Vrije Universiteit Amsterdam

Masters Thesis

Understanding user behavior in e-commerce with Long Short-term Memory (LSTM) and Autoencoders

Author:
Leidy Esperanza Molina Fernández

VU Supervisor:
Prof. Dr. Sandjai Bhulai

VU Second Reader:
Prof. Dr. Rob de Jeu

MeasureWorks Supervisor:
Jeroen Tjepkema

Faculty of Science

August 2018

Abstract

Over the past few years, it has been proved that a fast website has a positive influence on the conversion rate of e-commerce websites. However, it is not yet known what other reasons related to performance explain the fact that some users convert and others do not. In this thesis, we analyze the behavior of users in three e-commerce websites, in order to analyze the path of the user in the site, and estimate a series of features that impact the behavior of the user. Two approaches were proposed: Sequence Modeling and Anomaly Detection. The former is an application of Long Short-term Memory (LSTM) in the path of the user as sequential data, aiming to predict whether the user purchases an item on the site. The latter is an alternative use of Autoencoders to detect anomalies, where the buyers are considered anomalies, this approach allows to analyze the influence of other features related to performance in the final decision of the user. We found that the path of the user does not influence the decision of purchasing a product. In contrast, we show that the user becomes more likely to purchase after a certain number of steps on the website. In addition, we are capable of showing the influence of the performance on each e-commerce website for specific page groups. Based on the results a decision-making tool was developed to estimate the possible impact caused by a positive increment of speed.

Key Words: performance, e-commerce, conversion, conversion rate, page group, LSTM, Autoencoders.

Chapter 1

Introduction

1.1 Company Description

MeasureWorks is a (predictive) Web Performance Solutions provider. They provide performance tooling and consultancy for a wide range of international companies to make websites faster, more reliable, and ultimately deliver more conversion (revenue). In the last years, MeasureWorks has been selected as 1 of the top 3 Analytics & Optimization companies by Emerce100. Originally founded in 2008 by Jeroen Tjepkema, MeasureWorks currently has 14 people working for more than 40+ clients (ING, Bol.com, Adidas).

Due to the international nature of their clients, MeasureWorks has large sets of real-time tracking data from clients such as Bol.com and ING, but, also social data. The latter data refers to the performance of the website during the visit of the user, from the moment they start surfing on the site until they finalize the session by closing the browser.

MeasureWorks has focused its activities on increasing the performance of e-commerce websites, which are known as websites where customers can purchase products. They believe that a fast website is a fundamental requirement for every user interaction on the web. Therefore, they offer different methodologies to keep the websites fast and then optimize the rate conversion, which is the ratio of converted sessions over the total number of session.

1.2 Problem Description and Motivation

Each e-commerce website generally has a predefined logical structure, in the sense that all of them have a home page that is linked to other logical categories, such as Product, Listing, or Offers. These logical groups of pages are clustered by some meaningful reason, for example, similar application or functionality. At the same time, logical categories can be subdivided into more specific arrangements, called page groups.

Once a customer logs on to the site, a session is recorded and it gathers useful information about what pages have been visited during the navigation. The session can be understood as a linked list of pages, each one belonging to only one page group. When a purchase is performed, the session is converted, which generates a profit. Each session is recorded for 30 minutes, but if the user keeps browsing after those minutes or if the same user logs in multiple times, the system considers it as a new session. For this reason, it is not possible to link multiple sessions to one single user, so when user behavior is discussed, it refers to the session instead of the single user.

E-commerce data collection includes massive numbers of clickstreams per session (Kohavi and Provost [2001]), which means that there are millions of page group combinations that might lead to a conversion. Throughout the thesis, I will use the term path to refer to the combination of page groups, where the path for each user can be represented as a sequential data. In order to know sequential data, the fundamental property consists in recognizing the importance of the order of information (Lang and Rettenmeier [2017]). Each click in the session that redirects to a Page Group is considered a step in the path. As an example, a user may log in to the Page Group Home and then go the Page Group Listing, after which the session is closed or expired. This means that the session did not convert. However, it can also happen that the user followed a different path to purchase an item, indicating that the session converted, as it is shown in Figure 1.1.

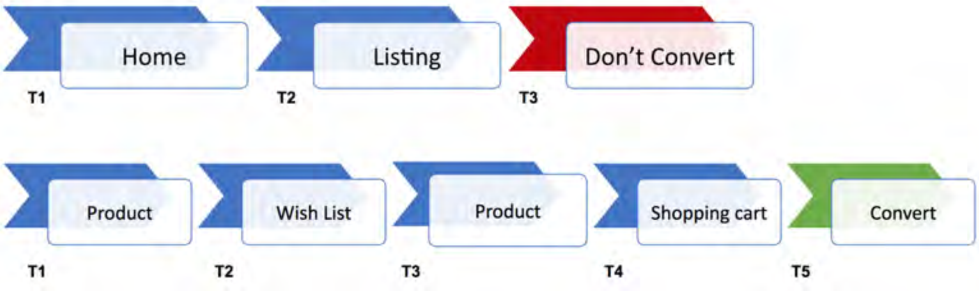

FIGURE 1.1: Example of the path of different sessions in a e-commerce website

Furthermore, for every e-commerce company, Conversion Rate optimization is important not only because it minimizes Customer Acquisition costs, but also because it is transcendental to transforming passive website visitors into active users that will purchase their products. For the past 5 years, it has been proved that performance has a positive influence in conversion (Kurylets [2012], Nygren et al. [2010], Rempel [2015]). Nonetheless, it is not yet known whether other features as the length of the session, or the performance of each page group are also transcendental for this process.

