Evaluation of Machine Learning and Deep Learning Models for Customer Journey Prediction in E-Commerce - An Executive Perspective

Alexander Merdian-Tarko

April 5, 2019

1. Introduction (2 pages)

   - Introduction, Motivation and (Business) Relevance
   - Topic: Evaluation of Machine Learning and Deep Learning Models for Customer Journey Prediction in E-Commerce - An Executive Perspective
   - Research question: How do different machine learning and deep learning models perform on the problem of customer journey prediction in e-commmerce 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?
   - 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)
   - Cover page, citation style + Citavi, abstract, eidesstattliche Erklärung, note on folders, files and how to generate pdf from jpnb in readme
   - Sheil et al. (2018):
     - Which users are most likely to purchase (predict purchasing intent).
     - Which elements of the product catalogue do users prefer (rank content).

Section 2 explains the methodology chosen to conduct the present thesis; Section 3 analyzes related studies and motivates the choices of models to be implemented and examined in the 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 in more detail; 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.

2. Methodology (2 pages)

- meta-analysis of comparative literature for motivation of choice of models, methods and metrics -> leveraging existing knowledge and sound foundation for comparing models
- evaluation framework based on adapted attribution model evaluation framework developed by Anderl et al. (2014), including marketing, management and comparative machine learning literature -> sound foundation of evaluating models across different dimensions apart from accuracy only

3. Related Work (4-6 pages)

3.1. Comparative Machine Learning Literature

- criteria for literature selection due to the meer mass of literature available (max. 20 studies):
  - 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 literature
- summary of most notable studies and results (Michie et al. 1994 - user guide, complete statlog project, King et al. 1995 another study analyzing statlog)
- explain dimensions of meta analysis (columns)
- table 1: authors, algorithms/models, data, methods, metrics, results (pull automatically from CSV, wrangle and transform to LaTeX)
- table 2: top 10 most frequently used and recommended models (pull automatically from CSV, wrangle and transform to LaTeX)

3.2. Comparative Marketing Literature dealing with Customer Journey Prediction using E-Commerce Clickstream Data

- criteria for literature selection due to the meer mass of literature available (max. 10-15 studies):
  - 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
- explain dimensions of analysis (columns)
- table 1: authors, algorithms/models, data, methods, metrics, results (pull automatically from CSV, wrangle and transform to LaTeX)

- table 2: top 10 most frequently used and recommended models (pull automatically from CSV, wrangle and transform to LaTeX)
- 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)

Table 1 - 3.1. Comparative Machine Learning Literature

Study	Models	Datasets	Instances	Type	Classes	Evaluation metrics
Weiss and Kapouleas (1989)	DA, KNN, BM, NN, RL, DT (9)	4	106-3772	R	B/M	Accuracy
Shavlik et al. (1991)	DT, NN (3)	5	226-11500	R	B/M	Accuracy, time
Michie et al. (1994)	DA, DT, RL, LR, KNN, BM, NB, NN (23)	22	270-58000	R/S	B/M	Time, accuracy, ranking, cost, complexity, qualitative evaluation
King et al. (1995)	DT, RL, NB, KNN, DA, LR, BM, NN, MISC (17)	12	270-58000	R	B/M	Accuracy, cost, time, qualitative evaluation
Bradley (1997)	DT, NN, KNN, DA (9)	6	117-768	R	B	Accuracy, AUC, ROC, ANOVA, tests
Bauer and Kohavi (1999)	DT, NB, BAG, BOOST (7)	14	1000-58000	R	B/M	Accuracy, bias variance decomposition
Lim et al. (2000)	DT, RL, DA, LR, NN (33)	32	151-4435	R/S	B/M	Accuracy, time, complexity
Huang et al. (2003)	NB, DT, SVM (4)	18	132-8124	R	B/M	Accuracy, AUC
Perlich et al. (2003)	DT, LR, BAG (8)	36	700-1000000+	R	B	Accuracy, learning curve, ranking, AUC
Provost and Domingos (2003)	DT, BAG (5)	25	unclear, but incl. large-scale datasets	R	B/M	Ranking, AUC
Tan and Gilbert (2003)	DT, RL, NB, KNN, SVM, NN, STC, BAG, BOOST (17)	4	106-1484	R	B/M	Accuracy, precision, recall

