Correlation of Domain Knowledge and Success of Data Scientists

Bachelorarbeit

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

vorgelegt von

Name: Naehring Vorname: Kira Marie

[redacted] [redacted] [redacted] [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Hannover, 10.08.2018

Table of Content

Abstract

This thesis deals with the connection between the attributed Domain Knowledge of a Data Scientist and their success, as well as the quality of their results. The ongoing discussion whether and how Domain Knowledge influences a Data Scientist in his work will be investigated and the opposing positions will be examined in more detail. The main focus is on three defined hypotheses on how Domain Knowledge and the effectiveness of a Data Scientist could be related. Then possible approaches are developed to test these hypotheses and to answer the main question which is wheter a data scientist is in the need of Domain Knowledge or not. This thesis thus prepares the ground for further experimental- or analytical investigations of the connection between Domain Knowledge and the effectiveness of a Data Scientist and has no claim to answer the main question or to give an overall answer to the discussion. However this paper provides an overview of the current state of research in the field and opens up new approaches.

Key Words:

Data Science, Domain Knowledge, Big Data, Data Sets, Effectiveness, Translators

1. Introduction

This scientific work focuses on a still young field of research: the Data Science, and an area that has not yet been fully illuminated: the connection between Domain Knowledge and the performance of a Data Scientist.

The research question covers how the efficiency of a Data Scientist is related to his Domain Knowledge and how this relationship could be tested.

Knowledge of the effects of Domain Knowledge on the effectiveness of a Data Scientist would have far-reaching consequences. On this basis, the training and further education of Data Scientists could be adapted and their use in the company optimised. In many scientific papers, for example, a connection is assumed or hypothesised and

recommendations for action and statements are made on the basis of this assumption (Waller and Fawcett 2013: 78). Clarity about the actual context could thus bring new and reliable findings to research or call old findings into question. At the moment, there are no scientific papers or researchs to investigate or trace this research gap. This paper provides first approaches how this research gap could be closed.

This paper is challenging the overall outcome of an earlier paper from Waller and Fawcett. - "Data Science, predictive analytics, and big data: a revolution that will transform supply chain design and management.", from 2013 in the aspect of the relevance of Domain Knowledge for Data Science. In which it is assumed that Domain Knowledge has a positive influence on the efficiency of a Data Scientist. In this paper, possible negative correlations are also questioned, and for the first time options for examining this correlation are presented.

The question to be answered is whether Domain Knowledge is relevant for Data Scientists to work more effectively or not. In the following, the basic terms are defined and a summary of the relevant research results is given. First, Big Data is viewed as such, then the resulting relevance for data scientists and the special requirements and competencies of this profession are described. Next, Domain Knowledge will be examined and research results from the social sciences will be used which are partially applicable to the Domain Knowledge of the Data Scientist. The basic part is followed by a literary overview and an overview of the current papers and opinions on the connection between the effectiveness and the Domain Knowledge of a data scientist. After arguments for and against relevance, and arguments for a negative and positive influence of Domain Knowledge on the performance of the Data Scientist, three basic hypotheses are formulated on how domain knowledge could have an effect.

In chapter 4 the research gap will be addressed and possibilities will be presented how the relationship between the efficiency of a Data Scientist and his Domain Knowledge could be tested. Altogether four abstract experiment designs are presented which could test such a connection. Subsequently, possible experiment results are discussed and their effects and recommendations for action are explained.

are not Data Scientists and are not necessarily Analysts. Simplified, they combine technical expertise with operational expertise and make their Domain Knowledge available to the Data Scientist (Henke et al. 2018). Translators not only work one-sided but also translate the concerns of the management for the Data Professionals as well as the results of the Data Professionals for the management (Henke et al. 2018). Henke et al. name Domain Knowledge as the most important skill of a Translator. They have to be experts in both their industry and their company and must understand the key operational metrics of the business and their impact (Henke et al. 2018). In addition to comprehensive Domain Knowledge, a Translator should also possess quantitative analytical skills and structured problem solving skills (Henke et al. 2018). The focus is less on the modeling of quantitative tools than on the interpretation of results (Henke et al. 2018).

However, the Translator is not a universal remedy for companies. Although the Data Scientists themselves no longer need domain knowledge, which makes it easier to work with external Data Scientists and to train newly hired Data Experts, the Translator requires comprehensive domain and deep company knowledge. To fill the position of a Translator, it is therefore advisable to train long-term employees and use them as Translators (Henke et al. 2018). Another problem is that there are no standardized programs for Translators. For example, there are currently no harmonized certificates or degrees that would distinguish Translators (Henke et al 2018). A problem that until recently also occurred in the job profile of Data Scientist. As a result companies are currently training their own Translators to help increase the return on investment from their analytical initiatives (Henke et al. 2018).

5. Summary, Limitation and Outlook

In the previous chapters, the necessity of closing the research gap regarding whether Domain Knowledge is relevant for Data Scientists or not was discussed in detail. It becomes clear how much the handling of Data Scientists in companies depends on whether the effectiveness can be increased by domain expertise or whether it even has a negative influence. Ending the ongoing discussion would have consequences not only for companies but also for data scientists and their training and further education. This question needs to be answered in order to enable a more optimal collaboration with data scientists. Another interesting research gap has also been discovered which needs to be investigated. Do data analysts work better when they work with unlabeled data and allow it to affect them only if they are free of experience

and prior knowledge? An affirmation of this question by conducting experiments and the corresponding results would have far-reaching effects on future research and working methods in all disciplines.

This work does not provide an exact answer to the question of whether Domain Knowledge leads to increased effectiveness among Data Scientists. It summarizes the current literature and highlights the problems of Domain Knowledge from different perspectives. Hypotheses have been developed which cover the possible answers to the research question and illustrate their consequences. Furthermore, a variety of alternatives are presented which could experimentally get to the bottom of the research question. Furthermore, a rapid execution of the experiments listed above would be a good step towards a better understanding of the optimal handling of big data and the emerging career profile of the Data Scientist as well as the requirements that must be placed on him.