Karlsruher Institut für Technologie

Schätzung der Ausfallwahrscheinlichkeit von P2P-Krediten

Sylvia Schumacher

Dr. Markus Höchstötter

Inhaltsverzeichnis

Abbildungsverzeichnis

1. Einführung

Mit der Finanzierung des Sockels der Freiheitsstatue auf Ellis Island wurde bereits vor mehr als 100 Jahren erfolgreich ein erstes bekanntes Crowdfundingprojekt durchgeführt. Auf ganz ähnliche Weise, wie Joseph Pulitzer bereits 1885 eine Spendenkampagne zur Finanzierung des Sockels der Statue initiierte, finanzieren heute zahlreiche Privatpersonen über die sogenannte „Crowd“ ihre Kredite. Im Fokus dieser Arbeit wird das sogenannte **P2P-Lending** stehen, dessen Grundidee die Kreditvergabe von Privatpersonen an Privatpersonen ist (vgl. [10], [6]).

1.1. P2P Kreditmärkte

Crowdfunding liegt generell die Idee zugrunde, dass ein Kreditsuchender seinen Finanzierungsbedarf auf einer Plattform vorstellt. Über die (Online-)Plattform wird eine große Zielgruppe an potenziellen Kreditgebern, die „Crowd“ erreicht, die die Möglichkeit haben, für das vorgestellte Projekt finanzielle Mittel bereit zu stellen. Die Crowd besteht in seiner ursprünglichen Idee aus einer Vielzahl von Menschen.

Damit eine Finanzierung von Krediten oder verschiedenen Projekten wirtschaftlich ablaufen kann, müssen die Transaktionskosten für die Investoren und Kreditnehmer gering gehalten werden. Dies konnte durch moderne Informationstechnologie, insbesondere der Nutzung des Internets erreicht werden (vgl. [6]).

Bereits im März 2005, über drei Jahre vor Beginn der Finanzkrise, wurde die erste P2P-Plattform namens „Zopa“ in Großbritannien gegründet. Seitdem ist weltweit ein starkes Wachstum der P2P-Kreditmärkte zu beobachten. Bis 2025 sollen die internationalen P2P-Märkte der Venture Capital Firma *Foundation Capital* zufolge gar einen Umsatz von einer Trillion Dollar generieren. Als Investoren treten dabei zunehmend auch institutionelle Anleger in den Markt ein. Bei *Prosper* und *LendingClub*, den beiden großen

P2P-Plattformen in den USA, beträgt der Anteil institutioneller Anleger bereits bis zu *90* Prozent des Umsatzes (vgl. [5]).

1.2. Zielsetzung der Arbeit

Zielsetzung dieser Arbeit ist es, ein **Modell zur Schätzung der Ausfallwahrscheinlichkeit von P2P-Krediten** zu entwickeln, welches dem potenziellen Investor vor der tatsächlichen Kreditvergabe eine Einschätzung über das Verhalten des Kreditsuchenden vermittelt. Dadurch erhält der Kreditgeber eine Entscheidungshilfe und kann seine Chancen und Risiken bei der Kreditvergabe besser einordnen.

2. Modell zur Prognose der Ausfallwahrscheinlichkeit

Im Folgenden sollen mit Hilfe von maschinellen Lernverfahren zwei Modelle zur Prognose der Ausfallwahrscheinlichkeit von P2P-Krediten der Plattform *Bondora* entwickelt werden. Softwareseitig erfolgt die Modellentwicklung über das OpenSource-Statistikprogramm *R*.

Dabei wird auf zwei Ansätze des maschinellen Lernens zurückgegriffen, die im Kontext des Creditscorings stark verbreitet sind: die logistische Regression (vgl. [9]) und RandomForests (vgl. [7]).

2.1. Aufbereitung der exogenen Daten

Die P2P-Plattform Bondora wurde 2009 gegründet und bringt Investoren sowie Kreditnehmer in Europa zusammen. Insgesamt konnten über Bondora knapp zehn Prozent des von Kreditantragsstellern angefragten Volumens, nämlich *35* Millionen Euro, erfolgreichen finanziert werden (vgl. [1]).

Zur Modellentwicklung liegt ein historischer Datensatz der angefragten und vermittelten P2P-Kredite der Plattform vor (vgl. [4]). Der Datensatz besteht aus n=*35.123* Beobachtungen mit jeweils m=*173* Merkmalen und wurde am 09. April 2015 entnommen.

Vorselektion und Bereinigung der Daten Informationen über den Ausfall eines Kredits liegen nur bei erfolgreich durch Investoren finanzierten Kreditanträgen vor. Für die Datenbasis der Modellentwicklung wurden darüber hinaus lediglich diejenigen Kredite genutzt, deren geplantes Finanzierungsende in der Vergangenheit liegt. Dadurch verkleinert sich der vorliegende Datensatz zur Modellanpassung auf n=*7.306* Beobachtungen.

Aufgrund der *173* vorliegenden Merkmale des Datensatzes erscheint auch eine Selektion der Modellvariablen sinnvoll. Auf Basis von vermuteteten Zusammenhängen von Modellvariablen und Zielgröße verbleiben *26* potenzielle Variablen, darunter *13* metrische und *13* kategorielle. Eine übersichtliche Darstellung sowie Beschreibung der vorselektierten Merkmale befindet sich in den Abbildungen A.1 und A.2.

