

**Universität Paderborn**

**Fakultät für Wirtschaftswissenschaften**

**Department Wirtschaftsinformatik**

**Lehrstuhl für Wirtschaftsinformatik, insb. Data Analytics**

<Bachelor-/Master-/Seminararbeit>

<Titel der wissenschaftlichen Arbeit>

von

<Name des Autors>

<Matrikelnummer>

<Adresse>

<E-Mail>

vorgelegt bei

**Erstkorrektor:**<Titel und Name des Erstkorrektors>

**Zweitkorrektor:**<Titel und Name des Zweitkorrektors>

Eingereicht am <Abgabedatum>

**Sperrvermerk**

**Vom Verfasser zu unterschreiben:**

Die <Bachelor-/Master-/Seminararbeit> wurde im Rahmen eines laufenden Projektes bei der <Firmenname> durchgeführt. Um eine Geheimhaltung[[1]](#footnote-1) zu gewährleisten, verpflichte ich mich – von Beginn bis mindestens für 5 Jahre nach Abgabe der <Bachelor-/Master-/Seminararbeit> – Veröffentlichungen jeglicher Art der <Bachelor-/Master-/Seminararbeit> oder ihres wesentlichen Inhalts nicht vorzunehmen. Aus diesem Grund muss auch eine Aufnahme eines Exemplars der <Bachelor-/Master-/Seminararbeit> zur Einsicht und Ausleihe in der Bibliothek der Universität Paderborn während des genannten Zeitraums unterbleiben.

Ich habe keine Einwände, dass das Thema meiner <Bachelor-/Master-/Seminararbeit> mit meiner Namensnennung in einem Verzeichnis der Hochschule genannt wird.

Ort/Datum: .......................................... Unterschrift: ..........................................

**Von der Fakultät Wirtschaftswissenschaften der Universität Paderborn zu unterschreiben:**

Aus den oben genannten Gründen verpflichtet sich die Fakultät entsprechend, Veröffentlichungen der <Bachelor-/Master-/Seminararbeit> oder ihres wesentlichen Inhalts nicht vorzunehmen und die <Bachelor-/Master-/Seminararbeit> auch nicht zur Einsicht und Ausleihe in die Bibliothek aufzunehmen.

Ort/Datum: .......................................... Unterschrift: ..........................................

**Vom Betreuer der <Firmenname> zu unterschreiben:**

Oben genannter Sperrvermerk zur Kenntnis genommen.

Ort/Datum: .......................................... Unterschrift: .........................................

**Eidesstattliche Erklärung:**

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig, ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken, Tabellen, Skizzen, Zeichnungen, bildliche Darstellungen usw. sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

Paderborn, <Abgabedatum eintragen>

|  |
| --- |
| (Vorname Nachname) |

Abstract – deutsch

<Hier den Text der Zusammenfassung in deutscher Sprache eintragen>

Stichworte: <Stichworte in deutscher Sprache hier eintragen>

Abstract – englisch

<Hier den Text der Zusammenfassung in englischer Sprache eintragen>

Keywords: <Stichworte in englischer Sprache hier eintragen>

{Diese Formatvorlage ist für den einseitigen Druck gemacht, kann aber durch eigene Anpassungen auch für den zweiseitigen Druck benutzt werden.}

{Die Verwendung der Formatvorlage entbindet nicht von der Pflicht, sich eigene Gedanken über Aufbau und Aussehen der schriftlichen Ausarbeitung zu machen, sowie die Aktualität und Richtigkeit zu prüfen. Die Vorlage ist nur eine Hilfestellung. Insbesondere sind nicht benötigte Teile der Vorlage zu entfernen, sowie noch zusätzlich benötigte Teile hinzuzufügen.

