Paper 350-2012

Including the Salesperson Effect in Purchasing Behavior Models Using PROC GLIMMIX

Philippe Baecke
Faculty of Economics and Business Administration, Department of Marketing, Ghent University, Belgium

Dirk Van den Poel
Faculty of Economics and Business Administration, Department of Marketing, Ghent University, Belgium

ABSTRACT

Nowadays, an increasing number of information technology tools are implemented in order to support decision making about marketing strategies and improve customer relationship management (CRM). Consequently, an improvement in CRM can be obtained by enhancing the databases on which these information technology tools are based. This study shows that a salesperson's personal attitudinal and behavioral characteristics can have an important impact on his sales performance. This salesperson effect can be easily included by means of a generalized linear mixed model using PROC GLIMMIX. This can significantly improve the predictive performance of a purchasing behavior model of a home vending company.

INTRODUCTION

In an increasingly competitive business environment, a successful company must provide customized services in order to gain a competitive advantage.[1] As a result, many firms have implemented information technology tools to customize marketing strategies in order to build up a long-term relationship with their clients.[2] This study will try to improve such customer relationship management (CRM) models by taking the salesperson effect into account.

Traditional CRM models are typically based on variables related to the individual such as socio-demographics, lifestyle variables and the individual past purchasing behavior of the customer. This study suggests that the purchasing behavior of a particular customer can also depend on social surroundings that have an influence during the purchase occasion. In a home vending environment the most important social surrounding is the interaction between the customer and the salesperson. A salesperson's personal attitudinal and behavioral characteristics have an important impact on his sales performance.[3] because a home vending company decides in advance which salesperson will visit which customer at what time. This makes it possible to already include this knowledge in a highly dynamic model that scores the customers on a daily basis. Hence, PROC GLIMMIX in the SAS® 9.2 program is introduced to capture this effect. This procedure makes it possible to estimate a generalized linear mixed model (i.e. a multilevel model) with a binomial outcome variable.

This study will investigate whether data augmentation with the salesperson effect will result in better purchasing behavior prediction. These predictions generated daily can be used for several applications. For example, when the demand is too high to visit every client, these predictions can help to select the most profitable ones. On the other hand, in a situation of overcapacity the salesperson has extra time left, in this situation the predicted probabilities can be used to generate revisit suggestions of the most profitable clients that were not home during the first visit.

METHODOLOGY

DATA DESCRIPTION

For this study, data is collected from a large home vending company, specialized in frozen foods and ice cream. This company uses about 180 salespeople to distribute their products to approximately 160,000 clients, visited on a regular basis in a biweekly schedule. Transactional data is used from February 1st, 2007 to November 30th, 2007 to build and validate the model. The same period in 2008 is used for out-of-period testing. Because a lot of promotional activities take place during the holiday period of Christmas and New Year, the months December and January are excluded and should be scored with a different model.

The data from the home vending company has been captured in explanatory variables. In Table 1, an overview of all variables used in this study can be found. The purpose of the proposed model is predicting whether a customer will buy at least one product conditional on him/her being at home. Therefore, only observations where the customer is at home are retained in the model. In a next step, this model can be combined with a second model predicting the probability a client will be at home, but this is beyond the scope of this research. In order to avoid correlation between purchase occasions of the same customer, only one visit per customer is randomly selected. If the customer was at home during the visit, (s)he bought at least one product in 46% of the purchase occasions. This signifies that the analysis table for this study is rather equally balanced between events and non-events.

Variable name	Description
Dependent variable:	
Sales	A binary variable indicating whether the customer purchased at least one product
Independent variables:	
Transactional variables:	
Recency visit	The number of days since the last visit
Recency bought	The number of days since the last purchase
Frequency visit	The number of visits in the last 8 weeks
Frequency bought	The number of purchases in the last 8 weeks
Monetary value	Total monetary value spent in the last 8 weeks
Sales ratio	The percentage of purchases based on all visits in the last 8 weeks
Avg. monetary value	The average amount spent per visit
Last time visit	A binary variable indicating whether the customer was visited in the last 21 days
Last time bought	A binary variable indicating whether the customer purchased at least one good at the last visit within 21 days
Last time amount	The amount spent on the last visit within 21 days
Sales person variables:	
Salesperson	A categorical variable indicating the sales person

Table 1. Model variables

As independent variables several traditional variables are created based on historical transactional information of the individual customer. Based on these variables a basic model will be constructed as benchmark model. Though, this study suggests that because every one of the 175 salespeople in the model has unique attitudinal and behavioral characteristics, correlation between the outcomes of the purchase occasions with the same salesperson can be expected. Therefore, a multilevel model is introduced to capture this effect. Next, the results of this model will be compared with the benchmark model in terms of predictive performance.

