfolgreich einstuft, erreicht im Schnitt einen Gewinn von *16.670.643,11* Dollar, was dem Mittelwert von Gewinn und Verlust entspricht.

Ein automatisch generiertes lineares Modell mit einer Vorhersagegenauigkeit von *61,7%* sagt in mehr als *62%* der Fälle einen erfolgreichen Film korrekt voraus und würde somit einen durchschnittlichen Gewinn von *25.352.240,18* Dollar erzielen. Mit einem solchen Modell würde die Firma knapp neun Millionen Dollar im Schnitt pro Film mehr verdienen.

Aus Gründen der Nachvollziehbarkeit wird das Logit-Modell II mit den 27 besten Features geprüft. Mit einer Genauigkeit von *63,88* Prozent wird durchschnittlich ein Gewinn von *25.823.564,61* Dollar erzielt. Das einfache Modell wird hier durchschnittlich um mehr als 9 Millionen Dollar Gewinn übertroffen und auch das automatische Modell wird mit fast einer halben Millionen Dollar übertroffen.

Das primäre Ziel besser bestimmen zu können, ob ein Film erfolgreich wird oder nicht, ist zumindest im Vergleich zum einfachen Modell erreicht. Mit einer *Precision* von über *63* Prozent auf erfolgreiche Filme und einer Genauigkeit von ebenfalls über *63* Prozent schneidet dieses Modell unter der zusätzlichen Bedingung der Erklärbarkeit am besten ab.

Deployment Phase

Das *Deployment* für die logistische Regression wird im nächsten Kapitel detailliert beschrieben.

Preparation

In diesem Abschnitt soll einerseits beschrieben werden, wie die Umsetzung des Modells in der Praxis aussehen soll und wo die Grenzen der unterschiedlichen Modelle, vor allem aber die der automatischen Modelle liegen.

Abschließend werden die Ergebnisse vorgestellt und diskutiert.

Umsetzung des Modells

Alle verfügbaren Daten, die bei der Vorstellung des Films beim Produktionsstudio verfügbar sind, können auf einer eigens dafür entwickelten Webseite manuell oder über eine CSV-Datei eingegeben werden. Auch eventuelle Schätzungen sollten so genau wie möglich eingespeist werden.

Anschließend werden die Daten eingelesen und über ein Python-Skript auf Vollständigkeit und darauf, ob die Daten verarbeitet werden können, geprüft. Falls dies der Fall ist, wird die Eingabe so aufbereitet, dass das oben beschriebene und trainierte Modell die Daten einlesen und verarbeiten kann, sodass bspw. der Schauspieler als binäre Features abgebildet wird.

Wenn dies ohne Fehler erfolgte, wird das Modell über eine Schnittstelle an *RapidMiner* übermittelt und das Ergebnis gespeichert und auf der zuvor beschriebenen Weboberfläche dargestellt. Alternativ kann der Vektor der Features in RStudio in das Logit II Model eingespeist werden. Wird eine Wahrscheinlichkeit grösser als 50% vorhergesagt, so handelt es sich voraussichtlich um einen erfolgreichen Film. Wenn also ein Erfolg vorausgesagt wird, sollte der Film - unter Berücksichtigung der Verfügbarkeit finanzieller Mittel und weiterer üblicher Kriterien der Machbarkeit - realisiert werden. Der Blick über das Modell hinaus ist dabei wichtig. Gerade in Zeiten der Corona Pandemie sollte überhaupt kein Film für die nahe Zukunft avisiert werden – egal was das Modell prognostiziert!

Mindestens einmal im Jahr sollte das Modell erneut auf seine Genauigkeit geprüft werden. Dazu sollten aktuellere Filme hinzugezogen werden und das Modell erneut getestet werden. Dies soll verhindern, dass neue Features eventuell im Laufe der Zeit relevant werden oder aktuelle Features irrelevant. Des Weiteren soll die laufende Kontrolle dazu dienen, dass das Modell weiterhin aktuell bleibt, denn die Welt selbst ist nicht statisch!

Grenzen der Modelle und Programme

Die Grenzen eines Modells liegen darin, dass es keine 100-prozentige Genauigkeit bei der Vorhersage gibt und ein Filmstudio nur eine kleine begrenzte Zahl an Filmen produzieren kann. Die größten Filmstudios produzieren im Median vier Filme im Jahr, was eine relativ kleine Auswahl ist. Dementsprechend wichtig ist es, dass diese Entscheidung möglichst gut getroffen wird und hier kann ein solches Modell unterstützen.

