Digital Business University of Applied Sciences
Data Science & Business Analytics
ADS41 - ADS-04: Machine Learning
Prof. Dr. Marcel Hebing

Analyse und Vorhersage der Click-Through-Rate in Google Ads-Kampagnen: ein maschinelles Lernmodell

Studienarbeit

Eingereicht von Sabine Wildemann
Matrikelnummer 190297
Datum 8.06.2024

Zusammenfassung

Digitales Marketing hat sich in den letzten Jahren zu einem zentralen Element in den Geschäftsstrategien vieler Unternehmen entwickelt. Insbesondere die Optimierung von Online-Werbekampagnen durch datengetriebene Ansätze gewinnt stetig an Bedeutung. Vor diesem Hintergrund untersucht die vorliegende Studienarbeit, inwieweit maschinelle Lernmodelle eingesetzt werden können, um die Click-Through-Rate (CTR) und die Kosten pro Klick (Avg. CPC) in Google Ads-Kampagnen zu optimieren. Diese Aspekte sind entscheidend, da sie direkt die Effizienz und Kostenwirksamkeit von Online-Werbemaßnahmen beeinflussen.

In der Studienarbeit werden zwei Hypothesen getestet: Hypothese 1 konzentriert sich auf die Fähigkeit der Modelle, Anzeigenkombinationen zu identifizieren, die zu einer hohen CTR führen. Die zweite Hypothese prüft die Identifikation von Kombinationen, die den Avg. CPC minimieren. Für die Analyse werden zwei Algorithmen verwendet, der Random Forest und der Gradient Boosting Machines (GBM). Beide Methoden bieten eine hohe Leistungsfähigkeit in der Mustererkennung und der Vorhersage in komplexen Datensätzen.

Die Ergebnisse zeigen, dass der Random Forest-Algorithmus besonders effektiv in der Vorhersage der CTR ist, während das GBM-Modell sowohl für die CTR als auch für die Avg. CPC robuste Vorhersagen liefert. Die Modelle belegen Einflussfaktoren wie spezifische Wochentage und Ad-Gruppenzugehörigkeiten, die erhebliche Auswirkungen auf die Leistung der Kampagnen haben.

Zusammenfassend zeigt diese Arbeit, dass maschinelles Lernen genutzt werden kann, um tiefere Einblicke in die Performance von Online-Werbekampagnen zu gewinnen. Es wird empfohlen, die Modelle weiter verfeinern und um zusätzliche Datenquellen erweitern, um die Präzision der Vorhersagen zu verbessern und die Anpassungsfähigkeit zu erhöhen.

Inhaltsverzeichnis

1) Einleitung und Forschungsfrage

In der heutigen digitalisierten Wirtschaft spielt Performance Marketing eine zentrale Rolle im digitalen Marketing-Mix vieler Unternehmen. Durch zielgerichtete Werbekampagnen und die stetige Optimierung von Anzeigen und Keywords können Unternehmen signifikante Verbesserungen in der Lead-Generierung und Kundenbindung erreichen.
Die Dynamik einer zunehmend kostspieligen Werbelandschaft, neue regulatorische Rahmenbedingungen und volkswirtschaftliche Unsicherheiten zwingen Marketing-Fachleute, sich verstärkt auf leistungsorientiertes digitales Marketing zu verlassen. In einer Zeit, in der Verbraucherinnen ihre Kaufentscheidungen sorgfältig abwägen, sind effektive Strategien im unteren Bereich des Verkaufstrichters (Lower Funnel) vor allem bei B2C-Unternehmen ausschlaggebend für die Steigerung von Conversions und Umsatz.
In diesem Kontext sind Google Ads-Kampagnen neben Anzeigen auf Social Media ein mögliches Werkzeug, um relevante Zielgruppen gezielt und effizient anzusprechen.
Vor diesem Hintergrund untersucht die vorliegende Studienarbeit, inwiefern die spezifischen Merkmale von Ads innerhalb solcher Kampagnen genutzt werden können, um die Click-Through-Rate und den Avg. CPC vorherzusagen.

