UNIVERSITEIT GENT

FACULTEIT ECONOMIE EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Random Multiclass Classification: Generalizing Random Forests to Random MNL and Random NB

Anita Prinzie [1]

Dirk Van den Poel [2]

June 2007

2007/469

[1] Postdoctoral Fellow of the Research Foundation, Flanders (FWO Vlaanderen), Department of Marketing, Ghent University
[2] Prof. Dr. Dirk Van den Poel, Department of Marketing, Ghent University

www.crm.ugent.be

D/2007/7012/40

Random Multiclass Classification: Generalizing Random Forests to Random MNL and Random NB

Anita Prinzie[1] and Dirk Van den Poel[1]

[1] Department of Marketing, Ghent University, Tweekerkenstraat 2, 9000 Ghent, Belgium, {Anita.Prinzie, Dirk.VandenPoel}@UGent.be}

Abstract. Random Forests (RF) is a successful classifier exhibiting performance comparable to Adaboost, but is more robust. The exploitation of two sources of randomness, random inputs (bagging) and random features, make RF accurate classifiers in several domains. We hypothesize that methods other than classification or regression trees could also benefit from injecting randomness. This paper generalizes the RF framework to other multiclass classification algorithms like the well-established MultiNomial Logit (MNL) and Naive Bayes (NB). We propose Random MNL (RMNL) as a new bagged classifier combining a forest of MNLs estimated with randomly selected features. Analogously, we introduce Random Naive Bayes (RNB). We benchmark the predictive performance of RF, RMNL and RNB against state-of-the-art SVM classifiers. RF, RMNL and RNB outperform SVM. Moreover, generalizing RF seems promising as reflected by the improved predictive performance of RMNL.

1 Introduction

Random Forests (RF), introduced by Breiman [Breiman] to augment the robustness of classification and regression trees, have been applied successfully in many domains. RF is a bagged classifier building a 'forest' of decision trees splitting at each node on the best feature out of a random subset of the feature space. Whereas bagging enhances the robustness of the original base classifier, random feature selection improves the accuracy in domains characterized by many input variables, with each one containing only a small amount of information. It seems logical, that methods other than decision trees, might also benefit from the exploitation of these two sources of randomness, i.e. random inputs (bagging) and random feature selection.

Firstly, most algorithms suffer the curse of dimensionality and therefore they will benefit from random feature selection. Given the susceptibility of many methods to Hughes phenomenon (on increasing the number of features as input to the classifier over a given threshold, the accuracy decreases) and the tendency towards huge input spaces, most algorithms benefit from *any* type of feature selection. However, as the dimensionality of the feature space grows, a complete search is infeasible. Hence, random feature selection might be an acceptable solution. Secondly, building an ensemble, e.g. bagging, typically results in significant improvements in accuracy.

Therefore, this paper investigates the performance improvement of classification algorithms by injecting randomness by adopting a randomized ensemble approach. We generalize the RF framework to two classification algorithms, MultiNomial Logit, a Random Utility (RU) model explaining unordered multiple choices using a random utility function, and naive Bayes, a probabilistic classifier simplifying Bayes' Theorem. We propose Random MNL (RMNL) as a new bagged classifier combining a forest of MNLs estimated with randomly selected features. Analogously, we introduce Random Naïve Bayes (RNB). The performance of RF, RMNL and RNB is benchmarked against state-of-the art SVMs [Vapnik]. We illustrate our RMNL and RNB on a multiclass classification problem; a cross-sell case. The results are promising as generalizing RF to MNL substantially improves predictive performance.

2 Methodology

2.1 Random Forests (RF)

Random Forests (RF) [Breiman] is a highly accurate machine-learning algorithm far more robust than decision trees and capable of modeling huge feature spaces. RF is a bagged classifier combining a collection of T classification or regression trees (i.e. forest of trees), here T classification trees. Each tree t is grown on a different bootstrap sample S_t containing randomly drawn instances with replacement from the original training sample. Besides bagging RF also employs random feature selection. At each node of the decision tree t, m features are selected at random out of the M features and the best split selected out of these m. Each decision tree is grown using CART methodology to the largest extent possible. An instance is classified into the class having the most votes of over all T trees in the forest, i.e. Majority Voting.
Breiman [Breiman] estimates the importance of each feature on *out-of-bag* (oob) data, cf. in each bootstrap sample about 1/e instances are left out. Randomly permute the feature m in the oob data and put the data down the corresponding tree. Subtract the number of votes for the correct class in the feature-m-permuted data from the number of correct votes in the untouched data and average over all trees T in the forest. This is the *raw importance score* for feature m from which the z-score is derived by dividing the raw score by its standard error.

