Verfahren des maschinellen Lernens zur Entscheidungsunterstützung

D I S S E R T A T I O N

zur Erlangung des akademischen Grades
doctor rerum politicarum
(Doktor der Wirtschaftswissenschaften)

eingereicht an der

Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Artem Bequé

Präsident/Präsidentin der Humboldt-Universität zu Berlin:

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan/Dekanin der Wirtschaftswissenschaftlichen Fakultät:

Prof. Dr. Daniel Klapper

Erstgutachter: Prof. Dr. Stefan Lessmann

Zweitgutachter: Prof. Dr. Wolfgang Härdle

Tag des Kolloquiums: 30.08.2018

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe nur unter Verwendung der angeführten Literatur angefertigt habe.

Artem Bequé

Erklärung zum Promotionsvorhaben

Hiermit erkläre ich, dass ich zuvor noch keiner Promotionsprüfung unterzogen wurde sowie ich mich noch um keine Zulassung an der Humboldt-Universität zu Berlin bzw. einer anderen Universität beworben habe. Weiterhin habe ich noch keiner Universität oder ähnlichen Einrichtung eine Dissertation vorgelegt.

Artem Bequé

- Meinen Eltern -

Zusammenfassung

Erfolgreiche Unternehmen denken intensiv über den eigentlichen Nutzen ihres Unternehmens für Kunden nach. Diese versuchen, ihrer Konkurrenz voraus zu sein, und zwar durch gute Ideen, Innovationen und Kreativität. Dabei wird Erfolg anhand von Metriken gemessen, wie z.B. der Anzahl der loyalen Kunden oder der Anzahl der Käufer. Gegeben, dass der Wettbewerb durch die Globalisierung, Deregulierung und technologische Innovation in den letzten Jahren angewachsen ist, spielen die richtigen Entscheidungen für den Erfolg gerade im operativen Geschäft der sämtlichen Bereiche des Unternehmens eine zentrale Rolle.

Um die Entscheidungen zu treffen, welche zum Erfolg führen, sammeln die Unternehmen riesige Datenbestände über ihre Kunden, die Konkurrenz oder allgemein die Lage auf dem Markt. Die Verfügbarkeit dieser großen Datenmengen ergibt sich aus dem umfassenden Einsatz von Informations- und Kommunikationssystemen in den unterschiedlichen Unternehmensbereichen. Diese Daten werden analysiert. Basierend auf diesen Analysen werden Reports erstellt, welche die operativen Entscheidungen unterstützen.

Entscheidungen spielen beispielsweise bei Klassifikationsproblemen eine entscheidende Rolle. Dort ist es häufig notwendig, anhand der Datenmenge über die Gruppenzugehörigkeit der Kunden zu entscheiden. Zum Beispiel wird im Bereich des Credit Scoring mithilfe von historischen Daten täglich entschieden, ob ein Kunde seinen Kredit inklusive die dazugehörigen Zinsen zurückzahlt oder nicht. Ein weiteres Beispiel ist das Direktmarketing, wobei hier die Kunden in zwei Gruppen klassifiziert werden müssen: in Kunden, die auf eine bestimmte Marketing Campaign reagieren und in eine andere, die nicht reagieren.

Die Entscheidungen, welche auf Prognosemodellen basieren, führen ggf. zum besseren Erfolg. Methoden der klassischen Statistik oder die moderne Verfahren des maschinellen Lernens repräsentieren solche Progrnosenmodelle. Verfahren des maschinellen Lernens sind in der Lage, die Analyse existierender Datenbestände praktisch unbeaufsichtigt durchzuführen, mögliche Zusammenhänge zu erkennen und die Wahrscheinlichkeiten zu ermitteln, welche die Grundlage für die Entscheidungen darstellen. Außerdem verfügen gerade die Verfahren des maschinellen Lernens einen hohen Automatisierungsgrad, was sie sehr geignet für die Integration in die existierenden Systeme macht.

Vor diesem Hintergrund entstammen die in der vorliegenden Arbeit zur Evaluation der Methoden des maschinellen Lernens untersuchten Entscheidungsprobleme vornehmlich der *Entscheidungsunterstützung*. Hierzu gehören Klassifikationsprobleme wie die Kreditwürdigkeitsprüfung im Bereich Credit Scoring und die Effizienz der Marketing Campaigns im Bereich Direktmarketing. In diesem Kontext ergaben sich Fragestellungen für die korrelativen Modelle, nämlich die Untersuchung der Eignung der Verfahren des maschinellen Lernens für den Bereich des Credit Scoring, die Kalibrierung der Wahrscheinlichkeiten, welche mithilfe von Verfahren des maschinellen Lernens erzeugt werden sowie die Konzeption und Umsetzung einer Synergie-Heuristik zwischen den Methoden der klassischen Statistik und Verfahren des maschinellen Lernens. Desweiteren wurden kausale Modelle für den Bereich Direktmarketing (sog. Uplift-Effekte) angesprochen. Diese Themen wurden im Rahmen von breit angelegten empirischen Studien bearbeitet.

Zusammenfassend ergibt sich, dass der Einsatz der untersuchten Verfahren beim

derzeitigen Stand der Forschung zur Lösung praxisrelevanter Entscheidungsprobleme sowie spezifischer Fragestellungen, welche aus den besonderen Anforderungen der betrachteten Anwendungen abgeleitet wurden, einen wesentlichen Beitrag leistet. Dieser besteht darin, dass der Entwurf eines ganzheitlichen, methodisch konsistenten Vorgehensmodells bei der Lösung betriebswirtschaftlicher Klassifikationsprobleme mittels Verfahren des maschinellen Lernens gegenüber anderen Verfahren wie den klassischen Methoden der Statistik hinsichtlich der Prognosegüte und anderer Dimensionen signifikant überlegen ist.

Abstract

Nowadays right decisions, being it strategic or operative, are important for every company, since these contribute directly to an overall success. This success can be measured based on quantitative metrics, for example, by the number of loyal customers or the number of incremental purchases. These decisions are typically made based on the historical data that relates to all functions of the company in general and to customers in particular. Thus, companies seek to store an enormous amount of data in databases and data cubes, analyze it and apply obtained knowlegde in decision making. Classification problems represent an example of such decisions. For instance, in credit scoring it is necessary to classify the customers into "bad" or "good" category, where the former represent customers who are not able to re-pay their credit lines, whereas the latter do. Another example is direct marketing where customers are classified for the purpose of marketing campaings. Classification problems are best solved, when techniques of classical statistics and these of machine learning are applied, since both of them are able to analyze huge amount of data, detect dependencies of the data patterns, and produce probability, which represents the basis for the decision making. Especially, the techniques of machine learning are quite popular, as they have high potential being completely automated and integrated into the existing systems of the company. In this study, I apply these techniques and examine their suitability based on correlative models for decision making in credit scoring and further extend the work by causal predictive models for direct marketing. In detail, I analyze the suitability of techniques of machine learning for credit scoring alongside multiple dimensions, I examine the ability to produce calibrated probabilities and apply techniques to improve the probability estimations, and seek for the best combination between the last two. I further develop and propose a synergy heuristic between the methods of classical statistics and techniques of machine learning to improve the prediction quality of the former, and finally apply conversion models to turn machine learning techqiques to account for causal relationship between Marketing Campaigns and customer behavior in direct marketing. The study has shown that the techniques of machine learning represent a suitable alternative to the methods of classical statistics for decision making and should be considered not only in research but also should find their practical application in real-world practices.

Inhaltsverzeichnis

Teil I

Begründung des thematischen Zusammenhangs

1 Verfahren des maschinellen Lernens zur Entscheidungsunterstützung

1.1 Thematische Einordnung

Heutzutage finden die Verfahren des maschinellen Lernens, aber auch die Methoden der klassischen Statistik in unterschiedlichsten Bereichen ihren Einsatz. Zum einen werden diese Verfahren für die Analyse existierender Datenbestände genutzt, um mögliche Zusammenhänge zu erkennen. Zum anderen generieren sie Wahrscheinlichkeiten, welche eine Grundlage für Entscheidungen darstellen, die in unterschiedlichsten Bereichen getroffen werden müssen. Gerade aufgrund der Rolle, welche die Verfahren des maschinellen Lernens bei der *Entscheidungsunterstützung* besitzen, sind sie von großer Bedeutung. Das aktuelle Interesse an diesen Verfahren nimmt ferner dadurch zu, dass diese Methoden über einen hohen Automatisierungsgrad verfügen und direkt in das operative Geschäft in sämtlichen Unternehmensbereichen integriert werden können.

Eine verbesserte *Entscheidungsunterstützung* ist vor dem Hintergrund der Wettbewerbsintensivierung, welche über die letzten Jahre stark zugenommen hat, von entscheidender Bedeutung. Der Wettbewerb ist insbesondere durch die Globalisierung, Deregulierung und technologische Innovation in den letzten Jahren angewachsen. Dadurch ergibt sich im operativen Geschäft ein starker Druck, Entscheidungen in *real-time* treffen zu müssen. Unternehmen, die nicht in der Lage sind, in *real-time* Entscheidungen zu treffen, die zu geschäftlichen Erfolgen führen - beispielsweise in der Erhöhung der Kundenbasis oder der Anzahl der Käufer -, verlieren dadurch ihre Konkurrenzfähigkeit. Dies wird sich zum Beispiel an der abnehmenden Kundenloyalität oder den steigenden Kosten bei der Akquise von neuen Kunden zeigen. Solche Entscheidungen sind insbesondere im *Onlinehandel* von Belang.

Als ein Beispiel solcher Bereiche ist das Credit Scoring zu nennen. Mit dem Credit Scoring können die Entscheidungen über die Kreditwürdigkeit der Kunden zu extremen Verlusten führen. Wenn die Kunden, welche laut der ermittelten Wahrscheinlichkeit ihre Kreditlinien tilgen, keine Kredite bekommen, verlieren die Banken oder Finanzinstitute ihre Profite. Falls diese aber laut der Wahrscheinlichkeit nicht in der Lage sind, ihre Kredit zurückzuzahlen, und dann trotzdem einen Kredit bekommen, erleiden die Banken oder Finanzinstitute ggf. größere Verluste. In beiden Fällen werden die Verfahren des maschinellen Lernens eingesetzt, um die Wahrscheinlichkeiten zu ermitteln, welche die Grundlage für die Entscheidungen darstellen, deswegen ist die Verständlichkeit der Prognosemodelle besonders gefragt. Im Bereich *Online peer-to-peer crediting* werden von den Verfahren zudem weitere Eigenschaften verlangt, nämlich ein hoher Automatisierungsgrad und eine große Geschwindigkeit.

Im Bereich des Direktmarketings spielt die *Entscheidungsunterstützung* ebenfalls eine besondere Rolle. Heutzutage arbeiten sowohl finanziell etablierte Unternehmen als auch junge Start-Ups mit einer breiten Palette an Tools, die eine Integration der Verfahren des maschinellen Lernens ermöglichen. Hierbei führen ihre Vorteile zu einer Kostenreduzierung sowie einer Umsatzerhöhung bedingt durch ein besseres Targeting von Kunden, das in dem Fall durch diese Verfahren personalisiert wird. Als personalisiertes Targeting kann beispielsweise eine Marketing-Kampagne bezeichnet werden, welche zum richtigen

Zeitpunkt, zur richtigen Person und mit dem richtigen Produktangebot durchgeführt wird. Diese Maßnahmen führen zur Erhöhung von Konversionsraten, einer höheren Kundenzufriedenheit und eventuell einem besseren Image des Unternehmens auf dem Markt.

Die Verfahren des maschinellen Lernens werden direkt in die Systeme der Unternehmen integriert. Unter Integration wird hierbei ein Prozess verstanden, welches direkt im Back-End entwickelt wird. Dieses beginnt mit der Erfassung, Speicherung und Harmonisierung der Daten der einzelnen Kunden in die Datenbank, geht über zur automatisierten Auswertung dieser Daten, wendet die Verfahren des maschinellen Lernens an und erzeugt schließlich eine Wahrscheinlichkeit, anhand derer eine Entscheidung getroffen wird.

Die *Entscheidungsunterstützung* bildet den betriebswirtschaftlichen Rahmen der vorliegenden Arbeit. Ein Großteil der empirisch untersuchten Fragestellungen entstammt diesem Anwendungsfeld. Dabei werden ausschließlich Klassifikationsprobleme modelliert, das heißt, dass eine Entscheidung jeweils durch die Einordnung eines Objekts, beispielsweise eines Kunden, in eine von mehreren vordefinierten Gruppen repräsentiert wird. Wie oben bereits beschrieben, bedingt der operative Charakter der untersuchten Problemstellungen einen hohen Automatisierungsgrad sowie eine hohe Geschwindigkeit. Ferner ist die Verständlichkeit der Verfahren gefragt. Die vorliegende Arbeit dokumentiert und untersucht empirisch die Relevanz dieser Themenstellungen. Die Verfahren des maschinellen Lernens basieren dabei auf der mathematischen Optimierung und untersuchen einen funktionalen Zusammenhang zwischen vorliegenden Beispieldaten und einer zu modellierenden diskreten Zielgröße. Die Lösung einer solchen Optimierung mittels exakter Verfahren oder intelligenter Heuristiken gehört zu den Kernkompetenzen der Forschung im Bereich des maschinellen Lernens.

Die Dissertation besitzt gemäß der vorangehenden Darstellung einen interdisziplinären Charakter. Es sollen betriebswirtschaftliche Fragestellungen als Klassifikationsproblem abgebildet und durch Einsatz von statistischen Verfahren und Verfahren des maschinellen Lernens gelöst werden. Entsprechend dem Kerngedanken der Wirtschaftsinformatik wird dabei ein prozessorientierter Ansatz verfolgt und versucht, die Belastung des eigentlichen Entscheiders durch Anwendung der Verfahren des maschinellen Lernens strikt zu begrenzen.

1.2 Zielsetzung und Motivation

Im Mittelpunkt der Arbeit stehen die Verfahren des maschinellen Lernens, welche auf ihre Eignung für die Unterstützung ausgewählter betriebswirtschaftlicher Entscheidungsprobleme untersucht werden. Alle Methoden gehören zu Prognoseverfahren und ermöglichen die Vorhersage einer Gruppenzugehörigkeit auf der Basis vorliegender Beispieldatensätze. Fragestellungen dieser Art sowie entsprechende Lösungsmethoden werden in der Statistik schon seit vielen Jahren untersucht.

Die Arbeit hat sich auf die beiden Domänen Credit Scoring und Direktmarketing konzentriert. Eine zentrale Fragestellung im Bereich des Credit Scoring bezieht sich darauf, ob ein Kreditnehmer seinen Kredit zurückzahlt (sog. non-defaulter) oder nicht

(sog. defaulter). Die Verfahren des maschinellen Lernens werden hierbei dazu benutzt, die Wahrscheinlichkeit für die Kreditrückzahlung zu ermitteln. Diese unterstützt die Entscheidung bei der Kreditwürdigkeitsprüfung, wobei hier zwei mögliche Fehler auftreten können: Wird ein Kredit abgelehnt, obwohl vorausgesagt wird, dass der Kunde den Kredit zurückzahlen würde, so verliert die Bank die möglichen Profite; wird ein Kredit trotz der Vorhersage der fehlenden Tilgung vergeben, so trägt die Bank offensichtlich höhere Kosten.

Eine konkrete Fragestellung im Direktmarketing ist beispielsweise, ob ein Kunde (wobei es keine Rolle spielt, ob er neu oder nicht neu ist) auf eine Marketing Campaign reagiert. Die Verfahren des maschinellen Lernens ermitteln die Wahrscheinlichkeit dafür, ob ein Kunde kontaktiert werden soll. Diese unterstützt wiederum die Entscheidung im Bereich Marketing. Hier kann man wieder von zwei Szenarien ausgehen. Wird ein Kunde durch eine Marketing Campaign trotz geringer Wahrscheinlichkeit für ein Reagieren kontaktiert, verliert das Unternehmen die Kosten, welche durch die Marketing Campaign verursacht werden. Im gegenteiligen Fall verliert das Unternehmen ggf. die Profite, welche dieser Kunde erbringen könnte.

Die Verfahren des maschinellen Lernens wie die Methoden der klassischen Statistik führen durch ihre direkte Einbindung in den Prozess der Entscheidung zu einer *Entscheidungsunterstützung* und demzufolge einen unabdingbaren Einfluss auf das Leistungsverhalten und Wachstum des Unternehmens sowie die Erhöhung der Profitabilität. Einbindung der Verfahren des maschinellen Lernens findet in Initiativen und Strategien wie dem *customer relationship management* sowohl im Credit Scoring als auch im Direktmarketing ihre zielgerichtete Anwendung.

Die wesentliche Motivation der Arbeit besteht in der Untersuchung, in wie weit die Verfahren des maschinellen Lernens zur Lösung ausgewählter Klassifikationsprobleme aus der Betriebswirtschaft zielführend eingesetzt werden können. Die Verfahren des maschinellen Lernens werden dabei durch den Vergleich mit etablierten Alternativen empirisch validiert. So sollen diese Verfahren mit den Methoden der klassischen Statistik in breit angelegten Experimentdesigns verglichen werden. Im Bereich Credit Scoring wird die logistische Regression als ein vorgegebener Standard angesehen, die also unabdingbar in die Vergleiche aufgenommen werden muss. Die Vergleiche werden durch statistische Testverfahren abgesichert. Diese anwendungsorientierte Potentialanalyse solcher Methoden ist die Methodik der vorliegenden Arbeit.

Zusammenfassend bietet die Arbeit durch den Methodenvergleich und den methodischen Erweiterungen einen wissenschaftlichen Erkenntnisgewinn. Dabei wird ein empirisch-induktiver Forschungsansatz verfolgt, welcher von einer konkreten Problemstellung ausgeht, entsprechend geeignete Experimente durchführt und Ergebnisse liefert, die im günstigen Fall zu verallgemeinerungsfähigen Erkenntnissen führen.

Konkret wurden in den vier Fachartikeln, die im Rahmen der Promotion eingebracht werden, folgende Fragestellungen behandelt:

1. Die Untersuchung des Potentials von *extreme learning machines* für den Bereich Credit Scoring anhand mehrerer Dimensionen: *ease of use* (d.h. wie leicht die Methode einzusetzen ist), *computational complexity* (d.h. wie aufwendig das Verfahren

im Betrieb ist) und *predictive accuracy* (d.h. die Güte der Prognosen). Die Methode wird anderen Verfahren des maschinellen Lernens sowie klassischen Methoden der Statistik gegenübergestellt. Außerdem wird das Verfahren in Rahmen von zwei ensemble Techniken untersucht.

2. Die unter 1. genannte Studie wird durch folgende Fragestellungen ergänzt: In wie weit sind die Wahrscheinlichkeiten der Verfahren des maschinellen Lernens kalibriert, welche Techniken können diese Wahrscheinlichkeiten kalibrieren und welche Kombination aus den Methoden der Kalibrierung und dem Verfahren der Modellierung der Klassifizierung funktioniert am besten? In dieser Studie werden weitere Verfahren des maschinellen Lernens herangezogen, die davor nicht angesprochen wurden.

3. Darauf aufbauend wird eine Heuristik der Synergie zwischen den Verfahren des maschinellen Lernens und Methoden der klassischen Statistik entwickelt. Die Vorteile der Verfahren des maschinellen Lernens werden in die Methode der klassischen Statistik auf verschiedene Weise integriert und empirisch hinsichtlich der Prognosegüte untersucht.

4. Während die ersten drei Fachartikel korrelative Modelle darstellen, werden im vierten Artikel kausale Modelle angesprochen. Kausale Modelle werden zur *Entscheidungsunterstützung* im Bereich Direktmarketing verwendet. Der Artikel befasst sich mit der Modellierung des Erfolges von Marketing Campaigns durch einen Vergleich der Verfahren des maschinellen Lernens mit den klassischen Methoden der Statistik. Darüber hinaus werden Konversions-Methoden für Uplift-Effekte herangezogen, welche die Kausalität zwischen einer Marketing Campaign und dem Verhalten der Kunden vorhersagen. Im Mittelpunkt der Studie steht die Frage, welche Konversions-Methoden mit welchen Klassifikatoren am besten funktionieren.

1.3 Die Untersuchung der Fragestellungen durch die vier Artikel

Wie bereits erwähnt, wird in Bequé and Lessmann (2017) die Alternative zu den klassischen künstlichen neuronalen Netzwerken - *extreme learning machines* - zur Lösung ausgewähltes Klassifikationsproblems im Bereich des Credit Scoring eingesetzt. Unter Klassifikation wird dabei eine prognostische Ausprägung verstanden. Die Zielvariable also, die in Bequé and Lessmann (2017) betrachtet wurde, stammte aus der Klassifikationsanalyse und lieferte für jeden Kunden eine zugehörige Kategorie: nämlich "hohes/niedriges Risiko" bei der Kreditwürdigkeitsprüfung. Dabei dient die Ermittlung der Wahrscheinlichkeiten als die Grundlage für die *Entscheidungsunterstützung* und soll in möglichst kurzer Zeit erfolgen, so dass ihre Anwendung in den anliegenden Entscheidungen von tatsächlichem Nutzen sein kann. Eine weitere wichtige Dimension ist die Verständlichkeit bzw. die Lesbarkeit der Ergebnisse und das Tuning der Prognosemodelle der Verfahren des maschinellen Lernens im Ganzen und von *extreme learning machines* im Konkreten. Vor diesem Hintergrund erfolgt in Bequé and Lessmann (2017) eine Charakterisierung dieser Dimensionen. Jede dieser logischen Dimensionen wird in einer Benchmark Studie anhand mehrerer Datensätze genau studiert.

Um die davor erwähnten Thesen zu prüfen, wird das Verfahren *extreme learning*

machines in Bequé and Lessmann (2017) sechs anderen Verfahren aus dem Bereich maschinelles Lernen gegenübergestellt, wie *k-nearest neigbour*, *artificial neural networks*, *support verctor machines*, *J4.8* und *CART* sowie *regularized logistic regression* aus der klassischen Statistik. Letztere gilt als absolute Standard-Methode im Bereich Credit Scoring. Die empirische Untersuchung wird anhand von drei verschiedenen Dimensionen betrachtet:

- *Ease of use* – d.h. wie leicht sind die Methoden anzuwenden. Gerade die Verfahren des maschinellen Verfahrens werden stark dafür kritisiert, dass sie nur schwer in das existierende System zu implementieren bzw. das sog. *Tuning* (Parametrisierung) der Verfahren oder die Ergebnisse der Wahrscheinlichkeitsermittlung nur schwer und bedingt interpretierbar sind. Diese Dimension wird anhand zweier Metriken untersucht, einmal anhand der Anzahl der Parameter des *Tunings* des jeweiligen Verfahrens, zum anderen anhand der Sensibilität zur Änderung der Einstellungen dieses Parameters. Das letzte wird mit zwei Metriken untersucht, dem *sensitivity index* und dem *coefficient of variance*.
- *Comptutational complexity* – d.h. wie schwer bzw. wie aufwendig es ist, diese Verfahren anzuwenden. Diese Frage ist besonders relevant, wenn man das Verfahren des maschinellen Lernens in ein System des Credit Scoring integrieren möchte. Gerade für einen Geschäftszweig wie das *online peer-to-peer crediting* ist diese Dimension immer mehr von Belang. Die Zeit und die Speichernutzung sowohl von der *training phase*, d.h. der Phase, in der die Verfahren lernen, als auch der *testing phase*, d.h. dem Zeitraum, in dem die gelernten Verfahren an einem nicht bekannten Datensatz angewandt (verwendet) werden, werden aufgenommen und verglichen.
- *Predictive accuracy* – d.h. die Güte der Vorhersagen der Verfahren. Dies ist die wichtigste Dimension, die betrachtet wird. Sie wird anhand zweier Metriken bewertet: der *Percentage Correctly Classified* und der *Area Under the Curve*. Die Bewertung dieser Dimension wird durch statistische Testverfahren unterstützt.

Extreme learning machines wurde entwickelt, um die Nachteile der neuronalen Netzwerke zu beheben. Dieses neue Verfahren benötigt deutlich weniger Zeit für das Lernen (im engl. training time) als klassische neuronale Netzwerke und das, ohne die Güte der Prognosen negativ zu beeinflussen. Die benötigte Lernzeit ist dabei eine relevante Größe insbesondere für Business Modelle wie das *online peer-to-peer crediting*. Dies führt zu einem empirischen Vergleich unter den Ensemble-Techniken. Es werden zwei populäre Techniken ausgesucht, nämlich *bagging* und *boosting*.

In der nachfolgenden Arbeit (Bequé et al. 2017) folgt zum einem eine vertiefende Betrachtung der Fähigkeit von Verfahren des maschinellen Lernens, die Wahrscheinlichkeiten in guter Qualität zu liefern, zum anderen werden neue Verfahren des maschinellen Lernens, die davor nicht angesprochen wurden, herangezogen. Es geht also um die Untersuchung, ob die Verfahren des maschinellen Lernens in der Lage sind, im Vergleich zu den Methoden der klassischen Statistik kalibrierte Wahrscheinlichkeiten zu liefern. Solche Wahrscheinlichkeiten werden vonseiten des Basel Accord im Credit Scoring gefordert, was die Bedeutung der Untersuchung unterstreicht. Deswegen wird die Studie Bequé et al. (2017) um die Methoden, die sog. Kalibratoren, welche die Klassifikatoren bzw. die Prognosemodelle - seien es klassische statistische oder moderne Verfahren des maschinellen

Lernens - kalibrieren, erweitert.

