Vorhersage des Kundenwerts in einer Versicherungsgesellschaft mittels Machine Learning

Marc Aurel Kennoe
Referent: Prof. Dr. Arnim Malcherek
Korreferent: Prof. Dr. Horst Zisgen
Hochschule Darmstadt
Fachbereiche Mathematik und Naturwissenschaften & Informatik

Motivation und Ziel

Die heutige wettbewerbsorientierte Wirtschaft zeichnet sich dadurch aus, dass sich Unternehmen zunehmend auf ein datengesteuertes Geschäftsmodell stützen. Eine solche Ausrichtung ist dadurch gekennzeichnet, dass Daten die Grundlage für die Entscheidungsfindung bilden. Kundendaten sind besonders wertvoll, vor allem in der Versicherungsbranche, in der der Kunde der Kern des Geschäfts ist. Daher ist es für jedes Versicherungsunternehmen entscheidend, seine Kunden gut zu kennen, um profitable Kunden langfristig zu binden und das begrenzte Budget effizient zu planen. In diesem Zusammenhang ist das Wissen um den zukünftigen Wert eines Kunden entscheidend für ein erfolgreiches Kundenmanagement.
In dieser Arbeit wurde die Effizienz von Machine Learning-Methoden bei der Vorhersage des Kundenwerts in einem Versicherungsunternehmen untersucht. Aufgrund der Asymmetrie der Verteilung des Kundenwerts wurde neben den üblichen Regressionsmodellen für den bedingten Mittelwert auch der Median des Kundenwerts im Rahmen einer Quantilsregression modelliert. Insbesondere soll in dieser Studie herausgefunden werden, wann der Einsatz von Machine Learning besonders wertvoll ist und ob sich die klassische Mittelwertregression und die Quantilsregression in Bezug auf die Prognosefehler unterscheiden.

Vorgehensweise

Zielvariable: Kundenwert des nächsten Jahres ($kundenwert_{t_1}$), Vorhersage mit Daten zum Zeitpunkt t_0

Verwendete Regressionsmodelle: Multiple lineare Regression, Random-Forest-Regression, XGBoost-Regression und Quantilsregression für das 0,5-Quantil (Median).

Modellierungsansätze

- Im ersten Ansatz wurden die Modelle ohne Manipulation von Ausreißern trainiert
- Im zweiten Ansatz wurden Datenpunkte mit einem Kundenwert oberhalb des 99-Quantils entfernt, um die Auswirkungen potenzieller Ausreißer zu kontrollieren
- Im dritten Ansatz wurde vor der Regressionsmodellierung eine Klassifizierung (mit einem Gradient Boosting Classifier) durchgeführt. Die Klassifizierung zielte darauf ab, diejenigen Kunden zu identifizieren, deren Wert sich potenziell nicht ändern würde. Die Regression wurde dann für die Kundengruppe durchgeführt, für die das Klassifikationsmodell eine Änderung des Kundenwerts vorhersagte

Ergebnisse

Die besten Regressionsmodelle (QUANTILSREG ($\tau = 0.5$) und MULTIPLE LINEAR REG) wurden mit einem naiven Status-quo-Modell basierend auf die HIT_RATE (Trefferquote auf der Grundlage von vordefinierten Quantilgruppen des Kundenwerts) und den RAE (Relative Absolute Error) verglichen. Das Status-Quo-Modell geht dabei einfach davon aus, dass der Kundenwert im Laufe der Zeit konstant bleibt. Die nachstehenden Tabellen zeigen, dass die Verwendung des Klassifizierungsschritts vor der Regressionsmodellierung eine positive Auswirkung auf die Prognosegüte der linearen Modelle hat.
Die Testdaten wurden ebenfalls auf der Grundlage der Entwicklung des Kundenwerts von einem Jahr zum nächsten in vier Gruppen eingeteilt, um zu schauen, bei welchen Kunden die Modelle besonders gute Prognose liefern. Die nachstehende Abbildung zeigt, dass das naive Modell im Vergleich zu den Machine Learning-Modellen die schlechtesten Vorhersagen für die vierte Gruppe liefert, wenn die Entwicklung des Kundenwerts stark positiv ist. Gerade in diesem Bereich hebt sich die Quantilsregression positiv von den anderen Modellen ab.

Modell	HIT_RATE (%)	RAE (%)
QUANTILSREG ($\tau = 0.5$)	70.7	14.074
MULTIPLE LINEAR REG	74.3	14.623
STATUSQUO	89.3	14.407

Tabelle 1: Vergleich von Regressions- und Status-quo-Modellen ohne vorherige Klassifizierung

Modell	HIT_RATE (%)	RAE (%)
QUANTILSREG ($\tau = 0.5$)	88.9	13.971
MULTIPLE LINEAR REG	88.2	14.188
STATUSQUO	89.3	14.407

Tabelle 2: Vergleich von Regressions- und Status-quo-Modellen mit vorheriger Klassifizierung

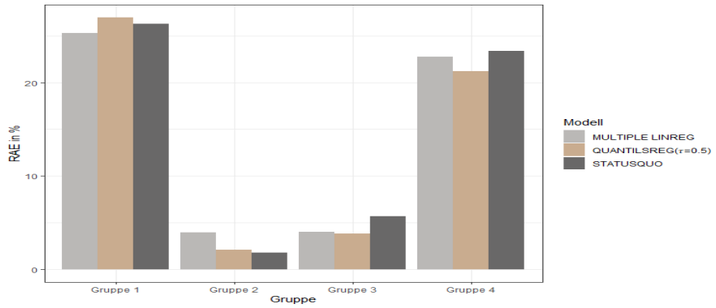

Abbildung 1: Vergleich der Modelle auf verschiedenen Gruppen mit zunehmender Entwicklung des Kundenwerts

Fazit und Ausblick

Auf den ersten Blick liefert das naive Status-quo-Modell zufriedenstellende Ergebnisse für den gesamten Testdatensatz. Die Effizienz der trainierten Modelle wird jedoch besonders deutlich, wenn Kunden isoliert werden, für die keine Veränderung des Kundenwerts zu erwarten ist. Die detaillierte Analyse der Prognosefehler in verschiedenen Kundengruppen hat ebenfalls die Wirksamkeit der Machine Learning-Modelle bewiesen, wenn sich der Wert eines Kunden tatsächlich von einem Jahr zum nächsten ändert. Insgesamt liefert die Medianregression geringere Vorhersagefehler als die klassischen Modelle, die den bedingten Mittelwert modellieren, insbesondere wenn der Kundenwert stark ansteigt.
Darauf aufbauend könnten sich weitere Studien stärker auf die Quantilsregression konzentrieren, um zu untersuchen, wie der Einfluss der Prädiktoren entlang der gesamten bedingten Verteilung des Kundenwertes variiert. Eine interessante Aufgabe könnte es auch sein, das optimale Quantil für die Modellierung entweder der gesamten Daten oder bestimmter Kundengruppen zu finden.

Literatur

[1] Dries Benoit und Dirk Van den Poel. “Benefits of Quantile Regression for the Analysis of Customer Lifetime Value in a Contractual Setting: An Application in Financial Services”. In: Expert Syst. Appl. 36 (Sep.2009), S. 10475–10484.

[2] Bas Donkers, Peter C. Verhoef und Martijn G. de Jong. “Modeling CLV: A test of competing models in the insurance industry”. In: Quantitative Marketing and Economics 5 (2007), S. 163–190.