Paper 1407-2017

Benefit of using Clustering as input to Propensity to Buy Predictive Model

Krutharth Peravalli, Sumit Sukhwani and Dr. Dmitriy Khots

West Corporation

ABSTRACT

Propensity to buy models comprise one of the most widely used techniques in supporting business strategy for customer segmentation and targeting. Some of the key challenges every data scientist faces in building predictive models are the utilization of all known predictor variables, uncovering any unknown signals, and adjusting for latent variable errors. Solving these challenges ensures the creation of a robust model. Often, the business demands inclusion of certain variables based on previous understanding of process dynamics. To meet such client requirements, these inputs are forced into the model resulting in either a complex model with too many inputs or a fragile model which may decay faster than expected.

WEST Corporation's Center for Data Science (CDS) has found a work around to strike a balance between meeting client requirements and building a robust model by making use of clustering technique in the data preparation process. A leading telecommunication services provider uses West Corporation's SMS Outbound Notification Platform to notify their customers about an upcoming Pay-Per-View event. WEST has built a propensity to buy predictive model using customer's account related attributes, demographic attributes and customer purchase history. As part of this process, client has identified a few variables as key business drivers and CDS used those variables to build clusters which were then used as inputs to the predictive model. In doing so, not only all the effects of the client mandated variables were captured successfully, but this also helped to reduce the number of inputs to the model, making it parsimonious. This paper illustrates how WEST has used clustering technique in the data preparation process and built a robust model.

INTRODUCTION

WEST Corporation Outbound Notification Platform

West Corporation offers customers of various entertainment companies the flexibility to order Pay Per View (PPV) movies and sporting events using IVR and SMS channels. In particular, when a customer wishes to order a PPV events, they can text keywords such as "MOVIES" or "PPV" to a specific short code using their mobile phone and the system will start interacting with the customer who can order a movie or even browse different titles while on the go. West uses its outbound platform to send Voice, SMS and Email notifications to its customers. Notification platform primarily consists of a campaign management tool, multi-channel integration environment, and channel-specific delivery mechanisms.. Notification process is initiated with loading of files consisting of customer's phone number/email address into the campaign management tool using either SFTP or a web service. Once files are loaded, all records are scrubbed against "do not call" lists and other filtering criteria and notification date/time/type are configured within the campaign management tool. Multi-channel integration process then pushes the contacts to the corresponding system depending on the type of notification set in the campaign management tool.

Figure 1 shows high level overview of outbound platform and process. The platform has an ability to initiate communication through one way or two way notifications. In one way notification, platform does not interact with the customers, whereas in two way, customers can reply back and interact with the platform.

