Paper 113-2009

Generalizations of Generalized Additive Model (GAM): A Case of Credit Risk Modeling

Wensui Liu, JP Morgan Chase
Chuck Vu, Acxiom
Jimmy Cela, Merkle

ABSTRACT

Famous for its easy implementation, Logistic Regression has been the primary model in credit scoring. Albeit attractively simple, it is criticized for failing to capture nonlinearity and nonmonotonicity and therefore possibly not leads to satisfactory results. Introduced by Hastie and Tibshirani, Generalized Additive Model (GAM) provides the ability to detect the nonlinear and nonmonotonic relationship between the risk behavior and predictors without the loss of interpretability. However, the semi-parametric nature of GAM makes it difficult to be used in a business. In this paper, we present an application of GAM in credit scoring as well as a class of hybrid models combining ideas of Logistic Regression and GAM in order to improve the generalizability of nonlinear modeling. Model performance is evaluated and compared among discussed models using the area under Receiver Operating Characteristic (ROC) curve. It is found that our proposed method is able to yield superior results in practice.

INTRODUCTION

The credit score aims to evaluate the credit worthiness of a potential borrower or the default likelihood of an existing customer with the purpose of minimizing financial costs and losses due to the credit failure. While FICO score from the credit bureau has often been used directly by most small- and mid-sized banking, credit card, and lending companies, major players tend to develop the internal scoring system by their own modeling team. The development of credit scoring models has been extensively studied in SAS community and is expected to be a continuously active area in the financial industry given the present turmoil of credit crisis. Fallen into the class of Generalized Linear Model (GLM), Logistic Regression is currently the Number One statistical technique in credit scoring models due to its simplicity. In recent researches, Generalized Additive Model (GAM), a more flexible statistical model with the semi-parametric functional form, has shown promising successes in the credit risk modeling by combining the flexibility with the interpretability. However, due to its semi-parametric nature, GAM has been found difficult to be understood in the business environment and be implemented in the production setting. In our paper, we should illustrate the superiority of GAM over Logistic Regression through the development of a credit risk model. Meanwhile, we would also demonstrate how to use other data mining techniques, namely Classification and Regression Tree (CART) and Multivariate Adaptive Regression Splines (MARS), to improve the usability of GAM and to bring this new modeling technique down to earth within the framework of GLM that is more familiar to the majority of our modelers.

METHODOLOGY

In a credit risk model, the adverse behavior is often formulated by an indicator function such that the bad are coded as Y = 1 and the good coded as Y = 0. Traditionally, Linear Discriminant Analysis or Logistic Regression has been used to model such binary outcome under the assumption that predictors are linearly related to the outcome variable. However, a potential risk of such assumption is model misspecification. While the effect of a predictor is often neither linear nor monotonic in the real world, it is always challenging to find an appropriate functional form between the response and the predictor. Consequently, Logistic Regression may not always be able to provide an adequate goodness-of-fit given the complex data structure.

As an alternative to Logistic Regression, GAM relaxes the linear restriction and assumes that the response is dependent on predictors in a flexible manner, which could be either linear or nonlinear. The nonlinear relationship is largely driven by data and can be estimated nonparametrically with a univariate B-spline or local regression smoother. In other words, instead of having a single coefficient for each predictor, GAM uses an unspecified nonparametric function to describe the relationship between each predictor and the response with the purpose of maximizing the predictive performance. Such nonparametric function is analogous to the coefficient in Logistic Regression and can be used to visualize the relationship between the response and the predictor. This ability to visualize the nonparametric function between each predictor and the response is an important feature of GAM and provides an intuitive way for modelers to explore the complex data structure and interpret the modeling result.

The data analyzed in this paper has come from a French bank and been used by Muller (Muller, 2000) to demonstrate an application of Generalized Partial Linear Model in credit risk modeling. The data set consists of 6,180 cases and 24 variables. The response variable Y reflects the status of loan and it has been coded as 1 for the bad and 0 for the good. Predictors include eight numeric variables (X2 - X9) and fifteen categorical variables (X10 – X24) with levels ranging from 2 to 11. However, all predictors have been unlabelled and no prior information about the data is known. Outliers in the data set have been removed such that X2 – X9 are all within the range between -3 and 3. Before the model development, we randomly divided the entire dataset into two parts: one for model development, and the other for model validation, as shown in Table 1.

Table 1: Response Summary

	Full Data	Development	Validation
Y = 1	372 (6.02%)	312 (6.02%)	60 (6%)
Y = 0	5,808 (93.98%)	4,868 (93.98%)	940 (94%)

A Logistic Regression is estimated for the conversion model using the development data with the inclusion of all predictors and a focus on seven numeric variables. Partial output of the model is shown in Table 2 below, in which predictors significant at 10% level are flagged.

```
proc logistic data = train desc;
  class X10 - X24 / params = GLM;
  model Y = X2 - X24 / lackfit;
  score data = train out = predict_train;
  score data = test  out = predict_test;
run;
```

