Projekt-Bericht und Dokumentation im Kurs Ausgewählte Kapitel sozialer Webtechnologien (SoSe15) Machine Learning

Claus Holland Richard Remus

25. Juli 2015

Contents

1 Einführung

Als wichtiger Teil des Wirtschaftslebens in einer Marktwirtschaft ist es die Aufgabe von Banken Kredite zu vergeben. Unternehmen können so zum Beispiel Investitionen finanzieren, wachsen und mit den hoffentlich erzielten Gewinnen ihren Kredit tilgen und wieder zurückzahlen.

Doch auch Privatbürger nehmen Kredite auf, um sich etwa Wohnmobilen zu kaufen oder andere Anschaffungen zu tätigen. Wohneigentum kann wiederum mit Hypotheken belastet werden, dem Abtreten von Rechten an einer Immobilie im Gegenzug für ein Darlehen. Unternehmen haben in der Regel Kapital in Form von Anlagen oder Ähnlichem. Gerade wenn eine Privatperson aber nicht mit viel Besitz für einen Kredit haften kann, stehen Banken vor einer schwierigen Aufgabe. Sie müssen bei Abschluss eines Kreditvertrages entscheiden, wie zuverlässig der Kreditnehmer in der Zukunft sein wird. Auf Basis dieser Entscheidung bestimmen sie die Konditionen des Kredites oder entscheiden sich bei einem womöglich sehr hohen Ausfallrisiko sogar gegen eine Kreditvergabe. Damit entgeht ihnen aber womöglich ein einträgliches Geschäft.

Mit den Methoden des Maschinellen Lernens stehen Werkzeuge zur Verfügung, um mit beschränkten Aufwand verhältnismäßig sichere Aussagen zu treffen, die einen potenziellen Bankkunden anhand bestimmter Eigenschaften, sogenannter Features, hinsichtlich seines Kreditzahlungs-Ausfallrisikos beurteilen helfen. Dies soll im Rahmen dieser Arbeit aufgezeigt werden.

Es liegen uns 150.000 Datensätze von Bankkunden vor. Diese umfassen neben verschiedenen finanziellen und demografischen Angaben auch Werte zum Kredit-Ausfallverhalten. Es handelt sich somit um Trainingsdaten im Sinne des Maschinellen Lernens. Unsere Aufgabe besteht daher darin, Klassifikatoren zu implementieren, die anhand der gegebenen Werte am zuverlässigsten, also mit einer möglichst hohen Trefferquote, das Ausfallrisiko für die Datensätze korrekt vorhersagen.

Dafür analysieren wir den Datensatz erst einmal hinsichtlich bestimmter Auffälligkeiten und auch bezüglich der vorhandenen Datenqualität, führen also die sogenannte explorative Datenanalyse durch. Aus ihr ergeben sich idealerweise bereits erste Erkenntnisse, welche der Features besser als andere geeignet erschienen, um mit dem Ergebnis “Guter Kreditnehmer” oder “Risikoreicher Kreditnehmer” zu korrelieren. Anschließend wird auf Basis des ersten Punkts von uns der Datensatz optimiert. Ermittelte Schwachstellen werden wenn möglich kompensiert und Skalierungen so gewählt, dass bildliche Darstellungen, die sogenannten Plots eine möglichst hohe Aussagekraft beinhalten. Darauf folgt die Klassifikationsphase. In ihr werden wir geeignete Klassifikationsalgorithmen ermitteln und zur Anwendung bringen.

Ergebnis soll ein Klassifikator sein, der den gewünschten Anforderungen, also einer hohen Trefferquote bei den Vorhersagen entspricht. Eine Evaluation und ein Fazit schließen sich an.