Eine weitere Begrenzung könnte sein, dass sich Features ändern beziehungsweise einige Features nicht mehr zur Verfügung stehen. In dem oben beschriebenen Modell sind viele Personen des Casts oder der Crew an einem Film beteiligt und ausschlaggebend für Erfolg und Misserfolg. Diese Features können abrupt nicht mehr zur Verfügung stehen, da bspw. ein Cast Mitglied aufhört oder gar verstirbt. In einem solchen Fall, wäre ein Modell, das weiterhin den Erfolg eines Films anhand dieses Features misst, nicht sonderlich gut in der Praxis anzuwenden.

Eine zusätzliche Grenze der verwendeten Modelle beziehungsweise Programme betrifft vor allem die automatisch erzeugten Modelle. Diese können meist nur einfachste Zusammenhänge erkennen und geben teils sehr wirre Ergebnisse aus. Hinzu kommt, dass die Ergebnisse schlechter werden, je ungenauer und schlechter die Daten vorverarbeitet sind. Ein weiterer Aspekt ist, dass nicht alle Modelle klar erklärt werden können, wodurch die Transparenz nicht gewährleistet werden kann.

Ergebnisse und Diskussion

Das primäre Ziel der Produktionsfirma CINEMAKE, welches das bessere Verständnis der Kunden und der erfolgsversprechenden Faktoren im Fokus hat, konnte in Abschnitt 4.4, in der Phase *Data Understanding*, erreicht werden.

Durch die reine explorative Datenanalyse wurde ein linearer Zusammenhang der Einnahmen und der Produktionskosten eines Films beobachtet, was auch durch die Literatur bestätigt werden konnte. Dieser Zusammenhang konnte durch die Modellbildung nicht belegt werden, was allerdings an weiteren Faktoren, welche in das Budget beziehungsweise in die Kosten eines Films miteinfließen, liegen könnte. Hierzu zählen unter anderem bekanntere und somit teurere sowie

eine größere Anzahl an Schauspielern, ein größeres Produktionsteam, eine längere Laufzeit sowie mehr verfügbare Berufe und Abteilungen. Genau diese Faktoren findet man allerdings in dem erstellten Modell wieder.

Ein weiteres erfolgsversprechendes Feature nach der reinen deskriptiven Analyse, ist der Veröffentlichungsmonat. Filme, welche in den Monaten Februar bis Juli beziehungsweise November und Dezember veröffentlicht werden, haben gute Erfolgschancen, da die Kunden in diesen Monaten häufiger Kinos besuchen.

Filme der Genres *Action*, *Adventure*, *Science-Fiction*, *Family*, *Fantasy* und *Animation* erfreuen sich nach der Datenanalyse großer Beliebtheit bei den Kunden. Dies wird auch durch die beliebten Keywords *Comic*, *Superhero*, *Alien* und *Magic* unterstrichen. Diese Genres und Keywords sind allerdings bei dem erstellten Modell nicht zu finden. Dies kann an den unterschiedlich betrachteten Zielvariablen liegen. Bei der explorativen Datenanalyse wurde als Zielvariable die Einnahmen eines Films betrachtet und nicht das Verhältnis von Produktionskosten und Einnahmen, wie bei der Modellbildung. Man betrachte beispielsweise Science-Fiction- oder Action-Filme, welche teurer in der Produktion als Komödien oder Dramen sind, unter anderem aufgrund von aufwendigen Spezialeffekten. Diese erzielen zwar auch hohe Einnahmen, sind aber in Bezug auf das Verhältnis von Kosten und Einnahmen nicht so erfolgreich, wie günstiger produzierbare Filme. Dies begründet die insgesamt leicht abweichenden Ergebnisse von EDA und Modellbildung.

Das Ziel der Erstellung eines Modells, welches vorhersagt, ob ein Film erfolgreich ist oder nicht, ist mit einer Genauigkeit von knapp über *63%* prinzipiell erreichbar. Damit sollte es möglich sein, weniger erfolglose Filme zu produzieren als ohne ein solches Modell. Des Weiteren sollte eine bessere Erklärung möglich sein, da die entsprechenden Features, die einen Einfluss haben, nun bekannt sind.

Dennoch sollten solche Modelle nicht ohne Umsicht verwendet werden, da es strukturelle Änderungen gegeben haben kann oder Brüche mit einem Blick über den Tellerrand absehbar sind – Stichwort Corona.