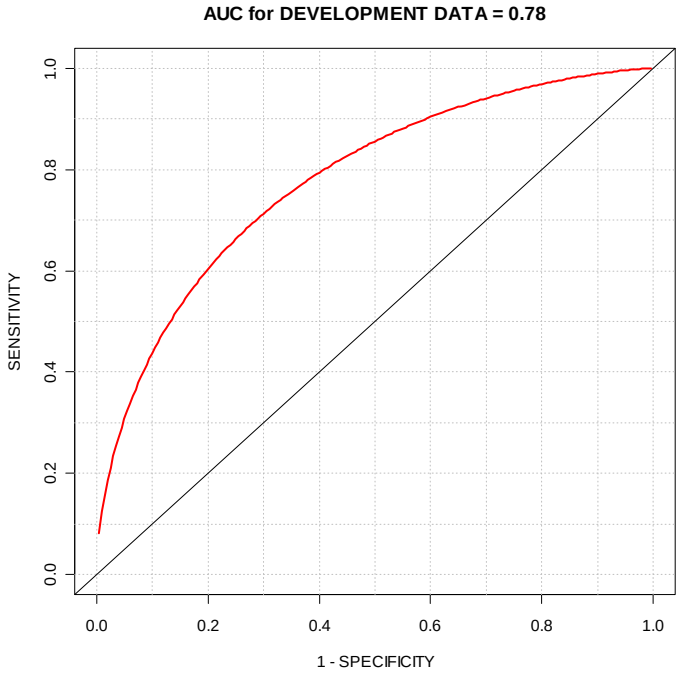
Table 2: Output of Logistic Regression

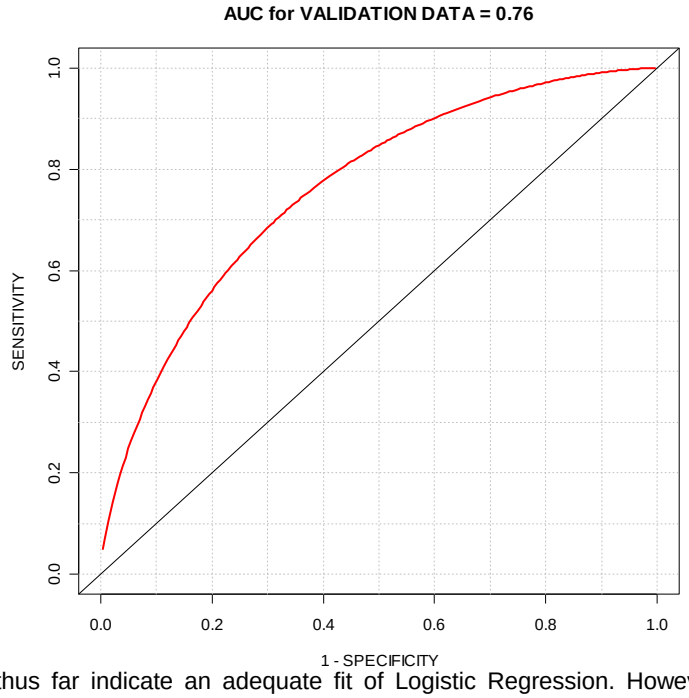
Variable	Estimate	Std. Error	z value	Pr(>|z|)	Sig
x_2	0.2140	0.1162	1.84	0.0656	*
x_3	-0.3242	0.0881	-3.68	0.0002	*
x_4	-0.0442	0.0935	-0.47	0.6367	
x_5	-0.0608	0.0895	-0.68	0.4968	
x_6	0.1431	0.0998	1.44	0.1514	
x_7	-0.1428	0.1756	-0.81	0.4160	
x_8	-0.8081	0.1976	-4.09	0.0000	*
x_9	-1.0458	0.4045	-2.59	0.0097	*
…	…	…	…	…	

Out of seven numeric variables, only four are shown to be statistically significant based on the linear assumption. Besides the illustrated output, all statistics suggest that Logistic Regression is able to provide an adequate fit for the development data.

Receiver Operating Characteristic curve, also known as ROC curve, is a graphical representation of the tradeoff between Type-I error (Sensitivity) and Type-II error (Specificity) for different possible cutoffs and is often used to compare predictive performance between different classification models. In a ROC curve, Sensitivity is placed on the Y-axis and 1 – Specificity on the X-axis. The area under ROC curve, also abbreviated as AUC, is a statistical measure often used to summarize the information of ROC curve derived from a predictive model. In brief, AUC can be interpreted as the probability that the predictive model is able to score a randomly selected positive response higher than a randomly selected negative response. A predictive model with the perfect performance has an area under ROC curve equal to 1, whereas this area is 0.5 for a predictive model as good as the random guess. In practice, the area under the ROC curve is between 0.5 and 1. In Figure 1, ROC curves of logistic regression for both development and validation data are plotted. The areas under ROC are equal to 0.78 and 0.76 respectively for the development and the validation data, suggesting a reasonable predictiveness for Logistic Regression.

Figure 1: ROC Curves for Development and Validation Data





All statistical evidences thus far indicate an adequate fit of Logistic Regression. However, whether the assumed linearity is the correct functional form for the analyzed data remains in question and still needs to be addressed. Since Logistic Regression indicates that X4, X5, X6, and X7 are statistically insignificant, we will pay extra attention to these four variables.

