Predicting revenues with the Multiplier heuristic

Florian M. Artinger

Center for Adaptive Rationality and Cognition, Max-Planck-Institute for Human Development, Lentzeallee 94, 14195 Berlin, Phone: +49 30 8240 6697, Email: artinger@mpib-berlin.mpg.de

Nikita Kozodoi

School of Business and Economics – Humboldt University Berlin, Spandauer Str. 1, 10178 Berlin, Telephone: +49 30 2093-99543, Email: nikita.kozodoi@hu-berlin.de

Julian Runge

School of Business and Economics – Humboldt University Berlin, Spandauer Str. 1, 10178 Berlin, Telephone: +49 30 20 93-99019, Email: julian.runge@googlemail.com

Abstract. Using machine learning to forecast revenues per customer, product, or store has become a major industry. It contrasts with a simple, intuitive managerial heuristic: multiply the revenue observed in the first *t* days by a constant. We test the predictive accuracy of this multiplier heuristic in 20 data sets. Surprisingly, the heuristic performs overall on par with machine learning models. We identify three central drivers when the heuristic can even outperform these models: limited sample size, time-to-prediction, and changes across time. The results provide insights when to rely on heuristics and managerial intuition or when on machine learning.

1. Introduction

At a rising tech company that sells mobile apps, a senior manager challenges a team of data scientist: can you beat a simple rule for predicting the annual revenue for each of the more than 500.000 customers? To target high value customers early on, such prediction takes place after customers have used an app for only seven days. The rule, or heuristic, the manager puts forth is to multiply the revenues an individual customer generates in the first seven days by a factor of six. At first sight, to the data scientists it seemed rather a trivial task to beat the manager's heuristic. Given the large amount of data available, companies have been increasingly shifting towards using machine learning tools for predicting revenues not only per customer but also per product or per store (Columbus 2018). However, it turns out that the manager at the rising tech company has intuitively developed a simple but powerful rule, which we refer to as the multiplier heuristic. Going beyond the initial challenge of the manger, we address under which conditions to rely on such simple and intuitive decision rules and when on more complex tools such as from machine learning.

The argument in favor of using statistical models[1] for prediction including machine learning, rather than human judgement, has been long made. Already in 1954, Meehl shows that models such as a linear regression systematically outperform human judgment including experts (Dawes et al. 1989). Moreover, research on heuristics and biases suggests that because of people's rather limited capacities for computation and information processing, decisions are error prone and far from optimal (Tversky and Kahneman 1974). Conducting a meta-analysis Grove et al. (2000) and find that that in 47% of 136 studies that the statistical model makes better predictions than human judgement, in only 6% human judgement prevails (in the remainder the two methods performed about equally). Given the increasing availability of data and advancements in machine learning, many businesses and researchers take it for granted that human judgement can ever only be second-best (e.g., Chen et al. 2012, Erevelles et al. 2016, Gandomi and Haider 2015, McAfee and Brynjolfsson 2012)."

At the same time, there are cases where heuristics, based on expert intuition, perform on par or even outperform complex statistical models (Katsikopoulos et al. 2008, for a review see Artinger et al. 2015). This includes identifying in which start-up to invest (Åstebro and Elhedhli 2006), predicting whether a customer will continue shopping at a store (Wübben and von Wangenheim 2008), diversifying among assets (DeMiguel et al. 2009), pricing used cars (Artinger and Gigerenzer 2016), predicting recidivism (Dressel and Farid 2018, Jung et al.

[1] In this paper, we consider machine learning models as a subcategory of statistical models, see for instance Hastie et al. (2001).

2017), or personnel selection (Luan et al. 2019, Luan and Reb 2017). Moreover, people are surprisingly apt in making early stage predictions, also referred to as decisions from 'thin slices' of data, a situation resembling that of the manager at the tech company (Ambady et al. 2000, Ambady and Rosenthal 1992). Such thin slice of information can take many forms including the data of the tech company that the senior manager was very familiar with. Judgements based on thin slices usually occur rapidly, unwittingly, and intuitively (Ambady et al. 2000). The challenge with intuitive judgements is to not only verbalize but also formalize them: the multiplier heuristic is a potential formalization of intuitive judgement based on thin slices.

Given such diverging findings on the effectiveness of heuristics, reporting either poor performance or surprisingly good performance, what is necessary is an understanding of the conditions under which either perspective holds. That is, when can and should one rely on human judgement and when on more complex statistical models such as machine learning? The contributions of this paper are two-fold. First it introduces the multiplier heuristic to the scientific literature and tests its performance using 20 different data sets where in each the goal is to predict revenues at the end of a year. The multiplier heuristic is the first heuristic that makes point-predictions for a continuous measure such as future revenue. The data sets cover three domains: customers, including those of the manager at the tech company, products, and shops. Across the three domains, we find that the multiplier heuristic performs on average on par in comparison to commonly used machine learning models. Second, the bias-variance trad-off has been proposed to be fundamental to the effectiveness of heuristics (Gigerenzer and Brighton 2009). The paper identifies conditions derived from the bias-variance trade-off under which to rely on human judgment and when on machine learning. Using simulations with the data sets show that sample size, time-to-prediction, and changes across time are crucial components to determine whether to rely on the heuristic or a more complex statistical model.

Next, we introduce the bias-variance dilemma and develop the hypotheses. The method section details the models and the data sets used. The result section commences by analyzing the aggregate performance of the models across data sets and then proceeds to systematically evaluate which factors favor the heuristic respectively the more complex statistical models. The discussion elaborates on factors that help or hinder the development of intuitive expert heuristics and how to combine this with machine learning.

2. The bias-variance trade-off

Seeking to better understand when heuristics that people use can be powerful tools, Gigerenzer and Brighton (2009) point to the bias-variance trade-off which has been popularized

in machine learning by Geman et al. (1992). The bias-variance trade-off is part of the broader study of ecological rationality (Gigerenzer et al. 1999). Here, the goal is to assess how well particularly human strategies perform in comparison to other strategies such as those derived via statistical procedures and to specify the conditions under which a given strategy is preferable (e.g., Baucells et al. 2008, Hogarth and Karelaia 2005, 2007, Katsikopoulos et al. 2010, Keller and Katsikopoulos 2016, Şimşek 2013, Simsek and Buckmann 2016).

