Abstract

The question of how to incorporate human domain knowledge in practical data science projects is still a major challenge. While machine learning tasks are usually carried out by technically skilled data scientists, these analysts do not necessarily have the required domain knowledge concerning a particular business problem to explain certain phenomena. In real-world data science applications, this may result in models that do not adequately reflect relationships in the data. We address this issue by introducing a heat map based technique for model error visualization to facilitate discussions of the results between data scientists and domain experts. By discussing model errors with domain experts during the iterative analysis process, the generated insights can be used for engineering new features (explanatory variables) which better represent the problem and therefore improve the results. We demonstrate the visualization approach based on artificial data and in the context of a real-world industry example.

Keywords: Feature Engineering, Domain Knowledge, Data Science, Visualization, Heat Maps, Model Evaluation

Introduction and Motivation

Data-driven decision-making is a steadily growing field in information systems research (Agarwal and Dhar 2014). The underlying concepts can be summarized under the term data science, which is often defined as a combination of statistics, programming and domain knowledge (Conway 2010). While statistics and programming knowledge are obvious requirements for valid and reproducible analyzes, the role of domain knowledge is more unclear in this discipline (Dumbill et al. 2013). In the literature, domain knowledge is often seen as part of an iterative feature engineering process for machine learning tasks (Domingos 2012). Features (independent variables), define the input space for an algorithm which learns the dependencies in...