The University of Edinburgh School of Mathematics

Testing a New Metric for Uplift Models

by

Oscar Mesalles Naranjo

Dissertation Presented for the Degree of
MSc in Statistics and Operational Research

August 2012

Supervised by
Dr Nicholas J. Radcliffe

Declaration:

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(Oscar Mesalles Naranjo)

Acknowledgements

I would like to thank Dr. Radcliffe for his help and guidance during this project.

Abstract

In this paper we evaluate the validity of a new metric (named Moment of Uplift, Υ) for assessing uplift models. Uplift models predict the change in behaviour as a direct result of a treatment, and they are most useful when they are both accurate and bold (audacious and clear to the eye) in their predictions, therefore Υ must capture those two characteristics. Currently there are no metrics capable of capturing both of these characteristics. To test the validity of Υ we analyze a given formulation for Υ and propose more useful variations. We also generate other metrics based on certain desirable characteristics of an uplift model. We produce a total of $36,000$ for thirty six datasets, use Υ and the rest of the metrics to help in choosing the best model and assess the models selected as proxy for assessing the metric. We conclude that one of the variations proposed for Υ (a quadratic form Υ_Q) is capable of identifying both accurate and bold models and makes a better model selection than the current metrics used, such as Qini coefficient and maximum uplift.

Contents

1 Introduction

1.1 Structure

Section 1 introduces the context for uplift modelling problems, the mechanics of its implementation, the current techniques used and the problems associated with assessing an uplift model. In section 2 we explain how we tackle the problem of evaluating a new quality measure or metric. We analyze its formulation and propose alternatives. In section 3 we include a considerable amount of models by presenting one of its possible graphical representation, the uplift by decile graph. We assess the results both subjectively and with a more objective approach. Finally, we conclude in section 5 summarizing our findings.

1.2 The Uplift Problem and Its Applications

Uplift modelling is a type of statistical predictive modelling, the aim of which is to detect true changes in the probability of an outcome given a treatment. This paper focuses on analyzing whether a new metric, named Moment of Uplift [24] (symbolized with the Greek letter Υ) can be used as a quality measure for uplift models, looking at the variations and refinements.

In the context of uplift modelling if we assume that we have a population X of entities which behave in a specific way with regards to an outcome, then for each individual entity we can decide if we treat it or not, in which case we expect to see a change in its behaviour.

This modelling technique was created in the environment of marketing campaigns [27] and customer relationship management and it is easier to understand within this context. In this case the population X is our current or prospective customers, a treatment is the campaign, for instance an offering of a product at a discounted price to a specific customer, and the outcome could be a binary variable indicating if the customer bought the product.

Uplift modelling, applied to marketing, challenges the traditional direct marketing view that a customer will not purchase unless s/he is convinced to do so with some kind of promotion. It recognises that direct marketing works for some customers better than others, and that not all customer purchases are a consequence of the treatment: a customer may have bought the product independently of the campaign. Even more, there will be individuals where the treatment may have a negative effect, which is a known fact within the advertising community [1].

Uplift modelling promises to identify the individuals that are the best target for a treatment, which can help to reduce cost and increase campaign efficiency. It does that calculating the uplift $u(x)$ for each individual which is defined as:

Definition 1.1 *We define the uplift, $u(x)$, for an individual x as*

$$u(x) := \mathrm{P}\left(O = 1 \mid T; x\right) - \mathrm{P}\left(O = 0 \mid \overline{T}; x\right)$$

where $\mathrm{P}(A \mid B)$ denotes the probability of A given B, O is a random binary variable referring to a positive customer response, $O = 1$, T indicates the event that the individual x was treated, and $\overline{T}$ indicates it was not treated. The event $O = 1$ will be called conversion and referred to as C.

In opposition, other modelling techniques try to predict the value of $\mathrm{P}(C \mid x; T)$, which does not take into account the effect of the campaign.

In definition 1.1 there are a few points to note. First, I have defined uplift at a given point x. Although this is conceptually easy to grasp (for a single individual, the fact that we treat or not treat him may modify his chances of behaving in a certain way), it is impossible to measure. This is because both actions on a single individual, treat and not treat, cannot be done at the same moment in time. To overcome this problem one may be tempted to predict $\mathrm{P}\left(C \mid \overline{T}; x\right)$ (see [4] for a discussion on this approach), but that would not give us the true uplift for the customer, but an uplift taking our prediction from historical records as the baseline.

Second, I have restricted the random variable O to be binary. This report focuses on this case, although it is possible to use uplift models for continuous outcome or in some cases combine uplift modelling with other modelling techniques to get the uplift for a continuous outcome [27]. In any case, the uplift for continuous outcome, for instance how much a customer spends, can be defined in terms of expected value instead of the probabilities:

$$u_c(x) := \mathrm{E}(O \mid T; x) - \mathrm{E}\left(O \mid \overline{T}; x\right) \forall x \in X \tag{1.1}$$

Third, I have restricted my research to situations where there are only two possibilities within the treatment, those are treat or not to treat, although uplift modelling has been extended to evaluate different possible treatments and helping in finding out what the best treatment is for an individual [32].

1.3 How Uplift Model Is Implemented

For all predictive modelling technique one must have a population sample that provides the information necessary to build the model. I will refer to this population sample as a dataset. Generally, one wishes to predict the outcome variable (or response) based on the value of other variables, called predictors or explanatory variables.

For implementing an uplift model the dataset must be made of two disjoint sets: a control set ($\overline{T}$) and a treatment (or treated) set (T), which are obtained sampling at random and without replacement from the overall population. After building the two sets, one applies the treatment *only* to the treatment set. The binary variable indicating whether the record belongs to T or $\overline{T}$, is not considered as an explanatory variable. Later, some time after the treatment – a period called the observation window – one collects data regarding the outcome variable for both treated and

control set. For instance, in marketing, one may wish to test the effectiveness of a campaign on a small set before launching it on a bigger scale. To assess how successful the campaign was, one control and one treatment group are randomly drawn and an uplift model is built based on the small-scale campaign. Hopefully, the uplift model identifies the individuals that are more likely to be influenced by the campaign, so it can be targeted to similar individuals from the overall population.

The outcome of an uplift model is a function $u(x) : X \rightarrow \mathbb{R}$ returning a prediction for the uplift on that record, which can be interpreted as a score. Records with higher scores will be the ones to target, and records with negative prediction values the ones to avoid.

As in any modelling process satisfying a business need, there will be a time spent designing the model and understanding its specific requirements [16, 20], which is the part only a human can do, and a time building the actual model, which can be done by a computer. In the process of building the actual model, a computer may have to decide from a range of possible options, and it is in this case where a quality measure for a model is most useful.

1.4 Overview of Current Uplift Modelling Techniques

A requirement for this dissertation was to use different uplift models and assess the adequacy of their predictions. In this section I briefly explain the models used. A detailed review of the different techniques can be found in [32] and [27].

There are currently two approaches to build uplift models: a) tree-based methods and b) regression-based methods. Tree-based methods assimilate CART (Classification And Regression Trees) methodology [6] to build a tree whose terminal nodes classify the population in segments with different predicted uplift. Algorithms based on CART need to define:

1. A splitting criteria to decide at each new node how to split the population in two segments that supposedly present different uplift;

2. A stopping rule indicating that the uplift in one node is considered the same for all the records belonging to it;

3. A pruning methodology to reduce the final tree size.

Different tree-based algorithms redefine the above three points, see for instance [27], [3] and [32]. To work out the predicted uplift on a terminal node k, which we refer to as u_{kp}, one applies the tree classifier algorithm on both the treated and on the control group and finds out:

- r_k^t: the number of conversions from the treated group that fall in node k;

- n_k^t: the total number of records from the treated group that fall in node k;

- r_k^c: the number of conversions from the control group that fall in node k;
- n_k^c: the total number of records from the control group that fall in node k.

Then estimates the probability of success, in each node k and group, as the proportion of successes, and subtracts the estimates between the treated and the control group for each node k:

$$u_{kp} = \frac{r_k^t}{n_k^t} - \frac{r_k^c}{n_k^c} \tag{1.2}$$

When applying the tree method to a new dataset (one that has not been used for constructing the tree) the above steps are used to calculate the actual uplift for each customer segment (u_{ka}), which can then be compared with the prediction u_{kp}. However, a single uplift tree is not always used as a model [23]. A method known as bagging [5] is more common: the bagged model consists of a set (or "bag") of uplift trees, each of which makes their predictions for an incoming record x. The actual prediction returned is the average of those values, which helps reducing over fitting the model to the specific dataset used for building it [6]. Using bagging one obtains a point wise predicted uplift for each record [1].

Regression-based methods use regression techniques (like ordinary and generalised linear models) to build an analytical function producing a score for each record. There are two methods known. The Two Model approach is applied to generate two predictive models, using both the control and treated set independently, and subtracting the results of each model[2]. Another method [17], which we will refer to as VLO method, builds a single regression from both the control and treated dataset using the explanatory variables, and adding an interaction term between each predictor variable and a dummy variable t, which indicates whether the record was treated ($t = 1$) or not treated ($t = 0$). In this case, naming f as the regression model function, the predicted uplift is given by $u(x) = f(x, t = 1) - f(x, t = 0)$, where x is the predictor variables for each individual.

To the author's knowledge no additional clustering techniques [15, 2], other than trees, are used to generate uplift models. However, uplift models could be interpreted as a clustering problem, where one clusters records that present the same uplift. It is straight forward to visualize it in two dimensions, see figure 1.1 for an example. The main problem is that you do not know the segment (or class) a record belongs to. There are clustering techniques that deal with unknown classes, namely latent class analysis [34] (LCA), but using this method for an uplift model is currently not available. One difference on the potential use of LCA would be that it would return the probability of a record belonging to a class, instead of assigning it to a class. However LCA is based on the assumption that the classes follow a known probability distribution function.

[1] In fact one obtains thinner segments.

[2] Actually, this approach can also be applied to any tree-based methods.

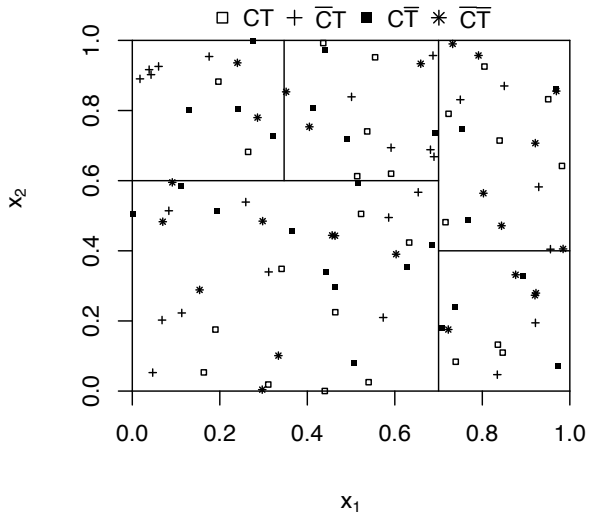

Figure 1.1: Uplift as a standard classification problem with two predictor variables x_1 and x_2, the different point types refer to the treated (T) or non-treated ($\overline{T}$), and conversion (C) or no-conversion ($\overline{C}$). The aim is *not* to separate the different points, but to group them in the right sets. In this case the sets are defined by a CART with four nodes and the following conditions: node 1: $x_1 < 0.7$, node 2: $x_2 < 0.6$, node 3: $x_2 < 0.4$, node 4: $x_1 < 0.37$. Once the sets are created one calculates the uplift for each set as the rate of successes in the treated minus the rate of successes in the control

Interpreting uplift models as a clustering problem could help in developing new quality measures from them. Unfortunately, the author's research in this direction has not produced any result.

1.5 Notation

Before proceeding any further we find it useful to compile the terminology used in this paper, as a reference for the reader. We use:

- X referring to a sample of records from an overall population. $X = \overline{T} \cup T$, where $\overline{T}$ is the control group and T the treated group;
- x refers to a specific record in X, and/or the predictor values taken by that record;
- $u(x)$ refers to the point wise predicted uplift given by a model;
- N the total number of records in X;
- N^t, R^t total number of records and total number of conversions in T;
- N^c, R^c total number of records and total number of conversions in $\overline{T}$;
- K the number of segments defined by an uplift model, and k a specific segment;
- n_k^t, r_k^t the number of records and the number of conversions for T in segment k;

- n_k^c, r_k^c the number of records and the number of conversions for $\overline{T}$ in segment k;
- u_{kp} the predicted uplift in segment k;
- u_{ka} the actual uplift in segment k, calculated as shown in equation 1.2;
- N_k^t, R_k^t the cumulative number of records and conversions in T from segment 1 to k;
- N_k^c, R_k^c the cumulative number of records and conversions in $\overline{T}$ from segment 1 to k;
- $U_{ka} = {}^{R_k^t}/_{N_k^t} - {}^{R_k^t}/_{N_k^t}$, the actual cumulative uplift up to segment k;
- μ the overall uplift for sample X, $\mu = {}^{R^t}/_{N^t} - {}^{R^c}/_{N^c}$;
- $u'_{kp} := u_{kp} - \mu$ and $u'_{ka} := u_{ka} - \mu$ refer to the uplift above or below the overall uplift;
- $\varepsilon := \frac{1}{N^t+N^c} \sum_{k=1}^{K} (n_k^t + n_k^c)|u_{kp} - u_{ka}|$ is an estimate of average error in predicting $u(x)$;
- $U_{max} := \max_k(U_{ka})$ is the maximum uplift achievable according to a model;
- r_s refers to the Spearman correlation coefficient between u_{kp} and u_{ka} for $k = 1, \cdots, K$
- $\nu_p := |\sum_{k=1}^{K} \text{sign}(u_{kp})|$ is a measure of the negative effect detected by a model;
- $s_p := \max_k(u_{kp}) - \min_k(u_{kp})$ is the range of predictions returned by a model. It is equal to $s_p = u_{1p} - u_{Kp}$ when the segments k refer to percentiles obtained after sorting $u(x)$.
- ROI^{pred} is the maximum return on investment achievable according to an uplift model predictions.
- ROI^{act} is the maximum actual return on investment achieved by an uplift model, if one follows its recommendations.

1.6 How to Assess an Uplift Model

After an uplift model is built one would like to evaluate its performance based on objective and easy to understand criteria. A single number working as a quality measure would be an ideal situation. This would help in identifying the best model out of a given list and could be used as an objective function in an algorithm to build a model.

Classifier algorithms applied to standard classification problems (one where the class of each item is determined) use quality measures based on the number of items that are wrongly classified [28], but that does not apply to our problem, as we do not know the segment (or class) a record should belong to.

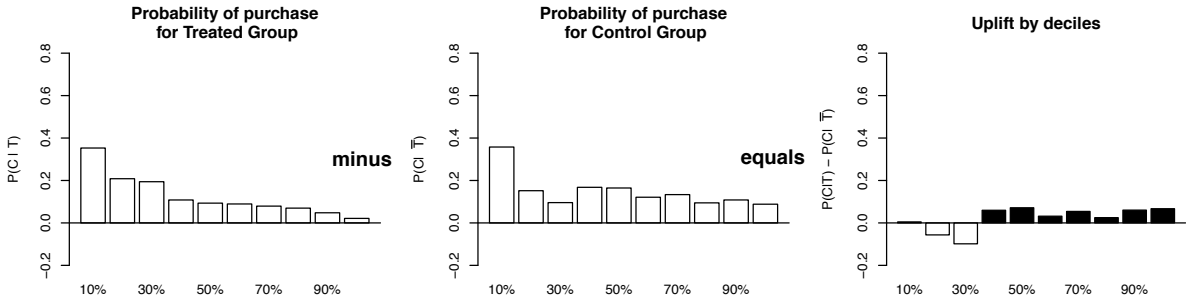

Figure 1.2: Conceptually the uplift by deciles graph is defined as the subtraction of two probabilities: $\mathrm{P}\left(C \mid T\right) - \mathrm{P}\left(C \mid \overline{T}\right)$. In the uplift by deciles graph on the right, black bars indicate positive uplift and white bars negative effect.

Regression methods use the residual error (or the deviance for the case of generalized linear models) between the prediction and the actual values, but it cannot be applied to uplift model as we do not know the actual point wise uplift.

Practitioners within the marketing community assess campaigns plotting the probability that a given group of customers will buy. The customer grouping is usually in percentiles, most commonly by deciles. This assumes that all customers within a segment have the same probability of buying. Using the estimated probability of success for T and $\overline{T}$, and applying formula 1.2 to each decile, one obtains a graph showing the uplift by deciles, which is a typical representation for an uplift model, see figure 1.2.

For regression-based uplift models, or any model that returns a point wise predicted uplift one can plot the uplift by deciles graph using the following procedure:

- Calculate the uplift $u(x)$ for all $x \in X = T \cup \overline{T}$;
- Sort both treated and control records by the uplift value $u(x)$;
- Find the list of uplift values $\{b_0 \cdots b_K, K = 10\}$ that define the deciles in X;
- Calculate the average predicted uplift per segment k, u_{kp}, using the point wise prediction $u(x)$. The following estimate for u_{kp} could be used:

$$u_{kp} = \frac{1}{n_k^c + n_k^t} \sum_{x: b_{k-1} < u(x) \leq b_k} u(x) \tag{1.3}$$

  Although this is the most reasonable approximation for u_{kp} other options are possible, such as the maximum ($u_{kp} = b_k$) or the minimum ($u_{kp} = b_{k-1}$) within the segment;

- From the values r_k^t, n_k^t, r_k^c and n_k^c for each segment, calculate u_{ka} for each decile using formula 1.2.

Once one has u_{kp} and u_{ka} for each segment k, a graph such as the one showed on figure 1.3 can be plotted.

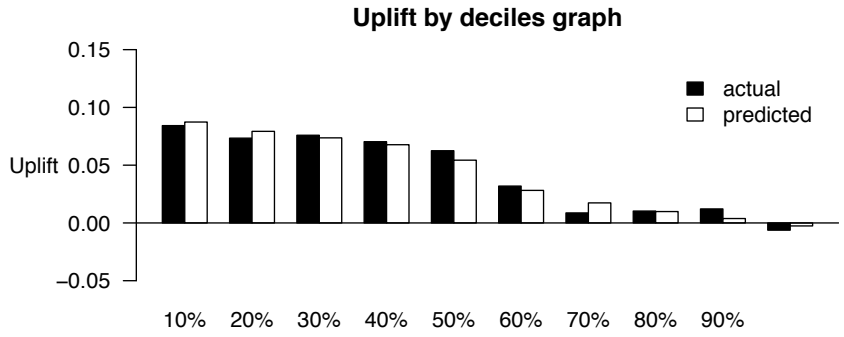

Figure 1.3: The uplift by deciles graph is a common representation for assessing the success of a marketing campaign using an uplift model. The different segments, in this case deciles, are sorted from higher to lower predicted uplift u_{kp}, and besides them the corresponding actual uplift u_{ka} is plotted.

When calculating the actual uplift for segment k, we assume that records in the treated group have a counterpart in the control group, and that this matching is given by our predicted uplift function. Although this is intuitive and all uplift models in literature seem to follow this process, we have not found any formal analysis of this approach.

Interestingly, a matching strategy between control and treated populations is also a requirement in causal inference for observational studies, [3] and within that field multiple techniques have been developed, such as kernel matching [13], propensity score matching [30] or Mahalanobis distance matching [31].

Presenting the predicted and the actual uplift by percentile side by side in a bar plot, where predicted uplift is sorted from higher to lower values, as in figure 1.3, one can identify if the model has certain desirable characteristics, characteristics that a quality measure should reflect.

Those desirable characteristics may depend on the specifics of the problem, but Radcliffe [27] proposes five generic features for evaluating an uplift model:

1. Monotonicity of incremental gains; are the segments with higher actual uplift presented before the ones with lower uplift? This is particularly relevant when there are substantial differences between segments uplift;

2. Prediction error or tight validation; is the prediction close to the actual value?

3. Spread or range of prediction, measurable as the subtraction of the maximum and the minimum predicted uplift;

4. Maximum cumulative uplift achievable, that is how much extra conversion the model predicts at its peak;

[3] In observational studies one does not control which items are assigned to the control or treatment group, and there could be a relationship between the variables, the outcome and the group assignment, while in uplift models we assume that both groups were a random sample of the population.

5. Impact at cut-off, or otherwise the cumulative uplift at a given percentile.

Both monotonicity and prediction error refer to how accurate the model is. Spread, maximum cumulative uplift and impact at cut-off, indicate how good the model is in separating segments with different behaviours. We will refer to that as **boldness**[4]. Boldness is related with performance: an indication of how much money a marketing campaign can return [19].

The ideal model optimizes all of the above features. Having only one is not an indication of a good model. For instance, a model accurately predicting the average uplift for all segments is not useful, because one does not know which segments perform better. Another example is when the model makes a wide range of prediction, but lacks accuracy.

Although boldness and accuracy are not conflicting objectives, it is more challenging to find accurate models that make bold predictions than finding conservative models that make accurate predictions. This is easier to understand using again the model predicting the same uplift for each segment. This model is not bold, but it is possibly accurate, as a random selection of individuals from the whole sample is likely to match its prediction. On the other hand, a model that detects a marked uplift in a segment, will fail to reproduce that prediction on another sample data if the uplift happened just by chance. A bold model is more likely to identify noise (extreme values consequence of the natural variability of the data) as an actual effect on the sample set. This is directly related to the well-known statistical phenomenon regression to the mean[11].

In the context of marketing, accuracy is not the most important characteristic, giving more preference to boldness. This is because marketers are generally not interested in predicting the behaviour of the whole population, but aim to find the top subset of the population with the highest uplift and that can provide higher return on investment [19, 21]. The prime characteristics they focus on are impact at cut-off or maximum uplift. Nevertheless, having better estimates improves the predictions for expected returns.

Negative effect is a sixth characteristic one may want to detect. Negative effect is the presence of a segment where the treatment causes fewer conversions than no treatment at all. Figure 1.4 illustrates the presence of negative effect in the last deciles.

The reader will realise that assessing an uplift model is not as straight forward as assessing more conventional models. Multiple models may be useful in the same case, for instance both models in 1.5 could be used, although the one on the left differentiates better the segments.

Monotonicity, prediction error and spread are easy to assess from the uplift by deciles graph but uplift at cut-off and maximum uplift are not. A second chart, named cumulative gains chart for uplift or Qini curve [23], can help to visualize the maximum cumulative uplift and the impact at cut-off. From the Qini curve one extracts a summary statistic that can be used as a metrics

[4]From the Online Oxford Dictionary, bold: 4a, In bad sense: Audacious, presumptuous, too forward; the opposite of modest; 8a Standing out to the view; striking to the eye.

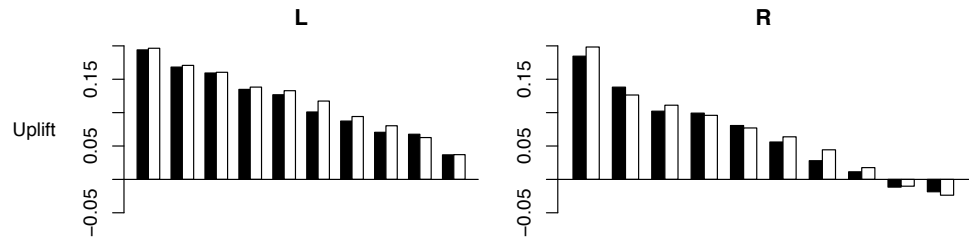

Figure 1.4: Models detecting a negative effect may be preferable in some cases. A campaigner may have the resources to target the whole population, but in this case targeting the last segments may be counterproductive.

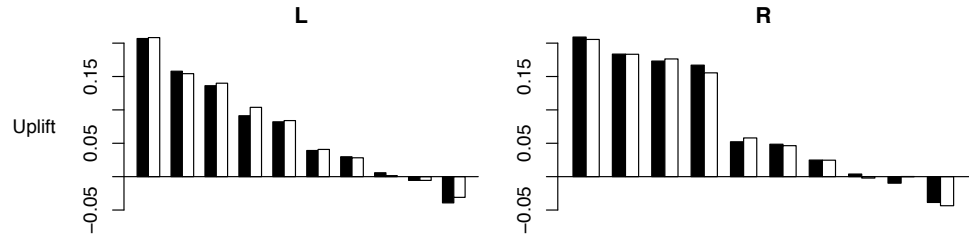

Figure 1.5: Assessing an uplift model is not a clear cut. In this case two accurate models with different degree of boldness, which one would be most useful?

for uplift models, but it does not reflect all the model characteristics, therefore another quality measure would be useful when assessing uplift models.

This is in fact the topic of discussion for this dissertation. In this paper, we evaluate if a new proposed metric for uplift modelling (named Υ, Moment of Uplift) is a reflection of model quality, where quality is considered based on the desirable characteristics explained above: monotonicity, prediction error, spread, maximum uplift and uplift at cut-off.

1.7 Qini Curve and Qini Coefficient

Qini curve, introduced by Radcliffe [23], is a generalization of the gains chart. It is easier to understand Qini curve starting by explaining the gains chart. The steps to draw a gains chart are as follow. Consider a model that tries to predict a binary outcome in a given population. Consider that the model produces a score for each individual in a way that higher values mean that the individual has more chances of success. In this scenario, one builds a gains chart first by sorting for decreasing score all individuals, and then plotting the number of successes versus the number of individuals targeted, assuming that one first targets individuals with high score. An example is presented in table 1.1, and a figure besides it, in that figure both X and Y axis can be transformed to represent percentage of the sample size.

From the figure besides table 1.1, it is clear that the optimum model should score x_4 below x_5, so the gains chart would go up to the top in a straight line with 45° angle. A model that makes

Individual	Is it a purchase	Score given by the model	Number of cumulative purchases	Number targeted
x_1	Yes	1.2	1	1
x_2	Yes	1.0	2	2
x_3	Yes	0.9	3	3
x_4	No	0.8	3	4
x_5	Yes	0.7	4	5
x_6	No	0.5	4	6
x_7	No	0.1	4	7

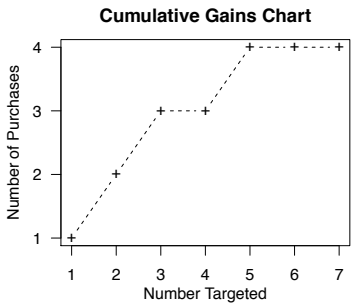

Table 1.1: Table used to plot the cumulative gains chart on the right.