The very nature of prediction requires that models generalize well to unseen data such as future revenues. That is, the task is to construct a function *f(x)* based on a training set $D = (x_1, y_1), \ldots, (x_N, y_N)$ to approximate *y* on some future observations of *x*. To be explicit about the dependence of *f* on the data, we write *f(x; D)*. A common measure of the effectiveness of *f(x; D)* as a predictor of *y* is the mean squared error

$$E_D[(f(x;D) - E[y|x])^2]$$

where E_D represents the expectation with respect to the training set, *D*, that is, the mean of the ensemble of possible *D* with *E[y|x]* as the generating function. A particular training set, *D*, *f(x;D)* may yield a good approximation of *E[y|x]* but it may also be the case that *D* yields a *f(x;D)* that is quite different and in general varies substantially with *D*. To understand what drives prediction error, it can be decomposed into bias, variance and an irreducible noise ϵ:

$$E_D[(f(x;D) - E[y|x])^2] = (E_D[f(x;D)] - E[y|x])^2 + E_D[(f(x;D) - E_D[f(x;D)])^2] + \varepsilon$$

If *f(x;D)* is on average different from *E[y|x]* then *f(x;D)* is biased which is captured by the first term. An unbiased estimator may still have a large mean-squared error due to the variance component, captured by the second term. That is, even if $E_D[f(x;D)] = E[y|x]$, the estimated function *f(x;D)* may be highly sensitive to the data, *D*, and far from the generating function *E[y|x]*. Given a complex problem even with a large sample, error due to variance can be substantial. Geman et al 1992 assert that "learning complex tasks is essentially impossible without the a priori introduction of carefully designed biases into the machine's architecture. (…) Despite a long pre-occupation with learning per se, the identification and exploitation of the "right" biases are the more fundamental and difficult research issues." (Geman et al. 1992, p. 3).

A number of methods have been developed to balance bias and variance. These seek to reduce the exposure of error due to variance and hence overfitting (see for a discussion Geman et al. 1992). Good results on the learning data do not necessarily imply good generalization performance on unseen data. This principle is used in *k*-fold cross-validation (commonly attributed to Stone 1974). Observations are separated into training and validation data set: *f(x;D)* is estimated on the training set, the validation set is used to compute the prediction error. The methods also include regularization which rely on a penalty function that is added to the mean squared error. The penalty function in LASSO regression (L1 regularization) modifies an ordinary linear regression by selecting only a subset of the provided variables for use in the final model rather than using all of them (Tibshirani 1996). Note that the multiplier heuristic is equivalent to an ordinary linear regression without intercept. A LASSO regression can in principle equate the multiplier heuristic. Another regularization technique is ridge regression (L2 regularization) where the penalty function shrinks the weights towards zero but maintains all dependent variables (Hoerl and Kennard 1970). All three methods are standard in machine learning.

A low prediction error also depends on the bias component. The literature in psychology and behavioral economics has identified many biases that are due to the use of heuristics (Tversky and Kahneman 1974). A prominent explanation why people use heuristics is the cost-benefit trade-off: heuristics are less costly in terms of computational effort but also less accurate (Payne et al. 1988, Sims 2003). Such conclusions are reached by comparing the heuristic to a rational choice benchmark, commonly the perspective of an omniscient observer for a stylized problem. Yet, in many real-world contexts such a benchmark is not available and one needs to construct a model, or function *f(x;D)*, from some training data. A competitive test which includes the evaluation of the bias-variance trade-off provides insights into the advantages or disadvantages that a bias provides. Note that if one does not estimate the multiplier from the data but uses the manager's judgement, the model's prediction error can only be due to bias but not variance. The study of human judgment provides a potential source for identifying the "right" biases that Geman et al. (1992) call for. The effect of the bias-variance trade-off on the performance of heuristics has so far been investigated on only one empirical data set for binary choice (Analytis et al. 2018). Heuristics for such a choice situation such as take-the-best (Gigerenzer and Goldstein 1996) tend to have a lower share of error due to variance in comparison to more complex models.

We propose three conditions based on the bias-variance trade-off which allow evaluating whether the multiplier heuristic makes use of bias in a way that is advantageous for

predictive performance. Geman et al. (1992) already highlight the role of sample size: given only limited data, a complex statistical model is likely exposed to significant error due to variance.

> *Proposition 1:* As data becomes scarcer, we expect the relative performance of the multiplier heuristic to improve as the heuristic incurs less error from variance, eventually outperforming more complex models.

Given a time series, the predictive strength of the variables changes with time-to-prediction, that is, as the time to the predicted event diminishes. Some variables might become strong predictors whereas adding further variables likely exposes the model to error from variance. For instance, consider predicting total annual sales at *t-1*, that is one day before the end of the year. Relying solely on all sales up to that *t-1* is likely a strong predictor and other variables can and should be ignored.

> *Proposition 2:* As time-to-prediction diminishes, we expect the relative performance of the multiplier heuristic to improve as the heuristic incurs less error from variance, eventually outperforming more complex models.

Changes in the data between training and prediction set, such as due to non-stationarity, will expose the models to error from variance. We compare two settings: in one setting we use conventional *k*-fold cross-validation. In the other, we evaluate the performance of the models across time where we estimate parameters from some initial period and make predictions for the remaining period. Note that *k*-fold cross-validation ignores the longitudinal dimension of the data disregarding the effect that changes across time can have on predictive accuracy. Such changes across time introduce error due to variance for all models.

> *Proposition 3:* We expect the relative performance of the multiplier heuristic to improve and outperform more complex models when evaluated across time as opposed to cross-validation as the heuristic incurs less error from variance.

3. Data and Methods

Table 1 contains a summary of the 20 data sets used. All data sets contain information on sales for each individual entity, either per customer (10 data sets), per product (5 data sets), or per store (5 data sets). Table 1 also lists the number of observations per data set and the number of variables (see Appendix for details on variable pre-processing). The variables contain individual entity characteristics (such as gender and age for customers) and purchase characteristics (such as time of purchase, number of items in basket) that can be used to predict the revenue that each entity generates. As can be seen, the three domains vary considerably in

terms of the number of observations but also in terms of the number of variables suggesting that there are different environments in which the models need to predict future revenues. The data sets cust_f1 to cust_f5 are those that the manager from the introduction faced. All stores in a data sets belong to a single retailer. For instance, the data set stor_wa contained the information of 6,417 individual Walmart stores including 12 variables for each store.