Der Datensatz entstammt der Web-Platform "www.kaggle.com" und dort dem bereits beendeten Wettbewerb "Give me some Credit" aus dem Jahr 2011. Ziel dieses Belegs ist es allerdings nicht, die Wettbewerbskriterien auf kaggle zu erfüllen. Uns geht es darum, die Mächtigkeit der Werkzeuge des Machinellen Lernens zu demonstrieren, um realwirtschaftliche Probleme zu erfassen. Da uns auf kaggle keine geeigneten Testdaten zur Verfügung stehen, in denen ebenfalls das Kreditausfallrisiko in einer Form gegeben ist, dass die erfolgreiche Anwendung eines geeigneten Klassifikators bewiesen werden kann, nutzen wir stattdessen randomisierte Datensätze aus den Trainingsdaten zur Verifizierung unserer Arbeit.

2 Explorative Datenanalyse

In den vorliegenden Daten finden primär zehn Features Verwendung, die die Bankkunden erfassen helfen und schließlich analysierbar machen:

RevolvingUtilizationOfUnsecuredLines Dieser Prozentwert gibt an, inwiefern eine Person ihre Kreditkarte oder ihren Kreditrahmen tatsächlich in Anspruch nimmt, geteilt durch die Summe der Kreditlimits. Immobilien- und Sachkredite werden dabei nicht berücksichtigt.

age Das Alter der Bankkunden in Jahren, ein ganzzahliger Wert.

NumberOfTime30-59DaysPastDueNotWorse Ein ganzzahliger Wert, der angibt, wie oft ein Bankkunde innerhalb der zwei vorangegangenen Jahre 30 bis 59 Tage mit einer Ratenzahlung in Verzug geraten ist.

NumberOfTime60-89DaysPastDueNotWorse Ein ganzzahliger Wert, der angibt, wie oft ein Bankkunde innerhalb der zwei vorangegangenen Jahre 60 bis 89 Tage mit einer Ratenzahlung in Verzug geraten ist.

NumberOfTimes90DaysLate Ein ganzzahliger Wert, der angibt, wie oft ein Bankkunde innerhalb der zwei vorangegangenen Jahre mindestens 90 Tage mit einer Ratenzahlung in Verzug geraten ist.

MonthlyIncome Ein float-Wert, der das regelmäßige monatliche Einkommen des Bankkunden beschreibt.

DebtRatio Ein Prozentwert der das Verhältnis von Tilgungszahlungen, Unterhaltsverpflichtungen und Lebenshaltungskosten im Vergleich zum monatlichen Bruttoeinkommen angibt.

NumberOfOpenCreditLinesAndLoans Ganzzahliger Wert, der die Anzahl offener Kredite und Darlehen, sowie Kreditkarten aufsummiert.

NumberRealEstateLoansOrLines Ein ganzzahliger Wert, der Kredite und Darlehen im Zusammenhang mit einer Wohnimmobilie zusammenfasst.

NumberOfDependents Die Summe der festen Haushaltsmitglieder des Bankkunden.

Ein elftes Feature stellt **SeriousDlqin2yrs** dar. Es handelt sich dabei um einen Binärwert, der die Datenmenge clustered in Personen mit bereits bewiesenen erheblichen Zahlungsschwierigkeiten und solche ohne.

Analyse der Datenkonsistenz

Im Rahmen der explorativen Analyse wird im Folgenden zunächst betrachtet, wie vollständig der Datensatz ist.

GRAFIK: In(?) [8] data-consistency

Die Analyse zeigt, dass in den meisten Features die Angaben tatsächlich komplett sind. Da es sich um einen Datensatz von Bankkunden handelt, die mindestens einen Kredit bedienen, ist es allerdings verwunderlich, dass fast 30.000 mal die Angabe zum monatlichen Einkommen fehlt. Das betrifft fast 20 Prozent der Datensätze. Sollten die späteren Analysen ergeben, dass Einkommen ein relevantes Feature ist, so müssten im Rahmen einer Datenbereinigung möglicherweise diese Datensätze komplett entfernt werden. Eine weitere Auffälligkeit im Datensatz gibt es nur bei den Angaben zu den Haushaltsmitgliedern, wobei hier die Fehlzahl mit knapp 4.000 vergleichsweise niedrig ausfällt.

Nun werden einige auffällige Features näher betrachtet.

