Prognose von Herstellkostenschwankungen mit Predictive Analytics

Alfred Stelzl
CGI Deutschland Ltd. & Co. KG
Sulzbach

Schlüsselworte
Predictive Analytics, Prognosen, Maschinelles Lernen, Analytics Lifecycle

Einleitung

Predictive Analytics (ein deutschsprachiger Begriff existiert bisher nicht) nutzt ausschließlich proaktive Analysetechniken, um Vorhersagen über zukünftige Ereignisse,Werte und Informationen zu tätigen. Dabei werden auf Basis umfangreicher Datenanalysen valide Prognosemodelle abgeleitet, um Geschäftsrisiken und Chancen frühzeitig zu erkennen. Maschinelles Lernen, Statistik und Data Mining spielen eine wesentliche Rolle bei der Predictive Analytics.

Im Folgenden wird ein Proof of Concept (PoC) beschrieben, um die Herstellkostenschwankung mit Predictive Analytics zu prognostizieren.

Der Use Case

Der Use Case bezieht sich auf den Angebotsprozess des Kunden. Die Herstellkosten (HK) bilden eine zentrale Kosteninformation und spiegeln die Produktionskosten von Transporten wider. In der Vorkalkulation werden die Herstellkosten auf Basis eines Transports berechnet, um eine erste Indikation für den Angebotsprozess zu liefern.
Instabile Produktionsstrukturen erzeugen schwankende Herstellkosten und führen zu einer Volatilität (Schwankung). Je stärker die Herstellkosten schwanken, desto unvorhersehbarer ist die letztendliche Marge. Es bedarf Frühindikatoren und Steuerungsinstrumente, um auf die Volatilität in der Angebotsphase entsprechend reagieren zu können.

Der PoC untersucht auf Basis von historischen Daten die Volatilität / Stabilität der Herstellkosten von Transporten sowie deren Vorhersagemöglichkeit. Die Volatilität wird als prozentuale Schwankung bzw. Streuung definiert. Für die Modellierung werden Methoden und Algorithmen des maschinellen Lernens eingesetzt, um prädiktive Modelle zu erstellen, die anschließend als Frühindikator für die Volatilität verwendet werden können.

Ziel ist es, die Rentabilität der Angebote durch die Vorhersage der prozentualen Schwankung der HK zu verbessern.
Kundennutzen:
• Risikobegrenzung dank Preisgleitklauseln
• Verbesserung des DB I durch preispolitische Maßnahmen bei volatilen Herstellkosten
• Größeres Vertrauen in die Vertriebsplanung

Die Architektur

Die Quelldaten werden aus einer Oracle Exadata und MS SQL Server extrahiert und anschließend mit RapidMiner weiterverarbeitet. RapidMiner ermöglicht eine grafische Modellierung von Prozessen, in denen einzelne Operatoren die Funktionalitäten widergeben. Als Erweiterungen werden die Extensions Feature Selection, Parallel Processing und WEKA eingebunden und verwendet.

Die Vorgehensweise

Als methodische Vorgehensweise wird der CGI Analytics Lifecycle verwendet, den CGI für Big Data Analytics Projekte empfiehlt. Er beschreibt die typischen Phasen eines Projekts inkl. der

auszuführenden Arbeiten. Als Prozessmodell bietet er einen Überblick über den Predictive Analytics Lebenszyklus. Der Prozess beinhaltet folgende Phasen:

Business Insights: Verständnis der fachlichen Aufgabenstellung und des operativen Kontextes
Data Understanding: Verständnis der benötigten Datenquellen
Data Modeling: Aufbereitung der Quelldaten zu einem analysefähigen Bestand
Analytics Modeling: Erstellung und Evaluierung der Modelle
Analytics Deployment: Bereitstellung der Ergebnisse für die Nutzung

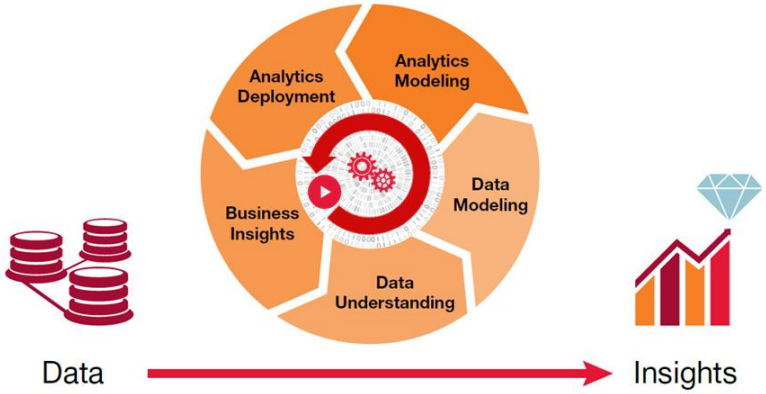

Abb. 1: CGI Analytics Lifecycle (Data to Diamonds)

