**Evaluation of Machine Learning and Deep Learning Models for Customer Journey Prediction in E-Commerce**

**An Executive Perspective**

**Contents**

List of Tables

List of Figures

Abstract

1. Introduction
2. Methodology
3. Related Work
   1. Comparative Machine Learning Studies
   2. Comparative Machine Learning Studies Focused on Customer Journey Prediction in E-Commerce
4. Evaluation Framework
5. Experiments
   1. Experimental Setup
   2. Data
   3. Descriptive Statistics
   4. Models
6. Evaluation
   1. Objectivity
   2. Predictive Accuracy
   3. Robustness
   4. Interpretability
   5. Versatility
   6. Algorithmic Efficiency
7. Discussion and Managerial Implications
8. Conclusion

Acknowledgments

References

Appendix

1. Introduction

* Introduction, motivation, (business) relevance
* Research question: How do different machine learning and deep learning models perform on the problem of customer journey prediction in e-commerce in comparison and particularly with regard to criteria that are relevant to marketing managers and executives?
* Hype about deep learning: Popularity of deep learning in computer vision, speech recognition and natural language processing - but does it really always have to be deep learning or do other models also suffice?
* Business problem: How can businesses predict customer behavior using clickstream data to timely react to customer behavior using personalization and couponing for example?
  + Which users are most likely to purchase (predict purchasing intent) (Sheil et al. 2018)
  + Which elements of the product catalogue do users prefer (rank content) (Sheil et al. 2018)
* Algorithmic trust and trust in models in general, Grahl's literature suggestions: Dietvorst et al. (2016), Brynjolfsson and Mitchell (2017), Hall et al. (2017), Logg et al. (2018)

Section 2 explains the methodology chosen to conduct the present thesis; Section 3 analyzes related studies and motivates the choices of models, metrics and methods to be implemented and examined in the subsequent experiments; Section 4 introduces the framework applied to evaluate the models used to predict customers' purchasing behavior; Section 5 explains the experiments, the data and the models; Section 6 evaluates the experiments along the dimensions of the evaluation framework; Section 7 discusses the findings and derives implications for marketing managers and executives; Section 8 concludes with a summary and an outlook for future research.

1. Methodology

The objective of the present thesis is to compare different machine learning and deep learning models for the application of customer journey prediction in e-commerce, applying a framework that allows the evaluation of the examined models with the regard to the fact that these models are supposed to support marketing managers and executive in their decision making. Therefore, the methodology for conducting a sound comparative study consists of a theoretical approach that is divided into two parts.

First, related studies that compare different machine learning and deep learning models are presented. The first part of this meta-analysis of comparative studies comprises general studies on comparing models for classifications tasks for a variety of different use cases and data sets, while the second part specifically focusses on studies that compare models for the application of customer journey prediction. To keep the meta-analysis concise, only the most relevant comparative studies have been considered based on a catalogue of specific selection criteria that will be later explained in more detail. This procedure results in 15 studies being selected for the general part of the meta-analysis and ten studies being selected for the second part of the meta-analysis. Both parts of the meta-analysis consider different dimensions of the examined studies, such as the choice of models, the size of the data sets, the type of classification tasks and the choice of evaluation metrics. The objective of this meta-analysis is to leverage experiences and findings of previous comparative studies on one hand and to identify frequently used and state-of-the-art models on the other hand to serve as the orientation for the model selection of the present thesis. This approach allows to answer the interesting question whether there are models or model families that tend to dominate others or whether there is no such model or model family that tends to dominate all others. There are other studies that choose a similar approach, analyzing and criticizing previous comparative studies and leveraging their learnings (e.g. Michie et al., 1994; Kind et al., 1995; Van den Poel and Buckinx, 2005; Baumann et al., 2018).