Each data set has at least two variables: (1) Total sales for each entity during the first t days of observation. The observation period starts with the first purchase of the entity recorded in the data. This variable serves as one of the predictors. (2) Total sales for the entity during the following $360 - t$ days. This serves as the dependent variable. Each data set contains at least two consecutive years of sales data. For all the analysis we partition the data into two subsets: the first year is the training set used to train the models and obtain parameter estimates, the second year is the prediction set used for performance evaluation. In this fashion, we directly emulate the situation of the firms that need to predict future revenue based on historical observations. We initially evaluate the models using an observation period $t = 7$ days and predict total cumulative revenue of each individual entity for the last $360 - t$ days as in the problem the manager in the introduction faced. We subsequently vary the observation period in order to assess the effect of time-to-prediction.

Table 1: Overview of the data sets

Data set	Description	Entity	Obser-vations	Variables
cust_sc	sales in a retail store	customers	220	23
cust_sa	sales in a retail store	customers	13	8
cust_fa	sales in a fashion store	customers	3,352	18
cust_ap	sales in an apparel store	customers	10,258	15
cust_go	online sales in Google merchandise store	customers	2,339	125
cust_f1	in-app purchases #1	customers	42,183	7
cust_f2	in-app purchases #2	customers	59,934	7
cust_f3	in-app purchases #3	customers	74,117	7
cust_f4	in-app purchases #4	customers	114,731	2
cust_f5	in-app purchases #5	customers	215,653	7
prod_cb	sales of carbonated beverages	products	16	8
prod_gr	sales of grocery products	products	5,718	2
prod_sp	sales in a retail store	products	295	27

prod_tv	sales of TV models	products	168	2
prod_hs	sales of household products	products	12,572	6
stor_br	sales in stores with food-related products	stores	154	8
stor_oc	sales in computer software stores	stores	105	4
stor_wa	sales in Walmart stores	stores	6,417	12
stor_rs	sales in Rossman stores	stores	223	8
stor_cl	sales in a chain of grocery stores	stores	756	2

We use six different models. The multiplier heuristic (M) predicts that the total sales of a given entity will be equal to the revenue in the first *t* days multiplied by a constant *m*, the multiplier. We compare the heuristic with the following five statistical models: linear regression (LR), LASSO regression (L1), ridge regression (L2), random forest (RF) and extreme gradient boosting (XG). Linear regression serves as a simple benchmark, LASSO and ridge regression rely on regularization and should be less prone to overfitting. They are all linear models, just as the multiplier heuristic. Using random forest and extreme gradient boosting we also include non-linear models. They are popular in machine learning due to their predictive power and ability to deal with overfitting.

The five statistical models all combine the observation of revenues from the first *t* days with other variables that the data sets provide. To ensure a fair comparison, we tune meta-parameters of the statistical models using grid search with 4-fold cross-validation (CV) on the training data. We select the combination of parameter values that minimizes CV-based mean RMSE. After parameter tuning, we train the statistical models on the full training set. For the multiplier heuristic we test not only the manager's intuition with a multiplier of $m = 6$ but we also identify a suitable multiplier value using grid search over the range of possible values between 1 and 100. We select the multiplier that minimizes RMSE on the training data and use it to predict total sales in the prediction subset (parameter grid for all models is reported in the Appendix).

4. Results

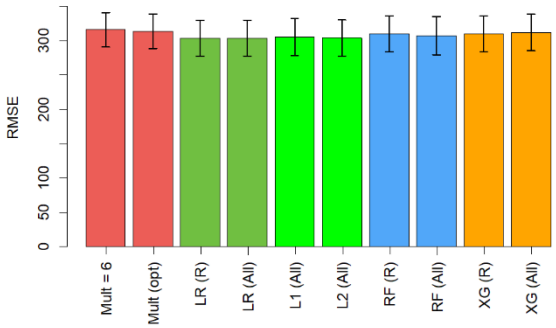
The analysis initially provides an overview on the performance of the individual models before analyzing the three propositions: the effect of sample size, time-to-prediction, and changes across time.

4.1 Model performance

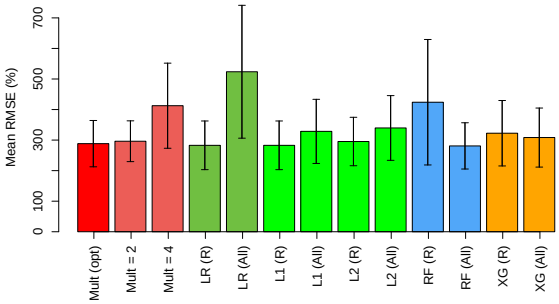
Figure 1 presents the performance of the models. Figure 1A in focuses on the situation of the manager in the introduction who challenged the company's data science team. Across the 5 data sets presented in Figure 1A, we see that a fixed multiplier of $m = 6$ that the manager suggested performs on par with the more complex statistical models. This holds for the linear models LR, L1, and L2, but also the non-linear models, RF and XG. That is, the company would not profit from using more sophisticated models. Estimating an optimal multiplier form the data as opposed to relying on the manager's intuition of a multiplier $m = 6$ yields very similar results. Going beyond the 5 app data sets, Figure 1B shows the aggregate results across all 20 data sets. Again, the multiplier heuristic performs well—none of the more sophisticated models can outperform it.

Figure 1. Model performance for RMSE.

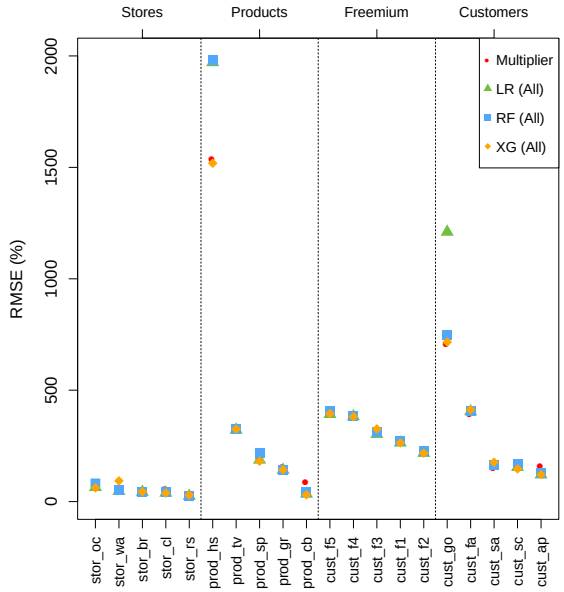
1A. Aggregate RMSE across 5 app data sets