Die Forschungsfrage lautet:
Wie kann ein maschinelles Lernmodell zur Identifizierung und Optimierung der effektivsten Kombinationen von Anzeigentexten, Keywords, Wochentagen und Stunden in einer Google Ads Kampagne verwendet werden, um die Klickrate (CTR) zu maximieren und die Kosten pro Klick (Avg. CPC) zu minimieren?

Zur Beantwortung dieser Frage werden zwei Hypothesen formuliert und getestet:
Hypothese 1 (H1): Ein maschinelles Lernmodell kann die Kombinationen von Anzeigentexten, Keywords, Wochentagen und Stunden identifizieren, die zu den höchsten Klickraten (CTR) führen.
Hypothese 2 (H2): Ein maschinelles Lernmodell kann die Kombinationen von Anzeigentexten, Keywords, Wochentagen und Stunden identifizieren, die zu den niedrigsten Kosten pro Klick (Avg. CPC) führen.

Durch den Einsatz von Machine Learning-Algorithmen zielt diese Arbeit darauf ab, tiefere Einblicke in das Nutzerverhalten zu gewinnen, um die Effektivität der Online-Marketingstrategie steigern zu können.

2) Daten und Methoden

2.1 Datensammlung

Als Datengrundlage werden quantitative Primärdaten eines Unternehmens aus der Schweiz analysiert, die für diese Studienarbeit zur Verfügung gestellt werden. Das Unternehmen, dessen Name nicht angeführt werden soll, hat die Nutzungsrechte für diese Daten eingeräumt und erhält im Gegenzug die Ergebnisse dieser Studienarbeit.
Die Daten wurden im Zeitraum vom 13.04.2024 bis 18.05.2024 im Rahmen einer Online-Werbekampagne über Google Ads erhoben.
Die Anzeigen wurden täglich zwischen 6:30 und 23:45 Uhr CET in der Zielregion “Schweiz” und in deutscher Sprache ausgespielt.
Es handelt sich um eine Werbekampagne mit dem Kampagnenziel “Leads”, die durch einen Sign up auf einer Landing page generiert werden.
Die Kampagne ist in 7 Ad Groups aufgeteilt, die thematisch insgesamt 121 Keywords gruppieren, die zuvor über die Software SE Ranking ermittelt wurden.
In Anhang sind die ausgewählten Entitäten und Attribute dargestellt.
Die Conversions bleiben wegen unzureichender Datenqualität unberücksichtigt.

Zunächst wird versucht, die Ergebnisse bezogen auf das Verhalten der Nutzerinnen auf der Landing page zu analysieren. In einer ersten Analyse der Daten wird festgestellt, dass zu wenige Datenpunkte vorliegen. Deshalb wird der Fokus auf Erkenntnisse auf Interaktionen mit den Anzeigengruppen verlagert.

2.2 Datenzugriff

Der Zugriff auf die Daten erfolgt über ein Google Ads-Konto des Unternehmens. Die Daten haben einen aktuellen Stand und eine hohe Qualität und wurden durch die Autorin selbst extrahiert und exportiert.
Im ersten Schritt werden Daten aus .csv-Dateien und .xls-Dateien eingelesen. Aufgrund auftretender Formatierungsprobleme wird im nächsten Schritt ein Export

in das Google-Tabellen-Format durchgeführt und überflüssige Daten entfernt. Die Dateien werden dann im .csv-Format abgespeichert und erfolgreich importiert. Die finalen Reports sind im Anhang näher erläutert.

2.3 Datenintegration

Die zur Verfügung stehenden Daten bieten eine solide Grundlage zur Beantwortung der Forschungsfrage und der Testung der beiden Hypothesen.
Die im Anhang beschriebenen Reports, Entitäten und Attribute werden wie im Folgenden beschrieben analysiert und auf ihre Relevanz zur Hypothesentestung hin geprüft.

