Decision Support Systems 72 (2015) 72–81

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector

Julie Moeyersoms *, David Martens

Faculty of Applied Economics, University of Antwerp, Belgium

ARTICLE INFO

Article history:
Received 11 August 2014
Received in revised form 15 December 2014
Accepted 6 February 2015
Available online 16 February 2015

Keywords:
Data mining
Predictive modeling
High-cardinality attributes
Churn prediction

ABSTRACT

High-cardinality attributes are categorical attributes that contain a very large number of distinct values, like for example: family names, ZIP codes or bank account numbers. Within a predictive modeling setting, such features could be highly informative as it might be useful to know that people live in the same village or pay with the same bank account number. Despite this notable and intuitive advantage, high-cardinality attributes are rarely used in predictive modeling. The main reason for this is that including these attributes by using traditional transformation methods is either impossible due to anonymization of the data (when using semantic grouping of the values) or will vastly increase the dimensionality of the data set (when using dummy encoding), thereby making it difficult or even impossible for most classification techniques to build prediction models. The main contributions of this work are (1) the introduction of several possible transformation functions coming from different domains and contexts, that allow the inclusion of high-cardinality features in predictive models. (2) Using a unique data set of a large energy company with more than 1 million customers, we show that adding such features indeed improves the predictive performance of the model significantly. Moreover, (3) we empirically demonstrate that having more data leads to better prediction models, which is not observed for "traditional" data. As such, we also contribute to the area of big data analytics.

1. Introduction

The increasingly widespread collection and processing of data enable companies to use these data assets to improve decision making. A popular way to do so is by using predictive modeling. Examples can be found in different domains like credit scoring, where predictive models are used to separate good from bad loan applications [5,26,44] and in marketing where likely adopters are identified [6,29].

Two types of data are commonly used in order to build predictive models: structured data (e.g. socio-demographic data, number of products purchased) on the one hand and relational or behavioral data (e.g. transactional invoicing data, phone call or other networked data) on the other hand. Previous work has shown that using behavioral data is extremely valuable and improves the performance of the model significantly [20,21,27]. Unfortunately, these behavioral data are very unique and are exclusively preserved and accessible to the banks, Telco operators and Googles of the world. Structured data, on the other hand, is widely available. Consider the example of an energy company: they have extensive information on customer's socio-demographics such as address, age, and family size but they do not have data on transactions between customers such as phone calls or payments.

The focus in this paper is on one specific type of structured data, namely high-cardinality attributes. These are categorical attributes with a very large number of distinct values, such as: bank account number, family names and ZIP codes. Surprisingly, high-cardinality features are rarely used in predictive modeling. The main reason is that high-cardinality attributes are difficult to handle as including them would imply that the dimensions of the data set will quickly explode. Consider the example of the energy company that has information on the family names of the customers. Including this attribute with dummy encoding implies that millions of dummies will be created: one for every family name. As a consequence, the computational effort will increase substantially, and it will even be impossible for most of the techniques to cope with such high dimensions.

This case study discusses churn prediction in an energy context, where the aim is to predict which customers are most likely to churn and thus switch to another energy supplier. The data is stemming from a large energy company in Belgium with more than 1 million customers and includes several high-cardinality attributes. The question remains whether or not it is possible to include this type of attributes without expanding the dimensionality of the data set and if it actually improves the prediction model significantly. This research is built around the following three research questions:

1. Is it useful to include high-cardinality attributes?
2. How to transform and include such high-cardinality attributes?

* Corresponding author at: Prinsstraat 13, 2000 Antwerp, Belgium.
E-mail address: Julie.Moeyersoms@uantwerp.be (J. Moeyersoms).

http://dx.doi.org/10.1016/j.dss.2015.02.007
0167-9236/© 2015 Elsevier B.V. All rights reserved.

3. Does adding more data yield a better generalization performance of the prediction model?

This last question refers to the issue of Big Data, which is defined as data that is so big that traditional data processing systems cannot cope with it [25]. In the case of behavioral or relational data, it has been shown that adding more data points effectively improves model performance [21]. For socio-demographic data on the other hand, traditional predictive analytics may not receive much benefit from increasing the amount of data beyond a certain point [32]. Whether or not the same principle applies for high-cardinality data will be shown later in the paper.

Note that although we focus on a case study in the energy sector, this research is relevant for all kinds of retailers as this high-cardinality data is largely available everywhere. An important final remark is that a sufficient amount of data is required when working with high cardinality attributes, in order to apply the techniques proposed in this work. The rest of the paper is structured as follows: Section 2 discusses churn prediction and the importance of the topic in the Belgian energy sector. In the next section, 'high-cardinality' data is explained in more detail as well as possible techniques to transform and include them in the data set. Section 4 defines the methodology of the experiments and provides information on the data set. Next, Section 5 describes the results of the experiments. Finally, the last section concludes the paper and identifies some interesting issues for future research.