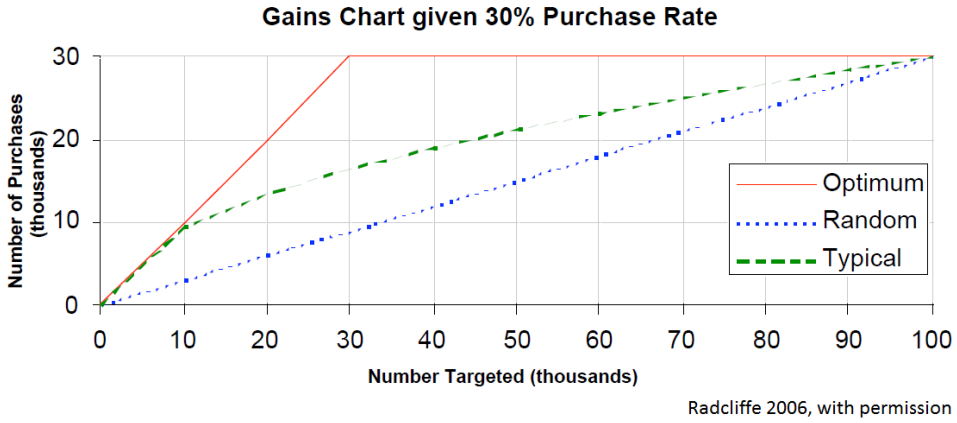

Figure 1.6: Gains chart for a typical model, the optimum model and the theoretical line for a random model.

a random sorting of the individuals would most likely result in a "straight" line going from $(0, 0)$ to $(7, 4)$ in our example. A more sophisticated example presenting all three curves (optimum, random and typical) is shown in figure 1.6.

From the gains chart, one defines the Gini coefficient as the ratio between two areas: a) the area above the diagonal and the cumulative gains chart; and b) the area between the diagonal and the optimum.

Instead of presenting the cumulative number of purchases in the vertical axis, one can plot the *incremental* number of purchases. Doing so, one obtains the Qini curve. The process to work out the Qini curve is similar to the gains chart, but now one works with customer segments. Table 1.2 shows an example of the calculations required. In this case the score is not presented, and the reader should assume that the different customer segments are already sorted by the predicted uplift from highest to lowest. For simplicity, table 1.2 presents the case where the number of cumulative targeted in both control and treated is the same for each segment. In general, those numbers differ, in which case one applies a correcting factor to work out the cumulative extra purchases. Using the notation from table 1.2, the cumulative extra purchase, when the control and treated segment sizes differ, becomes $U_{ka} = R_k^t - N_k^t R_k^c / N_k^c$. The Qini curve presents U_{ka}

versus N_k^t.

Table 1.2: Example of the calculations done for plotting the Qini curve. Segments are sorted by the uplift, that is uplift in segment A is higher than in B and so on. The subtraction of the cumulative purchases between treated and control is the cumulative extra purchases

	Treated		Control		Treated - Control
Segment k	Cumulative purchases (R_k^t)	Cumulative targeted (N_k^t)	Cumulative purchases (R_k^c)	Cumulative targeted (N_k^c)	Cumulative extra purchases (U_{ka})
A(10%)	10	100	3	100	7
B(20%)	31	200	11	200	20
C(30%)	45	300	20	300	25
...	...	...	...	...	...

From the Qini curve one can obtain a Qini coefficient in the same way as for the Gini curve. Although in this case the optimum uplift curve can have multiple forms because the treatment may present negative effect. First let us consider the case without any negative effect. The optimum Qini curve presents a similar shape to the optimum gains chart. The curve goes from $(0, 0)$ to (M, M), where M are the overall extra purchases in the treatment group versus the control group, and then it goes horizontally. Now consider what happens when negative effect is present. Negative effect can be of different size. For instance, suppose a control and treated group each with 100000 individual, and assume 30000 overall purchases in the treated group and 10000 overall purchases in the control group. The overall extra sales in the treated group are 20000. The campaign may have persuaded 20000 individuals in the treatment group, or it may have persuaded 30000 and deterred 10000 individuals (the ones that would have bought without the campaign, as indicated by the control group). Obviously any negative effect is limited by the overall purchases in the control group and must be compensated with extra purchases in the treated group. This limit in the negative effect determines the shape of the optimum Qini curve when negative effects are present. An example of Qini curve and the optimum Qini curve are presented in figure 1.7.

Qini curves allow identifying at a glance the maximum uplift achievable by a model, which corresponds to its maximum, and the uplift for any cut-off value.

Using the same method as for the Gini curve, one defines the Qini coefficients, only that in this case, as there are two possible optimum curves (the one without negative effect and the one considering the maximum negative effect possible) two definitions are possible:

$$q_0 = \frac{a}{b} \text{ and } Q = \frac{a}{B} \tag{1.4}$$

where a is the area between the random and the Qini curve, b is the area between the random and the optimum without negative effect, and B is the area between the random and the optimum with maximum negative effect.

Qini curve is not necessarily a piecewise function. It is possible to define it with more granu-

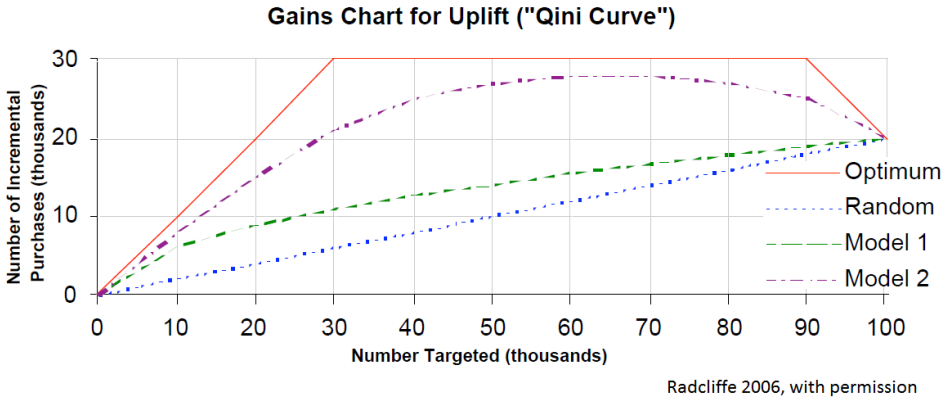

Figure 1.7: Qini curve presenting the extra purchases between the treated and control group in relation to the targeted population. A random model follows a straight line from $(0, 0)$ to the overall uplift. The optimum model identifies the best targets at the beginning and if negative effects are present, those happen at the end. The maximum negative effect possible is the overall purchase rate on the control group. The negative effect is compensated by extra purchases in the treated group, to obtain the overall purchases. A good model will be in between the optimum curve and the random curve.

larity simply by considering the cumulative uplift obtained when including a single individual at a time. However, in practice it will be plotted as piecewise function. Knowing U_{ka} in segment k, one can calculate the area underneath the curve using the trapezoidal rule to obtain that:

$$q_0 \propto \sum_{k=1}^{K} U_{ka} \tag{1.5}$$

where the proportionality constant is independent of the model. This formulation illustrates the fact that Qini coefficient does not take into account the value of the predicted uplift for each segment, but only how individual records have been sorted, which determines its allocation to a segment. Qini measures are rank based. Any monotone transformation of the score leaves the Qini coefficients unchanged (see figure 1.8).

Therefore, our aim is to find a measure that does not suffer from this problem, and can be combined with or replace Qini in the selection of a good model.

1.8 Assessing Marketing Campaign Profitability

The promise of an uplift model applied in marketing is that it can both reduce the cost of the campaign and increase the response rate, as targets are better defined. One would like to assess that statement in financial terms. There are plenty of performance indicators that can be applied [5] such as, total revenue, total profit or return on investment (ROI).

In this paper we use ROI as a possible quality measure for our models, being ROI = P/C with

[5] See for instance `http://www.lums.lancs.ac.uk/files/promo_promax_report.pdf`

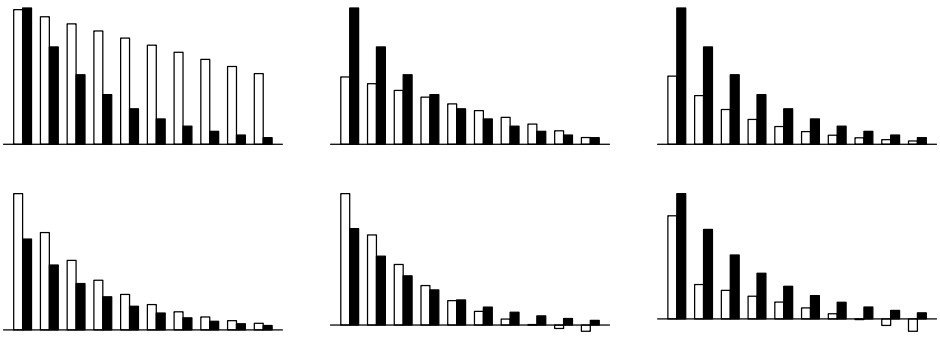

Figure 1.8: Uplift by deciles graph for different uplift models. They all have the same Qini coefficient. Starting from a single model, one can generate other predictive models applying a monotone transformation to the predicted uplift. Those transformation do not change the ranking order of the predicted values, therefore all Qini curves and Qini coefficients will be the same as the original model.

P being the extra profit from the campaign and C the cost of the campaign, and assume that maximizing ROI is a valid selection criterion for an uplift model. ROI focuses on the efficiency of the campaign (or minimizing the expenditure). We use the following assumptions to calculate it:

- We set the campaign fixed costs to be zero. We can consider any other value without increasing the complexity of the calculation, but we use zero to be more specific;
- c_k is the variable cost for a single individual when targeting the campaign to segment k;
- p_k is the expected monetary contribution to profit for a single individual in segment k;

For a given uplift model, one can calculate ROI either using the predicted uplift per segment u_{kp} or the actual cumulative uplift $U_k a$. P_j^{pred} and P_j^{act} refer to the predicted and expected extra profit from targeting the campaign up to segment j, similarly C_j refers to the cost of targeting up to segment j. Then,

$$C_j = \sum_{k=1}^{j} n_k c_k \tag{1.6}$$

$$P_j^{pred} = \sum_{k=1}^{j} u_k a n_k p_k \tag{1.7}$$

$$P_j^{act} = \sum_{k=1}^{j} (U_{kp} - U_{(k-1)p}) n_k p_k, \text{ with } U_0 = 0 \tag{1.8}$$

To be more specific if $n_k = {}^{N}/_{K}$, which corresponds to a segmentation by percentiles, $c_k = c$ and $p_k = p \; \forall k = 1 \dots K$, then:

$$\text{ROI}_j^{pred} = \frac{p}{jc} \sum_{k=1}^{j} u_{kp} \tag{1.9}$$

$$\text{ROI}_j^{act} = \frac{p}{c} U_{ja} \tag{1.10}$$

Therefore finding the maximum uplift for a given model is equivalent to maximizing ROI_j^{act}. We refer to $\text{ROI}^{act} := \max_j\{\text{ROI}_j^{act}\}$ and $\text{ROI}^{pred} := \max_j\{\text{ROI}_j^{pred}\}$

1.9 Moment of Uplift Motivation

The motivation for an alternative quality measure for uplift models is to identify if a model is both accurate and bold. At least the following terms are required for a formulation that aims to capture those properties:

- The overall uplift μ of the sample, which acts as a reference point to distinguish a segment that outperforms the rest;
- Both the prediction and actual values for uplift.

A familiar formulation combining these values is $\sum_{i=1}^{N}(u_{ip}-\mu)(u_{ia}-\mu)$, where in this case u_{ip} and u_{ia} refer to the point wise predicted and actual uplift for each record $x_i \in X$.

As in uplift modelling one does not know the actual point wise uplift, it is necessary to work with segments, therefore one can define a metrics, which will be called Moment of Uplift and symbolized with Υ, as:

$$\Upsilon_1 := \frac{1}{N^t}\sum_{i=1}^{K} n_i^t(u_{ip}-\mu)(u_{ia}-\mu) \tag{1.11}$$

which is equal to the weighted covariance $\text{cov}(u_p, u_a)$ with weights n_i^t, as long as $\mu = \mu_p$, where $\mu_p := \frac{1}{N^t}\sum_{i=1}^{K} n_i^t u_{ip}$ is the weighted arithmetic mean of the predicted uplift.

Working with records in the treated group (n_i^k and N^t) for equation 1.11, instead of using $n_i = n_i^t + n_i^c$ and $N = N^t + N^c$, simplifies the equation when replacing the value of $u_{ia} = {}^{r_i^t}/_{n_i^t} - {}^{r_i^c}/_{n_i^c}$ in Υ_1:

$$\Upsilon_1 = \frac{1}{N^t}\sum_{i=1}^{K}(r_i^t - \frac{r_i^c}{n_i^c}n_i^t)(u_{ip}-\mu) \text{ if } \mu = \mu_p \tag{1.12}$$

This formulation has a simple interpretation when there are no negative effect in any segment k: it counts the average number of extra customer, as long as the uplift is above the overall uplift.

Equation 1.11 has the following basic properties:

- If $u'_{ip} = u_{ip} - \mu$ and $u'_{ia} = u_{ia} - \mu$ have the same sign, the segment makes a positive contribution;
- If u'_{ip} and u'_{ia} have opposite signs, the segment makes a negative contribution;
- The further away the uplifts (predicted and actual) are from the overall, the larger are the contributions to (so bold, correct predictions are rewarded);

Therefore models predicting that some segments are further way from the overall have a higher value for Υ_1. Unfortunately, it does not penalize inaccuracy as once both the predicted and the

actual uplift are over (or below) the overall uplift the segment contribution increases for any value of the uplift.

A refinement on that first formulation, which tries to penalize inaccuracy and maintains the basic properties above, is:

$$a_i := \begin{cases} \left[\min(|u'_{ip}|, |u'_{ia}|) - |u'_{ip} - u'_{ia}|\right]^2, & \text{if } u'_{ip}u'_{ia} > 0 \\ -(u'_{ip} - u'_{ia})^2, & \text{otherwise.} \end{cases} \tag{1.13}$$

$$\Upsilon_Q^{min} := \frac{1}{N^t} \sum_{i=1}^{K} n_i^t a_i \tag{1.14}$$

While a quadratic form seems the natural derivation from the initial proposal, a linear form may reveal useful:

$$c_i := \begin{cases} \left|\min(|u'_{ip}|, |u'_{ia}|) - |u'_{ip} - u'_{ia}|\right|, & \text{if } u'_{ip}u'_{ia} > 0 \\ -|u'_{ip} - u'_{ia}|, & \text{otherwise.} \end{cases} \tag{1.15}$$

$$\Upsilon_L^{min} := \frac{1}{N^t} \sum_{i=1}^{K} n_i^t c_i \tag{1.16}$$

When calculating u'_{ip} and u'_{ia} for the sample used for creating a model, the overall predicted uplift μ_p and the actual overall uplift $\mu_a := \mu$ are very close and we can use either μ_p or μ_a, but this is not the case for a new dataset evaluated with a model previously generated. Although one expects both values to be close if the new dataset is similar to the one use for building the model. In all cases, we will use $u'_{kp} = u_{kp} - \mu_p$ and $u'_{ka} = u_{ka} - \mu_a$. One could argue that using only μ_a can also be useful, as then only segments above and below the actual overall uplift for both actual and predicted segment uplift contribute positively to Υ.

Finally, in our calculations for μ_p and Υ, only the size of the treated population for each segment was taken into account. If both groups T and $\overline{T}$ are of similar size, it does not matter whether you use n_k^t or $n_k^t + n_k^c$, but for consistency with formula 1.12 using n_k^t is preferred.

2 Methodology

A typical scenario for a machine learning research problem is proving that a new method for modelling or extract information from a dataset is better than previous methods [8]. The common approach is to choose a variety of quality measurements for model assessment, run the new algorithm in some test data, optimise the value of the quality measure (assuming that this means an improved model) and compare the results with current state-of-the-art methods. In this paper, although the objective is not as above, the approach is similar. The requirement is to characterise a new measurement for uplift models, Υ Moment of Uplift, and decide if it can be used to select a model with desirable characteristics, specifically accuracy and boldness.

The first step we took was to understand what Υ tries to achieve and what its expected behaviour is. Recall that Υ is made up of the sum of individual contributions for each model segment, therefore a 3D graphical representation of an individual segment contribution was inspected. Based on the 3D representation, a discussion on the benefits of Υ is done and some changes to the analytical formulation are presented. In addition, more quality measures are proposed that intend to detect accurate and bold models.

Once the Υ formulation was satisfactory, the next step was to generate multiple models on various datasets, calculate Υ, sort them according to Υ and identify what differentiates models with increasing values of Υ. This approach for assessing a new quality measure has previously been used for finding new ways of assessing classifiers [12]. The difference with an uplift model, compared with a standard classification problem, is that in a standard classification problem the misclassification rate is always an indication of how good a classifier is, while in an uplift model we do not have a standard metric.

In traditional modelling the perfect model achieves $\sum_{i=1}^{N} |f(x_i) - y_i| = 0$, which is only possible if the mapping $x_i \longrightarrow y_i$ is injective. For uplift modelling one does not have $f(x)$, so can not know if a model is perfect in a general case, but one can create simulated data where the uplift pattern is well defined, and build a model that exactly matches that pattern. Initially, we considered assessing the quality measure by hiding the perfect model in a myriad of models. However, we discarded that approach in the end because, although the quality measure earmarked the perfect model, that did not demonstrate whether the quality measure would select a useful model for a dataset with a complex or unknown uplift pattern.

To help in characterising the quality measures, we quantified the desirable features explained in section 1.6. This was the key criteria we used to decide the usefulness of the different quality measures. Judging an uplift model has a high subjective component but by quantifying the desirable features we can remove the subjectivity and simplify the selection of a useful model.

Spread, maximum cumulative uplift and uplift at cut-off are already quantitative magnitudes.

For the cut-off, we used in fact the uplift at first decile, which is a common quality measure for uplift models in marketing [21]. We use the following measures for those characteristics:

- Spread, $s_p := \max_k(u_{kp}) - \min_k(u_{kp})$
- Maximum cumulative uplift, $U_{max} := \max_k(U_{ka})$
- Uplift at first decile, u_{1p} and u_{1a}

Therefore we only needed to quantify accuracy and monotonicity.

The Spearman correlation coefficient, r_s, is an established measure of monotonicity and we have used it here. r_s is only sensitive to the rank order for the variable values[6]. We calculate the correlation between u_{kp} and the actual uplift u_{ka}, *without* considering the number of records in each segment.

The obvious prediction error quantification for an uplift graph is the weighed sum of error, $\varepsilon := \frac{1}{N^t+N^c}\sum_{k=1}^{K}(n_k^t + n_k^c)|u_{kp} - u_{ka}|$. We used the sum $n_k^t + n_k^c$, instead of only n_k^t, as to reflect the sizes of both control and treated set when accounting for errors. Dividing by the total number of records in the sample gives us the average point wise error in the uplift model.

Another option for measuring accuracy uses the slope (β) of the linear regression between the prediction and the actual (weighed by segment size $n_k^t + n_k^c$). A tight model has a slope close to 1.0, therefore $|\beta - 1|$ gives a metric for the error. Unfortunately, one could also have a slope close to one, or exactly one and still have huge inaccuracies in the predicted values.

In addition, we considered that detecting a negative effect should also be taken into account. To measure negative effect we could have simple counted the number of segments indicating a negative effect, but that would not be applicable to datasets where the most common situation is a negative effect, instead we chose to use $\nu_p := |\sum_{k=1}^{K} \text{sign}(u_{kp})|$. The problem with ν is that it only takes values in the set $\{K, K-2, K-4, \cdots, 0\}$, thus it will present low variability between models, and may not add much information for the selection of a good model. Moreover, it only considers the prediction values.

The dataset simulations required, the construction of uplift models and further analysis presented in this report was conducted using the open source package R [22].

[6]Interestingly Gini coefficient is also a measure or monotonicity[29], and Qini coefficient is the subtraction of two Gini[26]

2.1 Understanding Moment of Uplift

2.1.1 Segment Contributions to Υ

Υ (both Υ_L^{min} and Υ_Q^{min} see 1.16 and 1.14) is the sum of individual segment contributions for an uplift model. In this section we will refer to the individual segment contributions when talking about the overall metric Υ. To understand better each individual contribution we present 3D figures and contour plots for both Υ versions, see figure 2.1. Υ is negative when u'_p and u'_a have different sign (quadrants[7] II and IV). Both quadrants I and III present the same shape for Υ. Each quadrant has two valleys where $\Upsilon = 0$ for $u'_p = 2u'_a$ and $u'_p = 1/2u'_a$. Υ value increases when moving towards line $u'_p = u'_a$, reflecting on the fact that our predictions get more precise, hence a higher contribution to Υ is done, but it also increases when moving towards lines $u'_p = 0$ and $u'_a = 0$, which does not have any obvious justification. This unexpected positive contribution happens when either actual or predicted uplift is close to the overall uplift, and the other takes an arbitrary value (with the same sign), see figure 2.2 for an illustration. When working with the initial models, and reviewing the uplift by decile graphs for the different datasets, it seemed better to change the formulation of Υ to the *mean* instead of the *min*. The formulation using *mean* gives more importance to the value of uplift than to the error from the prediction ($\epsilon := |u_a - u_p|$). This can be seen if we express the contribution in equation 1.15 as $c_i = |\min(|u'_a|, |u'_p|) - \epsilon|$. Replacing by *mean* it becomes $c_i = |(|u'_a| + |u'_p|)/2 - \epsilon|$ or, equivalently, $c_i = |\min(|u'_a|, |u'_p|) - \epsilon/2|$. In other words, we half the error.

We exemplify the impact of using *min* or *mean* in figure 2.3. For the case of *min*, the contribution from the "R" bars would be higher than the "L" bars contribution, using the (mean) this is reversed, which seems more logical if we want to praise bold predictions.

That change from *min* to *mean* does not change the topology of Υ. You can see the 3D representation and contour plot in 2.4, top two 3D graph and contour plots. We still have the anomaly of positive contributions when $u'_p \to 0$ or $u'_a \to 0$, although now the values taken by Υ are smaller than using *min*. Using the *min* we have that $\Upsilon = 0$ for $u'_p = 3u'_a$ and $u'_p = 1/3u'_a$. That observation and the desire to find a simpler three dimensional shape for Υ inspired the alternative formulation presented in equation 2.1:

$$r_i^c = \begin{cases} (cu'_{ip} - u'_{ia})/2, & \text{if } u'_{ia} \geq -u'_{ip}, u'_{ia} \geq u'_{ip} \\ (cu'_{ia} - u'_{ip})/2, & \text{if } u'_{ia} \geq -u'_{ip}, u'_{ia} < u'_{ip} \\ (u'_{ia} - cu'_{ip})/2, & \text{if } u'_{ia} < -u'_{ip}, u'_{ia} < u'_{ip} \\ (u'_{ip} - cu'_{ia})/2, & \text{otherwise.} \end{cases} \tag{2.1}$$

[7]When describing contour plots, we refer to the four quadrants in the plane as I, II, III and IV. Starting from the quadrant $x > 0$ and $y > 0$ and numbering them anti-clockwise

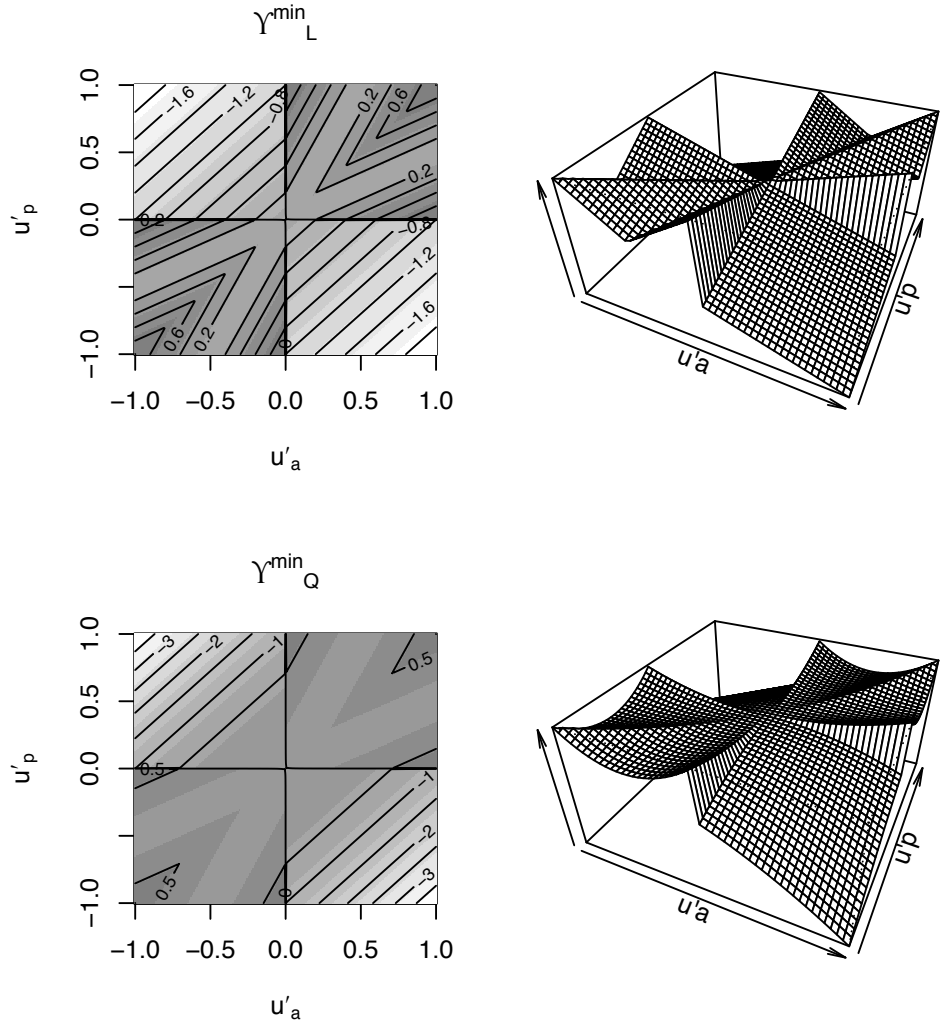

Figure 2.1: A contour plot on the left and the corresponding 3D representation on the right for both Υ_L^{min} and Υ_Q^{min} individual segment contribution. A grey scale has been used to indicate the values of Υ in the contour plot, the darker values are positive contribution, while clearer values are negative contribution. Contributions for Quadrants I and III are always positive

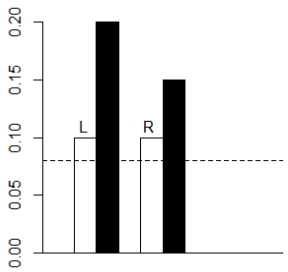

Figure 2.2: Once both u_p and u_p are above (or below) μ - dotted line - we have a positive contribution to Υ. However, if $u'_p > 2u'_a$ the contribution is even bigger. In the illustration bars "L" contribute more than bars "R".