CLASSIFICATION TECHNIQUES

Modeling whether a visited customer will purchase at least one product, results in a binary classification problem. This paragraph introduces two statistical techniques used throughout this study that are able to handle such problems. The basic model is based on logistic regression techniques whereas a multilevel model is introduced to capture the salesperson effect.

LOGISTIC REGRESSION MODEL

Logistic regression is a well-known technique frequently used in traditional marketing applications.[4] An important benefit over other methods (e.g. neural networks) is its interpretability. It produces specific information about the size and direction of the effects of independent variables. Moreover, in terms of predictive performance and robustness, logistic regression can compete with more advanced data mining techniques.[5] Logistic regression belongs to the

group of generalized linear models (GLM). GLMs adopt ordinary least square regression to other response variables, like dichotomous outcomes, by using a link function[6]. In logistic regression the parameters are estimated by maximizing the log-likelihood function. Including these estimates in the following formulae creates probabilities, ranging from 0 to 1, that can be used to rank customers in terms of their likelihood of purchase.[7]

$$\pi_i = \frac{e^{\eta}}{1+e^{\eta}} \quad (1)$$

$$\eta = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} \quad (2)$$

Whereby: π_i represents the *a posteriori* probability of purchase by customer i; X_{nj} represents the independent variables for customer i; β_0 represents the intercept; β_n represent the parameters to be estimated; n represents the number of independent variables.

Due to the high correlation between independent variables, it is possible that some variables, although significant in a univariate relationship, have little extra predictive value to add to the model. Hence, this study will include a backward selection technique that creates a subset of the original variables by eliminating variables that are either redundant or possess little additional predictive information. This should enhance the comprehensibility of the model and decrease the computation time and cost, which is very important in a highly dynamic model that must be scored on a daily basis.[8]

The SAS code used to estimate such a logistic regression model is shown below:

```
PROC LOGISTIC DATA = inputtable_train OUTMODEL = parest_train;
MODEL sales (EVENT='1') = &indepvars.
/SELECTION = backward SLSTAY = 0.01 STB;
OUTPUT OUT = predlog_train P = sales_pred;
ODS OUTPUT parameterestimates = log_paramest;
RUN;
```

PROC LOGISTIC is specifically designed for a logistic regression model. The procedure estimates parameters by means of maximum likelihood for a model with a binary dependent variable. The OUTMODEL option specifies the name of the data set containing sufficient information to score new data without having to refit the model. In the MODEL statement, the variable sales is defined as dependent variable and has to be modeled using a macro list of independent variables. In the SELECTION option the backward selection method is specified based on a significance level of 0.01. The STB option adds standardized parameter estimates to the output. The OUTPUT OUT option creates a new dataset, called predlog_train, identical to the input dataset but with an extra column containing the predicted sales probabilities. The parameter estimates of the model are saved using the ODS OUTPUT statement.

Next, based on this model, prediction for the validation sample and the out-of-time test sample can be made using the following code:

```
PROC LOGISTIC INMODEL = parest_train;
SCORE DATA = inputtable_val OUT= predlog_val (rename = (p_1 = sales_pred));
RUN;
```

This code uses the information from the parest_train dataset to make estimations based on the dataset defined in the DATA option. These predictions are saved in the dataset defined in the OUT option.

MULTILEVEL MODEL

Originally, multilevel or hierarchical models were often used in research disciplines as sociology to analyze a population structured hierarchically in groups or clusters. For example, in Ref 9 students on the lowest level are nested within schools on a higher level. In such samples, the individual observations are often not completely independent. As a result, the average correlation between variables measured on observations within the same group will be higher than the average correlation between variables measured based on observations from different groups. Standard statistical techniques, such as logistic regression, rely heavily on the assumption of independence of observations and a violation of this assumption can have a significant influence on the accuracy of the model.[10] In this study it is expected that due to the differences in personal attitudinal and behavioral characteristics between salespeople, purchase occasions of the same salesperson will have a higher correlation than average. In other words, purchase occasions can be nested within salespeople.