Bequé et al. (2017) stützt sich auf eine breit angelegte empirische Studie. Konkret werden die Verfahren des maschinellen Lernens (wie z.B. *artificial neural networks*, *ensemble techniques* wie *bagged hill-climbing ensemble selection* oder *random forest*) wiederum den Methoden der klassischen Statistik (logistische Regression) gegenübergestellt. Diese werden mit allen Methoden der Kalibrierung, die zu dem aktuellen Zeitpunkt bekannt sind, zusammen ausgeführt. Insgesamt werden fünf Verfahren für die Klassifikation (Klassifikatoren) und sechs Methoden der Kalibrierung (Kalibratoren) in der Studie herangezogen. Es wird jede mögliche Kombination der beiden Gruppen untersucht.

Zunächst wird der Unterschied zwischen zwei Größen der Güte der Wahrscheinlichkeiten festgelegt. Der konzeptuelle Unterschied zwischen der Kalibrierung und der Fähigkeit, einen Einzelfall richtig zuzuordnen, wird anhand zweier Metriken - *Brier Score* und *Area Under the Curve* - demonstriert. Ferner werden alle Kalibratoren, die zum aktuellen Zeitpunkt existieren, in die Studie einbezogen. Diese werden entsprechend erläutert und dokumentiert. Die Interaktion zwischen Klassifikatoren und Kalibratoren ist eine weitere Forschungsfrage, die bis jetzt nicht untersucht wurde. Zuletzt wird anhand von *Calibration Plots* und der Zerlegung des *Brier Score* untersucht, welche Determinanten der Kalibratoren wesentlich dazu beitragen, das gewünschte Ergebnis zu erzielen.

Aufbauend auf dem erworbenen Wissen wird in Bequé and Lessmann (2018) ein ganzheitliches Vorgehensmodell zur Lösung klassifikatorischer Fragestellungen aus dem Bereich Credit Scoring auf Basis der Verbindung von klassischen Methoden der Statistik mit modernen Verfahren des maschinellen Lernens konzipiert, implementiert und empirisch validiert. Die Heuristik stützt sich auf der einen Seite auf Verfahren des maschinellen Lernens wie *random forest* und *stochastic gradient boosting* und auf der anderen Seite auf die logistische Regression aus der klassischen Statistik. Man findet viele Publikationen, in denen Methoden bzw. ihre Derivate miteinander verglichen werden, allerdings findet man kaum etwas darüber, wo eine Synergie zwischen den Methoden vorgeschlagen wird. Genau mit dieser Frage beschäftigt sich dieser Artikel.

Zuerst wird die Differenz bzgl. der Prognose-Güte zwischen den Verfahren des maschinellen Lernens und der klassischen Statistik in unterschiedlichen Dimensionen dargestellt. Im Einzelnen geht es hier um die *Correctness of Categorical Prediction*, d.h., ob die Klassifikatoren in der Lage sind, die jeweilige Klassenzugehörigkeit zu kategorisieren. Ferner wird die *Qualitity of Probabilistic Prediction* geprüft, d.h. untersucht, inwieweit die Klassifikatoren in der Lage sind, Wahrscheinlichkeiten guter Qualität zu liefern. Darüber hinaus wird die Aufmerksamkeit darauf gerichtet, inwieweit die Klassifikatoren die Kunden zwischen den defaulter (d.h. der Kredit wird nicht getilgt) und den non-defaulter (d.h. der Kredit wird getilgt) unterscheiden. Außerdem wird eine weitere Dimension betrachtet, die in *Expected Maximum Profit* gemessen wird. Ferner werden Dimensionen wie *comprehensibility* und *justifibility* untersucht. Der erste Teil betrachtet die Dimension, inwieweit man die Ergebnisse bzw. das Tuning der Verfahren interpretieren kann. Der zweite Teil beschäftigt sich mit der Frage, ob die Wahrscheinlichkeiten, welche die Verfahren ergeben, gerechtfertigt sind. Wenn z.B. zwei Kreditnehmer ähnliche Profile in Bezug auf alle Merkmale (wie z.B. Alter, Ort etc.) aufweisen und sich nur in ihrem Einkommen unterscheiden, kann es nicht den realen Begebenheiten entsprechen, dass

einem Kreditnehmer, der weniger verdient, ein Kredit gewährt wird, während einem anderen, der mehr verdient, eine Absage für die Aufnahme eines Kredites erteilt wird.

Die vorgeschlagene Heuristik zwischen den Verfahren wird in allen diesen Dimensionen daraufhin geprüft, ob und inwieweit diese zur Verbesserung der Güte der Prognosen beiträgt. Die Heuristik wird wie folgt umgesetzt: Zuerst werden die Verfahren des maschinellen Lernens trainiert. Basierend darauf werden anhand von *variable importance measures* die wichtigsten Merkmale der Kreditnehmer definiert. Jedes Verfahren hat eigene Metriken, anhand derer die Merkmale bewertet werden. Diese werden später in die funktionale Form der logistischen Regression integriert. Die Integration wird in nicht linearer Form sowie in der Form der Interaktion dargestellt. Die diesbezügliche Hypothese ist, dass die Verfahren des maschinellen Lernens, in dem Fall random forest und stochastic gradient boosting, speziell die Interaktionen bzw. Nicht-Linearität zwischen den Merkmalen untersuchen, was die logistische Regression eben nicht leistet. Diese Vorteile werden dann später zugunsten der logistischen Regression integriert.

Während sich die ersten drei Fachartikel mit den korrelativen Modellen beschäftigen, welche eine Grundlage zur *Entscheidungsunterstützung* darstellen, werden im vierten Fachartikel kausale Prognosemodelle untersucht, welche die Qualität der *Entscheidungsunterstützung* erhöhen. Bei der Erweiterung des Themas um kausale Prognosemodelle standen Daten aus dem Bereich Direktmarketing zur Verfügung (Bequé et al., 2018). Im Fokus dieser Untersuchung steht das klassische betriebswirtschaftliche Problem einer Marketing-Abteilung: Ein Klassifikator soll die Kunden identifizieren, welche mit einer Marketing Campaign gezielt aufgespürt werden können bzw. es soll die Wahrscheinlichkeit ermittelt werden, ob ein Kunde auf eine Marketing Campaign reagiert. Betont werden soll, dass die Studie die Konversionmethoden, welche die Kausalität zwischen einer Marketing-Kampagne und dem Verhalten des Kunden (sog. Uplift-Effekte) modellieren, einschließt, was die Klassifikatoren nicht leisten können. Zum Zeitpunkt der Abfassung des Fachartikels fehlten die Empfehlungen, welche Konversions-Methoden bzw. welche Kombination aus Konversions-Methoden und Klassifikatoren am besten funktionieren. Das Ziel der Studie ist es, diese Forschungslücke zu schließen.

Folgende Fragestellungen werden in der empirisch breit angelegten Studie angesprochen: Zuerst wird die Literatur, welche aus verschiedensten Quellen und Domänen stammt, auf konzeptuelle Unterschieden untersucht (erster Teil des Fachartikels). Die ausgewählten Konversions-Methoden werden dann detailliert beschrieben (zweiter Teil). Ferner wird die Leistung der Konversions-Methoden in einer Benchmark-Studie geprüft. Die Studie stützt sich auf 27 Datensätze, welche aus verschiedenen Ländern und verschiedenen Bereichen der E-Commerce kommen. Ferner wird untersucht, wie die Klassifikatoren (aus dem maschinellen Lernen und der Statistik) mit Konversions-Methoden für die Uplift-Modellierung funktionieren. Dazu werden konkrete Empfehlungen ausgearbeitet, welche Klassifikatoren mit welchen Konversions-Methoden am besten funktionieren. Ferner wird die Frage diskutiert, wie die Konversions-Methoden (also zusammen mit den Klassifikatoren) zum Unternehmenswert (*business value*) beitragen. Unter *business value* wird hier die wachsende Anzahl der Verkäufer verstanden. Zuletzt wird die Frage erläutert, wie sich das *response modeling* (d.h. die Anwendung der Klassifikatoren) von der Kombination aus Konversions-Methoden und Klassifikatoren unterscheidet, d.h., wie die Konversions-

Methoden durch die Modellierung der Kausalität zwischen Marketing-Kampagnen und dem Verhalten von Kunden zur Erhöhung des *business value* beitragen. Die Studie verwendet sog. *Qini-Plots* und *Uplift-Gain-Charts*, um die Differenzen zwischen der Güte der Prognose der Methoden genauer zu studieren.

1.4 Ergebnisse

Bequé and Lessmann (2017) hat gezeigt, dass *extreme learning machines* als ein Verfahren des maschinellen Lernens tatsächlich eine denkbare Alternative zu anderen Verfahren darstellt. In Bezug auf *ease of use* hat das Verfahren Nachteile durch die höhere Anzahl der Parameter für das Tuning im Vergleich z.B. zur logistischen Regression oder z.B. zu *k-nearest neigbours*. Aber bei richtiger Parametrisierung hat das Verfahren einen Vorteil in Bezug auf die Sensibilität der Parametrisierung gegenüber z.B. künstlichen neuronalen Netzwerken. Hervorragende Ergebnisse hat das Verfahren in Bezug auf die *computational complexity* gezeigt. Das Verfahren hat eine schnellere Lernphase als alle anderen untersuchten Verfahren; dies gilt insbesondere im Vergleich zu *support vector maschines* und *artificial neural networks*. Von Bedeutung ist, dass es *extreme learning machines* gelingt, diese schnelle Leistung ohne Verlust der Güte der Prognosen zu erzielen. Das Verfahren zeigt somit vergleichbare Ergebnisse in der Dimension *predictive accuracy*. Des Weiteren hat das Verfahren sehr gute Ergebnisse im Rahmen des Regimes der Ensemble-Techniken gezeigt. Das alles spricht dafür, dass das Verfahren des maschinellen Lernens - *extreme learning machines* - in der Tat eine denkbare Alternative für das Credit Scoring darstellt.

Bequé et al. (2017) hat gezeigt, dass die Verfahren des maschinellen Lernens im Vergleich zu klassischen Methoden der Statistik schlecht kalibrierte Wahrscheinlichkeiten erzielen, diese Wahrscheinlichkeiten jedoch mithilfe von Kalibratoren verbessert werden können. Die Kalibratoren führen in der Tat zur Verbesserung der Wahrscheinlichkeiten (gemessen anhand des *Brier Score*), und zwar ohne Verlust der Zuordnungsfähigkeit (gemessen mittels *Area Under the Curve*). Dabei wurden *generalized addivitive models* als der beste von allen existierenden Kalibratoren ermittelt. Dieser besitzt die Fähigkeit, mit allen Klassifikatoren gut zu funktionieren. Die Kombination zwischen *generalized addivitive models* und *random forest* wird besonders wegen der guten Ergebnissen mittels beider Metriken empfohlen.

Bequé and Lessmann (2018) hat gezeigt, dass die logistische Regression eine Methode darstellt, welche über alle untersuchten Dimensionen hinweg gute Ergebnisse erbringt. Allerdings wird immer ein Leistungsunterschied zwischen den Verfahren des maschinellen Lernens und der logistischen Regression festgestellt. Die Studie zeigt, dass die vorgeschlagene Heuristik zur Verbesserung der Güte der Prognosen beiträgt. Insbesondere die Interaktionsterme haben sehr gute Ergebnisse erzielt und zur Verbesserung Prognose-Güte geführt. Die nicht-lineare Integration hat zu keiner Verbesserung geführt. Außerdem hat die Studie deutlich demonstriert, dass *variable importance measures* von *stochastic gradient boosting* deutlich geeigneter für eine Heuristik sind als von *random forest*, was von großer Relevanz für die verwendeten Heuristiken ist.

Bequé et al. (2018) hat gezeigt, dass die neuen Derivate der Konversions-Methoden

nicht zwangsläufig bessere Ergebnisse in der Modellierung der Uplift-Effekte erbringen. Deswegen wird in der Studie empfohlen, bei der Entwicklung neuer Methoden der Konversions-Methoden eine breite Palette an Methoden zu untersuchen, um bessere (d.h. im engl. *comptetitive*) Vergleiche zu ermöglichen. Außerdem zeigt die Studie, dass die Methoden des maschinellen Lernens gegenüber den Methoden der klassischen Statistik in Bereich der Uplift-Modellierung besser abschneiden. So zeigen z.B. *random forest* oder *k-nearest neighbours* bessere Ergebnisse als die logistische Regression. Jedoch erbringen nicht alle Methoden des maschinellen Lernens ausgezeichnete Ergebnisse. Das *stochastic gradient boosting* - eine Methode, welche bevorzugt angewandt wird - hat beispielsweise keine empfehlenswerten Ergebnisse gezeigt. Die Studie zeigt außerdem, dass bei der falschen Wahl der Konversionsmethode bzw. der Kombination aus den Konversions-Methoden und den Klassifikatoren die Response-Modellierung (d.h. Anwendung der Klassifikatoren ohne Konversions-Methoden für Uplift-Effekte) erfolgreicher sein kann. Die Studie zeigt genau auf, welche Konversions-Methoden für Uplift-Effekte mit welchen Klassifikatoren zusammen am besten funktionieren.

1.5 Konklusion

Im Rahmen der Promotion erfolgte eine umfassende Evaluation von Verfahren des maschinellen Lernens. Sie wurden den Methoden der klassischen Statistik hinsichtlich ihrer Eignung zur Lösung betriebswirtschaftlicher Klassifikationsprobleme gegenübergestellt. Dabei standen Fragestellungen aus den Bereichen Credit Scoring und *Online* Marketing im Mittelpunkt. Um die Dimension der Evaluation weiter zu vergrößern, wurden ferner ausgewählte weitere Fragestellungen, beispielsweise die Fähigkeiten kalibrierte Wahrscheinlichkeiten zu liefern oder die Erhöhung der Uplift-Effekte durch die Anwendung der Konversions-Methoden, untersucht. Die einzelnen Teilschritte einer Implementierung des Verfahrens des maschinellen Lernens wurden individuell betrachtet und entsprechende Handlungsempfehlungen für einen effektiven Verfahrenseinsatz ausgesprochen. Diese wurden anschließend integriert, um ein ganzheitliches Vorgehensmodell zur Lösung betriebswirtschaftlicher Klassifikationsprobleme abzuleiten.

Als Ergebnis kann festgehalten werden, dass die Verfahren des maschinellen Lernens in der Tat eine gute Alternative zu den Methoden der klassischen Statistik darstellen. *Extreme learning maschines* ist eine lukrative Alternative zu anderen Verfahren des maschinellen Lernens, aber auch zur logistischen Regression für den Bereich Credit Scoring. Die Methoden des maschinellen Lernens zeigen in der Regel etwas schlechter kalibrierte Wahrscheinlichkeiten, welche sich mit Verfahren der Kalibrierung verbessern lassen. Die Vorteile der Verfahren des maschinellen Lernens können in die Methoden der klassischen Statistik integriert werden und führen zur Verbesserung der Güte der Prognose. Diese Methoden zeigen ausgezeichnete Ergebnisse im Bereich Direktmarketing, insbesondere, wenn bei Kombination mit Konversions-Methoden für Uplift-Effekte. Die Ergebnisse der vorliegenden Arbeit legen ferner nahe, dass es lohnenswert ist, die Methoden des maschinellen Lernens weiter zu entwickeln.

Es ist die erklärte Hoffnung des Verfassers, dass diese Referenzmodelle - über einen rein wissenschaftlichen Erkenntnisgewinn hinausgehend - auch einen wertvollen Beitrag

für die betriebliche Praxis leisten.

1.6 Literaturverzeichnis

Bequé, A. und Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications, 86,* 42-53.

Bequé, A., Coussement, K., Gayler, R. und Lessmann, S. (2017). Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems, 134,* 213-227.

Bequé, A. und Lessmann, S. (2018). Best of both worlds: Combining logistic regression and ensemble learners for accurate and interpretable credit risk models. *Arbeitspapier.*

Bequé, A., Gubela, R., Lessmann, S. und Gebert, F. (2018). Conversion uplift modeling in e-commerce: A benchmark study of recent modeling techniques. *Arbeitspapier.*

2 Dissertation

2.1 Veröffentlichung von Fachartikeln

Im Rahmen der vorliegenden Arbeit wurde die Veröffentlichung in zwei Fachzeitschriften angestrebt, um dem interdisziplinären Charakter der Wirtschaftsinformatik gerecht zu werden. Wissenschaftliche Zeitschriften in der Betriebswirtschaftslehre als Publikationsmedium wurden gegenüber z.B. Konferenzen präferiert. Zwei der Arbeit beigefügten Aufsätze sind wie folgt veröffentlicht:

- Bequé, A. und Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications, 86,* 42-53.
- Bequé, A., Coussement, K., Gayler, R. und Lessmann, S. (2017). Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems, 134,* 213-227.

Bequé, A. und Lessmann, S. (2018). Best of both worlds: Combining logistic regression and ensemble learners for accurate and interpretable credit risk models und *Bequé, A., Gubela, R., Lessmann, S. und Gebert, F. (2018). Conversion uplift modeling in e-commerce: A benchmark study of recent modeling techniques* werden dabei als Arbeitspapiere betrachtet, wobei das erste bereits im Dezember 2017 bei *Journal of Credit Risk* und das zweite im Juni 2018 bei *International Journal of Information Technology & Decision Making* eingereicht wurde.

2.2 Ko-Autorenschaft

Die beigefügten Fachartikel repräsentieren Ergebnisse von Forschungsprojekten und sind auf Grund dessen mit dem Namen aller beteiligten Personen unabhängig des Status (Student, wissenschaftlicher Mitarbeiter, Professor) veröffentlicht beziehungsweise eingereicht worden. Tabelle 1 setzt die Anzahl der Ko-Autoren pro Fachartikel zusammen:

Table 1: Ko-Autoren pro Fachartikel

Nr.	**Titel**	**Anzahl Autoren**
1.	Extreme learning machines for credit scoring: An empirical evaluation	2
2.	Approaches for credit scorecard calibration: An empirical analysis	4
3.	Best of both worlds: Combining logistic regression and ensemble learners for accurate and interpretable credit risk models	2
4.	Conversion uplift modeling in e-commerce: A benchmark study of recent modeling techniques	4

2.3 Substantieller Beitrag des Doktoranden

Die hier eingereichten Fachartikel stellen einen wesentlichen Bestandteil meiner wissenschaftlichen Forschung dar und wurden so ausgewählt, dass ein substantieller eigener Beitrag durchgängig gegeben ist. Dieser wird formal auch durch die Erst-Autorenschaft bei allen Fachartikeln repräsentiert und bezieht sich unter anderem auf die Initiation des Forschungsvorhabens, die Implementierung entsprechender Applikationen im Zusammenhang mit *R-Statistics* und die Durchführung empirischer Studien sowie den Anteil am Verfassen des Aufsatzes.

Keiner der hier eingereichten Beiträge ist zum aktuellen Zeitpunkt Bestandteil eines laufenden oder abgeschlossenen Promotionsvorhabens.

Teil II

Literatur

Extreme learning machines for credit scoring: An empirical evaluation.

Referenz: Bequé, A. und Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications, 86, DOI: https://doi.org/10.1016/j.eswa.2017.05.050,* 42-53.

Approaches for credit scorecard calibration: An empirical analysis.

Referenz: Bequé, A., Coussement, K., Gayler, R. und Lessmann, S. (2017). Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems, 134, DOI: https://doi.org/10.1016/j.knosys.2017.07.034,* 213-227.

Best of both worlds: Combining logistic regression and ensemble learners for accurate and interpretable credit risk models

Authors, Affiliations, and Postal address:

Artem Bequé *
Stefan Lessmann

School of Business and Economics, Humboldt-University of Berlin,
Unter-den-Linden 6, 10099 Berlin, Germany

Email:
artem.beque@outlook.com
stefan.lessmann@hu-berlin.de

Tel.: +49 (0)30 2093 5742
Fax.: +49 (0)30 2093 5741

*corresponding author

Abstract

Credit scorecards are widely used by financial institutions to enhance decision making. A credit scorecard represents a data-driven model, also called classifier, that gathers information from historical data and predicts the entry probability of events of interest. In the domain literature we find multiple studies that oppose the relative merits of individual classifiers with those of ensemble frameworks. Often coming to a conclusion that ensemble learning outperforms more conventional methods, they advocate for throughout application of ensemble frameworks in general and random forest in particular. Unlike many previous studies this study seeks to empirically examine the performance of a synergy heuristic between the logistic regression that stands for individual classifiers and random forest (stochastic gradient boosting) that represent ensemble frameworks. The synergy heuristic opens a possibility to integrate advantages of more sophisticated techniques to logistic regression. We empirically examine the performance of the original classifiers and that of the synergy heuristic to see how it influences the original logistic regression. Empirical examination goes alongside multiple dimensions. The observed results suggest that original logistic regression demonstrates competitive results. The proposed heuristic never deteriorates the performance of logistic regression and might contribute to a higher predictive fit.

Keywords: Credit scoring, logistic regression, ensemble learners, synergy heuristic, probability of default

1 Introduction

In application scoring, a scorecard represents an instrument to support decision making. In detail, the scorecard provides an estimate of the probability that a specific discrete event will take place. An example of such an event could be either default or non-default on some obligation. The prediction of such probabilities is well-established in credit scoring (Hand and Henley 1997; Khashei and Mirahmadi 2015; Thomas 2000; Gurný and Gurný 2013; Waagepetersen 2010). Based on application forms' data, demographics information, customers' transactions records or other characteristics (Crook, Edelman, and Thomas 2007) of the application that is subjected to risk assessment (Hájek 2011; Hamerle and Rösch 2006), the prediction model provides an estimate of the default probability for a certain product (for example loan). That is to say, they assign a credit score to every novel applicant. Credit score is typically given by log odds of the model-estimated probabilities of an applicant being a good or bad risk (Thomas 2010).

An increased demand for crediting has led to the urgent need for developing sophisticated techniques to support lending decision (Hand and Henley 1997). For example, in the US in May 2013 the value of consumer loans was \$1,132.4 bn.[1]; in the UK in 2012 that number was £11,676 m.[2]. On global scale, the total number of general purpose credit cards circulating in 2011 was 2,039.3 m.[3]. Given these figures, it becomes obvious that business clearly depends on quantitative methods in lending decisions. These methods enhance decision making in the industry since they evaluate the expected performance of applicants, avoid selectivity and human bias (Kiefer and Larson 2006), and quantify expected losses (Blöchlinger and Leippold 2006). Credit scoring, therefore, results in an effective risk management, prevention of the loss of future profit, and correct pricing for financial services and products (Cole, Kanz, and Klapper 2015).

By cause of a high number of retail applications (Thomas 2010), predictive accuracy is especially asked in probability of default modeling. For this reason, authors target classifiers with high discriminative power. One of the most popular ways to model the binary outcome in the credit scoring is a logistic regression (LR) (Crook, Edelman, and Thomas 2007). LR has attracted much attention in financial applications (Dong, Lai Kin, and Yen 2010; Crook, Edelman, and Thomas 2007), is a clear industry standard (Irimia-Dieguez, Blanco-Oliveer, and Vazquez-Cueto 2014; Martínes and Lechuga 2015; Yu et al. 2015) and is often practiced while evaluating alternative learning methods (Baesens et al. 2013; Lessmann et al. 2015).

However, the relative merits of LR have been questioned in the benchmarking study by (Lessmann et al. 2015). This study examines the relative merits of LR and other

[1]Data from the Federal Reserve Board, H8, Assets and Liabilities of Commercial Banks in the United States (http://www.federalreserve.gov/releases/h8/current/)

[2]Data from ONS Online, SDQ7: Assets, Liabilities and Transactions in Finance Leasing, Factoring and Credit Granting: 1st quarter 2012 (http://www.ons.gov.uk)

[3]Nielsen. (2012). Global Cards - 2011. The Nielsen Report, April 2012 (Issue 992), Carpinteria, CA, USA.

classification algorithms along multiple performance indicators in a large scale benchmark and concludes that *outperforming LR can no longer be accepted as a signal for a methodological advancement; but outperforming* random forest *can*. Thus, the authors advocate to use more sophisticated methods in general and random forest (RF) in particular. Indeed, RF along with other ensemble and multiple classifier systems, e.g., stochastic gradient boosting (SGB), have attracted much attention in the domain of credit scoring. During the last few years, ensemble learning has proved its validity for the industry and its ability to be more accurate in predictions than single classifier algorithms. Multiple examples of researches that contrast novel and established frameworks to identify the scorecards with the better predictive performance can be found in (Ala'raj and Abbod 2016b; Florez-Lopez and Ramon-Jeronimo 2015; Kruppa et al. 2013; Paleologo, Elisseeff, and Antonini 2010; Van Gestel et al. 2005).