Umgang mit fehlenden Werten Die fehlenden Werte der potenziellen Modellvariablen wurden mit Hilfe von Imputationsmethoden vervollständigt. Dazu wurde je Modellvariable die bedingten Ausfallwahrscheinlichkeit betrachtet, die sich ergibt, wenn lediglich die Beobachtungen selektiert werden, die einen fehlenden Wert enthielten und anschließend wurden die fehlenden Werte mit verschiedenen statistischen Verfahren ersetzt. Die Wahl für eines der Maße wurde danach ausgelegt, welche Zielgröße der bedingten Verteilung der vollständigen Werte der Zielgröße der bedingten Verteilung mit unvollständigen Werten am Nächsten kommt:

1. Ersetzung der fehlenden Werte durch ein bereits existierendes Faktorlevel (*VerificationType*)
2. Erstellung eines neuen Faktorlevels für die fehlenden Werte (*credit_score*, *CreditGroup*, *marital_status_id*, *Gender*)
3. Ersetzung der fehlenden Werte durch den Median (*Employment_Duration_Current_Employer*,*income_total*, *nr_of_dependants*)
4. Ersetzung der fehlenden Werte durch das arithmetische Mittel (*SumOfBankCredits*)
5. Ersetzung durch einen anderen geeigneten Wert (*TotalNumDebts* (Wert: 0), *work_experience* (Wert: 0))

Aufteilung in Test- und Trainingsdaten Um sicherzustellen, dass das angepasste Modell letztlich auf seine Güte getestet werden kann, wurde der vorliegende Datensatz in einen Trainings- und einen Testdatensatz aufgeteilt.

Der Datensatz wurde mit Hilfe eines **Samplingverfahrens** zufällig im Verhältnis von ca. *2:1* in Trainings- (*5000* Beobachtungen) und Testdaten (*2306* Beobachtungen) geteilt. Eine Übersicht liefert Abbildung 2.1.

	Trainingsdatensatz	Testdatensatz
Anzahl Beobachtungen	5000	2306
Durchschnittliche Ausfallwahrscheinlichkeit	12,24 %	13,79 %

Abbildung 2.1.: Aufteilung des Datensatzes in Trainings- und Testdatensatz

2.2. Modellsimulation logistische Regression

Anpassung an die Trainingsdaten Im Rahmen dieser Arbeit erfolgte die Modellanpassung in einem zweistufigen Verfahren:

1. **Modellauswahl anhand des AIC-Informationskriteriums:** Zunächst wurde auf Basis aller vorselektierten, potenziellen Variablen ein Regressionsmodell anhand des AIC-Informationskriteriums ermittelt. Zur softwareseitigen Unterstützung der automatischen Variablenselektion wurde für die hier vorliegende Arbeit wurde auf das Packet ***glmulti*** (vgl. [3]) zurückgegriffen.

   Die Regressionsergebnisse des mit Hilfe der „consensus"-Methode des *20*-fach angewendeten genetischen Algorithmus ermittelten besten Modells sind in Abbildung A.3 dargestellt.

2. **Modellanpassung mit Hilfe der schrittweisen logistischen Regression:** Da das unter 1. generierte Modell vergleichsweise viele Variablen für eine Scorecard enthielt, wurde mit Hilfe der Rückwärtsselektion versucht, die Komplexität des Modells weiter zu reduzieren, ohne dabei ein statistisch signifikant schlechteres Modell zu erhalten.

   Als Kennzahl zur Auswertung der Signifikanz der einzelnen Variablen wurde der **p-Wert** verwendet. Die p-Werte der metrischen Variablen können direkt aus der „summary"-Funktion des Anpassungsmodells abgelesen werden und sind in der linken Spalte in Abbildung 2 dargestellt.

Variable	p-Wert (aufsteigend)
CurrentLoanHasBeenExtended	< 2e-16
Credit_score	4.81e-15
LoanDuration	6.23e-06
NewCreditCustomer	0,000867
CountOfBankCredits	0,012687
FundedAmount	0,01567
Age	0,015901

p-Werte der metrischen Modellvariablen (logistische Regression)

Variable	p-Wert (aufsteigend)
VerificationType	< 2.2e-16
CreditGroup	0,0005133
education_id	0,06524
employment_status_id	0,071
Country	0,509
language_code	0,5897
ApplicationType	0,8137

p-Werte der kategoriellen Modellvariablen (logistische Regression)

Abbildung 2.2.: p-Werte der Modellvariablen (logistische Regression)

Hierbei fällt auf, dass alle metrischen Variablen mit einer Irrtumswahrscheinlichkeit von nur *2 %* oder weniger bereits signifikant zur Erklärung der Zielgröße beitragen. Bei den kategoriellen Modellvariablen erreichen die p-Werte der Variablen *language_code* (p-Wert: 0,5897), *Country* (p-Wert: 0,5090) und *ApplicationType* (p-Wert: *0,8137*) kein statistisch signifikantes Niveau.

Beim Versuch ein neues Modell ohne alle drei Modellvariablen anzupassen ergibt sich jedoch eine statistisch signifikante Verschlechterung (p-Wert: *0,0018*). Eine mögliche Erklärung hierfür ist, dass die Variablen *language_code* sowie *Country* geographische und damit ähnliche Informationen über einen Kreditnehmer beinhalten. Aufgrund ihres geringeren p-Wertes wurde die Variable *Country* wieder in das Modell aufgenommen. Dieses Modell schneidet nicht signifikant schlechter ab (p-Wert: *0.6728*) und wird als final angepasstes Modell der logistischen Regression verwendet. Die Regressionsergebnisse für dieses Modell sind in A.4 abgebildet.

Interpretation der In Sample-Regressionsergebnisse

Wahl des Schwellenwerts In Abbildung A.5 ist die Trefferquote in Abhängigkeit verschiedener Schwellenwerte dargestellt. Wie aus Abbildung A.5 hervorgeht, kann durch eine Variation des Cutoff-Werts das Modell nach dem Kriterium der Trefferquote optimiert werden. So ist das **Optimum bei einem Cutoff-Wert von *0,46*** und einer **Trefferquote von *89,86 %*** erreicht.

