Digital Business University of Applied Sciences

Data Science & Business Analytics

17.DSBA ADS-04: Machine Learning

Prof. Dr. Marcel Hebing

Analyse des Superstore Datensatz

Studienarbeit

Eingereicht von Nico Bauer

Matrikelnummer 200003

Datum 19.12.2021

Inhaltsverzeichnis Seite 2

1. Executive Summary

Im Rahmen der Studienarbeit wurde mit Hilfe des Superstore Datensatz, der von Tableau zur Verfügung gestellt wird, gearbeitet. Der Datensatz beinhalten neben den verkauften Artikeln und deren Klassifizierung, wichtige Versanddaten wie zum Beispiel die Versandkosten, -dauer und die Versandart. Anhand diverser Analysen konnte festgestellt werden, dass zwischen den geprüften Daten und der Zielvariable Dauer keine, beziehungsweise nur geringe Zusammenhänge bestehen. Die Zielgröße Dauer ist nicht nur abhängig von einer der geprüften Variablen, sondern ergibt sich aus der Kombination verschiedener Variablen.

2. Einleitung

Die nachfolgende Studienarbeit befasst sich mit dem Superstore Datensatz, der von Tableau zur Verfügung gestellt wurde. In der Studienarbeit wird aufgezeigt, ob und wenn Ja, welcher Zusammenhang zwischen der Dauer vom Bestelleingang bis hin zum Versand und anderen, gegebenen Variablen wie zum Beispiel der Versandart, der Lieferpriorität und der Versandkosten besteht.

2.1 Datenanalyse des Superstore Datensatz in Bezug auf die Versanddauer

2.1.1 Hypothesen zur Versanddauer

Je höher die Versandkosten eines Artikels, desto schneller wird er nach dem Bestelleingang versendet.

2.1.2 Datenimport und Datentransformation

Bevor mit der Analyse und Bearbeitung des Datensatz begonnen wird, benötigt man verschieden, hilfreiche Python und Scikit-learn Pakete damit die Analysen und Algorithmen im Anschluss fehlerfrei ausgeführt werden können. Hierzu zählen unter anderem der DecisionTreeClassifier von Scikit-learn oder das Python Paket xlrd um den Datensatz als Excel Datei einlesen zu können. Um die Ergebnisse der Analysen im Anschluss visualisieren zu können wird zusätzlich matplotlib.pyplot geladen.

Nachdem die Pakete in Python geladen wurden, kann mit dem Einlesen der XLS-Dateien aus dem Input Ordner gestartet werden. Nachdem die Datei nun als Pandas Dataframe zur Verfügung steht, kann diese Datei direkt bearbeitet werden. Da es jedoch von Vorteil ist, die Rohdatei nicht abzuändern, werden diese im nächsten Schritt in einen neuen Dataframe kopiert. Mit dieser Kopie kann nun weitergearbeitet werden, ohne dass die Rohdaten verändert werden.

Im ersten Schritt der Datenanalyse ist es wichtig, dass die vorgegebenen Daten genau untersucht werden. Bei meiner Analyse fiel mir auf, dass ich für meine Zielvariable, welche die Differenz aus dem Datum des Bestelleingangs und des Versanddatum ist, kein direkter Wert vorliegt. Daher habe ich eine neue Spalte im Dataframe hinzugefügt, welche mir genau diesen Wert berechnet. Da dieser neu geschaffene Wert dem Objekttype „timedelta“ entspricht, wandelte ich diesen direkt in den Objekttype „float“ um, damit bei der Analyse einfacher damit gearbeitet werden kann. Somit wurde die Zielvariable, Dauer, die analysiert werden soll, dem Dataframe hinzugefügt.

Bevor mit der eigentlichen Datentransformation gestartet werden kann, muss der neu generierte Dataframe in einen Train- und einen Testdatensatz gesplittet werden. In der Regel werden die Daten im Verhältnis 80 zu 20 gesplittet. Mit Hilfe verschiedener Befehle kann man sich ein genaueres Bild der beiden aufgesplitteten Dataframes machen. Es können unter anderem die Anzahl von Spalten und Zeilen angezeigt werden oder man lässt sich die ersten 5 Datensätze komplett ausgeben. Da ich für die Analyse meine Zielvariable wissen wollte, wie viele verschiedene Einträge es in den Spalten Order Priority, Ship Mode und Region gibt, habe ich mir diese Daten anzeigen lassen.