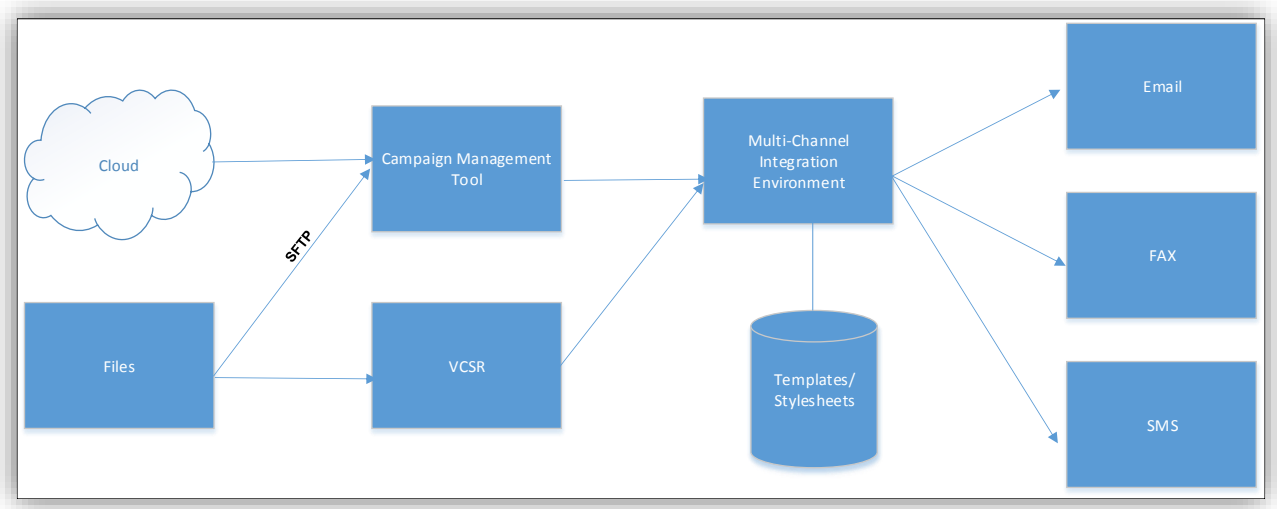

Figure 1. High Level Outbound Notification Process

DATA MINING PROCESS

Business Understanding

One of West Corporation's leading Telecom services client uses Outbound Notification Platform to send two-way SMS notifications to its customers for PPV events. Customers interested in buying PPV event have an option to respond to the notification and make a purchase through SMS or IVR. When purchases are made through IVR, data is stitched together with SMS data to create complete picture of purchase history. For more information on IVR data please refer Khots, Dmitriy. 2015.

As a business rule, notifications are sent only to the group of customers who have opted in to receive marketing messages. A total of ~9M notifications were sent out for 18 different PPV fight events in 2016 alone. Overall response rate to outbound SMS notifications was 0.63%, indicating percentage of people made purchase either through SMS or IVR. CDS team was involved in performing detailed descriptive analysis on the PPV SMS Campaigns and suggested to build a propensity to buy predictive model to improve the response rates.

Data Understanding

Propensity to buy predictive model has been built using historical SMS notification campaign data. There were a total of 6 SMS notification campaigns conducted in the time frame selected for the model build. Variables used for building model are collected from different data sources and are classified into one of the below categories:

Customer Specific Variables: Client provided account information such as Customer's Account Type, Account Status, Various value scores, and US Census Region were selected as input to the model.

Demographic Variables: Demographic variables such as Average Household Size, Average Household Income, Presence of children in the household, etc. are input to the model. This data is available to West from a third party vendor. Client was interested in looking at certain demographic characteristics, which were incorporated in the model using a clustering approach.

Data Preparation and Target Variable Identification

Step1: Extract Account level information of all the customers opted in to receive marketing messages.

Step2: Join Account data from step 1 with demographic information available at West using Zip9 or Zip5 codes of the accounts.

Step3: Build clusters on demographic variables using SAS clustering procedure “Proc Fastclus”.

```
/*PROC FASTCLUS Code*/

Proc fastclus data=dataprepstep3 out= cluster_data
maxclusters=10 maxiter=100;
var

Variable 1
Variable 2
Variable 3
Variable 4
Variable 5;
run;
```

Step4: Join the clustering output from step 3 to data from step 2.

Step5: Target Variable Creation - If an account has ordered a PPV event between August-2016 and November-2016, then target=1 else target=0 (Binary Target).

Predictive Modeling in SAS Enterprise Miner

Sampling: Finalized dataset is split into Model (70%) and Holdout (30%) sample using Proc Surveyselect procedure with option of Simple Random Sampling technique available in SAS

```
/*Create Holdout Data*/

Proc surveyselect data=ACCOUNT_info
method=srs n=XXXX
seed=YYYY out=Account_Info_Holdout;
run;

/*Create Model Data*/
Proc SQL;
Create table ACCOUNT_info_Model as
Select * from ACCOUNT_info where account not in
(Select distinct account from Account_Info_Holdout);
Quit;
```

Imputations & Transformations: Model Data is then loaded into Enterprise Miner. Data Partition node is used to split the data into Train (80%) and Test (20%) Sample. Less than 1% of data has missing values and they are imputed using impute node. Interval variables are imputed by replacing missing information with mean value and Class variables are imputed by replacing missing information with mode value of the respective variable. Interval variables are then normalized and dummy indicators are created for class variables using Transform Node available in SAS Enterprise Miner. Correlation statistics are then run using SAS procedure “PROC CORR” to quantify the degree of association between the continuous variables available in dataset. All the variables which show co-relation of greater than 0.4 or less than -0.4 are evaluated from business perspective and then a decision is taken to either remove or keep as an input to model.