Second, an evaluation framework is introduced that builds on what management and marketing literature found to be important key factors for models in general to be successfully used in practice by marketing managers and executives (e.g. Little, 1970; Lodish, 2001; Lilien and Rangaswamy, 2004; Little, 2004; Lilien, 2011). Anderl et al. (2014) condense this research into a framework that they use to evaluate online attribution models for mapping the customer journey. Apart from predictive accuracy their evaluation framework includes objectivity, robustness, interpretability, versatility and algorithmic efficiency as the criteria along which they evaluate online attribution models. The application of marketing models in practice is difficult and the most complex model does not necessarily turn out to be the one that has the largest impact on a company in terms of productivity (Anderl et al., 2014, p. 7). Besides, it is evident that it can be inadequate to compare models solely on the ground of a single quantitative metric such as accuracy or the area under the receiver operating characteristic curve (AUC). Anderl et al.’s model evaluation framework mitigates this issue by incorporating not only quantitative metrics but also stresses the importance of dimensions such as robustness and interpretability for instance (Anderl et al., 2014, p. 8). Although they designed their evaluation framework to evaluate attribution models, it can be easily adjusted to be a valid evaluation framework for evaluating machine learning and deep learning models for the application of customer journey prediction as well. The objective of using this adjusted evaluation framework is to allow for a comprehensive and thorough comparison of different models beyond the calculation of several quantitative metrics, including qualitative and those dimensions that are particularly relevant to marketing managers and executive as well.

1. Related Work

Section 3.1. presents a selection of related studies that compare different machine learning and deep learning models in general, mainly comparing their predictive performance on classification tasks for a variety of different use cases and data sets. Section 3.2. explicitly focusses on studies that compare machine learning and deep learning models for the application of purchase prediction in e-commerce. Both sections first introduce the criteria applied to select the most relevant studies from a large body of comparative machine learning literature. Then, the selected studies are briefly presented while the most notable are particularly highlighted. Finally, both sections present the most frequently used models, respectively, which form the choice of models that are compared in the later experiments. Since machine learning and deep learning have only gained widespread popularity in recent years, Section 3.2. mentions several studies from the marketing context that have applied models that have been typically used in marketing science before machine learning gained popularity in marketing science as well. The objective of Sections 3. is to conduct an analysis of existent comparative machine learning and deep learning studies on a meta-level to leverage previous research in this field to form a sound foundation for this thesis with regard to the choice of models, evaluation metrics and other analytical methods.

* 1. Comparative Machine Learning Studies
* criteria for selection of studies due to their mere mass
  + focus on supervised learning
  + focus on general studies rather than studies that focus on a specific use case or field
  + focus on studies that compare at least 3 different families of models or models (by different models it is not meant different variants of models with only some minor modifications, but rather variations that lead to fundamentally different models)
  + focus on studies that use real data
  + focus on studies that use more than 1 real-world dataset
  + focus on studies that appeared in renowned publications or conferences
* summary and criticism of comparative studies
* summary of most notable studies and results (Michie et al. 1994 - user guide, complete Statlog project, King et al. 1995 another study analyzing Statlog)
* Table 1: Literature overview (authors, models, data, metrics, methods, results)
* Table 2: Top 10 most frequently used and/or recommended models
  1. Comparative Machine Learning Studies Focused on Customer Journey Prediction in E-Commerce
* criteria for selection of studies due to their mere mass
  + focus on studies dealing with customer journey prediction using e-commerce clickstream data
  + focus on studies using supervised machine learning models
  + focus on studies that compare at least two models
  + focus on studies that appeared in renowned publications or conferences
* note that previous marketing literature focused on models such as Markov models and logit models
* summary and criticism of comparative literature
* summary of most notable studies and results
* Table 1: Literature overview (authors, models, data, metrics, methods, results)
* Table 2: Top 10 most frequently used and/or recommended models
* Interesting non-comparative Marketing Literature:
  + Moe et al. 2002 (model)
  + Montgomery et al. 2004 (multinomial probit model)
  + Sismeiro and Bucklin 2004 (model, managerial implications)
  + Van den Poel and Buckinx 2005 (features, literature review, managerial implications)
  + Stange and Funk 2015 (managerial implications)
  + Baumann et al. 2018 (model, literature review)

