Buy Prediction in E-Commerce with Deep Learning

Tasmin Herrmann

University of Applied Sciences Hamburg
Department of Computer Science
Hamburg, Germany
tasmin.herrmann@haw-hamburg.de

August 20, 2018

Abstract

The topic of the work is the prediction of buys in an e-commerce shop with the help of session data. This paper presents the research questions that will be dealt in future work. These are the following. I want to investigate whether deep learning methods are suitable for this type of prediction model. Furthermore, the results will be compared with the results from the RecSys Challenge 2015 and it should be answered whether deep learning methods are more suitable than the ensemble learning methods used in the challenge. To answer the questions, evaluation criteria have to be established and different deep learning methods have to be used for modeling. In this paper a few scientific papers on the topic are presented and the methodoligical design for investigating the questions is explained.

Keywords: Machine learning, buy prediction, RecSys Challenge 2015, deep learning

1 Introduction

In times of big data, companies ask themselves how they should store the large amount of data, but above all how they should process it in order to gain knowledge. In electronic commerce (e-commerce) companies store a lot of data about their customers. What they look at, in what order, for how long, on which day, how they rate and review products.

There are many ways for companies to increase sales by using this customer data. This can be done with the help of machine learning algorithms. Applications with machine learning are product recommendation, personalized search result lists and dynamic pricing.

Amazon patented anticipatory shipping in 2013 [2]. It is a predictive algorithm that starts boxing and moving products before the shopper clicks buy. The implementation of this approach leads to greatly shortened delivery times, which is a great advantage over the competition. To achieve this, it has to be predicted

which customer will buy which product and when. From an entrepreneurial perspective, this concept offers the motivation to deal with buy prediction. Even if this work only concentrates on predicting whether a visitor of an e-commerce shop will buy in this session and if so, which products.
The first section presents papers that have dealt with in-session buy prediction. Here I present papers of the RecSys Challenge 2015 [1], which are all based on one dataset. Afterwards, I will present papers that have created e-commerce applications using deep learning methods. From this, research questions are derived, which are presented in the following section. Next, I will explain which experiments will be answered and at the end follows a summary with risks of the work.

2 Literature Review

In this section, solutions for the RecSys Challenge 2015 are presented and papers on deep learning in e-commerce.

2.1 Recommender System Challenge

The task of the RecSys Challenge 2015 was to predict whether the user is going to buy something or not, and if he is buying, what would be the items he is going to buy. The data for the tasks contains sequences of click and buy events performed by some user during a typical session in an e-commerce shop.
The winners of the competition were [12]. In developing the model, they were guided by the two questions that needed to be answered. If the model predicted a user as a buyer, then it is also predicted what the user will buy. The model is a two-staged classification model with gradient boosting. The algorithm used is XGBoost [5] with hash tables.
In [4] are two approaches for the task described. One approach is to classify each item in a session as a buy or not. The second approach is a two-staged classifier like in [12]. In this solution the features are divided in features by session and features by item in session. The first set contains 13 features and the second set contains 22 features. The dataset is imbalanced, because there are much more non-buy sessions than buy sessions. They used undersampling for model training to balance the data. For 509,696 buy sessions are 509,696 non-buy sessions chosen.
The first approach achieved the best results with REPTree using Weka [15] and both feature sets. The REPTree is a decision tree algorithm that builds the tree using information gain and pruning it using reduced-error.
In the second approach two classifiers with REPTree were trained. One for buy prediction with the features by session and one for bought item prediction with features by item in session with the data of the buyers.
[16] did not take part in the RecSys Challenge 2015, but with the knowledge from the solutions of the competition he developed two models. These are described in detail, which is usually not the case with competitive solutions. This was his

motivation to develop a model with XGBoost like [12] and one with random forest.

All of this procedures are ensemble learning algorithms. In ensemble learning, an ensemble is formed to calculate a collective mean value. For example, if some classifiers run away from certain data entries in their results, other classifiers control it. Ensemble learning takes the approach that a group of algorithms produce a better average result than a single algorithm could.

2.2 Deep Learning in E-Commerce

Deep learning is a class of techniques that allows to develop models with multiple processing layers to learn representations of data with multiple levels of abstraction [10]. Deep learning has brought great advances for models in the domains speech recognition, visual object recognition and object detection. Deep learning methods used include convolutional neural networks (CNN) [10], recurrent neural networks (RNN) [10] and multilayer perceptron (MLP) [7].

