Paper 1713-2014

Role of Customer Response Models in Customer Solicitation Center's Direct Marketing Campaign

Arun K Mandapaka, Amit Singh Kushwah, Dr.Goutam Chakraborty

Oklahoma State University, OK, USA

ABSTRACT

Direct Marketing is the practice of delivering promotional messages directly to customers or prospects on an individual basis rather than using a mass medium. In this project, we build a finely tuned response model that helps a financial services company to select high quality receptive customers for their future campaigns and identify the important factors that influence the marketing campaign to effectively manage their resources.

This study was based on the customer solicitation center's marketing campaign data (45,211 observations and 18 variables) available on UC Irvine website with attributes of present & past campaign information (communication type, contact duration, previous campaign outcome etc.) and customer's personal & banking information. As part of data preparation, we had performed decision tree imputation to handle missing values and categorical recoding for reducing levels of class variables.

In this study we had built several predictive models using SAS® Enterprise Miner: Decision Tree, Neural Network, Logistic Regression and SVM Models to predict whether the customer responds to a loan offer by subscribing or not. The results showed that Stepwise Logistic Regression Model was the best when chosen based on the misclassification rate criteria. When the top 3 decile customers were selected based on the best model, the cumulative response rate was 14.5% in contrast to the baseline response rate of 5%. Further analysis showed that the customers are more likely to subscribe to the loan offer if they have the following characteristics: never been contacted in past, no default history, and provided cell phone as primary contact information.

KEYWORDS Direct Marketing, Response Model, Campaign, SAS® Enterprise Miner, SAS® Enterprise Guide, Decision Tree, Logistic Regression, Support Vector Machine, Neural Networks, Response rate.

INTRODUCTION

In this present market condition for financial services, it is very much important for the marketers to make most of the customer contact details. The marketer's goal is that no consumer should be receiving a junk mail or irrelevant call. This means that every contact that the company targets should be interested in at least attending or responding to the promotion. It has become really important that the responsive customers need to be selected for effective direct marketing. Saturated markets and increasing competition is raising lot of concern in the finance industry as it is reducing the responsiveness of the consumers and increasing the marketing costs. This concern has led to the requirement of better response models with a finely tuned approach, which will enable the companies to invest on direct marketing with effective and efficient selection of contacts.

In this project, we had built a finely tuned response model that helps a financial services company to select high quality receptive customers for their future campaigns and find the most important factors that influence marketing to effectively manage their resources. We have used SAS® 9.3 & SAS® Enterprise Guide 5.1 for data preparation. The ability to build effective predictive models and compare them is a major strength of the SAS® Enterprise Miner. So, in this research the SAS® Enterprise Miner 12.1 is used to build the predictive models which can improve the campaign efficiency and ultimately track the quantifiable factors that improve the customer response.

This paper will give an overview of how the SAS system can be used in building customer response models and show the importance of these predictive models in direct marketing campaigns. The data used for this study is masked due to proprietary issues.

DATA PREPARATION

The value of the database is directly proportional to its cleanliness and integrity. The data is an essential component of marketing industry and erroneous data can lead to serious ramifications. Especially the customer databases play a vital role in marketing industry today. So, poor data quality can lead to ineffective direct marketing and reduced efficiency and effectiveness of marketing efforts.

This study was based on the customer solicitation center's marketing campaign data (45,211 observations and 18 variables) available on UC Irvine website. The telemarketing data included the customer information (age, sex, and marital status), campaign history (last response result, no. of days since last contact), banking information (Balance, default status) and the customer solicitation information.

Original Variable	Description	Range of Values	Type of Information
Age	Age of the consumer	21-73	Personal
Marital Status	Marital status of the consumer	Married Divorced Single	Personal
Sex	Sex of the consumer	Male/Female	Personal
Job_type	Job description	admin. unknown unemployed management housemaid entrepreneur student blue-collar self-employed retired technician services	Personal
Education	Highest education of the consumer	Secondary Primary Tertiary	Personal
Housing loan	Does consumer has a housing loan or not?	Yes/No	Bank
Credit Default	Has credit in default?	Yes/No	Bank
Average Balance	Average yearly balance	-8019 to 102127	Bank
Personal Loan	Does consumer has a personal loan or not?	Yes/No	Bank
Communication Type	Communication type used to contact the consumer in the past	Cellular Telephone Unknown	Last Contact Information of current campaign
Day of Contact	The last contact day of the month	1-31	Last Contact Information of current campaign
Month of Contact	The last contact month of the year	Jan-Dec	Last Contact Information of current campaign
Duration of Contact	The duration of the contact in seconds	0-4918	Last Contact Information of current campaign
Contacts_Campaign	The number of contacts performed for the consumer	1-63	Previous campaign information
Contacts_previousdays	The number of days passed after the last contact. -1 indicates that the consumer has never been contacted	-1 to 871	Previous campaign information
contacts_previouscampaign	Number of contacts performed before this campaign	0-275	Previous campaign information
Outcome_previouscampaign	Outcome of the previous campaign	Failure Success Unknown Other	Previous campaign information