Anhang

Zusätzliche Diagramme aus Data Preparation

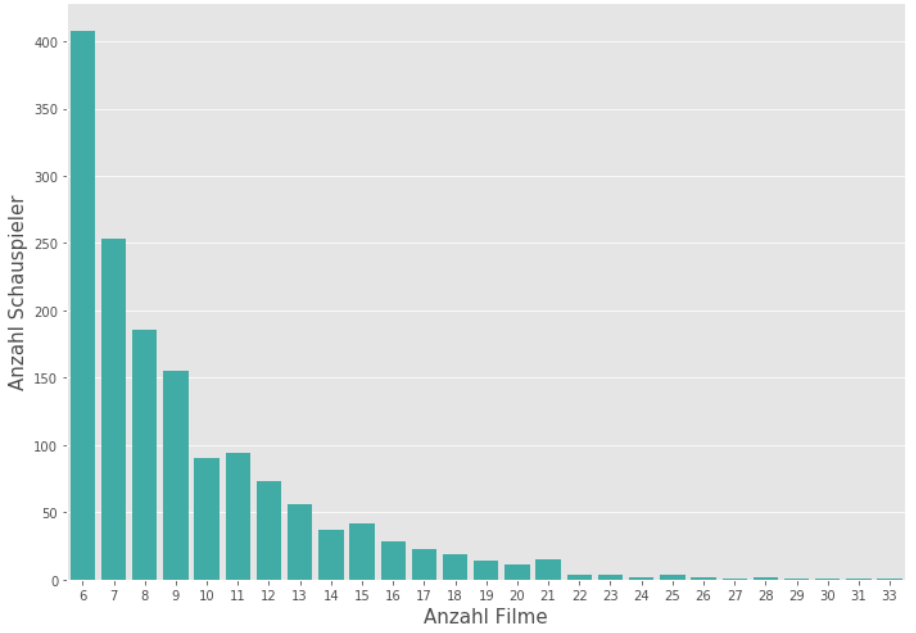

Abbildung 28: Anzahl Schauspieler mit n-Filmen (mehr als 5)

Verwendete Software

Für die Erstellung des Datensatzes und die Daten Präparation wurde *Python 3.7.4* mit den in Tabelle 9 aufgelisteten Bibliotheken verwendet. Als Entwicklungsumgebung kam dabei *JupyterLab* in der Version *1.2.4* zum Einsatz.

Zusätzlich wurde *RapidMiner* in der Version 9.5 und R in der Version 3.6.2 verwendet.

Bibliothek	Version
matplotlib	3.1.1
nltk	3.4.5
pandas	0.25.1
plotly	4.4.1
requests	2.22.0
seaborn	0.9.0
tmdbsimple	2.2.0
wordcloud	1.6.0

Tabelle 9: Verwendete Python Bibliotheken

Bibliothek	Version
logistf	1.23
MASS	7.3-51.5
Pscl	1.5.2

Tabelle 10:Verwendete R Bibliotheken

Zusätzliche Ergebnisse aus R

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.98649  -0.43210  -0.07114   0.45620   0.89945

