Contributions to Data Analytics Techniques with Applications in Forecasting, Visualization and Decision Support

Der Wirtschaftswissenschaftlichen Fakultät der

Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften

- Doctor rerum politicarum -

vorgelegte Dissertation

von

Dennis Eilers, M.Sc.

[redacted]

2018

Betreuer und Gutachter:	Prof. Dr. Michael H. Breitner
Weiterer Gutachter:	Prof. Dr. Hans-Jörg von Mettenheim
Vorsitzender der Prüfungskommission:	Prof. Dr. Stefan Wielenberg
Weiteres Mitglied (beratend):	Dr. Michelle Muraz

Für meine Oma.
In Erinnerung, dass bei allem was wir tun, die Liebe das Wichtigste im Leben ist.

Danksagung

Begegnungen sammelt man auf seinem Weg mit vielen Menschen. Die Richtung allerdings prägen nur Wenige. Den Menschen die meinen Weg geprägt haben, möchte ich an dieser Stelle Danke sagen.

Einer dieser Menschen ist mein Doktorvater Prof. Dr. Michael H. Breitner. Sie haben mir nach dem ersten Semester die Möglichkeit gegeben bei Ihnen als Studentische Hilfskraft zu arbeiten und mir so frühzeitig einen Blick hinter die Kulissen der akademischen Welt gewährt. Für das Vertrauen bin ich Ihnen sehr dankbar. In den vergangenen Jahren habe ich immer versucht etwas davon zurückzugeben, auch wenn ich zuweilen sicher ein unbequemer Student und Doktorand war, der gerne die fachlich kritische Auseinandersetzung gesucht hat. Ihre Geduld, uneingeschränkte Förderung aber auch klar formulierte Kritik werden mir im Gedächtnis bleiben. Jeder dieser drei Punkte hat mich gleichermaßen vorangebracht.

Einen akademischen Mentor, Förderer und guten Freund habe ich während meiner Zeit am Institut in Prof. Dr. Hans-Jörg von Mettenheim gefunden. Bei dir hatte ich die Freiheit meine Ideen umsetzen zu können und gleichermaßen Leitplanken, die meine Arbeit auf den richtigen Weg gebracht haben. Unsere vielen Gespräche waren für mich eine sowohl fachliche als auch persönliche Bereicherung, die ich nicht mehr missen möchte. Vielen Dank!

Über den Satz "Na, bist du auch am IWI HIWI" von Daniel Olivotti freue ich heute noch besonders, denn als Freund, Leidensgenosse, Reisehelfer, Co-Autor, Küchenbauer, Gesprächspartner, Berater und nebenbei Kollege hast du mir nicht nur mit deinem Eis das Leben süßer gemacht. ;-) Für deine Unterstützung ein großes Dankeschön! Von der Tapete bis zur Wand und darüber hinausdenken, konnte ich am besten mit meinem Bürogenossen Rouven Wiegard. Danke für eine tolle Zeit und eine super Freundschaft! Mit Jean-Henrick Schünemann im Büro waren dann die großen Fragen der Menschheit kein Problem mehr. Als Freund, Bürogenosse und philosophischer Sparringspartner unschlagbar. Danke Joi! Bei meinen bisher noch nicht genannten Co-Autoren Christian L. Dunis, Dennis Gercke, Christoph Gleue und Cornelius Köpp möchte ich mich ebenfalls ganz herzlich bedanken! Bei allen Kollegen die ich im Laufe der Zeit kennenlernen durfte bedanke ich mich genauso herzlich! Dankbar bin ich auch für die vielen Studenten mit denen ich in Vorlesungen, Übungen, Seminaren und Abschlussarbeiten zusammenarbeiten durfte. Eure Ideen und Feedback haben mich angetrieben auch immer mein Bestes in der Lehre und Betreuung zu geben!

Und abschließend möchte ich mich bei den Menschen bedanken, die mir privat die Kraft und Unterstützung haben zukommen lassen, ohne die eine solche Arbeit nicht möglich ist. Allen voran meine Eltern, die mir immer die Sicherheit gegeben haben, die ich brauchte. Und bei meinen Freunden, die mich unterstützt und ausgehalten haben.

Abstract

This cumulative dissertation summarizes and critically discusses seven peer-reviewed publications where I was involved as a co-author. All publications contribute to data analytics techniques. The dissertation consists of four main sections.

(1) Machine Leaning in Finance: In this section a Decision Support Algorithm based in Reinforcement Learning is introduced which filters rule-based trading decisions. We contribute to the literature by describing the implementation of the algorithm. We also provide empirical evidence of financial market anomalies.