Study	Models	Datasets	Instances	Type	Classes	Evaluation metrics
Caruana and Niculesci-Mizil (2006)	SVM, NN, LR, NB, KNN, RF, DT, BAG, BOOST (30)	11	9366-40222	R	B	Accuracy, F-score, lift, ROC, precision, recall, cross-entropy
Fernandez-Delgado et al. (2014)	DA, NB, BM, NN, SVM, DT, RL, BOOST, BAG, STC, RF, ENS, GLM, KNN, LR, REG, MISC (179)	121	10-130064	R	B/M	Rankinging, accuracy, Cohen k, tests
Khan et al. (2018)	KNN, SVM, ENN, LMNN (4)	11	~200 to ~5000	R	B/M	Accuracy
Olson et al. (2018)	NB, LR, SVM, KNN, DT, RF, BOOST, MISC (13)	165	mostly < 5000 and incl. large-scale datasets	R	B/M	Ranking, accuracy, tests

Table 2 - 3.1. Comparative Machine Learning Literature

Models	Occurrences
DT	14
KNN	9
NN	9
NB	8
LR	7
BAG	6
DA	6
RL	6
SVM	6
BOOST	5

Table 1 - 3.2. Comparative Marketing Literature

```
In [5]: from IPython.display import Latex
```

```
In [9]: %%latex
        \begin{table}[]
        \begin{tabular}{|l|l|l|l|}
        \hline
        Study                       & Models                            & Data
        Moe and Fader (2004)        & CM, LR, MISC (6)                  & 8 months, 4k visitors
        Suh et al. (2004)           & DT, NN, LR, ENS (4)               & 1 day, 1 mio. records
        Boroujerdi et a. (2014)     & DT, RF, LR, NN, SVM, DR, KNN (16) & 60k sessions, 23-82
        Wu et al. (2015)            & RNN, NN, BOOST (3)                & 11,8 mio. sessions, 6
```

Study	Models	Data
Moe and Fader (2004)	CM, LR, MISC (6)	8 months, 4k visitors, 11k visits
Suh et al. (2004)	DT, NN, LR, ENS (4)	1 day, 1 mio. records, 170k users, 21 feat
Boroujerdi et a. (2014)	DT, RF, LR, NN, SVM, DR, KNN (16)	60k sessions, 23-82 features
Wu et al. (2015)	RNN, NN, BOOST (3)	11,8 mio. sessions, 6 months (Ben-Shimo
Sarwar et al. (2015)	NB, RF, BM, LR, BOOST (5)	9,3 mio. sessions, 22 features
Vieira (2016)	LR, DT, RF, DBN, SDA (5)	1,5 mio. sessions in test set only, 6 montl
Iwanaga et al. (2016)	CM, LR, SVM (6)	44k customers, 1 month training, 1 mont
Zhao et al. (2016)	RF, LR, NB, SVM, MISC (5)	183k actions, 5 months, 27 features
Lang and Rettenmeier (2016)	LR, NN, RNN (3)	6 weeks, 20-23 features
Toth et al. (2017)	MM, RNN (2)	200k sessions, 2 weeks, include only sess
Nishimura et al. (2018)	CM, LR, RF, NN (6)	2 months, 4 mio. customers, top 1% purc
Sheil et al. (2018)	RNN, BOOST (4)	RecSys Challenge 2015 9,2 mio user sess
Sakar et al. (2018)	DT, RF, SVM, NN (4)	12k sessions

```
Sarwar et al. (2015)        & NB, RF, BM, LR, BOOST (5)          & 9,3 mio. sessions, 22
Vieira (2016)               & LR, DT, RF, DBN, SDA (5)           & 1,5 mio. sessions in
Iwanaga et al. (2016)       & CM, LR, SVM (6)                    & 44k customers, 1 mon
Zhao et al. (2016)          & RF, LR, NB, SVM, MISC (5)          & 183k actions, 5 montl
Lang and Rettenmeier (2016) & LR, NN, RNN (3)                    & 6 weeks, 20-23 featu
Toth et al. (2017)          & MM, RNN (2)                        & 200k sessions, 2 wee
Nishimura et al. (2018)     & CM, LR, RF, NN (6)                 & 2 months, 4 mio. cus
Sheil et al. (2018)         & RNN, BOOST (4)                     & RecSys Challenge 201
Sakar et al. (2018)         & DT, RF, SVM, NN (4)                & 12k sessions
\end{tabular}
\end{table}
```

```
In [10]: %%latex
         \begin{align}
         \nabla \times \vec{\mathbf{B}} -\, \frac1c\, \frac{\partial\vec{\mathbf{E}}}{\partial
         \nabla \cdot \vec{\mathbf{E}} & = 4 \pi \rho \\
         \nabla \times \vec{\mathbf{E}}\, +\, \frac1c\, \frac{\partial\vec{\mathbf{B}}}{\partia
         \nabla \cdot \vec{\mathbf{B}} & = 0
         \end{align}
```