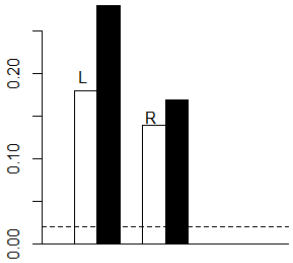

Figure 2.3: Both bar groups "R" and "L" predict a high uplift, in this situation we would like bar group "L" to contribute higher to Υ, than group "R", so we reward models that predict the highest uplift. We achieve that using the *mean*.

$$\Upsilon^s_{Lc} := \frac{1}{N^t}\sum_{i=1}^{K} n^t_i r^c_i \tag{2.2}$$

where c is a parameter. The formulation is a bit convoluted. It aims to eliminate the anomaly of having a positive contribution when $u'_p \to 0$ or $(u'_a \to 0)$ regardless of the value of u'_a (or u'_p). It also reduces the range of values for u'_a and u'_p with positive contribution, being the lines $u'_a = cu'_p$ and $u'_p = cu'_a$ the limits where the contribution starts to become negative. Those limits for positive contribution in quadrant I and III can be tuned by changing the parameter c. Higher values for c increases the surface in the plane (u'_a, u'_p) with positive contribution, which means that we sacrifice accuracy. The 3D representation for Υ^s_{Lc} for $c = 2$ and $c = 3$, can be seen in 2.4. It is close to a saddle shape, so it seems natural to propose the following quadratic version for Υ contributions: $r'_i = u'_p u'_a$. This is in fact the original form that motivated uplift (equation 1.11).

A modification on 1.16 and 1.14 definitions also correct the anomaly of having a positive contribution when $u'_p \to 0$ (or $u'_a \to 0$) regardless of the value of u'_a (or u'_p). Therefore we will redefine Υ^{min}_Q, Υ^{min}_L, Υ^{mean}_Q and Υ^{mean}_L as follows:

$$a_i = \begin{cases} \min(|u'_{ip}|, |u'_{ia}|) - |u'_{ip} - u'_{ia}|, & \text{if } u'_{ip}u'_{ia} > 0 \\ -|u'_{ip} - u'_{ia}|, & \text{otherwise.} \end{cases} \tag{2.3}$$

$$\Upsilon^{min}_Q = \frac{1}{N^t}\sum_{i=1}^{K} n^t_i a^2_i \text{sign}(a_i) \tag{2.4}$$

$$\Upsilon^{min}_L = \frac{1}{N^t}\sum_{i=1}^{K} n^t_i a_i \tag{2.5}$$

$$b_i = \begin{cases} |u'_{ip} + u'_{ia}|/2 - |u'_{ip} - u'_{ia}|, & \text{if } u'_{ip}u'_{ia} > 0 \\ -|u'_{ip} - u'_{ia}|, & \text{otherwise.} \end{cases} \tag{2.6}$$

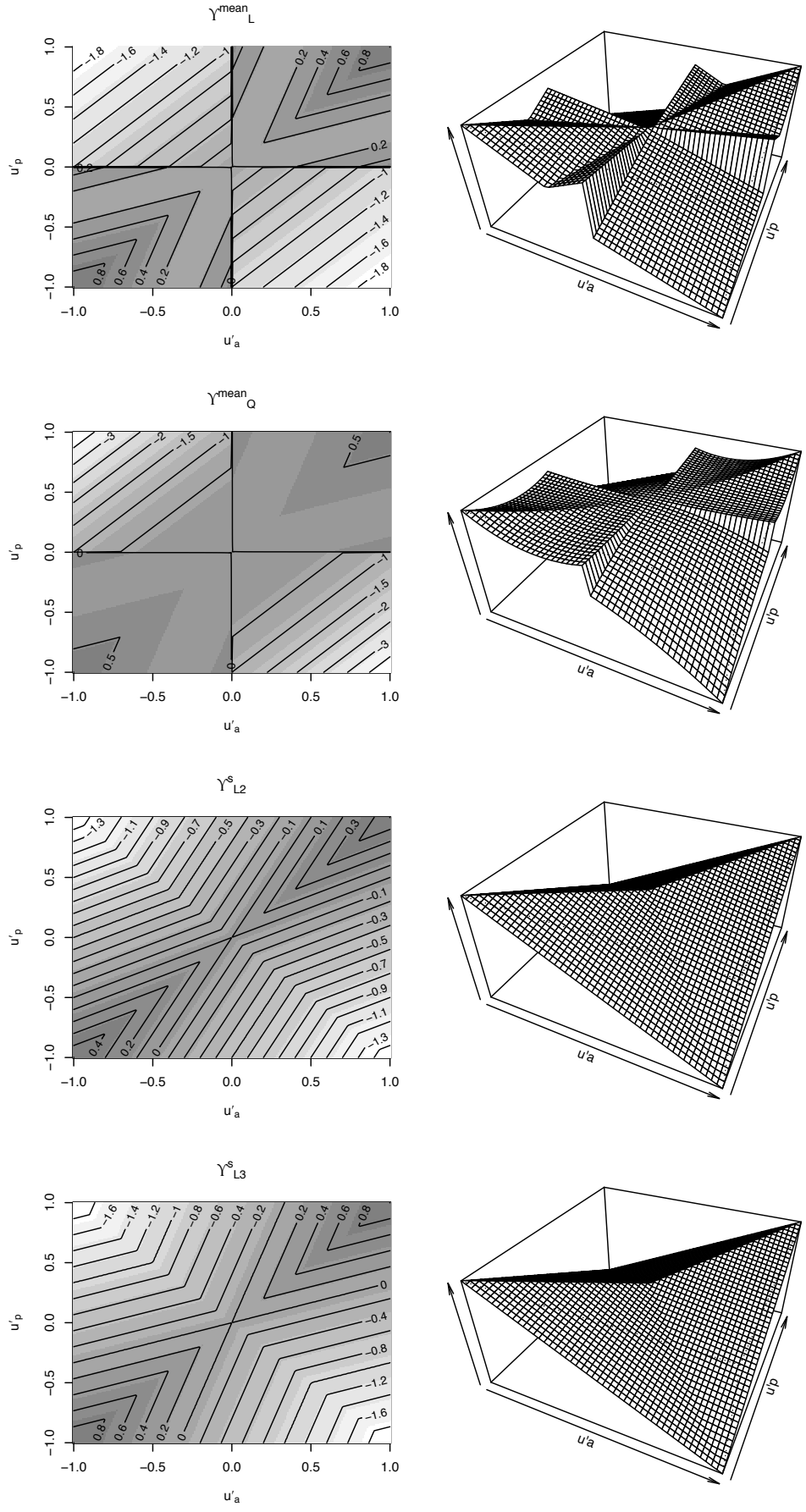

Figure 2.4: 3D representation and contour plot for different Υ versions

$$\Upsilon_Q^{mean} = \frac{1}{N^t} \sum_{i=1}^{K} n_i^t b_i^2 \text{sign}(b_i) \tag{2.7}$$

$$\Upsilon_L^{mean} = \frac{1}{N^t} \sum_{i=1}^{K} n_i^t b_i \tag{2.8}$$

Figure 2.5 illustrates the shape of the individual contributions for the new versions of Υ. The quadratic forms Υ_Q^{min} and Υ_Q^{mean} do not present the anomaly, but they have a flat surface in quadrants I and III. It is for this reason than a final quadratic form is proposed:

$$\Upsilon_{Qc,n}^s := \Upsilon_{Lc}^s \left((u_a')^2 + (u_p')^2 \right)^n \tag{2.9}$$

with n a parameter that could be adjusted if the quality measurement offers promising results. Reducing the value of n sacrifices prediction precision versus boldness; the smaller the value, the bolder the models. The starting value used was $n = 1/4$. Note that Υ_Q^{mean} and Υ_Q^{min} could have been defined using a parameter n instead of considering them quadratic.

To sum up, we will calculate the following Υ values for multiple uplift models and review their potential as quality measures:

Table 2.1: Different Υ versions that are assessed in this paper.

Symbol	Formulation
$\Upsilon_{Lc=2.5}^s$	2.2
Υ_Q^{min}	2.4
Υ_L^{min}	2.5
Υ_Q^{mean}	2.7
Υ_L^{mean}	2.8
$\Upsilon_{Qc=2.5,n=0.25}^s$	2.9

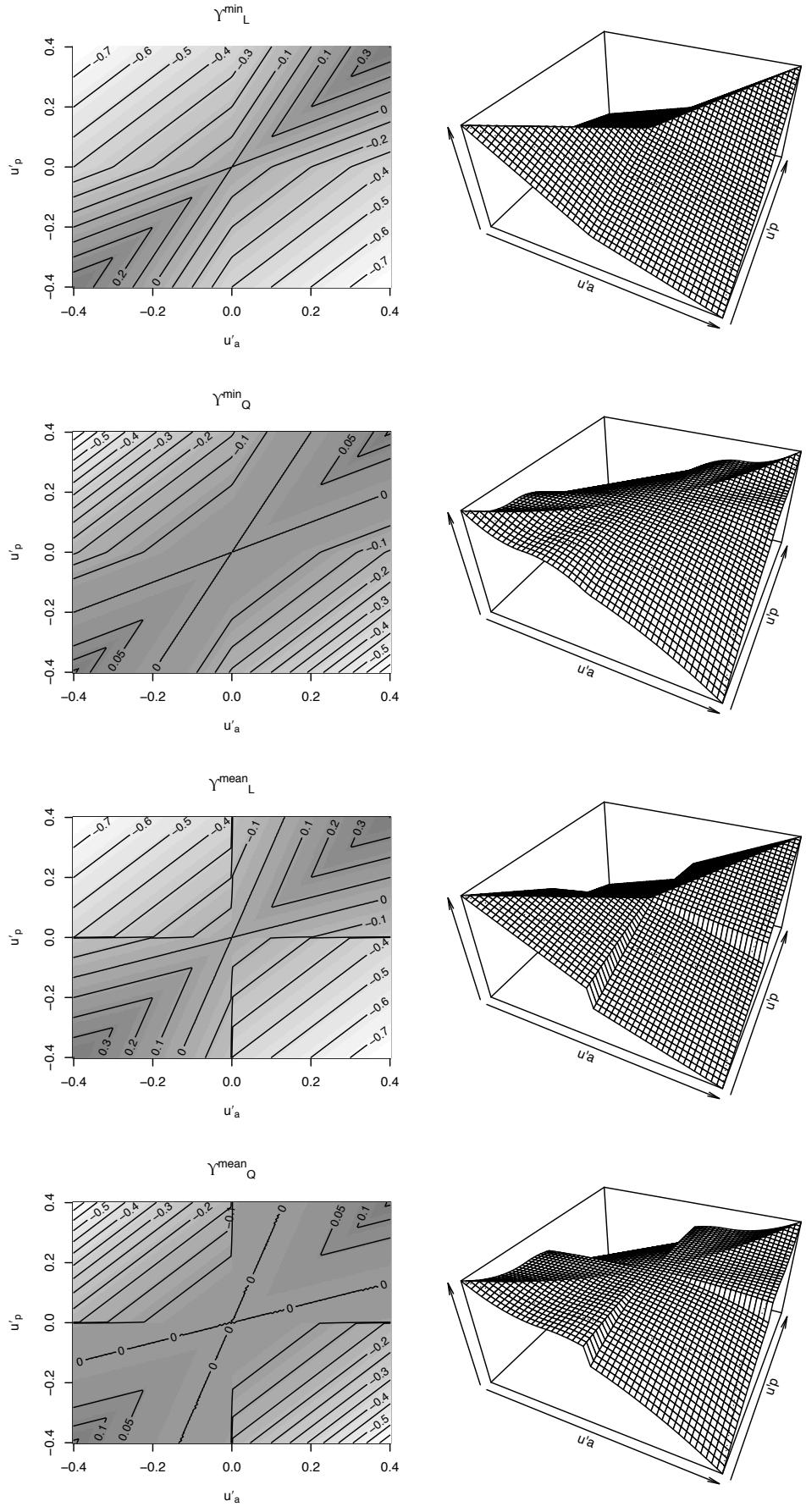

Figure 2.5: 3D representation and contour plot for different Υ versions

We will refer to $\Upsilon^s_{Lc}, \Upsilon^{min}_L, \Upsilon^{men}_L$ as Υ_L, and $\Upsilon^s_{Qc}, \Upsilon^{min}_Q, \Upsilon^{men}_Q$ as Υ_Q.

2.1.2 Sum of Contributions to Υ

It is difficult to make a prediction for the total value Υ as the sum of contributions, an increase or decrease in Υ can be caused by multiple segments increase and/or reductions in value. For a model with perfect accuracy and Υ^{min}_L version, the total contribution equals the $\sum_{k=1}^{K} |u'_{ka}|$. Therefore accurate and bolder models – one with predictions above or below the average uplift – have higher Υ^{min}_L. On the other hand, models mistakenly predicting that some segments are above or below the overall uplift will present negative or small Υ^{min}_L values. This general behaviour will also happen for any Υ version, as the shape of the individual contributions is similar.

As we explained in section 2.1, one could interpret Υ_L (and by extension all the other versions of Υ) as the total uplift achieved minus the error in the prediction. Given a dataset and a set of models it makes sense to compare the different Υ values. However such comparison should not be made between models created for different datasets, as those datasets may present a different "intrinsic" uplift or variability, which influences the minimum error achievable. In other words, one can compare the error made by models in different datasets and can also compare the uplift between two different datasets, but when combining error and uplift in a single measure, such as Υ, the comparison is difficult to interpret.

Υ values will be generally very small. This is not ideal for a quality measure but no normalization value has been found to present it in a more readable format, although the number of segments used for the model is an upper limit for Υ.

There may be some value in evaluating Υ for different segmentation schemes. Assuming that we start from an accurate model, if we increase the number of segments it is unlikely that the new segmentation will still match the actual uplift. Therefore we could expect a decrease in the value of Υ, as the error in each segment contribution will be higher than before. If that does not happen, it may be an indication of a good model or more accurate segmentation. We have not investigated that option because tree-based uplift models somehow optimize the segmentation and there is no need to find a better segmentation for them.

2.2 Composite Measures

Considering that we have intuitive measures for accuracy and boldness, as introduced at the beginning of this chapter, one could build composite measures for that combine those characteristics. For instance:

Definition 2.1 $M_1 := \frac{U_{max}}{1+\varepsilon}$

Which increases if the initial deciles present higher uplift, provided that the error stays the same, therefore bolder models may present higher values of M_1. It also increases if error decreases, thus M_1 could identify bold and accurate models. We could also consider monotonicity, using Spearman coefficient r_s

Definition 2.2 $M_2 := r_s M_1 = r_s \frac{U_{max}}{1+\varepsilon}$

If U_{max} and ε are similar for two different models, M_2 rewards the one with higher monotonicity, that is higher r_s values.

A third version includes spread, s_p, and the predicted uplift at 1^{st} decile, u_{1p}.

Definition 2.3 $M_3 := \frac{s_p}{u_{1p}} M_2 = \frac{r_s s_p}{u_{1p}} \frac{U_{max}}{1+\varepsilon}$

Note that $s_p/u_{1p} = 1 - u_{Kp}/u_{1p}$, therefore the smaller the ratio u_{Kp}/u_{1p}, indicating a higher first decile or a smaller last decile, the higher the value for M_3.

Finally, as Qini measures are more oriented towards boldness than accuracy, we could define a modified Qini that penalizes for inaccuracy:

Definition 2.4 $M_4 := \frac{q_0}{\varepsilon+1}$

and

Definition 2.5 $M_{5,\tau} := q_0 \exp(-\tau(\beta - 1)^2)$

being τ a parameter and β the correlation coefficient for the linear regression u_{kp}, u_{ka}. Modifying τ one gives more or less importance to accuracy over boldness, we used $\tau = 5$ as a starting point. Any other bell-shaped curve could be used, for instance Cauchy distribution, but we prefer the Gaussian curve as it has a shorter tail.

We will refer to M_1, M_2, M_3, M_4 and M_5 as the M family.

2.3 Datasets Used for the Models

To assess the usefulness of Υ as a quality measure we opted for generating multiple models on various datasets, ideally real-world datasets. We had at our disposal one real-world dataset made available on the web [14]. We refer to it as the Hillstrom dataset.

The Hillstrom dataset contains $64,000$ records each describing one customer. The data is divided in three disjoint sets, one set for three different marketing strategies:

- One third of the customers receive no e-mail and acts as the control group for the other two sets;
- Another third were sent an e-mail with one promotion type (we will refer to this set plus the control group as the Hillstrom ME dataset);

- And the last third received another e-mail promotion (Hillstrom WE dataset);

The outcome of the campaign was measured within a two-week period, and it was described from three different perspectives, but we focused only on two of the outcomes, as the third one was a continuous variable:

- Visit, a binary variable indicating whether the customer visited the site;
- Conversion, a binary variable indicating whether the customer bought anything.

Using the two independent e-mail campaigns and the two outcomes means that we had four different uplift problems to work with.

Unfortunately only one dataset was not enough to guarantee that we would have enough models with different characteristics, therefore we decided to generate simulated datasets. We did that by using the Hillstrom dataset as a starting point, changing only the outcome variable based on pre-defined customer characteristics. We effectively created a hidden segmentation. We used a Bernoulli trial simulation to decide on the value of the outcome. In most cases we increased the success rate in the treated group, but we also worked with segments were the treatment caused a negative effect, which resulted in lower success rate in the treated segment than in the equivalent control segment. In total thirteen artificial datasets were created, five using the $64,000$ records and eight using $42,000$ records (either the ME or the WE subsets).

We tried to obtain more real-world datasets searching the Internet. Sources such as the UCI machine learning data repository [10] and official statistical sites where consulted, but unfortunately it was not possible to find a suitable one having the control and treated subsets with enough records[8]. This is a key requirement to be able to apply uplift models. Nevertheless, we reused one of the UCI datasets (the "adult" dataset, which contains $48,842$ records) to generate more synthetic data that could be used for an uplift model. This required to create two extra information fields, one for a binary outcome and the other to identify if the record was or not treated. The separation in treated and control was done randomly using a 50% split. The outcome value was generated using a Bernoulli trial simulation.

In the end we had eighteen datasets available, which are listed in table 2.2.

2.4 Strategy for Generating the Models

We used two modelling techniques to generate different models in our datasets:

- Two Model approach;

[8] Reports from practitioners [33] suggest that and uplift model may not be applicable for datasets with less than 100000 records.

Source dataset	Alias for derived dataset	Records	Description
Hillstrom	dataset0ME - visit	42,000	Hillstrom dataset for ME campaign using visit as outcome
	dataset0WE - visit	42,000	Hillstrom dataset for WE campaign using visit as outcome
	dataset0ME - conversion	42,000	Hillstrom dataset for ME campaign using conversion as outcome
	dataset0WE - conversion	42,000	Hillstrom dataset for WE campaign using conversion as outcome
	dataset1ME	42,000	Simulated outcome using ME campaign predictors
	dataset2ME	42,000	Simulated outcome using ME campaign predictors
	dataset3ME	42,000	Simulated outcome using ME campaign predictors
	dataset4ME	42,000	Simulated outcome using ME campaign predictors
	dataset1WE	42,000	Simulated outcome using WE campaign predictors
	dataset2WE	42,000	Simulated outcome using WE campaign predictors
	dataset3WE	42,000	Simulated outcome using WE campaign predictors
	dataset4WE	42,000	Simulated outcome using WE campaign predictors
	hillstrom1	64,000	Simulated outcome and treated groups on the overall Hillstrom dataset
	hillstrom2	64,000	Simulated outcome and treated groups on the overall Hillstrom dataset
	hillstrom3	64,000	Simulated outcome and treated groups on the overall Hillstrom dataset
	hillstrom4	64,000	Simulated outcome and treated groups on the overall Hillstrom dataset
	hillstrom5	64,000	Simulated outcome and treated groups on the overall Hillstrom dataset
UCI repository	adult	48,842	Simulated outcome and treated groups on the adult dataset from UCI repository

Table 2.2: Total datasets used. Each dataset is separated in test and training set.

- The VLO approach;

As those are regression based methods, we could use three different regression types (linear, logistic and logistic with over dispersion using Poisson), giving a total of six possible modelling approaches, although we expected small differences between them.

For each dataset, we generated multiple models using combinations of the variables available for each record as predictors, and applying the six model types above. For instance, a dataset with seven variables allows generating 2^7 variable combinations and a total of $6 \times 27 = 512$ models could be created [9], but at least four variables were requested to build the models. From the point wise uplift $u(x)$ returned by the regression we obtained $b_k, u_{kp}, N_k^t, R_k^t, N_k^c, R_k^c$ and $u_{ka} = {}^{r_k^t}/_{n_k^t} - {}^{r_k^c}/_{n_k^c}$ as explained in sections 1.4 and 1.6, for $k = 1, \cdots, 10$.

Using above values we calculated the quality measures proposed for each model: Υ and M in its multiple versions, q_0 and ROI. We also collected the metrics that reflect the desirable characteristics: $\varepsilon, r_s, U_{max}, u_{1p}, s_p, \nu_p$.

When using regression methods, segmenting by deciles may increase the prediction error ε. Instead of working with deciles, one could find the optimal segmentation minimizing ε, q_0 or

[9] Obviously this is not a practical approach in a commercial situation, where it is common that hundreds of variables are available for each record [19]

U_{max}, but we did not proceed that way. In principle, this is not a problem for our work, as long as our models make valid predictions. Tree based methods already define an optimal dataset segmentation when building the model, although the segmentation becomes fine-grained when performing "bagging"

We used a common technique for machine learning modelling, which is separating the dataset in training and test (or hold out) subset. Both the training and the test are made up of two disjoint control and treatment groups. The training subset is used for building the model and the test subset for testing it. This strategy is also relevant for assessing quality measures that balance both accuracy and boldness, because the predictions on the test subset are less precise. Moreover, that would be the way one would choose a model using quality measures, and we are reproducing it in our assessment of the quality measurements.

The separation in training and test subset was done randomly and all the different models were built using the same training subset. We used 50% of the records for the training subset and 50% for the test subset. We repeated the separation and produced again the models multiple times, to have some confidence that the observations we made were not an artefact of the separation. This is similar to the k fold cross-validation technique used for machine learning, whose best known variation is the 5×2 cross validation [9], and aims to provide a more robust result. We applied the steps mentioned in sections 1.4 and 1.6 to obtain the predicted and the actual uplift for the test subset, and used those values to calculate the quality measures.

3 Results and Discussion

If one is confident that increasing values of a metric can capture the usefulness of a model, one would aim to build models that maximize it. We have argued in section 2.1.2 that inaccurate models have lower Υ values, and that accurate and bold models have a higher Υ value than inaccurate or conservative models. In this section, we assume that increasing values of Υ generally improve both accuracy and boldness of a model and we test this hypothesis.

We generated around two hundred different models for each dataset described in table 2.2 and collected the multiple measurements that reflect accuracy and boldness: $\varepsilon, r_s, s_p, U_{max}, u_{1p}, u_{1a}$ and ν_p (see section 1.5). In addition, we calculated q_0, the various Υ versions from table 2.1.1, M versions from section 2.2, ROI^{act} and ROI^{pred} (using $p = 1$ unit and $c = 0.05$) as explained in section 1.8. We sorted the models in descending order by each of the quality measures proposed, and inspected the uplift by deciles graph to identify relevant characteristics for models with increasing values of the quality measure. We review the models in the next sections, while we discuss them, we have taken the license of referring to the model selected by maximizing the quality measure Z, as the Z model.

We start presenting the uplift by deciles graph for the real world data, Hillstrom dataset. Each row of figures contains three uplift graphs (label max for the highest value obtained; mid for a random value in the 3^{rd} quintile and; and min for the lowest value) for different values of the quality measure:

There are two numbers on the graph, one refers to the value the quality measure takes on the model, and the other is an identifier of the model[10], which can be used to confirm if it two plots are the same model. The sequence of graphs in each row illustrates how the quality measure behaves when it increases in value. The graphs include two thin dotted lines, marking the value of μ_p and μ_a. If the two lines overlap or they are very close, this indicates that the training-test separation evenly distributed the uplift effect. Otherwise, it is likely that the models generated by the training set do not work well in the test set.

3.1 Hillstrom ME Dataset - Visit Rate

3.1.1 Training Subset

We start with the models constructed using the training set (see figure 3.1 to 3.4). The first thing to notice is that increasing values of all quality measures seem to produce models that are in general either more accurate or bolder.