(2) Mining Customer Reviews: Opinions from customers about certain products are more and more expressed on social media platforms. Here we provide the first study which analyses YouTube comments as a data source for an aspect-based Sentiment Analysis. We also contribute to the literature by proposing a filtering method based on Google Trends which sorts product aspects according to their relevance for the customers.

(3) Forecasting Resale Prices of Used Cars: In this section we show how to efficiently forecast resale prices of used cars with Artificial Neural Networks. We provide lessons learned about long-term forecasts. We also provide insights in the importance of certain independent factors which determine the resale price.

(4) Visual Model Evaluation: The research in this section is mainly driven by the question of how to better incorporate human domain knowledge in data science. We develop a visualization technique based on heat maps which provides a more intuitive view on errors of a machine learning model. The visualization technique allows domain experts to discuss the results of machine learning models with data science experts on the same level of complexity.

Keywords: Reinforcement Learning, Artificial Neural Networks, Sentiment Analysis, Leasing, Used Cars, Feature Engineering, Domain Knowledge, Visualization.

Zusammenfassung

Diese kumulative Dissertation fast sieben von mir mitverfasste peer-reviewed Publikationen zusammen und diskutiert diese kritisch. Alle Publikationen leisten einen Beitrag zu Data Analytics Techniken. Die Dissertation ist in vier Hauptbereich eingeteilt.

(1) Maschinelles Lernen in Finance: In diesem Kapitel wird ein Entscheidungsunterstützungsalgorithmus basierend auf Bestärkendem Lernen vorgestellt, welcher regelbasierte Tradingentscheidungen filtert. Wir leisten einen Beitrag zur Literatur indem wir die Implementierung des Algorithmus beschreiben und zusätzlich empirische Hinweise auf Finanzmarktanomalien liefern.

(2) Mining von Kundenrezensionen: Meinungen von Kunden zu einem bestimmten Produkt, werden zunehmend auf sozialen Netzwerken gepostet. Wir liefern die erste Studie, die YouTube Kommentare als Datenquelle für eine aspektbasierte Sentiment Analyse untersucht. Wir leisten zudem einen Beitrag zur Literatur, durch die Entwicklung einer Filtermethode basierend auf Google Trends zur Sortierung von Produkteigenschaften in Abhängigkeit ihrer Relevanz für die Kunden.

(3) Vorhersage der Wiederverkaufspreise von Gebrauchtfahrzeugen: In diesem Kapitel zeigen wir, wie die Wiederverkaufspreise von Gebrauchtwagen mit Künstlichen Neuronalen Netzen effizient prognostiziert werden können. Wir liefern Erkenntnisse über langfristige Prognosen. Wir geben auch einen Einblick in die Bedeutung bestimmter unabhängiger Faktoren, die den Wiederverkaufspreis bestimmen.

(4) Visuelle Modellevaluation: Die Forschung in diesem Abschnitt wird hauptsächlich von der Frage geleitet, wie menschliches Domänenwissen besser in Data Science Projekte integriert werden kann. Wir entwickeln dazu eine Visualisierungstechnik basierend auf Heatmaps, die eine intuitivere Sicht auf Fehler eines maschinellen Lernmodells bietet. Die Visualisierungstechnik ermöglicht Domänenexperten, die Ergebnisse von Modellen des maschinellen Lernens mit Datenexperten auf der gleichen Komplexitätsebene zu diskutieren.

Schlüsselwörter: Bestärkendes Lernen, Künstliche Neuronale Netze, Sentiment Analyse, Leasing, Gebrauchtfahrzeuge, Feature Engineering, Domain Knowledge, Visualisierung.

Management Summary

Analyzing data for example to make better decisions, find hidden insights or improve customer satisfaction is one of the major challenges in modern, digitized industries. Almost every part of today's economy is affected by the possibilities of data analytics. Hence, there exist a tremendous interest in this topic, both in research and practice. This cumulative dissertation contributes to four research streams within the broad topic of data analytics.

1. Machine Learning in Finance
2. Mining Customer Reviews
3. Forecasting Resale Prices
4. Visual Model Evaluation

(1) The first part of this dissertation (Section 2) empirically investigates the existence of calendar anomalies and empirical regularities on financial markets and propose a new decision support algorithm based on Reinforcement Learning (RL) which allows to build an intelligent trading system which self-learns and adapts to new market situations. The findings from the empirical part of the paper show some weak evidence that certain anomalies like the new year and Easter effect exist in major stock indices. But the results also show that exploiting these anomalies by a simple trading strategy is hardly possible. To improve trading decisions, the main part of the publication proposes a new algorithm which analyzes the market based on certain characteristics like past returns and technical indicators and provides recommendations of how to position oneself. The basis for this algorithm is the idea of RL which is inspired by the learning process of humans. By rewarding good decisions and punishing bad decisions, the algorithm learns how to act such that the own actions lead to the highest expected reward. The process is visualized in Figure 1.