After establishing Logistic Regression Model as a benchmark, we fit a Generalized Additive Model with the same development data and apply a flexible nonparametric estimation to those four insignificant predictors in Logistic Regression. With the flexibility provided by GAM, there is a strong temptation to over-fit the model to the development data using excess degrees of freedom in the model. Based upon our experience, the benefit of using conservative degrees of freedom in GAM is twofold. First of all, over-fitting can be prevented with low degrees of freedom.

Secondly, the computation cost can be reduced dramatically. Table 4 shows the partial output of the best GAM developed with all predictors included.

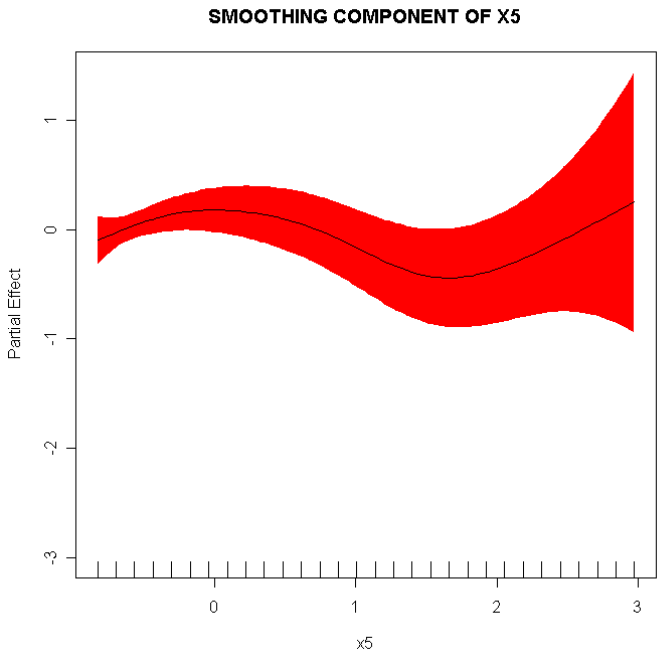
```
proc gam data = train;
  class X10 - X24;
  model Y = param(X2 X3 X4 X6 X8 - X24) spline(X5) spline(X7) / dist = binomial;
  score data = train out = predict_train;
  score data = test out = predict_test;
run;
```

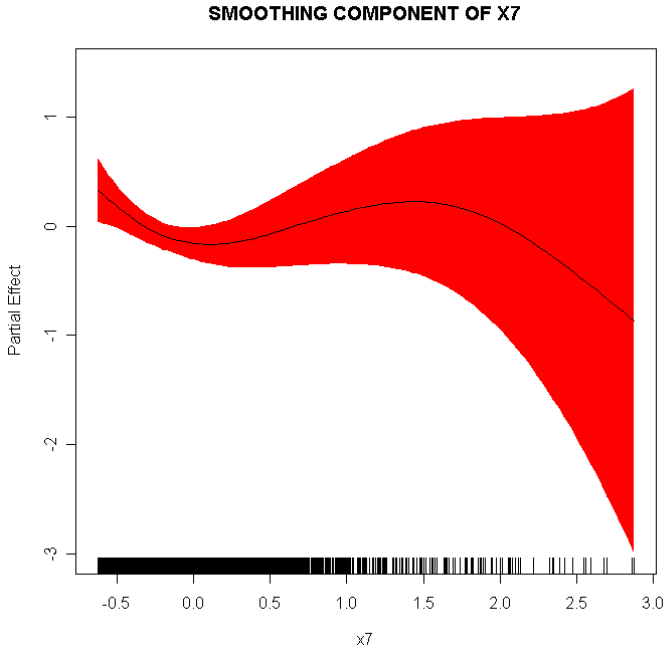
Table 4: Partial Output of Generalized Additive Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Sig
x_2	0.1953	0.1167	1.673	0.0944	*
x_3	-0.3224	0.0888	-3.633	0.0003	*
x_4	-0.0313	0.0939	-0.334	0.7386	
x_6	0.1238	0.1011	1.224	0.2210	
x_8	-0.8161	0.1971	-4.141	0.0000	*
x_9	-1.0547	0.4049	-2.605	0.0092	*
	D.F.	Rank	chi. sq	P-value	
$S(x_5)$	3.4520	7	12.31	0.0908	*
$S(x_7)$	3.3100	7	13.04	0.0712	*
…	…	…	…	…	

Note that X5 and X7 become significant after being estimated nonparametrically under the nonlinear assumption. Nonlinear effects of predictors are shown in Figure 2. It is clear that the relationship between X5 and Y is neither linear nor monotonic, a violation of the linear assumption in logistic regression. Instead, the risk goes up as X5 increases, starts decreasing once X5 reaches 0, and then picks up again after X5 exceeds 1.5.