The exploitation of randomness make RF accurate classifiers in several domains. While bagging increases stability of the original decision trees, the random feature selection enhances the 'noise' robustness, yielding error rates that compare even favorably to Adaboost [Freund and Shapire].

2.2 Random MultiNomial Logit (RMNL)

Within multinomial-discrete choice modeling [Ben-Akiva], RU models define a random utility function U_{ik} for each individual *i* for choice *k* belonging to choice set D_K with K > 2 (cf. multiclass). This random utility is decomposed into a deterministic and stochastic component (1):

$$\boldsymbol{U_{ik}} = \boldsymbol{\beta}' \boldsymbol{x_{ik}} + \boldsymbol{\varepsilon_{ik}} \quad (\mathbf{1})$$

where x is a matrix of observed attributes which might be choice (e.g. price of product) or individual specific (e.g. age of customer), β' is a vector of unobserved marginal utilities (parameters) and ε_{ik} is an unobserved random error term (i.e. disturbance term or stochastic component). Different assumptions on the error term of the random utility function U_{ik} give rise to different classes of models. In this paper, we apply the MultiNomial Logit (MNL, independent and i.i.d. disturbances). To date, the MultiNomial Logit (MNL) model is the most popular RU model due to its closed-form choice-probability solution [Baltas JBR 2001]. The probability of choosing an alternative *k* among K_i choices for individual *i* can be written as in (2). The classifier utilizes the maximum a posteriori (MAP) decision rule to predict the class for individual *i*. MNL exhibits great robustness but is susceptible to multicollinearity.

$$\boldsymbol{P_i}(\boldsymbol{k}) = \frac{\exp(\boldsymbol{x'_{ik}}\ \boldsymbol{\beta})}{\sum_{\boldsymbol{k}\in\boldsymbol{K}} \exp(\boldsymbol{x'_{ik}}\ \boldsymbol{\beta})} \quad (\mathbf{2})$$

We will estimate a MNL model incorporating all features. This model might serve as a benchmark for the Random MNL.

Just like the instable decision trees (cf. RF), even a robust classifier like MNL could benefit from injecting randomness by random input selection and random feature selection. Where decision trees performance improves mainly because bagging enhances stability, we hypothesize that MNLs performance will increase because random feature selection reduces the estimation bias due to multicollinearity. We prefer to combine this random feature selection with bagging as it can still improve the stability of an even robust base classifier like MNL. Therefore, inspired by RF, we propose Random MNL (RMNL) as a new bagged classifier combining a forest of R MNLs estimated with *m* randomly selected features on the *r*-th bootstrap sample. Firstly, just like RF builds *T* classification trees on bootstrap samples S_t, in RMNL each MNL *r* is estimated on a different bootstrap sample S_r containing randomly drawn instances with replacement from the original training sample. Secondly, this bagging is used in tandem with random feature selection. To classify an observation put the input vector 'down' the *R* MNLs in the 'forest'. Each MNL votes for its predicted class. Finally, unlike RF, we assess the predictive value of the bagged predictor using the adjusted Majority Vote (aMV) as each *r*th MNL delivers continuous outputs, i.e. posterior probabilities.

We utilize the *out-of-bag* (oob) to assess the feature importances [Breiman].

2.3 Random Naïve Bayes (RNB)

Naive Bayes (NB) is a probabilistic classifier simplifying Bayes' Theorem by *naively* assuming class conditional independence. Although this assumption leads to biased posterior probabilities ((3), Z is a scaling factor), the ordered probabilities of NB result in a classification performance comparable to that of classification trees and neural networks [Langley].

$$p\left(C|F_i,...,F_m\right)=\frac{1}{Z}p(C)\prod_{i=1}^{n}p\left(F_i|C\right) \quad (3)$$

