Zusammenfassung

Für viele Unternehmen stellt die Beurteilung der Zahlungsmoral ihrer Kunden eine wichtige Aufgabe dar. Probleme hinsichtlich der Bonität bzw. Zuverlässigkeit können sich auf einem weiten Spektrum zwischen Zahlungsverzug und komplettem Ausfall befinden. Zwar ist die Schwere des Verzugs auf Ebene eines individuellen Kunden gering, allerdings handelt es sich um ein relativ häufiges Problem. Zinsen, die sich hätten erzielen lassen, wenn das Geld rechtzeitig eingegangen wäre, können auf Unternehmensebene hohe Summe erreichen. Diese Thematik wird aufgrund steigender Zinssätze in Zukunft noch weiter an Brisanz gewinnen.

In meiner Fallstudie habe ich analog zum typischen Vorgehen für Credit-Scoring das Problem des Zahlungsverzugs anhand von realen Daten eines Versicherungsunternehmens untersucht. Hierbei habe ich sowohl mittels logistischer Regression als auch Random Forests Modelle mit Daten vor der Covid-19-Pandemie trainiert. Als Zielvariable habe ich verschiedene Varianten des Zahlungsverzugs (von allgemein bis hin zu begrenzt auf längere Verzugsdauern) untersucht.

Ich konnte zeigen, dass beide Modellarten mit einem AUC von mehr als 0,8 sehr gute Vorhersagen auf einem Testdatensatz treffen. In beiden Fällen konnten als wichtigste Einflussgrößen für den Zahlungsverzug vergangene Verzüge und Mahnverfahren, aber auch die Zahlweise identifizieren werden.

Für fast alle Zielvariablen ließen sich auch bei Anwendung der Modelle auf Zeiträume während der Covid-19-Pandemie hohe AUC-Werte (mehr als 0,8) erzielen. Insbesondere für spätere Zeitpunkte (Zahlungsverzug in 2021) war die Güte der Vorhersage wieder ähnlich gut wie vor der Pandemie. Dies deutet auf die andauernde Gültigkeit der trainierten Modelle und der identifizierten Einflussgrößen hin.

Schlagwörter: Data-Science, Credit-Scoring, Zahlungsverzug, logistische Regression, Random Forest, Modellvergleich, Covid-19-Pandemie

Abstract

Analyzing and predicting a customer's payment morale is crucial to a variety of companies. The range of problems concerning solvency or responsibility is very wide and includes both payment delays and default. Regarding a single customer, delays might be seen as a minor inconvenience, however, they constitute a rather frequent problem. If all customers had paid on time, the firm could have invested this money earlier, resulting in possibly high interest earnings at company level. Due to rising interest rates this topic will continue to gain in importance.

In this case study, I examine payment delays using real world data from an insurance company. My approach builds upon the one typically applied in credit scoring. Methodologically, I train logistic regression models as well as random forests with data preceding the COVID-19 pandemic. Concerning the target variable, I consider a variety of payment delays (with a general scope as well as focused on longer delays).

I could show that both model types yield very good AUC values of more than 0.8 when applied to a test data set. In both cases previous delays and requests for payment, but also the mode of payment turned out to be the most important influencing variables.

Using these models for prediction during the pandemic also yielded high AUC values of more than 0.8 for most target variables. In particular for 2021 (in comparison with 2020), prediction quality reached similar levels as seen before the pandemic. This finding hints at the models' persisting applicability.

Key words: Data science, credit scoring, payment delays, logistic regression, random forest, model comparison, COVID-19 pandemic