Hochschule Anhalt

Anhalt University of Applied Sciences

Programmierung für Data Science:

Marketing Campaign Projekt

Qais Doofesh

5050068

Prof. Dr. Michael Cebulla

WiSe 2021/2022

Einführung

Auf der Suche nach einem interessanten aber herausfordernden Datensatz, um meine Python und Daten Analyse Kenntnisse sowohl zu testen als auch zu erweitern, bin ich auf einen Marketing Datensatz unter https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign gestoßen.

Da handelt es sich um Marketingkampagnen. Der Datensatz besteht also aus Kunde Informationen (zunächst 2241 einzigartige Kunden) mit den zusammen verbundenen Eigenschaften wie Alter, Bildungsniveau und unterschiedliche Metriken von Konsumverhalten.

Der Datensatz besteht zunächst aus 29 Spalten bzw. Variablen und zwar:

1- AcceptedCmp1 bis 5 – Ob der Kunde das entsprechende Marketingangebot akzeptiert hat (1: Ja, 0: Nein)
2- Response (target) - Ob der Kunde das letzte Marketingangebot akzeptiert hat (1: Ja, 0: Nein)
3- Complain – Ob der Kunde sich in den letzten 2 Jahren beschwert hat (1: Ja, 0: Nein)
4- DtCustomer – Datum der Eintragung des Kunden
5- Education - Das Bildungsniveau des Kunden
6- Marital – Der Familienstand des Kunden
7- Kidhome – Die Anzahl an Kinder zu Hause
8- Teenhome - Die Anzahl an Jugendliche zu Hause
9- Income - Das Einkommen des Kunde (Jährlich)
10- MntFishProducts - Ausgabe für Fischprodukte in den letzten 2 Jahren
11- MntMeatProducts - Ausgabe für Fleischprodukte in den letzten 2 Jahren
12- MntFruits - Ausgabe für Obst in den letzten 2 Jahren
13- MntSweetProducts - Ausgabe für Süßigkeiten in den letzten 2 Jahren
14- MntWines - Ausgabe für Wein in den letzten 2 Jahren
15- MntGoldProds - Ausgabe für Gold in den letzten 2 Jahren
16- NumDealsPurchases – Anzahl an Käufe mit Rabatt
17- NumCatalogPurchases - Anzahl an Käufe mit einem Katalog
18- NumStorePurchases - Anzahl an Käufe im Laden
NumWebPurchases - Anzahl an Käufe auf der Webseite
19- NumWebVisitsMonth - Anzahl an Besuche der Webseite pro Monat
20- Recency - Anzahl an Tage seit dem letzten Kaufen
21- ID - Die Kennnummer des Kunden
22- Year_Birth - Geburtsjahr des Kunden
23- Z_CostContact - Nicht im Scope des Projekts
24- Z_Revenue - Nicht im Scope des Projekts

Herangehensweise

Mit solchen verfügbaren Informationen, kann man die Kundenverhalten analysieren, vorhersagen und unser hypothetische Laden ermöglichen, Datengetrieben zu werden, in dem es identifiziert werden kann, welche Kunden mit welche Kampagnen bzw. Produktangebote ins Visier genommen werden sollen.

Es werden also folgender Herangehensweise gefolgt:

1- Data Exploration: Identifizierung von Outliers und Duplikate und die Exploration der Variablen anhand Matplotlib Diagrammen.
2- Data Cleansing and Prepping: Die Aufbereitung des Datensatzes durch die Entfernung von Outliers, Duplikatte und Variablen, die keinen Mehrwert haben.

3- Data Analysis: Die Untersuchung der Zusammenhänge zwischen den Variablen und die Darstellung davon anhand Matplotlib Diagrammen.
4- Multiple Linear Regression: Die Erstellung eines Multiple Linear Regression Modells, um die gesamte Ausgaben (Zielattribut, erstellt durch die Summierung alle Ausgaben Spalten) vorhersagen zu können basierend auf die ausgewählte Attribute Year_Birth, Education, Marital_Status, Kidhome, Teenhome, Income, MntTotal.