The (single) model selected by all Υ_L versions and M_2, (figures 3.1 and 3.5) makes in general accurate predictions and it distinguishes the different deciles, although it presents the smallest

[10]The identifier refers to the variables used for the model, but it does not specify what model type was used.

spread (s_p), see table 3.1

Similarly all Υ_Q versions in figure 3.2 select the same model, although the prediction error is

	U_{max}	$-\varepsilon$	r_s	u_{1p}	u_{1a}	$-\nu_p$	s_p
Υ_L^{min}	0.124	-0.006	0.939	0.124	0.115	-10	0.06
Υ_L^{mean}	0.124	-0.006	0.939	0.124	0.115	-10	0.06
Υ_L^{s}	0.124	-0.006	0.939	0.124	0.115	-10	0.06
Υ_Q^{min}	0.124	-0.008	0.745	0.124	0.122	-10	0.069
Υ_Q^{mean}	0.128	-0.007	0.782	0.128	0.121	-10	0.068
Υ_Q^{s}	0.128	-0.007	0.782	0.128	0.121	-10	0.068
q_0	0.109	-0.007	0.857	0.109	0.104	-10	0.104
ROI^{pred}	0.124	-0.008	0.745	0.124	0.122	-10	0.069
ROI^{act}	0.134	-0.009	0.842	0.134	0.118	-10	0.062
M_1	0.134	-0.009	0.842	0.134	0.118	-10	0.062
M_2	0.124	-0.006	0.939	0.124	0.115	-10	0.06
M_3	0.107	-0.007	0.888	0.107	0.102	-10	0.102
M_4	0.109	-0.007	0.857	0.109	0.104	-10	0.104
M_5	0.109	-0.006	0.857	0.109	0.102	-10	0.102

Table 3.1: Hillstrom ME dataset; visit rate; training subset. Values for maximum uplift, prediction error, monotonicity, predicted and actual uplift at 1^{st} decile, negative effect and spread for "max" models in figures 3.1 to 3.5. Υ_L, M_2 and M_5 make the most accurate predictions. The Υ_L family makes the most monotone predictions (highest r_s) and with one of the highest actual uplift at first decile (u_{1a}) and maximum uplift U_{max}. q_0 presents the highest spread (s_p).

slightly higher, but it successfully marks segments above and below the overall uplift. Lower values of Υ_Q are indeed less accurate or more conservative.

The q_0 model is acceptable, figure 3.3, although less bold than the ones selected by Υ.

There are little differences between M_3, M_4 and M_5 models, the later is slightly more precise and with higher U_{max}. M_1 selects a bold model with a high first uplift, although the prediction for the fifth decile is completely wrong. The model selected by Υ_Q^{min} makes fairly accurate predictions, although it predicts a smaller first uplift than M_1.

The model selected by ROI^{act}, see 3.4, is the same as the one selected by M_1 and the one selected by ROI^{pred} is the same as the one selected by Υ_Q^{s}.

There are some variations in the models selected by M family, in figure 3.5. M_1 predicts a high first decile uplift and M_3 makes very accurate predictions for the first five segments, although it is slightly more conservative (lowest U_{max}) than the rest of the models so far.

To sum up, all quality measures have chosen acceptable models, with similar overall shape but different accuracy and subtle differences in terms of boldness. Υ_Q family makes the best selection, presenting the highest U_{max} and good prediction error. The models with lower values (columns mid and min in each of the figures) are less accurate, non-monotone and presenting a smaller range of predictions.

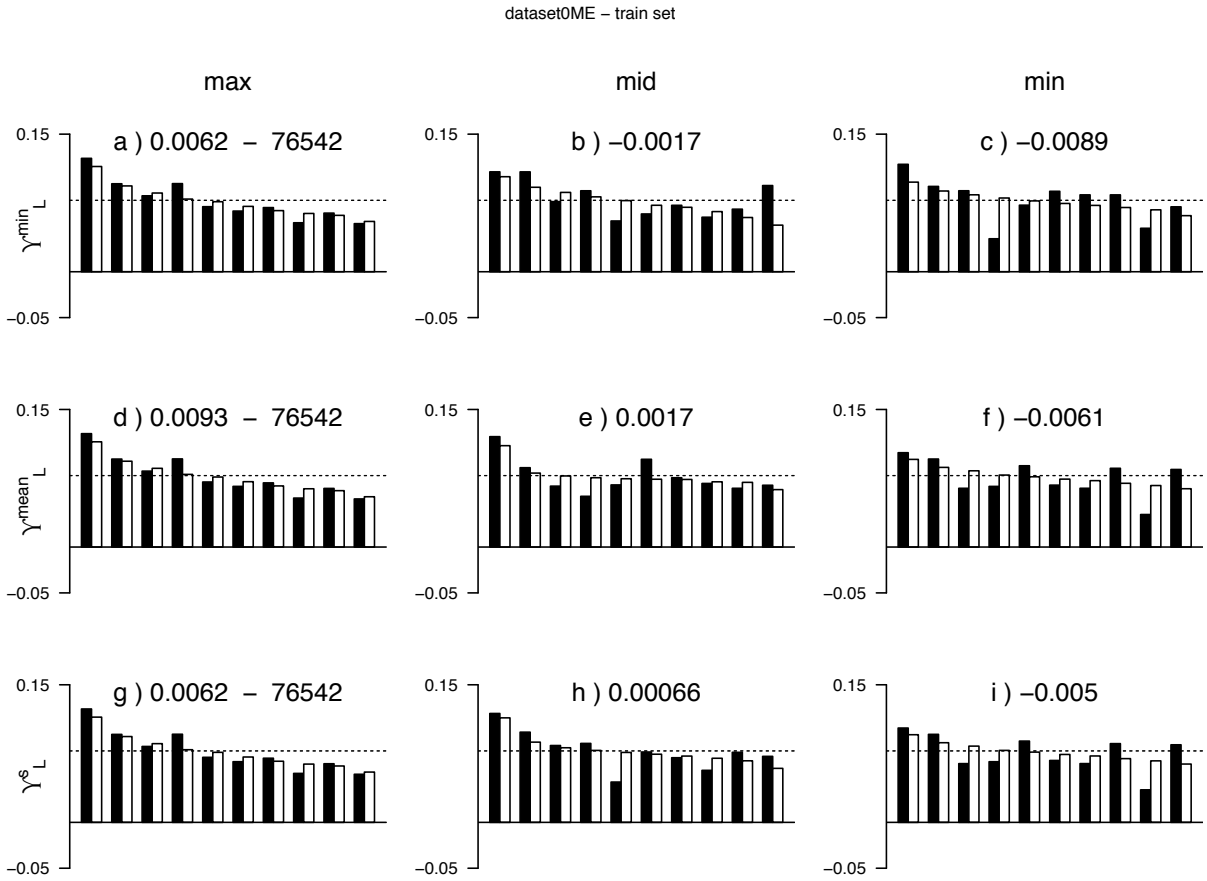

Figure 3.1: All Υ_L select the same best model. The model is accurate, except for the fourth decile, and bold.

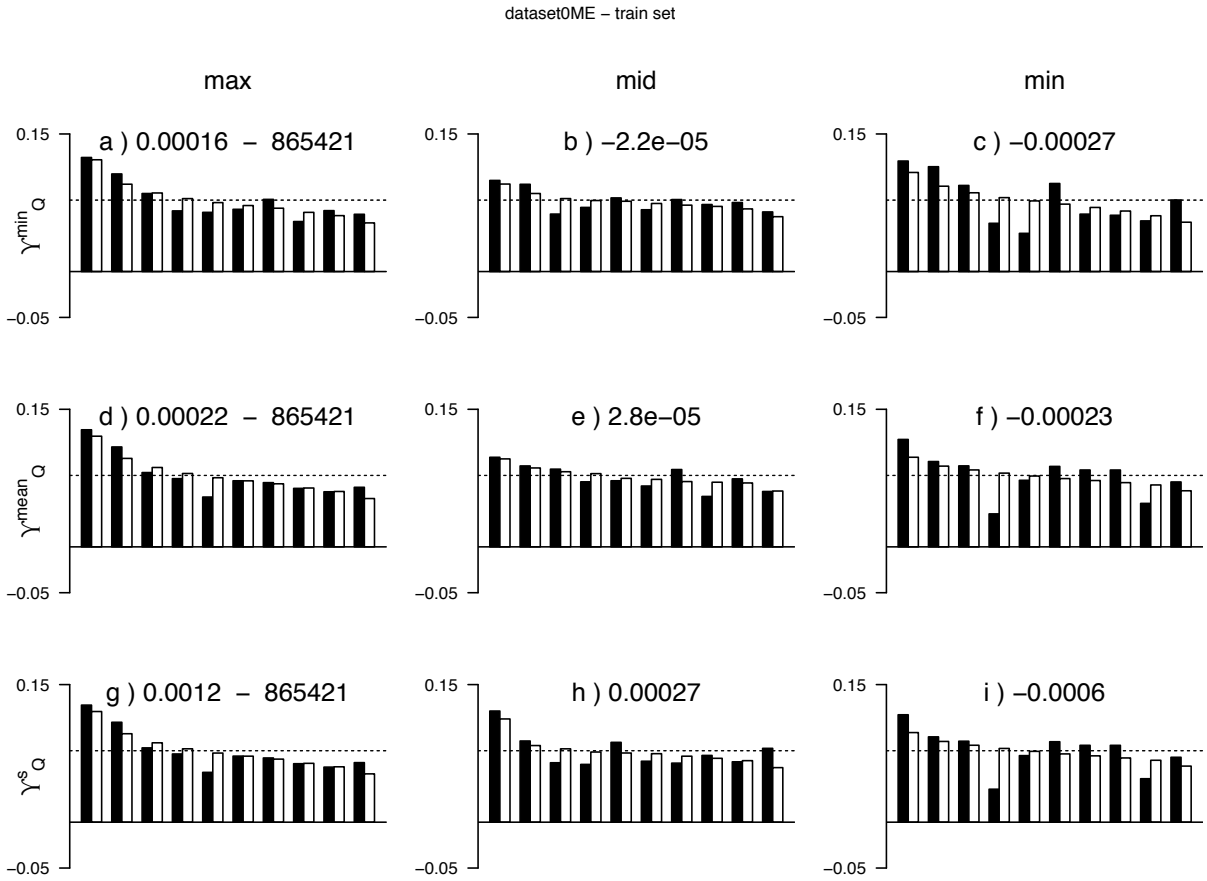

Figure 3.2: All Υ_Q family selects accurate and bold models. Increasing values of Υ_Q indicate either a bolder or more accurate model.Υ^{mean}_{Q} and Υ^{s}_{Q} select the best model, with high U_{max} and low ε (see table 3.1)

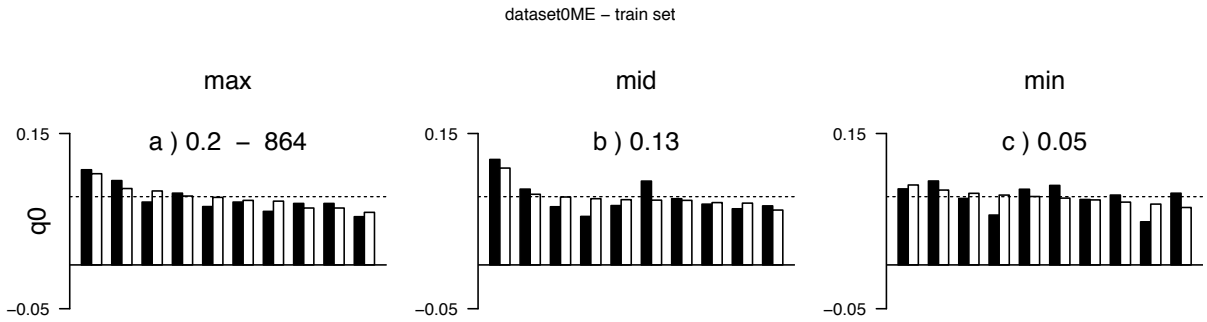

Figure 3.3: The model selected by q_0 seems to take into account the accuracy of the predictions. In this case it predicts a smaller first and second decile than the Υ family.

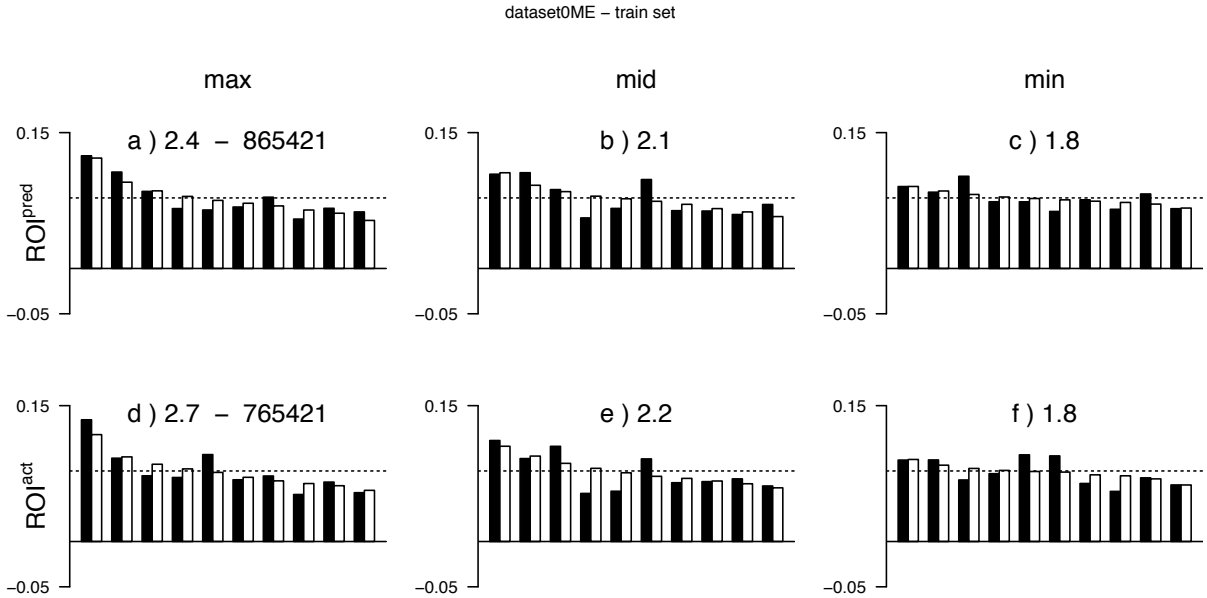

Figure 3.4: Although ROI does not account for accuracy, the predictions are accurate for ROI^{pred}. Note that the uplift detected by the models is not much higher than the one identified by Υ.

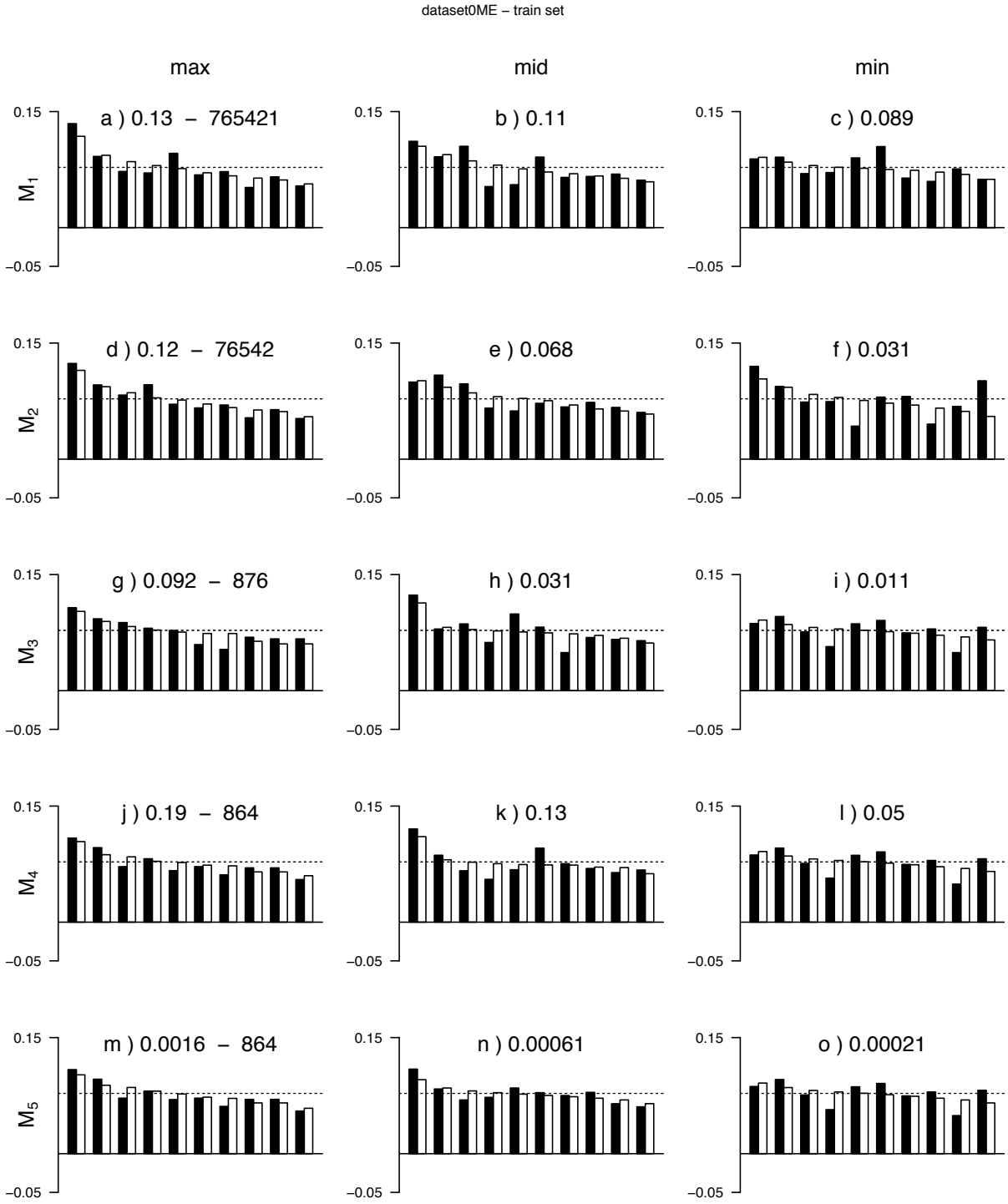

Figure 3.5: M metrics select four different models, being M_3 arguably the best, as it is very accurate on the first five deciles, although M_2 has the lowest ε (table 3.1)

3.1.2 Test Subset

The test subset is more challenging because the models were built on the training subset, therefore they may not accurately predict the behaviour of the test subset.

	U_{max}	$-\varepsilon$	r_s	u_{1p}	u_{1a}	$-\nu_p$	s_p
Υ_L^{min}	0.124	-0.01	0.503	0.124	0.117	-10	0.06
Υ_L^{mean}	0.117	-0.009	0.648	0.117	0.118	-10	0.062
Υ_L^{s}	0.117	-0.008	0.576	0.117	0.116	-10	0.061
Υ_Q^{min}	0.117	-0.009	0.648	0.117	0.118	-10	0.062
Υ_Q^{mean}	0.117	-0.009	0.648	0.117	0.118	-10	0.062
Υ_Q^{s}	0.117	-0.009	0.648	0.117	0.118	-10	0.062
q_0	0.155	-0.015	0.204	0.155	0.113	-10	0.113
ROI^{pred}	0.098	-0.016	0.406	0.098	0.123	-10	0.07
ROI^{act}	0.16	-0.015	0.406	0.16	0.114	-10	0.051
M_1	0.16	-0.014	0.418	0.16	0.114	-10	0.051
M_2	0.144	-0.011	0.6	0.144	0.114	-10	0.053
M_3	0.136	-0.015	0.511	0.136	0.109	-10	0.109
M_4	0.155	-0.015	0.204	0.155	0.113	-10	0.113
M_5	0.118	-0.011	0.492	0.118	0.109	-10	0.109

Table 3.2: Hillstrom ME dataset; visit rate; test subset. Values for maximum uplift, prediction error, monotonicity, predicted and actual uplift at 1^{st} decile, negative effect and spread for "max" models in figures 3.6 to 3.10. The Υ family of metrics reports the smallest ε, but also the smallest U_{max}, except for Υ_L^{min}. M_2 reports slightly higher ε but considerable higher U_{max} than the Υ family.

In this case (see figures 3.6 to 3.10), one clearly sees that "mid" and "min" models are more inaccurate, or make predictions close to the overall uplift than the model on the left. The Υ family selects similar models and all the Υ_Q metrics choose the same model, which is visible different from the selection done by M measures, q_0 and ROI.

M measures, q_0 and ROI models present a higher first decile predicted uplift than Υ measures do, but they are less accurate than the Υ_Q model. Υ family are the models with lowest prediction error (ε) and scoring highest in monotonicity, Υ_L^s is actually the best on that respect.

There is no best model amongst the M family (figure 3.10), but M_5 is the worst out of the five, with a smaller first decile uplift than the rest, but without gaining in accuracy. ROI^{pred} (figure 3.9) performs badly in this case. q_0 selects a model with a high U_{max}, but it has one of the highest ε. M_1 performs better than q_0 in all aspects (see table 3.2) except spread.

In summary, for this dataset, increasing values of the quality measures clearly indicate a better model, except for ROI^{pred} and M_5, which do not select good models.

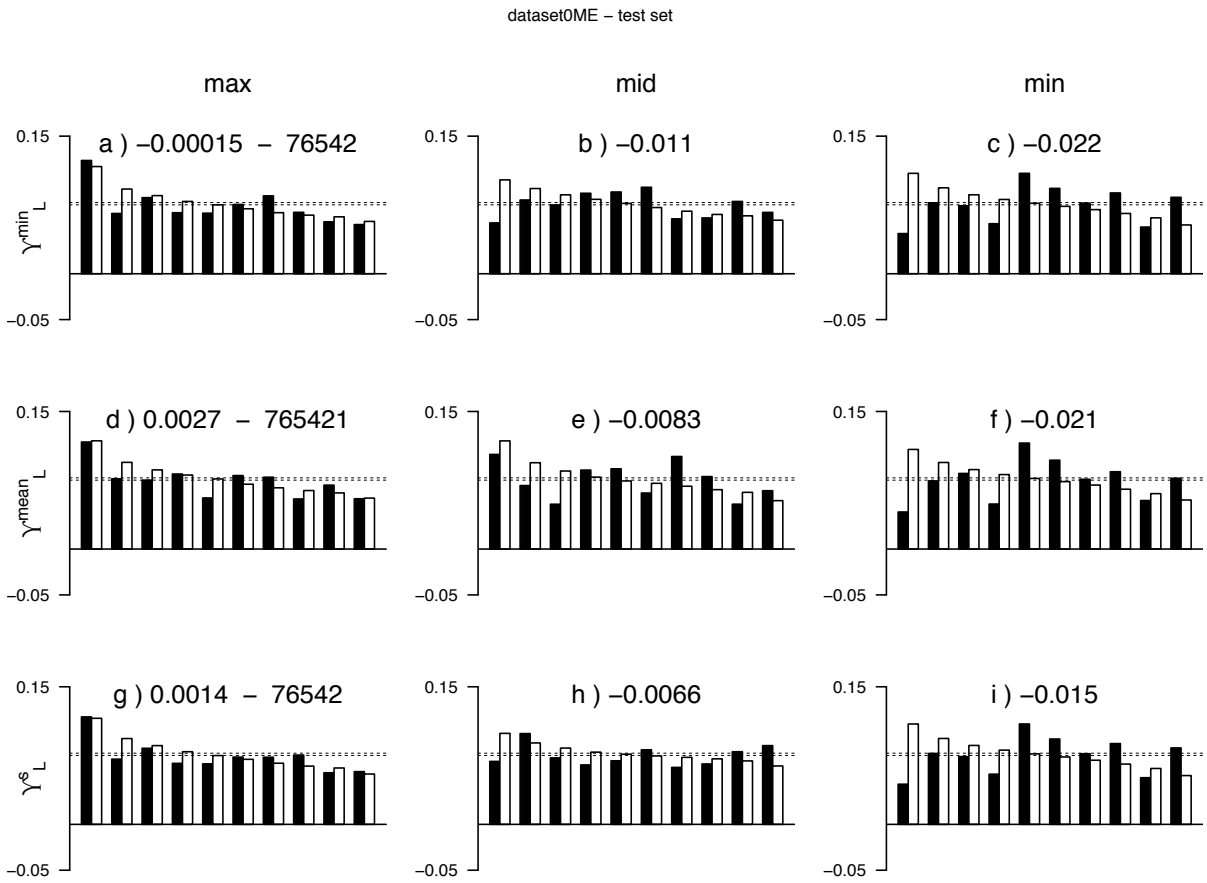

Figure 3.6: Generally speaking models are usually less precise in the test set, but Υ still selects models with acceptable predictions.

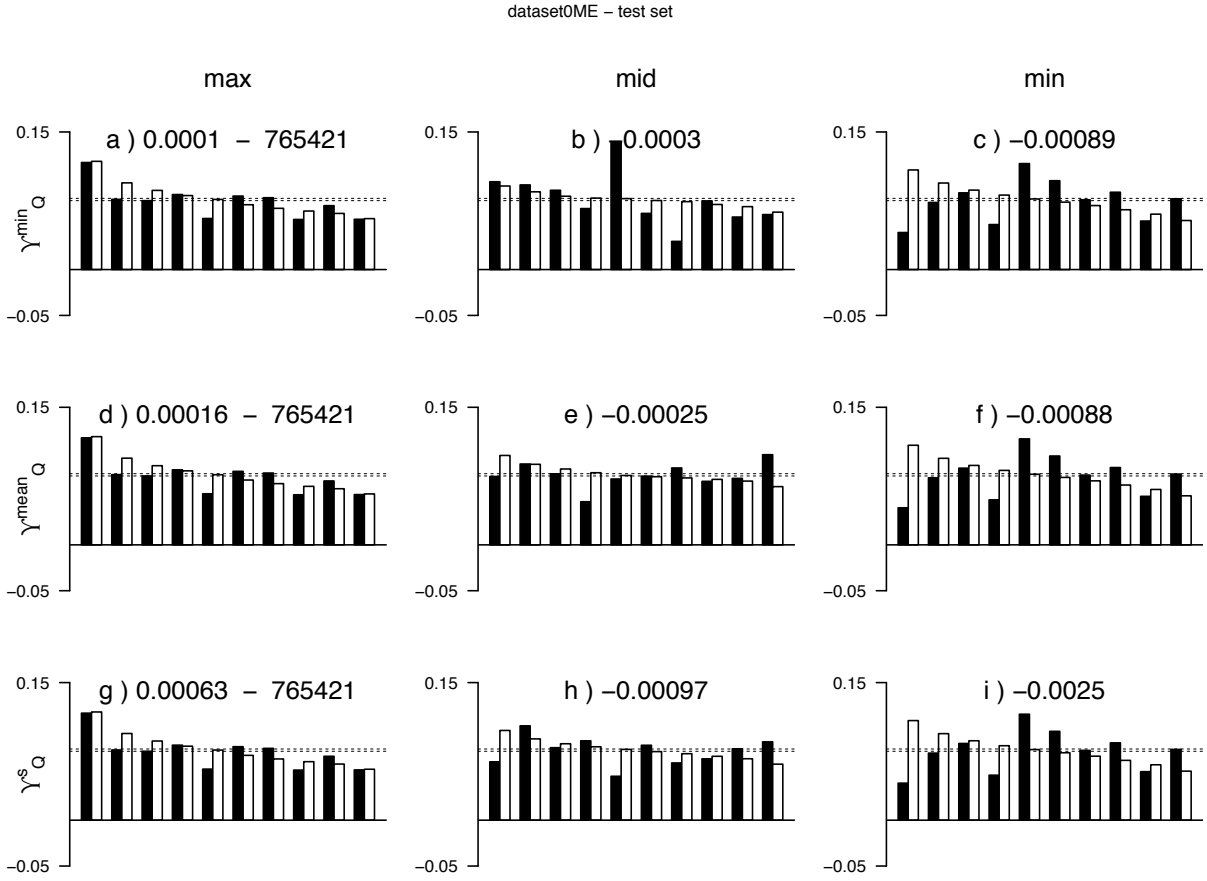

Figure 3.7: An accurate and bold single model is selected by all Υ_Q. Plots on the right correspond to lower values of the quality measure, and they are clearly unacceptable models.

