Machine Learning Based Model Evaluation in Big Data Analytics Applications

Masterarbeit

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

vorgelegt von

Name: Gercke Vorname: Dennis

[redacted] [redacted]

Prüfer: Prof. Dr. M. H. Breitner

Hannover, den 02.10.2017

Contents

9 Conclusion

In the introduction in Chapter 1 of this thesis, it was pointed out that ML "(…) has progressed dramatically over the past two decades, from laboratory curiosity to a practical technology in widespread commercial use." (Jordan & Mitchell, 2015: 255). In these two decades, data scientists were able to utilize their ML knowledge to detect and investigate new application areas in diverse scientific fields. This was a major reason for ML's rapid progress. They accomplished this task even though they lacked one very integral determinant, i.e., the respective expert knowledge in these areas. The increasing difficulty in identifying new, more sophisticated application areas, however, will require more and more of the missing domain/expert knowledge in the respective scientific fields. This thesis proposed that this challenging situation for further ML research progress could be solved best by the experts in the respective scientific fields. The understanding of the fundamental mechanisms in their scientific areas and the data related to them gives them a major advantage. Compared to this advantage, they lack only a minor determinant, i.e., some basic knowledge of how ML methods work and are applied. Therefore, the thesis aimed to provide these experts with the basic ML knowledge needed to be able to detect and solve ML problems in their areas.

This task was conducted by means of an in-depth representation of a scientifically and practically oriented, complete Big Data Analytics ML problem-solving process in a case study approach, which effectively imparted the necessary knowledge in a target-group-focused manner. The self-developed representation of the origin and the theoretical development and categorization of decision-tree-based ML models in Section 2.3.1, comprehensively delivered important theoretical background information. This section did not only evaluate the mechanics of XGBoost, but also of all of its predecessors. This strong theoretic foundation was important, as it created a much deeper understanding of ML methods in general, instead of providing only the method knowledge absolutely needed for the problem at hand. In combination with the holistic ML data analysis process, this facilitated the comprehension of the subject as a whole, which, in turn, lead to the desired increase in ML problem detection and solving capability.

In the case study, XGBoost was shown to be easy to use, yet sophisticated, as was promised in Section 2.3. For the data preprocessing task, particularly in Section 3.3.1, which was concerned with cleaning missing and faulty data, its sparsity awareness was a major advantage in dealing with the low-quality data set. In the model building step in Section 4.1, XGBoost was operational after an exceptionally few lines of code, which did not require very sophisticated programming skills. In the applied FE part in Section 5.2, it was of major assistance in the

feature selection task, where its “Feature Importance Score” enabled the user to declare specific feature extraction measures (un-)successful. In the model tuning part of Section 6.2, XGBoost was found to be relatively independent of this task. Thanks to its well-working regularization term, a sophisticated HP tuning is often not needed. For the specific case study problem, a tuning nevertheless was conducted since the default HPs did not work sufficiently well on the unusually low-quality data set. Though computationally demanding, the HP tuning process was not as challenging as in other ML methods, such as in ANNs. For optimal problem-solving results, this section also featured the configuration and utilization of a sophisticated AWS EC2 cloud computing solution, which delivered optimal HP results for the extensive fine-tuning-task.

While it was easy to use during the entire data analysis process of the case study, it also delivered state-of-the-art results, as described in Section 6.3. The score in the Kaggle competition, which was the basis for the case study, was just 0.03 behind the winning RMSLE score of 0.30. In addition, XGBoost was utilized by many other competitors in the disclosed top 100 solutions at the Kaggle competition site, while the two best solutions used another tree-based approach (LightGBM). Overall, the ML algorithm used in this thesis proved to be the optimal model for ML entrants, as its balanced nature made it easy-to-use while delivering sophisticated results. Thus, XGBoost can be said to have optimally supported the desired learning process.

Of major interest in the course of the thesis was Chapter 5. Initially, the thorough literature review in Section 5.1 identified FE as a highly under-conceptualized term. FE is often used as kind of a buzz-word, but what is conducted differs markedly from FE; in a lot of works, it consists of only one subtask. Further, these subtasks are defined very inconsistently in different works and are often used interchangeably. It is this under-conceptualization which results in the current comprehension of FE as a mere data preprocessing task, which is the reason why it cannot be utilized to its full potential. Therefore, this thesis proposed a new understanding of the currently under-conceptualized “FE” term in the data sciences, one that turns it into a more central concern in the data analysis process. As such, it could improve the representation of the data for a better understanding by an algorithm through working as a tool for incorporating domain/ expert knowledge.

The application of FE measures in this proposed sense led to intriguing examples for the case study data in Section 5.2. The thesis started with a quote by Mark Twain, stating that “A man who carries a cat by the tail learns something he can learn in no other way.”. At least for the task of detecting and solving ML problems in specialized scientific areas, this section

proved this to be true. It was shown that the domain/ expert knowledge the targeted researchers hold is indeed the most important determinant for this task, as this knowledge, learned from years of experience, cannot be replaced by whatever extent of ML expert knowledge. In the special Kaggle competition case, the expert knowledge was crowdsourced by 3,274 contestants which were able to provide at least some of the required domain/ expert knowledge, which delivered insights not achievable by ML knowledge alone.

Examples for these insights were given in Section 5.2. Expert knowledge in the realty business revealed, e.g., that a change in the dependent variable can greatly improve the prediction accuracy, as the more common and relatable measure for realty prices is the price per m^2. Another issue that was only explainable by expert knowledge in the realty area was a dependent variable distribution deviating from the desired normal distribution in terms of unimodality. Expert knowledge showed that this was caused by under-reported market values of certain realties due to tax fraud reasons, possibly affecting up to 2,000 realty observations in the training set. A Sberbank realty expert for the Russian market pointed out that the under-reporting of the market value of an apartment is a common practice for "Investment" product type realties. Conclusively determining the realties affected by under-reported prices, remains a problem for data scientists and Sberbank. The domain knowledge that this problem only occurred for the "Investment" but not for "OwnerOccupier" product type led to the insight that buyer types represented by these feature values have substantially different requirements regarding realties. Thus, they also have a completely different willingness to pay for apartments with specific characteristics. This led to the approach of developing separate models for the distinct product types, which heavily improved ML results.

These examples showed, most impressively, that the domain/ expert knowledge the targeted researchers hold is indeed the most important determinant for the task of detecting and solving ML problems in highly specialized scientific fields. Thus, the proposed approach of this thesis to employ those researchers for the challenging task ahead, by giving them the basic ML knowledge they need, can be deemed an adequate solution. The in-depth representation of the origin and the theoretical development and categorization of decision-tree-based ML models – in combination with the holistic, step-by-step ML data analysis process – can be considered a well-working measure to provide this ML knowledge to researchers. This knowledge combination will successfully enhance the ability of these researchers to perceive known matters and problems in their areas from a data science point of view. An important tool for leveraging the advantages of expert knowledge were found to be FE measures. Thus, it can be said that this thesis' approach accomplished the intended objective, as this will enable them to contribute,

not only to their own fields, but also to ML, in general. It will lead to new and more sophisticated application areas, which will ensure ML's further progress in the next decades.