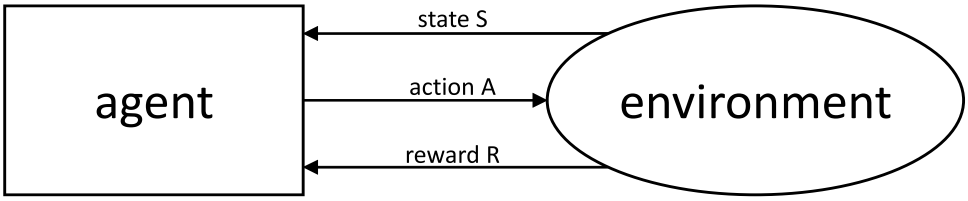


Figure 1: Idea of Reinforcement Learning

The results show that the trading strategy which is supported by RL, outperforms buy & hold as well as the naive anomaly strategies. Figure 2 provides an illustration of how the algorithm works in practice.

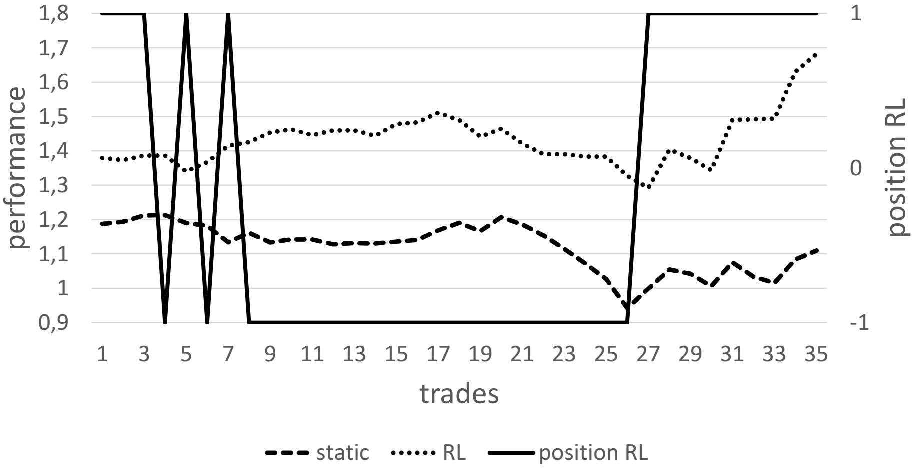


Figure 2: Trading Performance Improvements with Reinforcement Learning (RL)

The solid line represents the position of the trading strategy. A 1 indicate a long strategy, a -1 a short strategy and 0 a neutral position. The two dashed lines compare the return development of a static strategy which simply follow the anomaly exploitation rules from the literature and a RL strategy which is also based on the same rule but filters the results based on past experience. It can be seen that higher returns and also better reward to risk ratios (smaller maximum drawdowns) can be achieved by applying the RL filter.

(2) The next study (Section 3) contributes to the text mining literature. In e-commerce it becomes more and more important to understand what customers think about products or services to improve their experience. Opinions are often available in text form like conventional online reviews but also from an increasing number of social media data where users post their experience. Automatically analyzing the

unstructured text data in the context of product evaluations is known as Sentiment Analysis (SA) or Opinion Mining. SA is divided into several subtasks. Figure 3 illustrates the structure.