As a result, there is ample evidence that more advanced techniques are able to predict better than the traditional ones (Lessmann et al. 2015; Rodriguez, Kuncheva, and Alonso 2006; Caruana, Munson, and Niculescu-Mizil 2006). This suggests that the development of the classification algorithms takes place on the side of sophisticated algorithms. We pursue the goal to identify synergy between more established and modern techniques in the credit scoring. Certainly, we find multiple studies that concentrate on, e.g., balancing between accuracy and complexity (Zhu et al. 2013) or offering new multiple classifier systems (Ala'raj and Abbod 2016a), but scarcely something devoted to the synergy between the techniques. That is why we argue that the relevance of synergy between the modeling techniques is still not adequately addressed in the credit scoring literature. We find many scholars who refute the value of the advanced learning methods, criticizing, for example, a lack of comprehensibility (Hand 2006), whilst others promote them by, e.g., developing neural networks (Angelini, Di Tollo, and Roli 2008). Standing in-between these two positions, we seek to discover possibilities to integrate the advantages of more advanced approaches to LR to achieve favorable balance between predictive accuracy, comprehensibility, justifiability, and other quality criteria in credit scoring.

The goal of this paper is, therefore, to discover possibilities to integrate the advantages of more sophisticated modeling techniques to LR and to see how this integration influences the performance of the latter in multiple dimensions, whereby balancing between the predictive performance and comprehensibility. In pursuing this objective, we make the following contributions. First, we confirm the predictive performance difference between LR and RF (SGB). Evaluation of predictive performance goes alongside multiple dimensions: (i) correctness of categorical predictions; (ii) quality of probabilistic predictions; (iii) discriminatory ability; and (iv) performance measured in expected maximum profit. Meanwhile, we try to quantify comprehensibility and justifiability to account for an equally, yet often overlooked, important dimensions of building and developing of modeling techniques. Furthermore, we propose a synergy heuristic that opens a possibility to integrate the advantages of RF (SGB) to LR. Through empirical examination, we capture the influence of this integration on the performance of LR in every experimental setup. We rely upon multiple performance measures that are further backed by robustness procedure. We evaluate performance of all techniques

and provide specific recommendations regarding which techniques work better.

The remainder of the paper is organized as follows. We start by outlining the synergy heuristic we propose in this study. Next, we elaborate the experimental design, including the underlying data and the performance indicators. This is followed by the experimental results. We conclude by discussing the limitations and potential extensions of our study.

2 Synergy heuristic between the modeling techniques

In this study, we discover possibilities to integrate the advantages of RF (SGB) to LR and examine how this integration influences the performance of the latter. In the following, we outline the framework of the synergy heuristic between the modeling techniques. Figure 1 presents the overall flow of the framework. It implies that we undertake three steps. First, we screen the attributes by application of the variable importance measures of RF (SGB) and define most important features. We then manipulate the functional form of LR by integration of interaction and non-linear terms of the most important features in multiple setups (see Figure 1). Finally, we build and apply models of the manipulated LR. While the results of model building and prediction are presented later, here, we focus on attributes screening and functional form manipulation.

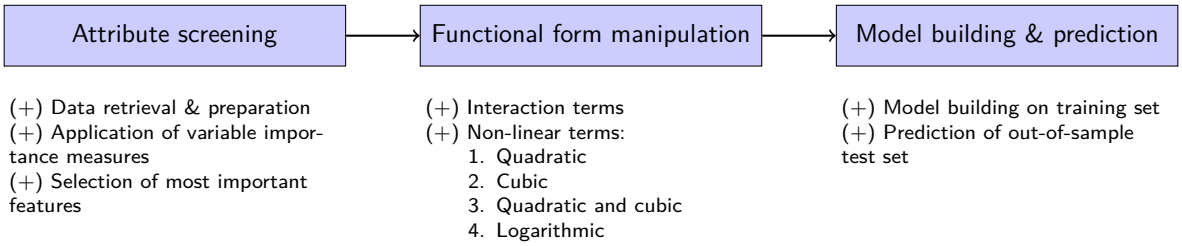

Figure 1: Overall flow of the synergy heuristic

2.1 Screening of the attributes

There are many studies (e.g., Breiman 2004; Biau 2012; Geurts, Ernst, and Wehenkel 2006) that have investigated different variants of tree-based ensembles methods and proved their performance consistency in applied research. By cause of the capability to build accurate predictive models and to deliver variable importance measures, tree-based ensembles, especially RF (Breiman 2001), have become a popular data analysis technique used with success in various areas. Despite the growing interest to the variable importance measures, we find studies (e.g., Ishwaran 2007) that specifically denote the examination of theoretical properties and mathematical mechanisms behind them. Thus, the tree-based ensembles

possess advantages the properties of which can be applied to achieve higher prediction accuracy.

One of the main advantages of the tree-based ensembles (Breiman 2001; Ishwaran 2007) is the ability to handle interaction and non-linear terms that makes them more competitive to LR. Put differently, RF (or SGB) manage interaction between the variables and non-linear terms automatically, which is further strengthened through introduction of random perturbations into the learning procedure by RF (and SGB). As a result, RF (SGB) defines the most important variables for solving a given problem. LR, on the contrary, does not handle interaction or non-linear terms by itself. Thus, we make use of the given advantages of RF (SGB) and integrate them to the functional form of LR. To do so, we define the importance of every explanatory variable as per importance measure of RF (SGB) that represent classification trees in the context of ensemble learning.

A binary classification tree (Breiman et al. 1984) represents a tree structure T of the input-output model, from a random input vector (x_1, ..., x_i) with values in x_1, ..., $x_i = X$ to a random output variable Y. Any node t in the tree represents a subset of the space X, with the root node being X itself. Internal nodes t are labeled with a binary test $s_t = (x_m < c)$ dividing subset in two children t_L and t_R subsets, while the terminal t are labeled with the majority class $j_{(t)}$ guess value of the output variable. The predicted output $\hat{Y}$ for a new instance is the label of the node reached by the instance when it is propagated through the tree. The tree learns from a sample size N drawn from $P(x_1, ..., x_p, Y)$ using a recursive procedure, which identifies at each t the split $s_t = s^*$ for which the partition of the N_t node samples into t_L and t_R maximizes the decrease of some measure $i(t)$ (e.g., mean decrease in accuracy). This measure is, thus, used to judge about the importance of every variable in X. Construction of the tree stops when, e.g., nodes become pure in terms of Y or when all variables X_i are locally constant.

To increase prediction accuracy and to avoid high variance, in the context of ensemble trees, practitioners introduce random perturbations into the learning procedure. Thus, modelers obtain multiple decision trees from a single learning set and aggregated predictions across all these trees (Breiman 2001). Therefore, some measures are used to evaluate the importance of variables aggregated across these perturbations. In this study, we rely upon the mean decrease in accuracy (MDA), retrievable from RF, where the values of X_m are randomly permuted in the out-of-bag samples (Hastie, Tibshirani, and Friedman 2011); and we also exercise the reduction of squared errors (RSE) retrievable out of SGB. These two measures represent the error rates for classification problems (like one we describe in this study) that are subjected to minimize while considering the importance of the variables. That is why we consider both of them in our experimental setup.

2.2 Functional form manipulation of LR

RF (or SGB) are able to identify and manage interaction between the variables and non-linear terms, which is not given by LR. That is why we consider interaction and non-linear terms for the functional form manipulation of LR. More specifically, we define most important features as per MDA (RSE) and then integrate them to the original LR. To elaborate the manipulation techniques, consider Y_i as the dependent variable (default probability), which we seek to explain by means of three explanatory variables related to the i^{th} applicant, X_i income, Z_i number of children, and W_i income of spouse. Thus, the regression equation has the following formal presentation:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 W_i + e_i \tag{1}$$

where β_0 represents the intercept, $\overrightarrow{\beta_i}$ is the vector of coefficients, and e_i is the error term.

First, we consider interaction terms for the functional form manipulation. We assume that there is interaction between X_i and Z_i. The original (1) will take, thus, the form as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 W_i + \beta_4 X_i Z_i + e_i \tag{2}$$

An interaction occurs when the magnitude of the effect of one feature on the dependent variable varies as a function of a second feature (Bauer and Curran 2005). This phenomenon is also known as the moderation effect and can be met in situations that involve univariate and multivariate analysis of variance and covariance or, e.g., in path analysis (Aiken and West 1991). The interaction between two terms is also known as two-way interaction and β_4 can be interpreted as the amount of change in the slope of Y_i on X_i when Z_i changes by one unit (Aiken and West 1991). One could also go with three-way or so-called higher-order interaction terms. This means, we will add the product $\beta_5 X_i Z_i W_i$, i.e., among all explanatory variables in our example. Thus, interaction terms contribute to a higher modeling fit when the effect of one explanatory variable on the dependent variable is different at different values of other explanatory variables.

There are many examples of integration of non-linear terms to the functional form in different fields (McGwin, Jackson, and Owsley 1999; Li et al. 2015). In all these studies, researchers add to the functional form quadratic or even cubic terms of the explanatory variables. Assume that income has been identified as important as per MDA (or RSE). We will thus consider it for non-linear integration. The original (1) will now take the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 W_i + \beta_4 (X_i)^2 + e_i \tag{3}$$

We might also include $\beta_4 X_i^3$ or $\beta_4 X_i^4$ to the functional form (1). The rationale behind the integration of non-linear terms is that a person i who has high income has more certainty to pay off the credit line thus lessen the probability to default (consider an example of applicants with big differences in income). Taking this into consideration, we enhance the knowledge of the non-linear relationships between the explanatory and dependent variables, and, thus, improve the predictive performance.

Sometimes the logarithm to some other transformation is preferred (Tukey 1997). There are several reasons for this. First, the residuals have a skewed distribution. Logarithmic transformation obtains residuals that are approximately symmetrically distributed. Second, the spread of the residuals changes systematically with the values of the dependent variable. The logarithmic transformation in this case will remove the systematic change in spread. Another example is when the scientific theory requires such kind of transformation (Tukey 1997). The original (1) will then take the form as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 W_i + \beta_4 log(X_i) + e_i \tag{4}$$

In this case we linearize the relationship between the variables by including $\beta_4 log(X_i)$ to the functional form (1). Again, we first identify the important explanatory variable as per MDA (RSE) and then consider these for non-linear manipulation.

3 Experimental setup

We seek to confirm the difference in predictive performance between RF (SGB) and LR as well as examine how the synergy heuristic influences the performance. Our experimental design involves a real credit scoring data set. This data set belongs to the field of application scoring, indicating the goal to categorize credit applicants into good and bad risks. More specifically, the data set comes from the 2010 PAKKD data mining challenge.[4] This data set has been used in prior work and can be considered as established in the literature, e.g., (Bahnsen, Aouada, and Ottersten 2014; Xie et al. 2009; Lessmann et al. 2015).

The data set entails a binary response variable that indicates the observed event, i.e., good or bad risk, of a granted credit and a number of attributes concerning the loan. The attributes can be categorized into several groups. For example, debtor attributes (e.g., marital status or education level), loan attributes (e.g., loan amount or product type), ability of debtor to pay back (e.g., personal income or other income), and other (e.g., a flag of having a visa card or quantity of bank accounts). In total the data set includes 50,000 credit applicants, 37 attributes and the prior default rate of .261.

To prepare the data for subsequent analysis, we employ standard operations for the attributes screening. In particular, we exclude the applicants with missing values, standardize

[4] http://sede.neurotech.com.br/PAKDD2010/

numeric variables and use the dummy coding technique to convey all the necessary information of the categorical attributes (Crone, Lessmann, and Stahlbock 2006; Kuhn and Johnson 2013).

Another important concern relates to data partitioning. Based on industry recommendations (Dietterich 1998) we apply k-fold cross-validation. We randomly split the data set to equal size training and out-of-sample testing set. We then randomly partition the training set into k equal size subsamples. Of all k subsamples, a single subsample is reserved as the validation data for testing the classifiers, and the remaining $k-1$ subsamples are used as a training data. Thus, the cross-validation process is repeated k times (i.e., number of the folds), where every k subsample is only ever used once as a validation data. The rationale behind this approach is that all observations in the given data set are used both for classifier training and validation, and every observation is used for validation exactly once. In our experiment we set k to 10 and report later on the results of every k to cross check the performance robustness of the classifiers.

The experimental design includes LR, RF and SGB. The experiment is performed in the R-Statistics environment. To secure more robust results, we consider a wide range of the meta-parameters for both RF and SGB, presented in Table 1. The choice is motivated through (Lessmann et al. 2015). Every model is automatically tuned and evaluated using 10-fold cross validation applied to the training set. The random seed is set before every algorithm is trained to ensure that every algorithm gets the same data partitions and repeats.

Table 1: Meta-parameters of the classifiers

Acronym	**No. of models**	**Meta-parameter**	**Candidate settings**
RF	30	No. of CART trees	[100, 250, 500, 750, 1000]
		Randomly sampled variables	m * [3, 5, 7, 9, 12, 15]
SGB	72	No. of trees to grow	[50, 200, 500]
		Depth of variable interactions	[1, 2, 3, 4]
		Shrinkage parameter	[0.2, 0.4, 0.6]
		Observations in terminal nodes	[8, 10]
LR	1	-	-

We are interested in how the synergy heuristic influences the performance of the original LR across multiple dimensions. That is why to assess the ability of classifiers to generate accurate predictions, we employ four different performance metrics. All of them embody a different notion of predictive accuracy and therefore measure different dimensions of the predictive performance. To judge the correctness of the scorecard's categorical prediction, we consider the *percentage correctly classified* (PCC). To measure the quality of probability estimates of the classifiers, we use *Brier Score* (BS). We involve *the area under a receiver-operating curve* (AUC) to judge the ability of classifier to rank high and low risk applicant in the right order. Finally, monetary value is an equally important dimension of classifier

performance, which we measure in terms of *expected maximum profit* (EMP) (Verbraken et al. 2014).

In credit scoring threshold metrics (PCC) (Atish and Jerrold 2004; Ong, Huanga, and Tzeng 2005; Twala 2010; Lessmann et al. 2015) get particularly high attention. All of them are derived from the confusion matrix, which presents the actual *versus* predicted class labels; whereby PCC is defined as the fraction of correctly classified labels over total number of labels. Threshold metrics ignore the absolute values of the estimates of posterior probability. That is why we also consider BS, which represents the mean squared-error between probabilistic predictions and a zero-one-coded target variable (Thomas, Edelman, and Crook 2002; Bequé et al. 2017; Ala'raj and Abbod 2016b, 2016a). AUC is well-established in credit scoring (Lessmann et al. 2015; Chawla 2005; Wang, Kun, and Shouyang 2012) and represents an aggregated measure of the classification performance averaged over all possible thresholds on the ROC-curve (Flach, Hernández-Orallo, and Ramirez 2011; Fawcett 2006). A recently proposed profit-based classification performance measure (Verbraken et al. 2014) - EMP - was applied in credit scoring to find a trade-off between the expected losses and the operational income by the loan. Developed and optimized from the average classification profit per borrower to maximum profit measure. See Appendix A for more details on the performance metrics.

4 Empirical results

The experimental results consist of the performance estimates of LR versus RF (SGB) along multiple dimensions. We first consider the comprehensibility and justifiability of the modeling techniques. We then analyze the predictive performance which includes correctness of categorical predictions, quality of probabilistic predictions, and discriminatory ability. The performance measures capture the degree to which the synergy heuristic influences the performance of LR when compared to RF (SGB). Finally, we measure the performance of the classifiers based on expected maximum profit.

4.1 Exemplification of the comprehensibility and justifiability

Comprehensibility and justifiability of the model are a key requirement, especially for the industry of credit scoring, since the models need to be validated and in line with the domain knowledge before they can be implemented. For example, the Equal Credit Opportunity Act of the US requires financial institutions to provide specific reasons why a customer's credit application was rejected, whereby unclear reasons for denial are considered as illegal.[5]

We stick to the definitions of comprehensibility and justifiability that have been already mentioned previously (Martens and Baesens 2010). More specifically, there are two main

[5]Federal Trade Commission for the Consumer. Facts for consumers: Equal credit opportunity. Technical Report, FTC, March 1998

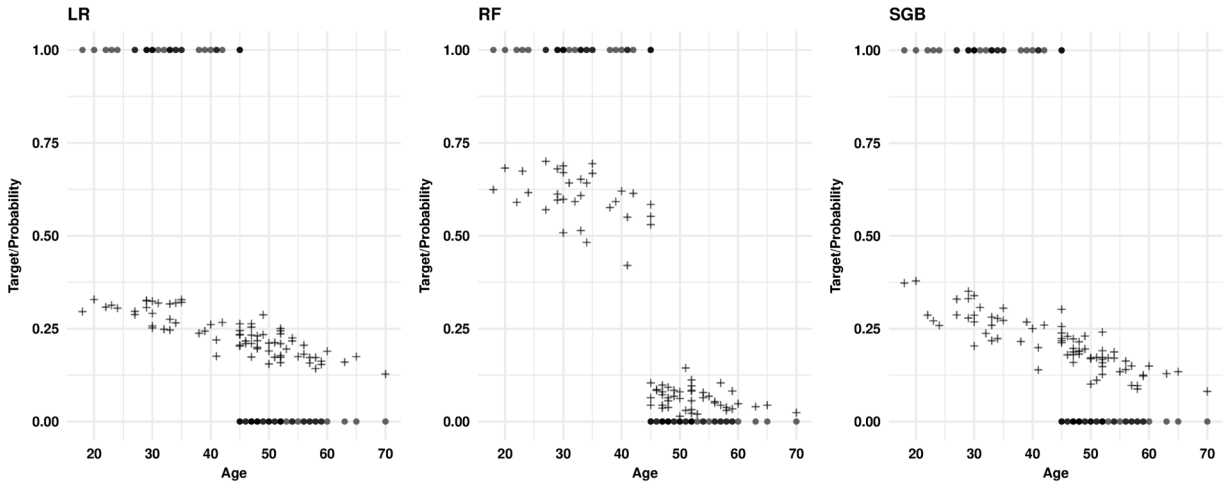

Figure 2: Justifiability test over age

drivers for comprehensibility: (i) the type and size of input parameters, meaning that the more meta-parameters a model relishes, the more a practitioner should invest to comprehend and tune model; and (ii) the algorithm behind the model. Given Table 1, we can easily conclude that LR is the winner in this dimension, since it requires no meta-parameters to tune. On the contrary, RF requires two meta-parameters that results in higher complexity behind the algorithm, which is even more complicated for SGB. As a result, the comprehensibility decreases with the size of the algorithm behind the predictive models (Askira-Gelman 1998). Occam's razor principle is also motivated in the domain (Domingos 1999), which says that simpler models should be preferred over more complex. All these signals in favor of LR in terms of comprehensibility.

Again, to define the justifiability, we rely upon (Martens and Baesens 2010). For example, it cannot be that applicants with higher age (that theoretically implies more working experience, better job opportunities, and possible savings) are rejected, while applicants with similar characteristics but younger in the their age are granted credit, all else being equal. To define it more formally, assume a data set $D = \{x^i, y^i\}_{i=1}^n$, with $x^i = (x_1^i, x_2^i, ..., x_m^i) \in X = X_1, X_2, ..., X_m$, and a partial ordering $\leq$ defined over input space X, and Y of class values y^i. Then, a classifier f is monotone, $f : x^i \mapsto f(x^i) \in Y$, if:

$$x^i \leq x^j = f(x^i) \leq f(x^j), \forall i, j \tag{5}$$

holds. To exemplify this, we retrieve a subsample of credit applicants with similar characteristics, but of different age for both classes (i.e., defaulter and non-defaulter). We then train LR, RF, and SGB on the training set and predict the default scores on the out-of-sample test set to examine our assumption. Figure 2 summarizes these scores. In detail, it represents the defaulters and non-defaulters, whereby 1 signifies a defaulter; it presents the probabilities across the classifiers on y-axis, and finally the age of the applicant on the x-axis.

Figure 2 reveals that our assumption - the older an applicant, the bigger chance of credit repayment - is realistic. We can make several additional conclusions. First, we observe that all classifiers assign right partial ordering of the probabilities, meaning the higher the age, the lower the default probability. However, we can observe that the spread of probabilities is not equal. We see a visibly higher disperse of probabilities obtained from RF than that of LR or SGB. LR, on the contrary, demonstrates the steady decrease in probabilities with increase in age. We can only conclude that in our experiment this finding gestures in favor of LR when compared to RF and LR shows competitive results when compared to SGB in terms of justifiability. Doing our best to approximate empirical judgment of justifiability, we caution that this conclusion, however, should be further authenticated in well-rounded experimental setups.

4.2 Exemplification of the explanatory variables

We now exemplify the importance of the attributes according to variable importance measures. Recall that we assess importance of explanatory variables based on MDA and RSE retrievable from RF and SGB, respectively. In particular, we examine the suitability of the most important attributes for the functional form manipulation of LR. Table 2 presents the relative variable importance. To obtain Table 2, we train both RF and SGB on the training set and retrieve the variable importance as per MDA (RSE). Recall that the higher the value of MDA (RSA) of a variable the more that variable contributes to the reduction of the corresponding error. In other words, the higher the value of the importance measure the more important it is. We also compare the importance of variables as per MDA (RSA) with their relative importance according to LR. Recall that the *p-values* higher than .05 indicate that the attribute is not significant in prediction of the response variable. To capture the relative importance according to LR, we estimate the *p-values* on the same training set as for the variable importance measures and present them in Table 2 as well. We mark the important attributes across the techniques in bold face.

Table 2: Relative importance of the attributes

Attribute	MDA	RSE	*p-values*
Age	**52.92**	**82.55**	**.000**
Residential zip 3	**39.24**	0.05	.087
State of birth	**37.78**	0	.183
Marital status	12.50	**11.68**	**.000**
Flag residencial phone	16.20	**5.44**	**.000**
Sex	8.62	0	**.000**
Flag mastercard	6.51	0	**.001**
Occupation type	28.41	0	**.003**
Application submission type	22.25	0	**.020**
Company	13.38	0	**.042**
Payment day	12.09	0.26	.526
Quant dependents	8.55	0	.587
Residential state	37.63	0	.143
Residence type	13.04	0	.663
Months in residence	11.02	0	.358
Flag email	14.43	0	.691
Personal monthly income	15.23	0	.467
Flag visa	2.22	0	.554
Quant banking accounts	21.68	0	.388
Quant cars	22.42	0	.880
Flag professional phone	19.53	0	.112
Profession code	24.38	0	.853
Product	7.41	0	.614

Table 2 reveals several important findings. First, Table 2 indicates that RF and SGB identify the importance of the variables differently, apparently, by cause of the different importance measures and techniques the algorithms exploit. Another important observation is that the *top three* attributes are by far more important than all other. This is especially relevant for SGB and less applicable to RF. However, we decide to involve the *top three* attributes for the manipulation logic we pursue; for both RF and SGB. The top three according to MDA are *Age*, *State of birth* and *Residential ZIP* of the applicant, according to RSE are *Age*, *Flag residential phone* (having or not a residential phone) and *Marital status*. As a result, both variable importance measures have defined only one continuous variable (i.e., *Age*) that can be involved to the non-linear manipulation. Thus, there will be no difference in the performance estimates between MDA and RSE in this regard. However, the interaction between the variables will be different and will result in different predictive performance. Table 2 also shows that important variables according to MDA (RSE) are not necessarily important as per LR and vice versa. For example, *State of birth* is important as per MDA but seems to be insignificant as per LR. On the contrary, the variable *Sex* is significant as per LR and not important as per MDA. It is interesting that all three techniques have identified only one attribute (i.e., *Age*) as important in common. It is also remarkable that top three attributes as per RSE are also significant as per LR which is not the case for MDA.

This indicates that LR might not have captured the non-linear effects. That is why we will consider the important variables as per MDA (RSE) for functional form manipulation going further.

4.3 Examination of predictive performance

Now, we assess the predictive performance of the classifiers. Recall that assessment goes along multiple dimensions: (i) correctness of categorical prediction measured in PCC; (ii) quality of probabilistic estimates measured in BS; (iii) discriminative ability measured in AUC; and (iv) performance measured in expected profit - EMP. Table 3 presents the results across the original classifiers (i.e., LR without functional form manipulation) to capture difference in predictive performance. To obtain Table 3, we execute LR, RF, and SGB on the out-of-sample test set and gain the probabilities of the default. To get PCC values, we need to compare the probability of default of every applicant i to a threshold. In practical applications (Bravo, Maldonado, and Weber 2013), a proper threshold is obtained from such attributes as the costs associated with granting credit to default customers. Lacking such information, we find a more generic approach that has been already applied in the credit scoring (Baesens, Roesch, and Scheule 2016; Lessmann et al. 2015). In particular, we estimate the prior default rate in the training set (e.g., 35%), rank credit applicants in the test set according to their model-estimated default probabilities, and classify the top 35% of observations with the largest default probability as bad risk, and other applications as good risk. Our data set does not entail the necessary information to calculate the input parameters (see Appendix A (4)) to obtain EMP, thus, we rely upon the default parameters provided by (Verbraken et al. 2014). Recall that we present the expected maximum profit of the ROC curve at the optimal cutoff fraction, whereby higher values indicate better performance.