Die Modellerstellung
Die Modellerstellung basiert auf dem Ansatz des maschinellen Lernens. Maschinelles Lernen bezeichnet einen Prozess, bei dem künstliche Systeme Wissen aus Erfahrungen generieren. Da künstliche Systeme über keine Erfahrungen wie Menschen oder Tiere verfügen, müssen diese mithilfe von Daten "erlernt" werden. Die Algorithmen des maschinellen Lernens erkennen in den historischen Daten (komplexe) Muster und Regelmäßigkeiten, die in Form eines mathematischen Modells gespeichert werden.
Die Modellierung wird mit einer multiplen linearen Regression (MLR) und dem M5Prime Modellbaum (M5') durchgeführt und anschließend miteinander verglichen. Als Feature Selection für die MLR wird eine Correlation Based Feature Selection mit anschließendem t-Test durchgeführt, für den M5' Modellbaum wird und eine Forward Selection eingesetzt. Beide Algorithmen werden jeweils mit einer Kreuzvalidierung verarbeitet.

Die Evaluierung
Für beide Algorithmen werden die Gütemasse in den Trainings-, Validierungs- und Testprozessen berechnet. Folgende Tabelle zeigt eine Übersicht der Ergebnisse:

Algorithmus	R^2 **Training**	*RMSE* **Validierung**	*RMSE* **Test**
MLR	0,44	9,026	8,905
M5'	0,48	8,927	8,734

Abb. 2: Übersicht der Ergebnisse

Der M5' Modellbaum schneidet bei allen drei Bewertungsmaßen geringfügig besser ab, als die multiple lineare Regression. Der R^2 für die Trainingsdaten liegt beim Modellbaum bei 0,48 und bei der multiplen linearen Regression bei 0,44. Ein R^2 von 0,5 kann bei einer Querschnittsbetrachtung bereits

als guter “fit“ eingestuft werden. Die RMSE Werte liegen bei beiden Modellen deutlich unter der Standardabweichung der abhängigen Variablen mit 12%.

Die Zusammenfassung

Die Ergebnisse des PoC zeigen, dass die Modelle nahezu identische Werte in der Genauigkeit der Prädiktion innerhalb der Trainings- und Testdaten liefern. Das M5’ Modell ist bei allen Bewertungsmaßen besser, als die multiple lineare Regression. Beide Modelle weisen kein Overfitting auf, was durch eine genauere Prädiktion in den Testdaten festgestellt werden kann und somit den Einsatz der aufwändigen Variablenselektion bestätigt. Beim Bestimmtheitsmaß ist der Unterschied zwischen den Modellen größer. Das M5’ Modell liefert hierbei für die Erklärung der Varianz gute Werte.

Der PoC zeigt weiterhin auf, dass Predictive Analytics Projekte sehr umfangreich, kreativ und iterativ sind, bei dem Fail Fast erlaubt sein muss. Die Daten verlangen besonders viel Aufmerksamkeit, bevor sie mit Algorithmen weiterverarbeitet werden können. Folgende Aspekte sind dabei wichtig:

• Inhalte und Qualität der Quelldaten müssen geprüft und verstanden werden. Was bedeuten die einzelnen Merkmale fachlich und welche Ausprägungen sind vorhanden? Gibt es Ausreißer, fehlende oder ungültige Werte? Sind Nachbesserungen erforderlich, z.B. das Korrigieren fehlerhafter Ausprägungen oder Ersetzen fehlender Werte mit sinnvollen Inhalten.

• Ist der Umfang der vorhandenen Dateninhalte ausreichend? Benötigt man für die Analyse neue, aus den Daten abgeleitete Merkmale?

• Müssen Datenquellen zusammengeführt oder verdichtet werden, um ein brauchbares Ergebnis zu erzielen?

Kontaktadresse:

Alfred Stelzl
CGI Deutschland Ltd. & Co. KG
Am Limespark 2
D-65843 Sulzbach am Taunus

Telefon: +49 (0) 6196-77420
E-Mail Alfred.Stelzl@cgi.com
Internet: http://www.de.cgi.com