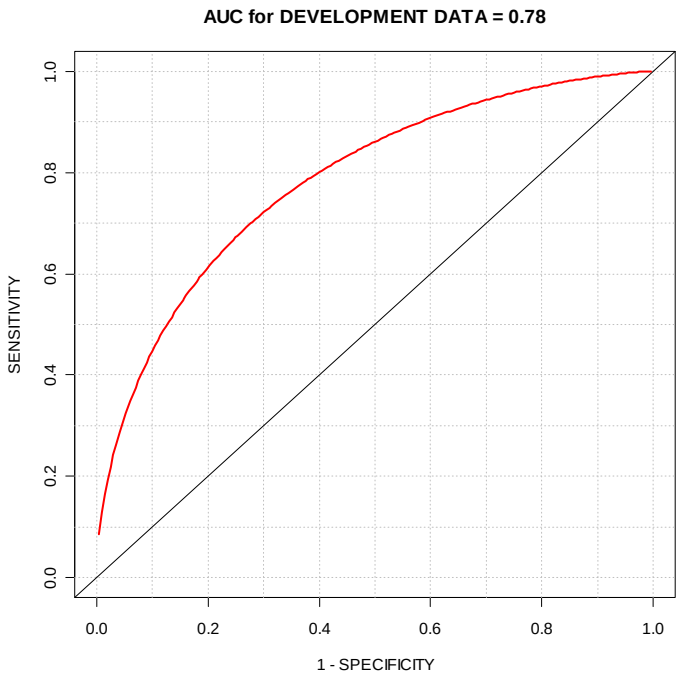
Figure 2: Partial Prediction Plots of X5 and X7

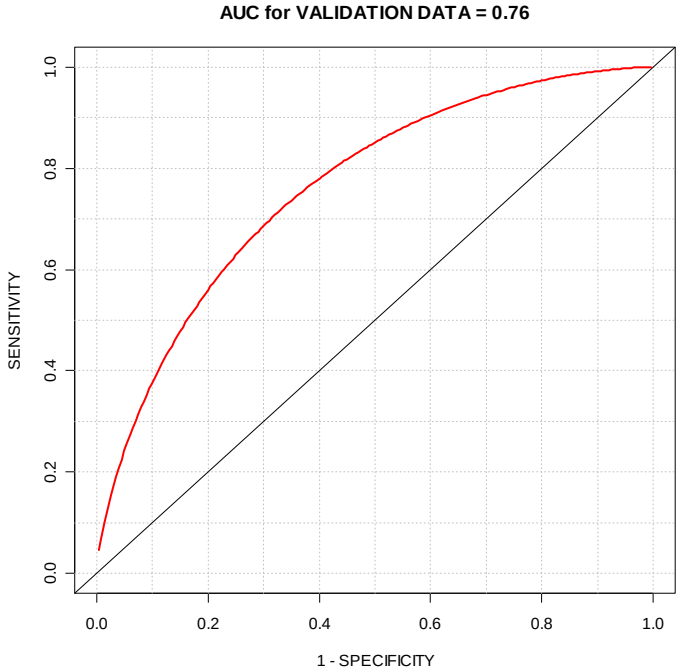




For the comparison purpose, ROC of GAM for both development and validation data are also plotted in Figure 3 to evaluate the predictive performance based upon values of Area under Curve (AUC).

Figure 3: ROC Curves for Development and Validation Data



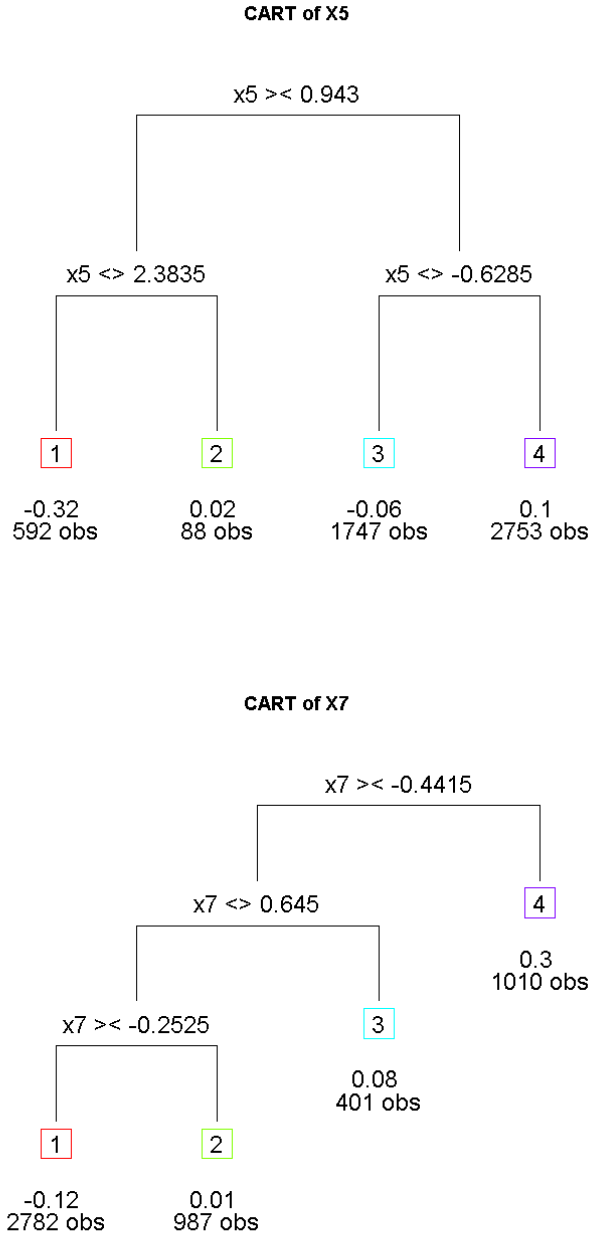


While AUC for the development data is marginally higher than AUC from Logistic Regression, both models perform comparably well for the validation data. However, it is important to note that GAM provides a better insight into the relationship between the response and the predictor without imposing the linear assumption and the pre-specified functional form on the model. This feature is particularly helpful in credit risk modeling, where the data structure is always complex and little knowledge about hundreds of variables is known before the model development.