2. Churn prediction in an energy setting

The market value of the Belgian electricity and gas market was estimated at $22.4 billion in 2011 [13] and is expected to grow steadily over the next years [37]. This clearly indicates the size and importance of this market for the Belgian economy. Belgium fully liberalized its energy market in July 2007, thereby allowing customers to choose their natural gas and electricity supplier. Since this liberalization, the energy landscape in Belgium is in a continuous change [53]. A recent report of the European Commission [13] shows that the rate of people who switch suppliers has increased rapidly. In 2011 the churn already reached 10% for the electricity retail market and 11.2% for the entire gas retail market. Due to these recent developments, customer churn prediction has received a large amount of attention from energy suppliers.

The goal of churn prediction is to indicate which customers are most likely to leave the company. In this way, those customers can be offered incentives to stay and the churn rate can be reduced [45]. This also allows companies to decrease the costs of customer retention campaigns, because they can target more efficiently the customers with the highest probability to churn [49]. The economic value of customer retention has received much attention in the literature [45,49]. A decrease in customer switching can lead to many benefits for the company because of the following reasons [45,51]: first, retaining existing clients is five to six times less expensive than to acquire new customers [7,9]. Second, it is shown that long-term customers are more willing to recommend the company to other people [9,16]. This positive word-of-mouth associated with increased customer retention can, in turn, lead to lower marketing costs to acquire new clients [22]. Finally, long-term customers are less sensitive to competitive pull. They pay less attention to competitor's advertising and are less likely to compare prices of their own supplier with those of other suppliers [9,42].

Many research has been done on customer churn prediction and its usage [8,30,39,50,52], which proves the importance of the topic. For an extensive overview of literature on churn prediction modeling we refer to [51]. None of the studies listed in [51] includes high-cardinality data. Yet, such data is readily available (think of the ZIP code or last name of customers) and including these data can lead to a superior predictive performance. Although in data mining research the focus is often on benchmarking different classification techniques in order to find the best performing technique in terms of predictive performance, it is argued that data quality is at least equally important [4]. Good data quality is often the best way to augment the performance of a prediction model. Therefore, the focus in this paper is on data inclusion rather than the classification technique applied.

3. Including high-cardinality features

The inclusion of attributes with a high cardinality in prediction models constitutes the subject of this case-study. First, a more detailed explanation of high-cardinality attributes is given. Next, possible transformation techniques are listed that allow us to include categorical features with or without a high cardinality.

3.1. High-cardinality versus traditional nominal data

Structured attributes can be continuous (implying that they contain real numbers and are defined over a continuous range) or nominal (meaning that they can take only a finite number of values).[1] Examples of continuous attributes are the *amount of the invoice* or *kWh used*. An example of a nominal feature is *type of contract*, that can take three different values: *electricity*, *gas* or *electricity and gas*. The cardinality of a nominal feature can be defined as the number of distinct values that attribute can take [31]. Thus, for the example of the feature *type of contract*, the cardinality is 3. The features with a small cardinality are referred to as '*traditional*' nominal attributes. The features with a very high cardinality on the other hand, are named '*high-cardinality*' attributes. The latter are generally removed from the data as including them using dummy encoding will expand the dimensions of the data set quickly, thereby impeding the model building. Based on literature [49], it can be noted that features with more than 100 different values are mostly discarded from the analysis, hence we consider this to be the threshold and name all features with more than 100 distinct values high-cardinality features. Based on this assumption, three features of the energy data set are identified as high-cardinality data: *family name*, *bank account number* and *ZIP code*.[2]

3.2. Transformation techniques

In data mining literature, nominal attributes are usually included in the data set using dummy encoding or grouping methods. These methods can also be applied on high-cardinality features and are discussed in this section. Moreover, three other methods are proposed that allow the transformation of high-cardinality features into continuous attributes. All methods are listed in this section and illustrated with an example in Table 1.

3.2.1. Dummy encoding

A common way to include nominal features is to use dummy encoding where the M categorical values are transformed into M new dichotomous variables that are coded as 1 or 0. This method allows the adding of these variables to the model and provides an easy interpretation of the output since one variable matches with one value of the original variable. The main drawback of this method is that if M becomes large, the computational effort will increase significantly.

In the case of traditional nominal features, dummy encoding is an appropriate method to use. Applying dummy encoding to high-cardinality attributes, on the other hand, could create millions of dummies. Since most of the predictive modeling techniques do not scale to such dimensions, dummy encoding cannot be used for high-cardinality data. For this reason, previous studies did not take into account this type of attributes. The first part of Table 1 shows an example where dummy encoding

[1] A third variable type is ordinal, which, like nominal variables, is categorical, but with an order in the values. For example age encoded as young, middle-aged and old. The proper way to handle such variables is through thermometer encoding.

[2] In our data set, these features include even a larger amount of distinct values, going from 1000 up to more than 1 million.