Data Exploration

In diesem Schritt wird hauptsächlich jede Spalte untersucht anhand Matplotlib Histogram oder Scatter Plots :

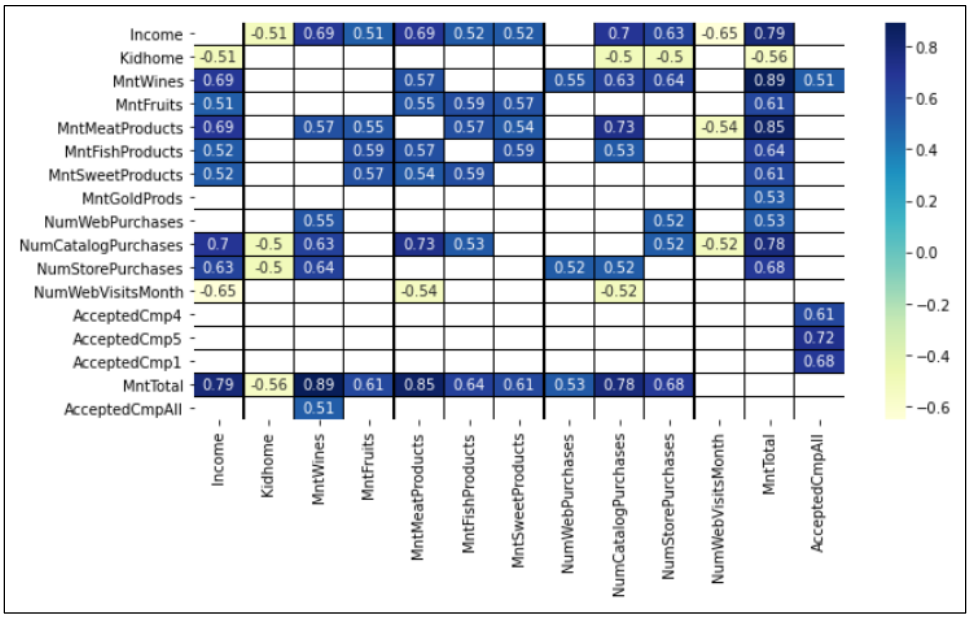
- ID hat keine Duplikate
- Year_Birth enthält Werte, die ca. 1900, was sehr unwahrscheinlich in einem Kundendatensatz zu finden ist
- Education enthält die typischen Bildungsniveaus
- Marital Status enthält zwar die typischen Familienstände aber auch sinnlose Werte bzw. Kategorien wie z.B. Absurd, Alone und Yolo
- Income hat ein einziges deutliches Outlier und zwar 666666. Der Rest der Werte liegen unter 200000
- Kid Home und Teen Home haben Werte zwischen 0, 1 oder 2
- Dt_Customer besteht aus Daten zwischen 2012 und 2014 und allen Tagen und Monaten
- Recencey hat keine Auffälligkeiten und schwankt zwischen 0 und 100 Tagen
- Die Ausgabe Spalten und die der Anzahl an Käufe waren hauptsächlich “Right-Skewed”; Häufigkeit ist hoch bei niedrigen Werten und steigt ab mit ansteigenden Werten
- AcceptedCmp 1 bis 5 waren entweder 0 oder 1

Weitere Spalten hatte keine Rollen in diesem Scope gespielt.

Data Analysis

A- Pearsons Correlation

Um die Correlation zwischen jeder Spalte zu identifizieren, wird Pearson’s correlation verwendet. Zur Vereinfachung der daraus entstehenden Tabelle, werden nur spezifische Spalten betrachtet, deren Correlation mehr als 0,5 und kleiner als -0,5 (signifikante Correlation).