The advantages of deep learning methods are they scale much better with more data than classical machine learning algorithms and there is no need for complex feature engineering because one can just pass the data directly to the network and usually achieve good performance.

Recommender systems [11] provide suggestions for items that are most likely of interest to a specific user. Software tools and techniques are used for this purpose. Machine learning methods are also used for recommender systems, which is why they are also an application case in e-commerce. Matrix factorization and neighborhood-based methods are often used to develop recommender systems.

[9] applied RNN on the domain of recommendation. They propose an RNN-based approach for session-based recommendations on the RecSys Challenge 2015 click dataset among others. They found that RNN-based models performed 20% to 30% better than a set of baseline models.

[13] improved the session-based recommendations on the RecSys Challenge 2015 dataset. They applied data augmentation, privil information and embeddings for the output layer on the models. The recall value has risen by 10% to the result of [9].

In [3] are phone prices on European markets predicted using Long Short Term Memory Neural Network (LSTM) and Support Vector Regression (SVR). The LSTM is the most accurate model to predict the next day's phone price.

3 Research Question

I would like to investigate whether deep learning models are suitable for buy prediction. Not only RNN, but also other deep learning methods such as CNN like LSTM and MLP should be considered.

Then I would like to investigate whether the models are better than ensemble learning models as used in the RecSys Challenge 2015. Here I would like to test whether the accuracy of deep learning models is better than that of ensemble models. As the work presented in chapter 2.2 Deep Learning in E-Commerce has produced better results than their comparative models. I will also compare the steps for model creation. Here I would like to see which model processes require less pre-processing and how comprehensible the models are. Exact criteria for the comparison of the procedures will have to be worked out.

4 Methodological Design

In this section it is described how the experiments for the master thesis are conducted.

To answer the two research questions, the models are created using the Knowledge Discovery in Databases (KDD) process [6]. In addition, criteria must be defined when a procedure is considered suitable for buy prediction and when a model is better than an ensemble learning model.

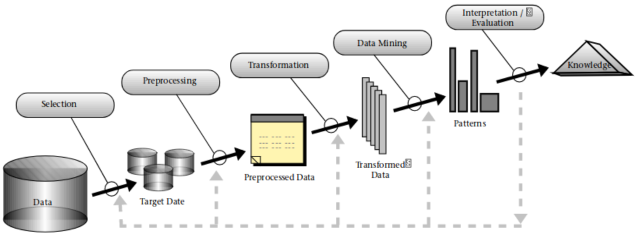

Fig. 1. KDD

It will be described which tasks have to be done in the different phases of the KDD process and which technologies are used.

4.1 Data Selection

In the phase of Data Selection, the datasets for the experiments are chosen. We use the datasets of the ACM RecSys Challenge 2015 [1]. There are two useful datasets. One with click events of an e-commerce shop and another one with buy events.

In table 1 some examples of the click data are showed. A click consists of a Session ID, a Timestamp, an Item ID and a Category. A click is one row in the dataset and a session consists of one or more clicks. The data file has a size of 1.5 GB and has 33,003,944 rows.

In table 2 some examples of the buy session data are showed. A column in the dataset represents one buy in a session. The file has 1,150,753 rows.

Session ID	Timestamp	Item ID	Category
1	2014-04-07T10:51:09.277Z	214536502	0
1	2014-04-07T10:54:09.868Z	214536500	0
1	2014-04-07T10:54:46.998Z	214536506	0
1	2014-04-07T10:57:00.306Z	214577561	0
2	2014-04-07T13:56:37.614Z	214662742	0
2	2014-04-07T13:57:19.373Z	214662742	0
2	2014-04-07T13:58:37.446Z	214825110	0
2	2014-04-07T13:59:50.710Z	214757390	0