Loanoffer_response	Did the consumer respond for the loan offer by subscribing or not?	Yes/No	TARGET

Table 1: The list of the original variables used for the study

***Note: The data used for this study is masked due to proprietary issues.**

As part of the data preparation, there were many issues that were dealt with the data. It is really important that the data is cleaned in the initial phase as it does effect the further modeling process.

1. ***Extract the data***: The data was obtained from the public databases of UC Irvine website.

2. ***Initial Exploratory Analysis***: The data consisted of 45,000 observations with 18 variables. The data collected was related to the banks customer solicitation center direct marketing campaign. During these phone campaigns, an attractive long-term deposit application, with good interest rates, was offered. For each contact, a large number of attributes (personal, banking & campaign contact) was stored and if there was a success (the target variable). The data consisted of attributes such as age and average balance which were continuous variables and campaign contact information which were mostly categorical variables. The exploratory analysis consisted of descriptive statistics, graphs and frequency tables. The descriptive statistics showed that the variables such as age, average balance had missing values. It was noticed that about 120 observations were duplicated which was dealt with SAS® programming. The duration of the contact was ranging between 0 seconds – 4918 seconds, had few outliers due to which the variable skewness was quite high.

3. ***Missing Values***: The variables such as age and average balance of the customer had missing values. The missing values are generally dealt well with models such as decision trees. Models such as logistic regression and Support vector machine do not work well with these missing values. The missing values are replaced with their means by using the SAS® programming. The missing values for the age and the average balance were replaced by their respective means.

```
/*Replacing Missing Values*/

%macro replace_missing(dataset);
      proc means data=&dataset mean ; ;
            output out=t (drop=_type_ _freq_) mean=/autoname;
    run;
      proc transpose data=t out=tr;
      run;
data _null_;
      set Tr end=last ;
      call symput('mean'||left(_n_),col1);
      if last then call symput('maxi', _n_);
run;
data &dataset;
      set &dataset;
      array Ages(*) _numeric_;
            %do n=1 %to &maxi;
                  if Ages(&n)=. then Ages(&n)= &&mean&n;
            %end;
run;
%mend;
%replace_missing(age_f);
run;
```

Code 1: Missing values replaced by their respective means using SAS® 9.3

4. ***Duplicate Observations***: Duplication of information within data sets is a common occurrence. Initial exploratory analysis showed that about 120 observations are duplicated. SAS® programming is used to remove these observations. A standard/accepted solution for removing duplicates i.e. NODUPKEY option of PROC SORT is used.

```
/*Removing the Duplicates*/

proc sort data=dataset nodupkey out=dataset_nodups;
run;
```

Code 2: Removing Duplicate Observations using SAS® 9.3

5. ***Recoding the Variables***: The categorical variable job type had 12 levels which were replaced with 4 levels such as management, services, unemployed and entrepreneur. The months of the contact had 12 months which was recoded into 4 quarters. The continuous variable average balance had negative balances. The negative balances have been replaced with zero balance.

Job_type	blue-collar	**services**	9730	C	blue-collar	.
Job_type	management	**management**	9454	C	management	.
Job_type	technician	**services**	7597	C	technician	.
Job_type	admin.	**management**	5168	C	admin.	.
Job_type	services	**services**	4153	C	services	.
Job_type	retired	**unemployed**	2261	C	retired	.
Job_type	self-employed	**entrepreneur**	1577	C	self-employed	.
Job_type	entrepreneur	**entrepreneur**	1484	C	entrepreneur	.
Job_type	unemployed	**unemployed**	1303	C	unemployed	.
Job_type	housemaid	**entrepreneur**	1240	C	housemaid	.
Job_type	student	**unemployed**	938	C	student	.