$$\nabla \times \vec{\mathbf{B}} - \frac{1}{c}\frac{\partial \vec{\mathbf{E}}}{\partial t} = \frac{4\pi}{c}\vec{\mathbf{j}} \tag{1}$$

$$\nabla \cdot \vec{\mathbf{E}} = 4\pi\rho \tag{2}$$

$$\nabla \times \vec{\mathbf{E}} + \frac{1}{c}\frac{\partial \vec{\mathbf{B}}}{\partial t} = \vec{\mathbf{0}} \tag{3}$$

$$\nabla \cdot \vec{\mathbf{B}} = 0 \tag{4}$$

Table 2 - 3.2. Comparative Marketing Literature

Models	Occurrences
LR	6
NN	5
RF	4
RNN	4
DT	3
BOOST	3
SVM	2
KNN	1
NB	1
RL	1

4. Model Evaluation Framework (2-3 pages)

    - based on attribution model evaluation framework proposed by [Anderl, E., Becker, I., Wangenheim, F. V., & Schumann, J. H., 2014], including following dimensions:
        - objectivity
        - predictive accuracy (e.g. accuracy, AUC, precision, recall, F-score)
        - robustness
        - interpretability (e.g. model type and complexity)
        - versatility
        - algorithmic efficiency (e.g. training and testing times)
    - management and marketing literature used by [Anderl, E., Becker, I., Wangenheim, F. V., & Schumann,
        - [Lilien, 2011]
        - [Little, 1970]
        - [Little, 2004]
        - [Lodish, 2001]

5. Experiments (5-6 pages)

5.1. Experimental Setup and Data

- explain target: purchase within current visit, within next visit, within next 24 hours or within next 7 days -> focus on purchase within next visit and within 24 h
- features: crafting features by hand, feature engineering and selection (e.g. feature importance or recursive feature elimination) -> explore levels, distribution and potential bias of categorical features, binarize/discretize/normalize numerical features + overview table with features and their data types, # levels etc.
- sampling method and size: stratified sampling with 6 weeks and 250k visitors, 12 weeks and 500k visitors and 24/max. weeks and 1 mio. visitors
- class imbalance: depends on choice of target
- hyperparameter optimization: random search
- use literature for legitimation of choices and decisions when it comes to models, samples, dropping of visits/visitors -> drop bounce visits
- workstation (operating system + version, memory/RAM, CPU, GPU)
- scikit-learn +version (LR, DT, NB, KNN, RF, SVM, BOOST, BAG) (pip freeze)

- keras +version (NN, RNN)
- data processing: aggregation from hit to visit level, remove visits with only 1 hit (bounce), first vs last
- Features (Boroujerdi et al. 2014, Sawar et al. 2015, Romov and Sokolov 2015, Wu et al. 2015, Zhao et al. 2016, Lang and Rettenmeier 2017, Sakar et al. 2018)
- visits, page views, product views, purchases in last n hours/days (binarize)
- correlation between device type, brand, browser, search engine and os
- first-last aggregation: static features such as age, gender, device type, brand, os and connection type
- feature importance/selection (e.g. add to basket, checkout, cart value, age, gender, nps etc. appear to be correlated with the target)
- Software and packages (Ubuntu, Anaconda, Python, regex, DeviceDetector, numpy, pandas, scikit-learn, keras, matplotlib, seaborn, Jupyter Notebook, MikTeX, pandoc)