$\widehat{POD}$ \ $\overline{POD}$	0	1
0	1928	182
1	60	136

Abbildung 2.3.: Wahrheitsmatrix (Schwellenwert: *0,46*), (Testdaten logistische Regression)

Anwendung des Modells auf die Testdaten Um die Güte des Anpassungsmodells zu testen, wird für die *2306* Beobachtungen der Trainingsdaten zunächst ein Scorewert prognostiziert. Die zugehörige Wahrheitsmatrix ist Abbildung 2.3 zu entnehmen. Der **positive Vorhersagewert**, der vor allem für Investoren relevant ist, beträgt *91,37 %*.

2.3. Modellsimulation RandomForests

Anpassung an die Trainingsdaten Die Modellanpassung eines Zufallswalds gestaltet sich mit Hilfe des R-Packets „RandomForest“, entwickelt von Breiman and Cutler [2] sehr einfach.

Als potenzielle Modellvariablen für die Generierung des Zufallswalds wurden alle *26* vorselektierten metrischen und kategoriellen Inputvariablen bereitgestellt. Die Anzahl der Entscheidungsbäume wurde variiert, um den Lernprozess des Zufallswalds besser nachvollziehen zu können. Bei der Anzahl der für jeden Split bereitgestellten Variablen wurde der Default-Wert von $\frac{p}{3}$ beibehalten. Dies bedeutet, dass bei p potenziellen Modellvariablen bei einer Regression $\frac{p}{3}$ Variablen pro Split zur Verfügung stehen.

Interpretation der In Sample-Regressionsergebnisse Der Beitrag der einzelnen Variablen zum Modell wird bei RandomForests in *R* automatisch erfasst. Hierfür stehen zwei verschiedene Alternativen zur Verfügung:

- **Verbesserung im Splitkriterium:** Diese Alternative prüft, welchen Beitrag die jeweilige Variable in den einzelnen Knoten zur Knotenreinheit liefert.
- **Permutation der OOB-Beobachtungen:** Hierbei werden die einzelnen Modell-

variablen permutiert, um deren Einfluss auf die Zielgröße zu eliminieren. Anschließend erfolgt ein Vergleich des modifizierten Modells mit dem urspünglichen Modell (vgl. [2]).

Die Variablenwichtigkeit der Modellsimulation ist Abbildung A.3 zu entnehmen.

Anwendung des Modells auf die Testdaten Analog zur Modellsimulation mit Hilfe der logistischen Regression wird auch für den RandomForest die Modellgüte berechnet.

Die ROC-Kurve für die Modellanpassung an die Testdaten bestätigt zunächst die gute Diskriminierungsfähigkeit der Scorecard (vgl. Abbildung A.9). Der zugehörige AUC-Wert von *0,94* spiegelt ebenfalls eine gute Vorhersagegüte der Scorecard wieder.

In Abbildung A.7 ist die Trefferquote in Abhängigkeit verschiedener Cutoffwerte dargestellt. Ein Modell, welches allen Beobachtungen den häufiger vorkommenden binären Wert, im vorliegenden Fall also den Nichtausfall eines Kredits zuordnen würde, würde insgesamt *86,21 %* der Kredite korrekt einordnen. Die Trefferquoten für das simulierte Modell erscheinen daher mit Werten von über *91 %* vergleichsweise hoch. Die höchste Trefferquote wird bei einem Cutoff von *0,48* erzielt.

Abbildung A.8 ist zu entnehmen, dass ein sehr großer Teil des maschinellen Lernens der Simulation bereits mit den ersten *50* Entscheidungsbäumen erfolgt (Trefferquote: *91,37 %*). Der Zufallswald, der aus *1000* Entscheidungsbäumen besteht kann mit der Trefferquote von *92,24 %* nochmals eine deutliche Verbesserung vorweisen.

Eine Wahrheitsmatrix für das hinsichtlich der Trefferquote beste Modell ist Abbildung 2.4 zu entnehmen. Für das Simulationsmodell beträgt der positive Vorhersagewert *0,9326*. Das bedeutet, dass im Falle der Investition in die nach dem Modell klassifizierten „guten“ Kredite, lediglich *6,74 %* insgesamt ausfallen.

$\widehat{POD}$ \ $\overline{POD}$	0	1
0	1950	141
1	38	177

Abbildung 2.4.: Wahrheitsmatrix (Testdaten RandomForest, Schwellenwert: *0,48*)

2.4. Vergleich der Vorhersagegüte der beiden Regressionsmodelle

Gesamtperformance Ein Vergleich der Wahrheitsmatrizen der beiden Modelle zeigt, dass das RandomForest-Modell das logistische Simulationsmodell dominiert (vgl. Abbildung 2.5). So gelingt es dem RandomForest eine deutlich geringere Falsch-Positiv Rate zu erreichen (RandomForest: *6,12 %*, logistische Regression: *7,89 %*). Rechnerisch werden somit bei den RandomForests *22,5%* weniger Fehler ersten Grades begangen. Zusätzlich ist auch die Quote der korrekterweise als „gut“ klassifizierten Kredite im RandomForest höher als beim logistischen Regressionsmodell (RandomForest: *84,56 %*, logistische Regression: *83,61 %*). Bei den negativ klassifizierten Krediten spiegelt sich erneut die hohe Güte des RandomForest Modells wieder. So werden sowohl mehr Kredite korrekterweise falsch klassifiziert (RandomForest: *7,68 %*, logistische Regression: *5,90 %*), als auch weniger Kredite fehlerhaft in die Klasse der schlechten Kredite eingeordnet (RandomForest: *1,64 %*, logistische Regression: *2,60 %*).

