Hochschule Darmstadt, Fachbereich Informatik
Prof. Dr. S. Karczewski
Prof. Dr. P. Muth
M.Sc. C. Schnellbacher

Praktikum 4

Ziel des Praktikum

In diesem Praktikum beschäftigen Sie sich mit drei wichtigen Themen:

1. Data Mining
2. Machine Learning
3. Daten-Transformation

Ziel dieses Praktikum ist es, dass Sie einen ELT-Prozess entwerfen, in dem Sie Passagierdaten der Titanic einlesen verarbeiten und Vorhersagen darüber treffen, ob ein Passagier überlebt oder nicht.

Aufgabe 1 (Kaggle Titanic Tutorial)

Unter

`http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/`

finden Sie ein Tutorial dazu. Sie nutzen die im statistischen Bereich verbreitete Sprache R. Im Tutorial wird die Sprache R am Beispiel eingeführt. Führen Sie dieses Tutorial mindestens bis Part 4 (inklusive) durch. Im der nächsten Aufgabe werden wir dann weitere Daten hinzunehmen und in einem Data Warehouse ablegen. Sie können Ihre Vorhersage mit Hilfe dieser Daten weiter verbessern.

Um das Ergebnis ihrer Vorhersagen mit anderen vergleichen zu können, nutzen wir einen sog. Kaggle-Wettbewerb. Sie erstellen für Ihre Gruppe einen Kaggle-Account und reichen Ihre Vorhersage dort ein. Sie bekommen sofort das Ergebnis und können sich mit anderen Gruppen messen. Näheres dazu finden Sie ebenfalls im Tutorial.

Aufgabe 2 (ETL-Prozess)

Bisher haben Sie nur ein einfaches .csv-File als Datenquelle für Ihren Vorhersagealgorithmus benutzt. Sie haben im Tutorial gelernt, dass man die Vorhersagequalität verbessern kann, wenn man sich weitere, vorhersagerelevante Daten erschließt. Dazu finden Sie in Moodle eine Passagierliste der Titanic. Sie haben nun mehrere Dateien mit Daten über das Unglück der Titanic: Die Passagierliste aus der vorigen Aufgabe mit Informationen wer überlebt hat, und eine Datei mit weiteren Daten zu den Passagieren. Entwerfen Sie für die Daten aus beiden Dateien ein gemeinsames Star-Schema und modellieren Sie es.

Hinweis: Für die Durchführung dieser Aufgabe nutzen Sie KNIME und Power BI.

Erstellen Sie mit KNIME einen ETL Prozess, befüllen Sie Ihre Datenbank und führen Sie zur Kontrolle einfach Abfragen mit Hilfe von Power BI durch.

Tipp zum Star Schema: Überlegen Sie zunächst, welche Daten Fakten sind und in eine Faktentabelle gehören und welche Daten dimensionale Daten sind, aus denen Sie Dimensionen ableiten können. Auf den ersten Blick scheinen die meisten Daten Fakten zu sein.

Sie können aber sogar aus den Namen eine Dimension über den Rang einer Person in der Gesellschaft ableiten. Aus den Informationen über Kabinen und Decks können Sie eine Ortsdimension bauen. Legen Sie in Ihrem Schema mindestens drei Dimensionen an.

Tipp zum ETL-Prozess: Die Daten beider Dateien passen nicht ohne Bearbeitung im ETL-Prozess zusammen. Lösen Sie dieses typische Data Warehouse-Problem in angemessener Weise. Wenn Sie nicht alle Daten zuordnen und nutzen können, ist das kein Problem.

Aufgabe 3 (Optional)

Erzeugen Sie aus Ihrem Data Warehouse Input-Daten für eine verbesserte Vorhersage. Versuchen Sie, Ihr bisher bestes Ergebnis in Kaggle zu übertreffen.

Einzige Rahmenbedingung: Die Daten müssen aus Ihrem Data Warehouse kommen. Nutzen Sie hierfür KNIME.