Ads per Day und Engagement Report: Daten über das tägliche Engagement bieten Einsicht in die CTR und Avg. CPC in Abhängigkeit vom Wochentag, den Headlines und Descriptions. Diese sind direkt relevant für beide Hypothesen, H1 und H2.

Hour of the Day Report: Die Analyse nach Tageszeit kann aufzeigen, wie die Nutzeraktivität und die Kosten pro Klick (CPC) über den Tag verteilt sind, was für beide Hypothesen relevant ist.
Die Metriken Impressions, Clicks, CTR bieten Einblicke in die Effektivität der Anzeigen zu verschiedenen Tageszeiten (H1). Die Kosten pro Klick/CPC nach Stunden bieten eine präzise Möglichkeit, die Wirtschaftlichkeit der Ads zu verschiedenen Tageszeiten zu bewerten (H2).

Ad Performance Report nach Assets: Diese Daten liefern eine Indikation über die Performance von “Headlines” und “Description”, da sie die Anzahl der Impressionen in der jeweiligen Ad group zeigen. Wesentliche Merkmale wie die Wortanzahl können aufschlussreich sein, um deren Einfluss auf die CTR oder den AVG. CPC zu untersuchen.

Keywords Performance Report: Diese Daten sind zentral für beide Hypothesen, da Keywords ein kritischer Faktor bei der Bestimmung der Relevanz und damit der CTR und Avg. CPC von Anzeigen sind. Keyword-spezifische Daten wie

Impressions, Clicks, CTR helfen, die Beziehung zwischen Keyword-Effektivität und CTR zu verstehen (H1).
Die Analyse der Kosten pro Klick für verschiedene Keywords hilft, die Wirtschaftlichkeit der Keyword-Auswahl zu bewerten (H2).

Zur Beantwortung der Forschungsfrage und der Hypothesen werden der Random Forest Algorithmus und der Gradient Boosting Machines Algorithmus (GBM) ausgewählt.
Der Random Forest ist ein Ensemble-Lernalgorithmus und kann sowohl numerische als auch kategoriale Daten verarbeiten. Er bietet eine Methode zur Schätzung der Wichtigkeit von Merkmalen und kann daher genutzt werden, um die Bedeutung von Wochentagen, Tageszeiten, Anzeigentexten und Keywords für die CTR und den Avg. CPC vorherzusagen.
Der Gradient Boosting Machines-Algorithmus bietet die Möglichkeit, verschiedene Hyperparameter zu optimieren, um die Modellleistung zu verbessern und ist für die Vorhersage der ‘CTR’ geeignet, da er gut mit heterogenen und nicht-linearen Beziehungen in den Daten umgehen und eine hohe Genauigkeit erreichen kann.

2.4 Explorative Datenanalyse (EDA)

Mithilfe des SweetViz Profiling-Reports werden die importierten Daten visualisiert.
Es zeigen sich im ersten Schritt zahlreiche “--” Werte. Über die Funktion ‘load_data’ werden diese als fehlende Daten (NaN) identifiziert. Dieser Schritt automatisiert die Datenbereinigung und gewährleistet, dass solche Einträge in den Data Frames sofort als NaN markiert werden, was die Konsistenz des Datensatzes für die nachfolgenden Analysephasen sicherstellt.
Der "Hour of the day"-Report zeigt zahlreiche Nullwerte, die Zeiten ohne Benutzerinteraktion darstellen. Diese Nullwerte werden in der Analyse beibehalten, um keine irreführenden Signale über die Nutzeraktivitäten zu erzeugen.
Da die Anzeigen nur im Zeitfenster 6:30 bis 23:45 geschaltet wurden, weisen die Stunden 0 bis 5 zahlreiche missing values auf, die durch “0” bzw. “0%” ersetzt werden, damit die Datenintegrität erhalten bleibt.
Im “Keywords Performance”-Report werden “Missing Values” bei verschiedenen Attributen mit Null ersetzt, wenn die zugehörigen Impressions und Clicks sehr

niedrig oder Null sind. Dies reflektiert, dass ohne Impressions keine Qualitätseinschätzung oder Benutzerengagement stattfinden kann. Durch das Ersetzen mit Null wird sichergestellt, dass das Fehlen von Interaktionen präzise im Modell abgebildet wird, ohne durch künstliche Daten verzerrt zu werden.