Table 1. Examples of the click dataset

Session ID	Timestamp	Item ID	Price	Quantity
420374	2014-04-06T18:44:58.314Z	214537888	12462	1
420374	2014-04-06T18:44:58.325Z	214537850	10471	1
281626	2014-04-06T09:40:13.032Z	214535653	1883	1
420368	2014-04-04T06:13:28.848Z	214530572	6073	1
420368	2014-04-04T06:13:28.858Z	214835025	2617	1
140806	2014-04-07T09:22:28.132Z	214668193	523	1
140806	2014-04-07T09:22:28.176Z	214587399	1046	1
140806	2014-04-07T09:22:28.219Z	214586690	837	1

Table 2. Examples of the buy dataset

The data files are saved in the Hadoop Distributed File System (HDFS) to have replications on different data nodes in the cluster [8].
In this phase data samples are formed. The data is split into training and test data. The training data is used for the learning phase of the model and the test data to evaluate the generalization of the model. The training data are divided into smaller quantities to quickly test settings on the models.
Not only the rows are selected, but also the columns. Depending on the situation, only a subset of columns can be selected here. In this case, there are only a few columns and will probably be needed for the model calculation. Another point is balancing the data.
In binary classification, whether a user buys or not, the data of non-buyers is available in a much larger quantity in the dataset. Thus, the model will always predict non-buyers instead of learning the patterns for both groups.
Two approaches are undersampling and oversampling to make a balanced dataset out of an imbalanced one. [17] describes even more ways to deal with the imbalanced data set in modeling.

4.2 Preprocessing

In preprocessing, the data is examined for noise and outliers. Strategies for handling missing data must be considered.
However, this task is very difficult if you are not from the domain and know what the data looks like.

4.3 Transformation

During the transformation, the data is brought into the form so that it serves as input for the respective model procedure. For a feedforward network, the feature extraction can be very similar to that described in [16]. Some features are shown in table 3.

Features by Session
Session time in seconds
Average time between two clicks
Day of the week
Month of the year
Features by Item in Session
Whether the item appears more than once in the session
Whether the item was clicked first in the session
Number of appearances in the session

Table 3. Examples of features

4.4 Data Mining

In the data mining phase, the algorithms for the models are selected and adapted over several iterations. Various algorithms are used to test the suitability of deep learning methods for buy prediction. First, a MLP is tested and then procedures such as CNN and RNN.
For the implementation of the models the python library of TensorFlow [14] is used. The machine learning library offers GPU calculation, which distinguishes it from other libraries. Other useful libraries are pandas and scikit-learn, which are used more for data preparation than for model calculation.

4.5 Interpretation and Evaluation

In the interpretation phase, all created models with their results are considered. This means the results that were obtained with the help of the test data.
The confusion matrix and Jaccard index methods are used here. The final model must have a better Jaccard index than 0.765 to beat the prediction model in [12]. The final model must have an accuracy of over 88.53% to beat the best prediction model in [16]. It is also important to consider here how the test set looked like and, if necessary, to ensure that a similar or the same test set of data is used for a comparison.
At this point, further model results from the RecSys Challenge 2015 can be used for comparison with the models created.
This evaluation is important to answer whether the deep learning methods are better for buy prediction than ensemble learning methods.

5 Conclusion and Outlook

Predicting which products a customer buys is needed to implement such ideas like anticipatory shipping. Machine learning models are used for this purpose. In the RecSys Challenge 2015 models for purchase prediction were developed. The best solutions were implemented with Ensemble Learning procedures. Deep learning methods have also been successfully used in e-commerce for modeling, and RNNs have also been created on the data set for RecSys Challange 2015.
The two questions that I would like to examine in my master thesis were presented. I will examine whether deep learning models are suitable for buy prediction and whether they are better than ensemble learning models as used in the RecSys Challenge 2015. In order to answer these questions, the procedure was then presented. On the one hand, evaluation criteria must be defined to compare the modeling methods and on the other hand, the deep learning models must be created. For this purpose, the KDD was introduced, because the models are to be constructed with this process.
The goal of the first project is to set up the development environment for the KDD and create a feedforward network with it. In the second project follows the development of CNN and RNN on the data set. In the master thesis the

evaluation criteria for the suitability are then established and the models are checked for the criteria. Now it comes if necessary again to adjustments in the KDD process.

The risks of the work are that the quality of the datasets from the challenge are not good. However, the risk is low, as others have already created good models on the dataset. One risk is that the results are no better than those of the ensemble learning models. However, this is also an insight for buy prediction.

