Fachhochschule
Münster University of
Applied Sciences

Business Intelligence

Prüfungsaufgabe

Bachelor Wirtschaftsinformatik

Prof. Dr. Wolfgang Wicht
Dennis Cosfeld-Wegener

Business Understanding

Kontext & Verkaufserlösprognose

➔ Sie sind Mitarbeiter in einem Elektronik-Handelsunternehmen, das seine Waren zu einem Großteil über eine Auktionsplattform vertreibt. Der Verkaufsleiter hat festgestellt, dass die Auktionen zum Teil sehr unterschiedliche Verkaufserlöse erzielen. Ein Muster ist jedoch nicht erkennbar. Zum Zweck der Erlösmaximierung soll nun mittels Data Mining ein Vorschlag für eine optimale Auktionslistung (Start- und Endzeitpunkte, Laufzeit etc.) erarbeitet werden.

Im ersten Schritt ist hierfür ein Modell zu erstellen, welches für jede neu eingestellte Auktion eine Vorhersage treffen kann, ob der Verkaufserlös des Artikels über dem durchschnittlichen Verkaufserlös der jeweiligen Produktkategorie liegen wird.
Die Grundlage der Analyse bilden die Daten von 8.000 Online-Auktionen aus der Kategorie „Audio&Hi-Fi:MP3-Player:Apple iPod“.

Business Understanding

➔ Der Variable *category_avg_gms* kann der mittlere Verkaufserlös der jeweiligen Produktkategorie entnommen werden, während die Variable *gms_greater_avg* (Zielvariable) aussagt, ob der Verkaufserlös der jeweiligen Auktion über dem mittleren Verkaufserlös liegt.

➔ Die folgende Gewinnmatrix bildet die Grundlage für die Optimierung der Klassifikation in die Klassen **Hochpreis** (=Verkaufserlös über dem Mittelwert) und **Niedrigpreis** (=Verkaufserlös unter oder gleich dem Mittelwert):

Gewinnmatrix		Echter Wert	
		Hochpreis	Niedrigpreis
Vorher-sage	Hochpreis (*category_avg_gms* = 1)	G(1|1)= 1	G(1|0)= -1
	Niedrigpreis (*category_avg_gms* = 0)	G(0|1)= -1	G(0|0)= 1

Business Understanding

Bewertung Verkaufserlösprognose

- Die Gesamtpunktzahl wird durch den Vergleich Ihrer Prognosen mit den tatsächlichen Werten entsprechend der Gewinnmatrix errechnet:
  - 1|1 oder 0|0: Jede richtig eingestufte Auktion gibt 1 Punkt
  - 0|1 oder 1|0: Jede falsch eingestufte Auktion gibt -1 Punkt

- Ziel ist eine Punktzahlmaximierung durch richtige Klassifizierung
  - Gesamtpunktzahl = 1*[*1*|*1*] - 1*[*0*|*1*] + 1*[*0*|*0*] - 1*[*1*|*0*]
  - mit:
    - [0|0] – Anzahl der richtigen Zuordnungen zu Klasse 0
    - [1|1] – Anzahl der richtigen Zuordnungen zu Klasse 1
    - [0|1] – Anzahl der Zuordnungen zu Klasse 0 aus realer Klasse 1 (Falschzuordnung)
    - [1|0] – Anzahl der Zuordnungen zu Klasse 1 aus realer Klasse 0 (Falschzuordnung)

Business Understanding

Marketingmaßnahmen

➔ Unabhängig von den Optimierungsmaßnahmen für die Auktionslistung ist eine tiefgehende Evaluierung der bisherigen Vertriebsaktivitäten in Planung. Um hierfür Ansatzpunkte zu generieren, erhalten Sie den Auftrag zusätzlich eine Segmentierung der Auktionen für die Produktkategorie „Audio&Hi-Fi:MP3-Player:Apple iPod“ durchzuführen. Die hierdurch ermittelten unterschiedlichen Segmente von Auktionen sollen anschließend eingehender hinsichtlich ihrer Effektivität zur Erreichung der Vertriebsziele bewertet werden.

Zur Adressierung insbesondere der jüngeren Zielgruppe ist darüber hinaus vor kurzem eine spezielle Webpräsenz geschaffen worden, auf der aktuelle Medieninhalte kostenlos erhältlich sind. Das Klickverhalten der Nutzer dieser Plattform soll ebenfalls auf marketing-relevante Ansätze in Form von Inhaltskombinationen überprüft werden.