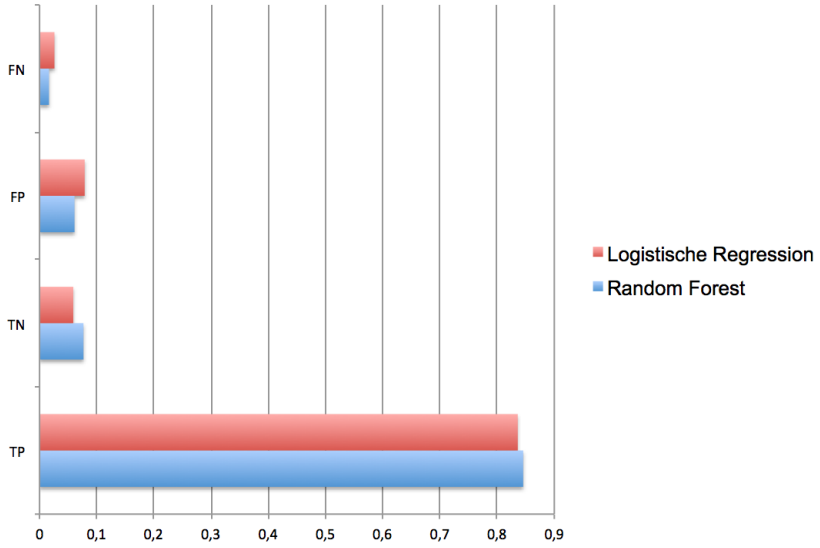

Abbildung 2.5.: Vergleich der Wahrheitsmatrizen (Testdaten) von logistischer Regression (rot) und RandomForest (blau)

Auch die Trefferquote bei der Anwendung der Modelle auf den Testdatensatz spiegelt die Überlegenheit des Zufallswaldmodells gegenüber dem logistischen Regressionsmodell

wieder (RandomForest: *92,24 %*, logistische Regression: *89,51 %*). Ein Vergleich der beiden ROC-Kurven zeigt, dass sogar unabhängig vom Cutoff-Wert, also der Wahl des diskreten Klassifikators das Zufallswaldmodell besser abschneidet.

Modellvariablen Zunächst fällt hier auf, dass die Anzahl der Modellvariablen sich deutlich unterscheidet (logistisches Regressionsmodell: *13*, Zufallswald: *26*). Dennoch scheinen sich ähnliche Variablen als sehr wichtig herauszustellen:

- **Schuldnerbezogene, persönliche Merkmale:** Insgesamt trägt dieses Cluster bei beiden Simulationsmodellen nur mäßig zur Erklärung der Zielgröße bei. Regionale Indikatoren haben bei beiden Modellen eine mittelmäßige bis geringe Bedeutung. Variablen, die die Jobsituation des Antragstellers wiederspiegeln, haben ebenfalls in beiden Simulationsmodellen eine eher geringe Bedeutung. So wird im logistischen Regressionsmodell gar keine der drei Variablen *Employment_Duration_Current_Employer*, *Employment_status_id* und *work_experience* verwendet.

  Eine weitere Übereinstimmung der Modelle zeigt sich im Geschlecht (*Gender*) und Alter (*Age*) des Antragstellers. Während das Geschlecht statistisch kaum signifikant ist, spielt das Alter in beiden Modellen eine essentielle Rolle.

  Darüber hinaus zeigen beide Modelle die niedrige Erklärungskraft der Indikatoren, die den familiären Hintergrund des Antragstellers wiederspiegeln (*Marital_status_id*, *nr_of_dependants*). Unerwarteterweise scheint der Bildungsstand eine gute Erklärung für die Zielgröße zu liefern. So wird der Indikator *education_id* in beiden Modellen als Modellvariable verwendet und erzielt hohe Kennzahlenwerte bei der Messung der Variablenwichtigkeit im Zufallswald.

- **Schuldnerbezogene, finanzielle Merkmale:** Auch in diesem Cluster weisen beide Regressionsmodelle ähnliche Ergebnisse vor. Bonitätsbezogene Merkmale werden in beide Regressionsmodelle einbezogen und haben eine relativ hohe Bedeutung. Während das logistische Regressionsmodell sowohl die Variable *credit_score* als auch die *CreditGroup* als Modellvariablen einbezieht, scheint im RandomForest überwiegend der *credit_score* wertvolle Informationen zum Regressionsmodell beizutragen. Alle anderen Merkmale dieses Clusters sind vergleichsweise unbedeutend für die beiden Regressionsmodelle.

- **Finanzierungsbezogene Merkmale:** Dem Cluster der finanzierungsbezogenen

Merkmale kommt in beiden Modellanpassungen die höchste Bedeutung zu. Die wichtigsten Variablen des Clusters stimmen dabei in beiden Modellen überein.

Dabei stammen nach den Kriterien der Quadratsumme der Residuen sowie der permutierten Fehlerfreiheit sogar jeweils die wichtigsten drei Variablen aus dem Cluster der finanzierungsbezogenen Merkmale. Das Simulationsmodell scheint also sehr gut über die finanzierungsbezogenen Aspekte abbildbar. Die wichtigsten drei Variablen sind *CurrentLoanHasBeenExtended*, *LoanDuration* sowie *VerificationType*. Während es absolut nicht verwunderlich ist, dass die Laufzeitdauer essentiell für die Modellanpassung ist, bringt die Variable *CurrentLoanHasBeenExtended* doch eine etwas größere Überraschung mit sich. Ob ein Kredit ausgeweitet wird oder nicht, scheint die Bonität des Gläubigers also wesentlich zu beeinflussen. Darüber hinaus spiegelt die hohe Bedeutung der Variablen *VerificationType* die große Unsicherheit der Investoren hinsichtlich wahrheitsgemäßen Angaben der Kreditnehmer wieder. Dies scheint nachvollziehbar unter dem Aspekt, dass es sich auf dem P2P-Kreditmarkt um unbesicherte Kredite handelt, die bei einem Ausfall häufig zu einem fast kompletten Verlust der verbliebenen Forderung führen.