Nachdem sich ein erster Überblick über die einzelnen Datensätze verschafft wurde, kann mit der ersten Datentransformation begonnen werden. Hierbei wurden zuerst die Spalten festgelegt, welche aus dem Datensatz entfernt werden können, da sie für die weitere Analyse nicht relevant sind. Des Weiteren wurde die Zielvariable, alle nummerischen Spalten und die Spalten mit Float Werten festgelegt. Anschließend wurden verschiedene Befehle definiert, welche auf den Trainingsdatensatz angewendet werden können. Ziel dieser Befehle war es, die Spalten Order Priority, Ship Mode und Region zu Normalisieren. Dies bedeutet, dass alle gleichen Werte mit derselben Nummer überschrieben werden. Nachdem alle Befehle ausgeführt

wurden, beinhaltet der Trainingsdatensatz nun ausschließlich nummerische Spalten. Dies kann mit dem „Info-Befehl“ auch noch einmal überprüft werden. Diese Werte lassen sich wie in Abbildung 1 – 4 visualisiert darstellen.

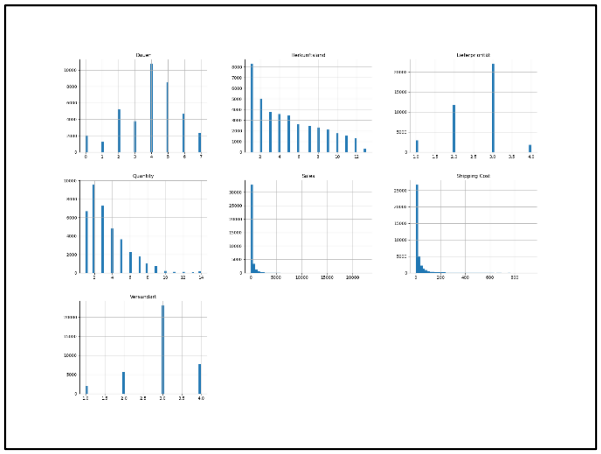

Abbildung 1 - Histogramm

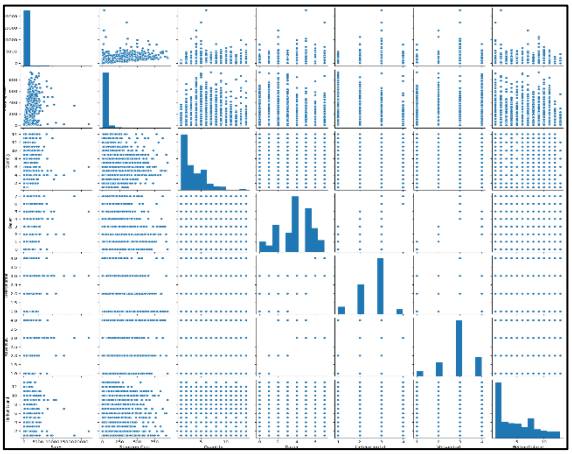

Abbildung 2 – Pair Plot

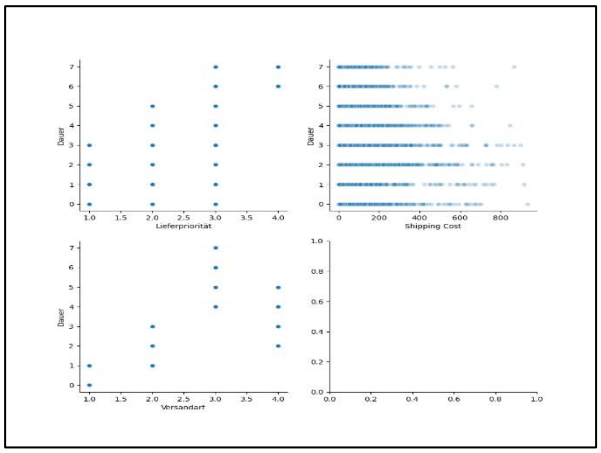

Abbildung 3 - Scatterplot

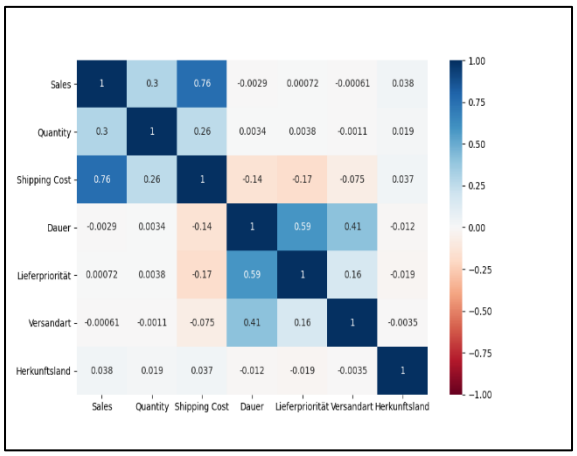

Abbildung 4 - Heatmap