GRAFIK: In [10], Out[10]

Der bereits erwähnte Binärwert schlägt immerhin bei etwas mehr als 10.000 Datensätzen an. Dies sind die Kunden, die den Banken Kopfzerbrechen bereiten und am besten durch eine passende Einschätzung rechtzeitig erkannt und mit entsprechenden Kreditkonditionen als Risiko manage-bar gemacht werden.

GRAFIK: Out [11] financial distress

Ein Histogramm dieser Personengruppe zeigt noch einmal deutlich, wie klein der Anteil der Bankkunden mit erheblichem Zahlungsverzug ist. Im Datensatz sind etwa sieben Prozent betroffen.

Explorative Analyse der relevanten einzelnen Features

RevolvingUtilizationOfUnsecuredLines

GRAFIK: Out[13],Out[14] nebeneinander

Bei diesem Feature ist der Input in einem Intervall von null bis zwei zu sehen. Dieses Intervall wurde auch aufgrund von Ausreißern gewählt. Es zeigt sich, dass die meisten Werte zwischen null und eins liegen, während die darüberlegenden Einträge anzeigen, dass die Kreditlimits überzogen wurden. Es lässt sich bereits an dieser Stelle der explorativen Analyse festhalten, dass es eine hohe Korrelation zwischen einem hohen Wert dieses Features und unserem Zielwert, den Ratenzahlungsproblemen, gibt.

DebtRatio

GRAFIK: Out[17], [18]

Auch hier verteilen sich die meisten Werte im Bereich zwischen null und eins. Darüber liegende Werte zeigen an, dass die Kunden im Monat mehr für ihre Kredite bezahlen müssen, als sie an

monatlichem Einkommen haben. Hier wäre also eine starke Korrelation zu Zahlungsproblemen zu vermuten. Ein Ausreißer der deutlich höher als zwei auf der Skala liegt und deshalb im Plot aus Gründen der besseren Visualisierung nicht berücksichtigt werden kann, hat den Wert 329.664. Dies könnte bedeuten, dass die Person sehr hohe Schulden bei gleichzeitig minimalem Einkommen hat, oder einfach ein Fehler bei der Dateneingabe sein. Da aber 31.045 der Datensätze einen DebtRatio-Wert von über zwei haben, wird es im nächsten Schritt herausfordernd sein, Ausreißer in den Daten sinnvoll zu eliminieren. Die Daten scheinen bei diesem Feature schlecht ausbalanciert zu sein. Als Ergänzung zu den Plots ergibt die Berechnung von Mittel- und Medianwerten eine mögliche Korrelation zwischen der DebtRatio und Zahlungsschwierigkeiten der Klienten. GRAFIK: In: [21],[22] (?)

MonthlyIncome

GRAFIK: Out:[23], Out[26]

Der Plot der Verteilung des monatlichen Einkommens zeigt viele Ausreißer, was auf ein Ungleichgewicht in den Daten schließen lässt, sonst aber wenig Auffälliges. Hier kommen die sehr vielen fehlenden Einträge zum tragen, die bereits bei der Datenkonsistenzanalyse festgestellt wurden. Der Datensatz selbst deckt ein großes Einkommensspektrum ab, so dass er insofern relevant ist, dass verschiedene Bevölkerungsgruppen berücksichtigt wurden. Die distress-Kurve folgt ebenfalls einfach der Einkommensverteilung und zeigt da höhere Werte, wo auch mehr Bankkunden gezählt wurden. Annähernd normal-verteilt sieht es beim Feature Age aus. Bei genauerer Betrachtung fällt aber auf, dass es erneut Lücken im Datensatz gibt. Trotz des großen Datensatzes von 150.000 Einträgen, sind einzelne Altersgruppen nicht im Datensatz zu finden, was sich durch Lücken im Plot zeigt.. Man sieht dort weder einen Y-Wert für “distressed” noch für “non-distressed”. Ursachen könnten Fehler bei der Dateneingabe oder wiederum ein schlecht ausbalancierter Datensatz sein. In der Altersgruppe der 40- bis 60-Järhigen zeigt der Plot, dass trotz ähnlich hoher Zahl der Alterskohorten die jüngeren Kunden eher zu Zahlungsschwierigkeiten neigen. Vielleicht sind sie Risiko-freudiger als die älteren Bankkunden.