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.2237272  0.0645314   3.467 0.000535 ***
runtime                         0.0017741  0.0004780   3.711 0.000211 ***
release_date_weekday           -0.0221049  0.0077455  -2.854 0.004353 **
genre_drama                    -0.0430143  0.0211121  -2.037 0.041710 *
genre_comedy                    0.0976895  0.0226961   4.304 1.74e-05 ***
genre_thriller                 -0.0724292  0.0250724  -2.889 0.003900 **
genre_crime                    -0.0623646  0.0289325  -2.156 0.031214 *
genre_family                   -0.0054430  0.0367167  -0.148 0.882162
genre_horror                    0.1562512  0.0345353   4.524 6.33e-06 ***
num_cast                        0.0006955  0.0004949   1.405 0.160062
cast_name_scarlett_johansson    0.0702918  0.1215289   0.578 0.563048
cast_name_don_cheadle           0.2547564  0.1604669   1.588 0.112501
cast_name_benedict_cumberbatch  0.1816983  0.1155104   1.573 0.115841
cast_name_stan_lee              0.1862237  0.1989337   0.936 0.349305
cast_name_jess_harnell          0.0064870  0.1539200   0.042 0.966386
cast_name_danny_mann            0.0867988  0.1731498   0.501 0.616208
cast_name_frank_welker          0.1060680  0.1215025   0.873 0.382761
cast_name_conrad_vernon         0.3261240  0.1812544   1.799 0.072095 .
cast_name_hugh_jackman          0.2415150  0.1317487   1.833 0.066896 .
crew_name_stan_lee             -0.0498526  0.1682754  -0.296 0.767059
crew_name_kevin_feige           0.2389636  0.3837029   0.623 0.533482
crew_name_daniel_sudick         0.1545009  0.2092213   0.738 0.460304
crew_name_dave_jordan          -0.0430518  0.1308052  -0.329 0.742085
crew_name_victoria_alonso      -0.6865105  0.6651915  -1.032 0.302146
crew_name_dan_oconnell         -0.2246178  0.1542841  -1.456 0.145551
crew_name_andy_nelson           0.2059391  0.0991271   2.078 0.037852 *
crew_name_gwendolyn_yates_whittle  0.1648553  0.1139779   1.446 0.148193
crew_name_louis_desposito       0.3095346  0.5826498   0.531 0.595289
crew_name_andy_park             0.1175576  0.2669822   0.440 0.659743
crew_name_luca_marco_paracels   0.2282565  0.1401521   1.629 0.103513
crew_name_luke_dunn_gielmuda    0.1193478  0.1583098   0.754 0.450986
crew_name_john_lasseter         0.0084214  0.1676807   0.050 0.959949
crew_name_tom_macdougall        0.0068674  0.2163835   0.032 0.974684
crew_name_wade_culbreath        0.1212576  0.1700056   0.713 0.475752
crew_name_john_t_cucci          0.2891783  0.1612832   1.793 0.073093 .
crew_name_ethan_van_der_ryn     0.1627761  0.1262816   1.289 0.197517
crew_name_rj_kizer              0.1128604  0.1479319   0.763 0.445580
crew_name_gary_rizzo            0.1511789  0.1225406   1.234 0.217426
crew_name_susan_dawes           0.1751613  0.1725492   1.015 0.310136
crew_name_leslee_feldman       -0.4173740  0.3763675  -1.109 0.267554
crew_name_christi_soper         0.4739723  0.3910076   1.212 0.225554
jobs_animation                 -0.0025341  0.0020737  -1.222 0.221824
jobs_editor                    -0.0053921  0.0146170  -0.369 0.712237
jobs_art_direction              0.0055501  0.0083848   0.662 0.508079
department_num_crew             0.0057101  0.0015960   3.578 0.000353 ***
department_num_directing        0.0143557  0.0064789   2.216 0.026796 *
department_num_editing          0.0155288  0.0064712   2.400 0.016481 *
department_num_lighting         0.0066307  0.0033112   2.002 0.045338 *
department_num_sound            0.0057032  0.0023312   2.446 0.014494 *
department_num_visual_effects   0.0022106  0.0012696   1.741 0.081772 .
department_num_writing          0.0096185  0.0056480   1.703 0.088688 .
num_crew                       -0.0033077  0.0008032  -4.118 3.94e-05 ***
num_keywords                    0.0070205  0.0019700   3.564 0.000372 ***
keyword_duringcreditsstinger    0.0245980  0.0357887   0.687 0.491948
keyword_woman_director         -0.0978883  0.0405339  -2.415 0.015806 *
keyword_sequel                  0.1669513  0.0441944   3.778 0.000162 ***
keyword_3d                      0.0192120  0.0497115   0.386 0.699181
keyword_aftercreditsstinger     0.0480296  0.0542428   0.885 0.375993
keyword_based_on_comic         -0.0142254  0.0873339  -0.163 0.870622
lang_ru                        -0.1286354  0.0541356  -2.376 0.017566 *
lang_fr                        -0.2959231  0.0653496  -4.528 6.22e-06 ***
lang_ta                         0.3743690  0.0704256   5.316 1.15e-07 ***
lang_ko                         0.4342834  0.0808137   5.374 8.40e-08 ***
has_homepage                    0.0603392  0.0205328   2.939 0.003326 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2185355)