Die vorbereiteten Daten werden schrittweise über einen Inner-Merge auf die Entität “Ad group ID” zusammengeführt.
Im Profiling report zeigen sich für 3 Metriken dieselbe Anzahl missing values (4.318). Die fehlenden Werte werden imputiert und mit "0" bzw. 0%” aufgefüllt. Bei den fehlenden Werten für ‘Engagement-Dauer’ und ‘Engaged sessions’ ist es sinnvoll anzunehmen, dass es kein Engagement gab und bei der Engagement-Rate können fehlende Werte als 0% Engagement interpretiert werden.
Im Anschluss erfolgt der Split von Trainingsdaten und Testdaten im Verhältnis 80/20 und der Export im .pickle-Format.
Nach dem Import des Trainingsdatensatzes werden die Features ausgewählt und in numerische und kategoriale aufgeteilt und die Zielvariablen (‘CTR’ und ‘Avg. CPC’) festlegt.

2.5 Feature Engineering

Spezifische Spalten, die Prozentwerte enthalten ('% Engaged sessions (GA4)', 'Interaction rate'), werden bereinigt, indem das Prozentzeichen entfernt und der verbleibende Wert in einen numerischen Typ konvertiert wird. Dies stellt sicher, dass die Modelle mit korrekten numerischen Werten arbeiten können.

Alle numerischen Features werden mittels verschiedener Skalierer (StandardScaler, MinMaxScaler, RobustScaler) standardisiert bzw. normalisiert, sodass alle numerischen Daten auf einer vergleichbaren Skala liegen.
Kategoriale Features werden durch den OneHotEncoder verarbeitet und in eine binäre Matrix gebracht.
Zunächst wird versucht, textbasierte Daten, wie 'Search keyword' und 'Asset' über den ‘TfidfVectorizer’ zu bearbeiten. Dieser Ansatz bringt trotz Einschränkung auf max_features=1000 kein nutzbares Ergebnis und wird daher verworfen.

Alle vorherigen Schritte werden in ColumnTransformer integriert, der die verschiedenen Transformationsprozesse für numerische und kategoriale Features kombiniert. Dieser wird dann zusammen mit den Regressionsmodellen (RandomForestRegressor, GradientBoostingRegressor) in Pipelines eingebettet. Diese Pipelines automatisieren den Prozess von der Datenvorbereitung bis zur Modellanwendung, was eine effiziente Wiederverwendung und systematische Evaluierung ermöglicht.

Nach der Transformation der Daten werden diese genutzt, um die Modelle zu trainieren und zu bewerten. Dabei wird die Modellleistung auf Trainings- und Testdaten gemessen und eine Kreuzvalidierung durchgeführt.

3) Ergebnisse

3.1 Random Forest-Modell

Der Random Forest erzielt perfekte Ergebnisse sowohl auf den Trainings- und Testdatensätzen, was potenzielle Probleme nahelegt. Die hohe Bedeutung der Features “Interaktionsrate”, “Cost”, “Clicks” und “Avg. CPC” deuten darauf hin, dass das Modell fast ausschließlich auf diese Features für Vorhersagen angewiesen ist. Diese Features werden deshalb entfernt und die Leistung des Modells erneut getestet. Außerdem wird auf mögliche Überschneidungen und fehlende Werte überprüft. Das Modell liefert nach dieser Anpassung folgende Ergebnisse für die beiden Ziel-Variablen ‘CTR’ und ‘Avg. CPC’:

Train score ‘CTR’:	0.9949331096382821
Test score ‘CTR’:	0.98473866147443
Cross-Validation Scores ‘CTR’:	0.98691169, 0.98651108, 0.98660366, 0.98576975, 0.98572793
Avg Cross-Validation Score ‘CTR’:	0.9863048215433412
Train score ‘Avg. CPC’:	0.7654983986390831
Test score ‘Avg. CPC’:	0.2878342998111084
Cross-Validation Scores ‘Avg. CPC’:	0.34769077, 0.36422311, 0.3686099, 0.38124669, 0.33618796
Avg Cross-Validation Score ‘Avg. CPC’:	0.3595916854848065

Angesichts des niedrigen Test-Scores für das ‘Avg. CPC’-Modell wird ein Hyperparameter-Tuning aufgesetzt, um die Modellleistung zu verbessern. Mittels *GridSearchCV* wird die beste Parameter-Kombination gefunden.

Die Verbesserung des Test-Scores für "Avg. CPC" auf 0.5630841796249113 nach dem Tuning zeigt, dass das Modell besser in der Lage ist, die Kosten pro Klick vorherzusagen. Der Negativ MSE ("neg_mean_squared_error") ist mit -0.0847 relativ klein und zeigt, dass das Modell eine relativ genaue Vorhersageleistung aufweist.

Als die 3 wichtigsten Features werden “Day_of_week_Friday” (0.3409), “Ad_group_x_4. Verhaltensbeobachtungen” (0.1571) und “Ad_group_x_6. ADHS Beratung & Coaching” (0.0644) ermittelt.

