Master-Thesis-Projekt

Vorhersage des Upselling-Potenzials von Kundinnen und Kunden der BLS AG mittels Supervised Machine Learning

Autorin	Zeliya Schär Bahnhofstrasse 16 6130 Willisau zeliya.schaer@hslu.ch
Referat	Hochschule Luzern Institut für Kommunikation & Marketing IKM Prof. Dr. Ingo Gächter Zentralstrasse 9 6002 Luzern ingo.gaechter@hslu.ch
Auftraggeber / Korreferat	BLS AG Jonas Bachmann Genfergasse 11 3001 Bern jonas.bachmann@bls.ch

Hochschule Luzern - Wirtschaft
Master of Science in Applied Information and Data Science (MScIDS)
Frühlingssemester 2023

Abgabedatum: 2. Juni 2023

Management Summary

Upselling bringt sowohl für Unternehmen als auch für Kundinnen und Kunden zahlreiche Vorteile mit sich. Zum einen kann das Unternehmen durch gezieltes und individuelles Upselling seinen Umsatz steigern und die Kundenzufriedenheit sowie -bindung erhöhen. Die Kundschaft hingegen profitiert von höherwertigen Dienstleistungen und einem besseren Angebot.

Bei der BLS AG, welche Auftraggeberin der vorliegenden Arbeit ist, stehen seit rund zwei Jahren NOVA-Transaktionsdaten zur Verfügung, welche bisher nicht näher analysiert wurden. Die Auftraggeberin hat sich jedoch zum Ziel gesetzt, den Umsatz durch den Verkauf von mehr 1. Klasse- und Klassenwechseltickets zu steigern. Diese Umsatzsteigerung soll bei bestehenden Kundinnen und Kunden durch gezieltes und personalisiertes Upselling erreicht werden. Aus diesem Grund möchte die BLS AG die vorhandenen Transaktionsdaten analysieren und Ergebnisse für das Marketing ableiten.

Für die vorliegende Arbeit wurden von der BLS AG anonymisierte Transaktionsdaten der Personenmobilität von 2020 bis 2022 zur Verfügung gestellt. Der Fokus der vorliegenden Thesis liegt auf der Analyse der Daten sowie der Vorhersage des Potenzials für ein Upselling in die 1. Klasse bzw. für den Kauf eines Klassenwechseltickets aufgrund der vergangenen Transaktionen einer Person. Darüber hinaus soll ermittelt werden, welche Merkmale das Upselling-Potenzial besonders beeinflussen. Aus der definierten Zielsetzung der Auftraggeberin wird für die Arbeit die folgende Forschungsfrage abgeleitet:

Mit welcher Genauigkeit kann das Upselling-Potenzial von in der Schweiz wohnhaften Kundinnen und Kunden der BLS AG aufgrund von internen Transaktionsdaten der Jahre 2020 bis 2022 mit Supervised Machine Learning vorhergesagt werden und durch welche Faktoren wird dieses Potenzial am stärksten beeinflusst?

Zudem wurden für die Arbeit die folgenden Ziele definiert:

- Die Daten sollen in einer ausführlichen explorativen Datenanalyse (EDA) analysiert und visualisiert werden.
- Es sollen unterschiedliche Supervised-Machine-Learning-Algorithmen angewendet und validiert werden.
- Aufgrund der Ergebnisse der unterschiedlichen Machine Learning (ML)-Modelle soll eruiert werden, welche Features für die Vorhersage des Potenzials von zentraler Bedeutung sind.

Methodisch folgt die Arbeit dem Modell Cross Industry Standard Process for Data Mining (CRISP-DM), welches branchen- und applikationsneutral ist und somit für verschiedenste Data-Science-Projekte eingesetzt werden kann. Das Modell gliedert sich in die sechs Phasen Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation und Deployment, wobei der letzte Schritt nicht mehr Bestandteil dieser Arbeit ist. Im Schritt Business Understanding geht es darum, die Ausgangssituation sowie die Ziele des Projektes zu definieren. Darauf folgen im Schritt Data Preparation die Evaluation und Definition der Datenquellen, die im anschliessenden Modelling verwendet werden sollen. Bei der Evaluation geht es darum, das generierte Modell mit den zu Beginn definierten Zielen und Erfolgskriterien abzugleichen.