Payment Delays

Unter Payment Delays lassen sich drei ausgesprochen ähnliche Features zusammenfassen. Sie entstammen einem Feature, den Zahlungsproblemen, und wurden in drei Kohorten aufgeteilt. Sie beschreiben alle de Häufigkeit von Zahlungsverspätungen der Bankkunden und unterscheiden sich nur nach der Art der Verspätung (kürzerer oder längerer Zeitraum). Für die Bank sind dabei die Kunden mit den gravierendsten Zahlungsproblemen am schwierigsten zu managen.

GRAFIK: [34],[35],[36]

Die Daten zeigen wenig überraschend, dass Kreditnehmer mit bedeutenden Zahlungsschwierigkeiten tendenziell auch ein Ausfallrisiko darstellen. Man könnte auch interpretieren, dass Personen, die immer wieder Zahlungsprobleme haben, schlecht mit ihrem Geld umgehen können.

NumberRealEstateLoansOrLines

GRAFIK: [37],[38]

Die Plots zeigen erwartbare Ergebnisse. Die Zahl der Menschen, die viele Kredite im Zusammenhang mit ihrer Immobilie bedienen müssen ist recht begrenzt. Da mit jeder Hypothek ein Anteil an der Immobilie an die Bank überschieben wird, ist ihre Anzahl beschränkt durch den Wert der Immobilie. Die Berechnung von Mittel- und Medianwerten zeigt allerdings eine nicht sofort erklärbare Verschiebung. (RICHARD???????)

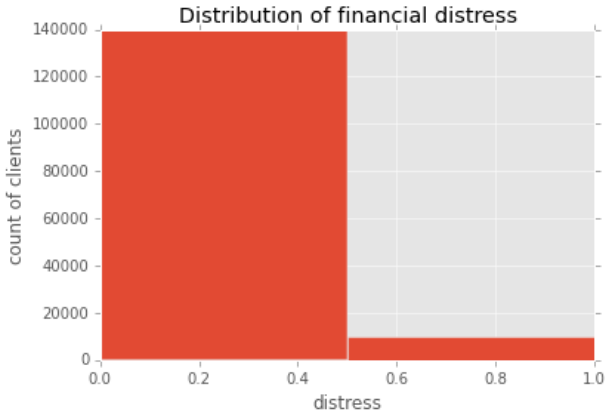
NumberOfDependents

GRAFIK: [41],[42]

Bei der Anzahl der Haushaltsmitglieder könnte man a priori vermuten, dass kinderreiche Haushalte eher Probleme mit Schulden haben, da das Einkommen auf mehr Köpfe verteilt werden muss. Dies lässt sich anhand der Daten jedoch nicht bestätigen, da auch Haushaltsmitglieder erfasst werden konnten, die zum Beispiel eigenes Einkommen haben und den Kreditnehmer daher finanziell nicht belasten. Die Korrelation zwischen distress und Zahl der Haushaltsmitglieder scheint durch die Verteilung von distress auf die Gesamtzahl der Datensätze überlagert zu werden.

3 FRAGMENT(ENTFERNEN?)

Die Zielklasse ist *SeriousDlqin2yrs*, welche Aussage darüber trifft, ob der Kreditnehmer innerhalb zweier Jahre nach Vertragsschluss in ernsthafte Zahlungsschwiereigkeiten kam. Zunächst betrachten wir die Verteilung dieser Klasse auf den Trainingsdaten.



#FRAGMENT ENDE

4 Data Wrangling

5 Erstellung der Modelle

5.1 Logistic Regression

5.1.1 Cross Validation

5.2 Support Vector Machine

5.2.1 Grid Search

5.3 Random Forest

6 Performancevergleich der Modelle

7 Ergebnisse

8 Fazit