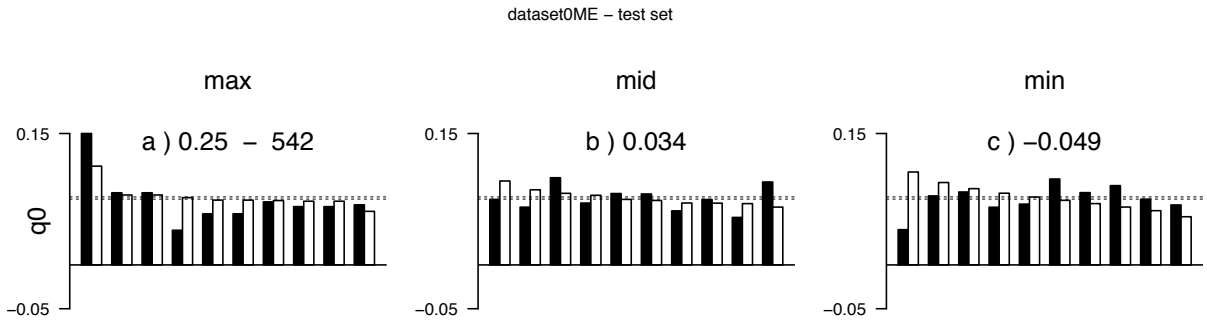

Figure 3.8: q_0 fails to select an accurate model, with high ε and low monotonicity. Both M and Υ measures outperform q_0

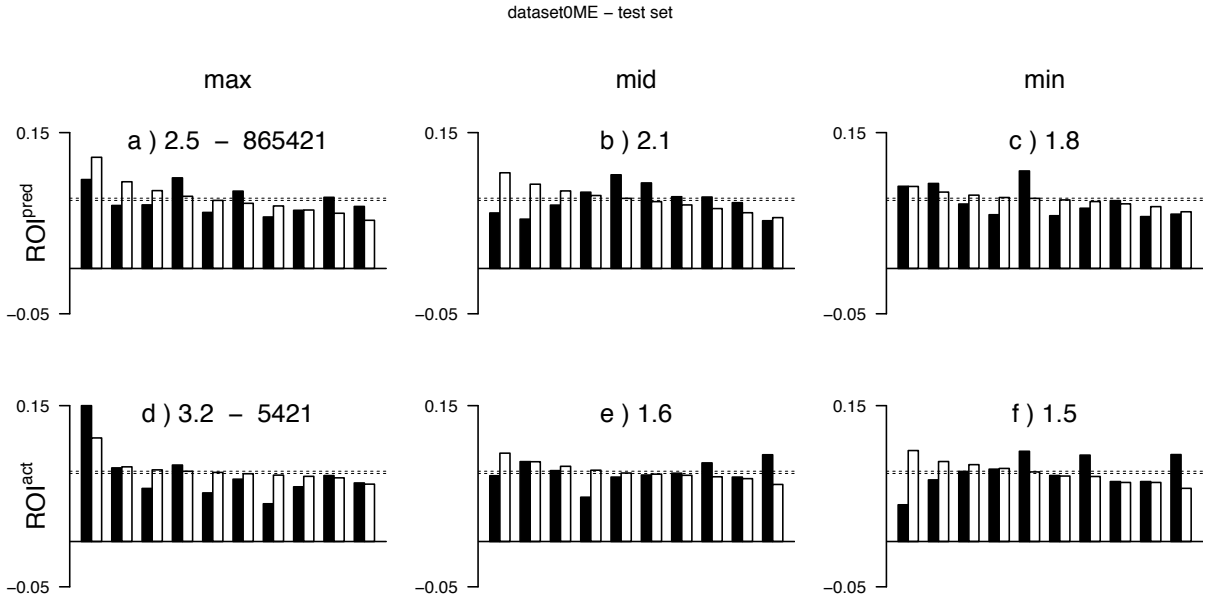

Figure 3.9: ROI metrics also fail to make a valid model choice in terms of overall accuracy. However ROI^{act} selects a model that accurately differentiates the first two segments.

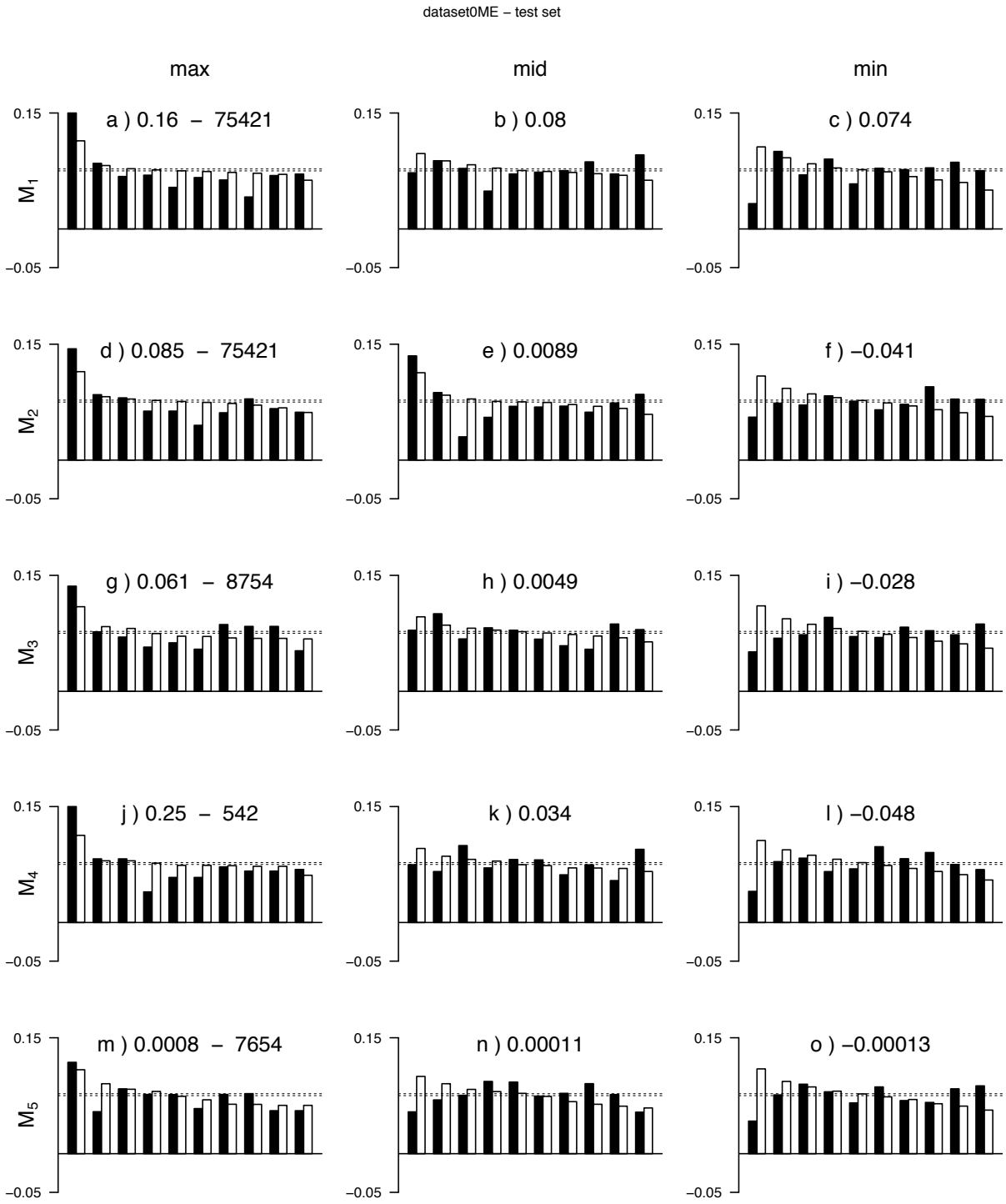

Figure 3.10: In this case, M measures give more weight to uplift at first decile than accuracy. Models with lower value for the quality measure are more conservative, with most segments close to the overall uplift.

3.2 Hillstrom WE Dataset - Visit Rate

3.2.1 Training Subset

All the metrics from the Υ family select models that make very accurate and bold predictions (see 3.11 and 3.12). M_1 in figure 3.15 selects what initially could be the most useful model from all shown (figures from 3.11 to 3.14), with a very high first uplift and fairly accurate. However, M_1 fails to identify a negative effect in the last decile, while Υ family does it. We see here an example where models with a high first uplift is present but Υ does not select it, sacrificing boldness in favour of accuracy.

	U_{max}	$-\varepsilon$	r_s	u_{1p}	u_{1a}	$-\nu_p$	s_p
Υ_L^{min}	0.084	-0.004	0.939	0.084	0.087	-8	0.09
Υ_L^{mean}	0.084	-0.004	0.939	0.084	0.087	-8	0.09
Υ_L^{s}	0.084	-0.004	0.939	0.084	0.087	-8	0.09
Υ_Q^{min}	0.091	-0.005	0.952	0.091	0.088	-8	0.094
Υ_Q^{mean}	0.091	-0.005	0.952	0.091	0.088	-8	0.094
Υ_Q^{s}	0.084	-0.004	0.939	0.084	0.087	-8	0.09
q_0	0.093	-0.005	0.679	0.093	0.087	-6	0.095
ROI^{pred}	0.136	-0.004	0.687	0.136	0.099	-6	0.105
ROI^{act}	0.136	-0.004	0.687	0.136	0.099	-6	0.105
M_1	0.136	-0.004	0.687	0.136	0.099	-6	0.105
M_2	0.099	-0.009	0.952	0.099	0.094	-8	0.108
M_3	0.093	-0.011	0.952	0.076	0.091	-8	0.107
M_4	0.093	-0.005	0.679	0.093	0.087	-6	0.095
M_5	0.087	-0.004	0.903	0.087	0.085	-10	0.079

Table 3.3: Hillstrom WE dataset; visit rate; train subset. Values for maximum uplift, prediction error, monotonicity, predicted and actual uplift at 1^{st} decile, negative effect and spread for "max" models in figures 3.11 to 3.15. M_1 reports low ε, the highest U_{max} and u_{1a}. The Υ does not perform comparatively well in this dataset, but it returns acceptable models.

The model selected by both ROI versions is the same as M_1, which is indeed the best model from all available, scoring the highest values in all the desirable characteristics (see table 3.3).

The single model chosen by q_0 and M_4 seems, at first instance, as useful as the model selected by Υ, but at a closer look one notices that segments two to four have the same uplift, and the negative effect is not detected. Therefore M_4 model would be discarded in favour of Υ model. From the rest of M models (figure 3.15), M_2 is the best in terms of giving accurate predictions and distinguishing the segments. M_3 returns an inaccurate prediction in the first segments, which are the most important ones for uplift modelling, although it improves for the last bits. Finally, the model selected by M_5 is very close to q_0 and M_4.

In this dataset the Υ family did not score the highest values in the desirable characteristics, but overall presented a model that was both accurate and bold.

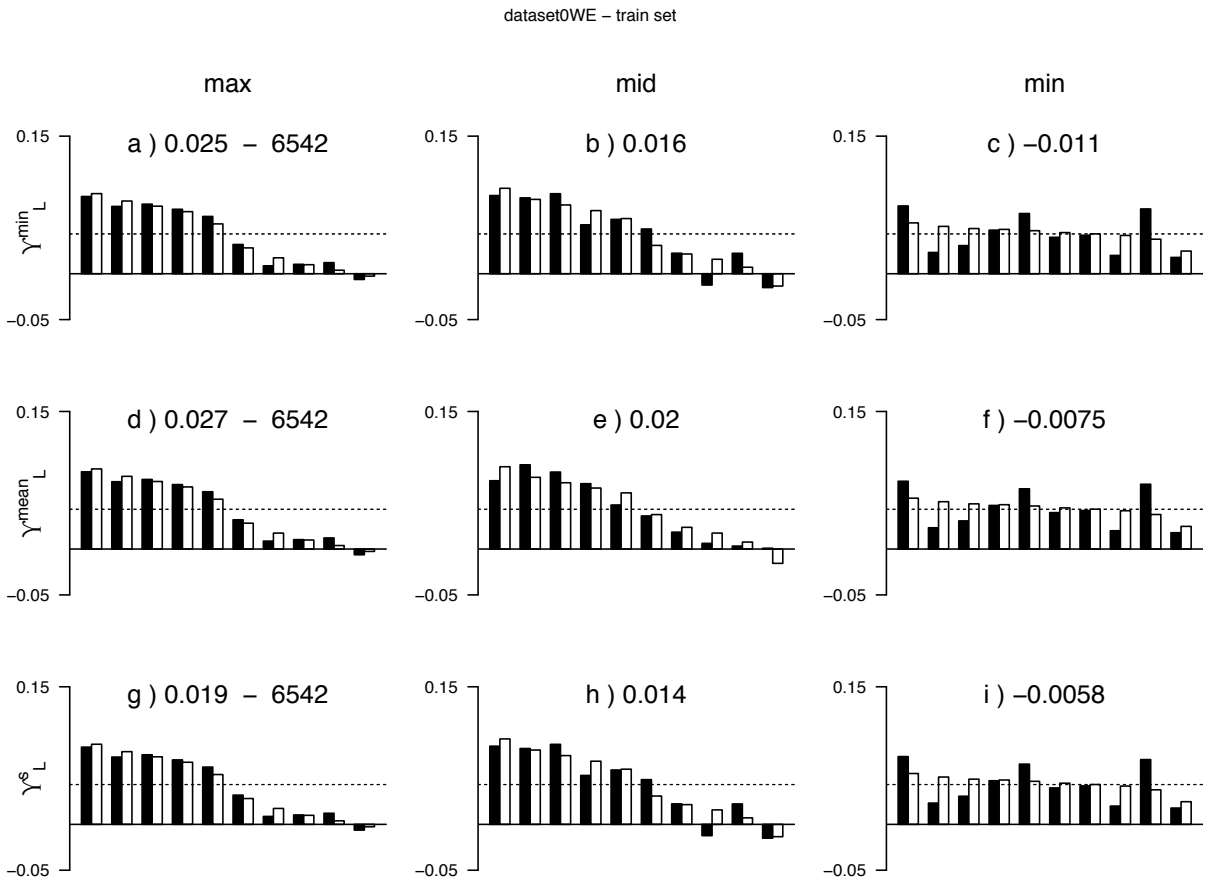

Figure 3.11: Accurate and bold models are available for this dataset and Υ_L is capable of selecting one that is a useful model. Lower Υ_L values correspond to less accurate models.

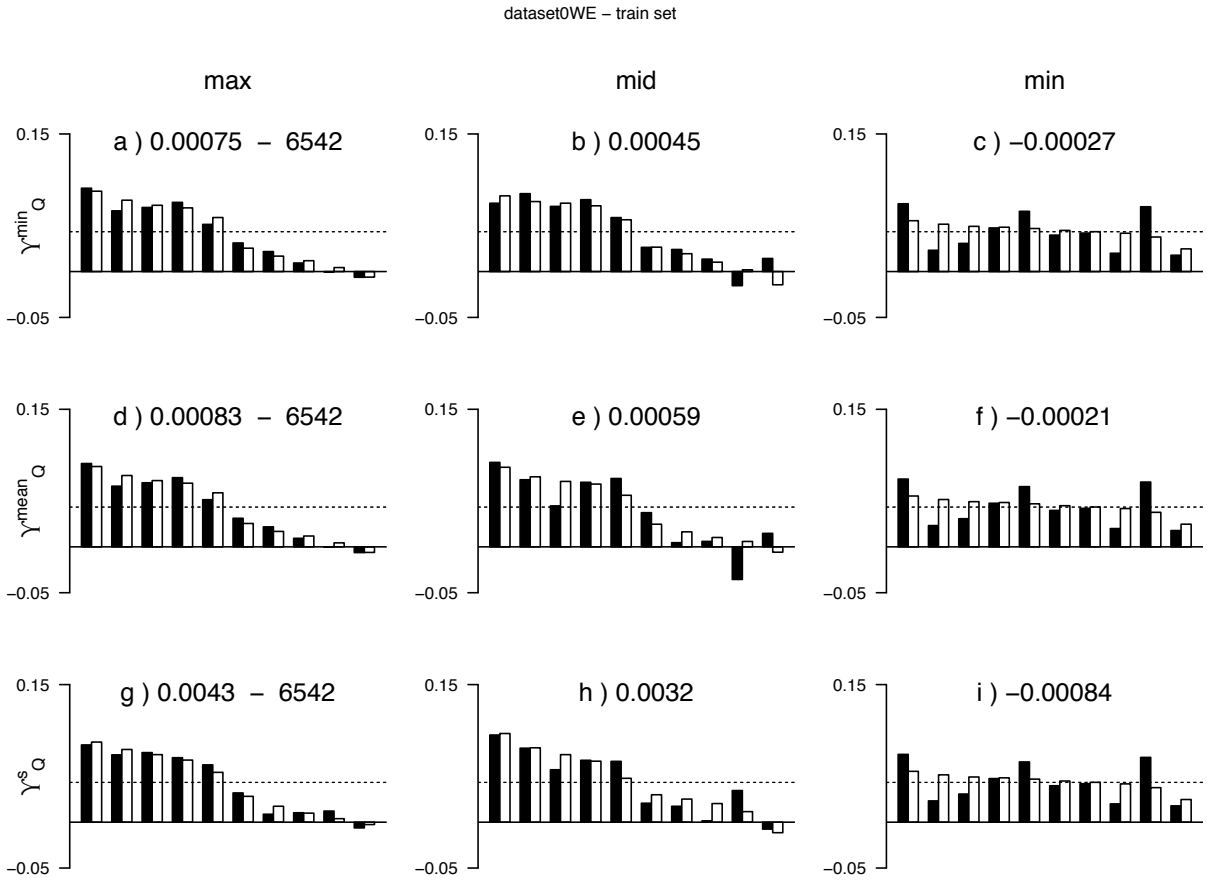

Figure 3.12: Increasing values for Υ_Q point to more accurate and bold models. Υ_Q models are mostly the same than the models selected by Υ_L.

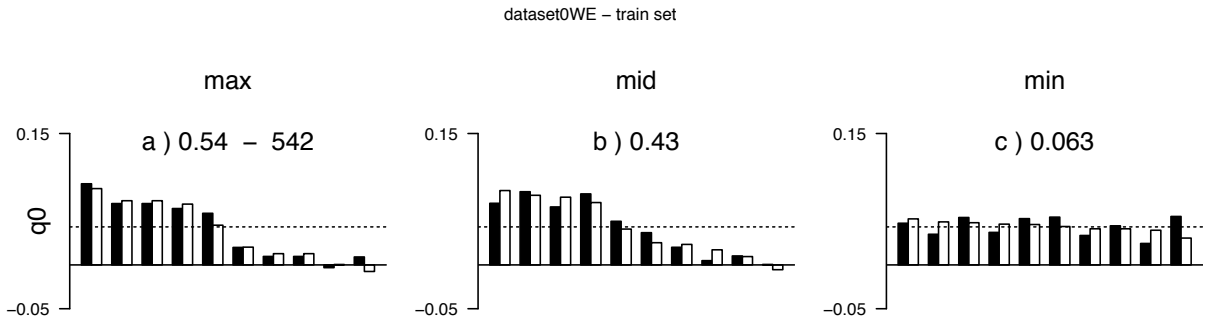

Figure 3.13: Although q_0 selects an acceptable model, it does not distinguish segments two to four between them, and it also fails in the prediction for the last deciles. The Υ family slightly outperform q_0 in this case, although q_0 scores higher in all the desirable characteristics in table 3.3

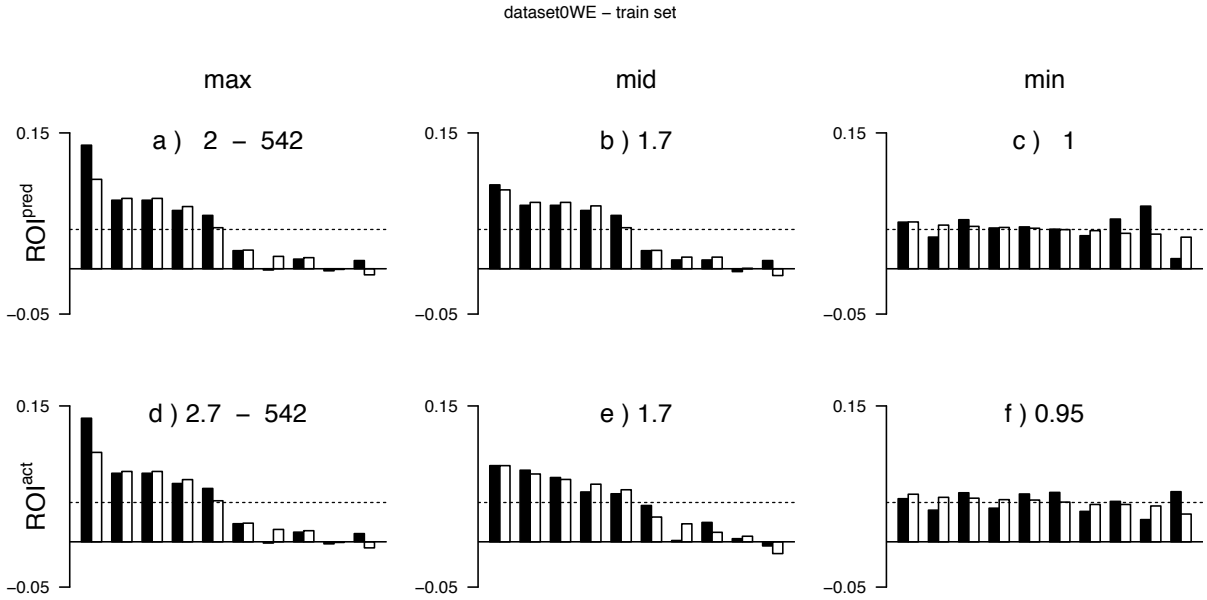

Figure 3.14: ROI metrics also fail to make a valid model choice in terms of overall accuracy, but praises uplift at first decile.

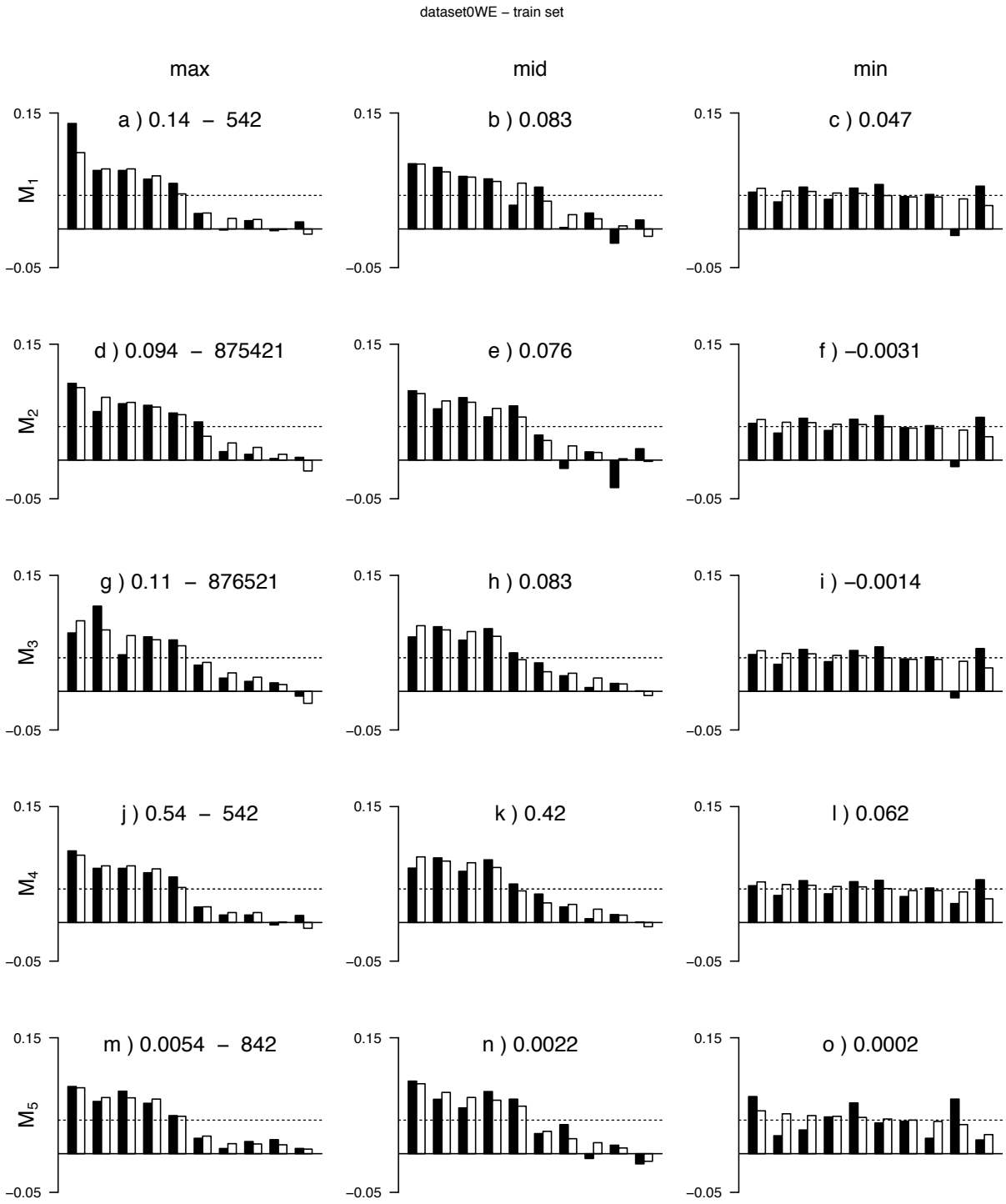

Figure 3.15: M metrics returns a variety of models, M_1 focuses on first decile uplift, M_3 fails to make a valid selection, while the rest of the measures try to compromise accuracy and boldness, but with less success than the Υ family.