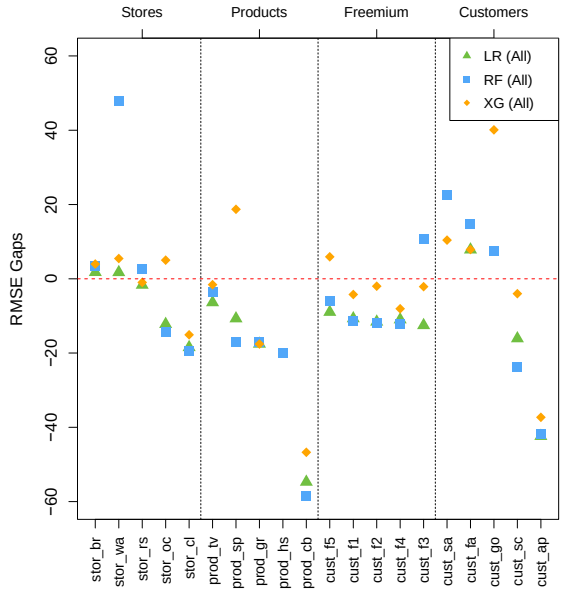
1B. Aggregate RMSE across all 20 data sets



1C. RMSE for indidivual data sets



1D. RMSE gap for indidivual data sets



Note: Figure 1A and 1B bars in indicate mean RMSE values, arrows refer to +/- one standard deviation. In Figure 1D, the gaps are computed as RMSE of a model minus RMSE of the multiplier heuristic. The gap of LR (All) on data sets 1 is 396 and is not included in the diagram to allow a smaller scale.

Considering the individual data sets in Figure 1C, one sees that there are substantial differences between the categories in terms of the RMSE. Whereas for stores one observes a relatively low RMSE, products and customers are more widely dispersed. Nevertheless, comparing the model's performance for a given data set one sees that only for two out of 20 data sets are there substantial differences in terms of performance (for the product data set of prod_hs and for the customer data set cust_go). In both cases does the multiplier heuristic perform rather well. Figure 1 D provides a relative comparison between the performance by showing the error of respectively LR, RF, and XG minus the error of the heuristic. Positive values indicate that the heuristic performs better than its competitors. The multiplier heuristic performs best on 5 data sets out of 20. At the same time, the ranking of the algorithms changes substantially from one data set to another. This result emphasizes the need to investigate conditions in which the heuristic is able to perform well.

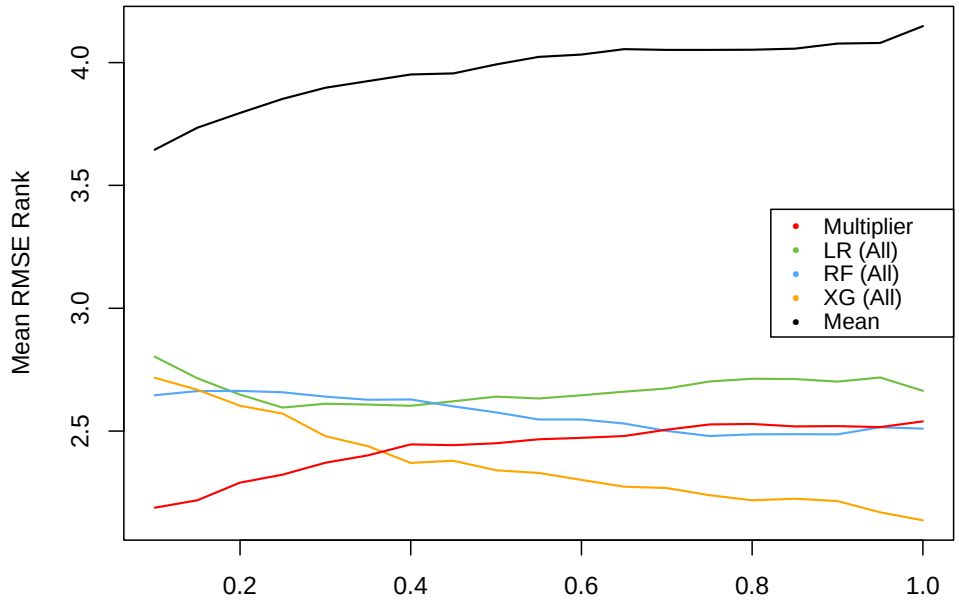
4.2 Conditions that shape performance

We first manipulate the sample size of the training data to investigate the relationship between the relative predictive performance of the multiplier heuristic and the number of observations in the historical training data. We only use data sets with at least 1,000 observations (N = 12) to ensure robust results. For each data set, we iteratively select random subsamples of the training data ranging from 10% to 100% of the original sample size, whereas the validation data remains intact. This is done by removing a random cohort of entities (1%) from the training data on each iteration. The statistical models and the multiplier heuristic are trained on a given subsample of the training data, following the modeling framework described in Section 4.1. Next, the models are used to predict the sales in the validation data. Predictive performance based on different percentages of the training data is aggregated across the data sets. The procedure is repeated 100 times to ensure that the order in which entities are dropped from the training sample is not affecting the results. The aggregated performance is measured using RMSE ranks. This indicates the relative rank of a model in terms of RMSE. The ranks are averaged across 12 data sets and 100 trials.

Figure 2 shows the average model ranks and their confidence intervals. Results suggest that the multiplier heuristic performs best when sample size is small (up to 35% of the original

size of the data set). The heuristic is outperformed by the statistical models on larger samples. Both XG and RF benefit from higher volume of training data. On larger samples, XG is the best model in terms of mean RMSE rank. This is in line with proposition 1: as data becomes scarcer, the relative performance of the multiplier heuristic improves as the heuristic incurs less error from variance and eventually outperforms more complex statistical models. It should be noted that all models benefit from additional information (for the results on the absolute RMSE for each individual data set see in Appendix Figure A1).

