RiskDataScience
Applying concepts

Machine Learning-basiertes Kreditrating-Frühwarnsystem

Übersicht Problemstellung und Lösungsansatz

Als wichtige Risikoart werden Kreditrisiken mit anspruchsvollen **Rating-Verfahren** quantifiziert. Aufgrund der aufwendigen Erstellung und fehlender aktueller Bilanzdaten liegen Ratings jedoch nur zeitverzögert vor. Für aktuelle Kreditrisikosignale wurden von Banken daher bereits **marktdaten-basierte Frühwarnsysteme** eingeführt, die aber keine Indikationen im Falle fehlender Marktdaten liefern können.

Andererseits liefern im Internet vorhandene **Unternehmensnachrichten** oft wichtige Informationen über Probleme und Schieflagen.

Hierfür existieren bereits leistungsstarke Algorithmen zur **automatischen Ermittlung und Klassifizierung von Nachrichten-Texten hinsichtlich Insolvenz-Relevanz (*News-Based Early Warning*)**.

Damit können Banken bzw. Industrieunternehmen aus Nachrichtentexten wertvolle Zusatz-Informationen über drohende Insolvenzen von Kunden bzw. Lieferanten gewinnen. Eine Früherkennung von Kreditrisiken ist damit auch für nichtgelistete Unternehmen ohne direkte Marktdaten möglich.

Kreditrisiko-Messung

Allgemeines

Unter *Kreditrisiken* versteht man Risiken durch Kreditereignisse, wie Zahlungsausfall, Zahlungsverzug, Herabstufung der Kreditwürdigkeit oder Einfrierung der Währung. Eine weitere Unterscheidung betrifft die Einteilung in Emittenten- (bei Anleihen), Kontrahenten- (bei Derivate-Geschäften) und – die im Folgenden betrachteten – Kreditausfallrisiken von Kreditnehmern i.e.S.

Kreditrisiken bilden oft das größte Bank-Risiko und müssen – neben Markt- und operationellen Risiken – gemäß Basel II/III mit Eigenkapital unterlegt werden.

Eine häufig herangezogene Kennzahl zur Quantifizierung von Kreditrisiken ist der *erwartete Verlust* (*Expected Loss*) eines Kredits. Dieser ergibt sich im einfachsten Fall als Produkt aus

- *PD*: Probability of Default, Ausfall-Wahrscheinlichkeit
- *LGD*: Loss Given Default, eins minus Wiederverwertungsrate
- *EaD*: Exposure at Default, ausstehendes Kreditvolumen

Externe und interne *Kreditratings* messen hauptsächlich die PD (und z.T. den LGD) und werden mit aufwendigen Verfahren ermittelt.

Ermittlung und Früherkennung

Die Verfahren zur Ermittlung der PD erfordern fundierte statistische Analysen auf Basis von

- *quantitativen Bilanzkennzahlen* wie Verschuldungsgrad, Eigenkapitalquote und EBIT
- *qualitativen Analysten-Kennzahlen* wie Qualität des Managements, Zukunftsaussichten und Marktstellung
- *allgemeinen Marktdaten* wie Zinsen, Inflation und Wechselkursen.

Die Ratingmodelle müssen regelmäßig anhand tatsächlicher Kreditereignisse *validiert* und gegebenenfalls angepasst werden.

Kreditratings liegen deshalb meist verzögert – oftmals nur jährlich – vor.
Zur Behebung dieses Problems wurden marktdatenbasierte Frühwarnsysteme eingeführt, die Signale auf der Basis signifikanter Änderungen von Aktienkursen, Credit Spreads oder weiterer mit dem Rating korrelierter Marktdaten liefern. Im Allgemeinen können damit allerdings nur systematische bzw. Risiken gelisteter Unternehmen erkannt werden.

Informationen aus Nachrichten

Allgemeines

Die Gründe für Insolvenzen sind oft unternehmensspezifisch (idiosynkratisch) und können nicht aus allgemeinen Marktentwicklungen abgeleitet werden. Beispiele hierfür sind

- Betrugsfälle durch das Management
- Insolvenz eines wichtigen Kunden bzw. Lieferanten
- Auftreten eines neuen Konkurrenten

Negative Ereignisse wie Werkschließungen, Kurzarbeit, Ermittlungen und Anklagen gehen dabei der eigentlichen Insolvenz zum Teil um mehrere Monate voraus.