Pearson's Correlation (Dunkle Farbe: Starke positive Correlation, Helle Farbe: Starke negative Correlation)

Als Beispiel kann man sehen, dass Income eine starke Correlation mit MntTotal (Alle Ausgaben summiert) während Acceptedcmpall (Summieren aller Cmp Spalten) eine positive Correlation mit der Ausgabe für Wein hat.

B- Pivoting und Visualisierung

Als Nächstes wird der Zusammenhang zwischen den aussagekräftigsten Variablen Year_Birth, Education und Marital_Status und jede der 3 Hauptgruppierungen von Variablen:

		AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2	Ob Kunde die entsprechende Marketingkampagne akzeptiert 0 oder 1 (Ja oder Nein)
Year_Birth	*Geburtsjahr*	NumDealsPurchases NumWebPurchases NumCatalogPurchases NumStorePurchases NumWebVisitsMonth	Anzahl an Online-Käufe, Laden-Käufe, Webung-Käufe usw... Numerisch
Education	*Bildungsniveau*		
Marital_Status	*Familienstand*		
		MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts MntGoldProds	Ausgaben für entsprechende Produktkategorie Numerisch

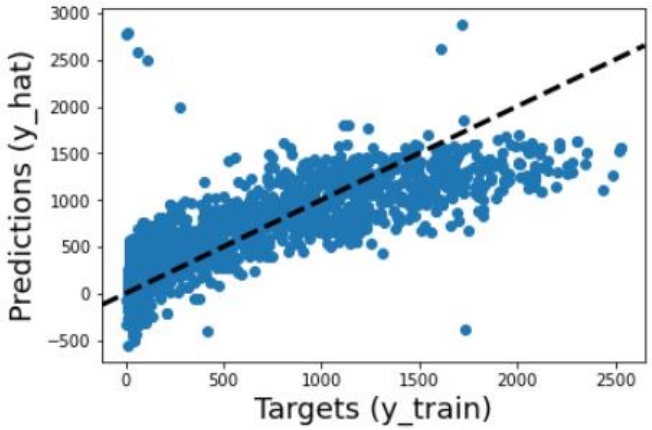
Das heißt für jede der 3 Variablen:

- Die Dataframe wird gruppiert (Pivoting), um die Werte aggregieren zu können (Mittelwert).
- Die Pivot wird mit Farben codiert mithilfe einer Heatmap
- Für jede Hauptgruppierung von Variablen wird ein Line Plot erstellt.

Multiple Linear Regression

um die gesamten Ausgaben (Zielattribut, erstellt durch die Summierung alle Ausgaben Spalten) vorhersagen zu können basierend auf die ausgewählte Attribute Year_Birth, Education, Marital_Status, Kidhome, Teenhome,

Income, MntTotal, wird ein Multiple Linear Regression angewendet. Nach weiteren Datenaufbereitung, Skalierung, Train-Test-Split und dem Training des Modells, wird ein gutes Modell erreicht, dessen Leistung wie folgt aussieht:



Die Targets (Zielattribut) und die Vorhersage davon sollen übereinstimmen, d.h. der Plot soll im idealen Fall eine 45 Grad Linie. Es ist festzustellen, dass diese Modell bei niedrige Werte eine bessere Leistung hat, da die dargestellte Punkte mehr geclustert und auf der 45 Grad Linie sind.

Conclusion

Mit diesen Insights, mithilfe der Data-Mining, kann der Betrieb datengetriebene Entscheidungen treffen, um die Prozesse zu optimieren, Marketingmöglichkeiten besser zu schöpfen und folglich die Kundenzufriedenheit zu erhöhen. Die Insights stellen dar, welche Rollen das Bildungsniveau, Alter und Familienstand beim Einkaufen spielen. Dadurch kann der Betrieb den Vorteil gewinnen, Ressourcen besser zuzuweisen und auf den richtigen Kunden mit dem richtigen Produkt und der richtigen Kampagne fokussieren.