Eine hohe Bedeutung im oberen Drittel kommt auch der Variablen *FundedAmount* zu. Da mit höherer Rückzahlungssumme die Gefahr eines Ausfalls steigt, scheint dies gut nachvollziehbar.

Weniger wichtig für die Modellanpassung sind hingegen der Tag und die Stunde der Antragsstellung. Eine Ausnahme stellt hier die Kennzahl der permutierten Fehlerfreiheit dar, nach der die Variable *ApplicationSignedHour* sich gerade noch im oberen Drittel befindet. Eine Erklärung hierfür ist schwer zu finden. Die Variable *NewCreditCustomer* kann nach den beiden Kennzahlen fast vernachlässigt werden. Eher mittelmäßig schneiden auch die beiden Variablen *UseOfLoan* sowie *NoOfPreviousApplications* ab.

3. Zusammenfassung

Im letzten Kapitel sollen wichtige Optimierungsansätze vorgestellt und abschließend ein Fazit gezogen werden.

3.1. Optimierungsansätze der Modelle

Die Anpassungsmodelle bieten noch zahlreiche Potenziale, die die Modellgüte möglicherweise weiter erhöhen würden:

- **Korrelationen zwischen den Modellvariablen:** Bislang wurden Abhängigkeiten zwischen den einzelnen Variablen nicht berücksichtigt. Strobl et al. [8] haben bereits gezeigt, dass das RandomForest-Verfahren eine Verzerrung bei der Variablenselektion vorweist. So werden bei kategoriellen Merkmalen Indikatoren mit vielen Kategorien gegenüber Variablen mit weniger Kategorien bevorzugt. Als Lösung entwickelte Strobl et al. [8] die *cforest*-Funktion des R-Packets *party*.

  Bei der logistischen Regression hingegen hätte eine detaillierte Varianzanalyse vorgeschaltet werden können, um Abhängigkeiten zu identifizieren und diese entsprechend im Modell zu berücksichtigen.

- **Imputationsmethoden:** Im Rahmen dieser Arbeit wurde auf eher weniger komplexe Methoden zurückgegriffen. Dies wurde insbesondere damit gerechtfertigt, dass nur potenzielle Modellvariablen mit vergleichsweise wenigen fehlenden Werten vorselektiert wurden.

- **Auswahl der Modellvariablen:** Die Vorselektion der *179* Variablen erfolgte im Rahmen dieser Arbeit auf Basis von vermuteten Zusammenhängen. Somit ist nicht auszuschließen, dass wichtige Variablen, die die Regressionsmodelle signifikant verbessert hätten, bereits in der Vorselektion als Modellvariablen ausgeschlossen wur-

den.

Außerdem wurde die Möglichkeit ausgelassen, weitere Variablen zu konstruieren. Diese hätten einerseits aus den bestehenden Daten abgeleitet werden können. Andererseits hätten externe Indikatoren in das Modell mit aufgenommen werden können, wie beispielsweise makroökonomische Indikatoren.

- **Vorselektion des Datensatzes:** Es hätte eine höhere Anzahl an Test- und Trainingsdaten erreicht werden können, wenn zur Modellanpassung weitere Kredite berücksichtigt wären worden, deren Laufzeitende noch nicht erreicht ist.
- **Zu optimierende Kennzahl des Modells:** Bei der Modellanpassung wurde der Schwellenwert so ausgewählt, dass die Trefferquote maximiert wurde. Eine Festlegung auf einen einzigen Schwellenwert erscheint jedoch nur begrenzt sinnvoll, da die Fehlklassifizierungskosten der einzelnen Kredite nicht bekannt sind.
- **Trendanalyse:** Eine Analyse der historischen Ausfallraten nach Jahren zeigt, dass eine höhere Jahreszahl auch höhere Ausfallraten impliziert. Eine Trendanalyse und die Integration dieser Information in die beiden Modelle hätte die Vorhersagegüte möglicherweise stark erhöht.

3.2. Fazit und Ausblick

In der hier vorliegenden Arbeit wurde gezeigt, dass mit Hilfe von maschinellen Lernverfahren gute Prognosen hinsichtlich der Ausfallwahrscheinlichkeit von P2P-Krediten möglich sind. Dabei konnte mit einem RandomForest-Modell eine deutlich bessere Vorhersagegüte als mit einem logistischen Regressionsmodell erzielt werden. Eine mögliche Erklärung hierfür ist, dass bei der Entwicklung der einzelnen Entscheidungsbäume des Entscheidungswalds eine deutlich höhere Flexibilität gegeben ist.

Insgesamt konnte gezeigt werden, dass insbesondere Indikatoren, die die Rahmendaten der Finanzierung beschreiben von hoher Relevanz für die Modellanpassung sind. Dies ist verwunderlich vor dem Hintergrund, dass es sich bei P2P-Krediten um unbesicherte Kredite handelt. Eine intuitive Vermutung würde eher nahe legen, dass schuldnerbezogene Merkmale, seien sie von privater oder finanziell beschreibender Natur, aufgrund der großen Unsicherheit hinsichtlich des Verhaltens der Schuldner einen höheren Stellenwert

einnehmen. Dass eine Kreditvergabe letztlich insbesondere von den Rahmendaten der Finanzierung abhängt, bietet viel Freiraum und Potenziale für Schuldner mit schlechter Bonität. Andererseits haben Schuldner insbesondere bei den privaten und finanziellen schuldnerbezogenen Merkmalen theoretisch die Möglichkeit, nicht wahrheitsgemäße Angaben zu machen, weshalb eine geringe Bedeutung dieser Merkmale auch wiederum als Qualitätssiegel für P2P-Märkte interpretiert werden kann.

Mit Hilfe dieser Arbeit wurde eine Grundlage für die Prognose der Ausfallwahrscheinlichkeit einer spezifischen P2P-Kreditplattform, in diesem Falle Bondora, geschaffen. In den nächsten Jahre wird sich vermutlich herausstellen, wie groß das Potenzial dieser innovativen Finanzierungsform tatsächlich ist.