Fig 1. Job Type variable replacement

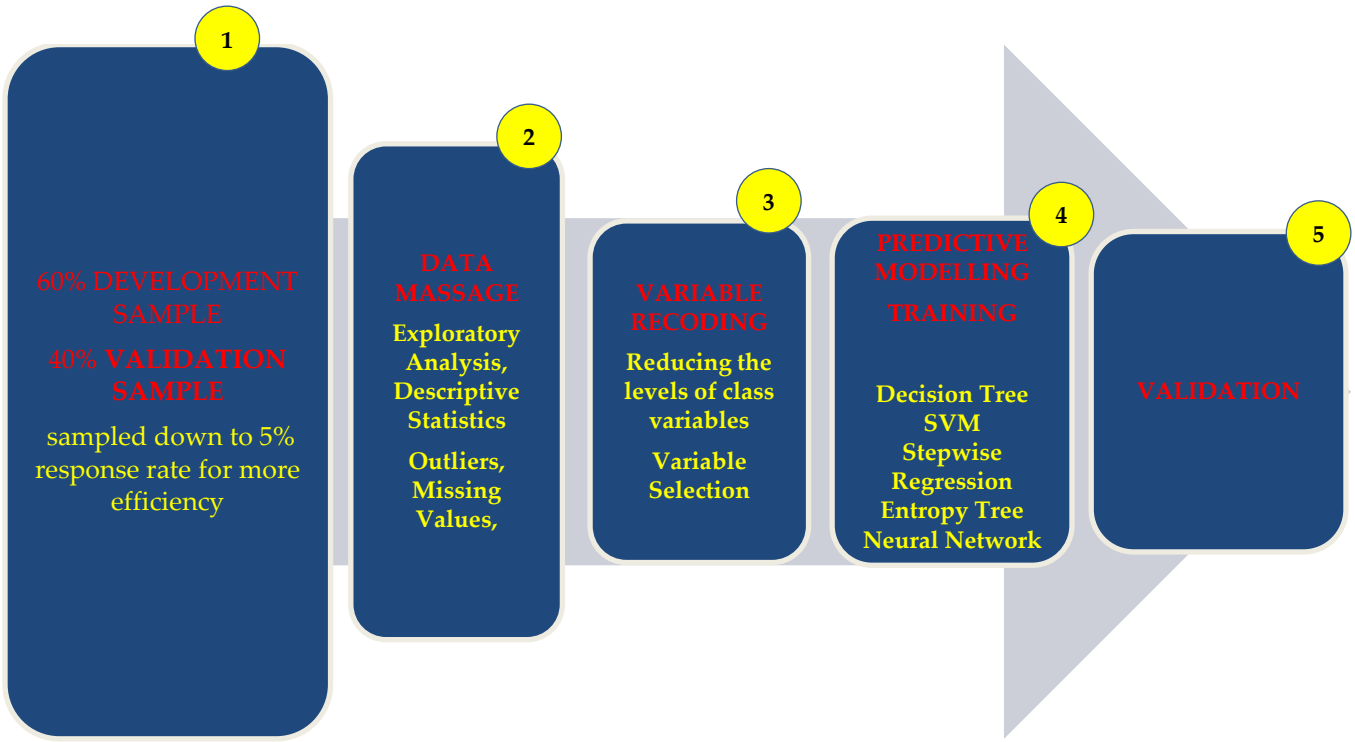
Month_of_Contact	may	2	13733	C	may	.
Month_of_Contact	jul	3	6862	C	jul	.
Month_of_Contact	aug	3	6179	C	aug	.
Month_of_Contact	jun	2	5249	C	jun	.
Month_of_Contact	nov	4	3954	C	nov	.
Month_of_Contact	apr	2	2924	C	apr	.
Month_of_Contact	feb	1	2635	C	feb	.
Month_of_Contact	jan	1	1388	C	jan	.
Month_of_Contact	oct	4	726	C	oct	.
Month_of_Contact	sep	3	569	C	sep	.
Month_of_Contact	mar	1	474	C	mar	.
Month_of_Contact	dec		212	C	dec	.