Although conceptually attractive, GAM is not beyond criticism. Lack of parametric functional form makes it difficult to score the new data directly from the database in a production environment. Also, the nonlinear effect without an estimated parameter might not be easily adopted by non-technical audiences such as business managers. A possible workaround might be to estimate a parametric approximation for each nonlinear term derived from the GAM, such as a piece-wise constant approximation. While there are numerous ways to come up with such an approximation based upon either percentile or experience, we propose a model-based method using Classification and Regression Tree (CART) to develop such approximation.

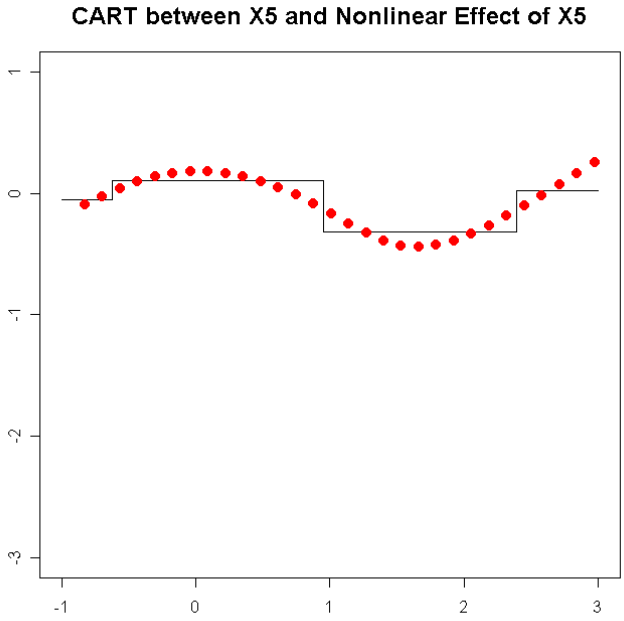
In our case, CART is developed with only one independent variable and one dependent variable, which are the predictor and the corresponding nonlinear term respectively. Such a tree-based model is constructed through a process of recursive binary partitioning. At each partition, an “if-then” splitting rule is generated to divide the nonlinear term into several homogeneous groups based upon the value of the predictor. Figure 4 shows diagrams of CART for X5 and X7.

Figure 4: Classification and Regression Trees for X_5 and X_7



After the development of CART, we then use the piecewise constant approximation as a categorical variable to replace the nonlinear term from GAM. As a result, GAM collapses to become a Logistic Regression. Figure 4 shows such piecewise constant approximation for X5 and X7.

Figure 5: Piecewise Constant Approximation for X5 and X7



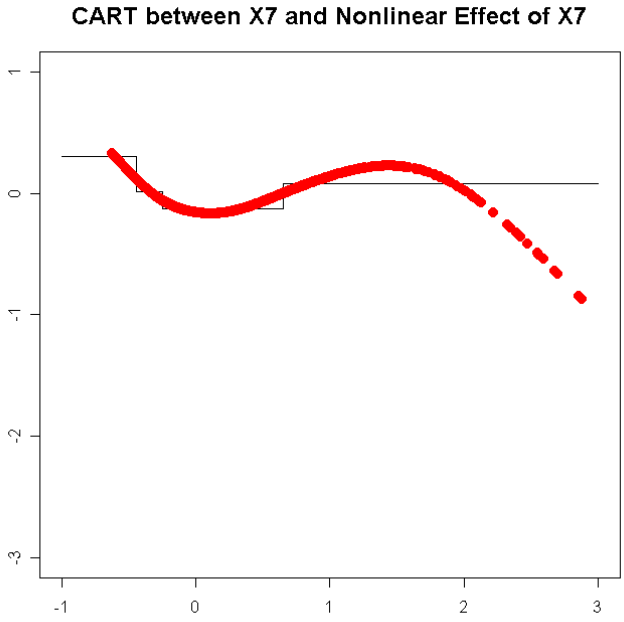


Table 5 displays the statistical output from revised Logistic Regression with the inclusion of piecewise constant approximation. While X5 shows statistical significance, X7 is at the border of 10% significant level. Parameter estimates and statistical significances of the other numeric variables are also very close to the ones in GAM.