Figure 2. Effect of sample size

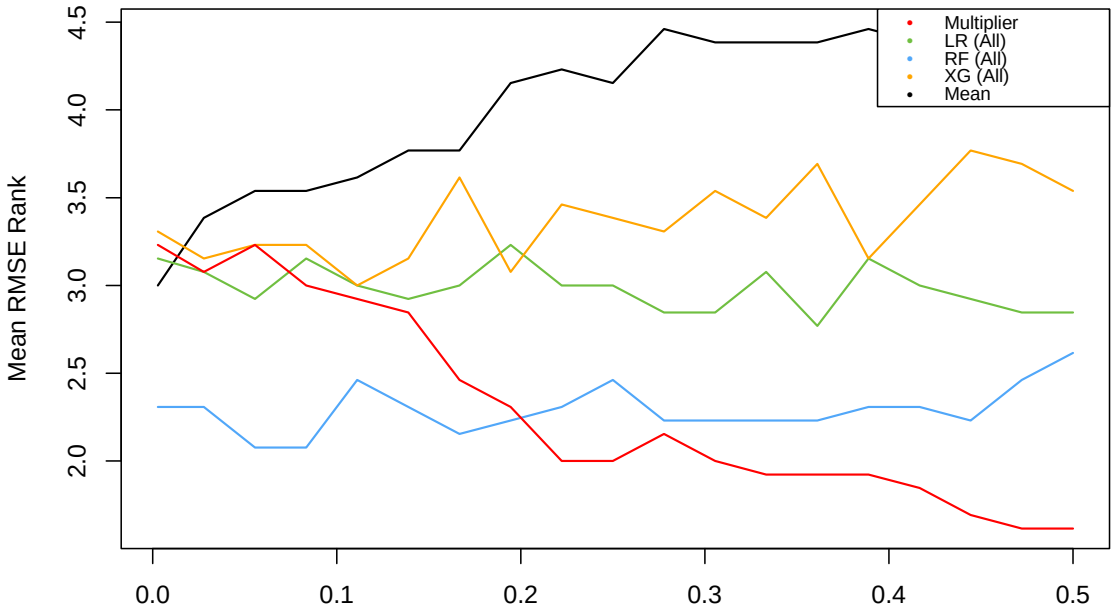


Note: "Mean" refers to a naive benchmark by predicting total sales as mean total sales in the training data.

Next we consider the effect of time-to-prediction. The analysis is limited to data sets that contain purchase behavior of an entity for at least 6 months (N = 13). This condition does not hold for all data sets since some of them only have detailed data on the purchases during the first week and the total sales after one year. The models train on the purchase behavior in the first t days, $t = 1, \ldots, 180$, and predict the amount of sales between day 180 and 360. This allows to keep the same target variable for models with different observation periods. The meta-parameters of the models are tuned separately for each of the considered observation periods to adjust the statistical models and the optimal multiplier to the changing prediction task. The tuning is performed using grid search presented in Table A1 in the Appendix.

Figure 3 reports the aggregated results in terms of the mean RMSE ranks as time-to-prediction becomes smaller. Observation period is measured as the percentage of the original time period (360 days). The mean rank of the heuristic tends to decrease for smaller time-to-prediction window. Statistical models that are based on multiple variables start to overfit, while the heuristic becomes more accurate because of its simplicity and focus on the right information. This is in line with proposition 2: as time-to-prediction diminishes, the relative performance of the multiplier heuristic improves as the heuristic incurs less error from variance and eventually outperforms more complex statistical models. It should be noted that all models benefit from having more information in terms of a longer observation period. Yet, the statistical models stop improving at some point whereas the heuristic continues to improve as time-to-prediction becomes smaller (for the results on the absolute RMSE for each individual data sets see in Appendix Figure A2).

Figure 3. Effect of time-to-prediction.



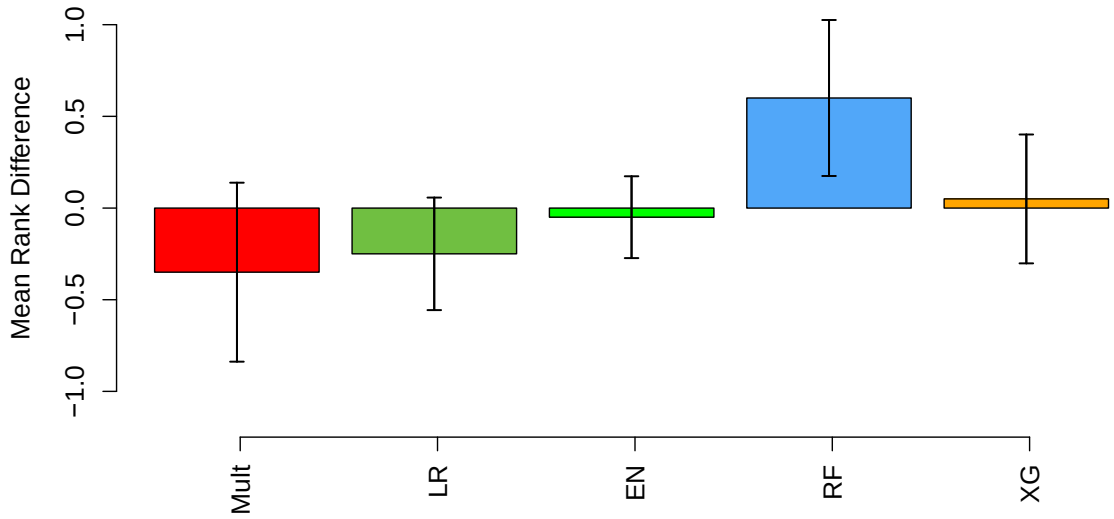
Note: "Mean" refers to a naive benchmark by predicting total sales as mean total sales in the training data.

Last, we consider the effect of changes across time. To do so, we computed mean RMSE of the models for *k*-fold cross-validation and compare it to the temporal validation approach used in the analyses above. We then subtract the rank of a model in temporal validation minus the rank in of the model in cross validation. For instance, if the heuristic is the best performing

model in a given data set in temporal validation but only the 5th best model in cross validation, the heuristic gets a value of -4 (= 1-5) for this data set.

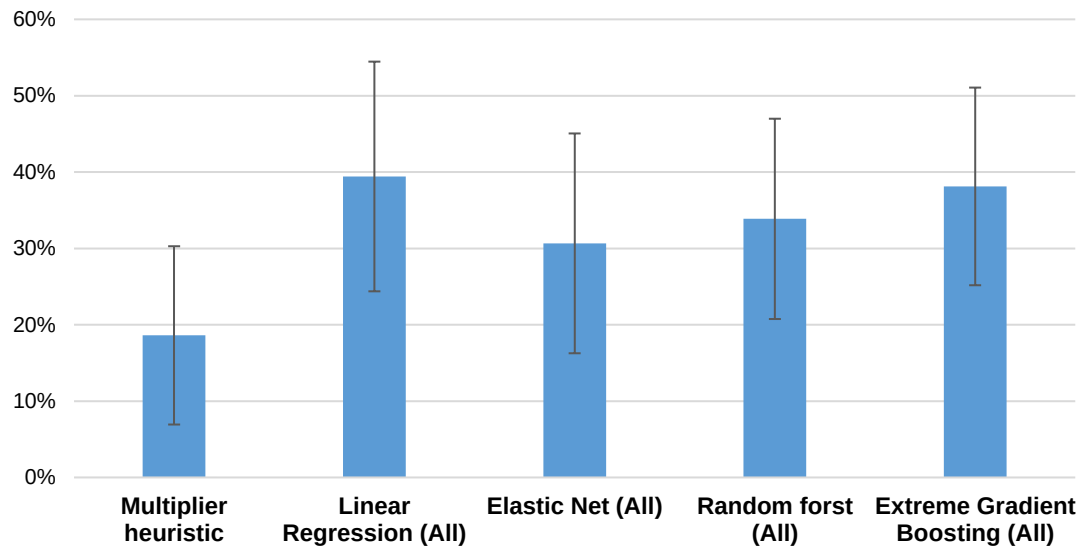
Figure 4 shows the mean rank difference between temporal and cross-validation. The multiplier heuristic but also LR perform relatively better and achieve a higher rank than the non-linear models RF and XG. The result for the heuristic is in line with proposition 3: the relative performance of the multiplier heuristic is better in comparison to the more complex models when evaluated across time as opposed to cross-validation as the heuristic incurs less error from variance.