Im Falle nichtgelisteter Unternehmen ist dennoch keine marktdatenbasierte Frühwarnung möglich. Hingegen liefern Nachrichten auch in diesen Fällen aktuelle und oftmals insolvenzrelevante Informationen.
Nachrichtenportale, Blogs, Soziale Medien und insbesondere Lokalzeitungen informieren dabei online über Probleme von Unternehmen.

Durch die effiziente Nutzung von Texten ist somit eine Erweiterung der Frühwarnung auf nichtgelistete Unternehmen möglich.

Effiziente Nachrichten-Analyse

Verfahren zur effizienten Analyse von Texten sind Voraussetzung um die relevanten Nachrichten zu identifizieren und darauf aufbauend mögliche Insolvenzen zu antizipieren. Hierfür notwendig sind

- eine rechtzeitige *Identifizierung* relevanter Datenquellen (Zeitungen, RSS-Feeds, etc.)
- ein *Einlesen* der relevanten Nachrichten zu allen Kunden anhand vorgegebener Muss- und

Ausschlusskriterien

- eine zeitnahe *Klassifikation* der relevanten Texte anhand möglicher Insolvenzrisiken
- eine sofortige *Analyse* und *Visualisierung* der Ergebnisse zur Erkennung von Risiken

Bereits realisierte *Machine Learning-Algorithmen* dienen *als Basis* für diese zunächst unmöglich erscheinende Aufgabe.

Wissensnutzung durch Machine Learning-Verfahren

Einlesen

Im ersten Schritt müssen alle relevanten Nachrichtenquellen anhand einer hinreichend großen Stichprobe zu untersuchender Unternehmen identifiziert und irrelevante Quellen möglichst ausgeschlossen werden.

Die Gewinnung der relevanten Texte aus diesen Quellen kann z.B. über folgende Verfahren erfolgen

- Bezug von Pressetexten über entsprechende Dienstleister
- direktes Abgreifen freier RSS-Feeds

Die Nachrichten sind dabei nach Relevanz zu filtern. Zur Vermeidung von Verwechslungen aufgrund des Namens oder irrtümlicher Textbausteine (z.B. bzgl. Aktien) sind Wortfilter und ggf. komplexe Textanalysen notwendig.

Klassifikation

Für die Klassifizierung der gewonnenen Nachrichtentexte kommen verschiedene *Text Mining-Methoden* aus dem Bereich *Data Science / Machine Learning* in Betracht. Beim *Supervised Learning* wird dabei wie folgt vorgegangen

- zunächst werden manuell die Wörter ermittelt, die für die Klassifikation irrelevant sind („*Stopwords*“)
- die Algorithmen werden dann mit bekannten Datensätzen darauf „*trainiert*“ Texte Kategorien zuzuordnen
- neue Texte können anschließend bekannten Kategorien mit bestimmten *Konfidenzen* zugeordnet werden

Methodisch sind dabei folgende Schritte durchzuführen

- aus den gefilterten Texten werden signifikante Wortstämme/Wortstamm-Kombinationen („*n-grams*“) ermittelt
- die Texte werden als Punkte in einem *hochdimensionalen Raum* (mit den n-grams als Dimensionen) abgebildet
- Machine Learning-Verfahren ermitteln Gesetzmäßigkeiten zur Trennung der Punkte nach Kategorien. Hierfür bieten sich dezidierte Algorithmen wie *naive Bayes*, *W-Logistic* oder *Support Vector Machine* an.

Die Analysen erfordern Programme auf der Basis entsprechender Analysetools, wie z.B. *Python*, *R* oder *RapidMiner*.

Anwendungsbeispiel

Für ca. 50 insolvent gegangene Unternehmen und 50 nicht-insolvente Referenzunternehmen wurden Nachrichten-Snippets für einen mehrmonatigen Zeithorizont (3M–3W) vor der jeweiligen Insolvenz gesammelt.