3.2.2 Test Subset

All Υ measures select a similar model (see 3.16 and 3.17), except for Υ_Q^{mean}, which makes an unacceptable choice, as the most distinguishable uplift is in the fifth segment. Υ_Q^{mean} presents the highest ε from all Υ (table 3.4).

Υ_Q^{mean} is a version of the Υ family which gives less weight to accuracy than the other Υ versions do. Moreover, being a quadratic version, it gives more weight to first deciles. Therefore it does not penalize the mistake in the fifth decile sufficiently to discard it. Specially given that it has been accurately predicted above the overall uplift.

	U_{max}	$-\varepsilon$	r_s	u_{1p}	u_{1a}	$-\nu_p$	s_p
Υ_L^{min}	0.084	-0.008	0.867	0.08	0.077	-8	0.075
Υ_L^{mean}	0.084	-0.008	0.867	0.08	0.077	-8	0.075
Υ_L^{s}	0.084	-0.008	0.867	0.08	0.077	-8	0.075
Υ_Q^{min}	0.087	-0.008	0.915	0.087	0.079	-8	0.077
Υ_Q^{mean}	0.078	-0.009	0.559	0.076	0.083	-6	0.088
Υ_Q^{s}	0.084	-0.008	0.867	0.08	0.077	-8	0.075
q_0	0.082	-0.009	0.62	0.082	0.075	-8	0.075
ROI^{pred}	0.081	-0.014	0.879	0.081	0.098	-8	0.113
ROI^{act}	0.091	-0.013	0.855	0.084	0.082	-8	0.081
M_1	0.091	-0.013	0.855	0.084	0.082	-8	0.081
M_2	0.087	-0.008	0.915	0.087	0.079	-8	0.077
M_3	0.084	-0.013	0.879	0.058	0.09	-8	0.107
M_4	0.082	-0.009	0.62	0.082	0.075	-8	0.075
M_5	0.077	-0.006	0.83	0.077	0.08	-10	0.074

Table 3.4: Hillstrom WE dataset; visit rate; test subset. Values for maximum uplift, prediction error, monotonicity, predicted and actual uplift at 1^{st} decile, negative effect and spread for "max" models in figures 3.16 to 3.20. M_5 reports the lower ε but it is not the preferred model because it does not differentiate much the segments, and it has the lowest U_{max}. The preferred model, M_2 reports a similar ε than the *Upsilon* family, but higher s_p and r_s

q_0 and ROI^{act} (figures 3.18 and 3.19) do not select an acceptable model, either because they are inaccurate in their predictions, make no distinction between segments or their selection is not monotone. ROI^{pred} makes a better choice, although not very accurate. Again q_0 is outperformed by the Υ family in all the desirable characteristics.

Measures M_1 and M_3 should be discarded because they are inaccurate. Similarly M_4 and M_5 should be discarded because they do not differentiate much the first segments and make wrong predictions in the fifth segment. M_2 is an acceptable model. It is the same model selected by Υ_Q^{min}, and it would be the preferred choice for this dataset.

As expected for a test subset, the list of models were overall less accurate, but most of the Υ metrics were capable of filtering the bad performers and combined boldness and accuracy to select a useful model.

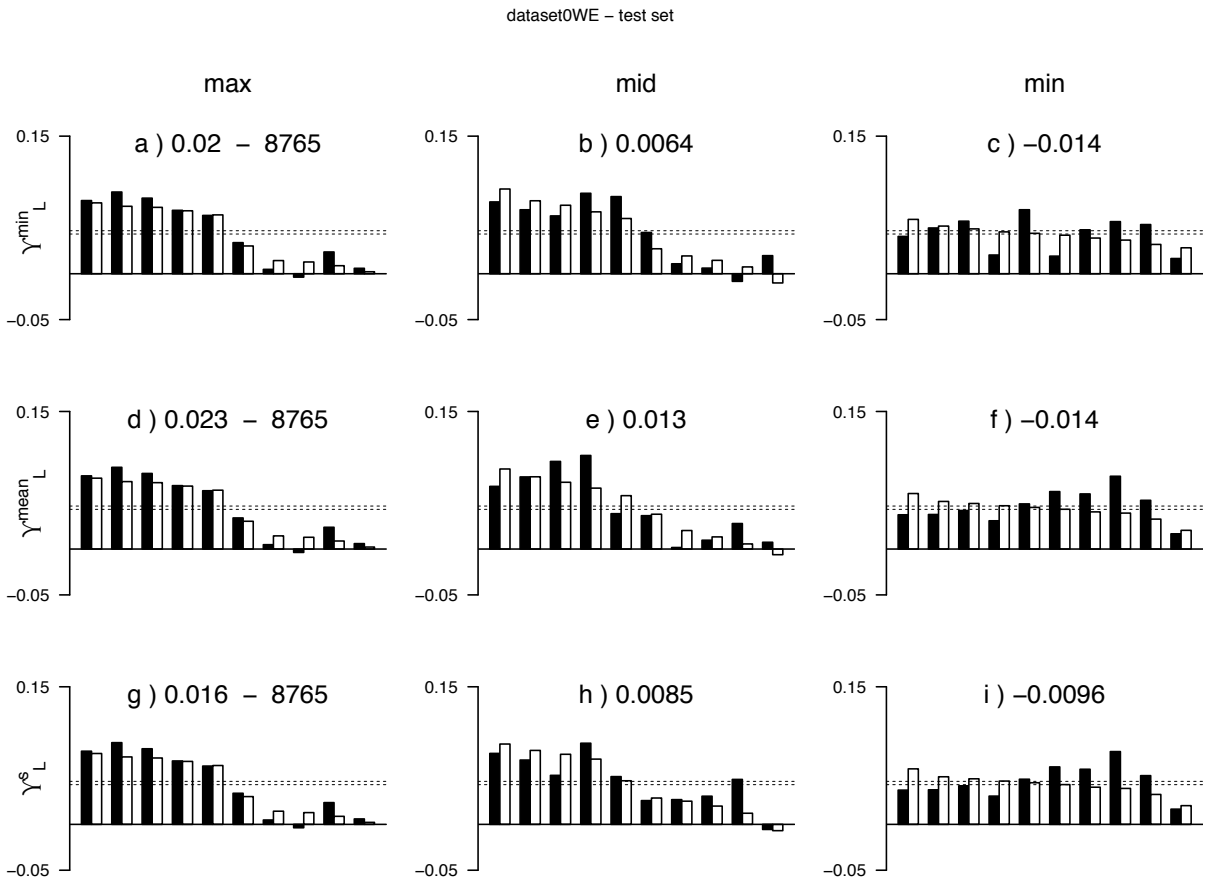

Figure 3.16: Υ_L family of metrics select a fair, although it does not account for the failure in the model, which is a second uplift segment higher than the first.

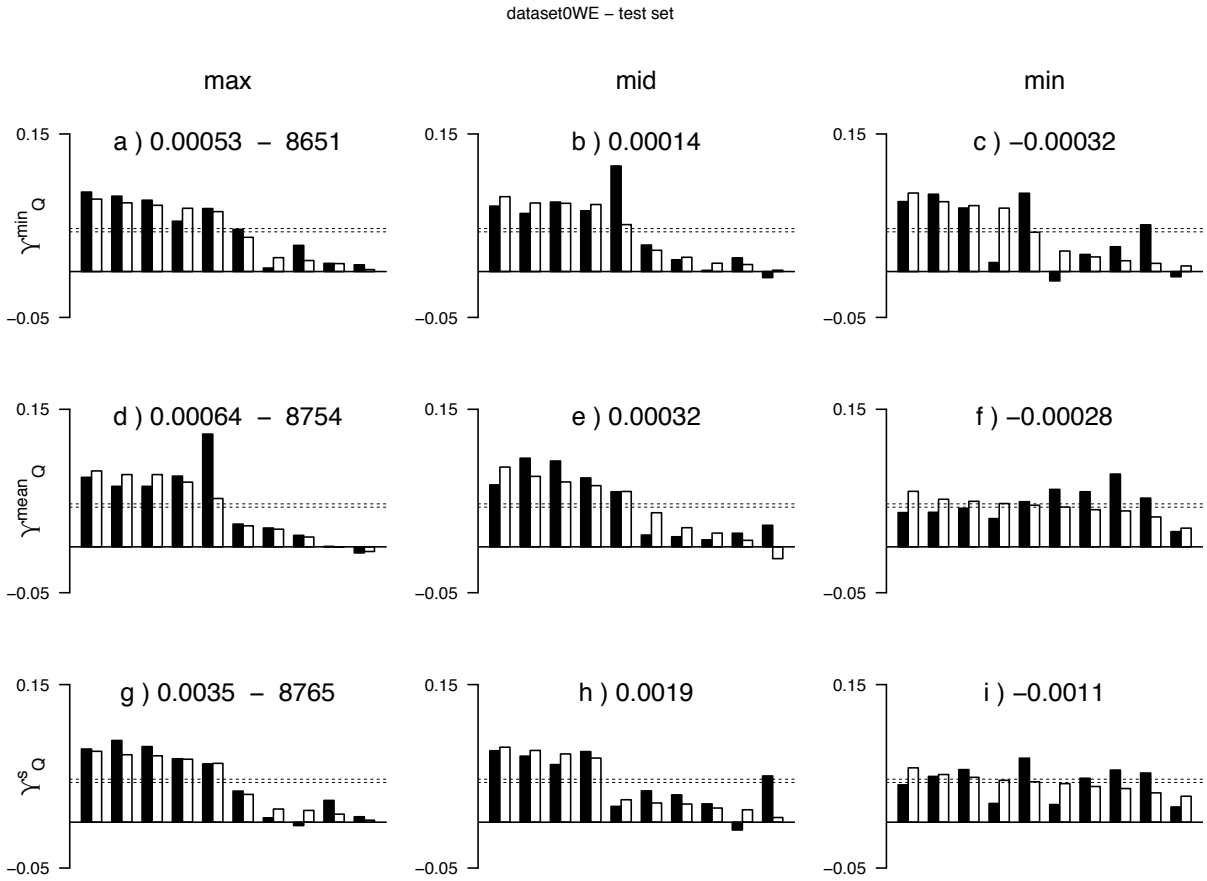

Figure 3.17: Υ_Q^{mean} makes a very poor choice, selecting a model that makes a highly inaccurate prediction in the fifth segment. Υ_Q^s chooses the same model as Υ_L, while Υ_Q^{min} is slightly more monotone.

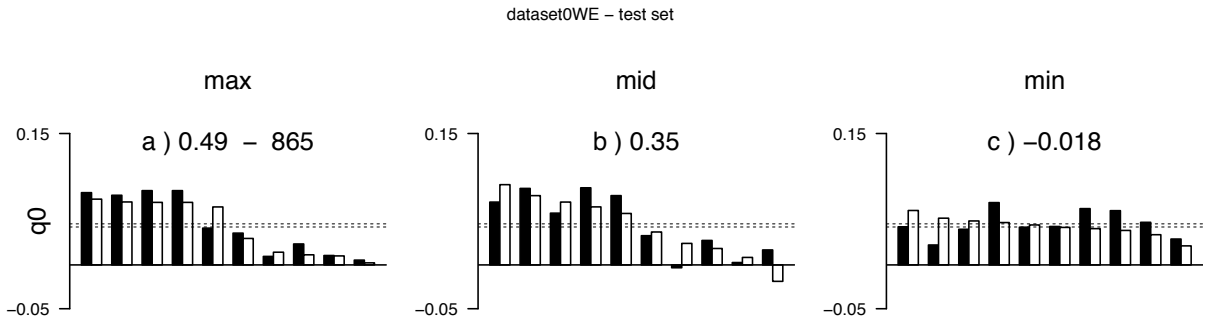

Figure 3.18: The model selected by q_0 is not as useful as the Υ_L family. Although it is not markedly inaccurate, as one would expect for q_0 in a test subset, its prediction are not monotone for the first five segments and completely wrong for the fifth segment.

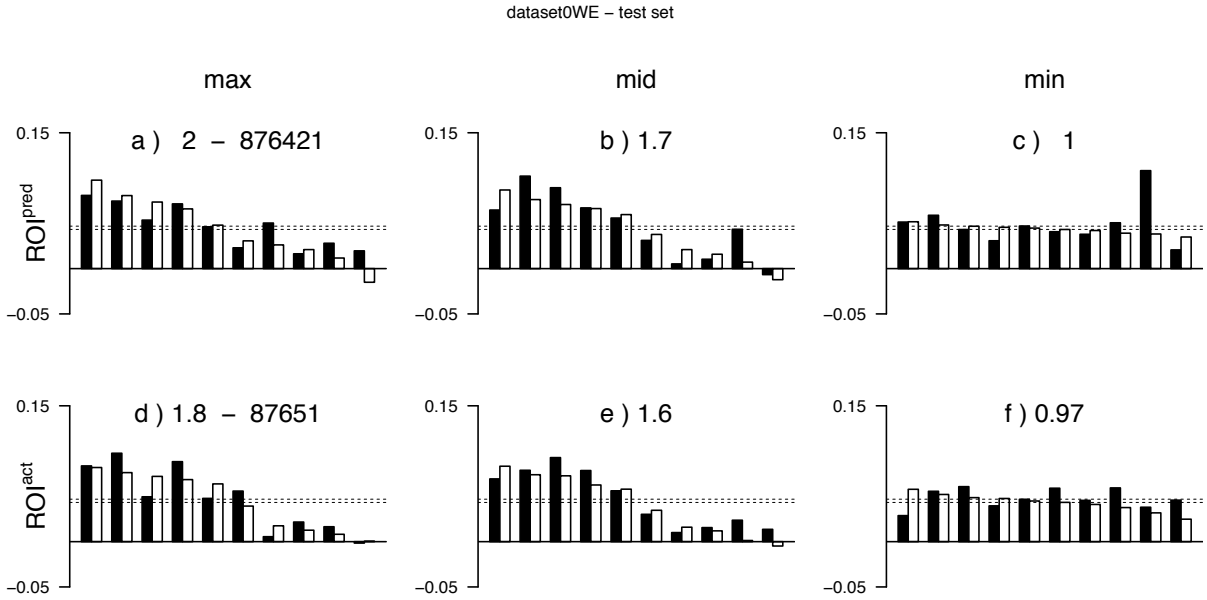

Figure 3.19: ROI metrics again fail to make a valid model choice in terms of overall accuracy. Although in this case ROI^{pred} selects a model that correctly ranks the first three segments.

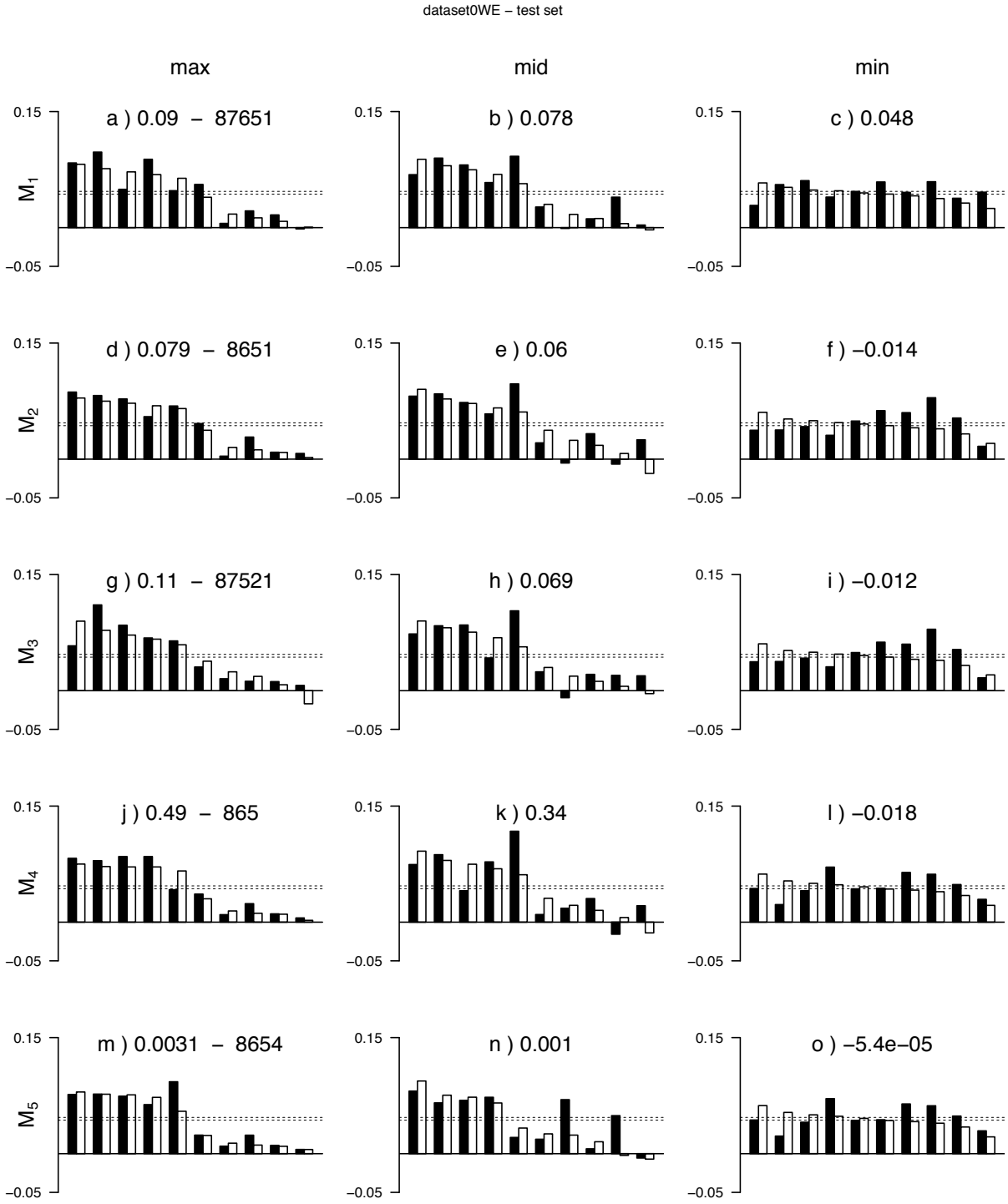

Figure 3.20: In this case M_2 is the only measure that selects a useful model. M_1, M_3 and M_5 metrics should be discarded in the view of that selection.

3.3 Hillstrom ME Dataset - Conversion Rate

The four previous sections illustrated that models chosen by high values of the different quality measure could indicate more accurate and/or bold models, or at least a model that cannot be judge to be worse than models with lower quality measure values, thus in the next sections, we only present the best model for each quality measure.

3.3.1 Training Subset

In this dataset, the best model is chosen by ROI^{act}, M_1 and M_2 (see figure 3.21). Υ_L^{min} does not make the best choice, as it selects a model that does not differentiate segments two to five. All the rest of the Υ family selects a different model, which is not exactly accurate, but it correctly identifies the segments that are above and below the overall uplift. q_0 chooses a similar model to the Υ family but with less accuracy on the last deciles. M_3, M_5 and ROI^{pred} make a poor selection. For instance, they do not identify the highest uplift. Reviewing M_3 performance so far, one can see that it systematically selects models with low u_{1a}, and very low last decile. This effect is caused by the ratio s_p/u_{1p} in its definition 2.2. An alternative definition of M_3 accounting for the actual spread and/or actual first decile uplift would probably be more convenient.

	U_{max}	$-\varepsilon$	r_s	u_{1p}	u_{1a}	$-\nu_p$	s_p
Υ_L^{min}	0.014	-0.001	0.83	0.014	0.014	-10	0.013
Υ_L^{mean}	0.017	-0.002	0.818	0.017	0.016	-8	0.016
Υ_L^{s}	0.017	-0.002	0.818	0.017	0.016	-8	0.016
Υ_Q^{min}	0.017	-0.002	0.806	0.017	0.016	-8	0.016
Υ_Q^{mean}	0.017	-0.002	0.806	0.017	0.016	-8	0.016
Υ_Q^{s}	0.017	-0.002	0.806	0.017	0.016	-8	0.016
q_0	0.017	-0.003	0.842	0.017	0.017	-8	0.016
ROI^{pred}	0.013	-0.003	0.818	0.011	0.017	-8	0.017
ROI^{act}	0.018	-0.002	0.83	0.018	0.016	-8	0.016
M_1	0.018	-0.002	0.83	0.018	0.016	-8	0.016
M_2	0.018	-0.002	0.83	0.018	0.016	-8	0.016
M_3	0.016	-0.004	0.758	0.01	0.016	-8	0.016
M_4	0.017	-0.003	0.842	0.017	0.017	-8	0.016
M_5	0.012	-0.001	0.906	0.012	0.012	-10	0.012

Table 3.5: Values for maximum uplift, prediction error, monotonicity, predicted and actual uplift at 1^{st} decile, negative effect and spread for models in figure 3.21.

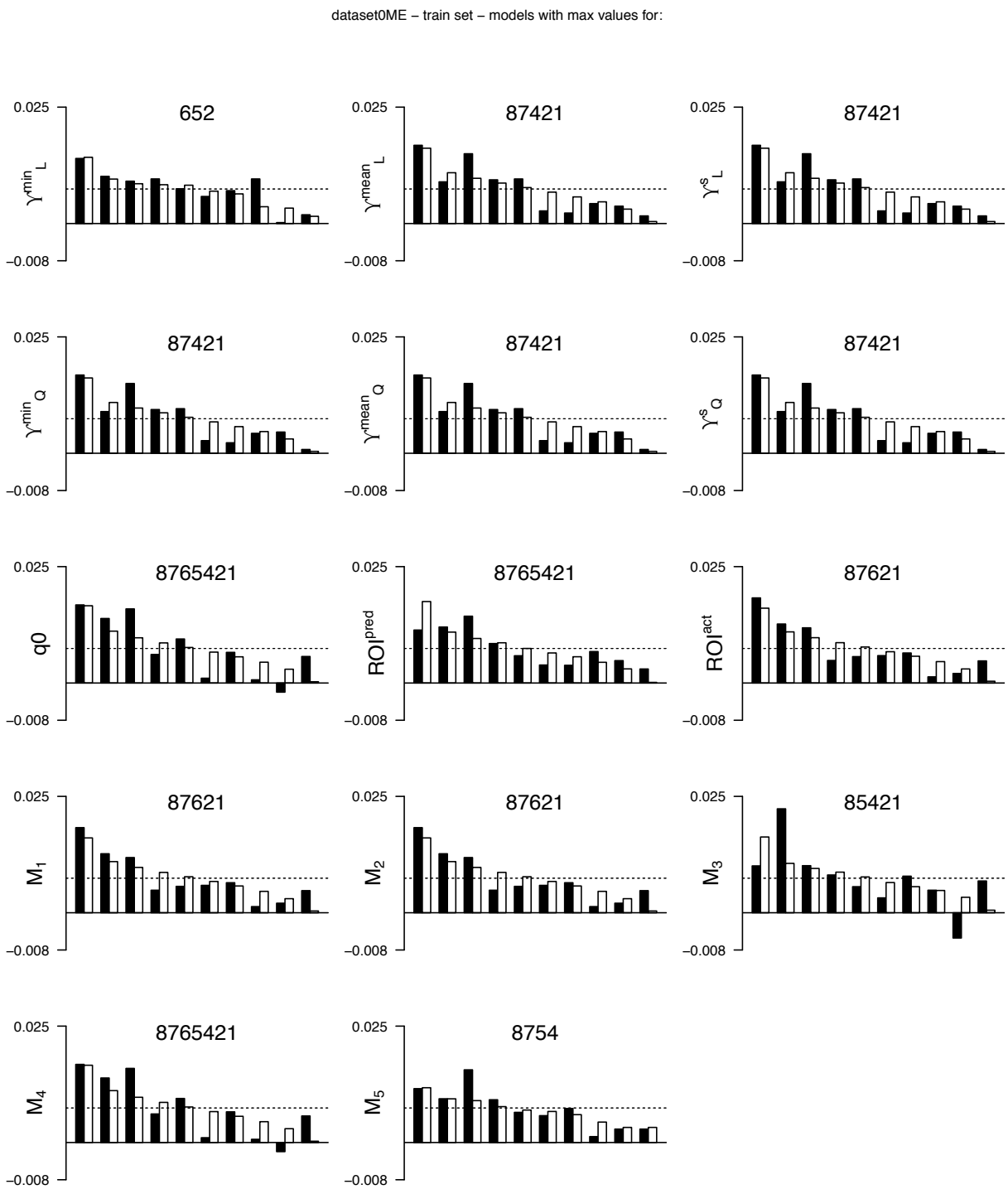

Figure 3.21: Models on a training subset. The number refers to a model identifier. M_3, M_5 and ROI^{pred} make a poor selection. They do not identify the highest uplift. Υ_L^{min} selects a model that does not differentiate segments two to five. All the rest of the Υ family selects a different model that correctly identifies the segments above and below the overall uplift. q_0 chooses a similar model to the Υ family but less accurate on the last deciles. ROI^{act}, M_1 and M_2 select the best model, with high uplift and accuracy in the first three segments.