Der gewählte Zitationsstil ist APA, 6th Edition.}

Inhaltsverzeichnis

[Abkürzungsverzeichnis III](#_Toc63501266)

[Abbildungsverzeichnis IV](#_Toc63501267)

[Tabellenverzeichnis 5](#_Toc63501268)

[1 Business Understanding 1](#_Toc63501269)

[1.1 Problematik 1](#_Toc63501270)

[1.2 Zielsetzung 1](#_Toc63501271)

[1.3 Vorgehensweise 1](#_Toc63501272)

[2 Data Understanding 2](#_Toc63501273)

[2.1 Collect Initial Data 2](#_Toc63501274)

[2.2 Explore Data 2](#_Toc63501275)

[2.3 Verify Data Quality 2](#_Toc63501276)

[2.4 Data Interpretation 4](#_Toc63501277)

[3 Data Preparation 6](#_Toc63501278)

[3.1 Select Data 6](#_Toc63501279)

[3.2 Clean Data 6](#_Toc63501280)

[4 Modeling 8](#_Toc63501281)

[4.1 Select Model Technique 8](#_Toc63501282)

[4.2 Build Model 8](#_Toc63501283)

[5 Evaluation 9](#_Toc63501284)

[6 Deployment 10](#_Toc63501285)

[Literaturverzeichnis 11](#_Toc63501286)

Anhang

[A1 <Titel von Kapitel A1 eintragen> A-1](#_Toc113163008)

[A1.1 <Titel von Kapitel A1.1 eintragen> A-1](#_Toc113163009)

[A2 <Titel von Kapitel A2 eintragen> A-3](#_Toc113163010)

[A2.1 <Titel von Kapitel A2.1 eintragen> A-3](#_Toc113163011)

[A3 <Titel von Kapitel A3 eintragen> A-5](#_Toc113163012)

[A3.1 <Titel von Kapitel A3.1 eintragen> A-5](#_Toc113163013)

# Abkürzungsverzeichnis

MDVSP Multi Depot Vehicle Scheduling Problem

# Abbildungsverzeichnis

**Es konnten keine Einträge für ein Abbildungsverzeichnis gefunden werden.**

# Tabellenverzeichnis

**Es konnten keine Einträge für ein Abbildungsverzeichnis gefunden werden.**

# Business Understanding

## Problematik

Der Gebrauchtwagenmarkt ist ein gutes Beispiel für asymmetrische Informationsketten, da der Käufer vor Vertragsabschluss die Qualität des Produktes nicht hundertprozentig kennt. Hinzu kommt, dass bei Auktionen zwischen zwei Bewertungskriterien unterschieden wird: defensive beziehungsweise offensive Bewertung. Die defensive Bewertung geschieht durch Mechaniker und Tester, welche alle Fehler oder Schwachstellen finden möchten. Die offensive Bewertung erfolgt durch das Marketing- und Verkaufspersonal. (siehe Quelle 6)

Trotz der defensiven Bewertungen besteht für den Käufer bei Auktionen von Gebrauchtwagen immer das Risiko eines Fehlkaufes. Daher besteht einer der größten Herausforderungen eines Käufers (z.B. Autohaus) das Risiko, dass das Fahrzeug schwerwiegende Probleme hat, zu verhindern. Der Fehlkauf kann dazu führen, dass die Reparaturkosten den Wert des Fahrzeuges überschreiten und somit ein Weiterverkauf an einen Kunden ausgeschlossen ist.

Ein Fehlkauf wird in Amerika als „lemon“ bezeichnet und insbesondere im Zusammenhang mit dem Gebrauchtwagenmarkt häufig erwähnt und analysiert. Genauer handelt es sich um ein Fahrzeug, das nach dem Kauf als Fehlkauf bewertet wird. Ein Fehlkauf ist genau dann vorhanden, wenn der tatsächliche Zustand des Fahrzeuges wegen zu reparierender Mängel nicht dem Kaufpreis gerecht wird. Genauer bewertet Eric W. Bond ein Fahrzeug als Fehlkauf, wenn innerhalb von zwölf Monaten ab Kauf eine große Reparatur in den Kategorien Motor, Getriebe, Bremsen oder Achsen erforderlich ist.