Fig 2. Month of the Contact variable replacement

Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace Method
Average_Balanc	**Default**	**User Specified**	**0**	**.**	**Default**

Fig 3. Average Balance of the customer variable replacement

METHODOLOGY



5-step process to develop response model using SAS® Enterprise Miner

Cross Industry Standard Process for Data mining (CRISP-DM) was followed to build the predictive models. CRISP-DM is an efficient methodology helping in building the effective and efficient predictive models to be used in real environment, assisting in business decisions.

The prior probabilities have been adjusted appropriately and the data is split into 60% development and 40% validation sample using the stratified sampling technique. After the sampling is done, the data preparation is performed by attending to issues such as missing values, duplicate values, outliers and then recoding the variables by reducing the levels of attributes. Several crosstabs, histograms and correlations have been run between the input variables and the target to get a better understanding of the attributes and their dependencies.

Using this prepared training sample the predictive models have been built using SAS® Enterprise Miner. Various predictive models such as Neural Networks, Logistic Regression, Entropy Tree, Support Vector Machine have been constructed to develop the best model which predicts whether customer responds to the loan offer by subscribing or not. Decision tree with the splitting measures for categorical variables as Entropy (Information Gain) and different combinations of maximum branches, maximum depth were used. These classification algorithms used as splitting criteria in classification trees by increasing the purity of categorical variables in child nodes. The logistic regression model with the model selection method as stepwise was built. Neural Network model was built using the multilayer perceptron architecture and by varying the number of hidden units. The support vector machine model was built which is a supervised machine learning method. Once the model is trained then it is validated with the validation sample.

The models were compared based on the validation misclassification rate. The best model was selected based on the least validation misclassification rate. ROC analysis and cumulative lift curve area have been analyzed with the baseline model to understand how the model is performing. The quantifiable factors have been analyzed with respect to the target variable to understand which were actually making an impact to improve the customer response rate.

RESULTS

Initial Exploratory Analysis

It had been observed that customers between ages 27 years to 43 years have responded more to the loan offer.

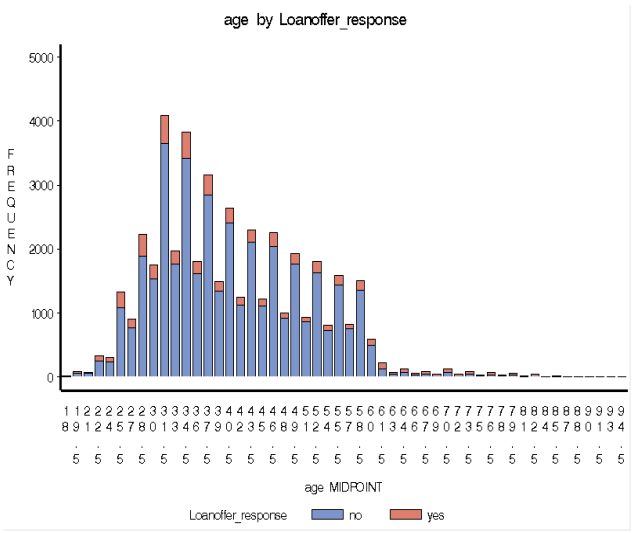


Fig 4. Age Vs Loan Offer Response

The customers who have been contacted less than once have responded more to the campaign.

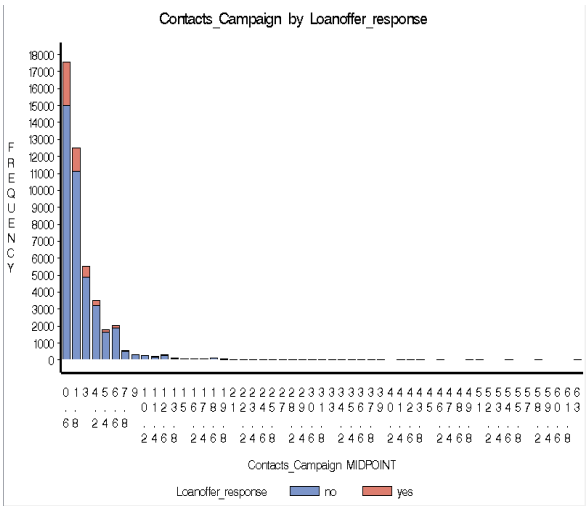


Fig 5. Number of contacts made for the campaign Vs Loan Offer Response

Customers with secondary education have responded more to the campaign than the customers with primary education.