There are several ways to extend a single-level model to a multilevel model. The easiest way to take the effects of higher-level units into account is by adding dummy variables so that each higher-level unit has its own intercept in the model. These dummy variables can be used to measure the differences between salespeople. The use of fixed intercepts, however, increases the number of additional parameters equal to the number of higher-level units minus one. Because this study includes 175 salespeople, this would result in a large number of nuisance parameters in the model. A more sophisticated approach is to treat the salesperson intercepts as a random variable with a specified probability distribution in a multilevel model. This method will lead to more accurate predictions.

Assuming that data is available from J groups with a different number of observations n_j in each group, a multilevel model can be estimated based on the following equation:[10]

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \quad (3)$$

In this equation, Y_{ij} and X_{ij} represent the dependent and one (or more) independent variables at the lowest level respectively. The residual errors e_{ij} are assumed to be normally distributed with a mean of zero and a variance, denoted by σ_e^2, that has to be estimated. The intercept and slope coefficients, β_{0j} and β_{1j} respectively, are assumed to vary across the groups. These coefficients, often called random coefficients, have a distribution with a certain mean and variance that can be explained by one or more independent variables at the highest level Z_j, as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \quad (4)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_j + u_{1j} \quad (5)$$

The u-terms u_{0j} and u_{1j} represent the random residual errors at the highest level and are assumed to be independent from the residual errors e_{ij} at the lowest level and normally distributed with a mean of zero and a variance of $\sigma_{u_0}^2$ and $\sigma_{u_1}^2$ respectively. The covariance between the residual error terms u_{0j} and u_{1j}, denoted as $\sigma_{u_{01}}^2$, is generally not assumed to be zero.

By substituting “Eq. (4)” and “Eq. (5)” into equation “Eq. (3)” and rearranging terms, a single complex multilevel equation is created:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + u_{1j}X_{ij} + u_{0j} + e_{ij} \quad (6)$$

This model can be split into a fixed or deterministic part $[\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij}]$ and a random or stochastic part $[u_{0j} + u_{1j}X_{ij} + e_{ij}]$. This illustrates that, in order to allow correlation between the observations, the generalized linear model (GLM) must be extended to a generalized linear mixed model (GLMM) with random effects that are assumed to be normally distributed.

In our study the dependent variable at the lowest level is the outcome whether the client purchased at least one product during the purchase occasion. Because this is a dichotomous variable, “Eq. (6)” needs to be transformed using a logit link function in the following way:[10]

$$Y_{ij} = \pi_{ij} \text{ ; } \pi \sim \text{Binomial}(n_{ij}, \mu) \quad (7)$$

$$\pi_{ij} = \text{logistic}(\gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + \gamma_{11}Z_jX_{ij} + u_{1j}X_{ij} + u_{0j}) \quad (8)$$

These equations state that the dependent variable is a proportion π_{ij}, assuming to have a binomial error distribution with sample size n_{ij} and expected value μ. If all possible outcomes are only zero and one, the sample sizes are reduced to one and dichotomous data is modeled. Due to the binomial distribution, the lowest-level residual variance is a function of the proportion:

$$\sigma_e^2 = \frac{\pi_{ij}}{1-\pi_{ij}} \quad (9)$$

Consequently, this variance does not have to be estimated separately and the lowest-level residual errors e_{ij} can be excluded from the equation. In Table 2 a summarized comparison between a logistic regression model and a logistic multilevel model can be found.

	Logistic regression model	Logistic multilevel model
Model family:	Generalized linear model (GLM)	Generalized linear mixed model (GLMM)
Regression equation:	$Y_i = \beta_0 + \beta_1 X_i + e_i$	$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$ $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$
Link function for dichotomous outcomes:	$\pi_i = \frac{e^{\eta}}{1+e^{\eta}}$	$\pi_i = \frac{e^{\eta}}{1+e^{\eta}}$
Correlation between observations:	Not assumed	Allowed
Relationship between dependent and independent variables:	Assumed to be linear	Assumed to be linear

Table 2. Comparison between a logistic regression model and a logistic multilevel model