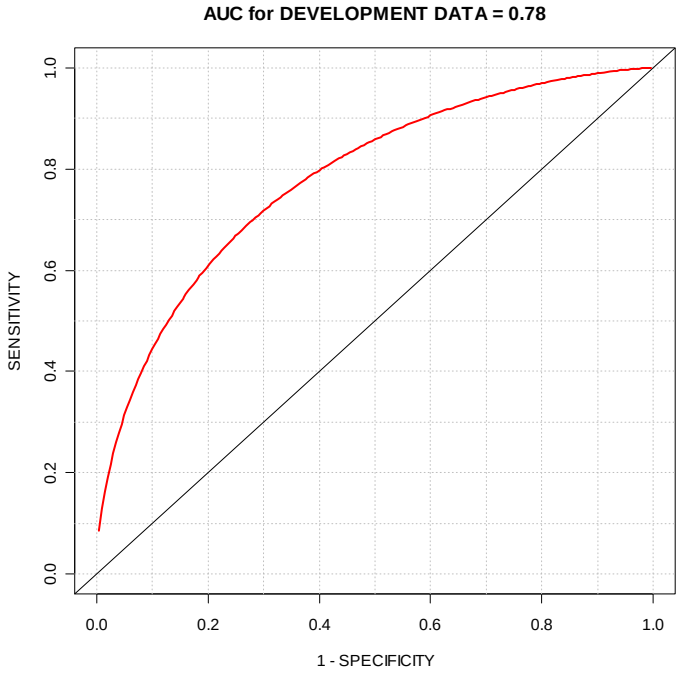
```
proc logistic data = train(drop = X5 X7) desc;
  class X10 - X24 X5_CAT X7_CAT / params = GLM;
  model Y = X2 - X24 X5_CAT X7_CAT / lackfit;
  score data = train out = predict_train;
  score data = test  out = predict_test;
run;
```

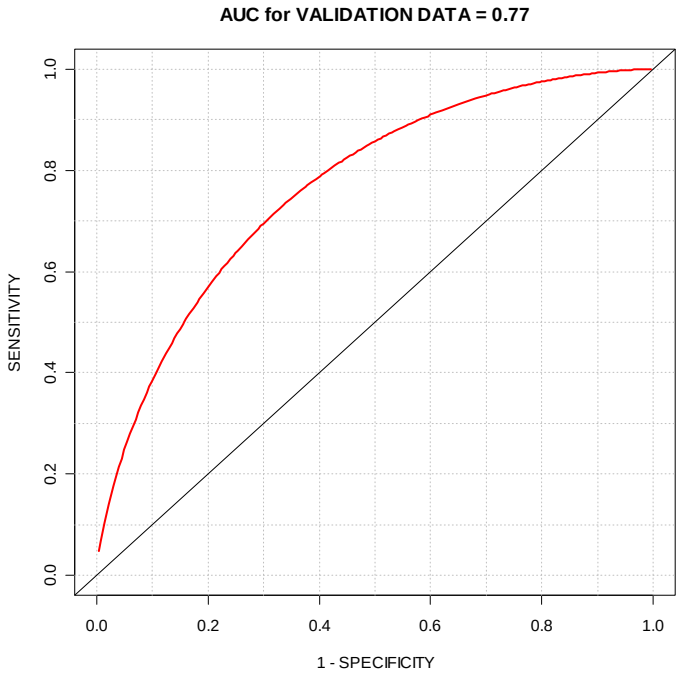
Table 5: Partial Output of Revised Logistic Regression with Piecewise Constant Approximation

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Sig
x_2	0.1911	0.1153	1.66	0.0974	*
x_3	-0.3280	0.0887	-3.70	0.0002	*
x_4	-0.0380	0.0936	-0.41	0.6846	
x_6	0.1254	0.1010	1.24	0.2141	
x_8	-0.8194	0.1970	-4.16	0.0000	*
x_9	-1.0603	0.4048	-2.62	0.0088	*
	D.F.		**chi. sq**	**P-value**	
x_5_cat	**3**		**6.78**	**0.0793**	*
x_7_cat	**3**		**5.98**	**0.1228**	
…	…	…	…	…	

In Figure 5, ROC of Logistic Regression with the piecewise constant approximation for both the development and the validation data are plotted to evaluate the efficiency of predictiveness. For the development data, this new hybrid model performs similarly compared with GAM illustrated early. However, it generalizes better than both Logistic Regression and GAM for the validation data with AUC = 0.77. Since one of the advantages in CART is the resistance to outliers, our understanding is that this marginal improvement might come from the reduction of over-fitting.