Notwithstanding NB's popularity due to its simplicity combined with high accuracy and speed, its conditional independence assumption rarely holds. There are mainly two approaches [Zhang, Jiang and Su] to alleviate this naivity: 1) Selecting attribute subsets in which attributes are conditionally independent (cf. selective NB; [Langley and Sage]), or 2) Extending the structure of NB to represent attribute dependencies [AODE, Webb etal 2005]. We adopt the first approach and hypothesize that NB's performance might improve by random feature selection. Analogous to AODE, we build an *ensemble*, but unlike AODE, we combine zero-dependence classifiers. To decrease the variance of the ensemble, we build a bagged NB classifier. Hence generalizing RF to NB, Random Naive Bayes (RNB) is a bagged classifier combining a 'forest' of *B* NBs. Each *b*th NB is estimated on a bootstrap sample S_b with *m* randomly selected features. To classify an observation put the input vector 'down' the *B* NBs in the 'forest'. Each NB votes for its predicted class. Finally, unlike RF, we assess the predicted class of the ensemble by adjusted Majority Vote (aMV) as each *b*th NB delivers continuous posterior probabilities.
We estimate the importance of each feature on oob data [Breiman].
The predictive performance of RF, MNL, RMNL, NB, RNB and a multi-class one-against-one SVM [Vapnik] with RBF-kernel function is evaluated on a separate test set. Given the objective to classify cases correctly in all classes *K* and the small class imbalance, a weighted PCC (each class-specific PCC_k is weighted with the relative class frequency f_k,) [Prinzie] is more appropriate than a PCC [Morrison, Barandela et al]. Secondly, we benchmark the model's performance to the proportional chance criterion Cr_{pro} rather than the maximum chance criterion Cr_{max} [Morrison]. A final evaluation criterion is the Area Under the receiver Operating Curve (AUC) [Fawcett]. A multiclass AUC results from averaging *K* binary AUCs (one-against-all).

3 A CRM Cross-sell Application

The methodological framework is applied on scanner data of a major home-appliances retailer to analyze customers' cross-buying patterns in order to support cross-sell actions. The objective is to predict in what product category the customer will acquire his next durable. We partition the home-appliance product space into nine product categories Hence, $Y \in \{1, 2, \ldots, 9\}$, K=9. Y has prior distribution $f_1 = 9.73\%$,

f_2 = 10.45, f_3 = 20.49, f_4 = 12.64, f_5 = 11.70, f_6 = 9.74, f_7 = 8.67, f_8 = 8.13 and f_9 = 8.45. We randomly assigned 37,276 (N_1) customers to the estimation sample and 37,110 (N_2) customers to the test sample. For each customer we constructed a number of predictors X building a general customer profile (e.g. purchase profile, brand loyalty, socio-demographical information) as well as capturing sequential patterns in customer's purchase behavior (the order of acquisition of durables - ORDER, and the duration between purchase events - DURATION) [Prinzie].

4 Results

4.1 Random Forests (RF)

We estimated RF with 500 trees (default), balanced (higher weights for smaller classes, [Breiman]), on a range of *m* values starting from the square root of M (default); $m=441^{1/2}$. We engage in a grid search with main step size 1/3 of the default setting. Table 1 reports some of the results and shows the sensitivity of RF to *m*. On the estimation data, a balanced RF with 500 trees, *m*=336 delivers the best performance: wPCCe=21.04%, PCCe=21.67% and AUCe=0.6097.

4.2 MultiNomial Logit (MNL) and Random MNL (RMNL)

MNL. We estimated a MNL model with M non-choice specific parameters (89 corresponding to RF's 441). A stable solution was not obtained. Alternatively, adopting a wrapper approach, we firstly selected the best features within three types of covariates (general, purchase order and duration). We subsequently compared four MNL models: 1) General, 2) General and Order, 3) General and Duration and 4) General, Order and Duration. The third MNL model delivered the highest performance (wPCCe= 19.75, PCCe=22.00 with Cr_{pro}=12.28% and AUCe=0.5973).

RMNL with R=100 (RMNL_100). Initially, we combine 100 MNLs (R=100) estimated on 100 bootstrap samples with *m* ($m \leq M$) randomly selected features. We take the square root of M; $m=89^{\wedge 1/2}$ as default parameter setting and, subsequently, engage in a grid search with main step size 1/3 with *m* in [3, 84]. Unfortunately, MNL models with more than 48 dimensions failed to estimate (cf. multicollinearity). Table 2, R_MNL (R=100) gives an overview of the results. The highest performance is observed for *m*=48 (wPCCe=21.25, PCCe=26.87, AUCe=0.6491, Cr_{pro}=12.28%).

RMNL combining MNLs with 10% highest wPCC (RMNL_10). Combining only the MNLs for a given *m* with the 10% highest wPCCe might improve the accuracy

[Dietterich]. We refrain from evaluating a) combining the 10% with highest PCCe or AUCe, as the wPCCe is the main criterion and b) the sensitivity to the number of classifiers combined. Table 2, column 10_R_MNL reports that, analogous to RMNL_100 (R=100), the highest predictive performance is attained for m=48.

