Klassifikation von Firmenkunden anhand eines Supervised Machine Learning Ansatzes

Studiengang: MAS Data Science

«Wird das ein guter Kunde?», fragt sich die Versicherungsfachfrau, welche gerade eine neue Offerte für die Firma FooBar erstellt und versendet. «Könnte diese Frage nicht durch das Wissen in unseren bereits vorhandenen Daten beantwortet werden?», antwortet ihr Team-Kollege, welcher bereits einmal etwas über Machine Learning Modelle gelesen hat.

Einleitung

Das Ziel ist es, einen Profitabilitätswert für die bereits versicherten Kunden zu berechnen. Der Profitabilitätswert ist ein Mass, welches ausdrückt, wie viel ein Kunde zum Versichertenbestand, in diesem Fall dem Mindestquoten-Bestand beiträgt. Anschliessend soll der, durch Kumulation über das versicherte Kollektiv ermittelte Wert, mittels Daten, welche nicht für die Berechnungen benötigt wurden, (Daten, welche bereits bei der Offertenerstellung vorhanden sind) vorhergesagt werden.

Methoden

In der Arbeit wurden verschiedene «supervised» Klassifikationsalgorithmen implementiert, ihre Parameter abgestimmt und die Resultate miteinander verglichen. Dabei wurden Binäre-Klassifikationen sowie Multiklassen-Klassifikationen durchgeführt. In der Multiklassen-Klassifikationen wurden die Modelle auf Genauigkeit (Accuarcy) optimiert. Bei der Binäre-Klassifikationen wurde zusätzlich auf dem unausgeglichenen Datensätzen aus verschiedenen Verfahren welche mit Über- und Unterabtastung (over- and under-sampling) umgehen können die optimale ermittelt (Bild 1). Die Modelle in der binären Klassifikation wurden auf den F1-Wert optimiert.

Ergebnisse

Algorithmen, welche auf Entscheidungsbäumen basieren, lieferten für diese Daten die besten Resultate. Multiklassen-Klassifikationen: Der Kunde sollte in eine von sechs Kategorien (A = trägt viel zum Bestand bei, - E = profitiert vom Bestand) eingeteilt werden. Bei der Kategorie E war die Treffsicherheit sehr schlecht, was sich negativ auf die Qualität des gesamten Modells auswirkte. Hier wurde auf dem Testdatensatz eine Genauigkeit von 0.6954 erreicht. Binäre-Klassifikationen: Auf einem ausgeglichenen Datensatz wurde ein Modell erstellt, welches gute von weniger guten Kunden sehr genau unterscheiden kann der F1 Wert auf dem Testdatensatz beträgt: 0.8628. Bei dem unausgeglichenen Datensatz wurde ein etwas schlechteres Resultat erreicht: 0.6881.

Diskussion

Die Abgrenzung in der Multiklassen-Klassifikation ist zu unscharf. Die binäre Klassifikation könnte gut in der Praxis eingesetzt werden. Wie anhand eines kurzen Versuchs aufgezeigt wurde, besteht durch die Anwendung einer Regression das Potential, eine noch genauere Vorhersage zu erzielen.

Alex Mosimann
alex.mosimann@gmail.com

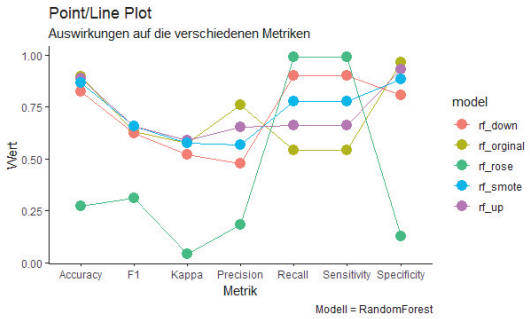

Bild: Auswirkung der verschiedenen Methoden auf die Metriken