Table 3: Predictive performance of original classifiers

	PCC	**BS**	**AUC**	**EMP**
LR	.66	.18	.60	.0146
RF	.67	.18	.62	.0149
SGB	.68	.18	.64	.0155

Table 3 reveals several important findings. First, we observe that LR demonstrates very competitive results across all performance metrics. This is especially relevant for BS. We see that the BS estimates are equal for all classifiers, indicating identical performance in terms of quality of probabilistic predictions. However, we also observe that there is a performance gap across other metrics. We see that SGB is the winner in handling the categorical prediction. That is to say, it gets the lowest number of FN and FP (see Appendix A (1)) when compared with the competing classifiers. Namely, the PCC values of SGB, RF, and LR are .68, .67, and .66. Furthermore, we observe performance difference when measured in AUC. More

specifically, we can observe the performance gap in the accuracy of predictions between LR and RF, indeed (.60 and .62, respectively). This gap gets even bigger, when we compare LR and SGB, .60 *versus* .64. Table 3 reveals that the performance gap between LR and RF as per EMP is borderline. This signals another time that LR can be regarded as very competitive to RF in this dimension as well. However, we have to conclude that there is a performance difference between LR and SGB. In detail, EMP of LR against SGB is .0146 to .0155, respectively. All this signals that LR might be inferior to more advanced techniques in the industry of credit scoring.

Having identified the performance difference across the metrics of the original classifiers, we now present the influence of the manipulation techniques in Table 4. Recall that this results in two groups of the predictions of LR on the out-of-sample test set. The first group of the predictions relates to RF (MDA) and the other to SGB (RSA). In every group we exercise manipulation to the top three attributes. Thus, we present three two-way and one high-order interactions and quadratic, cubic, quadratic and cubic, and logarithmic non-linear integrations. Recall that both MDA and RSA have identified only one continuous variable (i.e., *Age*) that is subjected to non-linear manipulation. In this respect there will be no difference in performance between MDA and RSA. Ultimately, Table 4 presents 12 manipulated functional forms of LR. It also echoes the performance estimates of the original LR across all metrics. As interaction is applied to all top-three attributes, type 1 stands for the interaction between the top-1 and top-2 attributes. In the same manner, type 2 means the interaction between top-1 and top-3 attributes, type 3 between the top-2 and top-3, and type 4 represents the higher-order interaction.

Table 4: Functional form manipulation (PCC)

Metric	**Measure**	**Original**	**I**	**II**	**III**	**IV**	**2**	**3**	**2/3**	**Log**
PCC	MDA	.66	.66	.66	.66	.66	.66	.66	.66	.66
	RSE		.66	.67	.66	.67				
BS	MDA	.18	.18	.18	.18	.18	.18	.18	.18	.18
	RSE		.18	.18	.18	.18				
AUC	MDA	.60	.61	.61	.61	.61	.61	.61	.61	.61
	RSE		.61	.62	.61	.62				
EMP	MDA	.0146	.0146	.0146	.0146	.0146	.0146		.0146	
	RSE		.0146	.0153	.0147	.0152				

Table 4 indicates several important findings. First, we observe that none of the functional form manipulation techniques diminishes the advantages of the original LR. This holds for all performance metrics and for every integration composition. Second, the performance estimates are equally stable across all types of integration. For example, non-linear integration

has firmly equal predictive performance (see, e.g., PCC). This indicates that the manipulation techniques have fairly equal influence on the performance of the original LR. Another important finding is that interaction is superior to non-linear integration. This can be clearly seen on the results provided by RSE. Type 2 interaction demonstrates the highest PCC value. The three-way interaction achieves the same result. Thus, interaction might have positive influence on the performance of the original LR and close the performance gap between LR and RF ad modum of correctness of categorical prediction. It is worth mentioning that RSE is superior to MDA. Table 4 also indicates that none of the manipulation techniques improves the BS values of the original LR. Thus, all manipulation techniques have literally no influence on BS. In general, we can further stress the importance of the quality of probabilistic predictions. Regulatory frameworks such as the Basel Accord require financial institutions to guarantee that internal rating systems produce well-calibrated risk predictions. Poorly calibrated risk predictions are penalized with higher regulatory capital requirements (Crouhy, Galai, and Mark 2000). Well-calibrated risk predictions are also relevant from a lending decision point of view (Cole, Kanz, and Klapper 2012). We find few studies that explicitly measure the performance of the classifiers in BS (Abdou et al. 2016; Ala'raj and Abbod 2016b; Tasche 2013). Table 4 illustrates that all manipulation techniques have contributed to the higher predictive fit of LR in terms of AUC. This is relevant for both variable importance measures. Thus, the manipulation techniques we perform might exhibit positive influence on the performance of LR. For example, the interactions as per MDA as well as non-linear manipulation techniques increased the AUC by 1 percent. Again, we see that interaction terms are superior to non-linear integration. More specifically, the performance gap between LR and RF could close the interaction type 2 and higher order interaction as per RSE. In particular, the corresponding AUC values get equal to those of RF (.62). This indicates that RSE is superior to MDA and we therefore can recommend it for similar manipulations we exhibit. A comprehensive review of 214 articles/books/theses on application credit scoring (Abdou and Pointon 2011) accepts the view that more advanced techniques (e.g., genetic algorithms) outperform conventional models (i.e., LR), but reports at the same time on studies that find similar performance in terms of predictive accuracy. We can further support the view of (Abdou and Pointon 2011) and conclude that LR can be regarded as competitive to RF (SGB), especially given that functional form manipulation closes the performance gap between LR and RF. However, we also admit that the gap between LR and SGB remains open. Table 4 reveals that functional form manipulation we pursue can be seen as successful when measured in EMP. More specifically, we observe that RSE outperforms MDA once again. Interaction terms of the second type proved its validity based on EMP, being superior to RF, and making the performance difference to SGB smaller. Slightly less successful is the interaction of the higher order. Interaction of the type three improved the performance of the original LR marginally. All these alarms in favor of interaction over other type of functional form manipulation. We can only assume that even marginal changes in EMP might result in substantial financial profits, however, this assumption requires proof in multiple experimental designs that entail corresponding information to capture the increase in revenue.

4.4 Examination of the predictive performance robustness

Finally, we verify the performance robustness of the classifiers as per performance metric on every validation fold. To do so, we perform 10-fold cross validation to catch the performance of the classifiers on every fold in separate. We present the performance robustness whereby we capture the performance of every classifier and functional form manipulation based on every fold in Table 9. Table 9 reveals some new findings. First, we observe that the BS values remain stable across all folds that strengthens the previous finding that LR is at least equal performer in this dimension. None of the functional form techniques improved the performance in terms of BS. However, we can also observe that the original LR outperforms the peer classifiers slightly on sample 10 when measured in BS. This signals another time in favor of LR that it is able to produce probabilistic prediction of high quality. Second, we can even further see that the original LR demonstrates very competitive results. This can be seen on sample 3, where LR achieves the same PCC values as both RF and SGB, and at the same time the same AUC value as RF does. We see that RSE is more successful than MDA. More specifically, the manipulation in terms of RSE shows improvement of PCC and AUC values on, e.g., sample 1, 2, or 4. We emphasize another time that the success of the interaction terms is even further validated. This technique closes the gap in predictive performance between LR and RF multiple times. See, for example, sample 2 where LR performs equally successfully as RF in terms of AUC. Other examples are samples 4, 6, or 9. Although non-linear interaction terms are less successful in general, we see that logarithmic term integration might also contribute to a higher predictive fit of LR as well. See, for example, sample 2 and 6. However, integration of logarithmic terms does not close the gap between LR and RF and it might deteriorate the performance. See sample 5 where the performance of the original LR diminishes after integration of logarithmic terms to the functional form. The decrease in performance is light, however, it takes place. We can also observe that all manipulation techniques might have literally no impact on the performance of the original LR. See the performance of LR in terms of AUC on sample 7. In general, we observe the tendency in positive influence of the manipulation techniques on the performance, especially when we are talking about RSE. Thus, we recommend to consider this measure for synergy heuristic we follow. We conclude this also when we pay attention at the performance as per EMP. More concrete, RSE (i.e., interaction of type II and higher order integration) improved the performance of the original LR on the samples 1, 3, 4, etc. We can see, for example, that the performance gap as per EMP between LR and RF is closed on sample 3 based on RSE. Thus, we conclude that based on EMP estimates the original LR is competitive to other modeling techniques and the synergy heuristic demonstrates clearly a positive potential in improvement of EMP.

Table 5: Examination of the performance robustness

					Manipulated LR											
Sample	Metric	Original			MDA				RSE				Non-linear			
		RF	SGB	LR	I	II	III	IV	I	II	III	IV	2	3	2/3	Log
1	PCC	.67	.66	.64	.64	.64	.64	.64	.64	.64	.64	.64	.64	.64	.64	.64
	BS	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.62	.62	.59	.59	.59	.59	.58	.59	**.60**	.59	**.60**	.59	.59	.59	.59
	EMP	.0168	.0170	.0147	.0147	.0146	.0147	.0147	.0147	**.0151**	.0146	**.0152**	.0147	.0147	.0147	.0149
2	PCC	.67	.67	.65	.65	.65	.65	.65	.65	**.66**	.65	**.66**	.65	.65	.65	.65
	BS	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.63	.64	.61	.61	.61	.61	.61	.61	**.63**	.61	**.63**	.61	.61	.61	**.62**
	EMP	.0159	.0171	.0166	.0166	.0165	.0166	.0166	.0165	.0165	**.0167**	.0165	.0166	.0166	.0166	.0165
3	PCC	.66	.66	.66	.65	.65	.65	.65	.65	.65	.65	.65	.66	.66	.66	.66
	BS	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.61	.62	.61	.61	.61	.61	.61	.61	.61	.61	.61	.61	.61	.61	.61
	EMP	.0168	.0173	.0160	.0161	.0160	.0160	.0161	.0160	**.0168**	.0162	**.0168**	.0160	.0160	.0160	.0165
4	PCC	.67	.68	.66	.66	.66	.66	.66	.66	**.67**	.66	**.67**	.66	.66	.66	.66
	BS	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.63	.64	.62	.62	.62	.62	.62	.62	**.63**	.62	**.63**	.62	.62	.62	.62
	EMP	.0154	.0166	.0152	.0152	.0152	.0150	.0152	.0152	**.0155**	.0152	**.0154**	.0152	.0152	.0152	.0155
5	PCC	.67	.67	.66	.65	.66	.66	.66	.66	.66	.65	.66	.66	.66	.66	*.65*
	BS	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.63	.63	.61	.61	.61	.61	.61	.61	**.62**	.61	**.62**	.61	.61	.61	.61
	EMP	.0150	.0160	.0142	.0141	.0141	.0141	.0140	.0142	**.0151**	.0141	**.0151**	.0142	.0142	.0142	**.0148**
6	PCC	.67	.67	.66	.66	.66	.66	.66	**.67**	.66	.66	.66	.66	.66	.66	.66
	BS	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.62	.63	.60	.60	.60	.60	.60	.60	**.62**	**.60**	.62	.60	.60	.60	**.61**
	EMP	.0153	.0155	.0140	.0140	.0140	.0140	.0140	.0140	**.0142**	.0139	**.0143**	.0140	.0140	.0140	.0138
7	PCC	.67	.68	.67	.67	.67	.67	.67	.67	.67	.67	.67	.67	.67	.67	.66
	BS	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.63	.63	.61	.61	.61	.61	.61	.61	.61	.61	.61	.61	.61	.61	.61
	EMP	.0176	.0151	.0170	.0152	.0150	.0150	.0152	.0152	**.0163**	.0154	**.0164**	.0151	.0151	.0151	.0151
8	PCC	.67	.67	.65	.65	.65	.65	.65	.65	.65	.65	.65	.65	.65	.65	.65
	BS	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.63	.63	.60	.60	.60	.60	.60	.60	**.61**	.60	**.61**	.60	.60	.60	.60
	EMP	.0136	.0143	.0130	.0131	.0132	.0130	.0130	.0131	.0131	.0130	.0132	.0130	.0130	.0130	.0136
9	PCC	.68	.68	.67	.67	.67	.67	.67	.67	.67	.67	.67	.67	.67	.67	.67
	BS	.18	.18	.18	.18	.18	.18	.18	.18	.18	.18	.18	.18	.18	.18	.18
	AUC	.62	.63	.61	.61	.61	.61	.61	.61	**.62**	.61	**.62**	.61	.61	.61	.61
	EMP	.0140	.0136	.0130	.0118	.0115	.0116	.0116	.0116	**.0121**	.0117	**.0121**	.0115	.0115	.0115	.0115
10	PCC	*.68*	.67	.66	.66	.66	.66	.66	.66	**.67**	.66	**.67**	.66	.66	.66	.66
	BS	.18	.18	**.19**	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19	.19
	AUC	.62	.62	.59	.59	.59	.59	.59	.59	**.60**	.59	**.60**	.59	.59	.59	.59
	EMP	.0172	.0180	.0166	.0166	.0167	.0167	.0167	.0167	**.0172**	.0167	**.0171**	.0166	.0166	.0166	.0168

5 Conclusion

We set out to empirically examine the performance of LR with more advanced techniques in credit scoring. Examination goes along multiple dimensions: comprehensibility and justifiability, predictive performance, and performance measured in maximum expected profit. We confirm performance different between the modeling techniques and seek to find a synergy heuristic to see how it influences the performance of LR. The heuristic involves two variable importance measures retrievable from predictive tree-based ensemble frameworks, whereby we identify the top three important variables based on every measure and then integrate them to the functional form of LR. The functional form manipulation is two-fold, we first study the interaction terms of lower and higher order and integration of non-linear terms.

In every dimension the original LR (i.e., the model with the full set of variables without functional form manipulation) establishes competitive results to the alternative predictive learners considered in the study. This can be seen not only in the main body of the experimental comparisons but also supported based on the considerations of comprehensibly and justifiability, and even further encouraged by the results obtained on cross validation folds in this study. This holds for all performance dimensions we consider and is especially valid for the quality of probabilistic prediction where we do not observe any performance gap between the modeling techniques. All this intensifies the view that more conventional modeling learners, like LR, (Abdou and Pointon 2011) are able to produce competitive results to more sophisticated techniques.

Most importantly, however, is that LR represents a feasible and sound alternative in terms of comprehensibility and justifiability predictive model. It is also important to underline that it is possible to close the performance gap with RF after the manipulation of the functional form. Especially, the interaction terms manifest most successful performance. More specifically, interaction terms were able to exhibit positive influence on the performance of LR in both PCC and AUC. When MDA and RSE are compared RSE has solidly outperformed the first and, therefore, we can only recommend to consider it for manipulation strategies similar to those we execute in this study. Given that embodiment of logarithmic terms illustrates positive signals of AUC improvements, we can only further recommend experimenting with it.

The composition of the above synergy heuristic makes LR even more appealing for credit scoring than the original LR and injects doubt on previous work of (Lessmann et al. 2015), where authors strongly advocate to exercise more sophisticated techniques. Given that many scholars refute the more advanced techniques (Hand 2006) due to, e.g., a lack of comprehensibility, another dimension that is equally important while building predictive learners and that has been addressed in this study, we encourage further experimenting with LR, since it has long verified its concept in numerous studies and practical applications (Crook, Edelman, and Thomas 2007; Dong, Lai Kin, and Yen 2010). All this can be only further promoted by the view that LR is the industry standard (Irimia-Dieguez, Blanco-Oliveer, and Vazquez-Cueto 2014; Martínes and Lechuga 2015; Yu et al. 2015).

The question whether LR and the synergy heuristic we follow are the modeling techniques that should be preferred over more sophisticated techniques is one that only researchers and corporate practice can answer. Absolutely certain is that this question requires further empirical experimentation. For example, the future work might include other modeling techniques for identification of most important variables and seek for novel opportunities for the functional form manipulation. Given that the advantages of the properties of variables importance measures have not been studied adequately yet (Ishwaran 2007), especially in the framework we propose, practitioners might find better performers or even identify novel derivatives among the existing importance measures. The scale of further studies should be extensive and multi-faceted especially in the credit scoring context. To generate our results for the industry domain there is a need for more case studies rooted to the research subject. In spite of what is often reported on performance of LR *versus* sophisticated techniques, we can only further promote LR and LR with the manipulated functional form.

Appendix A: Performance measures

PCC is derived from the confusion matrix, which presents the actual *versus* predicted class labels. To obtain this confusion matrix, the model-estimated probability of every i applicant is compared with a threshold. As a result, every observation is assigned to the positive class if the probability is higher than that threshold, and to the negative class otherwise. Thus, every applicant i can be marked to one of the 4 class labels, presented in the confusion matrix, as follows:

Table 1: Confusion matrix

	Actual negative $y = 0$	Actual positive $y = 1$
Predicted negative $\hat{y} = 0$	True negative TN	False negative FN
Predicted positive $\hat{y} = 1$	False positive FP	True positive TP

Given Table 1, PCC is defined as follows (e.g., Lessmann et al. 2015):

$$PCC = \frac{A}{B} \tag{1}$$

where $A = (TP + TN)$ and $B = (TP + TN + FP + FN)$.

BS represents the mean squared-error between probabilistic predictions and a zero-one-coded target variable and is defined as follows (Thomas, Edelman, and Crook 2002):

$$BS = \frac{1}{N}\sum_{i=1}^{N}(y_i - p_i)^2 \tag{2}$$

where p_i denotes the estimated default probability of case i and $y_i \in \{0, 1\}$ the actual class label.

AUC represents the classification performance averaged over all possible thresholds on ROC-curve and is defined as follows (Flach, Hernández-Orallo, and Ramirez 2011):

$$AUC = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} 1_{p_i > p_j} \tag{3}$$

where i runs over all M data points with the true label 1 (i.e., default event), and j over all N data points with the true label 0 (i.e., non-default event); p_i and p_j denote the probability

scores given to data point i and j, respectively. 1 is the indicator function. It outputs 1 if the condition $(p_i > p_j)$ is satisfied.

EMP finds a trade-off between the expected losses and the operational income by the loan and is defined as follows (Verbraken et al. 2014):

$$EMP = \int_{b_0} \int_{c_1} P(T(\theta); b_0, c_1, c^*) * h(b_0, c_1) dc_1 db_0 \tag{4}$$

where $h(b_0, c_1)$ is the joint probability density of the classification costs, θ is the cost-benefit ratio, T the optimal cutoff value, b_0 is the fraction of the loan amount, c^* is the cost of the action, and c_1 is the cost of incorrectly classified good applicants as defaulters. Thus, EMP accounts for the benefits generated by loans paid back and the costs caused by defaulters. EMP allows for profit-driven model selection, i.e., it identifies the credit model which most increases profitability.

References

Abdou, H. A., and J. Pointon. 2011. "Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature." *Intelligent Systems in Accounting Finance & Management* 18 (2-3): 59–88.

Abdou, H. A., M. D. Dongmo Tsafack, C. G. Ntim, and R. D. Baker. 2016. "Predicting Creditworthiness in Retail Banking with Limited Scoring Data." *Knowledge-Based Systems* 2013 (1): 89–103.

Aiken, L. S., and S. G. West. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Thousand Oaks: Sage.

Ala'raj, M., and M. F. Abbod. 2016a. "A New Hybrid Ensemble Credit Scoring Model Based on Classifiers Consensus System Approach." *Expert Systems with Applications* 64: 36–55.

———. 2016b. "Classifiers Consensus System Approach for Credit Scoring." *Knowlegde-Based Systems* 104: 89–105.

Angelini, E., G. Di Tollo, and A. Roli. 2008. "A Neural Network Approach for Credit Risk Evaluation." *The Quarterly Review of Economics and Finance* 48 (4): 733–55.

Askira-Gelman, I. 1998. "Knowledge Discovery: Comprehensibility of the Results." *Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences, IEEE Computer Society* 5.

Atish, P. S., and H. M. Jerrold. 2004. "Evaluating and Tuning Predictive Data Mining Models Using Receiver Operating Characteristic Curves." *Journal of Management Information Systems* 21 (3): 249–80.

Baesens, B., D. Roesch, and H. Scheule. 2016. "Credit Risk Analytics: Measurement Techniques, Applications, and Examples in Sas." Hoboken, New Jersey: John Wiley & Sons, Inc.

Baesens, B., T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen. 2013. "Benchmarking Stateof-the-Art Classification Algorithms for Credit Scoring." *Journal of the Operational Research Society* 54 (6): 627–35.

Bahnsen, A. C., D. Aouada, and B. Ottersten. 2014. "Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring." *13th International Conference on Machine Learning and Applications*, 263–69.

Bauer, D. J., and P. J. Curran. 2005. "Probing Interactions in Fixed and Multilevel Regression: Inferential and Graphical Techniques." *Multivariate Behavioral Research* 40 (3): 373–400.

Bequé, A., K. Coussement, R. Gayler, and S. Lessmann. 2017. "Approaches for Credit

Scorecard Calibration: An Empirical Analysis." *Knowledge-Based Systems*, 1–15.

Biau, G. 2012. "Analysis of a Random Forests Model." *The Journal of Machine Learning Research* 13 (1): 1063–95.

Blöchlinger, A., and M. Leippold. 2006. "Economic Benefit of Powerful Credit Scoring." *Journal of Banking and Finance* 30: 851–73.

Bravo, C., S. Maldonado, and R. Weber. 2013. "Granting and Managing Loans for Micro-Entrepreneurs: New Developments and Practical Experiences." *European Journal of Operational Research* 227 (2): 358–66.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

———. 2004. "Consistency for a Simple Model of Random Forests." *Technial Report, UC Berkeley*.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Taylor & Francis Ltd, Chapman & Hall, New York.

Caruana, R., A. Munson, and A. Niculescu-Mizil. 2006. "Getting the Most Out of Ensemble Selection." *In Proc. of the 6th Intern. Conf. on Data Mining. Hong Kong, China: IEEE Computer Society*, 828–33.

Chawla, N. V. 2005. *Data Mining for Imbalanced Datasets: An Overview*. In O. Maimon & L. Rokach, The Data Mining; Knowledge Discovery Handbook, New York: Springer Science+Business Media, Inc., 853-867.

Cole, S. A., M. Kanz, and L. F. Klapper. 2012. "Incentivizing Calculated Risk-taking: Evidence from an Experiment with Commercial Bank Loan Officers." *The Journal of Finance* 70 (2): 537–75.

Cole, S., M. Kanz, and L. Klapper. 2015. "Incentivizing Calculated Risk-Taking: Evidence from an Experiment with Commercial Bank Loan Officers." *Journal of Finance* 70 (2): 537–75.

Crone, S. F., S. Lessmann, and R. Stahlbock. 2006. "The Impact of Preprocessing on Data Mining: An Evaluation of Classifier Sensitivity in Direct Marketing." *European Journal of Operational Research* 173 (3): 781–800.

Crook, J. N., D. B. Edelman, and L. C. Thomas. 2007. "Recent Developments in Consumer Credit Risk Assessment." *European Journal of Operational Research* 183 (3): 1447–65.

Crouhy, M., D. Galai, and R. Mark. 2000. "A Comparative Analysis of Current Credit Risk Models." *Journal of Banking & Finance* 24 (1-2): 59–117.

Dietterich, T.G. 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning." *Neural Computation* 10 (7): 1895–1923.

Domingos, P. 1999. "The Role of Occam's Razor in Knowledge Discovery." *Data Mining and*

Knowledge Discovery 3 (4): 409–25.

Dong, G., K. Lai Kin, and J. Yen. 2010. "Credit Scorecard Based on Logistic Regression with Random Coefficients." *Procedia Computer Science* 1 (1): 2463–8.

Fawcett, T. 2006. "An Introduction to Roc Analysis." *Pattern Recognition Letters* 27 (8): 861–74.

Flach, P. A., J. Hernández-Orallo, and C. F. Ramirez. 2011. *A Coherent Interpretation of Auc as a Measure of Aggregated Classification Performance.* In L. Getoor & T. Scheffer (Eds.), Proc. of the 28th Intern. Conf. on Machine Learning (pp. 657-664). Bellevue, WA, USA: Omnipress.

Florez-Lopez, R., and J. M. Ramon-Jeronimo. 2015. "Enhancing Accuracy and Interpretability of Ensemble Strategies in Credit Risk Assessment. a Correlated-Adjusted Decision Forest Proposal." *Expert Systems with Applications* 42 (13): 5737–53.

Geurts, P., D. Ernst, and L. Wehenkel. 2006. "Extremely Randomized Trees." *Machine Learning* 63 (1): 3–42.

Gurný, P., and M. Gurný. 2013. "Comparison of Credit Scoring Models on Probability of Default Estimation for Us Banks." *Prague Economic Papers* 2: 163–81.

Hamerle, R., and D. Rösch. 2006. "Parameterizing Credit Risk Models." *The Journal of Credit Risk* 2 (4): 101–22.

Hand, D. J. 2006. "Classifier Technology and the Illusion of Progress." *Statistical Science* 21: 1–14.