Durch die oben genannten Fehlkäufe fallen für den Autohändler Kosten an, die das Geschäft des Händlers negativ beeinflussen und zu einem unattraktiven Bestandsangebot für Kunden des Autohändlers führen.

## Zielsetzung

Das Ziel besteht darin, das Risiko eines Fehlkaufes zu verringern. Käufer\*innen sollen somit die „Kicks“ vermeiden und den Bestand an qualitativen Fahrzeugen erhöhen können.

## Vorgehensweise

Um das oben beschriebene Ziel erreichen zu können, werden die im folgenden Absatz erklärten Kennzahlen des Datensatzes benutzt, um anhand dieser eine Vorhersage über den Kauf zu machen. Dabei steht „0“ für keinen Fehlkauf und die „1“ repräsentiert ein hohes Risiko eines fehlerhaften Fahrzeuges.

# Data Understanding

In diesem Kapitel wird der Datensatz analysiert, auf Zusammenhänge untersucht und mögliche Probleme werden herausgestellt. Der Datensatz wurde in Zusammenhang mit der Competition auf Kaggle durch die Firma „Carvana“ veröffentlich.

## Collect Initial Data

Der Datensatz wurde auf der Plattform Kaggle von dem Unternehmen „Carvana“ zur Verfügung gestellt und umfasst sechs Dateien. Nach dem Download aller Dateien in Form eines .zip Archives wurden die Daten entpackt.

Da die Daten nicht selbst erhoben wurden, umfasst der Datensatz eine Beschreibung aller Felder.

## Explore Data

Wie aus der Datei "Carvana\_Data\_Dictionary.txt" zu erkennen ist, bestehen die Datensätze aus 36 Spalten. Davon ist ein Attribut die ID (RefID) und eins die Bewertung, ob es ein schlechter Kauf war, oder nicht. Die weiteren Attribute beschreiben den Kauf des Fahrzeuges, also wann, wo und für wie viel es gekauft wurde.

Zwölf der 36 Attribute geben Werte über das Fahrzeug an sich an. Darunter fallen allgemeinere Daten wie das Baujahr, die Marke, das Modell und die Farbe. Darüber hinaus umfasst der Datensatz aber auch detailliertere Informationen über z.B. den Kilometerstand, das Getriebe und Ausstattungslevel.

Weitere acht Attribute beziehen sich auf die aktuelle Preislage des Fahrzeuges. Diese Preise wurden von dem Manheim Market Report erhoben und sind einige der genauesten Richtwerte für einen Gebrauchtwagenpreis. (siehe Quelle 8)

Die letzten Attribute geben Angaben darüber, wo das Fahrzeug gekauft wurde, wie hoch die Gewährleistung war und wann es zurück zur Auktion „gekickt“ wurde.

Diese Spalten finden sich in den Dateien „training.csv“, „test.csv“ und „example\_entry.csv“ wieder. Die mit Abstand größte Datei ist „training.csv“ mit 72983 Zeilen. Danach folgt die Datei „test.csv“ mit 48707 Zeilen und danach „example\_entry.csv“ mit ebenfalls 48707 Zeilen.

## Verify Data Quality

Betrachtet wird zunächst der Trainingsdatensatz mit 72983 Zeilen und 34 Spalten. Im Vergleich zu allen beschriebenen Spalten fehlen hier die Attribute „AcquisitionType“ und „KickDate“.

Im ersten Schritt, um die Daten richtig interpretieren zu können, wurden die Einträge, die keine Preisangabe hatten, ausgefiltert. Dies betrag 315 Datensätze und ist aufgrund der erheblichen Menge an Datensätzen zu vernachlässigen.

In dem Summary des Datensatzes fällt auf, dass das vorherzusagende Feld „IsBadBuy“ einen Mittelwert von 0.123 hat. Da es sich hierbei um ein Binärattribut handelt, besteht die Vermutung, dass der Trainingsdatensatz im Hinblick auf dieses Attribut nicht balanciert ist. Ein Balkendiagramm der gruppierten Daten bestätigt die Vermutung und zeigt, dass lediglich 8944 der insgesamt 72668 Einträge einen schlechten Kauf beinhalten. Lediglich ca. zwölf Prozent sind demnach schlechte Käufe.