    Null deviance: 657.41  on 2629  degrees of freedom
Residual deviance: 560.76  on 2566  degrees of freedom
AIC: 3529.1
```

Abbildung 29: Zusammenfassung glm

Literaturverzeichnis

Larose, D. T., Larose, C. D. (2015): Data Mining and Predictive Analytics, John Wiley & Sons, 2015

Kim, M. H. (2013): Determinants of revenues in the motion picture industry, in: Applied Economics Letters, 20.11, S. 1071-1075, 2013

Seiter, M. (2017): Business Analytics: Effektive Nutzung fortschrittlicher Algorithmen in der Unternehmenssteuerung, Vahlen, 2017

Wirth, R., Hipp, J. (2000): CRISP-DM: Towards a Standard Process Model for Data Mining, in: 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, S. 29-29, 2000

Ahmed, U., Waqas, H., Afzal, M. T. (2019): Pre-production box-office success quotient forecasting, in: Soft Computing, S. 1–19, 2019

De Vany, A., Walls, W. D. (1999): Uncertainty in the movie industry: Does star power reduce the terror of the box office? in: Journal of Cultural Economics, 23.4, S. 285–318, 1999

Menard, S. (2002): Applied logistic regression analysis, SAGE Publications, 2002

Internetquellen

U.S. Bureau of Labor Statistics (2019): Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCNS], https://fred.stlouisfed.org/series/CPIAUCNS (2019) [Zugriff 2019-12-22]

The SciPy community (2019): numpy.log1p, https://docs.scipy.org/doc/numpy/reference/generated/numpy.log1p.html (2019-07-26) [Zugriff 2019-12-22]

United Nations (2019): World Population Prospects 2019: Sex Ratio of Total Population, https://population.un.org/wpp/Download/Standard/Population/ (2019) [Zugriff 2019-12-22]

RapidMiner (2019): Generalized Linear Model, https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/functions/generalized_linear_model.html (2019) [Zugriff 2020-01-03]

Steinitz, D. (2020): Programmierte Erfolge: Hollywood setzt bei der Filmauswahl auf künstliche Intelligenz, https://www.sueddeutsche.de/politik/kino-programmierte-erfolge-1.4755081 (2020-01-13) [Zugriff 2020-02-17]

kostenloser Download
unter **fom-ifes.de**

Kladroba, A. (2019): Der Einfluss mathematischer Methoden auf das Ergebnis von Mannschaftswettkämpfen: Eine Simulationsrechnung, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 20, 2019, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-416-9

Raasch, A. / Lehrbass, F. (2019): Investmentstrategien im Rahmen von Übernahmen börsennotierter Gesellschaften – Merger Arbitrage und Maschinelles Lernen, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 19, 2019, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-414-5

Hagemann, D. / Lehrbass, F. (2018): Prognosemodelle für Länderrisiken: Logit- und Deep Learning-Methoden im Vergleich, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 18, 2018, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-412-1

Graalmann, M.-P. / Lehrbass, F. (2018): Eignung von Varianz-Kovarianz-Ansätzen und Copula-Modellen zur Risikoaggregation in bankaufsichtlichen Risikotragfähigkeitskonzepten, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 17, 2018, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-410-7

Cox, P. / Lehrbass, F. (2018): Determinanten der Replikationsgüte von Exchange Traded Funds, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 16, 2018, ISSN (eBook) 2569-5355, ISBN (eBook) 978-3-89275-408-4

Lehrbass, F. / Scheipers, N. (2017): Determinanten der Höhe von Wirtschaftsprüfungshonoraren am Beispiel von gelisteten Unternehmen im Prime Standard, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 15, 2017, ISSN 2191-3366, ISBN 978-3-89275-406-0

Schwarz, J. (2017): Ergebnisse der Analyse von Studienabbrüchen, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 14, 2017, ISSN 2191-3366, ISBN 978-3-89275-405-3

Lehrbass, F. (2016): Risikomessung für den globalen Kohlehandel: Einfache und fortgeschrittene Verfahren nebst Backtesting sowie ein Vergleich mit IFRS 7, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 13, 2016, ISSN 2191-3366, ISBN 978-3-89275-404-6

Godbersen, H. (2016): Die Means-End Theory of Complex Cognitive Structures – Entwicklung eines Modells zur Repräsentation von verhaltensrelevanten und komplexen Kognitionsstrukturen für die Wirtschafts- und Sozialwissenschaften, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 12, 2016, ISSN 2191-3366, ISBN 978-3-89275-403-9