Hand, D. J., and W. E. Henley. 1997. "Statistical Classification Models in Consumer Credit Scoring: A Review." *Journal of the Royal Statistical Society: Series A (General)* 160 (3): 523–41.

Hastie, T., R. Tibshirani, and J. Friedman. 2011. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer.

Hájek, P. 2011. "Municipal Credit Rating Modelling by Neural Networks." *Decision Support Systems* 51 (1): 108–18.

Irimia-Dieguez, A. I., A. Blanco-Oliveer, and M. J. Vazquez-Cueto. 2014. "A Comparison of Classifiction/Regression Trees and Logistic Regression in Failure Models." *Procedia Economics and Finance* 23: 9–14.

Ishwaran, H. 2007. "Variable Importance in Binary Regression Trees and Forests." *Electronic Journal of Statistics* 1: 519–37.

Khashei, M., and A. Mirahmadi. 2015. "A Soft Intelligent Risk Evaluation Model for Credit Scoring Classification." *International Journal of Financial Studies* 3: 411–22.

Kiefer, N. M., and C. E. Larson. 2006. "Specification and Informational Issues in Credit

Scoring." *International Journal of Statistics and Management Systems* 1: 152–78.

Kruppa, J., A. Schwarz, G. Arminger, and A. Ziegler. 2013. "Consumer Credit Risk: Individual Probability Estimates Using Machine Learning." *Expert Systems with Applications* 40 (13): 5125–31.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling*. New York: Springer Science+Business Media, Inc.

Lessmann, S., H.-V. Seow, B. Baesens, and C. L. Thomas. 2015. "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: A Ten-Year Update." *European Journal of Operational Research* 247 (1): 124–36.

Li, S., V. H. Nguyen, M. Ma, C.-B. Jin, C.-B. Do, and H. Kim. 2015. "A Simplified Nonlinear Regression Method for Human Height Estimation in Video." *EURASIP Journal on Image and Video Processing*. doi:10.1186/s13640-015-0086-1.

Martens, D., and B. Baesens. 2010. "Building Acceptable Classification Models." *Data Mining, Annals of Information Systems* 8: 53–74.

Martínes, S. J. F., and G. P. Lechuga. 2015. "Assessment of a Credit Scoring System for Popular Bank Savings and Credit." *Contaduría Y Administración* 61 (2): 1–27.

McGwin, G., G. Jackson, and C. Owsley. 1999. "Using Nonlinear Regression to Estimate Parameters of Dark Adaptation." *Behavior Research Methods, Instrument, and Computers* 31 (4): 712–17.

Ong, C. S., J. J. Huanga, and G. H. Tzeng. 2005. "Building Credit Scoring Models Using Genetic Programming." *Expert Systems with Applications* 29 (1): 41–47.

Paleologo, G., A. Elisseeff, and G. Antonini. 2010. "Subagging for Credit Scoring Models." *European Journal of Operational Research* 201 (2): 490–99.

Rodriguez, J. J., L. I. Kuncheva, and C. J. Alonso. 2006. "Rotation Forest: A New Classifier Ensemble Method." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (10): 1619–30.

Tasche, D. 2013. "The Art of Probability-of-Default Curve Calibration." *The Journal of Credit Risk* 9 (4): 63–103.

Thomas, L. C. 2000. "A Survey of Credit and Behavioral Scoring; Forecasting Financial Risk of Lending to Consumers." *International Journal of Forecasting* 16: 149–72.

———. 2010. "Consumer Finance: Challenges for Operational Research." *Journal of the Operational Research Society* 61: 41–52.

Thomas, L. C., D. B. Edelman, and J. N. Crook. 2002. *Credit Scoring and Its Applications*.

Philadelphia: Siam.

Tukey, J. W. 1997. *Exploratory Data Analysis*. Pearson.

Twala, B. 2010. "Multiple Classifier Application to Credit Risk Assessment." *Expert Systems with Applications* 37 (4): 3326–36.

Van Gestel, T., B. Baesens, P. Van Dijcke, J. A. K. Suykens, and J. Garcia. 2005. "Linear and Non-Linear Credit Scoring by Combining Logistic Regression and Support Vector Machines." *The Journal of Credit Risk* 1 (4): 31–60.

Verbraken, T., C. Bravo, Weber R., and B. Baesens. 2014. "Development and Application of Consumer Credit Scoring Models Using Profit-Based Classification Measures." *European Journal of Operational Research* 238 (2): 505–13.

Waagepetersen, R. 2010. "A Statistical Modeling Approach to Building an Expert Credit Risk Rating System." *The Journal of Credit Risk* 6 (2): 81–94.

Wang, J., G. Kun, and W. Shouyang. 2012. "Rough Set and Tabu Search Based Feature Selection for Credit Scoring." *International Conference on Computational Science*. doi:doi:10.1016/j.procs.2010.04.273.

Xie, H., S. Han, X. Shu, X. Yang, X. Qu, and S. Zheng. 2009. "Solving Credit Scoring Problem with Ensemble Learning: A Case Study." *Proc. of the 2nd International Symposium on Knowledge Acquisition and Modeling*, 51–54. doi:10.1109/KAM.2009.241.

Yu, L., X. Li, L. Tang, Z. Zhang, and G. Kou. 2015. "Social Credit: A Comprehensive Literature Review." *Financial Innovation*. doi:10.1186/s40854-015-0005-6.

Zhu, X., J. Li, D. Wu, H. Wang, and C. Liang. 2013. "Balancing Accuracy, Complexity and Interpretability in Consumer Credit Decision Making: A c-Topsis Classification Approach." *Knowledge-Based Systems* 52: 258–67.

Conversion uplift modeling in e-commerce: A benchmark study of recent modeling techniques

Authors, Affiliations, and Postal address:

Artem Bequé[1]
Robin Gubela[1]
Stefan Lessmann[1]
Fabian Gebert[2]

[1] School of Business and Economics, Humboldt-University of Berlin, Unter-den-Linden 6, 10099 Berlin, Germany
[2] Technology & Data Science Department, Akanoo GmbH, Mittelweg 121, 20148 Hamburg

Email:

artemlive@live.com
robin.gubela@hu-berlin.de
stefan.lessmann@hu-berlin.de
fabian@akanoo.com

Corresponding author:

Artem Bequé
School of Business and Economics, Humboldt-University of Berlin, Unter-den-Linden 6, 10099 Berlin, Germany
Email: artemlive@live.com
Tel.: +49 (0)30 2093 5742
Fax: +49 (0)30 2093 5741

Abstract

Uplift modeling is a combination of predictive modeling and experimental strategies to discern the differential effect of a treatment on individuals' behavior. Applications of such models are manifold and include marketing campaign planning, personalized medicine, and many more. This paper considers uplift models in the scope of targeting digital coupons in e-commerce. Using data from a broad set of online retailers, we perform a benchmarking experiment to compare a variety of alternative uplift modeling strategies. The study contributes to literature through i) consolidating prior work on uplift modeling approaches, which spread across diverse domains, ii) systematically comparing the predictive performance and utility of these approaches, and iii) examining the interaction between an uplift modeling strategy and its underlying learning algorithm, which facilitates making specific recommendations how to implement uplift model in e-commerce applications. Furthermore, the benchmark results allow us to quantify the degree to which targeting marketing communication using uplift models increases business value compared to conventional response models; an approach still in wide use in e-commerce.

Keywords: e-commerce analytics, machine learning, uplift modeling, real-time targeting

1 Introduction

Electronic commerce refers to the purchase and sale of goods and/or services via electronic channels (Mlelwa & Yonah, 2017). The business operates in all four of the major market segments, from B2B to C2C, e.g., (Bailey & Bakos, 1997; Jagtap, & Hanchate, 2017). The rapidly growing market and the increased internet availability give a vast opportunity for business to improve their relevance and expand their market presence in the online world. Nowadays, almost any product or service can be offered via electronic marketplace; from books and music (Brynjolfsson et al., 2010) to travel tickets (Escobar-Rodríguez & Carvajal-Trujillo, 2013) and financial services (Bakos, 1998). According to the Digital Commerce 360[1] report, researchers predict that by 2022 the digital commerce will be 17 percent of the U.S. retail sales and the U.S. will spend about $460 billion online in 2017.

Online marketplaces bring many positive effects with them. For example, they enable lower search costs (Baye et al., 2009) or overcome geographic isolation (Choi & Bell, 2011; Forman et al., 2009). However, they also suggest a fundamental shift to an increasingly competitive and challenging business environment. We observe that (i) lower search cost promote extreme price competition and can eliminate all seller profits (Bakos, 1998); (ii) there is little understanding of a link between a mobile user behavior and that on personal computers (Ghose et al., 2011); (iii) negative online consumer reviews can have a drastic influence on the perspective of information consumption (Lee et al., 2008); or (iv) customer retention orientation can be shaped by highly masked and knotty perception-based factors (Kang & Kim, 2017). To cope with all these challenges, companies execute digital marketing strategies to reach their market initiatives and to survive in the challenging environment.

Digital marketing uses scientific methods to convert massive amount of data into knowledge to drive the company to growth and profitability by, e.g., customer retention or development (Ascarza et al., 2017). Often marketing initiatives can be presented by intersection of departments (Doorn et al., 2017). For example, technology innovations and customer service approaches create exceptional customer experience (Thomas, 2017). Digital marketing is most successful when it is personalized and well-targeted, e.g., (Huang & Tsui, 2016; Simarmata & Ikhsan, 2017). That is marketers use response models (Coussement et al., 2015) to predict likelihood of customers to respond to a marketing campaign. There are many examples of response modeling in digital marketing from application of hybrid methods (Ahmed & Maheswari, 2017) and customer retention

[1] Data from the official website retrieved 22.09.2017 (https://www.digitalcommerce360.com/2017/08/09/e-commerce-grow-17-us-retail-sales-2022/)

modeling (Del Giudice & Peruta, 2017) to churn prediction using fuzzy classifiers (Azeem et al., 2017) and analysis of data sets with imbalanced nature (Gui, 2017). Given all merits, response modeling, however, suffers from some disadvantages.

The main disadvantage of response modeling is that it fails to differentiate between types of customer motivation. Put differently, response modeling does not take into account the behavior of customers (Michel et al., 2017; Kondareddy et al., 2016) who would take an action of interest irrespective of marketing campaigns, e.g., coupon targeting campaign (Daskalova et al., 2017; Ieva et al., 2017). This might have some unfavorable consequences for the entire company. For example, redundant marketing campaigns and communication might not only annoy the customers, but also result in a reduction of brand value. Recall that inaccurate targeting also implies higher direct marketing costs. To meet these challenges adequately, conversion methods for uplift modeling, e.g., (Jaroszewicz & Rzepakowski, 2014; Hansotia & Rukstales, 2002a), have been introduced to contact selected customers with the right offer through the right channel.

More specifically, conversion methods for uplift modeling identify customers who are likely to change their behavior in response to a marketing message (Kondareddy et al., 2016). This is equivalent to modeling the differential (i.e., causal) effect of a marketing incentive on customer behavior. The main advantage of conversion methods is that they can be paired with any response model technique, e.g., (Jaroszewicz & Rzepakowski, 2014; Hansotia & Rukstales, 2002a). This results, therefore, in an easygoing integration of these methods to the real-world business environments, no need of development of a new response model that accounts for uplift effects, and no sacrifice (eventually) of well-timed performance. Surprisingly, there have been made no attempt to systematically explore the potential of the conversion techniques for uplift modeling in well-rounded benchmark in prior work. The need of such a benchmark is clearly pronounced as all methods come from diverse disciplines and, therefore, there is no examination of the conceptual differences, comparison of predictive performance, and specific recommendations which techniques work better are missing.

The goal of this paper is, thus, to integrate previous literature on conversion methods for uplift modeling into one stream, to re-introduce available uplift methods to the marketing community, and to examine the degree to which alternative uplift methods contribute towards increasing the fit of marketing strategies for real-world practices through empirical experimentation. In pursuing this objective, we make the following contributions. First, we establish a thorough consolidation of state-of-the-art techniques for uplift modeling into combined unit. The comprehensive literature examination helps us understand and clarify the conceptual differences between the modeling techniques and through studying different streams of research provide an update on the modern

uplift modeling techniques. Second, we empirically evaluate the performance of the conversion methods for uplift modeling through large-scale experimentation. In particular, we benchmark the techniques by involving numerous data sets of different product lines from different geographies and provide a reference point for other academics and practitioners on the performance. Third, we consider multiple machine learning algorithms for the experimentation that we pair with every conversion method for the uplift modeling in a full-factorial setup. Thus, we shed light on the interaction between uplift conversion methods and underlying learning algorithms and provide specific recommendations which techniques work well together. Fourth, based on the benchmark results we quantify the degree to which targeting marketing campaigns using uplift modeling increases business value. That is, we explain which modeling technique contributes most (least) to business value. Finally, to clarify differences in performance between response modeling, a conventional method in the marketing applications, and uplift modeling methods, we compare the former with the latter in every experimental setup.

The remainder of the paper is organized as follows. We start by outlining conversion modeling using uplift and response models. We then summarize the related work coming from distanced strands of literature. Next, we elaborate the conversion methods for uplift modeling. We then describe the experimental design, including the campaign process, underlying data sets, model library, and performance indicators. This is followed by the experimental results. Finally, we conclude by discussing the main findings and providing an outlook for future research.

2 Conversion modeling using uplift vs. response models

Conversion modeling is used to tackle non-response rate and can be met in different marketing strategies and initiatives. For example, researchers develop measures to estimate the response rate based on multi-channel advertising (Zantedeschi et al., 2016); email questionnaires to investigate and combat non-response (Michaelidou & Dibb, 2006); or execute coupon targeting campaigns to increase the response rate (Daskalova et al., 2017; Ieva et al., 2017). We also observe that some scholars study purchase paths and conversion dynamics by comparing multiple websites (Park, 2017); develop complex multi-channel attribution modeling that is based on visitor journeys (Nottorf, 2014); and apply stochastic models to clickstream data (Lakshminarayan et al., 2016). Conversion modeling, therefore, finds broad application in marketing in general and e-commerce in particular to better understand the behavior of customers and to increase conversion rates.

In conversion modeling we differentiate between response and uplift modeling. The former estimates the probability of a customer to perform some action. Examples of such an action could

be a sign up for newsletter campaigns or a purchase of product. To that end, response models rely upon supervised classification algorithms (hereafter, base learners) which estimate a functional relationship between a binary class label, i.e., buyer or non-buyer, and a set of explanatory variables that capture customer characteristics. Such variables might include demographic, behavioral, and attitudinal information about the customer or, more generally, any piece of information, e.g., purchase paths, (Park, 2017), an analyst believes to be possibly linked to response.

To target customers by means of response modeling, candidate recipients are ordered according to model-estimated conversion probability and a fraction of top-ranked recipients is contacted; the size of the target group depends on the available budget and/or other business considerations. The subtle but crucial difference between a conventional response model and an uplift model is that the latter strives to identify customers the conversion probability of whom increases the most if they receive the marketing campaign (e.g., Rzepakowski & Jaroszewicz, 2012a). Therefore, an uplift model estimates a causal link between the action and how it alters customer behavior (i.e., conversion probability).

Uplift modeling, therefore, identifies customers who most likely responders (i.e., buyers) if being treated through a marketing campaign (Rzepakowski & Jaroszewicz, 2012a). Thus, uplift modeling makes it possible to measure success of a marketing campaign by quantifying causality between customer behavior and campaign process. Uplift modeling categorizes customer to four groups, presented in Table 1.

Table 1 Customer types as per uplift modeling

	Buyer without treatment Yes	Buyer without treatment No
Buyer being treated No	Do-Not-Disturbs	Lost Causes
Buyer being treated Yes	Sure Things	Persuadables

Table 1 indicates four types of customers. *Sure Things* respond regardless of any treatment, *Lost Causes* do not respond not even if being treated, *Do-Not-Disturbs* do not respond because of the treatment's negative impact on them and *Persuadables* respond because of the treatment (Rzepakowski & Jaroszewicz, 2012a). Table 1, thus, makes it clear that the only type of customers being targeted is the *Persuadables*. Targeting this type allows marketing managers to maximize the incremental number of purchases instead of gross purchases that is usually maximized by means of response modeling (Larsen, 2010). Apparently, targeting customers of different types than *Persuadables* induces a waste of marketing budget and, even more severe, targeting *Do-Not-*

Disturbs limits the source of revenue as the treatment unnecessarily makes these customers turn against their initial conversion intention.

Methodologically, response and uplift campaigns differ as follows: (i) in response modeling the whole customer base is subjected to probability estimation of being a buyer and a top fraction is targeted; whilst (ii) in uplift modeling the customer base is randomly partitioned to treatment and control group, separate probabilities of being a buyer are estimated in both partitions and based on their differences only those customers are targeted that most likely respond to a marketing pilot campaign if being treated.

3 Prior work in uplift modeling

Prior work in uplift modeling explores the development, application, and evaluation of the techniques for predictive modeling. All these techniques originate from diverse disciplines and model the relationship between the event of interest and a set of explanatory variables. We observe that some studies focus on the development of conversion methods for uplift modeling, e.g., Lo (2002), Tian et al. (2014), Rzepakowski and Jaroszewicz (2012a), which make it possible to couple these methods with any underlying base learner. Other authors manipulate the existing modeling techniques and convert them to account for uplift effects. For example, Radcliffe & Surry (1999), Hansotia & Rukstales (2002a), and Chickering & Heckerman (2000) manipulate specifically the decision trees. Other authors define uplift performance indicator (Kane et al., 2014) or concentrate strictly on variable important measures (Hua, 2016). Table 1 summarizes the existing prior work in uplift modeling and provides information on main topic of the studies, campaign experiments, industry and science, and data origin.

Table 2 Prior work in uplift modeling

Study	Main topic	Campaign (experiment)	Industry (science)	Data origin and access
Cai et al., (2009)	Two-stage estimation procedure for treatment differences for HIV-infected patients	Treatment 1: Therapy based on drug combination (zidovudine, lamivudine) Treatment 2: Therapy based on drug combination (zidovudine, lamivudine, indinarvir)	Clinical trials (medicine)	Licensed open-source real-world (AIDS Clinical Trials Group; see study ACTG 320 explained in Hammer et al., 1997)
Chickering & Heckerman (2000)	Greedy decision-tree learning algorithms (FORCE vs. NORMAL)	Mail advertisement for MSN subsription	Software	Private real-world (anonymized contracting authority)
Dost et al., (2014)	Willingness-to-pay (WTP) range-based targeting approach	Experiment 1: Discount offer Experiment 2: WOM (T1), visual (T2), information (T3) Experiment 3: Discount (T1), guarantee (T2) Experiment 4: Participation	various	Surveys in different settings Participants from Amazon Mechanical Turk Participants from Amazon Mechanical Turk Students from a German University Consumers from an Agency Panel
Guelman (2014)	Personalized treatment learning problem, uplift random forest and uplift causal conditional inference forest and uplift software description	E-mail promotion to buy a certain product at a bank	Financial services	Private real-world (anonymized contracting authority)
Guelman, et al. (2012)	Uplift random forests	Treatment 1: Letter (retention) Treatment 2: Letter plus outbound courtesy call (retention)	Insurance	Private real-world (anonymized contracting authority)
Guelman et al., (2014)	Causal conditional inference trees in personalized treatment learning	Direct mail campaign (cross-selling)	Insurance	Private real-world (anonymized contracting authority)
Guelman et al., (2015)	Uplift random forests	Information letter plus courtesy call (as one treatment)	Insurance	Private real-world (anonymized contracting authority)
Hansen & Bowers (2008)	Stratification to balance the distributions of pretreatment variables	Especially: GOTV field experiment (GOTV messages: personal visit, phone call, mailing) and simulation studies	Social and political sciences	Get-Out-The-Vote (GOTV) field experiment (Gerber & Green, 2000)
Hansotia & Rukstales (2002a)	CHAID decision tree with $\Delta\Delta P$ split criterion	Mail promotion ($10 off a purchase of at least $100 basket value)	Holiday retail	Private real-world (anonymized contracting authority)
Hansotia & Rukstales (2002b)	Concept of uplift tree-based approaches	-	-	-
Hua (2016)	Uplift random forests in capital market research with focus on results of embedded variable selection procedure	No campaign conducted	Banking	Licensed open-source real-world (heterogenous data sources)
Imai & Ratkovic (2013)	Estimation of heterogeneous treatment effects as a variable selection problem with modified support vector machines	GOTV field experiment (GOTV messages: personal visit, phone call, mailing) and simulation studies	Social and political sciences	Get-Out-The-Vote (GOTV) field experiment (Gerber & Green, 2000)
Jaroszewicz & Rzepakowski (2014)	Uplift modeling for survival analysis	Chemotherapy against colon cancer Treatment 1: Therapy with Levamisole Treatment 2: Therapy with Levamisole plus 5-FU (Fluorouracil)	Clinical trials (medicine)	Open-source real-world (UCI repository)
Jaroszewicz & Zaniewicz (2016)	Uplift support vector machines with Székely regularization	Therapy with right heart catheterization procedure (RCH)	Clinical trials (medicine)	Open-source real-world (Connors et al., 1996)
Jaskowski & Jaroszewicz (2012)	Response variable transformation	Experiment 1: Therapy with peripheral blood transplant Experiment 2: Therapy with tamoxifen plus radio therapy against breast cancer Experiment 3: Therapy with steroids against hepatitis	Clinical trials (medicine)	Open-source real-world (Pintilie, 2006) Open-source real-world (Pintilie, 2006) Open-source real-world (UCI repository)
Kane et al., (2014)	Comparison of four uplift model solutions; signal-to-noise (S/N) ratio	Experiment 1: Direct mail (paper) Experiment 2: E-mail Experiment 3: Direct mail (paper)	Financial services Online merchandise Retail office supplies	Private real-world (anonymized contracting authority)

Kuusisto et al., (2014)	Uplift support vector machines	Simulated marketing activity	-	Simulation data
Larsen (2010)	Uplift k-nearest neighbour and variable selection in uplift modeling	-	-	-
Lo (2002)	Interaction term approach	-	-	Simulation data
Lo & Pachamanova (2015)	Multiple treatment optimization approach for prescriptive uplift analytics	E-Mail campaign (men and women separately targeted)	Online retail	Open-source real-world (Hillstrom, 2008)
Manahan (2005)	Uplift neural network implementation with SAS modules	Contract renewal campaign	Wireless telecommunications	Private real-world (Cingular)
Nassif (2013)	Alternative uplift evaluation measures (ROC)	Therapy against breast cancer	Clinical trials (medicine)	Open-source real-world (Nassif et al., 2010)
Nassif (2013)	Multi-relational uplift modeling system for medical research (SAYL algorithm)	Therapy against breast cancer	Clinical trials (medicine)	Open-source real-world (Nassif et al., 2012)
Radcliffe (2007)	Uplift evaluation measures	Experiment 1: Catalogue mailing (deep-selling) Experiment 2: Retention offer (retention campaign) Experiment 3: Mail targeting (cross-selling)	Retail Telecommunication Banking	Private real-world (anonymized contracting authority)
Radcliffe & Surry (1999)	Fundamental idea of uplift modeling with reference to differential response analysis (first uplift paper ever published)	-	-	-
Radcliffe & Surry (2011)	Significance-based uplift decision trees with several key features; uplift evaluation measures	-	-	-
Rzepakowski & Jaroszewicz (2012a)	Uplift decision trees with different splitting criteria	E-Mail campaign (men and women separately targeted)	Online retail	Open-source real-world (Hillstrom, 2008)
Rzepakowski & Jaroszewicz (2012b)	Uplift modeling for multiple treatments	No campaign conducted (artificial allocation of observations to either treatment or control group in 16 datasets)	-	Open-source real-world (UCI repository)
Shaar (2016)	Pessimistic uplift modeling approach to minimize disturbance effects	Simulated campaigns/treatments in marketing and medicine	-	Open-source real-world (UCI repository; Hillstrom, 2008; Pintilie, 2006)
Siegel (2013)	Overview of uplift modeling methods and models	-	-	-
Sołtys et al., (2015)	Ensemble methods for uplift modeling (bagging, random forest)	Simulated campaigns/treatments in marketing and medicine	-	Open-source real-world (UCI repository; Hillstrom, 2008; Pintilie, 2006)
Su et al., (2012)	Causal inference trees and uplift k-nearest neighbor approach in assessing treatment effects	Synthetic data creation (uniform distribution)	Machine learning research	Simulation data
Tian et al. (2014)	Investigation of the effects of a transformation of input space on a certain outcome of interest in medical research	1. Study of the implications of ACE inhibitors on lowering cardiovascular risk for patients with stable coronary artery disease and normal or slightly reduced left ventricular function 2. Study of interactions between gene expression levels and Tamoxifen treatment in breast cancer patients	Clinical trials (medicine)	1. Preventive of Events with Angiotension Converting Enzyme Inhibition (PEACE) study (Braunwald et al., 2004) 2. Breast cancer dataset consisting of 414 patients in the cohort GSE6532 (Loi et al., 2007)
Yong (2015)	Prediction inference procedure with stratification to obtain valid and generalizable predictions for medical examinations	several	Clinical trials (medicine)	several; among them the Mayo liver study data
Zaniewicz & Jaroszewicz (2013)	Uplift support vector machines (USVM)	Simulated campaigns/treatments in marketing and medicine	-	Open-source real-world (UCI repository; Hillstrom, 2008; Pintilie, 2006)

Table 2 reveals several important findings. First, we observe high diversity in the research of uplift modeling. In particular, research and development are performed in multiple industries and sciences. For example, we see such industries as software, medicine, banking, insurance, and many other. All of them conduct various experiments and campaigns to further develop and examine the performance of the uplift modeling techniques. Also, all these studies seek to describe different events of interest. For example, Dost et al. (2014) are interested in the study of the willingness-to-pay when applying a range-based targeting approach, Hansen & Bowers (2008) examine the stratification to balance distributions of pre-treatment variables, or Jaroszewicz & Rzepakowski (2014) apply uplift modeling techniques in survival analysis. High diversity results in further progress of uplift modeling that goes alongside multiple domains.