Figure 6: ROC Curves for Development and Validation Data

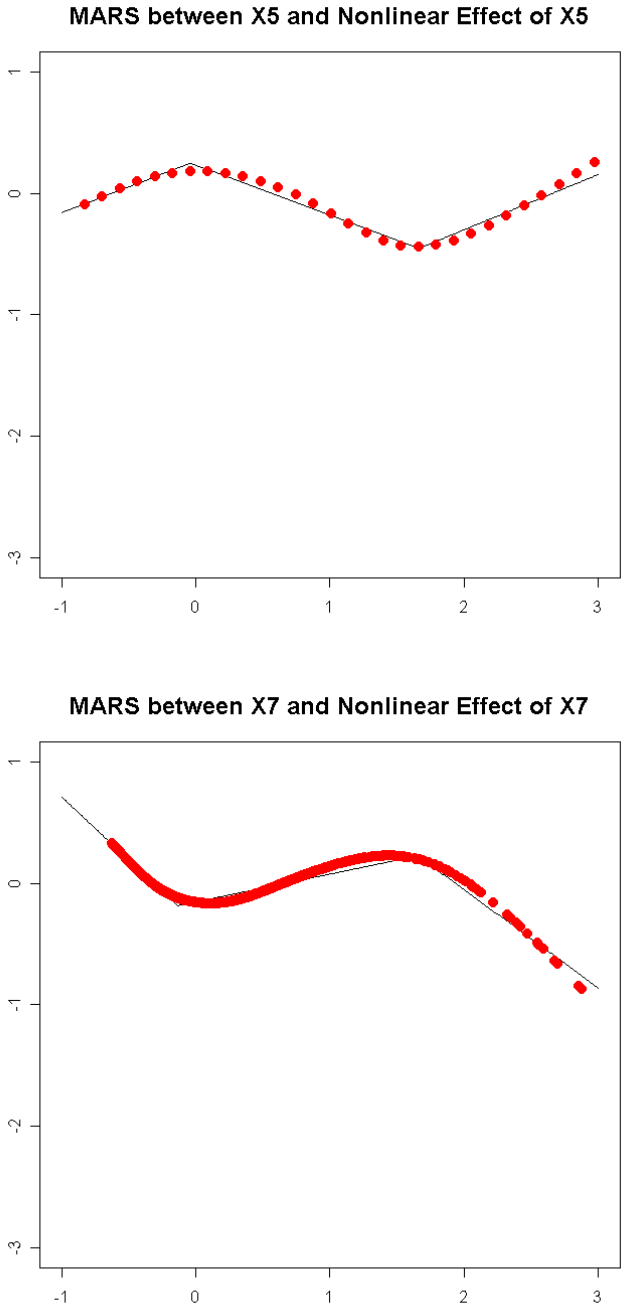




While CART provides a satisfactory solution through the piecewise constant approximation for the nonlinear effect, Multivariate Adaptive Regression Spline (MARS) is able to address the same problem with the piecewise linear approximation. Similar to CART, MARS is designed to partition the entire domain of a predictor into multiple sub-regions known as “Basis”. Within each sub-region, a regression with a different coefficient is used to define the relationship between the response and the predictor. As a result, such “divide and conquer” strategy provides the flexibility to approximate any nonlinear and nonmonotonic pattern with a sufficient number of basis functions. It is interesting to note that, when coefficients in each sub-region are equal to zero and only the intercept remains, then MARS will simply become a CART.

As the building block of MARS, the basis function can be considered a function with the “hockey stick” shape and takes the functional form of BF = max(X – K, 0), where K is the breaking point of the hockey stick. During the model development, a forward selection method is employed to include statistically significant basis functions, followed by a backward elimination process, which is a similar technique used in Logistic Regression for variable selection. Figure 7 shows the piecewise linear approximation for nonlinear terms of X_5 and X_7, indicating that each nonlinear pattern can be adequately approximated by three basis functions.

Figure 7: Piecewise Linear Approximation for X5 and X7



After the development of MARS, we can use derived basis functions to replace nonlinear terms from GAM and re-estimate a Logistic Regression with the inclusion of these basis functions. Table 6 illustrates the output of a revised Logistic Regression and shows that almost all basis functions are statistically significant in this new model.

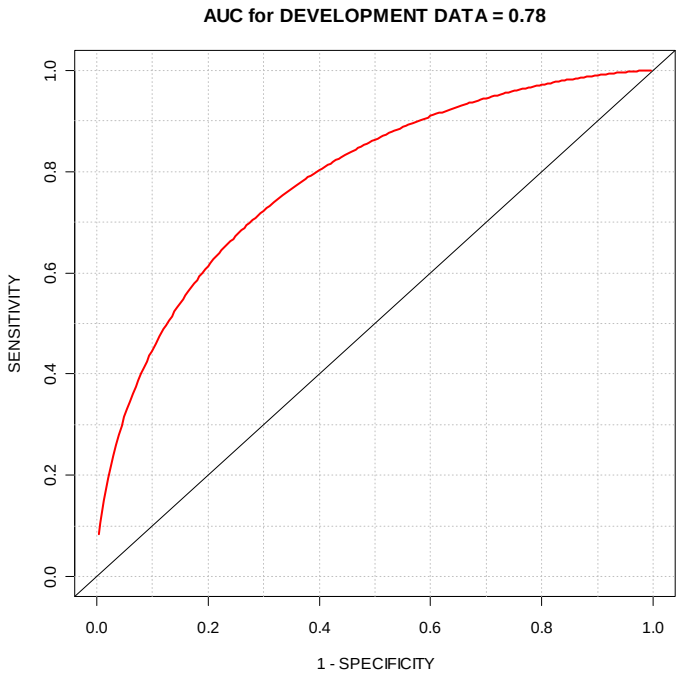
```
proc logistic data = train(drop = X5 X7) desc;
  class X10 - X24 / params = GLM;
  model Y = X2 - X24 X5_1 – X5_3 X7_1 – X7_3 / lackfit;
  score data = train out = predict_train;
  score data = test  out = predict_test;
run;
```