5.2. Descriptives

- definitions, e.g. hits vs events vs clicks, sessions vs visits vs journeys
- pull descriptives automatically from samples/output files, wrangle and transform to LaTeX table
- descriptives for each sample including the following information:
    - hits/events/clicks
    - sessions/visits/journeys
    - unique visitors
    - journey length/number of visits per visitor
        - thereof length >= 2
        - thereof length >= 5
    - average journey length
    - conversions
    - conversion rate
    - Verteilung visits, purchases pro Woche/Monat

5.3. Models

- LR
- DT
- NB
- KNN
- RF
- SVM
- BOOST
- BAG
- NN (hidden layers: 1, 3, 5) -> 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 (Lang und Rettenmeier 2017, Toth et al. 2017, Sheil et al. 2018, for parameter choice)

- no in depth explanations of models but provide references for further reading
- overview table
- optimal hyperparameters derived by random search

5.4. Methods

- train test split sufficient because of large dataset (Raschka, 2018)
- class imbalance (depends on choice of target and sample, computationally expensive) -> SMOTE ([Chawla et al., 2002])
- hyperparameter optimization (computationally expensive) -> random search ([Bergstra, J., & Bengio, Y., 2012])

5.5. Priorisierung und Zielsetzung (5 Wochen bis zur Abgabe) - Ordentliche Meta-Analyse der vergleichenden Machine Learning Literatur mit allgemeinem bzw. Marketing-Fokus (Fokus auf Purchase Prediction mit Machine Learning auf Clickstream Daten bei der Marketing Literatur) - Konzeptionell und methodisch saubere Implementierung der Purchase Prediction auf dem Siroop Clickstream mit verschiedenen Modellen und deren default Parametern inkl. der Begründung von bestimmten Entscheidungen in Bezug auf targets, features, sampling etc. - Schlüssige Evaluation der Modelle anhand der Dimensionen des Model Evaluation Framework von Anderl et al. (2014) (objectivity, predictive accuracy, robustness, interpretability, versatility, algorithmic efficiency) - Umfassende Diskussion der Managerial Implications, Limitations und Future Research (weitere Experimente mit verschiedenen Modifikationen von sampling, targets [e.g. predict purchase probability], und features, statistical und non-parametric tests für Modellvergleiche, Segmentierung der visitors und separate Modelle für einzelne Segmente, Recommender/Prediction für Produkte etc.) - *Fragen und Diskussionspunkte* - drop bounce visits: 6,5 Mio. -> 3,3 Mio. visits -> Kai kann sich nicht mehr wirklich an genaue Zahlen dazu erinnern - Korrelation zwischen categorical features: operating system, device type, brand, browser und search engine? - first – last Aggregation von categorical features: age, gender, connection type, device type, brand und os statisch/unveränderlich? - feature importance/engineering/selection (e.g. add to basket, checkout, cart value, age, gender, nps scheinen stark mit dem target zu korrelieren) - targets: purchase within current visit, next 24 hours, next 24 days and next visit -> erst mal Fokus auf purchase within next 24 hours und purchase within next visit - number weeks vs. number visitors sampling + Reihenfolge der visits beim train test split (check journey lengths, visits und purchases pro Monat) (s.u.) - SMOTE für class imbalance -> scheint erst mal doch nicht so prekär zu sein (conversion rate von 2,5-3,5% je nach target) - hyperparameter tuning -> erst mal allgemeiner Modellvergleich mit default parametern, dann die Vielversprechendsten auf Basis der Evaluation auswählen und optimieren (Outlook/Future Research) - *Gedanken zu Sampling und train test split* - Es geht im Grunde immer noch (bzw. wieder) darum, wie ich den gesamten Datensatz in kleinere Samples unterteile. Mache ich das über die Anzahl von Wochen (6, 12 und 24 Wochen Samples z.B.) oder über die Anzahl von visitors (125k, 250k, 500k, ... visitors z.B.) oder eine Mischung, so wie du es auch schon mal vorgeschlagen hast (6 Wochen und 125k visitors, 