Second, Table 2 makes it clear that most research is directed to the modification of the existing integrative modeling techniques. Often the researchers manipulate the tree-based uplift algorithms. The first uplift paper (Radcliffe & Surry, 1999) deals with decision trees and thereafter we observe a lot of studies that go in this direction. For example, Hansotia and Rukstales (2002a) grow the CHAID-type decision trees with uplift-specific split criteria, Chickering and Heckerman (2000) design modified greedy trees that incorporate cost-benefit analysis based on measures of expected profit, or Rzepakowski and Jaroszewicz (2012a; 2012b) develop in terms of pruning to maximize the difference between treatment and control class distributions. To achieve this, they adapt split criteria based on conditional divergence measures from information theory. Tree-based learning algorithms in Radcliffe and Surry (2011) differentiate themselves from others due to the assessment of the statistical significance of the differences in class probabilities. Guelman et al. (2015) propose to re-design random forest models so that they are capable to predict uplift. Trees are grown based on the determination of the best candidate split among a random subset of independent variables per node (Guelman et al., 2015). Besides uplift random forests, Guelman et al. (2014) have established uplift causal conditional inference trees and demonstrated their merits over uplift random forests on insurance data. The empirical work in marketing and medicine of Sołtys et al. (2015) suggests using uplift ensemble methods.

Third, only few studies (compared to modification) concentrate on the conversion methods for uplift modeling and these studies use yet few uplift techniques for empirical comparisons. Kane et al. (2014) provides uplift model comparisons and presents the generalized Lai's (2006) method to weighted uplift model. Given that the majority of models are developed for single treatments, Lo & Pachamanova (2015) and Rzepakowski & Jaroszewicz (2012b) develop models for multi-treatment settings. These techniques are either tree-based or related to multiple, treatment-related logistic regression models. Jaroszewicz & Rzepakowski (2014) were first to apply uplift methods

for survival analysis in a medical landscape and Lo (2002) as well as Tian et al. (2014) introduce uplift approaches based on transformed data input spaces. The methodology to modify the output space (i.e., response variable) to facilitate prediction of uplift effects is described by Jaskowski and Jaroszewicz (2012). Rzepakowski & Jaroszewicz (2012a) and Jaroszewicz & Rzepakowski (2014) seek to combat performance issues in the straightforward two-model uplift approach by pointing to the different behaviors of class probabilities. Shaar et al. (2016) refer to disturbance effects of uplift models that limit prediction reliability. To cope with these effects, authors combine diverse uplift techniques, including Lai (2006) uplift model and reflective uplift modeling in a weighted procedure to derive a pessimistic uplift score.

Fourth, most studies model only one base learner and do not consider any further for empirical examination. We observe a high interest to decision trees in many studies, e.g., Radcliffe & Surry, (1999); Hansotia and Rukstales (2002a); Chickering and Heckerman (2000); and many other. For example, Larsen (2010) suggests considering a variable selection approach for uplift based on an adjusted net information value measure. For capital market research purposes, Hua (2016) determines important variables by making explicit use of an uplift random forests' variable importance outcome. Other studies have shown interest in such base learners as logistic regression (Lo, 2012), neural networks (Manahan, 2005), and k-nearest-neighbors (Larsen, 2010). We also see that many authors execute support vector machines for uplift modeling, e.g., Jaroszewicz & Zaniewicz (2016), Kuusisto et al. (2014), and Zaniewicz & Jaroszewicz, (2013).

All these findings imply some concerns for e-commerce in general and marketing campaigns by means of couponing (targeting we describe in this study) in particular. Specifically, we see that all these uplift techniques come from diverse strands of literature and, therefore, a systematic comparison of the performance of these techniques is missing. Given that marketers use response modeling conventionally (Coussement et al., 2015), modification of these methods to techniques that account for uplift effects would mean additional efforts and, eventually, sacrifice of well-timed performance. Therefore, we regard conversion methods for uplift modeling as more beneficial for e-commerce since they make it possible to apply response techniques for uplift modeling without a need of modification. However, we also see that there are only few papers that focus on conversion methods and mostly they lack on empirical comparison. Given that uplift literature considered only limited amount of response models, specific recommendations which models work better are missing. In this study, therefore, we benchmark conversion methods for uplift modeling that we pair with multiple base learners and seek to close this research gap.

4 Conversion methods for uplift modeling

In this study, we empirically benchmark nine techniques for the conversion uplift modeling previously used in multi-faceted settings and present them in Table 3 (response model being discussed in 2). With this choice, we are confident to provide a wide portfolio of state-of-the-art conversion methods in uplift modeling. Recall that conversion methods enhance execution of the standard classification procedures in uplift modeling. As a result, the conversion methods can be practiced directly in e-commerce strategies and initiatives such as customer acquisition, customer development (Kane et al., 2014; Blattberg et al., 2001), or increase of customer retention rate (Guelman et al., 2015) without a need to modify base learners.

Table 3 Conversion methods for uplift modeling

Source	Conversion Method	Acronym
Jaskowski & Jaroszewicz (2012)	Class Variable Transformation	CVT
Lo (2002)	Interaction Term Method	ITM
Kane et al. (2014)	Lai's Generalized Weighted Uplift Method	LGWUM
Lai (2006)	Lai's Weighted Uplift Method	LWUM
Shaar et al. (2016)	Pessimistic	PESSIMISTIC
Shaar et al. (2016)	Reflective	REFLECTIVE
Various	Standard Conversion Response Modeling	RESPONSE
Tian et al. (2014)	Treatment-Covariates Interactions Approach	TCIA
Various	Two Model Uplift Approach	TWO_MODEL

Consider a training set $TRAIN_m = \{(x_i, y_i)\}_{i=1}^{m}$ of m customers gathered, for example, by means of a pilot campaign (see Figure 2). Every customer is characterized by a set of explanatory variables x_i and a binary variable $y_i \in \{0, 1\}$ that indicates whether a conversion has been observed for customer. We refer to y_i are the binary variable that we seek to explain. Let T_i and C_i indicate the membership of customer i to the treatment or control group, with prior probability distributions $P(T_i)$ and $P(C_i)$. Then, $P(Y_i = 1|T_i, X_i)$ and $P(Y_i = 1|C_i, X_i)$ denote the conditional probability of conversion for treatment and control group customers, respectively. For notational convenience, we refer to these conditional probabilities as $P(Y_i|T_i)$ and $P(Y_i|C_i)$ in the following. Furthermore, we define the four unconditional probabilities as follows: $P(T_i \cap Y_i)$ treated and response, $P(T_i \cap \bar{Y}_i)$ treated and non-response, $P(C_i \cap Y_i)$ non-treated and response, and $P(C_i \cap \bar{Y}_i)$ non-treated and non-response.

The ***two model*** uplift conversion method, e.g., (Rzepakowski & Jaroszewicz, 2012a) captures the difference in class probabilities by providing a mechanism to differentiate between structures of customers' motivation:

$$Uplift_i^{TWO\_MODEL} = P(Y_i|T_i) - P(Y_i|C_i) \quad (1)$$

Building and predicting with two equal learning algorithms given these two samples constitutes the methodology of the two model uplift approach. In contrast, the typical response model predicts $P(Y_i|T_i)$.

Lo (2002) presents a modification method - ***ITM*** - of the explanatory variables. In particular, a dummy variable D_i is obtained with $d_i \in \{0, 1\}$ for control and treatment group, respectively. D_i is then multiplied with the entire X_i input space to gain an interaction term that is used in model prediction:

$$Uplift_i^{ITM} = P(Y_i|D_i, X_i \cdot D_i) \quad (2)$$

To derive uplift effect, predictive scores of the models developed for control group are subtracted by their equivalents for treatment group. Evidently, the interaction term emphasizes treatment contrarily to control group observations.

Tian et al. (2014) present another approach - ***TCIA*** - to modify the explanatory variable. More specifically, TCIA mimics the same procedure rooted to ITM with slight adaptions to D_i. Namely, D_i^* is obtained to represent multi-dimensional vector of adapted baseline covariates (Guelman et al., 2014) and defined as $D_i^* = \frac{X_i^* \cdot D_i}{2}$, whereby X_i^* is a mean-centered dimension of explanatory variables values and values for $d_i \in \{-1, 1\}$. In essence, the uplift effect is now modeled as:

$$Uplift_i^{TCIA} = P(Y_i|D_i^*) \quad (3)$$

Tian et al. (2014) have developed this approach and implemented a multivariate regression learner to predict the modified data. Guelman et al. (2014) further validated this approach in simulation experiments.

Jaskowski & Jaroszewicz (2012) represent a transformation procedure - ***CVT*** - that assigns zero and one to the transformed variable which depends on the conversion indicator and the group membership status of every individual customer. Let Z_i be a binary transformed response variable on a customer level and $z_i = 1$ if $(T_i \cap Y_i) \cup (C_i \cap \bar{Y}_i)$ is given; otherwise $z_i = 0$. Thus, the uplift effect is defined as:

$$Uplift_i^{CVT} = 2 \cdot P(z_i = 1) - 1 \quad (4)$$

Thus, it is clear that modeling conditional uplift of the conversion variable equalizes modeling the conditional distribution of the transformed conversion variable.

Lai (2006) presents an extension - ***LWUM*** - of CVT that weights probabilities of positive and negative classes. LWUM assumes that the positive uplift lies in correctly identified *Persuadables* (i.e., treatment-group responders and control-group non-responders), whilst the negative uplift can be found in the *Do-Not-Disturbs* group (i.e., treatment-group non-responders and control-group responders). Therefore, let W be the number of positive observations divided by the total population. The uplift effect is then defined as:

$$Uplift_i^{LWUM} = P(z_i = 1) \cdot W - P(z_i = 0) \cdot (1 - W) \tag{5}$$

LWUM, thus, seeks to maximize the positive uplift while decreasing the negative one in the first decile.

Kane et al. (2014) present - ***LGWUM*** - the generalized version of LWUM with weighted probability scores that realize the influence of the fraction of treatment and control group customers on the lift measure and is defined as:

$$Uplift_i^{LGWUM} = P(Y_i|T_i) + P(\bar{Y}_i|C_i) - P(\bar{Y}_i|T_i) - P(Y_i|C_i) \tag{6}$$

The authors emphasize that an important merit of LGWUM compared to LWUM is that the former can be applied even if the ratio of treatment and control assignments is not approximately equal.

Shaar et al. (2016) present a method - ***reflective*** - by two separate models that are built to learn the treatment effect in the conversion and non-conversion groups. The authors recognize severe disturbance effects when applying uplift models. The first one is a response effect that takes place due to correlation between explanatory variables and a binary class label and the second - a partitioning effect - that appears when the treatment indicator depends on the covariates. To overcome these negative effects reflective uplift modeling has been introduced. The uplift effect is then calculated, whereas the groups are treated as positive and negative like in the CVT approach:

$$Uplift_i^{REFLECTIVE} = P(T_i \cap Y_i) \cup P(C_i \cap \bar{Y}_i) - P(T_i \cap \bar{Y}_i) \cup P(C_i \cap Y_i) \tag{7}$$

Thus, the probabilities for positive and negative groups are obtained from two different models. To determine a score in terms of pessimistic uplift modeling, LWUM is again considered. The final ***pessimistic*** approach is defined as:

$$Uplift_i^{PESSIMISTIC} = 0.5 \cdot \left(Uplift_i^{LWUM} + Uplift_i^{REFLECTIVE}\right) \tag{8}$$

5 Experimental setup

We examine the relative merits of the conversion methods for uplift modeling. We involve numerous data sets that belong to the field of e-commerce, indicating the goal to categorize customer base into two classes: buyer and non-buyer. In the following, we elaborate the campaign process and underlying data, base learners that we pair with the conversion methods for uplift modeling, and finally the performance metrics.

5.1 Campaign process and data

The experimental setup involves 27 data sets that come from a customer acquisition marketing campaign. This campaign has been carried out in multiple electronic marketplaces and designed so that customers who show specific behavior patterns during their store-related customer journey are identified by an uplift model and targeted with an online coupon. Customers that leave the respective store by having activated this coupon obtain a discount of 10% off their final basket value. A real-time targeting process (Ding et al., 2015) has been applied to identify customers to receive the coupon. Every customer has been assigned either to the treatment or control group by chance or by a model. In the latter case, the individual online behavior of new customers is considered after five pageviews and that of returning customers after three pageviews. The derived predictive scores determine whether the customer is likely to be *persuadable* (i.e., customer with high probability to respond if being treated with coupon). As a result, the model qualifies the customers to the treatment group that then receive a coupon. The systematic component of the targeting process creates a selection bias that leads to a quasi-experiment.

Table 4 summarizes the electronic retail data sets, including product line, geographical location, and number of cases as per data set.

Table 4 Summary of the e-retail data sets

Shop ID	Product Line	Geographical Market	No. of cases
1	Apparel	Poland	275,325
2	Apparel	Germany	171,936
3	Apparel	Germany	48,615
4	DIY products	United Kingdom	289,512
5	Apparel	Czech Republic	63,267
6	Apparel	Germany	11,610
7	Books and multimedia	Germany	12,033
8	Toys	Germany	1,200,129
9	DIY products	Germany	124,086
10	DIY products	France	12,780

11	Pharmaceuticals	Germany	6,999
12	Special apparel (hats)	Germany	22,314
13	Apparel and household items	France	63,864
14	Fan articles and toys	Germany	12,837
15	Apparel	Germany	24,450
16	Apparel	The Netherlands	7,326
17	Alcoholic beverages	Germany	9,396
18	Pharmaceuticals	Germany	8,697
19	Sports apparel and accessories	Germany	111,630
20	Pet food	Germany	22,482
21	Apparel	Germany	119,565
22	Shoes and accessories	Germany	326,232
23	Pharmaceuticals	Germany	5,343
24	Apparel	Austria	27,768
25	Shoes	Germany	3,204
26	Special apparel (hats)	The Netherlands	10,452
27	Outdoor apparel	Germany	60,138

Table 4 indicates that the consumer goods relate to different sort of apparel, toys, garden article, books and multimedia, pet food, and many other. In addition, sports and outdoor articles are also sold in a few stores. Businesses operate in Germany, France, Austria, the United Kingdom, the Czech Republic (CP) and many other. In total, the data we obtained from a partner in industry entails three million cases and over 60 features that profile the customers' behavior. Every observation relates to the store-based journey performed during a flexible time span (i.e., from entering to leaving the store). Cookie technology allows to differentiate between new and returning customers. Most features are numeric and the rest are factors. These features describe information on nearly every activity of a customer on website: How long the customer spends on certain page types, whether the customer has been interested in purchasing a certain product at the same store in the past, how much time has passed since the customer first added an item to the shopping cart, how many views the customer has made on a sale-related page and how many products lie in the customer's shopping basket for how long. Even technical information is collected such as the scroll height when the customer enters product-related pages for closer examination or the screen width of the customer's gadget. Inspiration on data collection has been gained from Van den Poel and Buckinx (2005). Furthermore, certain meta dimensions that are crucial for the uplift modeling have been collected: (i) a unique identifier of the respective store and time stamp; (ii) an indicator on group (treatment or control) assignment; and (iii) a variable that captures the purchase event.

Another important concern relates to data partitioning. We have created three partitions from all available data: 40% training partition that we use to train the techniques, 30% for a parameter-tuning partition that we use to validate the meta-parameter tuning, and another 30% for a test

partition. To guarantee a reliable evaluation, we apply a 10-fold cross validation scheme "through time" to reflect the situation in marketing practice and increase the size of observations by resampling. For all approaches the stated models first predict on the training and parameter-tuning partitions together. Those models per approach with the identified best candidate settings are then validated on the final validation sample to assure a reliable benchmark.

5.2 Base learners

The experimental design includes six base learners. Recall that we benchmark conversion methods for uplift modeling that can be paired with any base learner. Thus, we secure every possible combination between conversion methods and base learners. The experiment is performed in Python environment. To guarantee robust results, we execute a wide range of meta-parameters for every base learner, presented in Table 5. Every model is tuned automatically and transmitted to cross validation technique discussed previously. In total, we involve 280 models.

Table 5 Meta-parameters of the base learners

Base learner	Acronym	Models	Meta-parameter	Candidate setting
Logistic regression	LogR	34	Regularization term Regularization factor	[L1, L2] [1e-8, 1e-7, …, 1e8]
Support vector machines with linear kernel	SVC	42	Regularization factor Calibration method	[1e-10, 1e-9, …, 1e10] [Sigmoid, Isotonic]
k-Nearest-Neighbor	KNN	19	Number of nearest neighbors	[1, 5, 10, 100, …, 500, 1000, …, 4000]
Naïve Bayes	NB	1	-	-
Stochastic gradient descent for classification	SGDC	180	Loss function Regularization term Alpha Learning rate	[Log, Mod. Huber, Hidge, Percep.] [L1, L2, Elastic Net] [1e-6, 1e-5, …, 1e-1] [Optimal, Invscaling]
Random forest for classification	RFC	4	Max. no. of covariates Min. no. of samples	[8, 9] [1000, 2000]

We consider six base learners to ensure a vast benchmark study. In particular, we pair base learners and conversion methods in a full-factorial experimental setup. Recall that we involve nine conversion methods; this, thus, results in 2,520 models in grand total. We choose specifically these base learners due to their popularity not only in response but also in uplift modeling (see 2). In response modeling, for example, these base learners are often questioned in pivotal benchmark studies (Baesens et al., 2003; Lessmann et al., 2015). SGDC and RFC demonstrate excellent performance in real-world experiments (Guelman et al., 2015). Due to the fact that RFC is less sensitive to meta-parameter adaptations than SGDC (Ogutu et al., 2011), we consider for RFC a smaller

number of models. In uplift modeling, LogR (Lo, 2002), KNN (Larsen, 2010), and SVC (Kuusisto et al., 2014; Zaniewicz & Jaroszewicz, 2013) have gained a strong research interest. As a standard base learner without meta-parameters, we add a NB algorithm to the library of base learners.

5.3 Validation measures

Typically, the performance of predictive models grounds on a comparison of actual versus predicted outcomes. In uplift modeling, however, this is not reasonable since a customer cannot be part of both the treatment and control group. This phenomenon is known as the fundamental problem of causal inference (Holland, 1986). Consequently, today's best practice is a segment- or decile-based evaluation approach to identify uplift and to capture the performance results in terms of Qini coefficient Q and to visualize them in uplift gains charts by means of Qini curves (Radcliffe, 2007). This includes the underlying assumption that similarly scored cases behave likewise, i.e., the k percent highest scores on treatment out-of-sample test data are compared to the k percent highest scores on control out-of-sample test data and with the subtraction of the top gains from both groups a meaningful estimate of uplift can be derived (Jaskowski & Jaroszewicz, 2012). Q is, thus, defined as the area between a model's Qini curve and a random targeting line (Radcliffe & Surry, 2011). Because typically uplift gains charts display Qini curves that relate to a cumulative measure, we further consider uplift bar charts that mask the effect of cumulativeness to provide a decile-isolated analysis of model performance.

6 Empirical results

The experimental results consist of the performance estimates for every combination of 6 levels of base learners, 9 levels of uplift modeling techniques, and 27 levels of data sets. The performance measures capture the degree to which the marketing campaign strategy improves via application of uplift modeling techniques in terms of Qini coefficient and cumulative (non-cumulative) number of incremental purchases.

6.1 Examination of the interaction between uplift techniques and base learners

To identify the synergy effects between the modeling methods and to provide specific recommendations regarding which techniques work well together, we now examine the interaction between the base learners and conversion methods for uplift modeling. Table 6 summarizes the corresponding results. To obtain it, we pair every base learner with all uplift modeling techniques and

capture the predictive performance on the out-of-sample test set in terms of Qini coefficient. These values are averaged over the data sets. We express the Qini coefficient in percentage terms, i.e., Q_{pct}, by subtracting the control group response rate from the treatment group response rate for every decile. In contrast to the general Q coefficient (Radcliffe & Surry, 2011), Q_{pct} makes comparisons across the data sets with different number of observations possible and, thus, requires no normalization procedure. To increase the readability of Q_{pct}, we multiply its values with a factor of 1,000. We use bold face for every best combination (i.e., uplift technique coupled with base learner). For example, the very right value for CVT is marked in bold face indicating that CVT interacts best with RFC.

Table 6 Qini coefficient across the modeling techniques

Uplift method	**Base learner**					
	KNN	LogR	NB	SGDC	SVC	RFC
CVT	3.171	3.348	-0.951	-1.041	2.017	**6.145**
ITM	3.991	2.901	3.770	0.979	**8.017**	3.216
LGWUM	-0.230	3.767	-4.459	1.831	-0.932	**5.593**
LWUM	3.171	4.258	-0.945	0.203	2.049	**6.130**
PESSIMISTIC	1.418	4.269	-1.626	0.720	2.010	**6.606**
REFLECTIVE	-1.526	**3.310**	-2.914	0.868	-0.727	2.303
TCIA	1.043	-1.950	-2.821	1.222	**3.893**	0.403
TWO MODEL	**7.267**	4.305	3.297	0.688	2.806	5.401

Table 6 reveals multiple important findings. First, the best possible interaction is between ITM and SVC with Q_{pct} of 8.017. This is followed by two model coupled with KNN with Q_{pct} of 7.267 and CVT with RFC of 6.145. This strongly signals in favor of ITM as a conversion method for uplift modeling and of SVC as a base leaner. This view is only strengthened when we look at the pair of TCIA and SVC, where SVC is the best performer. However, we recommend RFC as a base learner for uplift modeling since it collects the biggest number of wins. More specifically, RFC is the best performer when coupled with CVT, LGWUM, LWUM, and PESSIMISTIC. We observe that KNN performs best when paired with the two model approach and the differences in the performance compared to other uplift methods are substantial. For example, the pair two model and KNN achieves Q_{pct} of 7.267 compared to the second best performer pair of ITM and KNN with Q_{pct} of 3.991 and the worst performer pair of reflective and KNN with Q_{pct} of -1.526. As a result, we can only recommend to consider KNN when coupled with two model method. We also observe that reflective method performs best coupled with LogR. However, LogR shows also high and better potential when interacting with other uplift techniques. For example, Q_{pct} of couples of

pessimistic, LWUM, and two model with LogR is higher than that of reflective with LogR. Thus, LogR seems to be more flexible than KNN for uplift modeling. On the contrary, NB and SGDC have no wins. This implies that these base learners present a very weak performance compared to other. Thus, we cannot recommend to execute them for uplift learning. This recommendation is supported by the fact that, e.g., NB collects a big number of negative Q_{pct} values. The same we see for the pair of CVT and SGDC. We also would like to stress that the best pessimistic uplift model outperforms all base learners related to LWUM and reflective methods. This is interesting since LWUM and reflective hold equal shares in creation of pessimistic method. LGWUM does not add more performance value to LWUM. SGDC being exception, all base learners paired with LWUM outperform their equivalents for LGWUM. Analogous picture we see for covariate transformations. All base learners but SGDC and SVC paired with CVT obtain higher Q_{pct} values when compared to respective TCIA counterparties.

To further facilitate findings obtained from Table 6, we compare the performance of the modeling techniques through a robustness procedure. In particular, we capture the performance of the uplift methods coupled with base learners in a 10-fold cross validation and visualize it in Figure 1. Every box plot portrays base learners on the *x-axis* and the performance measured in Q_{pct} on the *y-axis*. We scale the Q_{pct} values to ease comparability.