Bibliography

[1] *About Bondora.* `https://www.bondora.ee/en/about-us`. Entnommen am: 02.04.2015. 2015.

[2] Leo Breiman and A Cutler. "Random forests Classification manual". In: *http://www.math. usu.edu/adele/forests* (2008).

[3] Vincent Calcagno and Claire de Mazancourt. "glmulti: an R package for easy automated model selection with (generalized) linear models". In: *Journal of Statistical Software* 34.12 (2010), pp. 1–29.

[4] *Data Export.* `https://www.bondora.ee/en/invest/statistics/data_export`. Entnommen am: 09.04.2015. 2015.

[5] Emma Dunkley. *Royal Bank of Scotland to enter P2P lending market.* `http://www.ft.com/intl/cms/s/0/660447b0-5625-11e4-93b3-00144feab7de.html?siteedition=intl#axzz3Gimzs400`. Entnommen am: 25.10.2014. 2014.

[6] Dominik Faßbender. *P2P-Kreditmärkte als Finanzintermediäre: eine empirische Analyse deutscher P2P-Kreditmärkte zur Beurteilung der Eignung als Finanzintermediäre.* GRIN Verlag, 2012.

[7] Trevor Hastie et al. *The elements of statistical learning.* Vol. 2. 1. Springer, 2009.

[8] Carolin Strobl et al. "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8.1 (2007), p. 25.

[9] Lyn C Thomas, David B Edelman, and Jonathan N Crook. *Credit scoring and its applications.* Siam, 2002.

[10] Christina Waider. *Crowdfunding als alternatives Filminvestitionsmodell: Ist Crowdfunding und Crowdinvesting ein zukunftsfähiges Filmfinanzierungsmittel?* Diplomica Verlag, 2013.

A. Anhang

A.1. Potenzielle, vorselektierte Modellvariablen

	Variablendeklaration	Beschreibung
1	**Age**	Age of the borrower (years)
2	**ApplicationSignedHour**	Hour of signing the loan application
3	**ApplicationSignedWeekday**	Weekday of signing the loan application
4	**CountOfBankCredits**	Number of liabilities issued by banks
5	**Employment_Duration_Current_Employer**	Employment time with the current employer
6	**FundedAmount**	Amount the borrower received
8	**income_total**	Total income
9	**LoanDuration**	The loan term
10	**NoOfPreviousApplications**	Number of previous loan applications
11	**nr_of_dependants**	Number of children or other dependants
12	**SumOfBankCredits**	Sum of liabilities issued by banks
13	**work_experience**	Work experience in total

Abbildung A.1.: Vorselektierte metrische Modellvariablen

	Variablendeklaration	Beschreibung	Faktorlevels
1	**ApplicationType**		1 Timed funding (loan will be paid out after auction time runs out) 2 Quick funding (loan will be paid out as soon as the amount is full)
2	**Country**	Residency of the borrower	1 EE (Estonia) 2 ES (Spain) 3 FI (Finland) 4 SK (Slovak Republic)
3	**credit_score**		1000 No previous payments problems 900 Payments problems finished 24-36 months ago 800 Payments problems finished 12-24 months ago 700 Payments problems finished 6-12 months ago 600 Payment problems finished <6 months ago 500 Active payment problems
4	**CreditGroup**	Credit Group of the borrower	A Best Credit Group B Medium Credit Group C Bad Credit Group
5	**CurrentLoan HasBeenExtended**	Borrower has rescheduled current loan	0 No 1 Yes
6	**education_id**		1 Primary education 2 Basic education 3 Vocational education 4 Secondary education 5 Higher education
7	**employment_status_id**		1 Unemployed 2 Partially employed 3 Fully employed 4 Self-employed 5 Entrepreneur 6 Retiree
8	**Gender**		0 Male 1 Woman 2 Undefined
9	**language_code**	-	1 Estonian 2 English 3 Russian 4 Finnish 5 German 6 Spanish 7 Slovakian
10	**marital_status_id**		1 Married 2 Cohabitant 3 Single 4 Divorced 5 Widow
11	**NewCreditCustomer**	Did the customer have prior credit history in Bondora	0 Customer had at least 3 months of credit history in Bondora 1 No prior credit history in Bondora
12	**UseOfLoan**		0 Loan consolidation 1 Real estate 2 Home improvement 3 Business 4 Education 5 Travel 6 Vehicle 7 Other 8 Health 110 Other business loans
13	**VerificationType**	Method used for loan application data verification	1 Income unverified 2 Income unverified, cross-referenced by phone 3 Income verified 4 Income and expenses verified