1. Evaluation Framework

According to Lilien (2011), there appears to be a divide between the models developed for marketing decision support in academia and their actual application by practitioners in the field – despite the potential benefits marketing decision support systems (MDSS) might yield. He argues that MDSS improve consistency of decisions, which is desirable for frequently made decisions and they enable the exploration of a multitude of decision options that cannot be all properly evaluated individually by managers due to the mere mass of available options (Lilien, 2011, pp. 197-198). In contrast, in cases where there are only few decision options but many different variables that influence a decision, MDSS allow managers to assess the relative impact of individual decision variables (Lilien, 2011, p. 198). They also facilitate group decision making and help decision makers update their mental models of their markets they have built through experience (Lilien, 2011, pp. 198-199). For all the benefits MDSS offer, there are several reasons for why managers resist to apply them nonetheless. Lilien and Rangaswamy (2004) find four main reasons why managers lack to adopt generally beneficial models: (1) managers’ mental models are oftentimes satisfactory, particularly in environments that are highly predictable; (2) it is very likely that humans will always be involved in decision making in marketing environments because they solve problems and not models themselves; (3) since managers cannot generally quantify the opportunity costs of the decisions they make, it is hard for them to see the potentially superior outcome a model could have produced instead; and finally (4) because managers tend to prefer ambiguity and intuition over precision and analysis, which are decisive for applying models, they prefer to refuse to use them.

Little (1970, 2004) finds several more potential reasons why models appear to not be used by managers more widely. He states that, first, good models that satisfy managers are simply hard to find, second, good parametrization of the models (which crucially depends on quality data and measurements) is even harder to accomplish, third, managers that do not understand the models they are supposed to base their decisions on tend to refuse to use them and finally, models themselves are, in spite of attempts to optimize them, incomplete (Little, 1974, pp. 467-468; Little, 2004, pp. 1841-1842). To mitigate these issues and make managers apply models in their decision making more widely, Little (1974, 2004) suggests several requirements for models. According to his suggestions, models should be simple because “[s]implicity promotes ease of understanding (…)” which in turn favors adoption (Little, 1974, p. 470; Little, 2004, p. 1843). Models should also be robust, meaning they are unlikely to deliver output that is bad or useless (Little, 1974, p. 470; Little, 2004, p. 1843). Besides, managers should be able to control models in the sense that they understand how inputs and outputs are connected, so that model results are in line with what managers expect in general (Little, 1974, p. 470; Little, 2004, pp. 1843-1844). It is important that models are adaptive to allow for updating parametrization or structure, given there are changes in the model’s environment that affect it and its output (Little, 1974, p. 470; Little, 2004, pp. 1844). Furthermore, models are required to be complete in important situations, although contrasting on the notion of simplicity, implying that the inclusion of managers’ subjective judgements might be desirable in certain situations (Little, 1974, p. 470; Little, 2004, pp. 1844). Lastly, models should be “(…) [e]asy to communicate with (…)” so that managers are “(…) able to change inputs easily and obtain results quickly (…)”, which is made possible by technologies such as on-line, conversational and time-shared computing interfaces (Little, 1974, p. 470; Little, 2004, pp. 1844).

Lodish (2001) presents a personal summary on the relationship of models and productivity based on his academic and practical experience, i.e. issues that interfere with productivity and lessons learned on this relationship. He reports that a “(…) model may never be finished enough for someone (…)” to use it and even if it were finished, it may still not be used after all (Lodish, 2001, p. 46). He continues that a “(…) manager may not use a model’s results to improve (…)” decision making and even if a manager were to use the model’s results for decision making, “(…) the decisions may not improve productivity (…)” (Lodish, 2001, p. 46). Lodish (2001) derives the following general lessons from his experience: according to him, it is an art to build scientific models that are able to improve productivity; it is critical to consider implicit model trade-offs and phenomena that are relevant to managers’ decision making; it might make sense to take into consideration subjective estimates, risk, the balance between complexity and ease of understanding and strategy (Lodish, 2001, pp. 53-54). He also highlights several lessons explicitly related to subjective estimates, namely the importance of involving managers in the model building process and the efficient use of their time and resources, and empirical models, namely the importance of on-demand measures aggregated to the right level that are credible and therefore able to persuade a manager (p. 54). Lodish (2001) concludes his paper with a concise guiding principle: “The criterion for a good, productive model is not whether it is theoretically or empirically perfect. It is, will the manager’s decision, based on the model, improve productivity enough to justify the costs and resources devoted to developing and using the model?” (p. 54).