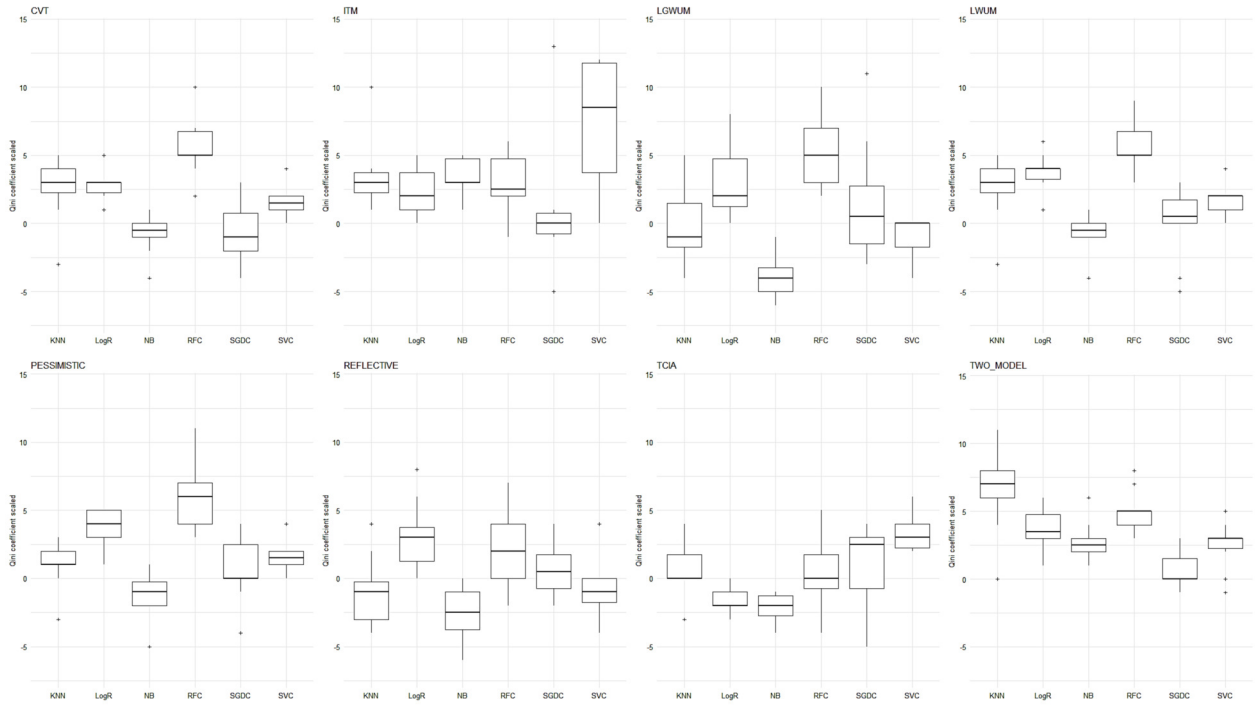

Figure 1 Scaled Qini coefficient across the modeling techniques

Figure 1 supports some previous findings but also reveals new ones. First, we would like to highlight remarkable performance of RFC. RFC is the best performer when coupled with, e.g., CVT, LGWUM, or PESSIMISTIC. Furthermore, RFC shows relatively small variance. This can be especially emphasized on the combination of RFC with two model. Thus, Figure 1 further supports the view that RFC is a very suitable alternative for uplift modeling. Second, very weak performance of NB and SGDC is further pronounced in Figure 1. We observe that the mean values of NB are negative for pessimistic, reflective, or CVT methods. The same we see on the couple of CVT and SGDC, whereby SGDC also exhibits higher variance than NB. These findings caution from execution these base learners for uplift modeling. Third, we see a very high variance of SVC when coupled with ITM. This finding injects doubt on the previous one where a couple ITM and SVC is the best performer (Q_{pct} of 8.017). Hence, we conclude that ITM paired with SVC does not provide a reliable estimate. In contrast, we observe that KNN paired with two model shows very low variance that makes this couple more promising than ITM and SVC (given findings from Table 5). To be more concrete, ITM-based SVC shows a standard deviation of 0.0083 whereas this of KNN is 0.0060. As a result, KNN has a 27% lower standard deviation than SVC. Note that the standard deviation values are percentages derived from taking the mean of all decile-wise values. At the same time, we have to conclude that KNN and SVC (ITM being exception) show very stable results in terms of variance when coupled with other uplift methods. Same conclusion can be drawn for LogR whereby it enjoys comparably high stable results across the uplift modeling techniques. In general, we would like to conclude that ITM and two model methods show the most promising results when interacting with all base learners (SGDC being exception). These methods do not show negative Q_{pct} values, relatively low variance, and comparable results among the base learners. Reflective and TCIA demonstrate opposite performance and, thus, can be regarded as worst uplift techniques involved to this study.

6.2 Examination of the impact of uplift modeling on business value

We now examine the potential of the conversion methods for uplift modeling to increase the business value. To do so, we analyze the weighted model performance for every targeting decile in terms of the cumulative (and later non-cumulative) incremental purchases. One can think of a marketing campaign similar or identical to that we describe in this paper that targets a certain fraction of customer base. The purpose of this targeting is the purchase customers perform. That is, we capture the degree uplift techniques contribute to the increase of those purchases. Again, we describe the effect of every uplift technique coupled with all base learners. Since the increased

number of the incremental purchases results in increased revenue, we argue that uplift modeling might contribute to the increase of business value. To quantify the impact of the uplift modeling techniques on business value, we first provide a tabular view of the decile-wise modeling. Table 7 presents the results obtained on the out-of-sample test set, across the uplift methods and base learners. We highlight in italic face the winner among the base learners within the uplift modeling technique and in bold face a global winner (i.e., across all uplift methods) in every decile. Consider the very left (upper) column. We contact a 10% fraction of the customer base via marketing campaign. CVT enhances RFC to achieve 883 purchases. We mark this estimate in italic face indicating that RFC is the winner within the 10% fraction across the classifiers paired with CVT. Another example (same column) is the pair of ITM and LogR. This pair achieves 3,596 purchases within the first decile and is marked in both italic and bold face indicating that LogR is the winner within the classifiers paired with ITM but also this pair (ITM and LogR) is the global winner in the first decile across the modeling techniques.

Table 7 Summary of cumulative number of incremental purchases

Uplift method / base learner	**Cumulative number of incremental purchases per decile**									
	1	2	3	4	5	6	7	8	9	10
CVT										
KNN	301	655	831	*1148*	1213	1303	1332	1423	1486	1671
LogR	819	611	712	795	1133	1352	1401	1444	1547	1671
NB	-234	278	421	213	874	1098	1281	1405	1533	1671
SGDC	66	225	422	602	754	918	1079	1300	1440	1671
SVC	-209	239	484	823	1158	*1601*	*1603*	1631	1571	1671
RFC	*883*	*983*	*1066*	1110	*1297*	1456	1597	*1641*	*1698*	1671
ITM										
KNN	418	673	868	901	1161	1381	1395	1722	1735	1671
LogR	***3596***	*4868*	*5360*	*5809*	*6116*	*6481*	*6732*	*6784*	*6194*	1671
NB	457	687	958	1564	1080	1058	1248	1445	1607	1671
SGDC	96	215	610	741	880	894	783	1751	2223	1671
SVC	395	1108	1292	1383	1602	1679	1885	1834	1838	1671
RFC	482	807	994	1043	1198	1205	1384	1341	1270	1671
LGWUM										
KNN	347	431	256	499	667	968	1146	1392	1655	1671
LogR	591	*867*	*878*	855	1127	1207	1289	1564	*1725*	1671
NB	364	373	285	11	16	728	901	746	1039	1671
SGDC	*434*	642	871	1002	1109	1139	1226	1407	1425	1671
SVC	271	229	301	609	720	1010	1030	1280	1432	1671
RFC	302	689	842	*1148*	*1501*	*1672*	*1701*	*1787*	1713	1671
LWUM										
KNN	301	655	831	*1148*	1213	1303	1332	1423	1486	1671
LogR	855	951	909	998	1143	1180	1422	1436	1543	1671
NB	-243	286	408	232	871	1097	1282	1407	1533	1671
SGDC	172	385	557	734	869	936	1150	1346	1511	1671
SVC	-208	245	474	808	1188	1602	1617	1625	1571	1671
RFC	*884*	*991*	*1071*	1103	*1294*	*1448*	*1598*	*1636*	*1694*	1671
PESSIMISTIC										
KNN	224	461	719	960	1065	1127	1261	1357	1319	1671
LogR	*935*	932	1023	934	1015	1121	1350	1547	1587	1671
NB	-148	226	325	674	867	1005	1110	1219	1127	1671
SGDC	216	425	655	788	967	1070	1124	1330	1438	1671

SVC	-149	226	486	787	1165	*1586*	*1626*	1577	1590	1671
RFC	876	*1017*	*1122*	*1271*	*1418*	1495	1581	*1581*	*1685*	1671
REFLECTIVE										
KNN	-123	-6	493	550	813	932	1140	1274	1399	1671
LogR	*403*	*821*	*970*	*1022*	1055	1236	1300	1398	1584	1671
NB	50	7	195	192	680	997	1273	998	1131	1671
SGDC	222	378	605	772	975	1076	1201	1374	1511	1671
SVC	170	257	368	444	490	816	1179	1462	1837	1671
RFC	-55	276	667	897	*1150*	*1464*	*1544*	*1500*	*1658*	1671
TCIA										
KNN	133	441	654	795	994	1220	1322	1305	1371	1671
LogR	103	60	88	-96	399	962	1171	*1573*	*1922*	1671
NB	11	229	84	64	463	710	965	1221	1838	1671
SGDC	*309*	*470*	711	843	980	1065	1235	1388	1482	1671
SVC	-17	261	*1033*	*1190*	*1281*	*1677*	*1578*	1498	1685	1671
RFC	249	423	642	802	961	1026	1085	1157	1454	1671
TWO MODEL										
KNN	321	732	*1202*	*1576*	*1730*	*1775*	*1810*	*1760*	1594	1671
LogR	*864*	*1126*	796	1002	1257	1059	1319	1503	1542	1671
NB	82	583	976	1183	1345	1364	1410	1347	1488	1671
SGDC	162	421	609	727	929	1010	1223	1378	1536	1671
SVC	-49	62	250	900	1285	1655	1785	1881	*1675*	1671
RFC	877	1111	1097	1117	1165	1273	1437	1502	1641	1671

Table 7 reveals several important findings. First, we would like to emphasize the performance of RFC another time. In particular, we observe that RFC performs quite well with multiple uplift techniques. For example, the pair CVT and RFC gets the biggest number of wins across the deciles in terms of the cumulative number of purchases. The same conclusion we can draw, e.g., for reflective and LWUM uplift methods. RFC is especially successful in the first deciles. Given this, we can only recommend RFC for the suggestion of Lo (2014) to limit the targeting to the top 10% most valuable customers. However, the success of RFC can be interrupted in the middle deciles. For example, the pair CVT and KNN compared to CVT and RFC gets 1,148 and 1,110 cumulative number of purchases, respectively. The pair CVT and SVC outperforms CVT-based RFC in the 6th and 7th deciles. Identical picture can be seen on the pessimistic approach, whereby SVC gets 1,586 and 1,626 cumulative number of purchases compared to 1,495 and 1,581 of RFC in the 6th and 7th deciles. Thus, we conclude that there might be differences in the impact on business value depending on the size of the targeted fraction of the customer base. In general, we see the bigger cumulative numbers of purchases in the middle deciles than in the first ones. To give an example, see a steady increase of cumulative purchases for the pair LWUM and SGDC from the 1st to the last decile. However, this does not indicate that targeting a bigger fraction results in a bigger cumulative number of purchases. See, for example, the pair two model and KNN in the 7th and 8th deciles (1,810 and 1,760 purchases, respectively). Therefore, our results show clearly that targeting the whole population of the customers – mail-to-all strategy according to Chickering and Heckerman (2000) – is not the best choice. Most importantly, we now are confident to identify the best

base learner - uplift modeling technique ensembles in terms of impact on business value. These pairs are CVT and RFC, ITM and SVC, LGWUM and RFC, LWUM and RFC, pessimistic and RFC, reflective and LogR, TCIA and SVC, and finally two model and KNN. They demonstrate the biggest numbers of wins within the deciles. This finding is also partially supported in terms of Qini coefficient (see 6.2). In the following, therefore, we concentrate on these (winner) pairs to identify which performs best.

To provide specific recommendations which modeling pair technique works best, we now present the uplift gain charts. These charts much resemble common gain charts. However, while the performance of models in gain charts in customer acquisition campaigns is typically illustrated by the number of purchases on the *y-axis*, uplift gain charts draft Qini curves that are by nature capable to signal incrementality. In our case this is an incremental number of purchases; a helpful indicator to support decision making in marketing practice. This implies that the number of purchases is replaced by the incremental number purchases in uplift gains charts. This number can be derived by comparing the purchase rate in the treatment with the purchase rate in the control group. In both the traditional and uplift case, the purchase indicator is a function of the fraction of people targeted from the campaign's total population, being mapped on the *x-axis* (Radcliffe, 2007). Qini curves summarize the decile-wise performance of their underlying specific uplift models. A diagonal line reflects random targeting and therefore presents a baseline for all approach-based combinations. Recall that we present the uplift gain charts only for the winner pairs identified before. We also draw the average performance line - AVG - across the winner pairs to better judge about the performance.

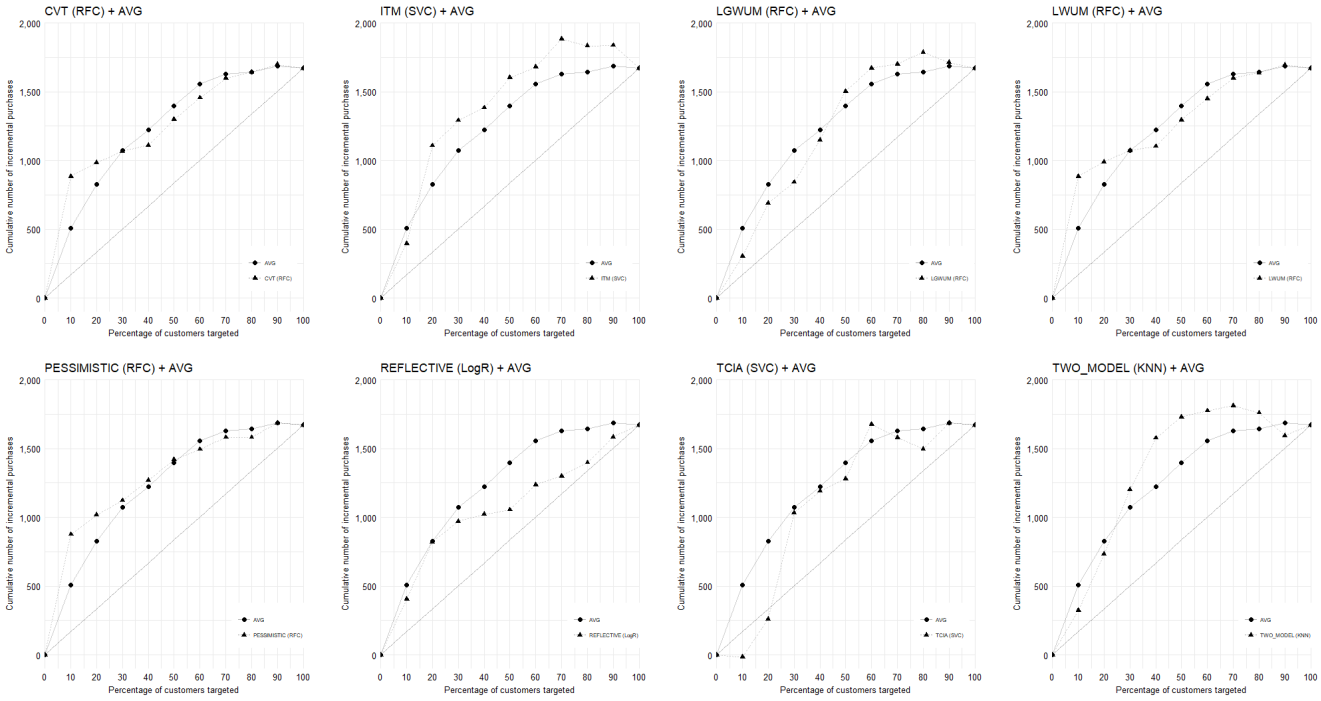

Figure 2 Uplift gain charts across the modeling techniques

Figure 2 provides new insights into the performance of the winner pairs identified before. First, we observe that all modeling pairs, even though unequally, contribute to higher cumulative number of purchases than the baseline. We see that the higher the fraction of customers targeted, the higher is the cumulative number of purchases. Every modeling pair is capable to increase that number right from the beginning. Only TCIA coupled with SVC fails to achieve that. Second, we now clearly see that ITM coupled with SVC and two model coupled with KNN outperform all other techniques. See, for example, that both couples perform better than the average performance starting from the 3rd decile. We also see that the performance of two model paired with KNN deteriorates starting from the 9th decile. This is not valid for the ITM-based SVC. However, we kindly remind that ITM-based SVC has shown extreme variance in the previous analysis (see 3.2). That is, we conclude that there are more signals in favor of the couple two model and KNN. KNN coupled with two model outperforms all other pairs (including ITM-based SVC) starting from 4th and ending with 8th deciles. Third, we regard CVT, pessimistic, LGWUM, and LWUM as the second choice since these techniques perform similar to the average. For example, pessimistic paired with RFC performs slightly better than the average in the first deciles, similar to average in the middle, and underperforms starting from the 7th decile. On the contrary, LGWUM coupled with RFC underperforms the average till the 5th decile and thereafter slightly outperforms the average. Fourth, we observe that combinations of reflective and LogR and TCIA and SVC show the weakest performance, whereby the former represents the worst choice. This is because both of them are clearly inferior to the average. This is especially relevant for the pair of reflective and LogR, since we observe the underperformance in every single decile. Thus, we cannot recommend these modeling techniques for the application in marketing campaigns similar or identical to that we describe in this paper. Given that RFC is the best choice in terms of base learners, Figure 2 suggests that it best performs coupled with pessimistic since it demonstrates till the 6th decile better or identical performance as average does; that is not given by other combinations.

To get more confidence in the findings obtained so far, we present in the next experiment the *non-cumulative* numbers of incremental purchases. More specifically, we collect non-cumulative numbers of incremental purchases based, again, on the out-of-sample test set. Figure 3 summarizes the respective results for the winner pairs on a decile-level.

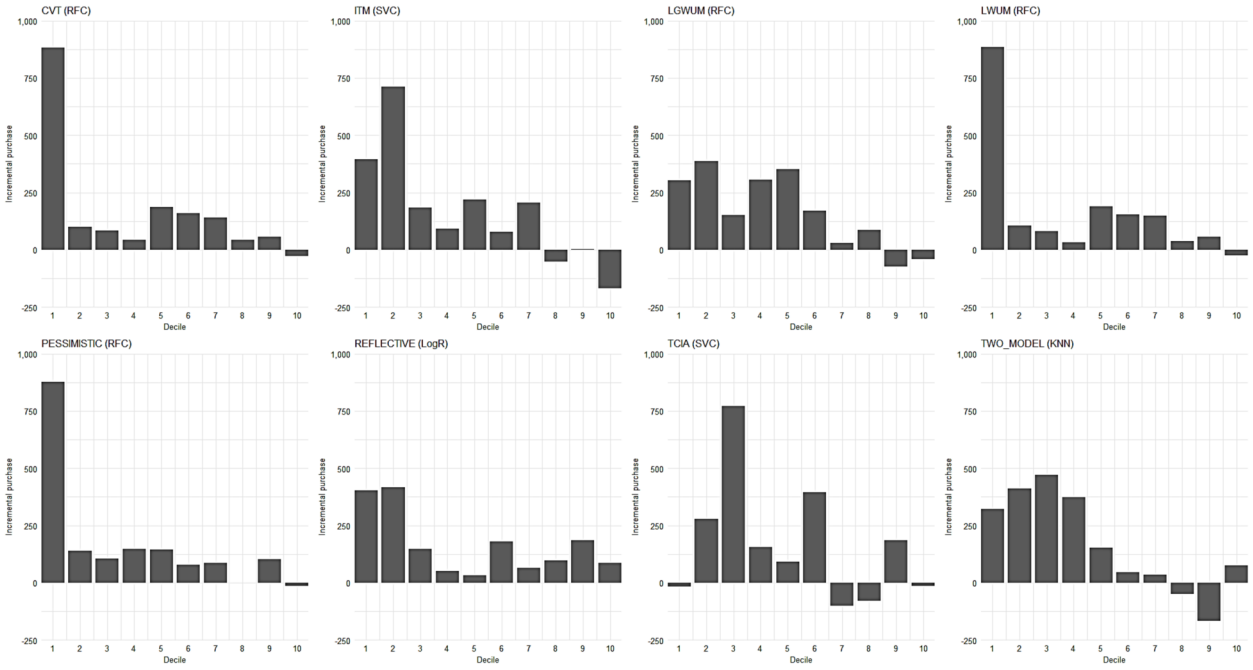

Figure 3 Non-cumulative number of incremental purchases

Figure 3 provides further findings. Given that truly valuable uplift models are capable to sort customers with high uplift to the first deciles and customers with comparably lower uplift or even negative to latter deciles (Kane et al., 2014), we first conclude that CVT, LWUM, and pessimistic perform quite well in the 1st decile. Second, comparing the winner pairs as per uplift gain charts – ITM coupled with SVC and two model coupled with KNN – we now are more confident that there are more signals in favor of the latter pair. This is because two model paired with KNN is able to assign customers who are likely to induce positive uplift to the first deciles and negative to the latter gradually. Although ITM-based SVC presents a powerful alternative achieving similar results, we observe that it assigns more customers in the latter deciles than two-model-based KNN. See, for example, the 5th, 6th, and 7th deciles. Beyond this, we observe that ITM-based SVC assigns less customers in the 4th decile than in the 5th, less in the 6th than in the 7th, indicating unstable results. Figure 3 provides more confidence in fact that the combinations TCIA and SVC as well as reflective and LogR present the worst alternatives. This is because the former allocates customers with negative uplift to the 1st decile and the latter presents a method with no negative uplift in any decile. Given that the pair of TCIA and SVC fails to allocate customers with positive uplift directly in the 1st decile and exhibits more variance in the latter deciles, we conclude that this pair presents the worst uplift modeling technique considered in this study. However, we caution from execution of both methods.

6.3 Performance comparison between response and uplift modeling

Our final experiment is devoted to the examination of the performance of response modeling, a conventional method in marketing applications, vis-à-vis *the best* – two model paired with KNN – and *the worst* – TCIA paired with SVC – uplift techniques. To provide a holistic picture on the performance of response modeling, we re-iterate all previous experiments, re-present the performance of the best and the worst uplift techniques, and extend these experiments by the estimates obtained from response modeling. To secure the fair empirical comparisons, we execute response modeling to the same out-of-sample test set for all experiments. We first examine the interaction between the modeling techniques. Recall that we involve Q_{pct} to find out the interaction synergies among the modeling techniques. Table 8 mimics the same setup for the interaction examination and presents response modeling as well.

Table 8 Qini coefficient of response modeling

Uplift method	**Base learner**					
	KNN	LogR	NB	SGDC	SVC	RFC
TCIA	1.043	-1.950	-2.821	1.222	**3.893**	0.403
TWO MODEL	**7.267**	4.305	3.297	0.688	2.806	5.401
RESPONSE	4.752	4.263	0.432	0.546	1.893	**5.679**

Table 8 shows that response approach outperforms TCIA. That is because it achieves higher Q_{pct} values for multiple base learners. See, for example, KNN, LogR, NB, or RFC. Furthermore, we observe that the highest Q_{pct} value of response coupled with RFC is bigger than that of TCIA coupled with SVC, 5.679 and 3.893, respectively. This all indicates that the *classical* response modeling might be more beneficial than modern uplift techniques. However, we also see that the response approach fails to outperform two model uplift technique. Apart from RFC, two model method is superior to response in every combination. We observe that two model interaction with KNN contributes to higher Q_{pct} value than the best combination of response, 7.267 and 5.679, respectively. We also see that response interacts best with RFC what generalizes our finding that RFC is the winner in terms of interaction with uplift techniques. On the contrary, NB and SGDC show worst results when interacting with response; finding that alerts to not execute these base learners for uplift modeling (and response modeling).

Next, Figure 4 presents the robustness procedure, aggregation of the results across the 10-fold cross validation, to judge about the variance in the results.

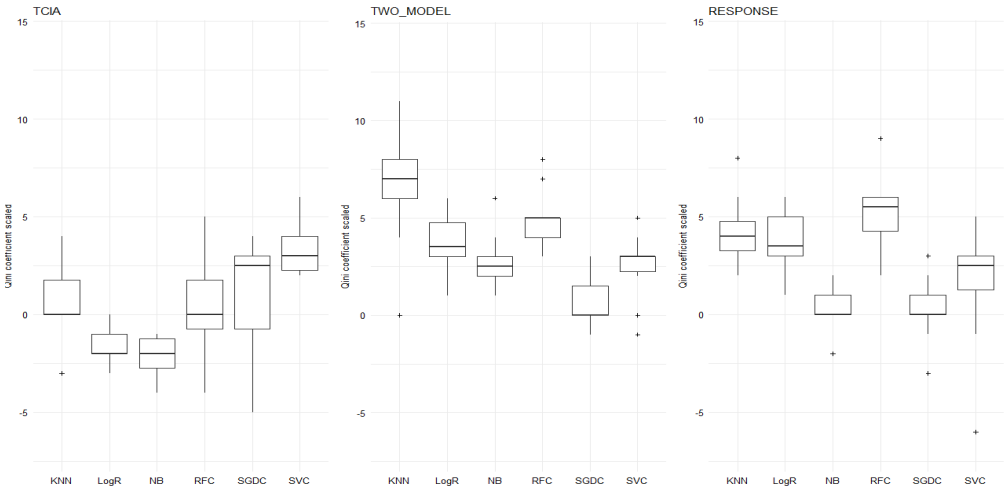

Figure 4 Scaled Qini of response modeling

Figure 4 shows another time clearly that response modeling is superior to TCIA since it exhibits smaller variance in the estimates (see, for example, RFC or SGDC) and better interacts with NB and SGDC than TCIA does. We now also see that response approach interacts with KNN and LogR comparably well to RFC and conclude that the former two base learners are promising when being paired with response. Figure 4 also confirms that response is inferior to two model method. We see that the big share of NB, SGDC, and SVC estimates shows negative scaled values for Qini, while this only the case for SGDC when paired with two model. We also observe that response interacting with SVC and RFC exhibits higher variance that two model with the same base learners.