The database from this study does not contain meaningful higher-level information about the salespeople. Furthermore, it is not expected that the slopes of any of the lower-level variables will vary across the salespeople. This makes it possible to reduce "Eq. (8)" to:

$$\pi_{ij} = \text{logistic}(\beta_{0j} + \beta_1 X_{ij}) \quad (10)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (11)$$

Combining "Eq. (10)" and "Eq. (11)" results into:

$$\pi_{ij} = \text{logistic}(\gamma_{00} + \beta_1 X_{ij} + u_{0j}) \quad (12)$$

This hierarchical logistic regression model still contains a fixed part $[\gamma_{00} + \beta_1 X_{ij}]$ and a random part $[u_{0j}]$.

The intraclass correlation coefficient (ICC), which measures the proportion of variance in the outcome explained by the grouping structure, can be calculated using an intercept-only model. This model can be derived from "Eq. (8)" by excluding all explanatory variables, which results in the following equation:

$$\pi_{ij} = \text{logistic}(\gamma_{00} + u_{0j}) \quad (13)$$

The ICC is then calculated based on the following formula:[10]

$$\text{ICC} = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e} \quad (14)$$

Because the variance of a logistic distribution with scale factor 1 is $\pi^2/3 \approx 3.29$ in a hierarchical logistic regression model, this formula can be reformulated as:[10]

$$\text{ICC} = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \pi^2/3} \quad (15)$$

The SAS code used to estimate such a multilevel model is shown below:

```
PROC GLIMMIX DATA = inputtable METHOD= MSPL;
CLASS salesperson_id;
MODEL sales (EVENT = '1') = &indepvars.
/DIST = binary LINK = logit SOLUTION;
```

```
RANDOM intercept / SUBJECT = salesperson_id SOLUTION;
OUTPUT OUT = predtable pred(BLUP ILINK) = sales_pred;
ODS OUTPUT covparms = ml_covparamest parameterestimates = ml_paramest
solutionR = randeff;
RUN;
```

PROC GLIMMIX is a procedure recently developed by SAS in order to fit generalized linear mixed models. The input table contains one row for each purchase occasion and also includes the variable salesperson_id that assigns each purchase occasion to one of the 175 salespeople. This model is estimated using the maximum log-likelihood and the subject-specific expansion principles (METHOD = MSPL). The class statement includes all variables that are categorical. Obviously, salesperson_id, which is the group variable, is also categorical. Just like in PROC LOGISTIC, the variable sales is defined in the MODEL statement as dependent variable and has to be modeled using a macro list of fixed effects. In the options of the MODEL statement it is defined that the distribution of the outcome variable is binary and a logit link function should be used for transformation. The significance of the effects can be evaluated using a t-test, which will be provided with the parameter estimates using the SOLUTION option. The RANDOM statement specifies that the intercept can vary across the salespeople. The OUTPUT OUT option creates a new dataset, called predtable, identical to the input dataset but with an extra column containing the predicted values based on the fixed and random effects (BLUP option), mapped onto the probability scale (ILINK option). The covariance parameter estimates and the solutions for fixed and random effects are saved using the ODS OUTPUT statement.

EVALUATION CRITERION

In order to be able to evaluate the predictive performance of each model the database, containing 162,424 observations, is randomly split into two equal parts. The first part, called training sample, is used to estimate the model. Afterwards, this model is validated on the remaining 50% of observations. It is essential to evaluate the performance of the classifiers on a holdout validation sample in order to ensure that the training model can be generalized over all customers of the home vending company. The analysis table is generated based on transactional information during the period between February 1st, 2007 and November 30th, 2007. Besides the training and validation sample, also an out-of-period test sample is created based on the same period in 2008, containing 161,462 observations. Using the model trained on data of 2007, predictions are made for all observations in the out-of-period test sample. This makes it possible to check the evolution of the accuracy of the model over time. If the performance does not drop significantly, the model can be generalized not only over all customers of the home vending company, but also over different time periods.