Das Ungleichgewicht kann bei späteren Vorhersagen zu einer sehr schlechten Performance führen und muss im folgenden Kapitel zur Data Preparation beachtet werden.

Durch die im Notebook erstellte Summary des gesamten Datensatzes konnte festgestellt werden, dass die Attribute VehYear, VehicleAge, VehOdo gleichverteilt sind. Die Attribute, die den Preis beschreiben sind teilweise gleichverteilt und teilweise verschoben.

Zusammenhänge innerhalb des Datensatzes lassen sich mit Hilfe einer Korrelationsmatrix feststellen. Dazu wurden die relevanten, numerischen Werte ausgewählt und mit einer Farbskala versehen.

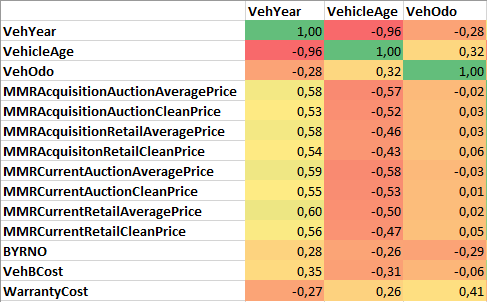


Abbildung 1: Ausschnitt Korrelationsmatrix mit Farbskala

Zwischen den Attributen VehYear und VehOdo besteht eine erhöhte Korrelation, was in Anbetracht des Zusammenhangs plausibel erscheint. Was auffällt ist, dass die Kosten weniger durch die gebrachte Laufleistung, sondern durch das Fahrzeugalter beeinflusst werden. Die Kosten für die Garantie hängen außerdem eher von dem Produktionsjahr des Fahrzeuges ab.

## Data Interpretation

Bezüglich der Interpretation der Daten besteht die Vermutung, dass im zunehmenden Alter der Fahrzeuge der Anteil der Fehlkäufe ansteigt. Um dies zu prüfen wurde das Fahrzeugalter auf der x – Achse und die Anzahl der Fahrzeuge, aufgeteilt nach Fehlkauf oder nicht, auf der y – Achse dargestellt.

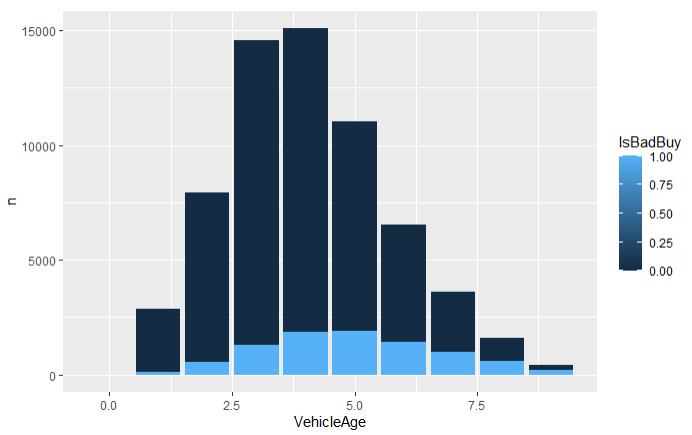


Abbildung 2: Anteil Good/Bad Buy im Fahrzeugalter

Die Grafik stützt die Vermutung, da der Anteil der Fehlkäufe immer weiter zu nimmt. Weiter verdeutlicht wird dieser Zusammenhang in folgender Grafik, in der auf der x – Achse wieder das Fahrzeugalter und auf der y – Achse der prozentuale Anteil der Fehlkäufe dargestellt ist.