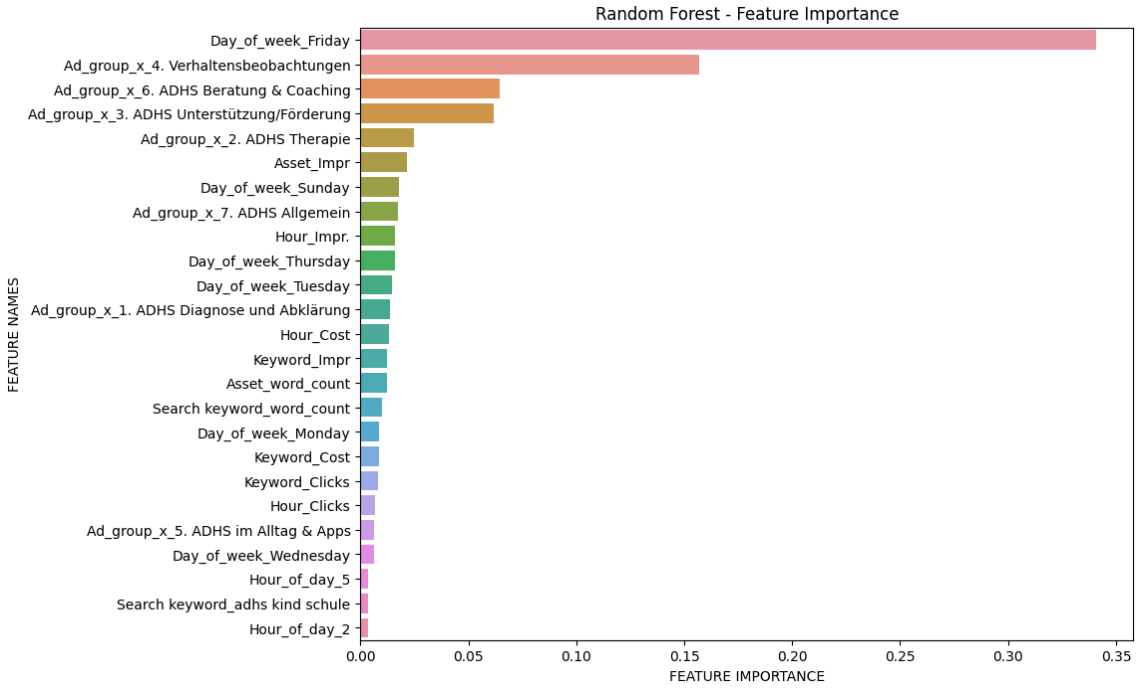

Abb. 1: Feature Importance der 25 wichtigsten Features des Random Forest-Modells

3.2 Gradient Boosting-Modell

Das Gradient Boosting-Modell zeigt ebenfalls starke Vorhersagen für die ‘CTR’ und bietet eine solide Basis zur Generalisierung auf neuen Daten. Bestimmte Tage wie Sonntage und Samstage stellen sich als besonders relevant für hohe Klickraten heraus, was auf veränderte Nutzerverhaltensmuster am Wochenende schließen lässt. Die Ergebnisse unterstützen somit Hypothese 1.

Die Vorhersage des ‘Avg. CPC’ ist genauer als beim Random Forest-Modell, was Hypothese 2 unterstützt, jedoch gibt es auch hier Raum für Verbesserungen.

Die Scores für Trainings- und Testsätze liegen recht nah beieinander, was ebenfalls eine gute Generalisierungsfähigkeit des Modells anzeigt.
Das Modell liefert folgende Ergebnisse:

Train score ‘CTR’:	0.9261082557385405
Test score ‘CTR’:	0.9261142764713529
Train Score ‘Avg. CPC’:	0.6379621707937067
Test Score ‘Avg. CPC’:	0.6394798707350944

Die Top 3 Features für ‘CTR’ sind “Day_of_week_Sunday” (0.168), “Ad_group_x_1. ADHS Diagnose und Abklärung” (0.1163) und “Day_of_week_Saturday” (0.1133).
Für den ‘Avg. CPC’ werden “Day_of_week_Friday” (0.4397), “Ad_group_x_4. Verhaltensbeobachtungen” (0.2119) und “Ad_group_x_3. ADHS Unterstützung/Förderung” (0.1202) ermittelt.

Die Ergebnisse aus der Modellierung mit dem Random Forest und Gradient Boosting-Modell bestätigen signifikante Einflussfaktoren, die sowohl die Click-Through-Rate als auch die durchschnittlichen Kosten pro Klick beeinflussen. In Bezug auf H2 zeigt sich, dass bestimmte Ad-Gruppen und Wochentage signifikant zur Reduzierung der durchschnittlichen Kosten beitragen. Insbesondere wurde festgestellt, dass Anzeigen, die an Wochenenden geschaltet werden, tendenziell geringere Kosten pro Klick verursachen, was auf ein höheres Nutzerengagement bei geringeren Wettbewerbskosten zurückzuführen sein könnte.

Diese Erkenntnisse sind direkt relevant für die Forschungsfrage, da sie zeigen, wie spezifische Merkmale von Ads innerhalb der Kampagnen optimiert werden können, um die Effizienz zu maximieren. Es wird deutlich, dass die Kombination aus bestimmten Tageszeiten, Wochentagen und spezifischen Ad-Gruppen die Kosten signifikant beeinflussen kann, was Marketern ermöglicht, ihre Strategien entsprechend anzupassen und somit die Rentabilität ihrer Kampagnen zu steigern.

4) Diskussion und Handlungsempfehlungen

Aus der Analyse der Daten und der Modellvorhersagen wird deutlich, dass eine gezielte Anpassung der Anzeigenausrichtung, insbesondere durch die Optimierung von Einsatzzeiten und die Fokussierung auf effektive Ad-Gruppen, das Potenzial hat, die Werbekosten signifikant zu senken und die Interaktion mit den Zielgruppen zu maximieren.