Abbildung A.2.: Vorselektierte kategorielle Variablen

A.2. Ergebnisse logistische Regression

Variable (Faktorlevel)	ß	s*(ß)	z-Wert	Pr(>|z|)	Sign.-level
Intercept	-1.188e+01	1.234e+03	-0.010	0.992320	
Country(ES)	-1.474e+01	2.118e+03	-0.007	0.994446	
Country (FI)	-9.442e-01	1.085e+00	-0.870	0.384320	
Country (SK)	-4.382e-01	1.042e+00	-0.421	0.673946	
CreditGroup (A)	1.307e-01	1.040e+00	0.126	0.899958	
CreditGroup (B)	-3.068e-01	1.065e+00	-0.288	0.773365	
CreditGroup (C)	-2.477e+00	1.454e+00	-1.704	0.088446	0.05
language_code(2)	2.918e-01	6.814e-01	0.428	0.668446	
language_code(3)	4.287e-01	1.922e-01	2.230	0.025717	0.01
language_code(4)	-1.392e+01	2.546e+02	-0.055	0.956398	
language_code(5)	5.990e-01	5.049e+03	0.000	0.999905	
language_code(6)	1.409e+01	2.118e+03	0.007	0.994691	
language_code(7)	-1.456e+01	6.523e+03	-0.002	0.998219	
language_code(15)	-4.409e-01	5.039e+03	0.000	0.999930	
language_code(17)	-7.080e-02	6.858e+03	0.000	0.999992	
language_code(20)	7.495e-01	6.858e+03	0.000	0.999913	
language_code(21)	-7.532e-01	6.858e+03	0.000	0.999912	
language_code(22)	-1.261e+01	6.523e+03	-0.002	0.998458	
employment_status_id(2)	1.460e+01	1.234e+03	0.012	0.990563	
employment_status_id(3)	1.443e+01	1.234e+03	0.012	0.990672	
employment_status_id(4)	1.457e+01	1.234e+03	0.012	0.990581	
employment_status_id(5)	1.374e+01	1.234e+03	0.011	0.991120	
employment_status_id(6)	1.469e+01	1.234e+03	0.012	0.990504	
NewCreditCustomer	4.217e-01	1.266e-01	3.331	0.000867	0
VerificationType(1)	-3.697e+00	5.717e-01	-6.467	1.00e-10	0
VerificationType(2)	-1.295e+00	4.600e-01	-2.814	0.004887	0.001
VerificationType(3)	-2.042e+00	4.863e-01	-4.199	2.68e-05	0
VerificationType(4)	-2.069e+00	4.561e-01	-4.537	5.71e-06	0
CurrentLoanHasBeenExtended	2.713e+00	1.924e-01	14.101	< 2e-16	0
education_id(1)	-8.885e-02	4.265e-01	-0.208	0.834991	
education_id(2)	2.839e-03	1.861e-01	0.015	0.987833	
education_id(3)	3.498e-01	1.774e-01	1.972	0.048649	0.05
education_id(4)	-1.315e-01	1.357e-01	-0.969	0.332728	
CountOfBankCredits	-1.657e-01	6.650e-02	-2.492	0.012687	0.05
FundedAmount	-1.725e-04	7.140e-05	-2.417	0.015670	0.05
Age	-1.294e-02	5.367e-03	-2.411	0.015901	0.05
LoanDuration	-2.131e-02	4.715e-03	-4.518	6.23e-06	0
credit_score	-2.163e-03	2.762e-04	-7.832	4.81e-15	0
ApplicationType	-3.310e-02	1.404e-01	-0.236	0.813665	

Abbildung A.3.: Regressionsergebnisse logistische Regression (Trainingsdaten) nach genetischem Algorithmus

Variable (Faktorlevel)	ß̂	s*(ß̂)	z-Wert	Pr(>|z|)	Sign.-level
Intercept	-8.919e+00	2.787e+02	-0.032	0.974474	
CreditGroup (A)	1.388e-01	1.040e+00	0.133	0.893867	
CreditGroup (B)	-3.007e-01	1.066e+00	-0.282	0.777874	
CreditGroup (C)	-2.429e+00	1.455e+00	-1.669	0.095053	0.05
Country(ES)	-7.467e-01	2.680e-01	-2.787	0.005327	0.001
Country (FI)	-3.020e+00	1.041e+00	-2.957	0.003109	0.001
Country (SK)	-4.957e-01	1.042e+00	-0.476	0.633869	
employment_status_id(2)	1.164e+01	2.787e+02	0.042	0.966696	
employment_status_id(3)	1.143e+01	2.787e+02	0.041	0.967282	
employment_status_id(4)	1.165e+01	2.787e+02	0.042	0.966661	
employment_status_id(5)	1.374e+01	1.234e+03	0.011	0.991120	
employment_status_id(6)	1.469e+01	1.234e+03	0.012	0.990504	
NewCreditCustomer	4.229e-01	1.264e-01	3.347	0.000818	0
VerificationType(1)	-3.672e+00	5.678e-01	-6.467	1.00e-10	0
VerificationType(2)	-1.295e+00	4.595e-01	-2.817	0.004840	0.001
VerificationType(3)	-2.047e+00	4.847e-01	-4.223	2.42e-05	0
VerificationType(4)	-2.064e+00	4.528e-01	-4.557	5.18e-06	0
CurrentLoanHasBeenExtended	2.709e+00	1.919e-01	14.118	< 2e-16	0
education_id(1)	-4.593e-02	4.274e-01	-0.107	0.914433	
education_id(2)	2.883e-02	1.842e-01	0.157	0.875638	
education_id(3)	3.768e-01	1.764e-01	2.136	0.032712	0.05
education_id(4)	-1.527e-01	1.344e-01	-1.136	0.255924	
CountOfBankCredits	-1.747e-01	6.585e-02	-2.652	0.007993	0.05
FundedAmount	-1.724e-04	7.110e-05	-2.425	0.015293	0.05
Age	-1.210e-02	5.327e-03	-2.272	0.023101	0.05
LoanDuration	-2.131e-02	4.718e-03	-4.514	6.37e-06	0
credit_score	-2.133e-03	2.749e-04	-7.758	8.65e-15	0

Abbildung A.4.: Regressionsergebnisse logistische Regression (Trainingsdaten) nach schrittweiser logistischer Regression

Modell	Cutoff	Trefferquote	Δ *a priori*
1	0,45	0,8982	0,1683
2	0,46	0,8986	0,1716
3	0,47	0,8982	0,1683
4	0,48	0,8982	0,1683
5	0,49	0,8984	0,1699
6	0,50	0,898	0,1667

Abbildung A.5.: Zusammenhang zwischen Trefferquote und Schwellenwert (logistische Regression)