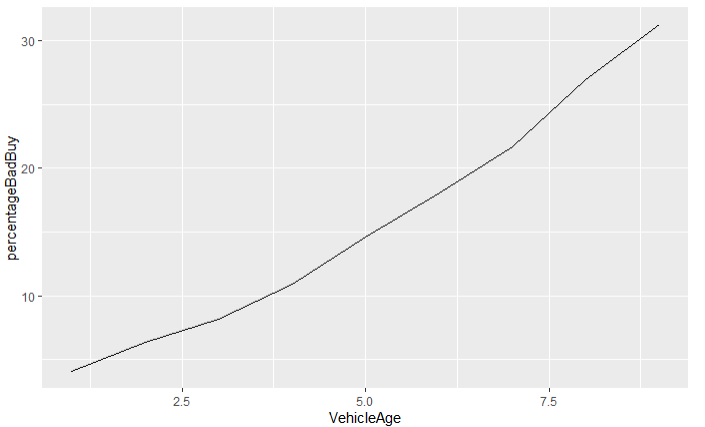
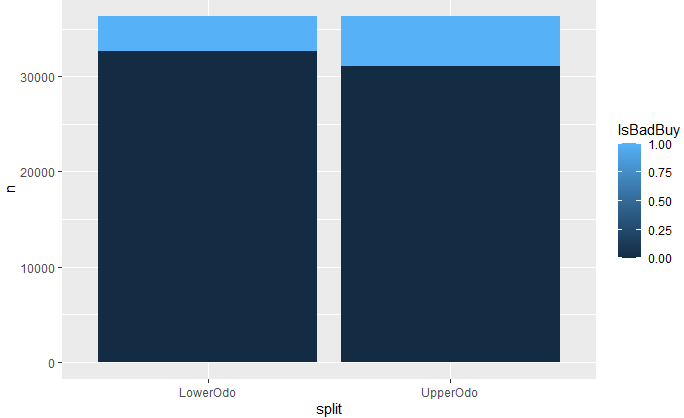


Abbildung 3: Verhältnis Good/Bad Buy im Fahrzeugalter

Weiter wurde die erbrachte Laufleistung im Zusammenhang mit den Fehlkäufen untersucht. Hierbei wurde zwar ein etwas erhöhter Anteil bei höherer Laufleistung festgestellt, im Hinblick auf den geringen Unterschied ist dies aber eher weniger aussagekräftig.

Abbildung 4: Verhältnis Good/Bad Buy bei unterschiedlicher Laufleistung

# Data Preparation

In diesem Abschnitt werden die Daten so vorbereitet, dass sie in der Modeling Phase benutzt werden können. Der erste Schritt ist die Wahl der relevanten Attribute für das Modell.

## Select Data

Die ausgewählten Daten sollen den größten Nutzen im Hinblick auf unser Modell geben. Daher können zunächst die ID’s entfernt werden, also die RefID, WheelTypeID und die BYRNO. Ebenso fällt das Attribut PRIMEUNIT heraus, da nur weniger als 4000 Daten vorhanden sind. Auch ist die Anzahl bei dem Attribut AUCGUART zu gering, damit es aussagekräftig sein kann. Das Attribut VNZIP1 wurde ebenfalls entfernt, da es sehr unwahrscheinlich ist, dass Fehlkäufe abhängig von nur einem einzigen Postleitzahlengebiet sind.

## Clean Data

Zunächst werden alle Attribute nach „NA“ Einträgen durchsucht. Es fällt die Spalte TRIM auf, in der es 2358 fehlende Werte gibt. Da es sich um ein kategorisches Attribut handelt, kann kein Durchschnittswert eingefügt werden. Deshalb werden die Reihen mit den fehlenden Werten aus dem Datensatz entfernt.

Außerdem befinden sich unter den importierten Daten kategorische Werte mit dem Eintrag „NULL“, welche auch entfernt werden.

Weiter werden die numerischen Attribute betrachtet. Hier werden fehlende Werte (= 0) mit dem Durchschnitt des jeweiligen Attributes ergänzt.