The area under the receiver operating characteristic curve (AUC) is used as evaluation metric of the classifiers.[11] The advantage of an AUC in comparison with other evaluation metrics, like the percent correctly classified (PCC), is the fact that PCC is highly dependent on the chosen threshold that has to be determined to distinguish the predicted events from non-events. The calculation of the PCC is based on a ranking of customers according to their *a posteriori* probability of purchase. Depending on the context of the problem of the home vending company (e.g. the amount of the capacity problem) a cutoff value is chosen. All customers with an *a posteriori* probability of purchase higher than the cutoff are classified as buyers and will be visited. All customers with a lower likelihood of purchase are labeled as non-buyers. This classification can be summarized in a confusion matrix, displayed in Table 3.[12]

		Predicted status	
		Buyer	**Non-buyer**
True Value	**Buyer**	True Positive (TP)	False Negative (FN)
	Non-buyer	False Positive (FP)	True Negative (TN)

Table 3. Confusion matrix

Based on this matrix the percentage of correctly classified observations can be formulated as:[13]

$$PCC = \frac{TP+TN}{TP+TN+FP+FN} \quad (16)$$

Besides the PCC, the following meaningful measures can also be calculated:

$$Sensitivity = \frac{TP}{TP+FN} \quad (17)$$

$$Specificity = \frac{TN}{TN+FP} \quad (18)$$

Sensitivity represents the proportion of actual events that the model correctly predicts as events (i.e. the number of true positives divided by the total number of events). Specificity is defined as the proportion of non-events that are correctly identified (i.e. the number of true negatives divided by the total number of non-events). It is important to notice that all these measures give only an indication of the performance at the chosen cutoff. In reality, the chosen cutoff will vary depending on the context of the problem of the decision maker, hence an evaluation criterion independent of the chosen cutoff, such as the AUC, is preferred.

The receiver operating characteristic (ROC) curve is a two-dimensional graphical representation of sensitivity and one minus specificity for all possible cutoff values used (e.g. Fig. 1). The AUC measures the area under this curve and can be interpreted as the probability that a randomly chosen positive instance is correctly ranked higher than a randomly selected negative instance.[11] This again illustrates that this evaluation criterion is independent of the chosen threshold. As a result, this criterion is often used as evaluation metric for the predictive performance of CRM models (e.g. Ref. 14). The AUC measure can range from a lower limit of 0.5, if the predictions are random (corresponding with the diagonal in Fig. 1), to an upper limit of 1, if the model's predictions are perfect.

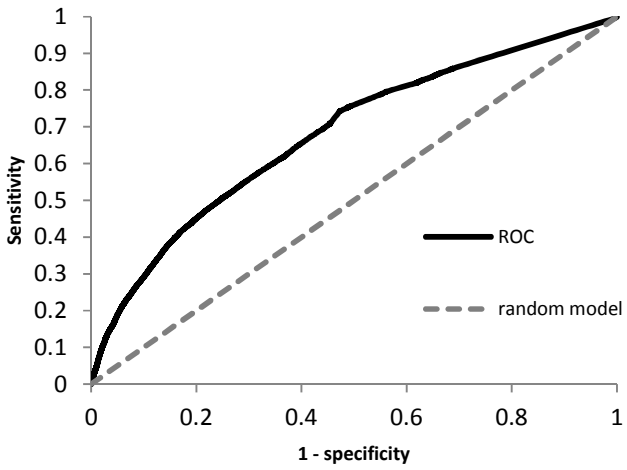

Fig. 1. AUC example

	Logistic regression model		Multilevel model (+ Salesperson)	
Variable	**Estimate**	**SE**	**Estimate**	**SE**
Intercept	-0.7425	0.0344	-0.6460	0.0447
Transactional variables:				
Recency visit	0.0062	0.0008	0.0023	0.0008
Frequency bought	0.4031	0.0184	0.4959	0.0194
Sales ratio	1.0153	0.0650	0.5762	0.0693
Avg. mon. value	0.0115	0.0021	0.0139	0.0021
Last time visit	-0.2697	0.0246	-0.2639	0.0255
Last time bought	-0.5984	0.0221	-0.6158	0.0223
Salesperson variables:				
Intercept variance ($\sigma^2_{u_0}$)			0.1208	0.0151

Table 4. Overview of the parameter estimates

RESULTS

The results of this study are clearly summarized in Table 4 and Table 5. Table 4 contains all parameter estimates of each model that are significance on a 0.01 confidence level. First, the basic model, based on only transactional data, will be discussed. Next, this model will be enhanced with the salesperson effect by means of a multilevel model. Because of the high number of observations, a significance level of 0.01 is preferred. In Table 5 the predictive performance, in terms of AUC, is displayed for the training, validation and out-of-period test sample.