Anderl et al. (2014) condense, among other, the research on the application and acceptance of marketing models mentioned above into an evaluation framework to assess online attribution models in a comprehensive and concise manner. Their evaluation framework comprises six criteria: objectivity, predictive accuracy, robustness, interpretability, versatility and algorithmic efficiency (Anderl et al., 2014, pp.7-10). The criterion of ***objectivity*** is defined as a model’s ability to assign credit to specific features in the data that factually contribute to the objective of the application the model is applied to, e.g. increasing the number of purchase events in an online shop (i.e. conversions) or revenue (Anderl et al., 2014, p. 7). Objectivity originates from Lilien’s claim for a model to allow for the computation of a variable’s relative impact and the objective evaluation of available decision options (Lilien, 2011, p. 198). ***Predictive accuracy*** is defined as a model’s ability to correctly predict conversions (Anderl et al., 2014, pp. 8), picking up Lodish’s lesson of the importance of a model’s credibility to persuade managers (Lodish, 2001, p. 54). ***Robustness*** is defined as a model’s ability to deliver “(…) stable and reproducible results (…)” after multiple runs of the model (Anderl et al., 2014, pp. 8), covering Little’s requirement for a model to return useful results (Little, 1974, p. 470; Little, 2004, p. 1843). According to Little, models should be simple (Little, 1974, p. 470; Little, 2004, p. 1843) and easy to communicate with (Little, 1974, p. 470; Little, 2004, pp. 1844), which Anderl et al. translate to the criterion of ***interpretability***, defined as the fact that a model’s structure and results should be transparent and understandable to all stakeholders involved with reasonable effort (pp. 8). ***Versatility*** incorporates Little’s requirements that models should be easy to control (Little, 1974, p. 470; Little, 2004, pp. 1843-1844) and to adapt (Little, 1974, p. 470; Little, 2004, pp. 1844), i.e. models should allow for the inclusion of novel information and data in rapidly and frequently changing environments through a high degree of flexibility (Anderl et al., 2014, p. 10). ***Algorithmic efficiency*** builds upon Lodish’s lesson that models should ideally deliver results on-demand (Lodish, 2001, p. 54), i.e. when managers need them, which is particularly important when dealing with large amounts of data (Anderl et al., 2014, p. 10).

Although Anderl et al. (2014) designed this framework to evaluate online attribution models, it generalizes well given that it builds upon research that explores the application and requirements of marketing models in general. Therefore, the evaluation framework presented above can be easily transferred to the application of customer journey prediction. The framework’s six criteria are applied in Section 6. to evaluate the experiments on predicting customers’ purchasing behavior in detail, using a multitude of machine learning and deep learning models and respective performance metrics.

1. Experiments

Section 5.1. presents the setup of the following experiments, covering the choice of models, targets, features and data samples in general as well as naming the tools and software packages used to conduct the experiments. Section 5.2. provides detailed information on the data used in the experiments, the choices made during processing the data and the techniques applied to transform the raw data to training and test sets ready to be used for modeling. To put the data and the application of customer journey prediction in context, Section 5.3. first defines important expressions and then analyzes the data in a descriptive manner to generate first insights for customer journey prediction. The targets and features used in the experiments are examined as well. Section 5.4. finally introduces the models used in the experiments along with some explanations, stating and justifying important choices that have been made with regard to implementation, parameter choice, training and testing. For the sake of brevity, the models are not explained in too much detail but secondary references are provided for the interested reader. The objective of Section 5 is to explain the experimental setup in detail, along with the approach and the reasoning on the choices of models, targets, features, samples and methods applied to process the raw data.