Im vorliegenden Projekt wird mit Supervised Machine Learning gearbeitet, da die generierten Personenprofile aufgrund der getätigten Transaktionen im Vorfeld gelabelt werden. Der Fokus liegt dabei auf dem prädikativen Modellieren, da aufgrund der definierten Forschungsfrage eruiert werden soll, mit welcher Genauigkeit das Upselling-Potenzial von Kunden und Kundinnen vorhergesagt werden kann. Jedoch soll auch eruiert werden, welche Personenmerkmale die beträchtlichste Relevanz für die Vorhersage aufweisen. Da es sich im vorliegenden Projekt um ein unausgeglichenes Datenset handelt, kann die Performance der Modelle

nicht nur anhand der Genauigkeit verglichen werden, weshalb für den Vergleich weitere Kennzahlen aus der Confusion Matrix wie Präzision, Recall, F1-Score, Kappa und AUC genutzt werden. Im vorliegenden Projekt liegt der Fokus auf den positiven Vorhersagen (True und False Positives), da diese Menschen vom Klassifikator als Personen mit Upselling-Potenzial klassiert werden und somit ein Subset einer möglichen Marketingkampagne bilden.

Die folgenden Erkenntnisse aus dem Training können präsentiert werden: Die Modelle, welche mit den balancierten Daten trainiert wurden, weisen eine tiefere Genauigkeit auf als die Modelle, welche mit nicht balancierten Daten trainiert wurden. Dies liegt daran, dass Letztere lernen, die Mehrheitsklasse vorherzusagen, und daher das Verhältnis aus den Trainingsdaten repräsentieren. Da insgesamt weniger Personen positiv klassiert werden, äussert sich dies in einer hohen Präzision. Der Recall hingegen ist eher tief, da das Modell zahlreiche Fälle als falsch negativ klassiert und somit der Anteil der korrekten positiven Vorhersagen an allen positiven Referenzwerten gering ist. Bei den Modellen, welche mit den ausbalancierten Daten trainiert wurden, zeigt sich das umgekehrte Verhalten. Da die Minderheitsklasse in den Trainingsdaten überproportional vorhanden war, werden insgesamt mehr positive Fälle klassiert, was dazu führt, dass die Präzision tiefer ist, da der Anteil der korrekten positiven Vorhersagen an allen positiven Vorhersagen kleiner ist. Es werden somit weniger Fälle als falsch negativ klassiert, womit der Anteil der korrekten positiven Vorhersagen an allen positiven Referenzwerten und somit der Recall hoch sind. Diese Modelle klassieren also mehr Fälle als falsch positiv, was im vorliegenden Projekt jedoch potenzielle Upselling-Chancen sein können.

Weiter kann festgehalten werden, dass Random Forest mit allen 81 Features auf ausbalancierten und nicht balancierten Daten über alle relevanten Kennzahlen am besten performt und die höchste Vorhersagegenauigkeit erzielt. Das mit den ausbalancierten Daten trainierte Modell erreicht eine Genauigkeit von 90 % und einen F1-Score von 0.34. Zudem liegt der Recall bei 88.2 %. Dies bedeutet, dass mehr als drei Viertel der Personen, welche bereits Upselling gemacht haben, vom Modell korrekt erkannt werden. Das Modell klassiert zahlreiche Personen als False Positive, womit die Präzision bei 21 % liegt. Das heisst, jede fünfte Person, welche als Upseller vorhergesagt wird, hat effektiv bereits ein Upselling gemacht. Das bedeutet im Umkehrschluss, dass 80 % der positiv vorhergesagten Menschen Potenzial haben könnten und daher als potenzielle Chancen betrachtet werden können.

Das Modell, welches mit den nicht balancierten Daten trainiert wurde, erreicht eine Genauigkeit von 97.9 % und einen F1-Score von 0.52. Es klassiert die Mehrheitsklasse besser und erzielt daher eine hohe False-Negative-Rate und somit einen tiefen Recall. Das heisst, dieses Modell erkennt 38 % der Personen, welche bereits ein Upselling gemacht haben, klassiert dafür weniger Personen als False Positive, womit die Präzision bei 82 % liegt. Im Umkehrschluss bedeutet dies, dass lediglich 18 % der positiven Vorhersagen potenzielle Chancen sein könnten.