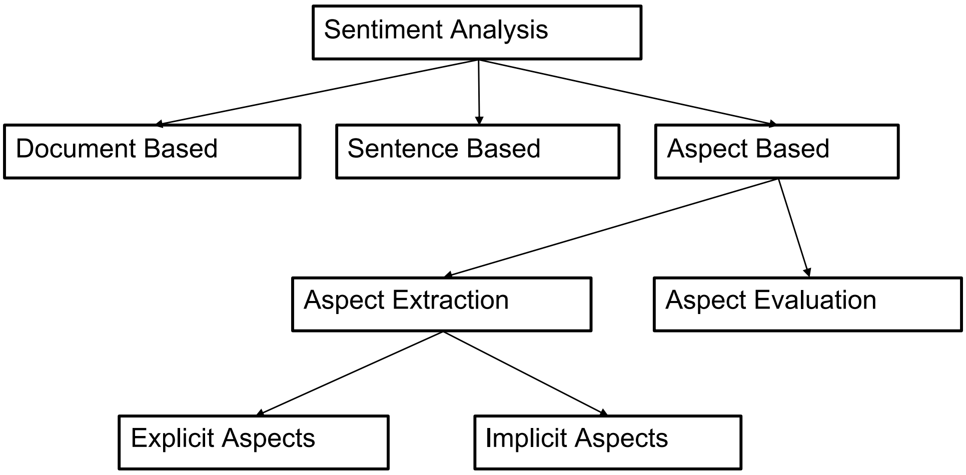


Figure 3: Structure of Sentiment Analysis

While document based and sentence based analysis try to extract the overall opinion about a product from a text/sentence, the aspect based analysis tries to extract opinions regarding certain important aspect of a product like the camera of a smartphone. This task is again divided into aspect extraction and aspects evaluation tasks. We are focusing on the aspect extraction task in our research which aims to find the aspect which are discussed in the text. Our first contribution to the literature is a comparison between conventional review data from Amazon reviews and noisier social media data from YouTube. We provide the first study which explicitly identifies YouTube comments to product related videos as a suitable source of information for an aspect based SA. Our comparisons show that a standard aspect based sentiment algorithm performs equally well on Amazon reviews and YouTube comments. Our second contribution is a filter mechanism which incorporates information from Google Trends about the search volumes of products in conjunction with their aspects. The assumption is that customers tend to search for products in conjunction with important aspects. Filtering potential aspects based on their search volume further increase the aspect extraction results. One particular problem remains namely the extraction of implicit aspects. There exist a difference between explicit aspect like a camera for a smartphone and implicit aspect like the weight of the smartphone which is only indirectly mentioned by adjectives like heavy and light.

(3) The third part of this dissertation (Section 4) discusses the findings from a research project with the goal to forecast resale prices of used cars. The initial situation is that a large car manufacturer faces the challenge of setting proper leasing rates for their cars. The main determinant for the leasing rate is the expected resale price of the car at the end of the contract. To improve the situation the car manufacturer started to collect all available contract data and merge them with the realized price on the used car market of the corresponding car. In the collaborative project we develop a forecasting model based on Artificial Neural Networks (ANN), which learns the dependencies. Figure 4 provides an overview of the implemented Decision Support System (DSS).

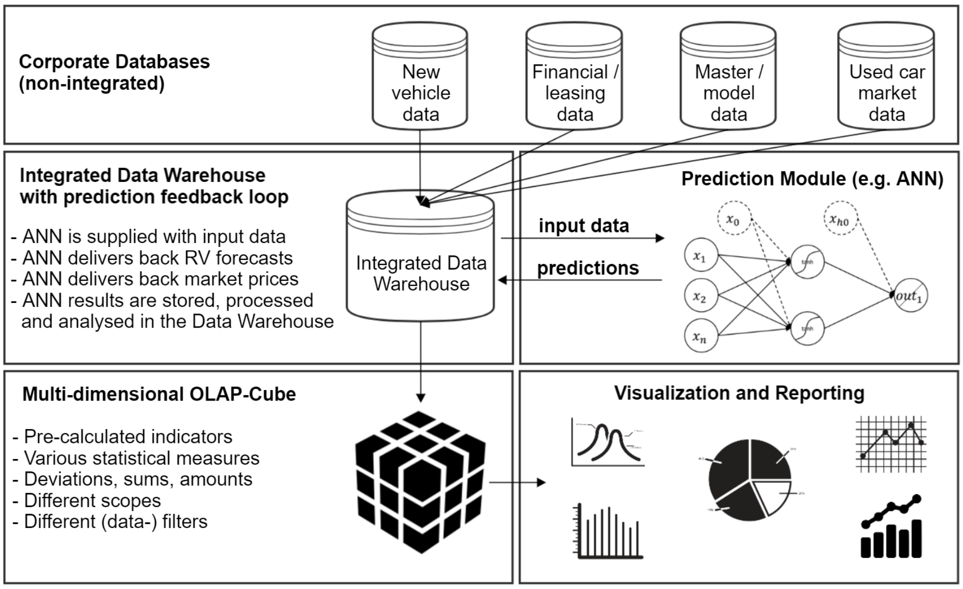

Corporate Databases (non-integrated)
New vehicle data
Financial / leasing data
Master / model data
Used car market data
Integrated Data Warehouse with prediction feedback loop
- ANN is supplied with input data
- ANN delivers back RV forecasts
- ANN delivers back market prices
- ANN results are stored, processed and analysed in the Data Warehouse
Integrated Data Warehouse
input data
predictions
Prediction Module (e.g. ANN)
x_0
x_{h0}
x_1
x_2
x_n
out_1
Multi-dimensional OLAP-Cube
- Pre-calculated indicators
- Various statistical measures
- Deviations, sums, amounts
- Different scopes
- Different (data-) filters
Visualization and Reporting

