Exploiting Time-Varying RFM Measures for Customer Churn Prediction with Deep Neural Networks

— Supplementary Information —

Gary Mena[1], Kristof Coussement[2], Koen W. De Bock[3], Arno De Caigny[2], Stefan Lessmann*[1]

[1] School of Business and Economics, Humboldt-University of Berlin, Unter den Linden 6, D-10099 Berlin, Germany

[2] IESEG School of Management, Univ. Lille, CNRS, UMR 9221 – LEM Lille Economie Management, 3 Rue de la Digue, F-59000, Lille, France

[3] Audencia Business School, 8 Route de la Jonelière, F-44312, Nantes, France

E-mail addresses: gary.mena@bdpems.de (Gary Mena), stefan.lessmann@hu-berlin.de (Stefan Lessmann), a.de-caigny@ieseg.fr (Arno De Caigny), kdebock@audencia.com (Koen W. De Bock), k.coussement@ieseg.fr (Kristof Coussement),

Corresponding author: Stefan Lessmann, School of Business and Economics, Humboldt-University of Berlin, Unter den Linden 6, D-10099 Berlin, Germany, stefan.lessmann@hu-berlin.de

Exploiting Time-Varying RFM Measures for Customer Churn Prediction with Deep Neural Networks

— Supplementary Information —

The supplementary information augments the empirical results presented in the main body of the paper. While the paper focuses on processing time-varying RFM customer variables using deep learning architectures, we acknowledge that standard supervised ML algorithms can also accommodate such features provided they are properly prepared. A range of feature transformations can be considered to aggregate temporal measures of a feature (e.g., recency) into a scalar value. For example, a vanilla churn classifier could include the k most recent previous realizations of a variable as a lagged feature. One can also compute the mean of a time-varying variable, perhaps considered time-weighing. Koehn et al. (2020) exemplify more common feature transformations.

To secure the conclusion that DNNs and especially RNNs are an effective approach to using time-varying RFM features in their native format and without losing information due to an aggregation step, we have compared deep learning architectures (RNN and Transformer) to regularized logistic regression and extreme gradient boosting models trained on various alternative aggregations of repeated RFM measures. Since there is no universal, agreed-upon standard to perform such aggregations, in these experiments, we deemed it useful to consider alternatives. To that end, we considered three basic aggregations and more advanced options, running experiments using the following 4 alternative aggregations:

- the average value of time-varying RFM measures ("Average RFM")
- the average of the deltas of time-varying RFM measures ("Average RFM Deltas")
- the most recent 6 values of time-varying RFM measures ("RFM lags")
- the fitted values for each RFM variable for each customer based on an exponential weighted moving average ("RFM EWMA")

The first three alternatives assume equal importance to all RFM measures and neglect the time dimension. In contrast, the fourth alternative, the more advanced exponential weighted moving average (EWMA) (Hunter, 1986; Holt, 2004) awards higher importance to more recent measures. For each RFM dimension (frequency, recency, and monetary value) and for every customer, the set of time-varying measures is substituted by a predicted smoothed estimate $\hat{X}_T$ through

$$\hat{X}_T = \frac{\sum_{t=1}^{T} w_j x_T}{\sum_{t=1}^{T} w_j},$$

which denotes a weighted sum where $w_j = (1-\alpha)^t$ is the weight associated with period t that depends on smoothing parameter α obtained through simple exponential smoothing (McCormick, 1969). For each RFM variable, we obtain α by assuming an additive error structure, applied to the series of the average values of the respective variable per period across all customers. Note that the EWMA estimate has seen applications for aggregating RFM measures in recent churn prediction literature (e.g. Perisic & Pahor, 2022) which motivated this particular aggregation.

The following tables summarize the results of these comparisons. Note that all experiments have been conducted using identical experimental choices and statistical testing procedures as adopted in the main paper. Table 1 summarizes the cross-validation results (averages and standard errors) in terms of AUC, top-decile lift (TDL), and maximum profit criterion (EMPC) for all classifiers and feature sets.

Tables 2 and 3 provide results of statistical comparisons between deep learning models depending on time-varying RFM measures and benchmark models trained on various aggregations of these measures. Table 2 shows comparison results for the Transformer model, while Table 3 presents results for RNN. These results demonstrate that the RNN models built on time-varying RFM features outperform benchmark algorithms in form of regularized logistic regression and extreme gradient boosting built on varying aggregations of these features. Even for the RFM aggregations based on an exponential weighted moving average, a more advanced transformation that considers a time effect, this conclusion holds for most performance metrics. The performance of the transformer model is more variable, which was to be expected since RNNs consistently outperform the transformer architecture in the main paper.