More than investigating the behavior of users within the website, MeasureWorks wants to know other features that influence their conversion. Research has tended to focus only on the effects of the average page load time rather than the length of the session or other features. An additional problem is that they are based on the behavior of the users in general, but no one has focused on the differences between the *viewer* and the *buyer* and the effect of different page groups in the final conversion. Here the *buyer* is the user that purchase an item and the *viewer* the one who does not. One of the main issues is related to the uneven number of converted sessions in comparison the number of sessions that does not convert.

The present report analyzes the behavior of the users in three different websites, in order to investigate what reasons related to performance explain the fact that certain sessions in a website convert and others do not, and analyze if those reasons are related to the page group that the user visits or to any other additional features of the website. To do so, we proposed two different approaches. The first one is to analyze the path of the user as sequential data, by investigating the page groups that the session has traversed in order to see if there is a successful session path. Based on that, the intention is to lead users from poorly converted session paths to others with a higher Conversion Rate. The second approach is to consider the clients that purchase as anomalies and identify the features that impact the final decision regarding purchasing or not a product. In order to do this, we use a series of KPIs (key performance indicators), for instance: session length, average page load time per page group. In addition, the behavior of the users in the three websites is analyzed and compared, to decipher if there is a difference between the e-commerce websites. The three websites were chosen because it is expected to find similarities between those related to the fashion commerce, but differences between e-commerce in fashion and supermarket. At the same time, by using these websites it is possible to see if the methodologies work in different contexts.

In order to analyze the path of the user as sequential data, we resort to Long Short-term Memory (LSTM) networks, which are a type of Recurrent Neural Networks (RNNs), capable of learning long-term dependencies (Hochreiter and Schmidhuber [1997]). In this context, we make use of the paths of the user as sequential data and we will classify the sessions between *buyers* or *viewers*, if the user purchase an item or do not. Besides its great success in classifying sequences, LSTM's inner operations are hard to be interpreted, so, to understand which were the page groups that influenced a specific path of the user, we use the Layer-wise Propagation (LRP), a method introduced by Montavon et al. [2017], to interpret and understand deep neural networks.

On the other hand, if we assume that the fact that a session converts is an anomaly, we can make use of methodologies related to anomaly detection. In this case, the anomaly will be the users that purchases an item, and the features will be characteristics in regards the performance of the session. It was decided to implemented Autoencoders, since they are based on neural networks and, therefore, have the ability to discover high quality, non-linear features. Autoencoders are also capable of eliminating outliers and noise without using a clean and balanced data set (Meng et al. [2017], Zhou and Paffenroth [2017]). In this sense, through the use of Autoencoders we are able to describe the behavior of the users in the websites, as well as their inner characteristics. The latter allows us to gain insights into the process of retaining current customers and to convert visitors into customers.

1.3 Contributions

In this paper, we evaluated the two proposed methodologies LSTM and Autoencoders on three e-commerce websites, to analyze the behavior of the users in each one. The contributions we obtained are summarized as:

1. The RNN is not capable of predicting whether the session will purchase or not, because the model is underfitted and it predicts every session as converted, as a result of the unbalanced instances. Thus, we show that there is no relation between the path of the user at a specific site and the decision of buying a product. However, by applying the LRP extended method to an LSTM trained on a classification task, it is possible to analyze the converted sessions and know which are the page groups that influenced their decision.

2. It is the first time that Autoencoders are applied to analyze the behavior of users in a website. The Autoencoders are capable of predicting whether a session will convert or not, and from the resulting model three classes are created: buyers, visitors and possible buyers. In addition, it is possible to analyze feature importance per class.

3. The Autoencoder showed better results than the LSTM for three different e-commerce websites. Showing that the methodology can be used for different commerces.

4. The document provides with a decision-making tool, that is capable of estimating the total passive visitors that can be converted into active users and the conversion rate if the site speeds up in a certain percentage.

1.4 Thesis outline

This thesis is organized as follows: the second chapter gives a brief overview of the literature review and related research, showing the way in which both methodologies have been used for similar problems, and introduces the necessary background for Long Short-term Memory (LSTM) and Autoencoders. The third chapter introduces the dataset, from three different companies. Chapter four describes the experimental setup. The next chapter reports and analyzes the results of the different models. Finally, chapter six concludes and proposes further actions.

Bibliography

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kohavi, R. and Provost, F. (2001). Applications of data mining to electronic commerce. *Data mining and knowledge discovery*, 5(1-2):5–10.

Kurylets, I. (2012). Web-analytics and performance evaluation of internet marketing.

Lang, T. and Rettenmeier, M. (2017). Understanding Consumer Behavior with Recurrent Neural Networks. In *Proceedings of the 3rd Workshop on Machine Learning Methods for Recommender Systems. http://mlrec. org/2017/papers/paper2. pdf.*

Meng, Q., Catchpoole, D., Skillicom, D., and Kennedy, P. J. (2017). Relational autoencoder for feature extraction. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 364–371. IEEE.

Montavon, G., Samek, W., and Müller, K.-R. (2017). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*.

Nygren, E., Sitaraman, R. K., and Sun, J. (2010). The Akamai network: a platform for high-performance internet applications. *ACM SIGOPS Operating Systems Review*, 44(3):2–19.

Rempel, G. (2015). Defining Standards for Web Page Performance in Business Applications. In *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering*, pages 245–252. ACM.

Zhou, C. and Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 665–674. ACM.