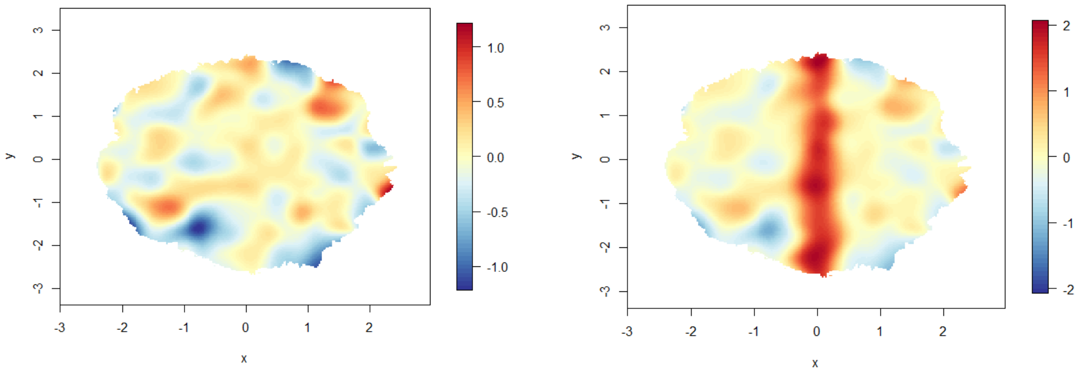
Figure 4: Decision Support System to Forecast Resale Prices

The contribution of this part is threefold. First, we show how the DSS is designed and implemented. Second, we investigate the influence/predictive power of the available independent variables. And third, we provide recommendations about how to design the forecasting application with ANNs while facing challenges like long-term forecast horizons, noisy data and time dependent variables.

The empirical findings show that the influence of external variables like oil price, stock index or consumer index are negligible. After controlling for all available information about the car and the contract specifications, none of the external factors show significant explanatory (linear mixed effects models and likelihood ratio test), nor predictive power (perturbation ranking algorithm). Another finding is the influence of the time factor on the forecasting performance. Our unique data set show that resale prices are subject to seasonal patterns and a trend. Again after controlling for all available information, the constructed time series of regression residuals are used for a time series decomposition which reveal higher resale prices during spring and lower resale prices during the end of the year. The implemented forecasting application then shows that ANNs with the capability of incorporating non-linear dependencies, outperform ridge linear regression models. The problem is to incorporate the time dependent variables. One possible solution is to combine ANNs to incorporate the non-linear dependencies in the data and incorporating the time dependent factors by a linear adjustment which provides a one and a half year unbiased out-of-sample performance in our tests.

(4) The forth part of this dissertation (Section 5) addresses the problem of the communication and understanding gap between data science experts and domain experts. In real world industry applications, the complex machine learning and data analytics task are carried out by highly skilled experts with a quantitative background. The problem is that these experts not necessarily have the required domain expertise to incorporate all relevant aspect and dependencies into the model. The result can be models which are biased or may lead to wrong conclusions in the worst case. On the other hand, domain experts who have worked many years in their field often lag understanding of the complex machine learning models which can result in mistrust and lack of acceptance. The consequence of both cases can be a non-optimal decision-making process. Therefore, our goal is to address this issue by proposing a new visualization technique for regression model performance based on heat maps. Heat maps are a familiar visualization tool for managers/decision

makers (domain experts) and data scientist alike. The idea is to use an intuitive visualization technique to enable a better communication between data scientist and domain experts about the current results of the model. With the help of this discussion, one can identify possible new independent variables (features) which might be previously overlooked. Figure 5 provides an impression of how the heat map visualization can provide an overview of the model performance in different regions of a data space.



(a) White Noise

(b) Bias

Figure 5: Heat Map Visualization of Model Performance

The heat map shows the error distribution (color scale) of a machine learning model depending on two different features. On the left-hand side (Figure 5a), we see a white noise with no clear patterns in the data. This is what can be expected if the models work reasonably well and no patterns remain in the data which are not explained by the model. On the right-hand side (Figure 5b), we see an example of a biased model which can be the result of a missing variable that is not properly reflected in the model specification.

The idea of model building, discussing and adjustment is an iterative process which should facilitate the so-called Feature Engineering (FE), a process of constructing a proper input space for a machine learning model. One example where the visualization technique is already successfully applied is the resale price forecast application from the previous part of the dissertation. Here we were able to identify three additional features which have improved the forecasting performance of the models in economically relevant magnitudes.

Contents