References

[1] ACM RecSys Challenge. *RecSys Challenge 2015.* 2015. URL: http://2015.recsyschallenge.com/challenge.html (visited on 06/12/2018).

[2] Alexa Cheater. *Amazons Supply Chain Innovation Delivers Results.* 2016. URL: https://blog.kinaxis.com/2016/04/17617/ (visited on 05/03/2018).

[3] Ghassen Chniti, Houda Bakir, and Hédi Zaher. "E-commerce Time Series Forecasting Using LSTM Neural Network and Support Vector Regression". In: *Proceedings of the International Conference on Big Data and Internet of Thing.* BDIOT2017. London, United Kingdom: ACM, 2017, pp. 80–84. ISBN: 978-1-4503-5430-1. DOI: 10.1145/3175684.3175695. URL: http://doi.acm.org/10.1145/3175684.3175695.

[4] Nadav Cohen et al. "In-House Solution for the RecSys Challenge 2015". In: *Proceedings of the 2015 International ACM Recommender Systems Challenge.* RecSys '15 Challenge. Vienna, Austria: ACM, 2015, 10:1–10:4. ISBN: 978-1-4503-3665-9. DOI: 10.1145/2813448.2813519. URL: http://doi.acm.org/10.1145/2813448.2813519.

[5] xgboost developers. *XGBoost Documentation.* 2018. URL: https://xgboost.readthedocs.io/en/latest/ (visited on 08/20/2018).

[6] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "Advances in Knowledge Discovery and Data Mining". In: ed. by Usama M. Fayyad et al. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. Chap. From Data Mining to Knowledge Discovery: An Overview, pp. 1–34. ISBN: 0-262-56097-6. URL: http://dl.acm.org/citation.cfm?id=257938.257942.

[7] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* http://www.deeplearningbook.org. MIT Press, 2016.

[8] Apache Hadoop. *HDFS Architecture.* 2018. URL: http://hadoop.apache.org/docs/r3.1.0/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html (visited on 06/13/2018).

[9] Balzs Hidasi et al. "Session-based Recommendations with Recurrent Neural Networks". In: (Nov. 2015).

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: 521 (May 2015), pp. 436–44.

[11] Francesco Ricci, Lior Rokach, and Bracha Shapira. "Recommender Systems: Introduction and Challenges". In: *Recommender Systems Handbook.* Ed. by Francesco Ricci, Lior Rokach, and Bracha Shapira. Boston, MA:

Springer US, 2015, pp. 1–34. ISBN: 978-1-4899-7637-6. DOI: `10.1007/978-1-4899-7637-6_1`. URL: `https://doi.org/10.1007/978-1-4899-7637-6_1`.

[12] Peter Romov and Evgeny Sokolov. "RecSys Challenge 2015: Ensemble Learning with Categorical Features". In: *Proceedings of the 2015 International ACM Recommender Systems Challenge*. RecSys '15 Challenge. Vienna, Austria: ACM, 2015, 1:1–1:4. ISBN: 978-1-4503-3665-9. DOI: `10.1145/2813448.2813510`. URL: `http://doi.acm.org/10.1145/2813448.2813510`.

[13] Yong Kiam Tan, Xinxing Xu, and Yong Liu. "Improved Recurrent Neural Networks for Session-based Recommendations". In: *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. DLRS 2016. Boston, MA, USA: ACM, 2016, pp. 17–22. ISBN: 978-1-4503-4795-2. DOI: `10.1145/2988450.2988452`. URL: `http://doi.acm.org/10.1145/2988450.2988452`.

[14] TensorFlow. *TensorFlow - An open source machine learning framework for everyone*. 2018. URL: `https://www.tensorflow.org/` (visited on 08/20/2018).

[15] Machine Learning Group at the University of Waikato. *Weka 3: Data Mining Software in Java*. 2018. URL: `https://www.cs.waikato.ac.nz/ml/weka/` (visited on 06/28/2018).

[16] Eduard Weigandt. "Auf Data-Mining basierende Personalisierung im E-Commerce mit implizitem Feedback". Masterarbeit. Hochschule fr Angewandte Wissenschaften Hamburg, 2016.

[17] Ye Wu and Rick Radewagen. *7 Techniques to Handle Imbalanced Data*. 2018. URL: `https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html` (visited on 08/03/2018).