

## Erweiterte Kreditrisikoanalyse für Schweizer Firmen

Wir untersuchen den Erklärungsgehalt von Zahlungsmeldungen auf die Wahrscheinlichkeit eines Kreditausfalls von Schweizer Firmen. Ziel ist es, ein KMU-Bonitätsrating mit Verfahren des Machine Learnings zu verbessern. Basierend auf einem zwanzigjährigen Datensatz wurde in einer vorhergehenden Projektarbeit für jede Firma ein Zahlungsprofil erstellt.

Aufbauend darauf werden in dieser Arbeit mit logistischer Regression, Decision Trees und Random Forests die Abhängigkeiten von diesen Merkmalen genauer untersucht. Zur Evaluation der einzelnen Methoden wird die Receiver Operating Characteristic (ROC) Kurve benutzt und die Areas under the Curve (AUC) werden miteinander verglichen. Zudem wird versucht, durch Under- und Oversampling die Qualität der Vorhersage zu verbessern.

Für alle Methoden hat im Fall einer univariaten Analyse der mittlere Zahlungsverzug mit einer AUC von 75 % den höchsten Erklärungsgehalt. Dabei spielt es keine signifikante Rolle, ob das gleichgewichtete Mittel oder das betragsgewichtete Mittel der Zahlungsverzüge betrachtet wird. Die Ausfallwahrscheinlichkeit steigt dabei linear an, bis sie bei einem Zahlungsverzug von etwa 100 Tagen eine Sättigung erreicht.

Bei einer multivariaten Analyse ergibt die logistische Regression keine Verbesserung der AUC und ist dementsprechend bei 75 %. Die Methode der Decision Trees gemäss CART Algorithmus ergibt sogar eine Verschlechterung der AUC.

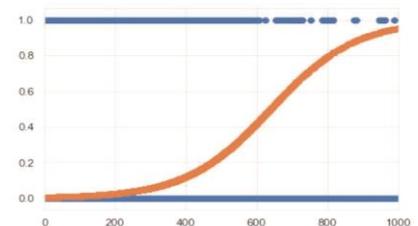
Der trainierte Random Forest konnte mit 76 % Genauigkeit das beste Resultat erreichen. Diese minimale Verbesserung des Erklärungsgehalts rechtfertigt jedoch nicht die zusätzliche Komplexität dieses Modells. Insbesondere verlangt das Einsatzgebiet im Credit Rating von Privatfirmen eine klare Begründung der Ergebnisse, was mit Random Forests nicht gegeben ist.

Zusammenfassend lässt sich sagen, dass die KMU-Ausfallwahrscheinlichkeit durch eine relativ einfache univariate Analyse vom mittleren Zahlungsverzug gut modelliert wird. Solange nur Zahlungsverzüge als erklärende Variablen betrachtet werden, ergeben multivariate Modelle unter Einsatz von Machine-Learning-Methoden keine statistisch signifikante Verbesserung. In einem nächsten Schritt sollte die Analyse daher auf andersartige erklärende Variablen ausgeweitet werden, wie Betriebsdaten oder finanzielle Kennzahlen.



Diplomierende  
Matteo Rezzonico  
Roman Sutter

Dozent  
Christoph Schmidhuber



Ein Teil der Auswertungen, welche mit der logistischen Regression entstanden sind. Zu sehen sind die logistische Regressionskurve und die Receiver Operating Characteristic (ROC) Kurve.