Seng, A. / Landherr, G. (2015): Vielfalt leben und Vielfalt gestalten – Diversity Management in der Lehre, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 11, 2015, ISSN 2191-3366, ISBN 978-3-89275-402-2

Gansser, O. A. / Schutkin, A. (2014): Studie zur Validierung der Persönlichkeitsmerkmale Abenteuerlust und Routineverhalten, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 10, 2014, ISSN 2191-3366, ISBN 978-3-89275-401-5

Gansser, O. A. (2014): Marketingplanung als Instrument zur Krisenbewältigung, in: Krol, B. (Hrsg.), ifes Schriftenreihe, Band 9, 2014, ISSN 2191-3366, ISBN 978-3-89275-400-8

Runia, P. M. / Wahl, F. / Rüttgers, C. (2013): Das Markenimage von Hersteller- und Handelsmarken: Eine empirische Analyse der Imagekomponenten von Körperpflegemarken auf der Grundlage eines Markenidentitätskonzeptes, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 8, 2013, ISSN 2191-3366

Naskrent, J. / Rüttgers, C. (2013): Sportmonitor Essen 2013: Eine empirische Analyse über das Image regionaler Sportvereine und ihre Sponsoring- und Promotionangebote, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 7, 2013, ISSN 2191-3366

Seng, A. / Fiesel, L. / Rüttgers, C. (2013): Akzeptanz der Frauenquote, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 6, 2013, ISSN 2191-3366

Naskrent, J. / Rüttgers, C. (2012): Wahrnehmung von Werbung mit Sportereignisbezug: Eine empirische Analyse der Einschätzung von Sponsoring und Ambush-Marketing im Rahmen der Fußball-Europameisterschaft und der Olympischen Spiele im Jahr 2012, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 5, 2012, ISSN 2191-3366

Seng, A. / Fiesel, L. / Krol, B. (2012): Erfolgreiche Wege der Rekrutierung in Social Networks, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 4, 2012, ISSN 2191-3366

Heinemann, S. / Krol, B. (2011): Nachhaltige Nachhaltigkeit: Zur Herausforderung der ernsthaften Integration einer angemessenen Ethik in die Managementausbildung, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 2, 2011, ISSN 2191-3366

Hermeier, B. / Rettig, P. / Krol, B. (2010): Marken- und Produktmanagement durch Nutzung von Sportgroßereignissen: Möglichkeiten und Grenzen für Industrie und Handel, in: Krol, B. (Hrsg.), KCS Schriftenreihe, Band 1, 2010, ISSN 2191-3366

ISBN (Print) 978-3-89275-417-6 ISSN (Print) 2191-3366
ISBN (eBook) 978-3-89275-418-3 ISSN (eBook) 2569-5355

Institut für Empirie & Statistik
der FOM Hochschule
für Oekonomie & Management

FOM Hochschule

FOM. Die Hochschule. Für Berufstätige.

Die mit bundesweit über 54.000 Studierenden größte private Hochschule Deutschlands führt seit 1993 Studiengänge für Berufstätige durch, die einen staatlich und international anerkannten Hochschulabschluss (Bachelor/Master) erlangen wollen.

Die FOM ist der anwendungsorientierten Forschung verpflichtet und verfolgt das Ziel, adaptionsfähige Lösungen für betriebliche bzw. wirtschaftsnahe oder gesellschaftliche Problemstellungen zu generieren. Dabei spielt die Verzahnung von Forschung und Lehre eine große Rolle: Kongruent zu den Masterprogrammen sind Institute und KompetenzCentren gegründet worden. Sie geben der Hochschule ein fachliches Profil und eröffnen sowohl Wissenschaftlerinnen und Wissenschaftlern als auch engagierten Studierenden die Gelegenheit, sich aktiv in den Forschungsdiskurs einzubringen.

Weitere Informationen finden Sie unter **fom.de**

ifes

Zunehmende Digitalisierung erfordert und ermöglicht datenbasierten Erkenntnisgewinn und fundiertes unternehmerisches Handeln. Um aus den allgegenwärtigen Daten die richtigen Schlüsse zu ziehen, ist überall eine kritische Methodenkompetenz erforderlich. Der wissenschaftliche Fokus der ifes-Akteure liegt dabei in den Bereichen der empirischen Unternehmens-, Markt- und Konsumentenforschung, der angewandten Statistik, des Data Minings und der Finanzstatistik.

Das ifes verfolgt das Ziel, empirische Kompetenzen an der FOM zu bündeln und die angewandte Forschung im empirischen Bereich der Hochschule weiter voranzutreiben. Damit nimmt das ifes eine zentrale Stellung im Bereich der Entwicklung und Unterstützung der Methodenausbildung in der Lehre der Bachelor- und Masterstudiengänge sowie im Promotionsprogramm der FOM ein.

Weitere Informationen finden Sie unter **fom-ifes.de**

SCIENCE BLOG

Im Forschungsblog werden unter dem Titel „FOM forscht" Beiträge und Interviews rund um aktuelle Forschungsthemen und -aktivitäten der FOM Hochschule veröffentlicht.

Besuchen Sie den Blog unter fom-blog.de