Data Understanding

Trainingsdaten

➔ 8.000 Datensätze

➔ Datei: e-auction_train.txt

Datenfeld	Beschreibung [Ausprägungen]	Datentyp
auct_id	ID Nummer der Auktion	Integer
item_leaf_category_name	Produktkategorie	String
listing_title	Titel der Auktion	String
listing_subtitle	Untertitel der Auktion	String
listing_start_date	Startzeitpunkt der Auktion	Date (MM/TT/JJJJ)
listing_end_date	Endzeitpunkt der Auktion	Date (MM/TT/JJJJ)
listing_durtn_days	Dauer der Auktion	Integer
listing_type_code	Typ der Auktion [normale Auktion, Multiauktion, …)	Integer

Data Understanding

Datenfeld	Beschreibung [Ausprägungen]	Datentyp
feedback_score_at_ listing_time	Feedback-Rating des Verkäufers bei Listung der Auktion	Integer
start_price	Startpreis (in EUR)	Decimal
buy_it_now_price	Sofortkauf-Preis (in EUR, bei Sofortkauf-Option)	Decimal
buy_it_now_listed_flag	Auktionslistung mit Sofortkauf-Option	Boolean
bold_fee_flag	Auktionslistung mit Fettschrift	Boolean
featured_fee_flag	Auktionslistung als Homepage-Top-Angebot	Boolean
category_featured_fee_ flag	Auktionslistung als Kategorie-Top-Angebot	Boolean
gallery_fee_flag	Auktionslistung mit Galerie-Bild	Boolean
gallery_featured_fee_fl ag	Auktionslistung mit Galerie (nur in Galerie-Ansicht)	Boolean
ipix_featured_fee_flag	Auktionslistung mit ipix (Additional, xxl, pic.show, pack)	Boolean

Data Understanding

Datenfeld	Beschreibung [Ausprägungen]	Datentyp
reserve_fee_flag	Auktionslistung mit Reserve-Preis	Boolean
highlight_fee_flag	Auktionslistung mit Hintergrundfarbe (in Listenansicht)	Boolean
schedule_fee_flag	Auktionslistung mit Festlegung des Startzeitpunktes	Boolean
border_fee_flag	Auktionslistung mit Rahmen	Boolean
qty_available_per_listing	Menge der angebotenen Artikel bei Multiauktionen	Integer
gms	Erzielter Verkaufserlös (in EUR)	Decimal
category_avg_gms	Durchschnittlicher Verkaufserlös (in EUR) der Produktkategorie (item_leaf_category_name)	Decimal
gms_greater_avg **(Zielvariable)**	0 wenn gms <= category_avg_gms 1 wenn gms > category_avg_gms	Boolean

Data Understanding

Trainingsdaten

➔ 2.234.558 Datensätze

➔ Datei: website_train.txt

Datenfeld	Beschreibung [Ausprägungen]	Datentyp
ID	IP-Adresse (durch ID-Nummer anonymisiert)	Integer
TARGET	Klickziel	String

Aufgaben

1. Führen Sie eine Klassifikationsanalyse durch, um aufgrund der Auktionsmerkmale in den Trainingsdaten (*e-auction_train.txt*) für die Auktionen in den Klassifizierungsdaten (*e-auction_class.txt*) eine Entscheidung zu treffen, ob diese als Hoch- oder Niedrigpreis einzustufen sind. Das Ziel ist hierbei eine Punktzahlmaximierung.
2. Führen Sie eine Segmentierungsanalyse der Auktionen anhand der Trainingsdaten durch und bilden Sie Cluster, welche in sich möglichst homogen, im Vergleich zu den anderen Clustern jedoch möglichst heterogen sind. Interpretieren Sie die Cluster und leiten Sie Implikationen für das Marketing ab.
3. Führen Sie eine Assoziationsanalyse hinsichtlich der angeklickten Servicedienste auf der neuen Website (*website_train.txt*) durch. Leiten Sie relevante Regeln für das Cross-Selling des Musik-Streaming-Services ab.

zu Aufgabe 1. Klassifikationsanalyse

➔ Form der Abgabe:

- PMML-Datei des Modells sowie eine Liste aller Auktionsnummern mit der Klassifizierung des Modells in folgender Form
  - **<auct_id>;<gms_greater_avg>**
  - **<auct_id>;<gms_greater_avg>**
  - …

    wobei <gms_greater_avg> den Wert „1“ für Hochpreiseinstufungen oder den Wert „0“ für Niedrigpreiseinstufungen enthält

- Alle Streams aus dem SPSS Modeler, welche für das Projekt erstellt wurden, gesammelt in einer .zip-Datei

- Ein maximal 12-seitiger Präsentations-Foliensatz, in dem Sie die wesentlichen Punkte Ihres Vorgehens beschreiben sowie – falls angebracht – Interpretationen vornehmen

zu Aufgabe 1. Klassifikationsanalyse

➔ Bewertungskategorien:

- Performance im Wettbewerb (Gewichtung: 25%)
  - Die Performance wird anhand der Gesamt-Klassifikationskosten (bzw. –gewinne) bestimmt.
  - Das beste Team erhält die höchste Bewertung hinsichtlich des Bewertungskriteriums „Performance“.
  - Die Abstände zum besten Team dienen der Einstufung der sonstigen Teams.

