Report der "Die Elite"-Gruppe

Einführung und Ziel

Die Gruppe "Die Elite" besteht aus den Studenten Jonah Jäger, Marius Kiskemper und Marvin Vielmeyer. Das Projekt "Data-Exploration-Profit-Prediction" besitzt das Ziel, basierend auf einem Input-Vektor, der aus verschiedenen Parametern besteht, den Profit einer Transaktion vorherzusagen. Um dies zu erreichen wird ein Modell aufgesetzt, auf dem der ausgewählte Datensatz trainiert wird.

Das Ziel ist es, den betriebswirtschaftlichen Zusammenhang umzusetzen. Dieser besteht darin, dass Unternehmen unsere Software zur Profit-Vorhersage nutzen können, um herauszufinden, welche Parameter des Datensatzes eine Auswirkung auf den Profit haben.

Related Work

Zum direkten Thema der Profitvorhersage bestehen keine direkten wissenschaftlichen Publikationen. Was dies jedoch am nächsten kommt, ist die Umsetzung von Microsoft in deren Tabellenkalkulations-Programm „Excel“.

Hier gibt es als einfache Version die „Schätzer-Funktion“. In dieser ist die Syntax: =SCHÄTZER(x,Y_Werte,X_werte). Dabei markiert X die Zelle an der der Wert vorhergesagt werden soll. Die Y-Werte währen hier beispielsweise der Profitverlauf im gleichen Jahr wie X und die X-Werte im vorherigen Jahr (vgl. Quelle1). Die nächste Version ist das Prognoseblatt in Excel, welches anhand eines Anfangs- sowie Enddatums und dem Profitverlauf in dieser Zeit eine graphisch dargestellte Prognose generiert. Hier können genauere Zeitintervalle in der die Prognose dargestellt werden soll eingestellt werden (vgl. Quelle2).

Durch diese bereits verwirklichten Ansätze von Microsoft lässt sich erkennen, dass eine Profit-Prognose definitiv einen direkten betriebswirtschaftlichen Zweck hat und auch schon von großen Konzernen angestrebt wird. Unsere Lösung erweitert und verbessert die von Microsoft um die Möglichkeit deutlich mehr Parameter übergeben zu können.

Datensatz

Der Datensatz ist von Kaggle und kann hier gefunden werden. Hier sind Daten von Nutzer Transaktionen auf E-Commerce Seiten gespeichert, die im Zeitraum des 1.Januar.2011 und dem 31.Dezember.2014 stattgefunden haben. Im Genaueren sind hier Informationen zu den Attributen "Row-ID", "Order-ID, "Order-Date", "Ship-Date", "Ship Mode", "Customer ID", "Costumer Name", "Segment", "City", "State", "Product ID", "Category", "Sub-Category", "Product Name", "Sales Quantity", "Discount", "Profit", "Shipping Cost" und "Order Priority". Diese Menge an Attributen wird sich jedoch, aufgrund von Bereinigung des Datensatzes, im Laufe des Projekts verkleinern.

Ebenfalls sind die Daten für unseren Sachzusammenhang gelabeled, da jede Zeile einen Wert für den Profit der Transaktion beinhaltet.

Wissenschaftliche Vorgehensweise

Die wissenschaftliche Vorgehensweise unserer Datenexploration sollte möglichst repräsentativ, vergleichbar, messbar und wiederholbar sein. Daher haben wir uns einen besonderen Ansatz für die Exploration überlegt. Das Ziel ist es ja, einen Algorithmus zu trainieren, der den Profit vorhersagt. Diese Vorhersage sollte anhand verschiedener

Parameter passieren, die den Profit-Wert beeinflussen. Um dies zu ermöglichen, müssen zuerst alle Attribute herausgestellt werden, die eine Korrelation mit dem Profit haben (deren Veränderung den Profit also auch verändert). Wir haben folgende repräsentative Methodik angewendet, um diese Attribute zu finden: Alle Attribute werden auf genau die gleichen Kriterien getestet. Also egal, wie offensichtlich eine vermeintliche Korrelation mit dem Profit ist, werden trotzdem die gleichen Schritte wie bei allen anderen Attributen durchgeführt. Dies erzeugt eine sehr gute Vergleichbarkeit und Nachvollziehbarkeit der Ergebnisse. Die Tests dieser Kriterien waren einerseits die Prüfung der grafischen Korrelation mittels Diagrammen, sowie andererseits die rechnerische Korrelation mittels statistischer Methoden. Zur genauen Umsetzung dieser Testmethoden jedoch später mehr. Der gesamte Quellcode unseres Projekts ist auf diese Vorgehensweise ausgelegt. Denn jedes der Notebooks der Datenexploration ist jeweils nummeriert. Hierbei spiegelt eine Nummer den Test eines Attributs wider. Dieser Test ist bei jedem Attribut gleich aufgebaut und durchgeführt. Außerdem hat auch jedes Attribut ein kleines Fazit im Dokument, welches begründet warum das Attribut entfernt oder beibehalten wird. Die wissenschaftliche Vorgehensweise findet sich also im Quellcode wieder.