Figure 4. Effect of changes across time: RMSE Rank (temporal) – RMSE Rank (cross-valiation)



The data so far provides evidence in terms of the performance changes for sample size, time-to-prediction, and changes across time. The explanation that has been suggested is that the heuristic has less exposure of error due to variance in comparison to the more complex models. Figure 4 displays the mean error due to variance of a given model across the 20 data sets. Note that we estimate the optimal multiplier which introduces also for the heuristic error due to variance. Using a multiplier of $m = 6$ of the manager implies no estimation and no error due to variance. The results are in line with the proposition that the multiplier heuristic has less exposure of error due to variance with 19% on average. All of the more complex models have a share of error due to variance that is above 30%.

Figure 4. Share of error due to variance.



Note: Displays the mean value of error due to variance for a model and one standard deviation.

Conclusion

The results clearly show that the intuition of the manager regarding the multiplier heuristic was powerful: the simple model performed on par with the more complex machine learning models that use significantly more data. The multiplier heuristic does not only perform well in the app data sets of the manager at the teach company but also in other data sets, including further data sets on customers, products, or stores. We present the results when the heuristic should be used: given a smaller sample size, shorter time-to-prediction, or changes over time, this can generate environments where the reliance on a single, powerful variable can be sufficient. Adding further variables or using more complex statistical models can weaken performance as it invites error due to variance. Guarding against such error via statistical tools such as regularization or cross-validation does by no means guarantee high performance of more complex statistical models. Rather, it is necessary to carefully assess whether best to rely on human judgement or machine learning. Note that this analysis has not yet accounted for the additional costs usually involved in setting up and running more complex statistical models.

This reaffirms the role of intuition in managerial decision making even in the context of a digital environment. Good intuitive judgment is frequently observed under conditions of uncertainty, that is given a limited sample size or a dynamic environment. These environmental properties also emerge from the analysis on when the multiplier heuristic performs well. In addition, there are two further requirements for good intuitive judgement (Hensman and Sadler-Smith 2011): the person must have sufficient experience in a domain that allows effective

learning with frequent and symmetric feedback linking action to outcome (Hogarth et al. 2015). The senior manager at the tech company was intimately familiar with customer's behavior and the revenue they generated providing a good source for learning. The third requirement is the organizational environment. In many organizations, decision maker's need to justify their action and list explicit reasons why they carried out certain actions. Yet, this frequently leads to the choice of the second-best option where decision makers cannot test their intuitions but rather rely on more formal, well-documented strategies (Artinger et al. 2019).

The analysis also points to future research in how to best combine human judgement and machine learning. Expert judgement can be a powerful source of identifying the right "biases" as Geman et al (1992) call for. This paper here presents a map of the environment and when to prefer a given model. A next step would be to actually combine human judgement and machine learning in one decision as Blattberg and Hoch (1990) for instance have shown. Such truly joint decision making has the power to generate decisions that are considerably better than when using either machine or human judgment alone.

Bibliography

Ambady N, Bernieri FJ, Richeson JA (2000) Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Adv. Exp. Soc. Psychol.* 32:201–271.

Ambady N, Rosenthal R (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychol. Bull.* 111(2):256–274.

Analytis PP, Barkoczi D, Herzog SM (2018) Social learning strategies for matters of taste. *Nat. Hum. Behav.* 2(6):415–424.

Artinger FM, Artinger S, Gigerenzer G (2019) C. Y. A.: Frequency and Causes of Defensive Decisions in Public Administration. *Bus. Res.* 12(1):9–25.

Artinger FM, Gigerenzer G (2016) The Cheap Twin: From the ecological rationality of heuristic pricing to the aggregate market. *Acad. Manag. Proc.* 1:1–6.

Artinger FM, Petersen M, Gigerenzer G, Weibler J (2015) Heuristics as adaptive decision strategies in management. *J. Organ. Behav.* 36(S1):S33–S52.

Åstebro T, Elhedhli S (2006) The Effectiveness of Simple Decision Heuristics: Forecasting Commercial Success for Early-Stage Ventures. *Manage. Sci.* 52(3):395–409.

Baucells M, Carrasco JA, Hogarth RM (2008) Cumulative Dominance and Heuristic Performance in Binary Multiattribute Choice. *Oper. Res.* 56(5):1289–1304.

Blattberg RC, Hoch SJ (1990) Database Models and Managerial Intuition: 50% Model + 50% Manager. *Manage. Sci.* 36(8):887–899.

Chen H, Chiang RHL, Storey VC (2012) Business Intelligence and Analytics: From Big Data To Big Impact. *Mis Q.* 36(4):1165–1188.

Columbus L (2018) 10 Ways Machine Learning Is Revolutionizing Sales. *Forbes* (December 26).

Dawes RM, Faust D, Meehl PE (1989) Clinical vs actuarial assessments. *Science (80-.).* 243:1668–1674.

DeMiguel V, Garlappi L, Uppal R (2009) Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy? *Rev. Financ. Stud.* 22(5):1915–1953.

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* 4(1):eaao5580.

Erevelles S, Fukawa N, Swayne L (2016) Big Data consumer analytics and the transformation of marketing. *J. Bus. Res.* 69(2):897–904.

Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manage.* 35(2):137–144.

Geman S, Bienenstock E, Doursat R (1992) Neural Networks and the bias-variance dilemma.

Neural Comput. 4(1):1–58.

Gigerenzer G, Brighton H (2009) Homo Heuristicus: Why Biased Minds Make Better Inferences. *Top. Cogn. Sci.* 1(1):107–143.