We now examine the potential of response modeling to contribute to business value in terms of cumulative and non-cumulative incremental purchases. We echo the same experiments in 6.3 and extend these by the estimates of response modeling. First, we examine the tabular view of the cumulative number of incremental purchases. Recall that figures marked in italic and bold face indicate the same as in 6.3.

Table 9 Summary of cumulative number of incremental purchases

Uplift method / base learner	**Cumulative number of incremental purchases per decile**									
	1	2	3	4	5	6	7	8	9	10
TCIA										
KNN	133	441	654	795	994	1220	1322	1305	1371	1671
LogR	103	60	88	-96	399	962	1171	*1573*	*1922*	1671
NB	11	229	84	64	463	710	965	1221	1838	1671
SGDC	*309*	*470*	711	843	980	1065	1235	1388	1482	1671
SVC	-17	261	*1033*	*1190*	*1281*	*1677*	*1578*	1498	1685	1671
RFC	249	423	642	802	961	1026	1085	1157	1454	1671
TWO MODEL										
KNN	321	732	***1202***	***1576***	***1730***	***1775***	***1810***	***1760***	1594	1671
LogR	*864*	***1126***	796	1002	1257	1059	1319	1503	1542	1671

NB	82	583	976	1183	1345	1364	1410	1347	1488	1671
SGDC	162	421	609	727	929	1010	1223	1378	1536	1671
SVC	-49	62	250	900	1285	1655	1785	1881	*1675*	1671
RFC	877	1111	1097	1117	1165	1273	1437	1502	1641	1671
RESPONSE										
KNN	295	626	933	1179	*1417*	*1509*	*1612*	1562	1641	1671
LogR	***917***	*1014*	733	1202	1154	1368	1287	1318	1445	1671
NB	-206	214	442	623	1285	1197	1349	1358	1555	1671
SGDC	154	388	601	724	946	1004	1195	1363	1521	1671
SVC	-87	-34	406	784	1158	1586	1507	*1684*	***1813***	1671
RFC	897	853	*1199*	*1307*	1340	1393	1457	1475	1486	1671

Table 9 confirms superiority of response modeling over TCIA in terms of increase of business value. We observe that response modeling holds two global wins, i.e., in the 1st and in the 9th deciles (i.e., 917 and 1,813 cumulative incremental purchases, respectively), while TCIA none. However, response modeling is inferior to two model method, since the latter holds all global wins starting from the 2nd and ending with the 8th deciles. Table 9 also reveals that response modeling might interact successfully with (apart from RFC) LogR, KNN, and SVC. See number of wins (marked in italic face). Although the pair of response and RFC holds only two wins compared to 3 wins of the pair of response and KNN, we conclude that the former is the best choice, since this finding is previously supported by the examination of Qini coefficient and robustness procedure. Therefore, we now examine the performance of this best pair compared to the two best other pairs. Recall that TCIA performs best with SVC and two model with KNN. Figure 5 presents uplift gain charts.

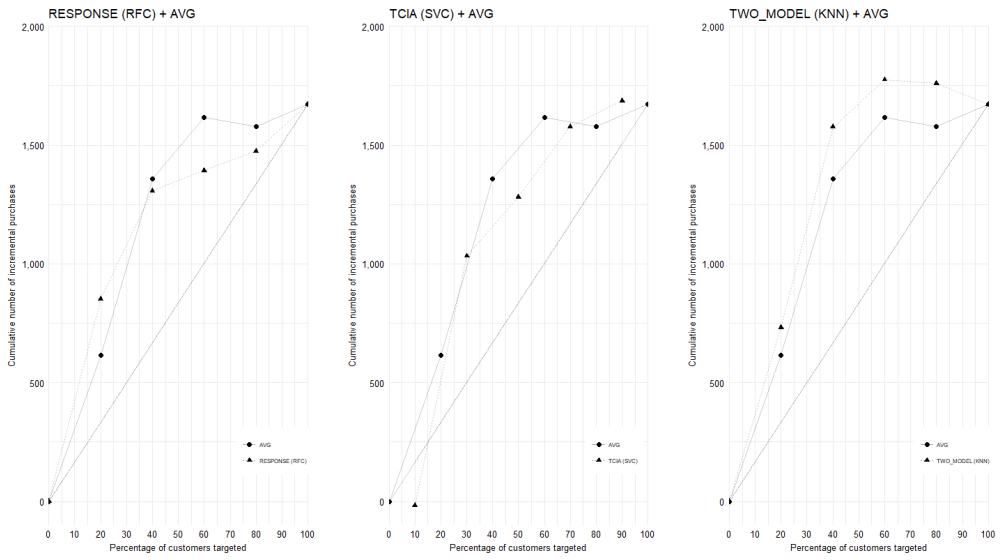

Figure 5 Uplift gain chart for response modeling

Figure 5 provides new insights. First, we see that response modeling is more successful in the first four deciles compared to the average. Recall that we now average the performance of only

these three winner pairs. However, the performance of response coupled with RFC deteriorates from the 4th decile. We that the pair TCIA and SVC outperforms response paired with RFC in the latter deciles. See, for example, the 7th, the 8th, and the 9th deciles. Figure 5, thus, indicates that TCIA-based SVC might be more beneficial when contacting a bigger fraction of customers than response-based RFC. Figure 5 also confirms that two model approach coupled with KNN is superior over response and RFC combination in every decile.

We now further examine the performance of the winning pairs as per non-cumulative number of incremental purchases. Figure 6 presents the corresponding results in box plots. We mimic again the same experimental setup as in 6.3.

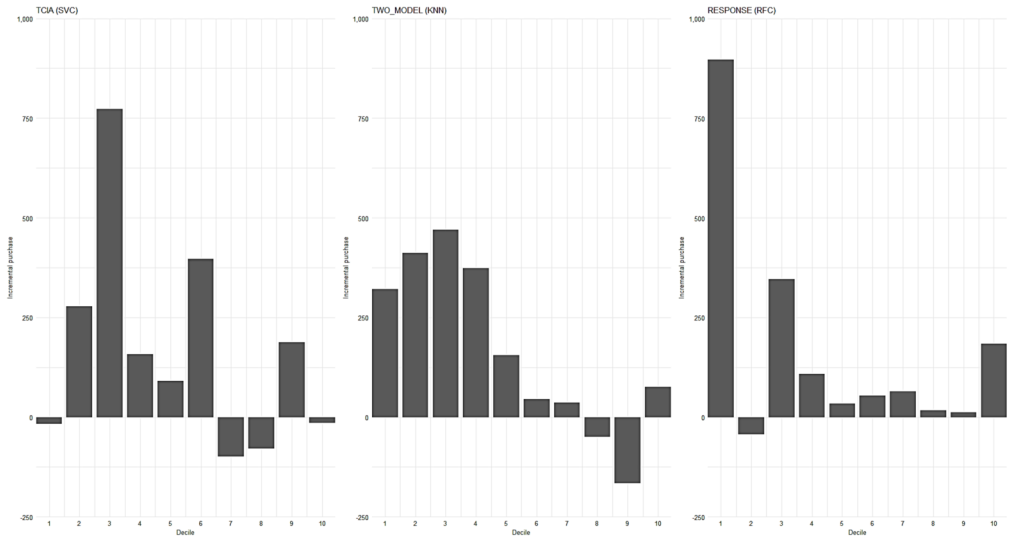

Figure 6 Non-cumulative number of incremental purchases of response modeling

Figure 6 provides new insights. First, we see that response-based RFC performs better than TCIA-based SVC in the 1st decile. However, we also observe that the former fails to perform in the 2nd decile. Recall that a good uplift technique aggregates a big number of non-cumulative purchases in the first deciles and small (or even negative) in the latter. We also see that response-based RFC fails to assign negative uplift in the latter deciles. We, therefore, conclude that response paired with RFC represents a weak alternative for uplift modeling (compared to two model coupled with KNN).

After all, we would like to conclude that 6.4 provides two important findings in general. First, response modeling that is conventionally practiced in marketing applications, e.g., (Coussement et al., 2015), represents a very powerful strategy that leads to success in such marketing campaigns that we describe in this study. We clearly see that it might easily outperform uplift techniques that

have been developed with the purpose to explain the causal relationship between marketing campaigns and an event of interest. And, second, most importantly, that the response modeling might be also inferior to uplift techniques in every experimental dimension. We, thus, conclude that our study makes it clear that marketers should be aware of the differences among the uplift techniques and apply the best choice in real-world practice.

7 Conclusion

We set out to examine how different conversion methods for the uplift modeling contribute towards increasing the fit of marketing strategies for real-world applications. Uplift modeling can be seen as a technique that patterns causal effect of a marketing incentive on customer behavior. Empirical examination goes alongside multiple dimensions and involves numerous data sets that come from different geographies and represent distinct product lines. Given that all conversion methods for uplift modeling have been proposed in different strands of literature and no attempt has been made to systematically explore and compare predictive performance of them, specific recommendations which techniques function better have been missing. This study aims to close this research gap through multi-faceted experimentation.

Our study consolidates previous work in conversion methods for uplift modeling and provides a holistic picture of the state-of-the-art in predictive modeling for retail electronic commerce; more specifically, personalized marketing targeting through couponing. From an academic viewpoint, an important question is whether efforts invested to the development of novel uplift techniques are worthwhile. Our study raises some critical concerns. We find the proposed method to generalize LWUM with weighted probability scores to account for the fraction of treatment and control group customers by Kane et al. (2014) fails to outperform the original LWUM developed by Lai (2006) in terms of Qini coefficient. Similar picture is obtained in the field of covariates manipulation. We find that TCIA method proposed by Tian et al. (2014) and that pretty much mimics the procedure of ITM with a slight adaption is inferior to the original ITM approach developed by Lo (2002). On the contrary, we find that ITM as well as the straightforward two model approach developed by Rzepakowski & Jaroszewicz (2012a) that captures differences in class probabilities of customers' motivation represent techniques of first choice for uplift modeling. Our study, therefore, implies that the progress has stalled and efforts invested to the methodological advancement must be accompanied by a rigorous assessment of new techniques vis-à-vis challenging benchmark. We find two model and ITM approaches which to compare novel conversion methods in the field of uplift modeling.

Another important question to answer in future research is the explanation behind interaction between uplift techniques and underlying base learners. We have identified the base learners that work well specifically for uplift modeling. However, our study does not seek to explain their success. We strongly believe this is a very fruitful avenue for future research. Nonetheless, our study can be regarded as a first move toward gaining insights to this question. For example, we find RFC to interact best with the majority of techniques that is not given by other base learners. Moreover, RFC performs quite well in the first deciles of targeting and, therefore, can be strongly recommended for limitation to the top 10% most valuable customers (Lo, 2014). We find SVC is a valid alternative, although it exhibits high variance in estimates as per robustness procedure presented in this paper. Surprisingly, KNN, a technique that is usually seen as a weak in predictive modeling, has shown appealing results, especially interacting with two model conversion method. On the contrary, SGDC and NB have shown poor results in every experiment, we, therefore, forewarn from considering these techniques for uplift modeling.

From a practitioner viewpoint, it is important to reason whether the results and findings in this study can be generalized to real world applications. However, we believe that numerous data sets coming from online shops, several cross-validation repetitions, and performance examination from different perspectives highly approximate real environment of the industry. We have conducted several experiments to judge about the increase of business value that we measure in cumulative (non-cumulative) number of incremental purchases. Ideally, the experiments should measure the business value in terms of revenue and profits. Lacking this information, we emphasize that our results do not guarantee external validity. Finally, examination of response vs. uplift modeling has clearly shown that the former can be superior to the latter. Therefore, we caution from executing such uplift techniques like TCIA. Practitioners should be aware of the fact that the wrong choice of uplift technique might result in poor performance of marketing campaigns and consider best performers, like two model accompanied with KNN, for their applications.

References

Ahmed, A. A. Q., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly particle swarm optimization algorithm based classification. *IOSR Journal of Computer Engineering, 19 (4),* 30-39.

Ascarza E., Fader P.S., & Hardie B.G.S. (2017). Marketing models for the customer-centric firm. In: *Wierenga B., van der Lans R. (eds) Handbook of Marketing Decision Models. International Series in Operations Research & Management Science, 254, Springer, Cham.*

Azeem, M., Usman, M., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. *Telecommunication Systems, doi.org/10.1007/s11235-017-0310-7,* 1-12.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54 (6),* 627–635.

Bailey, J., & Bakos, Y. (1997). An exploratory study of the emerging role of electronic intermediaries. *International Journal of Electronic Commerce, 1 (3),* 7-20.

Bakos, Y. (1998). The emerging role of electronic marketplaces on the Internet. *Communications of the ACM, 41 (8),* 35–42.

Baye, M., Gatii, R., Kattuman, P., & Morgan, J. (2009). Clicks, discontinuities, and firm demand online. *Journal of Economics and Management Strategy, 18 (4),* 935-975.

Blattberg, R. C., Getz, G., & Thomas, J. S. (2001). Customer Equity: Building and Managing Relationships As Valuable Assets. Harvard Business School Press, chapters 3-5.

Breiman, L. (2001). Random forests. Machine learning, 45(1), pp. 5-32.

Brynjolfsson, E., Dick, A. A., & Smith, M. D. (2010). A nearly perfect market? *Quantitative Marketing and Economics, 8 (1),* 1–33.

Cai, T., Tian, L., Wong, P. H., & Wei, L. J. (2009). Analysis of randomized comparative clinical trial data for personalized treatment selections. Harvard University, Biostatistics Working Paper Series. The Berkeley Electronic Press.

Chickering, D. M., & Heckerman, D. (June 2000). A decision theoretic approach to targeted advertising. Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence, pp. 82-88.

Choi, J., & Bell, D. (2011). Preference minorities and the internet. *Journal of Marketing Research, 48 (4),* 670-682.

Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Systems with Applications, 42 (22),* 8403-8412.

Daskalova, N., Bentley, F., & Andalibi, N. (2017). It's all about coupons: Exploring coupon use behaviors in email. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems,* 1152-1160.

Del Giudice, M., & Peruta M. R. D. (2017). A model of customer retention in business-intensive markets. In: *The Satisfaction of Change. Palgrave Studies in Democracy, Innovation, and Entrepreneurship for Growth. Palgrave Macmillan, Cham.*

Ding, A. W., Li, S., & Chatterjee, P. (2015). Learning User Real-Time Intent for Optimal Dynamic Web Page Transformation. Information Systems Research, 26(2), pp. 339-359.

Doorn, J., Onrust, M., Verhoef, P. C., & Bügel, M. S. (2017). The impact of corporate social responsibility on customer attitudes and retention - the moderating role of brand success indicators. *Marketing Letters, doi.org/10.1007/s11002-017-9433-6,* 1-13.

Dost, F., Wilken, R., Eisenbeiss, M., & Skiera, B. (2014). On the Edge of Buying: A Targeting Approach for Indecisive Buyers Based on Willingness-to-Pay Ranges. Journal of Retailing, 90(3), pp. 393-407.

Escobar-Rodríguez, T., & Carvajal-Trujillo, E. (2013). Online drivers of consumer purchase of website airline tickets. *Journal of Air Transport Management, 32,* 58–64.

Forman, C., Ghose, A., & Goldfarb, A. (2009). Competition between local and electronic markets: how the benefit of buying online depends on where you live. *Management Science, 55 (1),* 47-57.

Ghose, A., Goldfarb, A., Han, S. P. (2011). How is the mobile internet different? Search costs and local activities. *Information Systems Research, 24 (3),* 613 – 631.

Guelman, L. (2014). Optimal personalized treatment learning models with insurance applications. Doctoral dissertation. Universitat de Barcelona.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2012). Random forests for uplift modeling: an insurance customer retention case. Modeling and Simulation in Engineering, Economics and Management, pp. 123-133.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2014). Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study. (No. 2014-06).

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015). Uplift random forests. Cybernetics and Systems, 46(3-4), pp. 230-248.

Gui, C. (2017). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. *Artificial Intelligence Research, 6 (2),* 93-99.

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. Statistical Science, 23(2), pp. 219-236.

Hansotia, B., & Rukstales, B. (2002a). Incremental value modeling. Journal of Interactive Marketing, 16(3), pp. 35-46.

Hansotia, B., & Rukstales, B. (2002b). Direct marketing for multichannel retailers: Issues, challenges and solutions. Journal of Database Marketing & Customer Strategy Management, 9(3), pp. 259-266.

Holland, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81(396), pp. 945-960.

Hua, S. (2016). What Makes Underwriting and Non-Underwriting Clients of Brokerage Firms Receive Different Recommendations? An Application of Uplift Random Forest Model. International Journal of Finance & Banking Studies (2147-4486), 5(3), pp. 42-56.

Huang, E. Y., & Tsui, C.-J. (2016). Assessing customer retention in B2C electronic commerce: an empirical study. *Journal of Marketing Analytics, 4 (4),* 172-185.

Ieva, M., De Canio, F., & Ziliani , C. (2017). Daily deal shoppers: What drives social couponing? *Journal of Retailing and Consumer Services, https://doi.org/10.1016/j.jretconser.2017.03.005*

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics, 7(1), pp. 443-470.

Jackson, R., & Wang, P. (1996). Strategic Database Marketing. Chicago: NTC Publishing.

Jagtap, S. S., & Hanchate, D. B. (2017). Development of android based mobile app for Presta Shop eCommerce shopping cart (ALC). *International Research Journal of Engineering and Technology (IRJET), 5 (3),* 43-47.

Jaroszewicz, S., & Rzepakowski, P. (2014). Uplift modeling with survival data.

Jaroszewicz, S., & Zaniewicz, Ł. (2016). Székely Regularization for Uplift Modeling. In Challenges in Computational Statistics and Data Mining (pp. 135-154). Springer International Publishing.

Jaskowski, M., & Jaroszewicz, S. (2012). Uplift modeling for clinical trial data. ICML 2012 Workshop on Clinical Data Analysis.

Kane, K., Lo, V. S., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. Journal of Marketing Analytics, 2(4), pp. 218-238.

Kang, J.-Y. M., & Kim, J. (2017). Online customer relationship marketing tactics through social media and perceived customer retention orientation of the green retailer. *Journal of Fashion Marketing and Management, 21 (3),* 298-316.

Kondareddy, S. P., Agrawal S., & Shekhar S. (2016). Incremental response modeling based on segmentation approach using uplift decision trees. In: *Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects. ICDM 2016. Lecture Notes in Computer Science, vol 9728. Springer, Cham.*

Kotler, P. (1994). Marketing Managment, 8th edition, chapter 24. Prentice-Hall.

Kuusisto, F., Costa, V. S., Nassif, H., Burnside, E., Page, D., & Shavlik, J. (2014). Support vector machines for differential prediction. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 50-65). Berlin Heidelberg: Springer.

Lai, L. T. (2006). Influential marketing: A new direct marketing strategy addressing the existence of voluntary buyers. Master of Science thesis, Simon Fraser University School of Computing Science, Burnaby, BC, Canada.

Lakshminarayan, C, Kosuru, R., & Hsu, M. (2016). Modeling complex clickstream data by stochastic models: Theory and methods. *Proceedings of the 25th International Conference Companion on World Wide Web,* 879-884.

Larsen, K. (2010). Net Lift Models. Slides of a talk given at the A2010 - Analytics Conference, September 2-3, Copenhagen, Denmark.

Lee, J., Park, D.-H., & Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic Commerce Research and Applications, 7 (3),* 341-352.

Lessmann, S., Coussement, K., & De Bock, K. W. (2013). Maximize What Matters: Predicting Customer Churn With Decision-Centric Ensemble Selection.

Lessmann, S., Seow, H.-V., Baesens, B., & Thomas, C. L. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. *European Journal of Operational Research, 247 (1),* 124–136.

Lo, V. S. (2002). The true lift model: a novel data mining approach to response modeling in database marketing. ACM SIGKDD Explorations Newsletter, 4(2), pp. 78-86.

Lo, V. S., & Pachamanova, D. A. (2015). From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk. Journal of Marketing Analytics, 2, pp. 79-95.

Manahan, C. (2005). A proportional hazards approach to campaign list selection. SAS User Group International (SUGI) 30 Proceedings.

Michaelidou, N., & Dibb, S. (2006). Using email questionnaires for research: Good practice in tackling non-response. *Journal of Targeting Measurement and Analysis for Marketing, 14 (4),* 289-296.

Michel, R., Schnakenburg, I., & Martens, T. (2017). Effective customer selection for marketing campaigns based on net scores. *Journal of Research in Interactive Marketing, 11 (1),* 2-15.

Mlelwa, K., & Yonah, Z. O. (2017). A novel framework for secure e-commerce transactions. *International Journal of Cyber-Security and Digital Forensics, 6 (2),* 92-100.

Nassif, H., Kuusisto, F., Burnside, E. S., & Shavlik, J. W. (2013). Uplift Modeling with ROC: An SRL Case Study. Proceedings of the International Conference on Inductive Logic Programming (ILP'13), Rio de Janeiro, Brazil, pp. 40-45.

Nassif, H., Kuusisto, F., Burnside, E. S., Page, D., Shavlik, J. W., & Costa, V. S. (2013, September). Score as you lift (SAYL): A statistical relational learning approach to uplift modeling. In Joint European conference on machine learning and knowledge discovery in databases (pp. 595-611). Berlin Heidelberg: Springer.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd international conference on Machine learning (pp. 625-632). ACM.

Nottorf, F. (2014). Multi-channel attribution modeling on user journeys. *Communications in Computer and Information Science, 456,* 107-125.

Ogutu, J. O., Piepho, H. P., & Schulz-Streeck, T. (2011, May). A comparison of random forests, boosting and support vector machines for genomic selection. In BMC Proceedings (Vol. 5, No. 3, p. S11). BioMed Central.

Park, C. H. (2017). Online purchases paths and conversion dynamics across multiple websites. *Journal of Retailing, 93 (3),* 253-265.

Radcliffe, N. J. (2007). Using control groups to target on predicted lift: Building and assessing uplift models. Direct Market J Direct Market Assoc Anal Council(1), pp. 14-21.

Radcliffe, N. J., & Surry, P. D. (1999). Differential response analysis: Modeling true response by isolating the effect of a single action. Credit Scoring and Credit Control VI.

Radcliffe, N. J., & Surry, P. D. (2011). Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions.

Rzepakowski, P., & Jaroszewicz, S. (2012a). Uplift modeling in direct marketing. Journal of Telecommunications and Information Technology, pp. 43-50.

Rzepakowski, P., & Jaroszewicz, S. (2012b). Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems, 32(2), pp. 303-327.

Shaar, A., Abdessalem, T., & Segard, O. (2016). Pessimistic Uplift Modeling. ACM SIGKDD ’16 August-2016 San Francisco, California USA.

Siegel, E. (2011). Uplift Modeling: Predictive Analytics Can’t Optimize Marketing Decisions Without It. Prediction Impact white paper sponsored by Pitney Bowes Business Insight.

Simarmata, J., & Ikhsan, R. B. (2017). Building customer retention in on-line transportation, *Polish Journal of Management Studies, 15 (2),* 229-239.

Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modeling. Data mining and knowledge discovery, 29(6), pp. 1531-1559.

Su, X., Kang, J., Fan, J., Levine, R. A., & Yan, X. (2012). Facilitating score and causal inference trees for large observational studies. Journal of Machine Learning Research, 13(Oct), pp. 2955-2994.

Tambe, P. (2014). Big data investment, skills, and firm value. Management Science, 60(6), pp. 1452-1469.

Thomas, A. (2017). Multivariate hybrid pathways for creating exceptional customer experiences. *Business Process Management Journal, 23 (4),* 822-829.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. Journal of the American Statistical Association, 109(508), pp. 1517-1532.

Van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research, 166,* 557-575.

Yong, F. (2015). Quantitative Methods for Stratified Medicine. Doctoral dissertation. Harvard University, Graduate School of Arts & Sciences.

Zaniewicz, Ł., & Jaroszewicz, S. (December 2013). Support vector machines for uplift modeling. Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference, pp. 13

Zantedeschi, D., Feit, E. M., & Bradlow, E. (2016). Measuring multi-channel advertising response. *Plant Disease Management Reports, 63 (8),* 2706 – 2728.