Die Datenexploration

Wie erläutert, durchläuft jedes Attribut den Prozess der grafischen Analyse der Auswirkung auf den Profit. Hier wird grundlegend zwischen den kategorischen und numerischen Attributen unterschieden.

Kategorische Variablen haben den Vorteil, dass hier die Transaktionen gruppiert werden können (z.B. Gruppierung nach Produkt-Kategorie: entweder Technologie, Möbel). Es muss jedoch zunächst bei jedem gruppierten Attribut geprüft werden, ob die einzelnen Gruppen genug Transaktionen haben, um valide Vorhersagen treffen zu können. Denn wenn z.B. ein Attribut über 30000 verschiedene Ausprägungen hat, wobei jedes Attribut nur ein- oder maximal zweimal vorkommt, kann den Durchschnittswerten dieser Gruppe keine Aussagekraft zugewiesen werden. Denn die Gefahr wäre dann zu groß, dass diese ein oder zwei Werte für dieses Attribut nur Ausreißerwerte sind. Die optimalen kategorischen Attribute haben also möglichst wenige verschiedene Gruppen, die dafür dann jeweils viele Transaktionen haben. Anschließend an diese Prüfung wird für jede dieser Gruppen ein durchschnittlicher Profit gebildet. Im Idealfall ist der durchschnittliche Profit je nach Gruppe anders. Denn dann kann die Annahme getroffen werden, dass dieses Attribut den Profit beeinflusst.

Numerische Variablen hingegen erlauben keine Durchschnittsbildung. Daher werden hier vor allem Streudiagramme genutzt, um einen angenäherten linearen Zusammenhang zu überprüfen. Diese Art der grafischen Untersuchung kann also trotzdem genauso gut einen Aufschluss auf den Bezug zum Profit geben.

Die Interpretation der Diagramme ist jedoch subjektiv, sodass die Annahmen unterschiedlich ausfallen können. Um die Wissenschaftlichkeit beizubehalten wird jedes Attribut darüber hinaus noch einem statistischen Test unterzogen. Auch hier wird die Vorgehensweise zwischen kategorischen und numerischen Variablen leicht unterschieden.

Bei numerischen Variablen kann in Python Pandas einfach der Befehl "df.corr" ausgeführt werden, um die Korrelationen der Attribute untereinander herauszustellen. Die Interpretation dieses Werts ist naheliegend: ein Wert von -1 ist eine stark negative Korrelation, ein Wert von 0 keine Korrelation und ein Wert von 1 eine stark positive Korrelation. Die Ergebnisse dieses Befehls gleichen den Werten der numerischen Variablen der in unseren Notebooks verwendeten Korrelations-Heatmap von der Dython-library.

Kategorische Variablen müssen anders untersucht und analysiert werden. Hier ist es etwas komplizierter statistisch festzustellen, ob eine Korrelation zwischen dem Attribut und dem

Profit vorliegt. Daher haben wir uns dazu entschieden zwei verschiedene statistische Tests bei jedem kategorischen Attribut durchzuführen. Einerseits wurden die Werte aus der Korrelations-Heatmap von der Dython-library genutzt. (Wie diese berechnet werden ist hier erklärt) Darüber hinaus haben wir jedes kategorische Attribut dem Anova-Test unterzogen. (genaue Dokumentation dazu ist hier, hier und hier zu finden) Hierbei gilt, dass das Attribut eine statistisch signifikante Abhängigkeit zum Profit hat, wenn der p-Wert unter 0,05 und der F-Score möglichst hoch ist.

Wenn das jeweilige Attribut diese drei Schritte der Datenexploration (Prüfung auf Aussagekraft, grafische Analyse und statistische Analyse) durchlaufen hat, kann entschieden werden, ob das Attribut entfernt wird oder die Korrelation zum Profit stark genug ist, um für das folgende Training des Machine Learning Modells genutzt zu werden.