Gigerenzer G, Goldstein DG (1996) Reasoning the fast and frugal way: models of bounded rationality. *Psychol. Rev.* 103(4):650–69.

Gigerenzer G, Todd PM, the ABC Research Group (1999) *Simple Heuristics That Make Us Smart* (Oxford University Press, New York, NY).

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical versus mechanical prediction: A meta-analysis. *Psychol. Assess.* 12(1):19–30.

Hastie T, Tibshirani R, Friedman JH (2001) *The elements of statistical learning : data mining, inference, and prediction* (Springer, New York).

Hensman A, Sadler-Smith E (2011) Intuitive decision making in banking and finance. *Eur. Manag. J.* 29(1):51–66.

Hoerl AE, Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12(1):55–67.

Hogarth RM, Karelaia N (2005) Simple Models for Multiattribute Choice with Many Alternatives: When It Does and Does Not Pay to Face Trade-offs with Binary Attributes. *Manage. Sci.* 51(12):1860–1872.

Hogarth RM, Karelaia N (2007) Heuristic and linear models of judgment: Matching rules and environments. *Psychol. Rev.* 114(3):733–758.

Hogarth RM, Lejarraga T, Soyer E (2015) The Two Settings of Kind and Wicked Learning Environments. *Curr. Dir. Psychol. Sci.* 24(5):379–385.

Jung J, Concannon C, Shroff R, Goel S, Goldstein DG (2017) Creating simple rules for complex decisions. *Work. Pap.*:1–4.

Katsikopoulos K V., Pachur T, Machery E, Wallin a. (2008) From Meehl to Fast and Frugal Heuristics (and Back): New Insights into How to Bridge the Clinical--Actuarial Divide. *Theory Psychol.* 18(4):443–464.

Katsikopoulos K V., Schooler LJ, Hertwig R (2010) The robust beauty of ordinary information. *Psychol. Rev.* 117(4):1259–1266.

Keller N, Katsikopoulos K V. (2016) On the role of psychological heuristics in operational research; and a demonstration in military stability operations. *Eur. J. Oper. Res.* 249(3):1063–1073.

Kunc M, Malpass J, White L eds. (2016) *Behavioral Operational Research* (Palgrave Macmillan UK, London).

Luan S, Reb J (2017) Fast-and-frugal trees as noncompensatory models of performance-based personnel decisions. *Organ. Behav. Hum. Decis. Process.* 141:29–42.

Luan S, Reb J, Gigerenzer G (2019) Ecological Rationality: Fast-and-Frugal Heuristics for Managerial Decision Making under Uncertainty. *Acad. Manag. J.*:amj.2018.0172.

McAfee A, Brynjolfsson E (2012) Big Data. The management revolution. *Harvard Buiness Rev.* 90(October 2012):61–68.

Meehl PE (1954) *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* (University of Minnesota Press, Minneapolis).

Payne JW, Bettman JR, Johnson EJ (1988) Adaptive strategy selection in decision making. *J. Exp. Psychol. Learn. Mem. Cogn.* 14(3):534–552.

Sims CA (2003) Implications of rational inattention. *J. Monet. Econ.* 50(3):665–690.

Şimşek Ö (2013) Linear Decision Rule as Aspiration for Simple Decision Heuristics. *Adv. Neural Inf. Process. Syst.* 26(1):2904–2912.

Simsek Ö, Buckmann M (2016) On learning decision heuristics. *Proc. Mach. Learn. Res.* 58:75–85.

Stone M (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Ser. B* 36(2):111–133.

Tibshirani R (1996) Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* 58(1):267–288.

Tversky A, Kahneman D (1974) Judgment under Uncertainty: Heuristics and Biases. *Science (80-.).* 185(4157):1124–31.

Wübben M, von Wangenheim F (2008) Instant Customer Base Analysis: Managerial Heuristics Often “Get It Right.” *J. Mark.* 72(3):82–93.

Appendix

Data and Methods Preparation

Apart from sales, the data sets also contain additional variables. For each data set, we preprocess these variables as follows. All categorial variables with more than 5 unique values are binned such that only 5 most frequent categories are kept in the data. We use dummy encoding to create up to 5 binary indicators for each categorical variable. For each continuous variable x, we also create a new continuous variable x2 to help statistical models account for nonlinear relationships. All missing values are imputed with means (continuous variables) and modes (categorical variables).

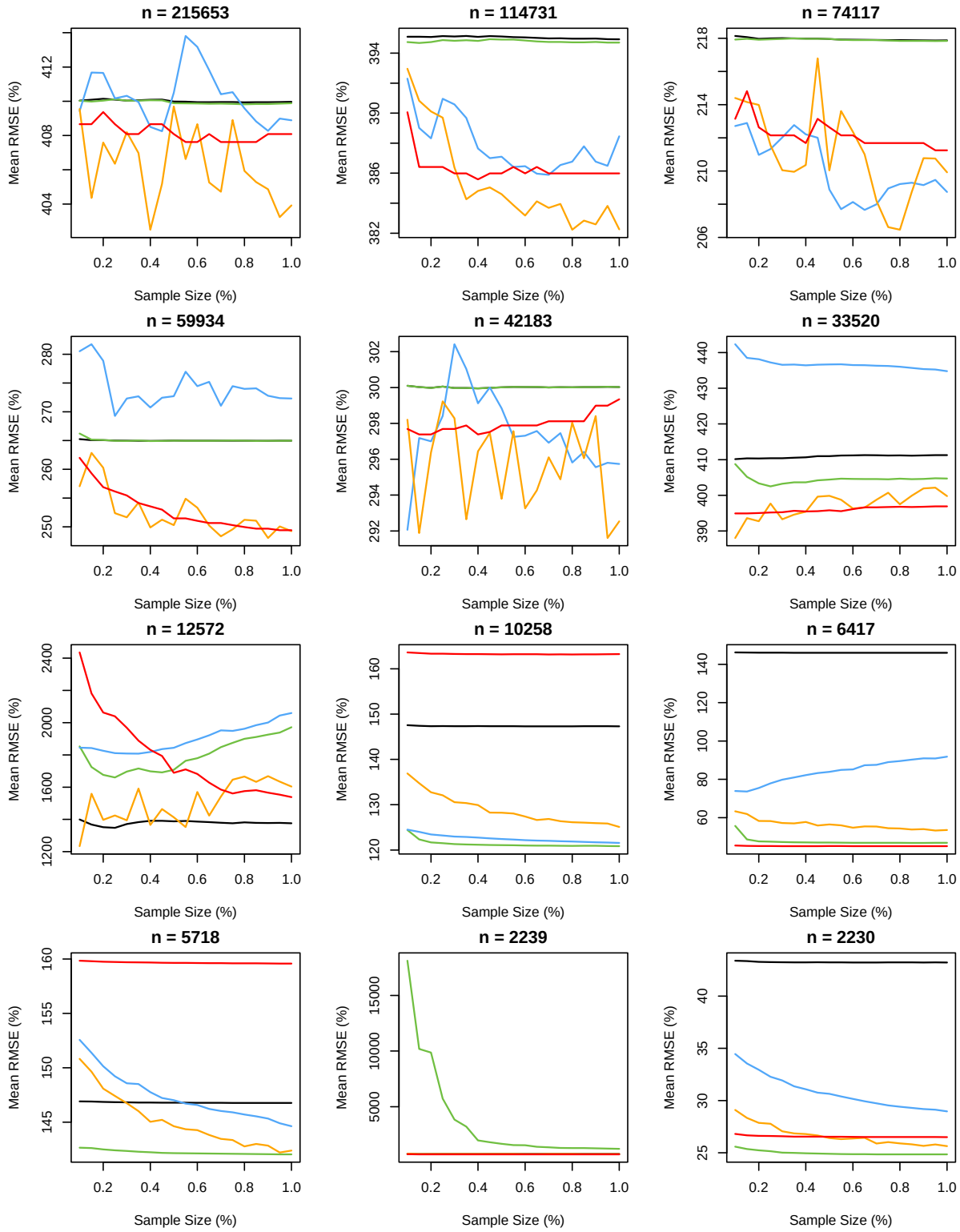
We tune meta-parameters of the statistical models using the parameter grid reported in Table A1 below.