3.3.2 Test Subset

In this case the split in training and test set did not divide the sample in two equivalent sets, which is reflected by the fact that the overall predicted and actual uplift (dotted lines) are at certain distance of each other (see figure 3.22). In addition, the overall uplift effect is very low and that makes the uneven separation between training and test set more noticeable. However, this can be a normal scenario in a real problem, so the quality measures should still make a wise selection out of the bad models. Although the selection of "the best of the worst" would depend on the particulars of the problem, we could argue that Υ measures, together with M_2 make the most accurate predictions, albeit not very bold, with a similar actual uplift for the first three segments. Υ^s_Q distinguishes itself by accurately predicting a high first decile uplift. Although if we focus only on boldness ROI^{act} and M_1 are the best models. Noticeable q_0, M_4 and M_5 models present a high first uplift (with the same overall prediction error ε than Υ), but the rest of the segment predictions are less accurate than Υ^s_Q predictions, so those models would not be preferred over the later.

	U_{max}	$-\varepsilon$	r_s	u_{1p}	u_{1a}	$-\nu_p$	s_p
Υ^{min}_L	0.011	-0.003	0.285	0.011	0.015	-10	0.015
Υ^{mean}_L	0.009	-0.003	0.624	0.009	0.013	-10	0.009
Υ^s_L	0.009	-0.003	0.624	0.009	0.013	-10	0.009
Υ^{min}_Q	0.009	-0.003	0.624	0.009	0.013	-10	0.009
Υ^{mean}_Q	0.009	-0.003	0.624	0.009	0.013	-10	0.009
Υ^s_Q	0.014	-0.003	0.333	0.014	0.013	-10	0.01
q_0	0.013	-0.003	0.274	0.013	0.012	-10	0.012
ROI^{pred}	0.011	-0.005	-0.079	0.011	0.019	-10	0.018
ROI^{act}	0.017	-0.004	-0.273	0.017	0.017	-8	0.016
M_1	0.017	-0.004	-0.273	0.017	0.017	-8	0.016
M_2	0.009	-0.003	0.624	0.009	0.013	-10	0.009
M_3	0.009	-0.004	0.394	0.009	0.016	-8	0.016
M_4	0.013	-0.003	0.274	0.013	0.012	-10	0.012
M_5	0.013	-0.003	0.274	0.013	0.012	-10	0.012

Table 3.6: Values for maximum uplift, prediction error, monotonicity, predicted and actual uplift at 1^{st} decile, negative effect and spread for models in figure 3.22.

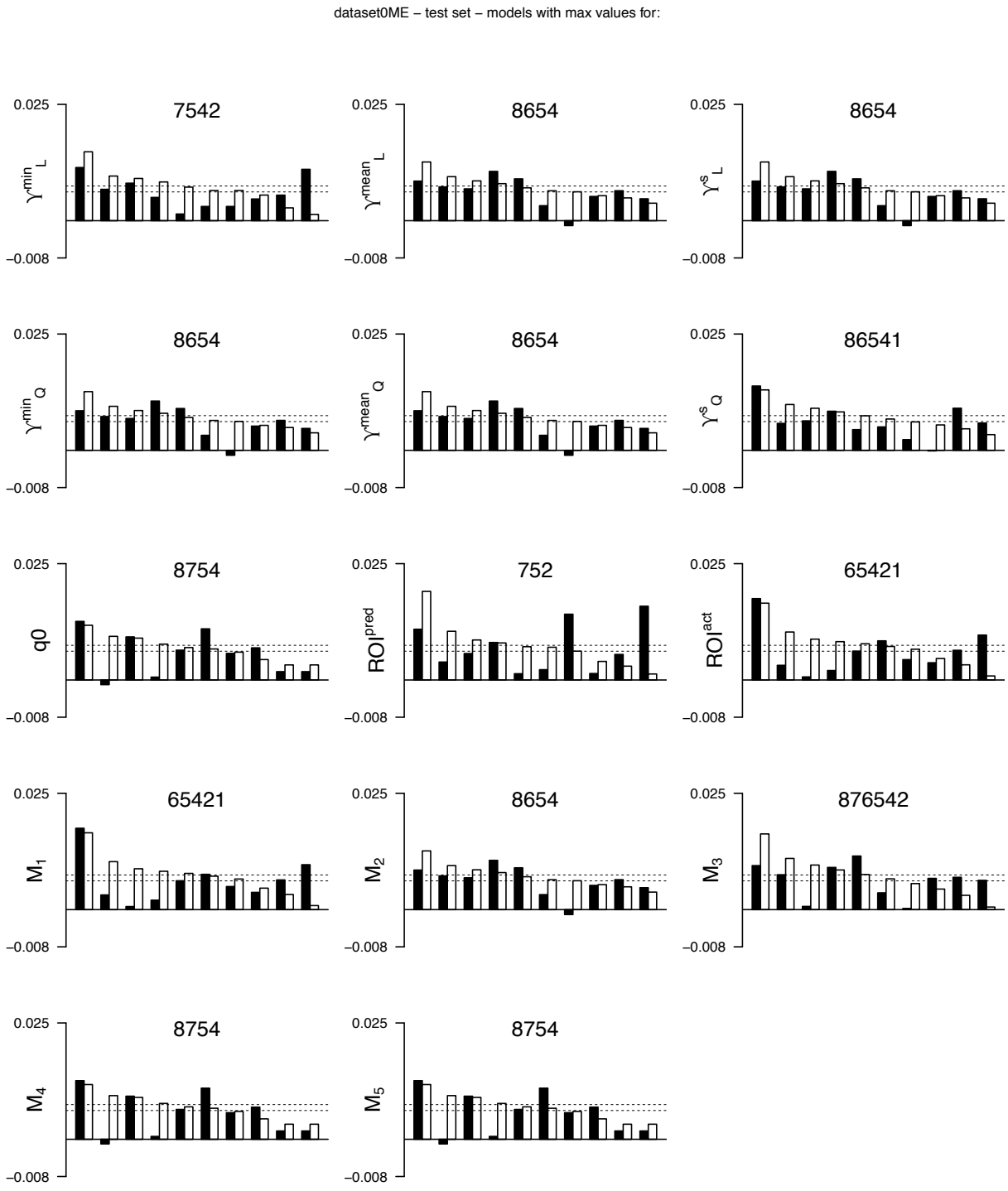

Figure 3.22: Models on a test subset. The number refers to a model identifier. Training and test subset have different overall uplift, as shown by the separation between the dotted lines. The models available are not overall accurate but the quality measures should still make a wise selection out of the bad models. It is difficult to consider any model better than the other unless we focus only on boldness. In this case ROI^{act} and M_1 are the best models, while Υ^{s}_{Q} selects a model with lower first decile uplift and comparatively accurate considering the rest of the models. In fact, all models from the Υ family (and M_2) are the same (except for Υ^{min}_{L} and Υ^{s}_{Q}), which seems more accurate than the models selected by other quality measures (although ε is not much different than for other models - table 3.6). Moreover, for this model the actual uplift for each segments tends to the overall uplift in most of the cases.

3.4 Hillstrom WE Dataset - Conversion Rate

3.4.1 Training Subset

For this dataset any quality measure makes similar selection and useful predictions (figure 3.23), except for ROI^{act} and M_1, whose models are inaccurate. q_0, M_2 and M_4 and M_5 are probably the best choices, as they make a clear distinction between segments. Υ^s_Q selects the same model as Υ_L.

	U_{max}	$-\varepsilon$	r_s	u_{1p}	u_{1a}	$-\nu_p$	s_p
Υ^{min}_L	0.014	-0.001	0.915	0.014	0.013	-4	0.017
Υ^{mean}_L	0.014	-0.001	0.915	0.014	0.013	-4	0.017
Υ^{s}_L	0.014	-0.001	0.915	0.014	0.013	-4	0.017
Υ^{min}_Q	0.014	-0.001	0.988	0.014	0.014	-6	0.019
Υ^{mean}_Q	0.014	-0.001	0.988	0.014	0.014	-6	0.019
Υ^{s}_Q	0.014	-0.001	0.915	0.014	0.013	-4	0.017
q_0	0.014	-0.002	0.939	0.014	0.014	-6	0.018
ROI^{pred}	0.013	-0.002	0.758	0.013	0.014	-6	0.017
ROI^{act}	0.016	-0.003	0.455	0.016	0.014	-6	0.016
M_1	0.016	-0.001	0.867	0.016	0.013	-6	0.016
M_2	0.015	-0.002	0.939	0.015	0.014	-6	0.019
M_3	0.014	-0.001	0.988	0.014	0.014	-6	0.019
M_4	0.014	-0.002	0.939	0.014	0.014	-6	0.018
M_5	0.014	-0.001	0.927	0.014	0.013	-6	0.016

Table 3.7: Values for maximum uplift, prediction error, monotonicity, predicted and actual uplift at 1^{st} decile, negative effect and spread for models in figure 3.23.

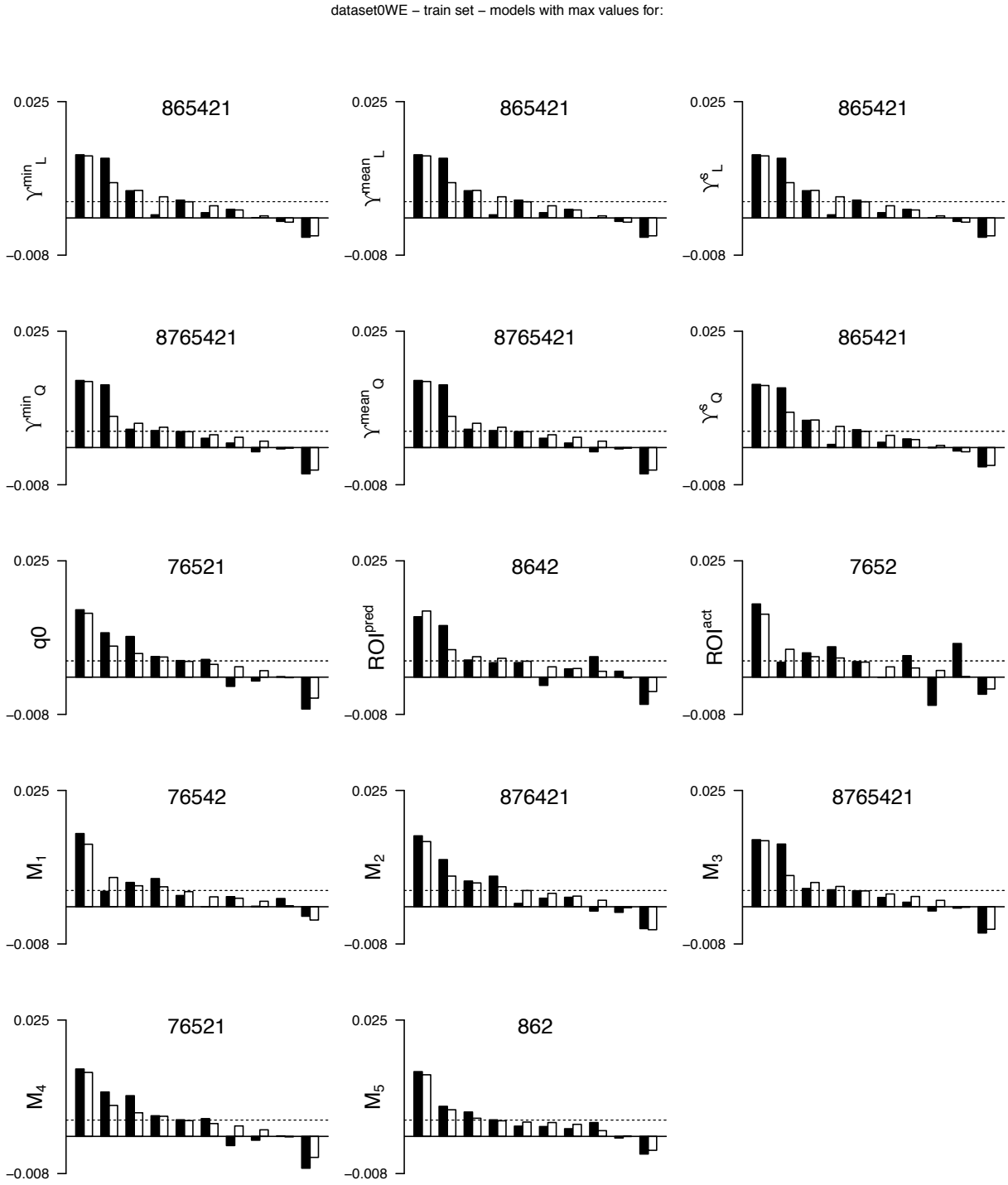

Figure 3.23: Models on a train subset. The number refers to a model identifier. The different quality measures select different models but they all have a similar shape. ROI^{act} and M_1 models are the least accurate. From the rest, any model on the Υ family, M_2 or q_0 are good choices, but M_5 is probably the most accurate and bold model, however M_1 presents the smallest ε, highest U_{max} (see table tab:TableWEconversiontrain).

.

3.4.2 Test Subset

We have again a situation where the separation between training and test data makes most models inappropriate for an accurate prediction in the test subset, and most of the quality measures fail to make a "less bad" model selection (see figure 3.24).

All Υ family selects the same model, which at least predicts correctly the ranking order of the first three segments, and accurately predicts a negative effect in the last decile. The Υ family presents the smallest ε in table 3.8. A similar model is selected by M_5. ROI^{act}, q_0 and M_4 make a less accurate prediction than the Υ family (the ranking order for the first three deciles is completely the opposite, which is confirmed by a low s_p coefficient) and their predictions are not bolder. M_3 model makes predictions that are not far from the actual values, except for the first and the sixth decile. Finally, the rest of the quality measures, $M_1, M_2, \text{ROI}^{act}$ make substantially wrong predictions in two segments, reflected by a low s_p coefficient in table 3.8.

	U_{max}	$-\varepsilon$	r_s	u_{1p}	u_{1a}	$-\nu_p$	s_p
Υ_L^{min}	0.008	-0.002	0.915	0.008	0.013	-6	0.015
Υ_L^{mean}	0.008	-0.002	0.915	0.008	0.013	-6	0.015
Υ_L^{s}	0.008	-0.002	0.915	0.008	0.013	-6	0.015
Υ_Q^{min}	0.008	-0.002	0.915	0.008	0.013	-6	0.015
Υ_Q^{mean}	0.008	-0.002	0.915	0.008	0.013	-6	0.015
Υ_Q^{s}	0.008	-0.002	0.915	0.008	0.013	-6	0.015
q_0	0.008	-0.003	0.794	0.008	0.013	-6	0.016
ROI^{pred}	0.008	-0.003	0.939	0.008	0.014	-6	0.018
ROI^{act}	0.012	-0.003	0.782	0.012	0.01	-8	0.011
M_1	0.012	-0.003	0.782	0.012	0.011	-8	0.011
M_2	0.012	-0.003	0.782	0.012	0.011	-8	0.011
M_3	0.009	-0.003	0.879	0.006	0.014	-6	0.018
M_4	0.008	-0.003	0.794	0.008	0.013	-6	0.016
M_5	0.008	-0.002	0.891	0.008	0.013	-6	0.016

Table 3.8: Values for maximum uplift, prediction error, monotonicity, predicted and actual uplift at 1^{st} decile, negative effect and spread for models in figure 3.24.

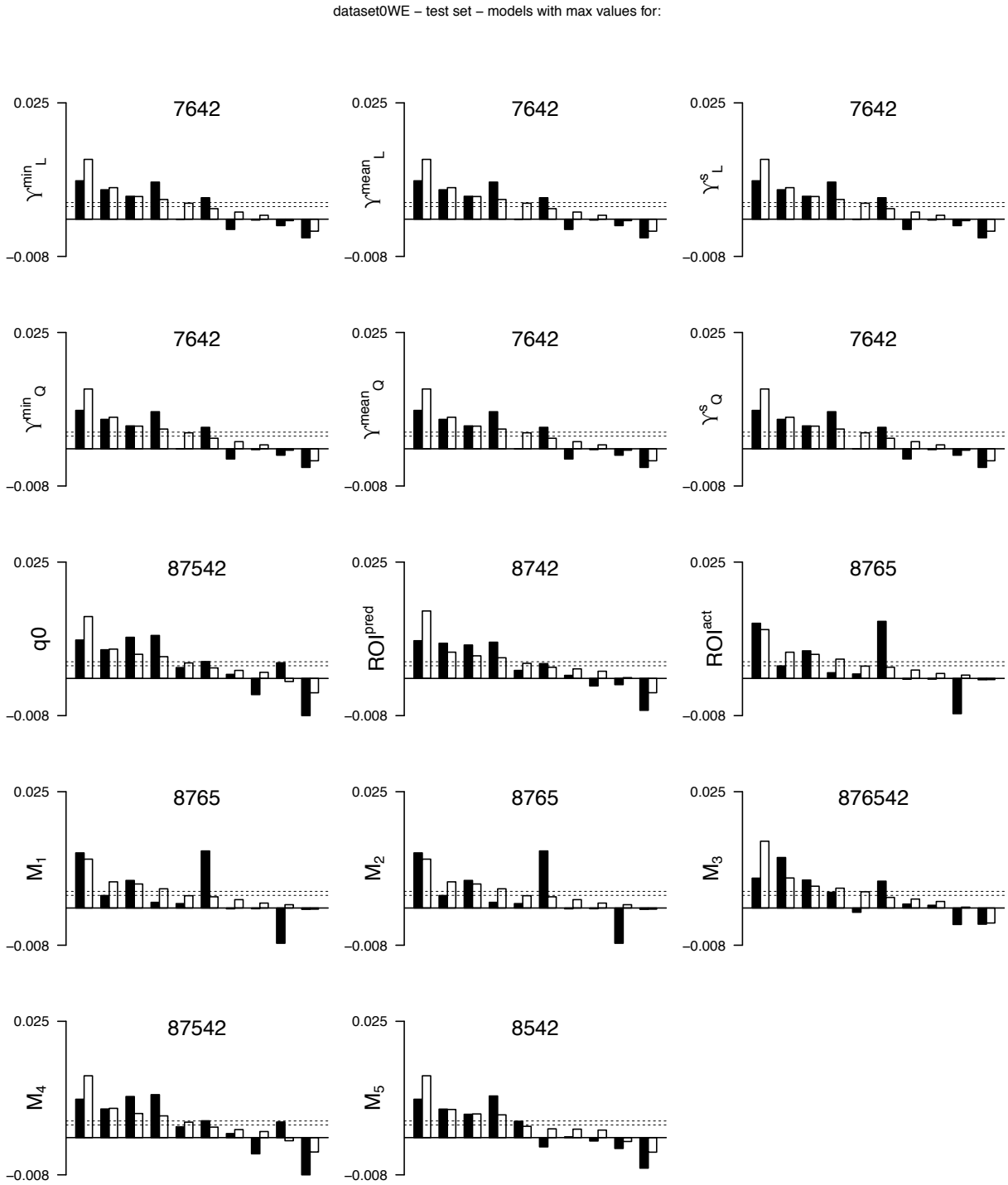

Figure 3.24: Models on a test set. The number refers to a model identifier. The Υ family of metrics selects the same model, which predicts correctly the ranking order of the first three segments, and accurately predicts a negative effect in the last decile. A similar model is selected by M_5. Models selected by the measures ROI^{act}, q_0 and M_4 are less accurate than the Υ family and their predictions are not bolder. M_3 model makes predictions that are not far from the actual values, except for the first and the sixth decile. Finally, the rest of the quality measures, $M_1, M_2,$ and ROI^{act} make substantially wrong predictions.

3.5 Synthetic Datasets

We present in this section, figures 3.25 to 3.27, the uplift by deciles graph for the models selected maximizing each of the quality measure, for three of the synthetic datasets generated. We present only the test subset as it is most challenging.

Figure 3.25 shows a dataset which was difficult to model, however the Υ family selects a model clearly more accurate than the rest, while still predicting an uplift above the overall uplift. On the other hand q_0 and ROI make predictions that are less accurate than the Υ selection. The models selected by the M family are more diverse, being M_2 possible the best, as it is monotone in the first five segments. One can conclude that when accuracy is not available, the Υ family still makes a selection that reduces error and at the same time achieves segment differentiation.

Figure 3.26 shows very similar models for all metrics. Interestingly, the Υ_L family groups the segments in two, while the Υ_Q family presents a smoother transition, which it should be preferred, as it distinguishes one segment from the other.

Finally, figure 3.27 shows the same model shape for all metrics, making it difficult what model is better than the other.

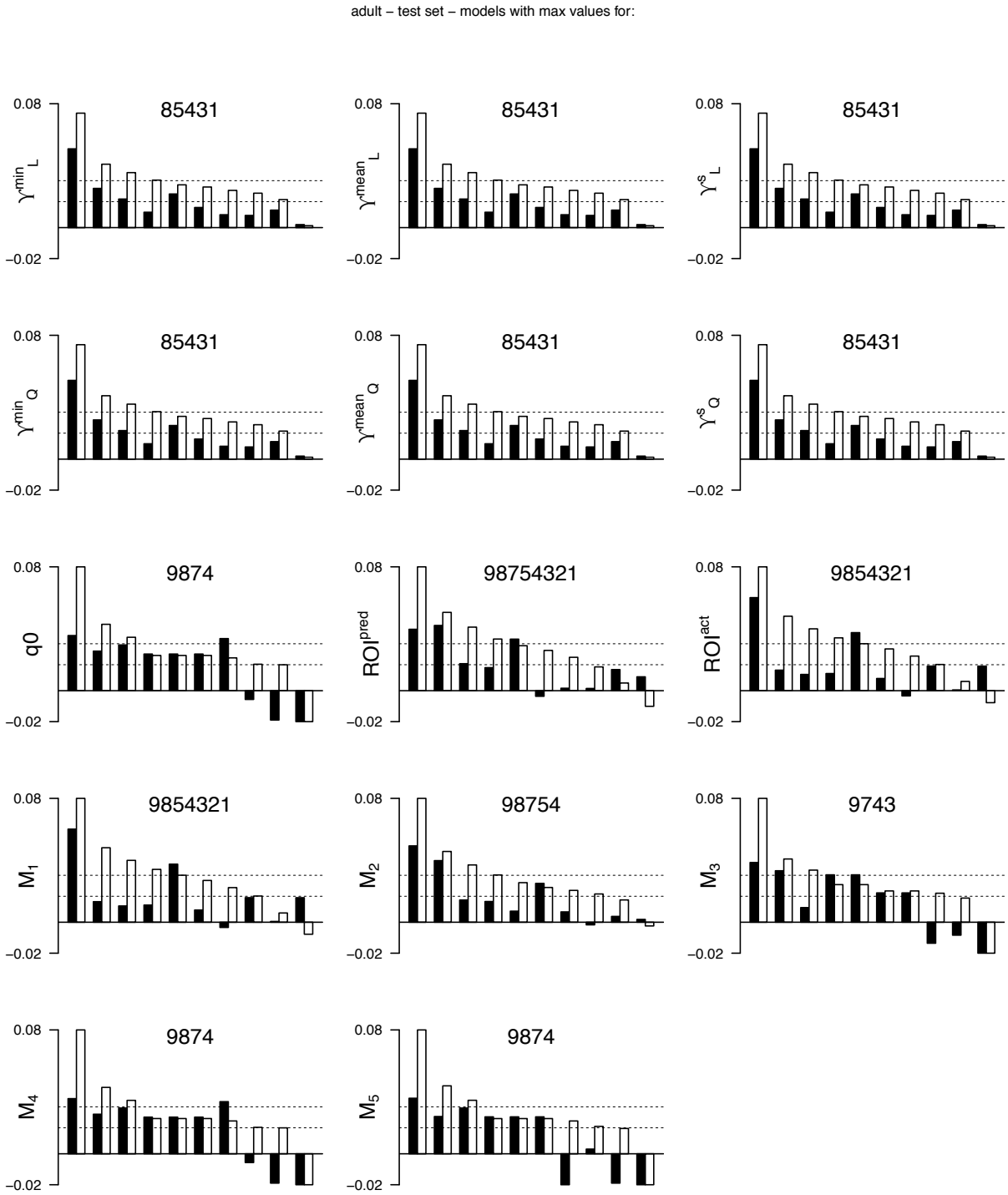

Figure 3.25: List of models selected by maximizing the different quality measures on a synthetic dataset ("adult"), for the test subset. The synthetic data was challenging to model and the quality measures did not have very good models available. However, the Υ family selects a model clearly more accurate, while still predicting an uplift above the overall uplift. On the other hand q_0 and ROI make predictions that are less accurate than the Υ selection. The models selected by the M family are more diverse, being M_2 possible the best, as it is monotone in the first five segments.