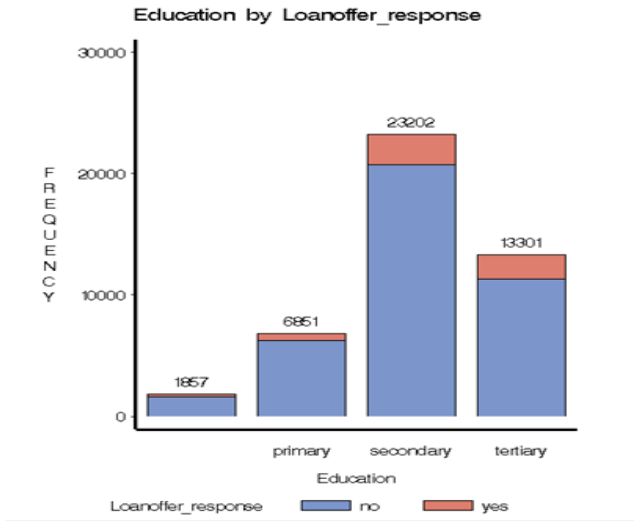


Fig 6. Education Vs Loan Offer Response

Customers who were married have higher response rate than the customers who were single or divorced.

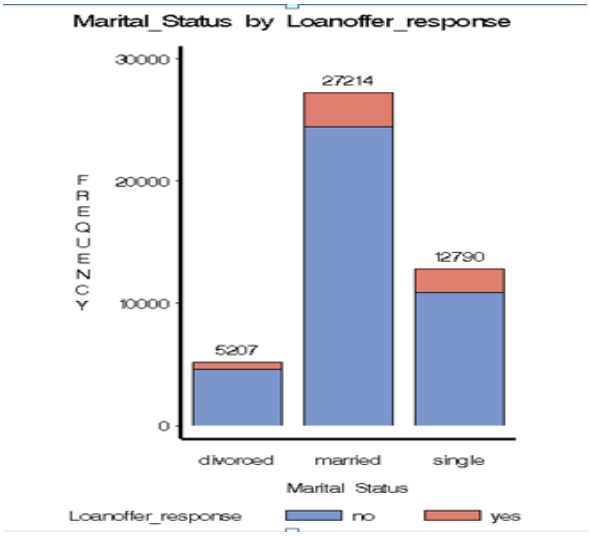


Fig 7. Marital Status Vs Loan Offer Response

Main Findings

1. The stepwise logistic regression model outperformed the other models in the validation data by predicting the target variable 88.94% correctly.

Calculation for Prediction Accuracy

Prediction Accuracy = (1-Misclassification rate)* 100

Prediction Accuracy for Stepwise Logistic Regression Model = 88.94%

Model Name	Validation Misclassification Rate
Stepwise Logistic Regression	0.110610
Entropy Tree	0.115676
Support Vector Machine	0.116678
Neural Network	0.116956
Decision Tree	0.116957

Table 1: Summary of Model Comparison using Validation Misclassification.

2. The final Stepwise Logistic Regression Model had the following significant predictors:

- Number of Contacts made after last campaign
- Credit Default Status of the Consumer
- Communication Type
- Housing Loan

Variable	Comparisons	Odds Ratio	P-Value
Contacts previous days	Less than month Vs Never	1.044	< 0.0001
Credit Default	yes Vs no	1.854	< 0.0001
Communication Type	Cellphone Vs Telephone	3.116	< 0.0001
Housing Loan	no Vs yes	2.715	< 0.0001

Table 2: Odds-Ratios of Significant Predictors in the Best Model.

Further analysis showed that the less the number of contacts made in the last campaign, more are the chances of customers responding to the campaign. The customer who had not defaulted in the past is more likely to have higher response rate when compared to the customer with bad credit status. The customer who had been contacted in the past via cellphone have responded more than the one's contacted via land line. The customer currently

having a long term housing loan is more likely to subscribe to the loan offer than the ones who are not currently having a housing loan.