Im zweiten Schritt der Datentranformation werden mit Hilfe der Scikit-learn Pakete sogenannte Pipelines erstellt, welche im Anschluss auf den Trainingsdatensatz angewendet werden. Mit dem StandardScaler lassen sich beispielsweise Werte im Datensatz standardisieren. Mit dem ColumnTransformer, der ebenfalls Bestandteil der scikit-learn Python Machine Learning-Bibliothek ist, lassen sich bestimmte Transformationen auf einzelne nummerische Spalten und parallel auf eine andere kategorische Spalte vornehmen. Die Erstellung der verschiedenen Pipelines dient als Vorbereitungsschritt für die im Anschluss durchgeführten Analysen. Abschließend werden neue Objekte, in diesem Fall der X_train erzeugt, auf den die erstellten Pipelines angewendet werden. Parallel dazu wird ein weiteres Objekt, y_train

erstellt, welches die Zielvariable des Trainingsdatensatz beinhaltet. Mit diesen beiden Objekten wird im späteren Verlauf das Machine Learning angewandt.

2.1.3 Datenanalyse

Bei der Datenanalyse wurden verschieden Analysemöglichkeiten aus der Scikit-learn Bibliothek genutzt. Als erste wurde der Datensatz auf den das Machine Learning angewendet wurde mit Hilfe einer Clusteranalyse analysiert. Wie man aus der Abbildung 6 ablesen kann, lassen sich die Werte nicht in Cluster unterteilen. Dies erkennt man daran, dass die Werte deutlich kleiner 1 sind und somit nicht miteinander korrelieren.