4.3 Naive Bayes (NB) and Random Naive Bayes (RNB)

NB. To assess the value of RNB, we benchmark it with a Laplace estimation of NB with all features M (441). We preprocessed numeric attributes by Fayad's and Irani's supervised discretization method [Fayad and Irani]. NB's wPCCe is comparable to MNLs (wPCCe= 19.74, PCCe=21.69, AUCe=0.5982).

RNB with B=100 (RNB_100). We investigate the value of injecting randomness for NB as an attempt to mitigate its class conditional independence assumption. We build an ensemble of 100 NBs (B=100) with m ($m \leq M$) randomly selected features. Similar to RF and RMNL, we engage in a grid search for m starting from the square root of M ($441^{1/2}$) with step size 1/3. Table 3, first column, reports some of the results. The highest performance is measured for RNB_100 with 42 (m) randomly selected features: wPCCe=19.83, PCCe=20.00, AUCe=0.6100.

RNB combining NBs with 10% highest wPCC (RNB_10). Analogous to RMNL, we address whether combining only the 10% best classifiers (based on wPCCe) of the ensemble might improve its accuracy [Dietterich 1997]. The results (Table 3, second column), show a maximum improved wPCCe (+ 0.74 pctp) for m=77, with corresponding PCCe (+ 0.62 pctp). Benchmarked against NB, an increase of almost 1 pctp is observed for RNB_10s AUCe with smaller comparable improvements for wPCCe and PCCe. Hence, generalizing RF to NB slightly enhances the accuracy of NB, but not as substantially as for generalizing RF to MNL.

4.4 Support Vector Machines (SVM)

We benchmark the performance RF, MNL, RMNL, NB and RNB against that of state-of-the SVM with RBF kernel [LIBSVM] determined by parameters (C,γ). Numerical attributes are scaled from [P1, P99] to the range [-1,+1]. Each SVM (C,γ) is estimated on the scaled training data omitting instances having at least one attribute's value outside the range [-1,+1]. The optimal (C,γ) is determined as the SVM with the highest cross-validation accuracy over 5 folds via parallel grid search. In a first step, using a coarse grid with $C=2^{-5}, 2^{-3}, \ldots, 2^{13}$ and $\gamma=2^{-15}, 2^{-13}, \ldots, 2^{3}$ (100 SVMs) we identified $(2^{7},2^{-11})$ as a "better" region on the grid. In a second step, we conducted a best-region-only grid search around $(2^{7},2^{-11})$ with $C=\{2^{6}, 2^{6.585}, 2^{7}, 2^{8}, 2^{8.585}\}$ and $\gamma=\{2^{-9.415}, 2^{-10}, 2^{-11}, 2^{-11.415}, 2^{-12}\}$ preserving the same difference between subsequent C/γ values. Taking the wPCCe as main criterion, the best predictive

performance is measured for (2^7, $2^{-11.415}$): wPCCe=18.66%, PCCe=25.01% and AUCe=0.6111.

4.5 Predictive Model Evaluation on Test Data

We assess the robustness of the results on the estimation sample by applying the best RF (*m*=336, balanced), the best MNL (general and order), the best RMNL (*m*=48, RMNL_10), the original NB, the best RNB (*m*=77, RNB_10) and the best SVM (2^7,$2^{-11.415}$) on a separate test sample (N_2=37,110). For SVM we first estimate the SVM (2^7,$2^{-11.415}$) on the full scaled training data (not cross validation data) excluding instances with at least one attribute with value outside [-1,+1] and applied this SVM on the full scaled test sample including instances with values outside [-1,+1].

Table 4 and Fig. 1 clearly corroborate the estimation findings. The arrows in Fig. 1 show the improvement in accuracy from MNL or NB as compared to RMNL and RNB. Both MNL and NB benefit from injecting randomness, but generalizing RF is most profitable for MNL. Clearly, random feature selection addresses the multicollinearity problem of MNL thereby sincerely enhancing predictive performance. The smaller advantage of injecting randomness for NB might stem from diminished (as compared to full NB), but still considerable dependence between the attributes in the subset of *m* randomly selected features. Overall, the highest test performance is measured for RMNL_10 (*m*=77). Note that only RMNL_10 achieves to combine a high PCC (26.41%) with a high wPCC (21.06%). For this best test model (RMNL_10), we determine whether its *k* AUCs are statistically different from those of RF, MNL, NB, RNB_10 and SVM. Per product category, we employ the non-parametric test by DeLong et al. [DeLong] to determine whether the areas under the ROC curves (AUCs) within a product category are significantly different. All AUCs on the test set are statistically significant at α=0.05 except for 10_R_MNL_SVM, k=4 and k=5. Finally, also the *k* AUCs for NB and RNB_10 are significantly statistically different at α=0.05.