3. Response Analysis

Decile	Cumulative % Response	Cumulative Lift
1 (TOP)	27.58	5.51
2	19.29	3.85
3	14.50	2.93
4	11.55	2.31
5	9.55	1.91
6	8.11	1.62
7	7.05	1.41
8	6.21	1.24
9	5.54	1.10
10	5.00	1.00

Table 3: Cumulative Lift of the Best Model

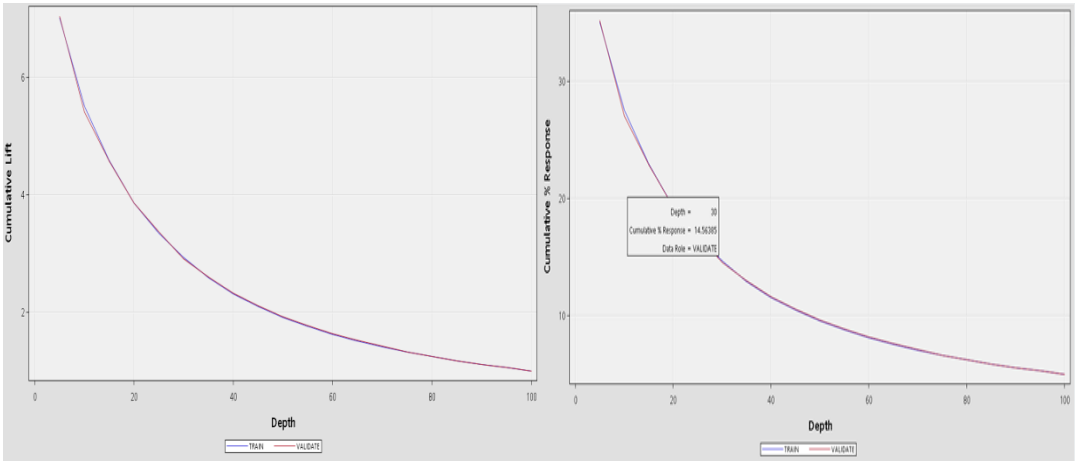


Fig 8. Cumulative Lift Chart **Fig 9. Cumulative Percentage Response Chart**

If the top 3 decile customers are selected based on the best model, the cumulative response rate is 14.5% in contrast to the baseline response rate of 5%. The cumulative lift chart signifies model's ability to beat the 'no model' case or average performance. In this case we have taken the model's average performance as 5%. For example, from Table 3 we see that the lift for the top two deciles is 3.85. This indicates that by targeting only these consumers we would expect to yield 3.85 times the number of responders found by randomly targeting the same number of consumers.

CONCLUSION & FUTURE WORK

Today marketers are trying to make most of their data for effective and efficient marketing campaigns. The response to the direct marketing in the finance industry is usually less than 2% which is typically very low. In this data, we find cumulative lifts of close to 4 at 2nd decile via a finely tuned logistic regression model.

In this study we have showcased the importance of the response models in direct marketing campaigns by measuring the cumulative response and as well as identifying the significant predictors for improving the response rate. In future we should use more of the client based data for analyzing the response rate of the customers to the marketing campaign. This will allow the companies to develop better strategies to promote their campaigns and target the right customer.

REFERENCES

[1] Lilien, Gary L., Philip Kotler, and K. Sridhar. Moorthy. Marketing Models. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[2] Sorger, Stephan. Marketing Analytics: Strategic Models and Metrics. S.l.: S.n., 2013.

[3] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:

Arun K Mandapaka, Oklahoma State University, Stillwater OK, Email: arun.mandapaka@okstate.edu

Arun K Mandapaka is a second year graduate student majoring in Management Information Systems at Oklahoma State University. He has three years' experience of using SAS® tools for Marketing Analysis, Credit risk analysis and Predictive Modeling. He is a SAS Certified Advanced Programmer for SAS 9 and SAS Certified Predictive Modeler using SAS Enterprise Miner 7. In April 2013, he received his SAS® and OSU Data Mining Certificate.

Amit Singh Kushwah, Oklahoma State University, Stillwater OK, Email: amit.kushwah@okstate.edu

Amit Singh Kushwah is a second year graduate student majoring in Management Information Systems at Oklahoma

State University. He has two years' experience of using SAS® tools for Marketing Analysis, Credit risk analysis and Predictive Modeling. . He is a SAS Certified Advanced Programmer for SAS 9 and SAS Certified Predictive Modeler using SAS Enterprise Miner 7. In April 2013, he received his SAS® and OSU Data Mining Certificate.

Goutam Chakraborty, Oklahoma State University, Stillwater OK, Email: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is a professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU business analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and Co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.