Sample	Logistic regression model	Multilevel model (+ Salesperson)
Training sample	0.6793	0.7014
Validation sample	0.6801	0.6996
Out-of-period test sample	0.6818	0.6996

Table 5. Model performance measured in term of AUC

BASIC MODEL

A logistic regression model that only uses transactional variables in order to predict purchasing behavior will be used as benchmark model. Because of the backward selection technique, only six of the initial ten input variables are retained. High correlation between some of the transactional variables results in the fact that four variables do not add extra predictive value to the model. Having a closer look at the parameter estimates in Table 4 gives interesting insights into the purchasing pattern of the home vending company's customers. All significant variables based on the past purchasing behavior in the last eight weeks (i.e. frequency bought, sales ratio and average monetary value) have a positive relationship with the future purchasing behavior. On the other hand, the transactional variables based on the last visit (i.e. last time visit and last time bought) all have a negative relationship with the probability to purchase on a next visit. Normally, a customer is visited in a biweekly schedule. This means that, if there are no capacity problems, there are 14 days between visiting the same customer again. These parameter estimates imply that the most attractive customers have high RFM scores in general, but if the customer was visited at a normal frequency the last time and moreover bought a product, his/her probability of buying the next time will drop. Although, if a customer was not visited due to capacity problems for example, the dummy variables last time visit and last time bought will be flagged zero, as a result his/her probability to purchase next time will rise and the chance that (s)he will

be excluded again will decrease. This illustrates the usefulness of a dynamic model that ranks customers on a daily basis in order to ensure that, at every moment, priority is given to clients with the highest purchase probability. With an AUC of 0.6793, 0.6801 and 0.6818 on the training, validation and out-of-period test sample respectively (Table 5), this study confirms that variables about the past purchasing behavior are still good predictors for future purchasing behavior. Notwithstanding this relative good performance based on transactional data, improvement can still be obtained by incorporating the salesperson effect.

DATA AUGMENTATION WITH SALESPERSON VARIABLES

In order to take the effect of social surroundings into account, a multilevel model is introduced. In this study the most important social surrounding at the purchase occasion is the personal influence of one of the 175 salespeople. First, the intraclass correlation coefficient is calculated based on an intercept-only model without independent variables. In this model, the intercept variance ($\sigma_{u_0}^2$) was estimated to be 0.1716. Using formula (15), this results in an ICC of 0.0496, meaning that 4.96% of the variation in the purchasing behavior can be explained by grouping the customers based on the salespeople who visit them. In a next step, a multilevel model including all independent variables is estimated. Table 4 shows that in such a model, the intercept variance drops to 0.1208 due to the inclusion of independent transactional variables, but this value is still significant. In other words, taking into account the salesperson that visits each customer can provide extra predictive value on top of the traditional independent variables. Figure 2 represents the intercepts for each of the 175 salespeople estimated by the final multilevel model. The values are ranked from lowest on the left side to highest on the right side. This figure illustrates that attitudinal and behavioral difference between salespeople result in a significant variation in the ability to sell products. These intercepts could even be useful during the evaluation process of the salesperson, because it gives an idea of the salesperson's performance controlled for the individual characteristic of the customers within the portfolio of each salesperson.

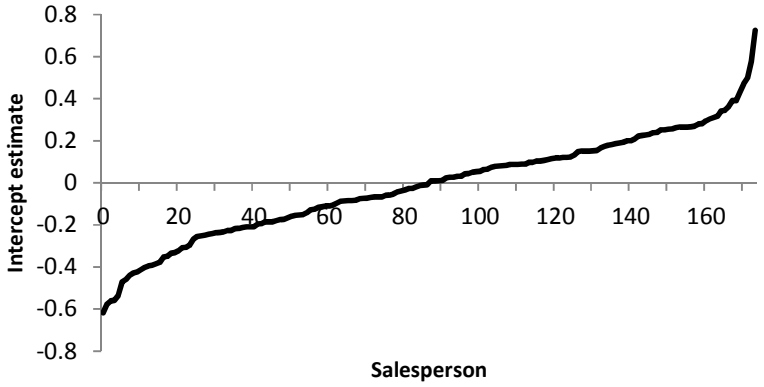

Fig. 2. Intercept estimates for each salesperson

Table 5 indicates that by structuring the purchase occasions by salesperson a strong increase in predictive performance can be obtained using the same transactional variables. On the training, validation and out of time test sample, this improvement is respectively 0.0221, 0.0195, 0.0178 which is not only statistical significant, but also economically relevant for the home vending company.