Für die zukünftige Kampagnenplanung empfiehlt es sich, die Schaltung von Anzeigen an Wochentagen mit nachweislich niedrigeren Avg. CPC zu intensivieren und die Budgetallokation für Zeiten hoher Nutzeraktivität zu optimieren. Um die Ergebnisse weiter zu vertiefen, kann detaillierter analysiert werden, wie die Anzeigenleistung durch die Kombination spezifischer Features wie Tageszeit, Keyword-Effektivität und Ad-Gruppen-Strategien beeinflusst wird. Dies kann durch Segmentierung der Daten oder durch Einsatz von Analysetechniken wie Interaktions- oder Polynommerkmale in der Modellbildung erreicht werden.

Gewinnbringend wäre zudem die Hinzunahme von Daten zu Conversions, um die vollständige Leistung der Anzeigen messen zu können.
Eine weitere Verbesserung in der Analyse kann durch die Erweiterung des Betrachtungszeitraums erreicht werden, um auch saisonale Effekte zu berücksichtigen. Durch gezielte A/B-Tests kann weiterhin untersucht werden, welche Elemente der Anzeigen den größten Einfluss auf die CTR und Avg. CPC haben.

Anhänge

Das Projekt ist in Github unter folgender URL abgelegt: https://github.com/pluzgi/studienarbeit-IV-wildemann.git

Begleitend zu dieser Studienarbeit wurden folgende Dateien erstellt:
- Jupyter Notebooks (01_import_02_eda_03_statistics.ipynb, 04_ml.ipynb")
- zip-Datei mit allen Dateien ("ADS-04_Hausarbeit_Wildemann.zip")

Datensätze

Hinweis: Die Abkürzung "Impr." steht für "Impressionen"

1. Ads per Day and Engagement (Report_Ads-per-day.csv)

Impressions und Clicks: Diese Daten bestimmen direkt die CTR und bieten Einblicke in die Nutzerinteraktionen mit den Anzeigen.

Kosten und CPC: Diese Metriken sind entscheidend für Hypothese H2, da sie direkt die Wirtschaftlichkeit der Kampagnen widerspiegeln. Durch die Analyse der Kosten und des CPC können wir verstehen, welche Faktoren zu höheren oder niedrigeren Kosten pro Klick führen.

Engagement und Interaktionsrate: Diese zusätzlichen Metriken bieten Einblicke in das Nutzerverhalten und die Reaktionen auf die Anzeigen, was indirekt zur Optimierung der Kampagnen beitragen kann.

Entitäten: Ad group, Ad group ID, Ad, Ad ID, Day of the week

Attribute: Impr., Clicks, CTR, Avg. engagement duration per session (seconds) (GA4), Interaction rate, % Engaged sessions (GA4), Events / session (GA4), Cost, Avg. CPC

2. Hour of the Day Report (Report_Hour-of-the-day.csv)

Nutzung: Untersuchung von Nutzerverhalten basierend auf Wochentag und Tageszeit auf AdGroup-Ebene

Entitäten: Ad group, Ad group ID, Day of the week

Attribute: Hour of the day, Impr., Clicks, CTR, Cost, Avg. CPC, Interaction rate

3. Ad Performance Report nach Assets (Report_AdGroup_AssetDetails.csv)
Nutzung: Leistungsanalyse auf Asset-Ebene
Entitäten: Ad group, Ad group ID
Attribute: Asset, Asset type, Impr.
Der „Asset type“ klassifiziert die Assets weiter in Kategorien “Headline” und “Description”.

4. Keywords Performance Report (Report_SearchKeyword.csv)
Keywordspezifische Daten, wie Impressions, Clicks, CTR helfen, die Beziehung zwischen Keyword-Effektivität und CTR zu verstehen (H1).
Die Analyse der Kosten pro Klick für verschiedene Keywords hilft, die Wirtschaftlichkeit der Keyword-Auswahl zu bewerten (H2).
Entitäten: Search keyword, Ad group, Ad group ID, Quality Score
Attribute: Impr., Clicks, CTR, Cost, Avg. CPC, Impr. (Abs. Top) %, Impr. (Top) %, % Engaged sessions (GA4), Avg. engagement duration per session (seconds) (GA4)