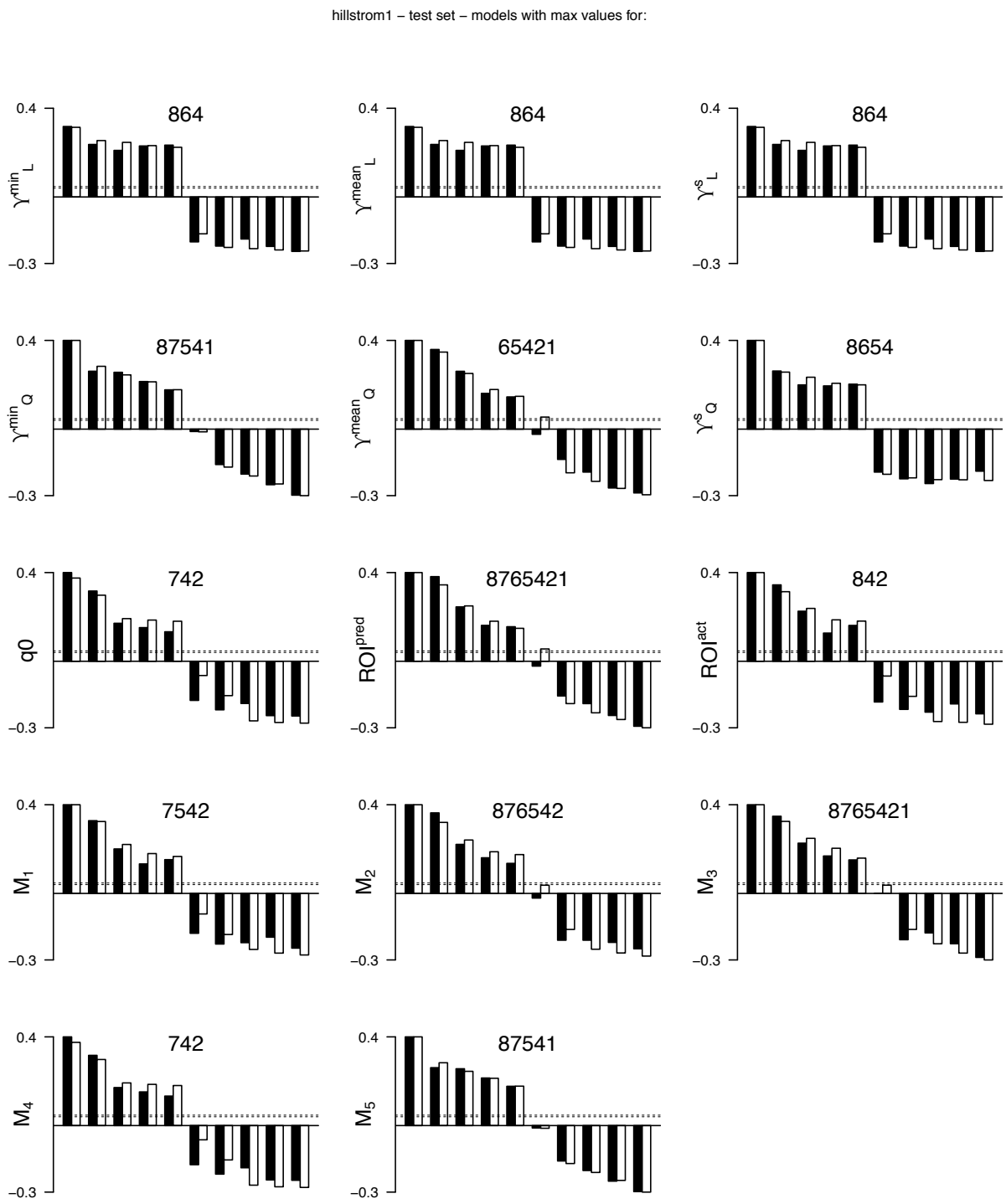

Figure 3.26: List of models selected by maximizing the different quality measures on a synthetic dataset ("hillstrom1"), for the test subset. Even though this is a test subset the predictions are accurate. In general, a synthetic dataset is easier to model and that made our task of characterising the quality measure more difficult. From all quality measures the Υ family, M_5 and ROI^{pred} select the most accurate models. Υ^{min}_{Q} makes a model selection that distinguishes all the segments, accurately predicting a different uplift for each decile.

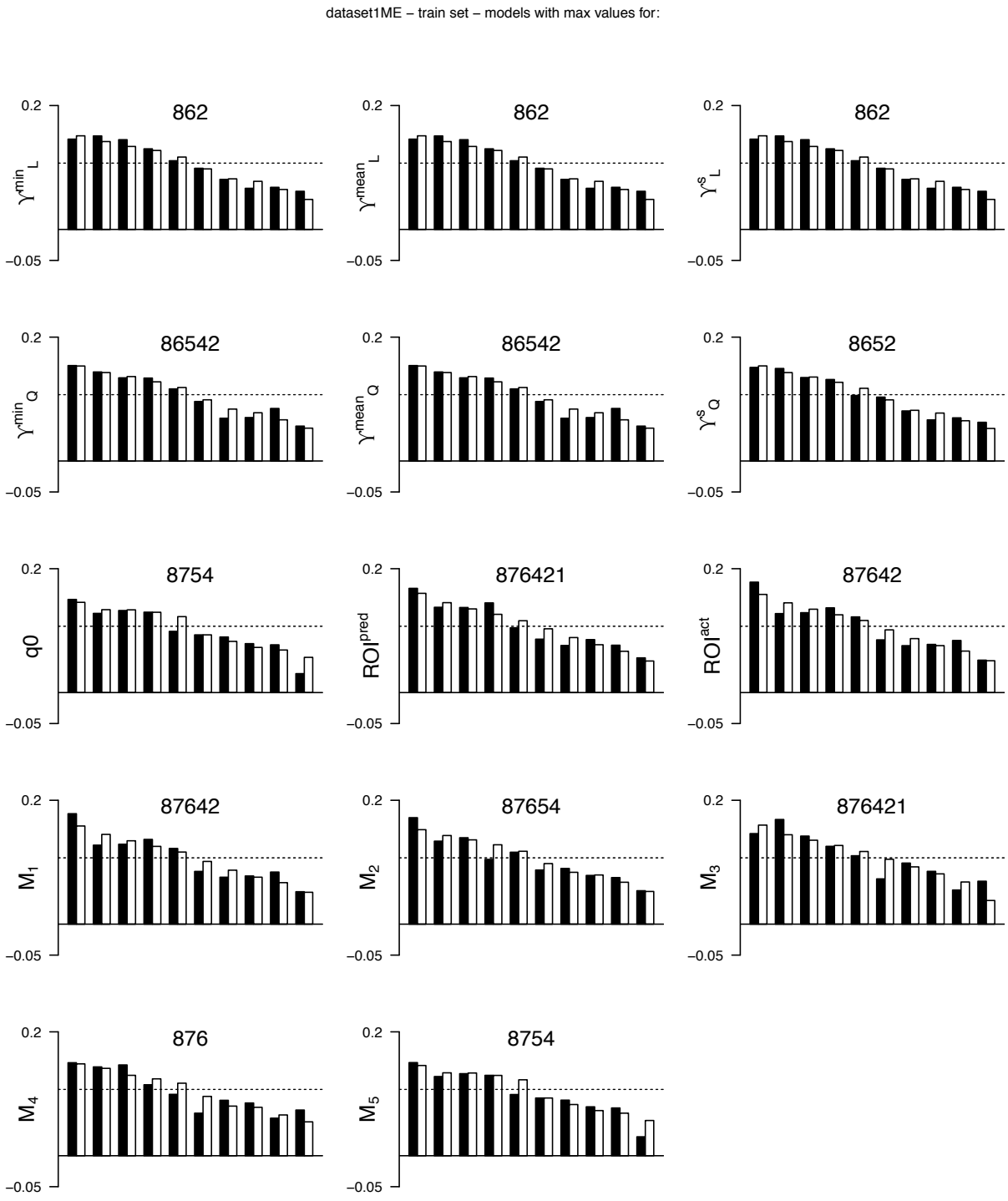

Figure 3.27: List of models selected by maximizing the different quality measures on a synthetic dataset ("dataset1ME"), for the test subset. Clearly the Υ family selects very accurate models, making precise predictions for all deciles. However most of the metrics perform well in this aspect and any difference with the Υ family is subtle and not significant. Synthetic datasets are not always useful on differentiating the characteristics of the metrics.

3.6 Ranking the Quality Measurements

So far, we have presented multiple models on eight real world datasets (four training sets and four test sets, sections 3.1 to 3.4) and made a subjective judgement on what model is best, hence assuming that the quality measure that pointed to that model is the most useful. We have also presented examples of the models selected by the quality measurements on synthetic data. We conclude that:

- Increasing values of the quality measures from the families Υ and M led to models that could not be considered worse than models with lower values for that quality measure, except for M_3 (figure 3.15) and Υ_Q^{mean} (see figure 3.17). M_3 account for the ratio s_p/u_{1k} is causing its low performance. Υ_Q^{mean} poor performance is due to the higher contribution to Υ from the accurate first deciles than the inaccurate middle deciles;
- When good models were in the list of choices, all quality measures scored them high;
- The different Υ versions behaved similarly, but Υ_Q differentiates the segments clearer than Υ_L (see figures 3.12, 3.17 or 3.26);
- ROI^{act} also selected accurate models when they were available, and bolder models when accuracy was not present in the models available;
- ROI^{pred} was not useful in at least three out of eight datasets;

However, we need a more objective way to understand what the different measures quantify. We decided to use a ranking system to capture how the quality measures relate to the desirable characteristics. To do that we proceeded as follows. For each dataset (and list of models):

- We selected the different fourteen models that maximise each of the quality measures: $\Upsilon_L, \Upsilon_Q, q_0, M$, and ROI.
- We ranked those fourteen models based on the different desirable features (see section 1.6) accounting for ties. Higher rank values indicate that the quality measure scores better for that characteristic.

This procedure returned a table with one quality measures per row, the desirable characteristics as columns and the rank in each cell (a number from one to fourteen).

We have done that for the 18 datasets in table 2.2 (each of which was divided into training and test set) and averaged the ranks for all datasets. We present the data separating test and training. This separation is done in order to identify whether the absence of accurate models changes the behaviour of the quality measures. Ranking the quality measurements implies loss of information, as we do not know the differences in the values ranked, but it allows comparing

the data obtained from different datasets. It would make no sense to compare the sum of errors for models in different datasets. It also allows using Mann-Whitney (MW) for testing whether the ranking difference between any two of the quality measures is statistically significant (see appendix A).

We have repeated the procedure above five times for five different training-test splits. The results were consistent for each individual table. We present the average of those five tables in 3.9 and 3.10, using a grey scale to indicate higher values. The results are nearly the same in both tables, and they agree with what we have seen so far in the subjective analysis of the models.

Υ measures are the best ones for detecting accurate models, together with M_5. Interestingly, M_5 does not use the sum of errors but the slope of the linear regression between predicted and actual uplift. Υ_L scores in the lower end for boldness but Υ_Q scores above the average in that respect. One can have confidence in the boldness reported by Υ_Q because both the subjective assessment and the ranking table report Υ_Q as metrics that capture accuracy. Bold predictions from M_1, M_2 and M_3 are more questionable as they score low on accuracy.

The strategy we used is in some aspects questionable; we defined measurements that identify the desirable characteristics, built composite quality measures (M) based on those measurements and assessed how they performed using the measurements. However, the important observation that validated the strategy is that Υ_Q performed better than M metrics in some aspects (accuracy, predicted 1^{st} uplift and spread) and close to average in others (U_{max} and actual 1^{st} uplift). q_0 appears slightly below the average rank in all respects.

	Maximum uplift (U_{max})	Accuracy ($-\varepsilon$)	Monotonicity (r_s)	Actual uplift at 1^{st} decile (u_{1p})	Predicted uplift at 1^{st} decile (u_{1a})	Negative effect ($-\nu_p$)	Spread (s_p)
Υ_L^{min}	4.8	12.1	8.8	5	6.5	7.7	5.3
Υ_L^{mean}	5	11.7	8.5	5.1	7.1	7.8	6
Υ_L^{s}	5	12	8.7	5.1	6.8	7.9	5.7
Υ_Q^{min}	6.9	9.5	7.3	7	9.2	7.6	8.2
Υ_Q^{mean}	7.5	8.9	7.9	7.6	10	7.5	9.5
Υ_Q^{s}	6.5	10.1	7.7	6.7	9	7.6	8
q_0	5.7	4.6	5.8	5.6	4.9	7.1	6.5
ROI^{pred}	6.2	4.5	7.2	6.1	13.9	7.9	12.1
ROI^{act}	13.4	3	5.6	13.4	6.5	7.7	6.4
M_1	13.1	3.9	5.9	13.1	6.9	7.5	6.6
M_2	11.8	5.9	10	11.8	7.2	7.4	6.7
M_3	8.7	5.5	9	8.2	6.9	7.5	11.1
M_4	5.3	5	5.6	5.2	5	7.3	6.8
M_5	5	8.3	7	5	5.2	6.7	6.3

Table 3.9: Average ranking for each quality measure with regards the desirable characteristics, which are quantified using $U_{max}, \varepsilon, r_s, u_{1a}, u_{1p}, \nu_p$ and s_p. *Higher* values indicate *more* of the desirable characteristic. The cells are coloured with a grey scale based on its the value. Accuracy is quantified as the inverse of ε. The average is taken from 90 different *train* subsets (eighteen datasets with five different training-test split each). Accuracy and monotonicity must be taken into account to have any confidence on the rest of the characteristics. Υ family scores high in accuracy, with Υ_L ranking higher than Υ_Q on that respect. Υ_Q scores over q_0 for U_{max}, s_p, u_{1a}. Most of the M family scores higher than q_0 in accuracy, particularly M_5 and M_2

	Maximum uplift (U_{max})	Accuracy ($-\varepsilon$)	Monotonicity (r_s)	Actual uplift at 1^{st} decile (u_{1p})	Predicted uplift at 1^{st} decile (u_{1a})	Negative effect ($-\nu_p$)	Spread (s_p)
Υ_L^{min}	5.5	11.7	7.2	5.7	6.5	7.8	5.5
Υ_L^{mean}	5.2	11.3	8	5.5	6.7	7.9	5.5
Υ_L^{s}	5	11.5	7.7	5.2	6.6	7.8	5.5
Υ_Q^{min}	7.4	9.6	8.6	7.3	9.1	7.5	8.2
Υ_Q^{mean}	6.7	9.1	8.7	6.8	9.5	7.6	9.2
Υ_Q^{s}	6.8	10.4	7.3	6.9	8.6	7.6	7.4
q_0	6.4	5	6	6.2	5.2	6.6	6.7
ROI^{pred}	5.1	4.2	7.2	5.4	14	8.6	12.7
ROI^{act}	13.3	3	5.7	13.3	6.6	7.7	6.7
M_1	13.1	4	5.4	13.1	7	7.6	6.7
M_2	10.5	6.2	10.8	10.5	7.3	7.4	7
M_3	7.5	5.1	9.7	6.8	7	7.7	10.9
M_4	6.5	5.8	6.1	6.4	5.2	6.7	6.1
M_5	5.9	8.1	6.6	6	5.7	6.6	7

Table 3.10: The *test* subset presents nearly the same colour scheme (see table 3.9). As in that table *higher* values indicate *more* of the desirable characteristic The Υ family of quality measures are the best in detecting accuracy and Υ_Q version ranks above or close to the average for qualities related to bold models.

4 Conclusion

We have answered the question of the validity of Υ as a metric for uplift models, identifying its capacity for earmarking a useful model from a given list.

To do that we first had to understand how one judges whether an uplift model is useful. Therefore we reviewed what practitioners considered a valid uplift model and stated the specific desirable characteristics they should have:

1. Tight prediction values;
2. Monotonicity;
3. Spread or range of predictions;
4. Uplift at cut-off (using the uplift at 1^{st} decile);
5. Maximum uplift;

We presented the tools, namely uplift by decile graph and Qini curve, used to judge those specific characteristics. We also established the criteria to quantify those characteristics, whenever they were not already quantified. That was required for characteristic 1 and 2 above.

Second, we required a rich group of datasets to generate uplift models for. Note that uplift models can not be applied to any kind of datasets, as it has to have a specific structure: a control and treated set that have been randomly allocated. Fortunately, we had access to a real world dataset, which effectively contained four workable datasets for generating uplift models, and simulate fourteen extra datasets. The lack of more actual datasets is a weak point in our research and the results could be dependant on the specifics of the dataset we used. However, we are confident that our synthetic datasets covered a comprehensive range of possible scenarios for uplift problems.

Third, we generated multiple models on those datasets and put our quality measure to test. Our model generation approach did not cover all current techniques, but that was not a problem, as the models were sufficiently valid (and invalid) to expose the characteristics of uplift models. Using other modelling techniques would have not added significant variety to our models.

In parallel to the above steps, we analyzed in detail the initial formulation for Υ (eq. 1.16), and identified issues on it. Therefore we proposed an alternative formulation, which led us to define a family of potential quality measures: Υ_L and Υ_Q. Both the Υ_L and Υ_Q family depend on a parameter. This parameter can be set in a way that model accuracy is considered more or less important. The Υ_Q family includes another parameter, which can be set to grant more weight to the accuracy of the predictions in the first deciles. In addition, we defined composite quality measures (M family) based on the desirable characteristics.

We used multiple quality measures ($\Upsilon_L, \Upsilon_Q, q_0, M$ and ROI) to select the best model from all the models generated on the different datasets. The models selected maximized each quality measure. We assumed that higher values of the quality measure meant better models. This is in fact a way to confirm the validity of a quality measure. We subjectively assessed the models selected and concluded that:

- Increasing values of the quality measures from the Υ and M families led to models that could not be considered worse than models with lower values for that quality measure;
- When accurate and bold models were in the list of choices, all quality measures from the Υ and M scored them high;
- When accurate models were *not* in the list of choices, Υ family selected models where all segments were closer to the overall uplift, but still differentiating the segments (see figures 3.16, 3.22 or 3.24);
- M_3 (definition 2.3) was not selecting valid models because of the ratio s_p/u_{1p};
- Υ_Q^{mean} performance was worse than the rest of the Υ family because it gives more weight to contributions from accurate first deciles than very inaccurate middle and last deciles (see figure 3.17). If one expects Υ_Q metric to be sensitive to the precision of middle decile predictions, it is better to reduce the range of positive contribution using Υ_Q^{min} or Υ_{Qc}^{s} with $c < 3$ versions;
- The different Υ versions behaved similarly, but Υ_Q differentiates the segments clearer than Υ_L does;

In order to remove the subjectivity in the assessment of the quality measures, we ranked them based on each of the desirable characteristics. We used a total of ninety datasets to average the rank. The ranking order marked the Υ_L family as the metric that detects more accurate predictions. Υ_Q ranked just below Υ_L in that respect. Υ_Q compromises accuracy and boldness, as it ranks above the median for predicted 1^{st} uplift and spread, and close to the median for maximum uplift and actual 1^{st} uplift.

The ranking system proposed, although useful for having a comparative view of the metrics performance, is only a reassurance on the subjective assessment done beforehand. Without the subjective assessment, the results from the ranking would be less relevant. This is because there is loss of information when you rank magnitudes, and one can not be sure about the relevance of the difference, even when there is a wide rank range.

Although Qini coefficient has not been our benchmark, it is unavoidable to use it as a reference. q_0 did not perform better than the Υ family in the subjective assessment of the models, and in

some cases it performed clearly worse, (see figures 3.24, 3.22 or 2.3 for examples). Moreover q_0 ranked below Υ_Q for maximum uplift, actual and predicted uplift at 1^{st} decile, and spread. We can conclude that the Υ_Q family of metrics is an improvement over the most commonly used q_0, as it accounts for accuracy and praises boldness, as long as one of the following versions is used: Υ_Q^{min} or Υ_{Qc}^{s} with $c < 3$.

The composite measurements proposed did not perform as initially expected. Being directly related with the overall error was assumed to be an advantage, however it did not mark more accurate models than the ones selected by Υ. Nevertheless, we believe that there is scope for improving the formulation of composite measures, and maybe combining them with Υ is a direction for future research.

Appendices

A Mann-Whitney Between Qini and Υ

The Mann-Whitney test between Υ and q_0 is presented below:

- Null hypothesis is that the distributions of Υ and q_0 differ by a location shift of 0, that is the medians are similar.
- Alternative is that Υ median is greater than the median of q_0).

The difference between Υ_L and q_0 ranking order is not significant for U_{max}, u_{1a} and s_p. For the Υ_Q, the ranking order is significantly different with q_0 for all desirable characteristics except ν. Υ_Q is clearly superior to q_0, as it detects more accurate and bold models.

	median Υ_L^{min}	median q_0	U statistic	p-value	confidence interval
U_{max}	4.5	5.5	3486	0.9473	(-1.5, ∞)
$-\varepsilon$	12.5	4.25	7938	0	(7.5, ∞)
r_s	9.75	5	5973	0	(2, ∞)
u_{1a}	5	5.25	3674.5	0.8595	(-1, ∞)
u_{1p}	6.25	3	5318	0.0001	(1, ∞)
$-\nu$	7.5	7.5	4543.5	0.0617	(0, ∞)
s_p	5	6.25	3456.5	0.9557	(-2, ∞)
	Υ_L^{mean}	q_0	U statistic	p-value	confidence interval
U_{max}	5	5.5	3596.5	0.9035	(-1.5, ∞)
$-\varepsilon$	12	4.25	7869.5	0	(7, ∞)
r_s	9.25	5	5793	0	(2, ∞)
u_{1a}	5	5.25	3774	0.786	(-1, ∞)
u_{1p}	7.25	3	5636.5	0	(2, ∞)
$-\nu$	7.5	7.5	4618.5	0.0379	(0, ∞)
s_p	5.5	6.25	3919	0.6468	(-1.5, ∞)
	Υ_L^s	q_0	U statistic	p-value	confidence interval
U_{max}	5	5.5	3624.5	0.889	(-1.5, ∞)
$-\varepsilon$	12.5	4.25	7931	0	(7.5, ∞)
r_s	9.75	5	5942.5	0	(2, ∞)
u_{1a}	5	5.25	3775.5	0.7848	(-1, ∞)
u_{1p}	6.5	3	5460.5	0	(1.5, ∞)
$-\nu$	7.5	7.5	4697.5	0.0216	(0, ∞)
s_p	5.5	6.25	3699.5	0.8428	(-2, ∞)

	median Υ_Q^{min}	median q_0	U statistic	p-value	confidence interval
U_{max}	7.5	5.5	4981.5	0.0038	(0.5, ∞)
$-\varepsilon$	9.5	4.25	7312.5	0	(4.5, ∞)
r_s	8	5	5028	0.0025	(0.5, ∞)
u_{1a}	7.5	5.25	5119	0.0011	(0.5, ∞)
u_{1p}	10	3	6630	0	(4, ∞)
$-\nu$	7.5	7.5	4423.5	0.1211	(0, ∞)
s_p	8	6.25	5008	0.003	(0.5, ∞)
	Υ_Q^{mean}	q_0	U statistic	p-value	confidence interval
U_{max}	8	5.5	5338.5	0.0001	(1, ∞)
$-\varepsilon$	9	4.25	7101	0	(4, ∞)
r_s	8	5	5379.5	0.0001	(1.5, ∞)
u_{1a}	8	5.25	5472.5	0	(1, ∞)
u_{1p}	10.25	3	6970.5	0	(5, ∞)
$-\nu$	7.5	7.5	4343	0.1795	(0, ∞)
s_p	10	6.25	5701	0	(2, ∞)
	Υ_Q^s	q_0	U statistic	p-value	confidence interval
U_{max}	6.5	5.5	4723	0.027	(0, ∞)
$-\varepsilon$	10.5	4.25	7524.5	0	(5.5, ∞)
r_s	8	5	5322.5	0.0001	(1, ∞)
u_{1a}	7.5	5.25	4891.5	0.008	(0.5, ∞)
u_{1p}	9.25	3	6585	0	(4, ∞)
$-\nu$	7.5	7.5	4500	0.0794	(0, ∞)
s_p	7.5	6.25	4931.5	0.0058	(0.5, ∞)

References

[1] M. Allen. *Direct Marketing.* Marketing in Action Series. Kogan Page, 1997.

[2] M. Aly. Survey on multiclass classification methods, 2005.

[3] B. R. Behram J. Hansotia. Direct marketing for multichannel retailers: Issues, challenges and solutions. *Journal of Database Marketing*, 9:259266, Jan. 2002.

[4] D. R. Bell, J. Chiang, and V. Padmanabhan. The decomposition of promotional response: An empirical generalization. *Marketing Science*, 18:504–526, 1999.

[5] L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123–140, Aug. 1996.

[6] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, 1984.

[7] D. Collett. *Modelling Survival Data in Medical Research, Second Edition.* Texts in statistical science. Taylor & Francis, 2003.

[8] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, Dec. 2006.

[9] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1923, 1998.

[10] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[11] F. Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15:246–263, 1886.

[12] D. J. Hand and R. J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.*, 45(2):171–186, Oct. 2001.

[13] J. J. Heckman, H. Ichimura, and P. E. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4):605–54, 1997.

[14] K. Hillstrom. The minethatdata e-mail analytics and data mining challenge, 2008. http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html.

[15] A. K. Jain and R. C. Dubes. *Algorithms for clustering data.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[16] K. Larsen. Flexible and practical modeling with the gnbc: A case study. SAS Workshop, oct 2006.

[17] V. S. Y. Lo. The true lift model: a novel data mining approach to response modeling in database marketing. *SIGKDD Explor. Newsl.*, 4(2):78–86, Dec. 2002.

[18] M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905.

[19] E. C. Malthouse. Ridge regression and direct marketing scoring models. *Journal of Interactive Marketing*, 13(4), 1999.

[20] S. Moro, R. Laureano, and P. Cortez. Using data mining for bank direct marketing: An application of the crisp-dm methodology. In P. N. et al., editor, *Proceedings of the European Simulation and Modelling Conference - ESM'2011*, pages 117–121, Guimaraes, Portugal, Oct. 2011. EUROSIS.

[21] G. Piatetsky-Shapiro and B. Masand. Estimating campaign benefits and modeling lift. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 185–193, New York, NY, USA, 1999. ACM.

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[23] N. J. Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal, An Annual Publication from the Direct Marketing Association Analytics Council*, pages 14–21, 2007.

[24] N. J. Radcliffe. Moment of uplift, version 2. Technical Report Technical Note, Stochastic Solutions Limited, August 2012.

[25] N. J. Radcliffe and R. Simpson. Identifying who can be saved and who will be driven away by retention activity. *Journal of Telecommunications Management. Henry Stewart Publications*, (to appear), 2008.

[26] N. J. Radcliffe and P. D. Surry. Quality measures for uplift models. manuscript.

[27] N. J. Radcliffe and P. D. Surry. Real-world uplift modelling with significance-based uplift trees. Technical Report Portrait Technical Report TR2011-1, Stochastic Solutions & Portrait Software, 2011.

[28] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[29] A. Raveha. Gini correlation as a measure of monotonicity and two of its usages. *Communications in Statistics - Theory and Methods*, 18(4):pages 1415–1423, 1989.

[30] P. Rosenbaum, D. Rubin, and W. C. C. C. Biostatistics. *The Central Role of the Propensity Score in Observational Studies for Causal Effects.* Technical report (Wisconsin Clinical Cancer Center. Biostatistics). Wisconsin Clinical Cancer Center, Biostatistics, 1982.

[31] D. B. Rubin. Bias reduction using Mahalanobis-metric matching (corr: V36 p752). *Biometrics*, 36:293–298, 1980.

[32] P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 441–450, Washington, DC, USA, 2010. IEEE Computer Society.

[33] P. Software. Optimal targeting through uplift modeling: Generating higher demand and increasing customer retention while reducing marketing costs. Technical Report Portrait White Paper, Portrait Software, 2011.

[34] J. K. Vermunt and J. Magidson. Latent class models for classification. *Computational Statistics & Data Analysis*, 41(3-4):531–537, 2003.