12 Wochen und 250k visitors, 24 Wochen und 500k visiors z.B.). Im Anschluss stellt sich dann aber die eigentliche Frage wie ich den train test split mache: (1) random, sodass theoretisch frühere visits eines visitors im test set sein können und spätere visits des gleichen visitors auch im training set sein können oder (2) die ersten n1 Wochen des Samples fürs training (z.B. 4) und die folgenden n2 Wochen (z.B. 2) des Samples fürs testing - In der Literatur gibt es beide Ansätze, manche machen einfach einen random train test split auf ihren Daten, andere nehmen die ersten paar Wochen zum training und die folgenden paar Wochen zum testing. Ich habe bisher so gedacht, dass die Reihenfolge der visits für training und testing egal ist. Wichtiger ist, dass ich, wenn ich mir einen bestimmten visit anschaue, dass ich das vorherige Verhalten des visitors kenne (z.B. visits, page views, product views, purchases etc. in the past 24 hours, 7 days...), weil das indikativ für einen Kauf ist. An der Stelle müsste man aber zwischen vektor-basierten Modellen, die solche features nutzen können, und sequentiellen Modellen wie RNNs unterscheiden glaube ich. - Zum diesem Thema habe ich mir die Länge der journeys schon angeschaut und möchte mir auch noch anschauen, wie sich die visits und die purchases auf die einzelnen Wochen im gesamten Datensatz verteilen. Wenn ich nämlich die ersten paar Wochen fürs training und die folgenden paar Wochen fürs testing verwende und insgesamt einen Zeitraum von mehreren Monaten betrachte, könnten die Verteilung der visits und purchases und auch Saisonalität ein Thema werden. Einerseits könnte ich diesen Effekten entgegenwirken, wenn ich meine training und test sets aus den 6 Monaten des gesamten Datensatzes zufällig wähle (auf jeden Fall stratifiziert wegen der class imbalance), andererseits sind diese Probleme (Schwankungen von visits und purchases sowie Saisonalität) vielleicht nicht so schlimm, wenn ich mir nur kürzere Ausschnitte wie z.B. 6 Wochen anschaue (wobei kürzere Ausschnitte eben auch nur kürzere Ausschnitte und nicht den gesamten Zeitraum wiedergeben). - Wenn ein solches System irgendwann mal produktiv und in real time eingesetzt werden sollte, dann mache ich eine Vorhersage für einen aktuell passierenden visit auch nur auf Basis der bereits passierten visits und des vorherigen Verhaltens dieses visitors, sodass die Reihenfolge "chronologisch" bleibt. - Und wie ist das bei den RNNs? Die brauchen ja keine features, die das vorherige Verhalten eines Nutzers abbilden, weil sie per Definition vorherige und folgende (im nicht real time Szenario) visits eines visitors für ihre prediction berücksichtigen. In diesem Fall müsste ich die ersten n1 Wochen fürs training verwenden und die folgenden n2 Wochen fürs testing und könnte nicht visits eine visitors mischen und sie zufällig auf training und test sets verteilen. Oder?

6. Evaluation (5-6 pages)

6.1. Objectivity
6.2. Predictive Accuracy

- accuracy
- AUC
- ROC
- precision
- recall
- F-score
- learning curve (e.g. F-score for different sample sizes)
- statistical (non-parametric) test for model comparison, e.g. t-test, Wilcoxon test, Friedman test ([Demšar, 2006], also Dietterich, 1998; Alpaydin, 1999; Lavesson and Davidsson, 2007; Raschka, 2018))
- pull descriptives automatically from samples/output files, wrangle and transform to LaTeX table
- select favorite model and optimize it (e.g. parameter tuning)

6.3. Robustness - cross validation -> discussion, to time and resource intensive
6.4. Interpretability

- model type and complexity
- overview of literature and methods
- overview table

6.5. Versatility
6.6. Algorithmic Efficiency

- training and testing times
- result table

7. Discussion and Managerial Implications (2 pages)

    - Entscheider sind doch auch Blackboxes und man weiSS nicht wie und warum sie entscheiden. Sie entscheiden auch nicht statistisch objektiv, sondern subjektiv.
    - 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?