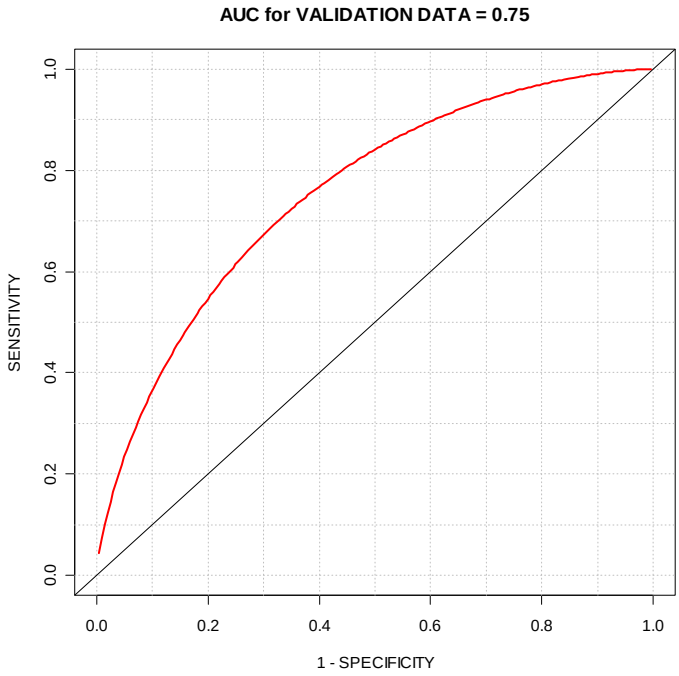
Table 6: Partial Output of Revised Logistic Regression with Piecewise Linear Approximation

Variables	Estimate	Std. Error	z value	Pr(>|z|)	Sig
x_2	0.1910	0.1171	1.63	0.1030	
x_3	-0.3215	0.0889	-3.62	0.0003	*
x_4	-0.0164	0.0942	-0.17	0.8616	
x_{5_1}	**0.5522**	**0.2703**	**2.04**	**0.0411**	*
x_{5_2}	**-1.1393**	**0.4039**	**-2.82**	**0.0048**	*
x_{5_3}	**1.5284**	**0.6629**	**2.31**	**0.0211**	*
x_6	0.1213	0.1013	1.20	0.2313	
x_{7_1}	**-1.3465**	**0.4545**	**-2.96**	**0.0031**	*
x_{7_2}	**1.7767**	**0.5784**	**3.07**	**0.0021**	*
x_{7_3}	**-2.3434**	**1.5222**	**-1.54**	**0.1237**	
x_8	-0.8210	0.1973	-4.16	0.0000	*
x_9	-1.0507	0.4041	-2.60	0.0093	*
…	**…**	**…**	**…**	**…**	

The predictive performance is summarized in Figure 8 using ROC and AUC again. While there is no noticeable difference compared with Logistic Regression and GAM for the development data, this new model doesn't seem to generalize well for the validation data with AUC = 0.75. In our view, this under-performance is likely to be arisen from over-fitting, a well-known drawback of MARS.

Figure 8: ROC Curves for Development and Validation Data





CONCLUSION

In this paper, we have demonstrated a new class of nonlinear modeling techniques and its application in credit risk modeling. In our experience, GAM outperforms Logistic Regression in two aspects. First of all, GAM relaxes the linear assumption between the response and predictors and therefore avoids the potential problem of model misspecification, which often occurs in a linear-based model such as Linear Discriminant Analysis or Logistic Regression. Secondly, by incorporating nonlinear effects, GAM helps discover hidden patterns between the response and predictors and consequently improves the predictive performance albeit at the risk of over-fitting.

Besides the original concept and implementation of GAM, we have also proposed two hybrid models based upon our learning experience, which are piecewise constant and piece linear approximation models. Our findings show that the hybrid model combining the flexibility of GAM and the resistance to over-fitting of CART is able to yield the best result.

REFERENCES

Franke, J., Hardle, W., and Stahl, G., Measuring Risk in Complex Statistical Systems, Springer Verlag, (2000).
Hastie, T. and Tibshirani, R., Generalized Additive Models, Chapman and Hall, (1990).
McCullagh, P. and Nelder, J., Generalized Linear Models, Chapman and Hall, (1989).
Muller, M., Semi-Parametric Extensions to Generalized Linear Models, Habilitationsschrift, (2000).
Liu, W. and Cela, J., Improving Credit Scoring by Generalized Additive Model, (2007).
McCullagh, P. and Nelder, J., “Generalized Linear Models,” Chapman and Hall, (1989).
Hastie, T., Tibshirani, R., and Friedman, J., “Elements of Statistical Learning,” Springer, (2001).
Breiman, L., Friedman, J., Olshen, R., and Stone, C., “Classification and Regression Trees,” Chapman & Hall, (1984)
Friedman, J., “Multivariate Adaptive Regression Splines,” The Annals of Statistics, Vol. 19, No. 1, 1-67, (1991).

CONTACT INFORMATION

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

Name:	Wensui Liu
Enterprise:	JP Morgan Chase
Phone:	(513) 295-4370
E-mail:	liuwensui@gmail.com
Web:	http://statcompute.spaces.live.com/

Name:	Chuck Vu
Enterprise:	Acxiom Corporation
Address:	1105 Lakewood Parkway, Suite 100
City, State ZIP:	Alpharetta, GA 30009
Phone:	(678) 537-6029
Fax:	(678) 537-6070
E-mail:	chuck.vu@acxiom.com
Web:	http://www.acxiom.com