- Erläuterung der Vorgehensweise und Lösung (Gewichtung: 25%)
  - Methodisch korrekte Durchführung der Analyse
  - Komplexität / Effizienz / Originalität der Streams
  - Dokumentation der Vorgehensweise und Lösung sowie ggf. Interpretation in Form von Präsentationsfolien

zu Aufgabe 2. Segmentierung

➔ Form der Abgabe:

- PMML-Datei des Modells sowie eine Liste **aller** Auktionen der **Trainingsdaten** und der Segmentzugehörigkeit in folgender Form
  - **`<auct_id>;<Segment>`**
  - **`<auct_id>;<Segment>`**
  - …
    wobei für die Nummer des ermittelten Segments steht, zu dem die Auktion zugeordnet wurde (Missing Value bei keiner Segmentzugehörigkeit)

- Alle Streams aus dem SPSS Modeler, welche für das Projekt erstellt wurden, gesammelt in einer .zip-Datei

- Ein maximal 12-seitiger Präsentations-Foliensatz, in dem Sie die wesentlichen Punkte Ihres Vorgehens beschreiben sowie Interpretationen der Cluster vornehmen

zu Aufgabe 2. Segmentierung

➔ Bewertungskategorien:

- Performance im Wettbewerb (Gewichtung: 15%)
  - Die Performance wird anhand des Silhouetten-Koeffizienten[1] (Gesamt-Silhouette) für die Segmentierungslösung bestimmt. Dem ausgewählten Modell sollten **mindestens 15 Variablen** für die Segmentierung zur Verfügung gestellt und es sollten **mindestens 3 Cluster** gebildet werden.
  - Das beste Team erhält die höchste Bewertung hinsichtlich des Bewertungskriteriums „Performance“.
  - Die Abstände zum besten Team dienen der Einstufung der sonstigen Teams.
- Erläuterung der Vorgehensweise und Lösung (Gewichtung: 15%)
  - Methodisch korrekte Durchführung der Analyse
  - Komplexität / Effizienz / Originalität der Streams
  - Dokumentation der Vorgehensweise und Lösung sowie ggf. Interpretation in Form von Präsentationsfolien

1) Quelle: IBM SPSS Modeler 17.0: Die standardmäßige Rangeinteilung mit Silhouetten verwendet einen Standardwert von 0, da ein Wert kleiner 0 (also ein negativer Wert) angibt, dass der durchschnittliche Abstand zwischen einem Fall und Punkten in seinem zugeordneten Cluster größer ist als der minimale durchschnittliche Abstand zu Punkten in einem anderen Cluster. Daher können Modelle mit einer negativen Silhouette einfach verworfen werden.
Die Rangeinteilung ist eigentlich ein modifizierter Silhouetten-Koeffizient, der die Konzepte von Cluster-Zusammenhalt (Favorisierung von Modellen mit eng zusammengehörenden Clustern) und Cluster-Abgrenzung (Favorisierung von Modellen mit stark separierten Clustern) kombiniert. Der durchschnittliche Silhouetten-Koeffizient ist einfach der Durchschnitt für alle Fälle der folgenden Berechnung für jeden Einzelfall:
$(B - A)/\max.(A, B)$
A ist die Entfernung zwischen Fall und Zentroid des Clusters, zu dem der Fall gehört und *B* ist der minimale Abstand zwischen Fall und Zentroid jedes anderen Clusters.
Der Silhouetten-Koeffizient (und sein Durchschnittswert) liegen zwischen -1 (stellvertretend für ein sehr schlechtes Modell) und 1 (stellvertretend für ein ausgezeichnetes Modell). Der Durchschnitt kann auf der Ebene aller Fälle (führt zu einer Gesamt-Silhouette) oder auf Cluster-Ebene (führt zu einer Cluster-Silhouette) durchgeführt werden. Entfernungen können mithilfe euklidischer Entfernungen berechnet werden.

zu Aufgabe 3. Assoziationsanalyse

➔ Form der Abgabe:

- PMML-Datei des Modells (enthält die generierten Regeln) sowie die **10 besten Cross-Selling-Empfehlungen** für den Musik-Streaming-Service (d.h. Musik-Streaming als **Regelfolge**) entsprechend des Supports. Formvorlage:
  - **`<sukzedens_musicstream>;<antezedens>;<support>;<konfidenz>;<regelunterstützung>`**
  - **`<sukzedens_musicstream>;<antezedens>;<support>;<konfidenz>;<regelunterstützung>`**
  - …
- Alle Streams aus dem SPSS Modeler, welche für das Projekt erstellt wurden, gesammelt in einer .zip-Datei
- Ein maximal 12-seitiger Präsentations-Foliensatz, in dem Sie die wesentlichen Punkte Ihres Vorgehens beschreiben sowie – falls angebracht – Interpretationen vornehmen

zu Aufgabe 3. Assoziationsanalyse

➔ Bewertungskategorien:

- Erläuterung der Vorgehensweise und Lösung (Gewichtung: 20%)
  - Methodisch korrekte Durchführung der Analyse
  - Komplexität / Effizienz / Originalität der Streams
  - Dokumentation der Vorgehensweise und Lösung sowie ggf. Interpretation in Form von Präsentationsfolien

Rahmenbedingungen

➔ Abgabefrist: 01. Februar 2016, 09:00 Uhr

➔ Abgabeform:

- E-Mail an
  wolfgang.wicht@fh-muenster.de und
  dennis.cosfeld@fh-muenster.de

sowie

- Ausdruck der Präsentations-Foliensätze an das Sekretariat (Frau Lauerwald)