Data Analytics: A Case Study

Masterarbeit

zur Erlangung des akademischen Grades „Master of Science (M. Sc.)“ im Studiengang Wirtschaftswissenschaft der Wirtschaftswissenschaftlichen Fakultät der Leibniz Universität Hannover

vorgelegt von

Dennis Eilers
[redacted]

Prüfer: Prof. Dr. Michael H. Breitner

Hannover, den 26. September 2016

Contents

1 Introduction and Motivation

In God we trust, all others bring data.
– William Edwards Deming

The success of a business is mainly driven by management decisions. A variety of research has been carried out to support these decisions. Since the 1970 so called Decision Support Systems (DSS) became popular in research and practice alike [Power, 2007]. With the increasing amount of available data about the business itself, from external sources, social media and user data, these systems more and more use this potential. Terms like Big Data[1], Business Intelligence[2] (BI) and Data Analytics/Science[3] emerged and changed the way of thinking about value oriented data analysis. Since just collecting and storing data generates no value for businesses, each data driven system has to have a data analytics core to process and analyses the data and finally derive valuable information from it. In this thesis I focus on methods and applications for the data analytics task in such systems. Therefore, three cases from real world problems are investigated based on a typical data analytics pipeline (section 5).

To motivate the first case, let's consider a company which introduces a new product on the market. A typical example could be a new coffee flavor created by Starbucks. A traditional approach to investigate the success of this new product is to analyze the sales data after a certain time and conduct surveys with customers about what they like and what they don't like. These procedure has a crucial disadvantage: Time. The results of the analysis are first available after weeks or even months since the market launch. A different approach uses web mining and social analytics. At the first day the new coffee is introduced, the company can collect data from social media like Facebook, Twitter or Instagram and analyze the sentiment of people who have bought the coffee and share their thoughts and feelings on the web. Such methods are able to analyze the sentiment about the product itself but also about specific aspects. A result after a few hours of data collection could be that people very much like the taste (one aspect of the product with good ratings) but are complaining about the high price (another aspect of the product with bad ratings) compared to other coffee blends. With that information, the company can take countermeasures, in this case lower the price and observe the reactions of the customers in nearly real time [Watson, 2014]. This stylized example motivates the relevance of the topic. In my study I illustrate data collection methods and algorithms to perform such an aspect based Sentiment Analysis (SA) with social media data. In view of the available data, I focus on analyzing sentiments of YouTube comments related

[1] described by the 4 V's Volume, Velocity, Variety and Veracity [Goes, 2014]

[2] "a broad category of applications, technologies, and processes for gathering, storing, accessing, and analyzing data to help business users make better decisions" [Watson, 2009, p. 491]

[3] an umbrella term for data analysis applications or the "getting data out" part of BI [Watson, 2014]

to three comparable smartphones. This study contributes to the literature in two ways. First, an aspect extraction algorithm originally designed for conventional (online) reviews is applied to much noisier and unstructured data from YouTube comments to illustrate the strengths and weaknesses of such algorithms with social media data. Second, the aspect extraction algorithm is enhanced by introducing a new filtering method using the search behavior of people at Google.

In the second case data from Twitter is used to investigate the influence of social media on financial markets. In the behavioral finance literature there exist some evidence that mood or sentiment of people actually can predict stock market returns [Da et al., 2015]. This question is a highly topical research field in financial econometrics because it challenges the widely accepted Efficient Market Hypothesis (EMH) of completely rational and informed market participants. There still exist open questions of how to measure the mood, who actually trades irrational and in which cases these irrational behaviors can be identified and exploited. Another approach is to investigate the effect of news events and how new information are incorporated in the asset price. There exists evidence for attention grabbing stocks and momentum effects based on news [Barber and Odean, 2008] which also challenges the EMH. In this study I investigate a combination of both aspects. Based on more than 2.6 million news related tweets collected from Twitter accounts of main news providers like Reuters and Bloomberg, I introduce a new hypothesis of how information from social media can reveal intraday misspricings of stocks in the German Stock Index (DAX). I investigate overreactions to good and bad news based on an exploratory analysis by identifying the fundamental value of an asset after a news event, using corresponding tweets. This study contributes to the existing literature by introducing a new theory of how to identify and measure intraday stock market misspricings induced by news events using social media data. It's worth mentioning that the research in the field of analyzing the influence of social media data (mainly twitter) on financial markets faces some crucial difficulties. The main concern is related to the data basis researchers use in their investigations. Since there are no standard data basis available each study relies on its own data collection and therefore the results are difficult or even impossible to reproduce. Tackling this challenge is an important part of any data driven research which I generally discuss in section 6.

In the third case I focus on the question of how to evaluate the results of a forecasting model. The problem statement is about a residual value forecast for used cars. Based on vehicle and leasing contract specific data, the residual value has to be determined. After the model building using linear regression, time series analysis and Artificial Neural Networks (ANN), according to the general data analytics pipeline, the question of how to properly evaluate the results arises. In many studies it is common practice to report the results using classical performance measures like Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE). Often the selected models are benchmarked against other popular

methods to illustrate the performance improvements of the proposed model. To compare the model accuracy tests like the Diebold-Mariano test [Diebold and Mariano, 2002] are available. Also statistical residual test for heteroscedasticity or autocorrelation are often used to construct and select the best fitting models. These methods have in common that they are unable to make a statement about the concrete regions where the models perform good/bad in the data space. They just accept or reject the models as a whole. In this study I propose a new residual visualization technique using heat maps. This method is able to visualize the regions in the data space where the model actually performs good or bad which provides hints on problems with the available data (e.g. outliers) or the model specifications (e.g. not incorporated nonlinearities). The method is demonstrated on a stylized example with standard normal distributed residuals and artificially introduced biases and afterwards used to evaluate the real world example of residual value forecasts. This study contributes to the literature by introducing a new method for model evaluation using visual residual analytics based on heat maps. The method can be adopted for any kind of regression or classification problem and any type of model.

This thesis is structured as follows. After this introduction and motivation the three cases are investigated in sections 2 – 4 respectively. An overall comparison is conducted in section 5 and a roadmap for future research is derived. Section 6 provides a discussion and limitations about data analytics in general and points out the importance of reproducible research. Section 7 concludes. The appendix provides the most important code segments which have a reference in the text.

7 Conclusion

Never trust anything that can think for itself if you can't see where it keeps its brain.
– Arthur Weasley (Joanne K. Rowling) [Rowling, 1998]

In this thesis three real world problems are investigated using data analytics methods. In a first case an algorithm for aspect based sentiment analysis is implemented and tested on YouTube data. Due to the bad performance compared to conventional online reviews, a new filtering method based on Google Trends is developed. In the second case the influence of Twitter on financial markets is discussed. An exploratory analysis reveals an interesting pattern of a possible overreaction in the market to news which can be identified by the first tweet (approximately the point in time when the information is publicly available) about triggering events. This theory is still under investigation and a detailed empirical study is planned for further research. In the third case a new method is developed for the visual evaluation of (big) data machine learning models. The residuals of the model on a test set are visualized in a two-dimensional weighted heat map which reveal previously hidden problems of the model specification.

In section 6 a critical discussion is provided about the chances and possible pitfalls of using data analytics in practice. Especially the standard of reproducible research is explained in detail and it is pointed out how important reproducibility in the field of data analytics actually is. A culture of openness in science is recommended to overcome the credibility crises of data driven research and to make the findings more trustable and better understandable for everyone. Any researcher has the privilege and the responsibility to communicate the findings in a way that actual knowledge can be derived from it.