Die Attribute „Model“ und „SubModel“ enthalten noch weitere Informationen, die über die eigentliche Modellbezeichnung hinaus gehen. Deswegen wurden die Informationen aus den Feldern in neue Attribute unterteilt und dem Data Frame hinzugefügt. Die so entstandenen Attribute sind „Doors“, „Cylinder“, „Liter“ und „WheelDrive“. Die so entstandenen Attribute besitzen viele fehlende Werte, sodass sie durch „Unknown“ ersetzt werden.

Im nächsten Schritt wurde der Datensatz mit Hilfe von Over- und Undersampling balanciert, sodass die Zielvariable „IsBadBuy“ gleich verteilt ist. Der daraus entstandene Datensatz beinhaltet insgesamt ca. 18000 Einträge, die in etwa zu gleicher Anzahl aus je Fehlkäufen und nicht Fehlkäufen bestehen. Damit wurden mehr Datensätze von erfolgreichen Käufen entfernt als hinzugefügt, um den Datensatz so weit es geht unverändert zu lassen.

Hierzu wurde das Paket ROSE verwendet.

Nachdem die Daten nun balanciert und angepasst wurden, beinhaltet der Datensatz 12890 Fahrzeugdaten, die 6433 Fehlkäufe (Lemons) beinhalten. Mit diesen Daten kann nun im nächsten Abschnitt das Modell entwickelt werden.

# Modeling

In diesem Abschnitt wird das Model zur Vorhersage der Zielvariable erarbeitet.

## Select Model Technique

Als Model Technik wird Random Forest aus dem Bereich des Supervised Learning gewählt. Das Klassifizierungs- bzw. Regressionsverfahren bietet die Vorteile wenig Rechenleistung auch bei größeren Datensätzen in Anspruch zu nehmen. Weiter kann das Verfahren mit einem Datensatz umgehen, der viele Merkmale besitzt. Da mir nur begrenzte Rechenkapazität zur Verfügung steht und der Datensatz aus vielen Merkmalen besteht, wird im Weiteren das Model mit oben genannten Verfahren entwickelt.

## Build Model

Um das Model zu entwickeln, wird das Paket „ranger“ installiert und importiert. Mit diesem Paket kann der Random Forest Algorithmus nach Breimann implementiert werden.

Da nach der Data Preparation der balancierte und bereinigte Datensatz vorliegt, kann mit dem Erstellen eines Modells begonnen werden. Dafür wird der Trainingsdatensatz in 80% Trainingsdaten und 20% Validierungsdaten aufgeteilt. Das Model wird auf den Trainingsdaten trainiert und hinterher mit den Validierungsdaten bewertet.

Um eine bestmögliche Vorhersage ermöglichen zu können, wird mit Hilfe von dem Paket „tuneRanger“ zunächst ein Klassifizierungstask angelegt. Mit Hilfe von diesem Task werden die Hyper Parameter „mtry“, „min.node.size“ und „sample.fraction“ getuned. Dabei werden 800 Bäume pro Durchgang erstellt und mit der Bewertungsmethode „Area Under The Curve“ bewertet. Es werden 30 „warm-up“ und 250 normale Durchgänge durchgeführt, um die bestmöglichen Parameter für das Modell zu berechnen. Die berechneten Parameter sehen wie folgt aus:

ABBILDUNG tuneRanger Ergebnisse

Die berechneten Parameter sind in dem Model „res$model$learner.model“ bereits hinterlegt und können direkt zur Vorhersage von Werten benutzt werden. Die Vorhersage erfolgt vom Typ „response“, sodass ein Wahrscheinlichkeitswert für die Vorhersage erstellt wird. In diesem Fall werden alle Werte > 0.5 zu einer 1 und damit einem Fehlkauf zugeordnet. Alle Werte <= 0.5 werden als kein Fehlkauf vorhergesagt.

# Evaluation

# Deployment

# Literaturverzeichnis

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Anhang

1. Bei Wegfall des Geheimhaltungserfordernisses ergeht eine gesonderte Information an Verfasser und Fakultät. [↑](#footnote-ref-1)