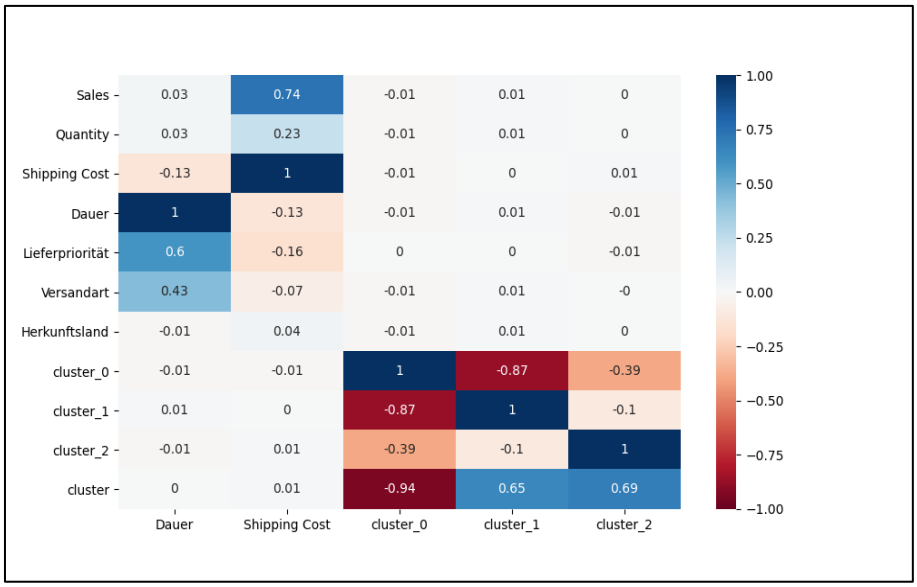

Abbildung 5 – Cluster

Nachdem mit Hilfe der Clusteranalyse die Hypothese nicht bestätigt werden konnte, wurde in der zweiten Analyse der DecisionTree Algorithmus auf die Datensätze angewendet. Hierbei wurde Entscheidungsbäume mit verschiedenen Tiefen erstellt, um herauszufinden, auf welcher Ebene die Daten am aussagekräftigsten sind. In meinem Fall war dies bei einer Tiefe von 3 Ebenen gegeben. Dieser Entscheidungsbaum wurde anschließend zum besseren Verständnis wie in Abbildung 6 visualisiert.

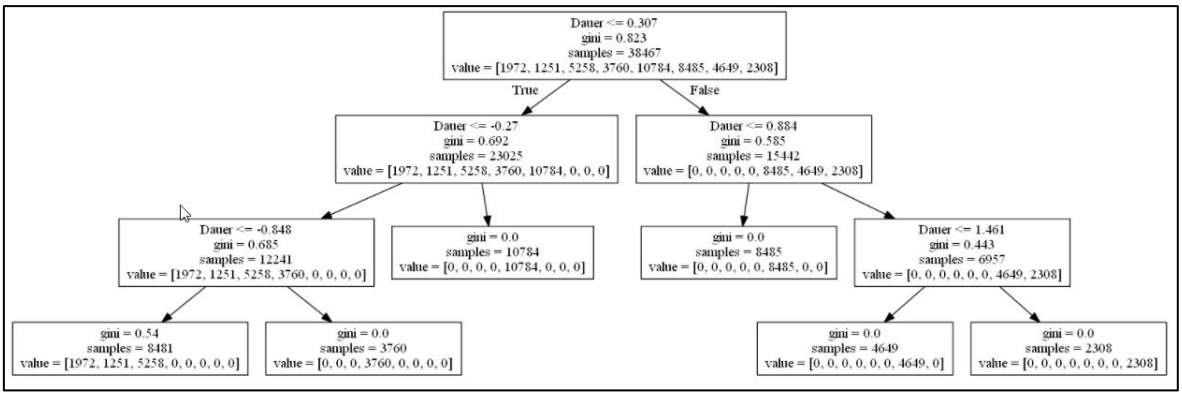

Abbildung 6 - DecissonTree

Als letzte Analyse wurde die OLS Regression auf den Datensatz angewandt. Dabei musste im ersten Schritt herausgefunden werden, welche Werte berücksichtigt werden müssen, damit eine hoher Ausprägung gegeben ist. Alle geprüften Variablen zeigten zwar eine Abhängigkeit an, jedoch war diese vor allem bei den Kombinationen Dauer und ShippingCost, Lieferpriorität und Versandart sehr gering und daher nicht aussagekräftig. Lediglich bei der Kombination aus den oben genannten Werten mit deren quintierten Werten konnte ein linearer Zusammenhang festgestellt werden. Dadurch wurde klar festgestellt, dass due Variable Dauer nicht nur von einer Variablen abhängig ist, sondern von mehreren.

3.Fazit

Zusammengefasst lässt sich sagen, dass die Variable Dauer nicht allein von einer der anderen, gegebenen Variablen abhängig ist. Die Kombination der verschiedenen Variablen hat Einfluss auf die Höhe der Dauer zwischen dem Eingang der Bestellung und dem Versand des Artikels. Empfehlenswert ist bei dieser Analyse daher, dass man sich zuerst die Zusammenhänge der anderen Variablen anschaut, bevor man diese im Zusammenhang mit der Dauer betrachtet.

4. Abbildungsverzeichnis