Bezüglich der Relevanz der Prädikatoren für die Vorhersagen wurden die vier Features Alter, Anzahl Tickets, Fahrausweis Einzelbillett sowie Bruttolohn in allen elf trainierten Modellen als Top-20-Feature genannt. Die Entscheidungsbäume lassen zudem eine weiterführende Interpretation zu. So nimmt die Wahrscheinlichkeit für ein Upselling zu, wenn eine Person mehr als ein Ticket gekauft hat. Weiter führen der Kauf von Einzeltickets, sowie das Alter über 28 Jahren zu einer höheren Wahrscheinlichkeit für ein Upselling. Betrachtet man zusätzlich die Koeffizienten der logistischen Regression, kann ergänzt werden, dass der Besitz eines Abonnements die Wahrscheinlichkeit für ein Upselling ebenfalls steigert. Aufgrund der Feature Importance der Modelle können die folgenden drei Hypothesen bestätigt werden.

- Ältere Personen tendieren eher zu einem Upselling als jüngere Personen.
- Personen, welche viele Transaktionen tätigen, haben eher das Potenzial für ein Upselling.
- Personen mit einem höheren Einkommen tendieren eher zu Upselling.

Aufgrund der genannten Ergebnisse kann die folgende Schlussfolgerung gezogen werden: Welches Modell für die Vorhersage des Upselling-Potenzials angewendet wird, hängt vom Ziel des Unternehmens ab. Da im vorliegenden Projekt die personalisierte Ansprache der Kunden und Kundinnen, welche ein Potenzial für Upselling in die 1. Klasse aufweisen, erreicht werden soll, liegt der Fokus auf den positiven Vorhersagen. Daher sind eine hohe Präzision sowie ein hoher Recall wünschenswert, was in einem hohen F1-Score resultiert. Da die trainierten Modelle unterscheiden, ob eine Person bereits ein Upselling oder ein 1. Klasse-Ticket gekauft hat oder nicht, sind Menschen, welche als False Positive klassiert werden, nicht zwingend falsch klassiert, sondern können als Upselling-Chancen und somit als Personen, welche aufgrund ihrer Merkmale eventuell 1. Klasse- oder Klassenwechseltickets kaufen könnten, betrachtet werden. Daraus resultiert, dass vor allem ein hoher Recall bedeutsam ist.

Wird davon ausgegangen, dass die BLS AG zurzeit kein Modell nutzt, um Personen zu selektionieren, und daher nicht weiss, welche Kundinnen und Kunden Potenzial für ein Upselling haben, würde bei einer geplanten Marketingmassnahme die gesamte Kundschaft als Zielgruppe einer Kampagne in Frage kommen. Durch den Einsatz eines Modells können die Kosten für die Kundenansprache gesenkt werden, da eine spezifische Zielgruppe selektioniert wird und somit eine höhere Konversationsrate erreicht werden kann. Ausserdem müssen nicht wahllos Kunden und Kundinnen angeschrieben werden, welche unter Umständen nicht an einem Angebot interessiert sind, was auch aus Sicht der Kundschaft besser ist.

Wie anhand der Daten der BLS AG ersichtlich ist, sind die Ausgaben in der 1. Klasse und für Klassenwechsel-Tickets über die letzten drei Jahre im Durchschnitt um CHF 280 pro Jahr höher als die in der 2. Klasse. Wird somit davon ausgegangen, dass pro Person, welche ein Upselling macht, Mehreinnahmen von CHF 280 generiert werden können und die Kosten für die Kundenansprache unter diesem Wert liegen, ist es sinnvoll, möglichst viele Personen zu kontaktieren und somit einen hohen Anteil an Personen mit potenziellem Upselling-Potenzial zu selektieren. Dies wird mit dem Random Forest Modell, welches mittels Upsampling ausbalanciert wurde, erreicht. Der nächste Schritt wäre nun ein Live-Test mit diesem Modell, um dessen Tauglichkeit im Marketing zu validieren und zu überprüfen, ob jene Personen, welche vom Modell als potenzielle Upseller klassiert werden, auch effektiv zu einem Kauf eines 1. Klasse- oder Klassenwechseltickets motiviert werden können.