A.3. Ergebnisse RandomForest-Modell

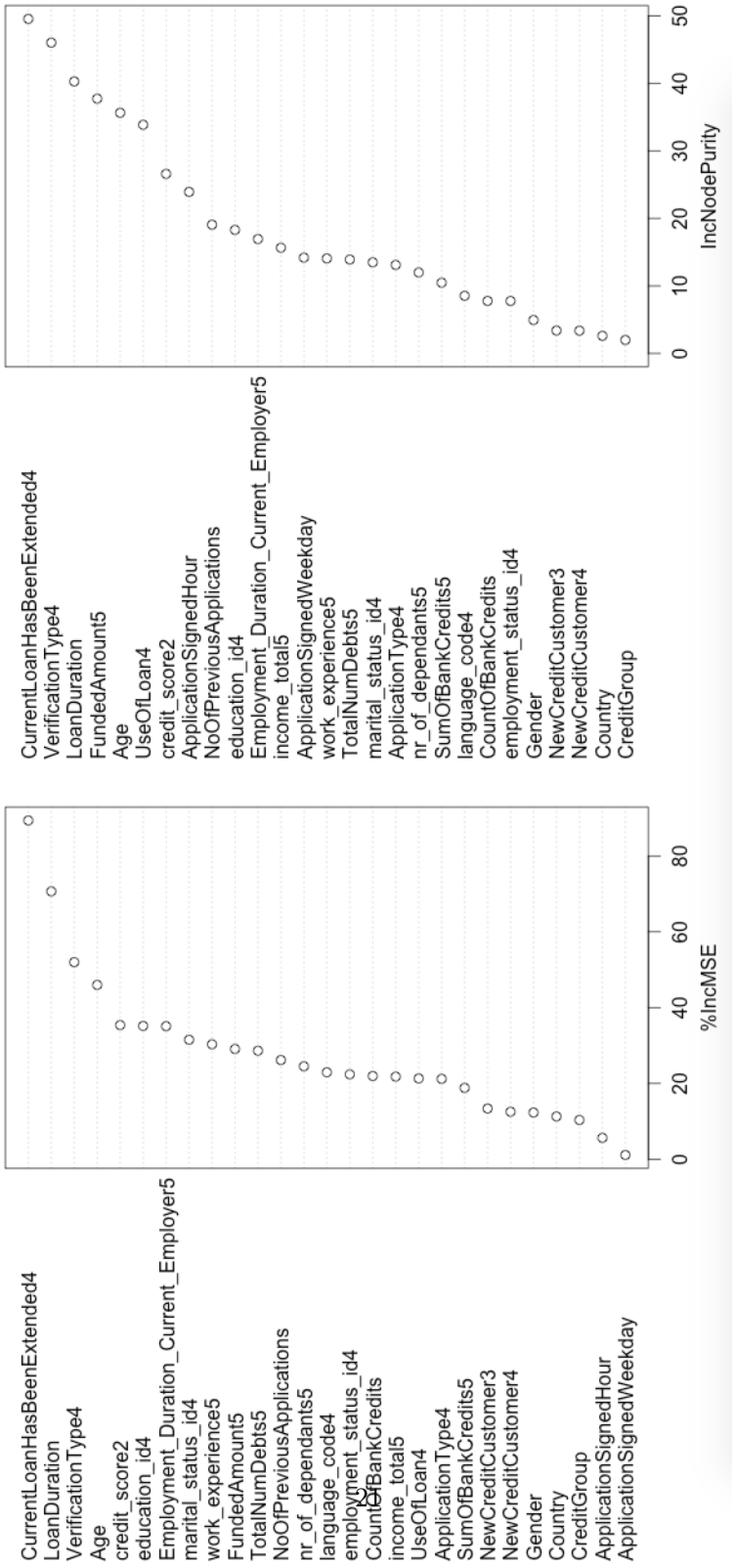

Abbildung A.6.: Variablenwichtigkeit im RandomForest-Modell

Modell	Cutoff	Trefferquote	Δ *a priori*
1	0,45	0,9198	0,0669
2	0,46	0,9206	0,0679
3	0,47	0,9206	0,0679
4	0,48	0,9224	0,0699
5	0,5	0,9189	0,0659

Abbildung A.7.: Zusammenhang zwischen Trefferquote und Schwellenwert (RandomForest)

Modell	Anzahl Entscheidungsbäume	Trefferquote	Δ *a priori*
1	50	0,9137	0,0599
2	500	0,9198	0,0669
3	1000	0,9224	0,0699

Abbildung A.8.: Zusammenhang zwischen Trefferquote und Anzahl der Entscheidungsbäume (Testdaten RandomForest)

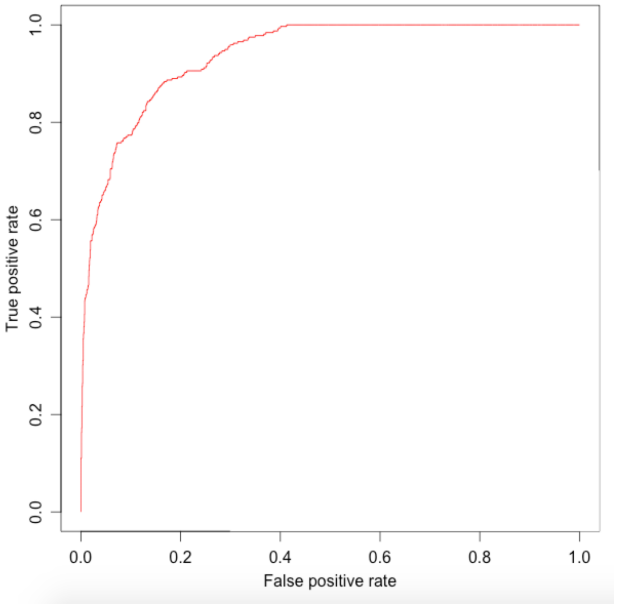

Abbildung A.9.: ROC-Kurve RandomForest (Testdaten)