Die dargestellten *Tagclouds* geben einen exemplarischen Überblick über den Inhalt der Texte. Mit einem *RapidMiner*-Prototypen wurden die Nachrichtentexte hinsichtlich möglicher Insolvenzen klassifiziert und die Resultate mit *In-* und *Out-Of-Sample-Tests* untersucht.

Abbildung 1: Tagcloud Nachrichten insolvent gegangene Unternehmen

Abbildung 2: Tagcloud Nachrichten nicht insolvent gegangene Unternehmen

Bereits anhand der Tagclouds ist somit ein deutlicher Unterschied zwischen den Nachrichten zu insolvent gegangenen und nicht insolvent gegangenen Unternehmen erkennbar.

Die *RapidMiner*-Lösung wurde mit einem Trainingssample (70% der Texte) trainiert und auf einem Test-sample (30% der Texte) angewendet.

Sowohl für das Trainingssample (In-Sample) als auch für das Testsample ergaben sich dabei **Trefferquoten** (*Accuracy*) von ca. **80%**. Die **Area Under the Curve** (*AUC*) lag zudem im In-Sample-Fall bei **90%**.

Anhand der *RapidMiner*-Konfidenzen und den tatsächlichen Insolvenzen konnte zudem eine PD-Kalibrierung durchgeführt werden.

Selbst mit dem relativ kleinen Trainingssample konnte damit eine **signifikante Früherkennung von Insolvenzen** erreicht werden. Weitere Verbesserungen sind mit einer Erweiterung der Trainingsdaten zu erwarten.

Kosteneffiziente Umsetzung

Ausgangslage

Da sich noch kein einheitlicher Markt für Internet-Nachrichten-Lieferungen gebildet hat, sind die *Preise* oft *uneinheitlich*. Unterschiedliche Anforderungen an die Bereinigungsroutinen und unterschiedliche technische Ansätze führen zu *großen Preisspannen*.

Hingegen sind qualitativ hochwertige Analyse-Tools wie *R* oder *RapidMiner* (Version 5.3) z.T. sogar *frei erhältlich*.

Zudem bietet ca. die Hälfte aller Online-Zeitungen ihre Schlagzeilen in Form standardisierter *RSS-Feeds* an.

Kostentreiber

Die Umsetzungs- sowie die laufenden Kosten von nachrichtenbasierten Frühwarnsystemen können sich insbesondere aus den folgenden Gründen z.T. deutlich erhöhen:

- Eine Auswertung *vollständiger Nachrichtentexte* erfordert aus *Urheberrechtsgründen* Gebühren an Verwertungsgesellschaften (*VG Wort*) bzw. einen direkten Kauf.
- Ein Crawling ist technisch *aufwendig*.
- Die Pflege fortschrittlicher NLP-Algorithmen (Natural Language Processing) zur Identifizierung relevanter Texte ist kostenintensiv.

Es ist daher zu prüfen, inwiefern die genannten Punkte – zumindest für eine Basis-Umsetzung – tatsächlich notwendig sind.

Urheberrechtliche Fragestellungen

Bei einer Realisierung nachrichtenbasierter Frühwarnsysteme müssen zwingend die rechtlichen Vorgaben beachtet werden, die sich insbesondere aus dem *Urheberrecht* (UrhG) ergeben.

Dieses setzt der Vervielfältigung und Bearbeitung von Nachrichten-Texten enge **Grenzen**.

Insbesondere im Falle von Datenbanken sowie Weiter-Veröffentlichungen können Probleme auftreten.

Demgegenüber stehen zahlreiche **Ausnahmen**, insbesondere in Bezug auf vorübergehende Vervielfältigungshandlungen sowie Zeitungsartikel und Rundfunkkommentare.

Hier wird aufgrund der hohen Komplexität des UrhG zur Absicherung anwaltlicher Rat empfohlen.

Kontakt

Dr. Dimitrios Geromichalos
Founder / CEO
RiskDataScience UG (haftungsbeschränkt)
Theresienhöhe 28, 80339 München
E-Mail: riskdatascience@web.de
Telefon: +4989244407277, Fax: +4989244407001

Internet: www.riskdatascience.net