4.8 Feature Importance

From a CRM cross-sell action perspective, it is vital to gain insight in which features drive cross-buying propensities. Therefore, we assess the importance of the seleceted features in the RF, RMNL_10 and RNB_10 models.

Table 5 lists the top-10 features for RMNL_10 (best overall model) together with their z-score calculated on oob data as well as their appropriate rank in RF and RNB_10. A serious loss in predictive accuracy occurs when dropping features on the number of (different) appliances acquired per product category, the gender of the customer, the order of acquisition of home appliances and the time until a first acquisition or between repeated acquisitions in a product category.

5 Conclusion

RU models are robust multiclass models dominating marketing choice modelling due to their micro-economic theoretical underpinnings. Unfortunately, these RU models are un-suited to model choice settings with many features as they suffer heavily from multicollinearity. The latter might prevent convergence and seriously distorts parameter estimate interpretation. As to date, RU models lack any feature selection, we propose Random MNL, employing bagging in tandem with random feature selection, as alternative.

Acknowledgments

Our thanks go to Ghent University for funding computer infrastructure (Grantno. 011B5901). Dr Anita Prinzie is a Postdoctoral Fellow of the Research Foundation, Flanders (FWO Vlaanderen).

References

1. Baltas, G., Doyle, P.: Random utility models in marketing: a survey. Journal of Business Research (2001) 51(2) 115-125
2. Barandela, R., Sánchez, J.S., Garcia, V., Rangel, E.: Strategies for learning in class imbalance problems. Pattern Recognition (2003) 36(3) 849-851
3. Ben-Akiva, M., Lerman, S.R.: Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press, Cambridge (1985)
4. Breiman, L.: Random Forests. Machine Learning (2001) 45(1) 5-32
5. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines (2001) Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
6. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics (1988) 44 837-845
7. Dietterich, T.G.: Machine-Learning Research – Four current directions. AI Magazine (1997) 18(4) 97-136
8. Fawcett, T.: ROC Graphs: Notes and Practical Considerations for Researchers. Technical Report HPL-2003-4, HP Laboratories (2003)
9. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, (1993) 1022-1027
10. Freund, Y., Shapire, R.: Experiments with a new boosting algorithm. Machine Learning: Proc. of the Thirteenth International Conference, (1996) 148-156
11. Langley, P., Iba, W., Thomas, K.: An analysis of Baysian classifiers. Proceedings of the Tenth National Conference on Artificial Inteligence AAAI Press (1992) 223-228
12. Louviere, J., Street, D.J., Burgess, L.: A 20+ retrospective on choice experiments. In: Wind, Y., Green, P.E. (eds.): Marketing Research and Modeling: Progress and Prospectives, Academic Publishers, New York (2003)

13. Morrison, D.G.: On the interpretation of discriminant analysis. Journal of Marketing Research, (1969) 6 156-163
14. Prinzie, A., Van den Poel, D.: Predicting home-appliance acquisition sequences: Markov/Markov for Discrimination and survival analysis for modelling sequential information in NPTB models. Decision Support Systems (accepted 2007) , http://dx.doi.org/10.1016/j.dss.2007.02.008
15. Zhang, H., Jiang, L., Su, J.: Hidden Naive Bayes. Proceedings of the Twentieh National Conference on Artificial Inteligence AAAI Press (2005)
16. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons, New York (1998)

Table 1. Estimation performance of RF

m	wPCCe	PCCe	AUCe
21	19.74	20.38	0.6007
42	20.20	20.88	0.6057
63	20.56	21.23	0.6071
84	20.63	21.25	0.6061
168	20.98	21.59	0.6089
231	20.86	21.56	0.6090
294	21.01	21.69	0.6114
336	21.04	21.67	0.6067