Das Modell

Nach dem alle Attribute, die bei der vorangegangenen Exploration wegen fehlender statistischer Abhängigkeit mit dem Profit entfernt wurden, kann ein Algorithmus (unser Modell) aufgesetzt werden. Anhand dieses Modells soll dann der Profit vorhergesagt werden können. Als erstes werden alle kategorischen Variablen in numerische Variablen umgewandelt. Dies ist nötig, da das Machine Learning Modell keine Zusammenhänge zwischen Text-Variablen finden kann. Durch die Methode „ df['Attribut'] =df['Attribut'].astype('category').cat.codes “ wird jeder kategorischen Ausprägung ein numerischer Wert zugeordnet. Davon sind die Attribute 'Sub-Category', 'Country', 'Market', 'Region' betroffen. Dabei hat als Beispiel das Land Australien die Zahl 6 und Deutschland die Zahl 47 zugeordnet bekommen. Danach wird der Datensatz in Trainings und Testdaten aufgeteilt. Das Splitting wird gemäß der Machine Learning Grundsätze mithilfe der library 'train_test_split' von sklearn durchgeführt. Nach dem Splitting wird die Lineare Regression angewendet, um mit dem Score die Genauigkeit des Modells widerzuspiegeln. Je näher der Score an 1 ist, desto besser ist das Modell. Das 'Weight' ist das Gewicht, welches den einzelnen Attributen zugeordnet wird, um den Profit vorherzusagen. Dies wird ebenfalls für die Ridge Regression durchgeführt. Darauffolgend wird die Aussagekraft der beiden Regressionen miteinander vergleichen, indem die beiden Regressionen in einer bestimmten Anzahl von Iterationen trainiert werden. Bei jeder Iteration wird der Datensatz neu 'geshuffelt' und davon zufällig 20% als Testdaten genutzt. Dies entspricht den Kriterien der Kreuzvalidierung und sorgt für ein valides Ergebnis. Ersichtlich wird, dass die Ridge Regression bei höherer Anzahl an Iterationen besser performt. Danach werden die gewünschten Inputs in Tabellen ausgegeben. Mit diesen Parametern kann dann im siebten Schritt mithilfe des Modells der Profit vorhergesagt werden.

Der betriebswirtschaftliche Bezug/Ergebnisse

Der betriebswirtschaftliche Bezug besteht, wie im Abschnitt Related Work erläutert, darin, dass es bereits tatsächlich Unternehmen gibt, die eine solche parameterorientierte Profitvorhersage nutzen wollen. Diese Funktion haben wir erfolgreich umgesetzt.

Ein praxisnahes Beispiel für diese Vorhersage haben wir ebenfalls im Notebook 5_Modell_aufsetzen.ipynb gegeben. Ein Unternehmen mit Sitz in den USA kennt schon Kategorie (Technologie), Sub-Kategorie (Handy), Quantity (1) und Discount (0,1) ihres neuen Produkts. Es nutzt unsere Software, um herauszufinden, was die beste Kombination aus den übrigen Attributen (Region, Country, Market, Shipping Cost) ist, um den Profit zu maximieren. Es wird also mehrmals der Befehl reg.predict() aufgerufen, dem jeweils eine andere Kombination aus den genannten Attributen übergeben wird. Das Ergebnis dieses Beispiels ist, dass die USA (auch wegen der niedrigeren Shipping Costs) der profitabelste Absatzmarkt ist.

Kritische Reflexion

Der Vorteil und große Mehrwert dieses Projekts ist, dass es eine detaillierte Datenexploration gab, die zu einem gut funktionierenden Modell geführt hat. Die Arbeits- und Vorgehensweise bei diesem Projekt lief sehr gut und hat entlang der Datenexploration zu zahlreichen Lerneffekten und guten Ergebnissen geführt.

Dennoch müssen wir das Projekt auch kritisch reflektieren: Offene Kritikpunkte sind die Größe der Datenbasis (ggf. wären für ein besseres Modell mehr Datensätze nötig gewesen), die Genauigkeit der Parameter (beim Testen ist aufgefallen, dass das Modell nicht gänzlich perfekt vorhersagt) und, dass keine weiteren Modelle (abgesehen von den zwei Regressionsarten) trainiert wurden. Da der Fokus dieses Projekts auf der Datenexploration und nicht dem perfekten Training liegen sollte, haben wir den letzten beiden Punkten eine niedrigere Priorität zugeordnet. Dementsprechend wären die nächsten Schritte zur Verbesserung der Software über dieses Projekt hinaus das Training anderer Machine Learning Modelle, um ggf. noch zuverlässigere Profit Vorhersagen machen zu können.

Weitere Quellen

Quelle1: https://www.controllingportal.de/Fachinfo/Excel-Tipps/Umsatzprognosen-ganz-einfach-erstellen-Nutzen-Sie-die-SCHAeTZER-Funktion.html

Quelle2: https://support.microsoft.com/de-de/office/erstellen-einer-prognose-in-excel-f%C3%BCr-windows-22c500da-6da7-45e5-bfdc-60a7062329fd