CONCLUSION

In order to remain competitive, a lot of firms implement information technology tools to improve their marketing strategies.[15, 16] Nowadays, an increasing number of software products are available to support decision making.[17] As a result, the company's database has become a valuable asset to support marketing decisions. This study shows that the predictors in CRM models should not only be restricted to variables that are related to the individual (e.g. the individual past purchasing behavior). Taking social surroundings such as the salesperson effect into account can already significantly improve purchasing behavior predictions. The PROC GLIMMIX procedure is an ideal SAS tool to construct a multilevel model that is able to incorporate this salesperson effect. This study demonstrates the added value of this procedure in a home vending context, but also in other industries where the salesperson plays an important role, a similar model can be implemented, such as real estate-agents, investment advisers, insurance agents, etc.

ACKNOWLEDGEMENT

Both authors acknowledge the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

REFERENCES

1. S. Lipovetsky, SURF - Structural Unduplicated Reach and Frequency: Latent class TURF and Shapley Value analyses, *Int. J. Inf. Technol. Decis. Mak.* **7**(2008) 203-216.
2. R. Ling, and D. C. Yen, Customer relationship management: An analysis framework and implementation strategies, *Journal of Computer Information Systems* **41**(2001) 82–97.
3. G. Albaum, Exploring Interaction in a Marketing Situation, *J. Mark. Res.* **4**(1967) 168-72.
4. R. E. Bucklin, and S. Gupta, Brand choice, purchase incidence and segmentation: An integrated modeling approach, *J. Mark. Res.* **29**(1992) 201–215.
5. N. Levin, and J. Zahavi, Continuous predictive modeling: A comparative analysis, *J. Interact. Mark.* **12**(1998) 5–22.
6. P. McCullagh and J. A. Nelder, *Generalized linear models (second edition)* (Chapman & Hall, London, 1989).
7. D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression (second edition) (*John Wiley & Sons, New York, 2000).
8. Y. S. Kim, Toward a successful CRM: Variable selection, sampling, and ensemble, *Decis. Support Syst.* **41**(2006) 542–553.
9. V. E. Lee, and A. S. Bryk, A multilevel model of the social distribution of high school achievements, *Sociol. Educ.* **62**(1989) 172-192.
10. J. Hox, *Multilevel Analysis: Techniques and Applications* (Taylor & Francis Group, New York, 2002).
11. J. A. Hanley, and B. J. McNeil, The meaning and use of area under a receiver operating characteristic (ROC) curve, *Radiology* **143**(1982) 29–36.
12. D. G. Morrison, On the interpretation of discriminant analysis, *J. Mark. Res.* **6**(1969) 156-163.
13. A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning
14. algorithms, *Pattern Recognit*. **7**(1997) 1145-1159.
15. S. Hill, F. Provost, and C. Volinsky, Network-based marketing: Identifying likely adopters via consumer networks, *Stat. Sci.* **21**(2006) 256-276.
16. C. T. Lin, C. Lee, and C. S. Wu, Fuzzy group decision making in pursuit of a competitive marketing strategy, *Int. J. Inf. Technol. Decis. Mak.* **9**(2010) 281-300.
17. M. Iwashita, K. Nishimatsu, T. Kurosawa, and S. Shimogawa, *The review of Socionetwork Strategies* **4**(2010) 17-28.
18. C. G. Sen, H. Baracli, and S. Sen, A literature review and classification of enterprise software selection approaches, *Int. J. Inf. Technol. Decis. Mak*. **8**(2009) 217-238.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Philippe Baecke
Faculty of Economics and Business Administration
Department of Marketing
Ghent University
Tweekerkenstraat 2
B-9000 Ghent
Belgium
E-mail : Philippe.Baecke@ugent.be

Dirk Vand den Poel
Faculty of Economics and Business Administration
Department of Marketing
Ghent University
Tweekerkenstraat 2
B-9000 Ghent

Belgium
E-mail : Dirk.VandenPoel@ugent.be
Website : http://www.crm.UGent.be