Algorithm category	Feature set	Algorithm	Feature set variation	Metric					
				AUC		TDL		EMPC	
DNN	***Time-varying RFM***	*RNN*	-	**0.7671**	(0.0211)	**4.4892**	(0.3061)	**0.3315**	(0.0327)
		Transformer	-	0.7130	(0.0154)	3.2447	(0.2232)	0.2023	(0.0284)
Benchmarks	***Aggregated time-varying RFM***	*Regularized logistic regression*	*Avg. RFM*	0.6650	(0.0241)	2.1955	(0.5923)	0.1264	(0.0462)
			Avg. RFM Deltas	0.6255	(0.0220)	3.0392	(0.3369)	0.1726	(0.0324)
			RFM Lags	0.7519	(0.0236	3.8784	(0.4178)	0.2785	(0.0353)
			RFM EWMA	0.7372	(0.0217	3.5049	(0.4932)	0.2417	(0.0413)
		Extreme gradient boosting	*Avg. RFM*	0.6972	(0.0224	2.8003	(0.3550)	0.1722	(0.0411)
			Avg. RFM Deltas	0.6666	(0.0237	3.0173	(0.1453)	0.1792	(0.0178)
			RFM Lags	0.7631	(0.0275	4.0444	(0.3489)	0.2987	(0.0403)
			RFM EWMA	0.7520	(0.0225	3.8644	(0.3135)	0.2742	(0.0422)

Table 1: Results (averages and standard errors) in terms of AUC, top-decile lift (TDL), and maximum profit criterion (EMPC) for the classifiers and feature sets considered in this comparison

Algorithm Comparison	Benchmark feature set	Metric					
		AUC		TDL		EMPC	
		t-value	p-value	t-value	p-value	t-value	p-value
Transformer vs. regularized logistic regression	*Avg. RFM*	-6.0963	0.0035***	-3.6184	0.0208**	-4.0711	0.0149**
	Avg. RFM Deltas	-4.1457	0.0136**	-0.7096	0.3294	-0.7724	0.3241
	RFM Lags	2.3687	0.0692*	2.2138	0.0798*	2.2903	0.0769*
	RFM EWMA	2.3313	0.0719*	1.0276	0.2560	1.5412	0.1644
Transformer vs. extreme gradient boosting	*Avg. RFM*	-1.4849	0.1699	-3.5928	0.0213**	-1.5454	0.1637
	Avg. RFM Deltas	-2.1588	0.0857*	-1.5281	0.1605	-0.8690	0.3011
	RFM Lags	3.8824	0.0169**	2.5777	0.0552*	2.9481	0.0401**
	RFM EWMA	2.8237	0.0440**	2.8610	0.0418**	2.4286	0.0668*

Table 2: Pairwise comparisons between the Transformer model based on repeated RFM measures versus (i) regularized logistic regression and (ii) extreme gradient boosting based on various aggregations of repeated RFM measures. Results are underlined when the absolute average performance is higher for the RNN model than for the benchmark model. Significant differences at the 90%, 95%, and 99% levels are indicated by *, **, and *, respectively.**

Algorithm Comparison	Benchmark feature set	Metric					
		AUC		TDL		EMPC	
		t-value	p-value	t-value	p-value	t-value	p-value
RNN vs. regularized logistic regression	*Avg. RFM*	-7.5186	0.0016***	-4.6094	0.0094***	-4.7526	0.0088***
	Avg. RFM Deltas	-6.2054	0.0033***	-8.6397	0.0009***	-12.8216	0.0002***
	RFM Lags	-2.3668	0.0693*	-2.1216	0.0877*	-3.3405	0.0278**
	RFM EWMA	-2.5417	0.0581*	-2.3411	0.0700*	-2.7356	0.0492**
RNN vs. extreme gradient boosting	*Avg. RFM*	-4.48980	0.0104**	-4.7203	0.0086***	-4.0825	0.0147**
	Avg. RFM Deltas	-5.3507	0.0056***	-7.0204	0.0020***	-9.4024	0.0007***
	RFM Lags	-0.2950	0.4306	-3.5818	0.0214**	-1.7338	0.1356
	RFM EWMA	-1.9260	0.1088	-2.9309	0.0390**	-2.3358	0.0733*

Table 3: Pairwise comparisons between the RNN based on repeated RFM measures versus (i) regularized logistic regression and (ii) extreme gradient boosting based on various aggregations of repeated RFM measures. Results are underlined when the absolute average performance is higher for the RNN model than for the benchmark model. Significant differences at the 90%, 95%, and 99% levels are indicated by *, **, and *, respectively.**

References

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting, 20(1), 5–10.

Hunter, J. S. (1986). The Exponentially Weighted Moving Average. Journal of Quality Technology, 18(4), 203–210.

Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, *150*, 113342.

McCormick, G. P. (1969). Communications to the Editor—Exponential Forecasting: Some New Variations. *Management Science*, 15(5), 311–320.

Perisic, A., & Pahor, M. (2022). RFM-LIR Feature Framework for Churn Prediction in the Mobile Games Market. *IEEE Transactions on Games*, 14(2), 126–137.