Table 2. Estimation performance of RMNL

	R_MNL (R=100)			10_R_MNL (R=10)		
m	wPCCe	PCCe	AUCe	wPCCe	PCCe	AUCe
3	11.53	21.41	0.6163	19.30	23.93	0.6232
9	15.60	23.62	0.6270	19.69	24.98	0.6315
15	18.36	24.98	0.6328	20.56	26.33	0.6403
21	19.33	25.56	0.6390	21.09	26.78	0.6436
27	19.74	25.90	0.6423	21.14	26.63	0.6435
33	20.37	26.53	0.6458	21.59	27.13	0.6468
42	20.91	26.69	0.6480	21.82	27.31	0.6477
48	21.25	26.87	0.6491	22.01	27.33	0.6489

Table 3. Estimation performance for RNB

	RNB_100			RNB_10		
m	wPCCe	PCCe	AUCe	wPCCe	PCCe	AUCe
7	17.46	23.46	0.6141	18.96	22.29	0.6071
14	19.41	22.74	0.6134	19.39	22.26	0.6078
28	19.74	22.31	0.6115	19.87	22.99	0.6081
42	19.83	22.15	0.6100	20.19	22.39	0.6122
56	19.78	22.05	0.6083	20.13	22.28	0.6115
77	19.78	21.97	0.6066	20.21	22.31	0.6096
133	19.73	21.78	0.5508	20.03	21.98	0.6048
294	19.78	21.75	0.6011	19.85	21.82	0.6008

Table 4. Predictive test performance

	wPCCt	PCCt	AUCt
RF	20,66	21,39	0.6090
MNL	19,75	21,84	0.5626
RMNL	21,06	26,41	0.6322
NB	19,27	21,05	0.5899
RNB_10	19,61	21,56	0.5983
SVM	18,92	25,24	0.6188

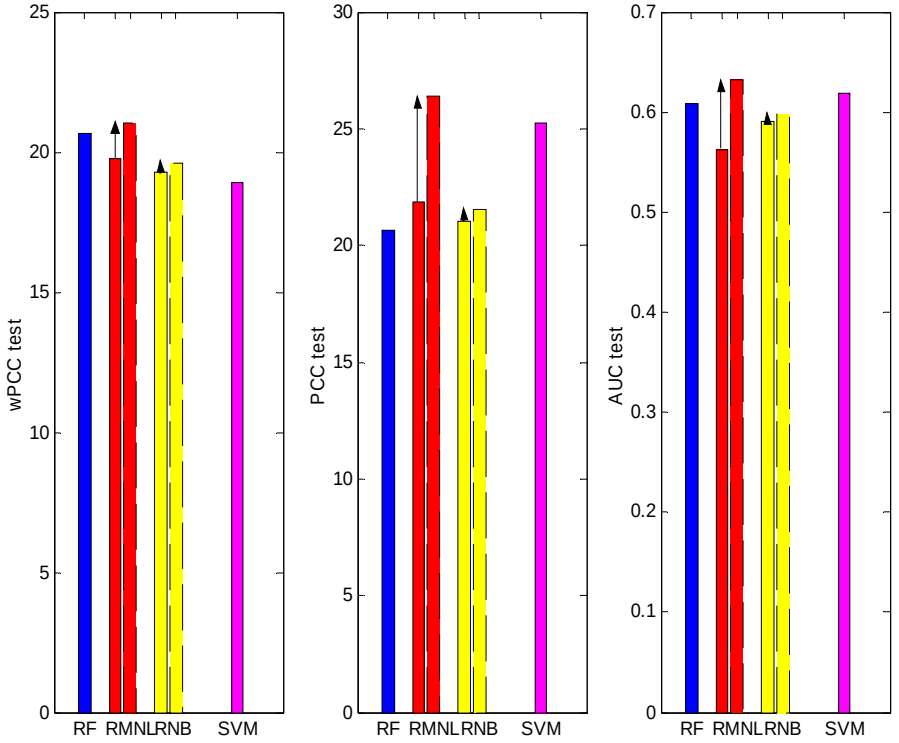

Fig. 1. Test set performance of RF, MNL, RMNL, NB, RNB and SVM

Table 5. Top-10 features

Rank	Description	z	RF	RNB
1	monetary, depth and width	29.27		18
2	monetary, depth and width	24.91		9
3	socio-demo	19.70	14	41
4	order	16.01		78
5	duration	9.48	4	63
6	order	9.21		81
7	order	7.69	51	15
8	order	4.86	7	34
9	socio-demo	4.84	35	
10	brand-loyalty	4.74	16	11