Table A1. Parameter Grid

Algorithm	Parameter	Candidate values
Linear Regression	-	-
LR with L1 penalty	lambda	10^{-5}, 10^{-4}, …, 10^{3}
LR with L2 penalty	lambda	10^{-5}, 10^{-4}, …, 10^{3}
Random Forest	ntree	100, 300, 500
	mtry	$\sqrt{k}, 0.5\sqrt{k}, 2\sqrt{k}$, k is a number of variables
Gradient Boosting	nrounds	100, 200, 500, 1000
	eta	0.1, 0.3
	max. depth	3, 6
	lambda	0, 1
	alpha	0, 1

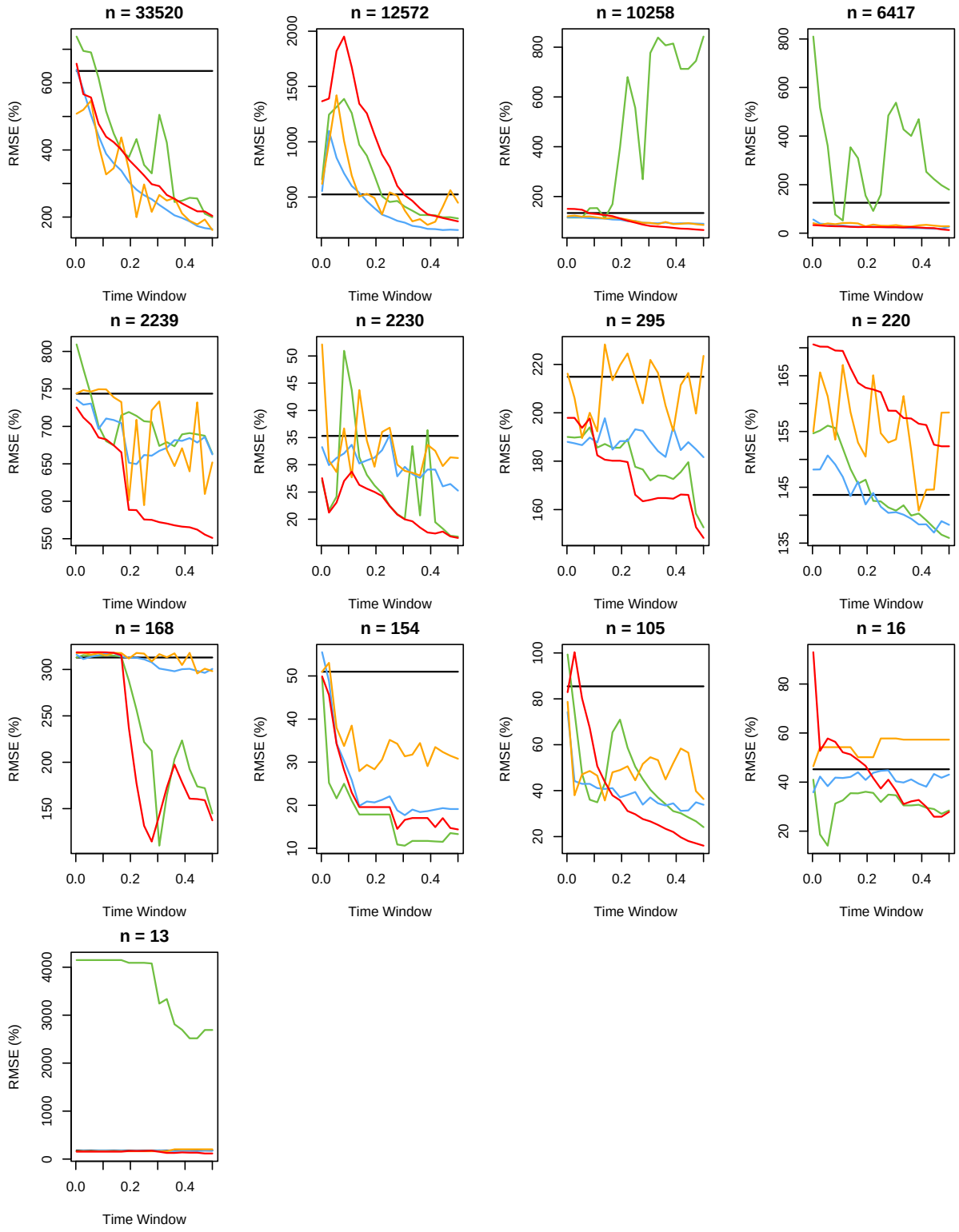
Results

Figure A1. Performance of models per data set as a function of sample size.



Note: Each diagram shows one data set. The total sample size of each data set is provided above each diagram. The color scheme is the same as in Figure 2.

Figure A2. Performance of models per data set as a function of time-to-prediction.



Note: Each diagram shows one data set. The total sample size of each data set is provided above each diagram. The color scheme is the same as in Figure 2.