* 1. Experimental Setup
* 13 models
* 2 targets: explain targets and choice using literature, mention class imbalance
* 5 samples: explain sampling and choice
* 130 experiments
* technical setup (tools and software packages, e.g. Linux workstation, scikit-learn, keras DeviceDetector, Jupyter Notebook, MikTeX, pandoc etc. -> pip freeze)
  1. Data
* origin, type, size, period
* cleaning, mapping, aggregation, preparing targets and features, sampling unique visitors, feature selection, preprocessing (incl. dimensions after each step if applicable, e.g. raw/clean hits, first vs last, bounces etc.)
  1. Descriptive Statistics
* definitions of visitors/journeys, sessions/visits, events/clicks/hits etc.
* Figure 1: visits, purchases, page and product views over time
* Table 1: descriptive statistics for 5 samples (unique visitors, visits, journeys with length >= 2, journeys with length >= 5, average journey length, unique days, features, conversions 24 hours, conversion rate 24 hours, conversions 7 days, conversion rate 7 days
* Table 2: targets, features, data types, levels, summary statistics
  1. Models
* train test split sufficient because of large dataset: Raschka (2018)
* default parameters and hyperparameter tuning, Random Search: Bergstra and Bengio (2012)
* class imbalance, SMOTE: Chawla et al. (2002)
* LR
* DT
* NB
* KNN
* RF
* SVM
* BOOST
* BAG
* NN1/NN3/NN5 -> justify params via literature (e.g. G. Hinton and dropout wegen overfitting), chosen for no specific reason, references for epochs (not too many due to resources) and batch size (default 32), rules of thumb, footnote on I tensorflow/core/platform/cpu\_feature\_guard.cc:141] Your CPU supports instructions that this TensorFlow binary was not compiled to use: AVX2 FMA
* RNN/LSTM -> not vector-based model like the others but sequence model, does not require features like days since last visit or purchase, for parameter choice: Lang und Rettenmeier (2017), Toth et al. (2017), Sheil et al. (2018)

1. Evaluation

Section 6.1. relates the models to the dimension of objectivity. Section 6.2. analyzes the experimental results in detail along the dimension of predictive accuracy, applying a variety of different metrics to evaluate model performance, such as accuracy, AUC and F-score. To test the robustness of the experimental results, Section 6.3. first explores learning curves as a means to visualize the importance of sample size and then presents results form a 5-fold cross validation for each model. The models used in the experiments vary widely in terms of complexity and therefore interpretability. Therefore, Section 6.4. stresses the importance of interpretability in different contexts and relates the notion of interpretability to the models under examination and the role of their complexity. Furthermore, Section 6.4. highlights different methods to identify the most relevant features and briefly mentions techniques used to make models and their predictions more interpretable. Section 6.5. considers the models from the viewpoint of versatility. Complexity and model-specific idiosyncrasies determine algorithmic efficiency, which is why Section 6.6. analyzes the models’ efficiency with regard to the time they required for training and testing on samples of different sizes. The evaluation of the models and the experiments in Section 6 constitutes the core of the present thesis and forms the foundation for the discussion and the derivation of managerial implications in Section 7.

* 1. Objectivity
  2. Predictive Accuracy
* Table 1: accuracy, AUC, true negatives, false negatives, true positives, false positives, precision, recall, F-score
* ROC
* statistical (non-parametric) tests, e.g. t-test, Wilcoxon test, Friedman test: Dietterich (1998), Alpaydin (1999), Demsar (2006), Lavesson and Davidsson (2007), Raschka (2018)
  1. Robustness
* Figure 1: learning curves
* 5-fold cross validation on 100k unique visitors sample for reasons of efficiency and resources
  1. Interpretability
* model type and complexity
* Table 1: k best features, weights from logistic regression, feature importances from decision tree, random forest and boosting ensemble
* overview of studies and methods
  1. Versatility
  2. Algorithmic Efficiency
* training and testing times

1. Discussion and Managerial Implications

* select one model and optimize it
* inspiration: Sismeiro and Bucklin, 2004; Van den Poel and Bucklin, 2005; Stange and Funk, 2015
* select model with regard to false positive and false negative: do you want to reach a large mass of potential buyers or do you want to reach visitors that are very likely to buy?

1. Conclusion

* Summary of research question, relevance, experiments and results
* Limitations and future research:
  + more data
  + different data
  + different use cases/applications
  + different models
  + further test effect of different targets/features
  + shorter time period
  + predict purchase probabilities rather than purchase or not purchase
  + more/better hyperparameter tuning
  + extend to item prediction/recommender system
  + expert interviews with marketing managers and executives on what really counts for them when it comes to models
  + use tools such as LIME for model interpretability
  + focus on the evaluation of decision support systems rather than models themselves
  + segmentation of visitors and separate models for individual visitor segments