8. Conclusion (2 pages)

    - summary of introduction, research question and (business) relevance
    - summary of experiments and most important results and contribution
    - limitations, potential for future research and outlook:
        - more data
        - different data
        - different use cases/applications
        - different models
        - test effect of different targets/features
        - different sampling techniques (period/weeks, # of visitors, # of visits)
        - predict purchase probabilities rathern 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
        - Literature on statistical tests for model comparison (Dietterich 1998, Alpaydin 1999, Demsar 2006, Lavesson and Davidsson 2007, Raschka 2018)

References
Appendix
Abbreviations - Models

Abbreviations	Algorithms/Models
BAG	Bagging
BM	Bayesian methods (incl. Bayesian network)
BOOST	Boosting
CM	Custom model
DA	Discriminant analysis
DBN	Deep belief network
DT	Decision tree

Abbreviations	Algorithms/Models
ENN	Extended nearest neighbor
ENS	Ensemble
GLM	Generalized linear model
KNN	k nearest neighbor
LMNN	Large margin nearest neighbor
LR	Logistic regression
MISC	Miscellaneous (less popular and
MM	frequently used models)
NB	Markov model (incl. Markov chain)
NN	Naive Bayes
REG	Neural network
RF	Regression (incl. linear)
RL	Random forest
RNN	Rule-based learning
SDA	Recurrent neural network
STC	Stack denoised autoencoder
SVM	Stacking

Abbreviations - Evaluation metrics

Abbreviations	Evaluation metrics
# of leaves	Number of leaves
# of neurons	Number of neurons
ACC	Classification accuracy
AUC	AUC
Average PREC	Average precision
Balanced ACC	Balanced accuracy
CM	Custom metric
FNCC	Fraction of actual negative examples classified correctly
Cohen k	Cohen k
Complexity	Complexity
Coverage	Coverage
Cross-entropy	Cross-entropy
Decision time	Decision time
ER	Error rate
Friedman ranking	Friedman ranking
F-score	F-score
Hamming loss	Hamming loss
Kappa	Kappa
Lift	Lift
MER	Mean error rate
MRER	Mean rank of error rate
MSE	Mean squared error
NLL	Negative log-likelihood
One-error	One-error
PAMA	Probability of achieving maximum accuracy

Abbreviations	Evaluation metrics
PBR	Probability-based ranking
PMA	Percentage of maximum accuracy
Precision/recall BEP	Precision/recall break even point
RAM MB hours	RAM MB hours
Ranking loss	Ranking loss
REC	Recall
ROC	ROC
SE	Squared error
Testing time	Testing time
Training time	Training time
Rank	Ranking
COM	Cost of misclassification
Learning rate	Learning rate
Cost	Cost
Cost matrix	Cost matrix
Confusion matrix	Confusion matrix
Comprehensibility	Comprehensibility
Ease of use	Ease of use
MR	Mean ranking

References

[Anderl, E., Becker, I., Wangenheim, F. V., & Schumann, J. H., 2014] Anderl, E., Becker, I., Wangenheim, F. V., & Schumann, J. H. (2014). Mapping the customer journey: A graph-based framework for online attribution modeling. *SSRN*, (2343077).

[Bergstra, J., & Bengio, Y., 2012] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[Demšar, 2006] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

[Lilien, 2011] Lilien, G. L. (2011). Bridging the academic–practitioner divide in marketing decision models. *Journal of Marketing*, 75(4):196–210.

[Little, 1970] Little, J. D. C. (1970). Models and managers: The concept of a decision calculus. *Management Science*, 16(8):B–466–B–485.

[Little, 2004] Little, J. D. C. (2004). Models and managers: The concept of a decision calculus. *Management Science*, 50(12_supplement):1841–1853.

[Lodish, 2001] Lodish, L. M. (2001). Building marketing models that make money. *Interfaces*, 31(3_supplement):S45–S55.