Model Building and Evaluation: Decision Tree, Logistic Regression and Neural Network techniques were tried on the transformed dataset available from the previous step. The output from the three model nodes were then fed into the Model Comparison node in SAS EM. Logistic Regression technique with Stepwise Selection turned out to be the best modeling technique based on KS Statistic. Final Model has a KS of 29. Stability of the model is tested by running the final model on Test Data (KS=30). Figure 2 shows the SAS EM screenshot of the entire data and model flow.

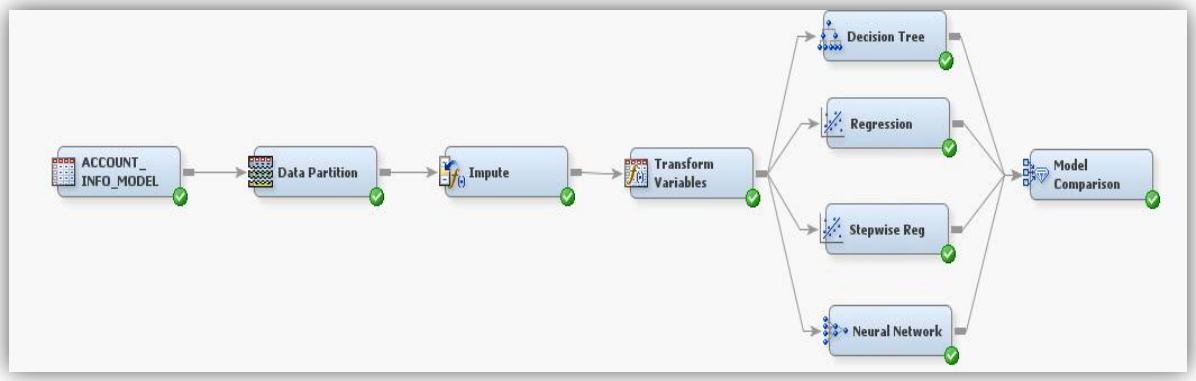

Figure2. Model Diagram

Logistic Regression Results

Variable importance is evaluated using the Chi Square statistics of the variables selected by the model, with variable having the highest Chi Square being the most important. Total of 12 variables are selected by the model, out of which Variable 1 is the most predictive variable with Chi - Square value of 1213.2254 while Variable 12 is the least predictive. Figure 3 lists all selected variable details and chi-square tests for each variable.

Summary of Stepwise Selection					
Step	Entered	DF	Number In	Chi - Sqaure	Pr > ChiSq
1	Variable 1	1	1	1213.2254	<.0001
2	Variable 2	1	2	1100.9932	<.0001
3	Variable 3	1	3	179.5216	<.0001
4	Variable 4	1	4	147.6527	<.0001
5	Variable 5	1	5	142.1543	<.0001
6	Variable 6	1	6	88.0957	<.0001
7	Variable 7	1	7	132.6869	<.0001
8	Variable 8	1	8	42.7401	<.0001
9	Variable 9	1	9	40.2705	<.0001
10	Variable 10	1	10	28.9108	<.0001
11	Variable 11	1	11	29.5868	<.0001
12	Variable 12	1	12	15.2267	<.0001

Figure3. Variable Importance

Cumulative Lift Chart - Accounts are divided into deciles based on their propensity to buy using Proc Rank procedure in SAS. Lift charts are used to demonstrate the model performance on the train and test dataset. It is evident from the cumulative lift charts of Train and Test datasets (Figure 4 and 5) that the top 5 Deciles contribute to ~80% of the sales.

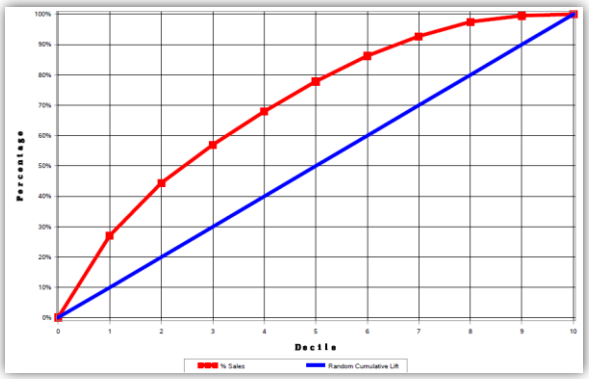

Figure4. Training Lift Chart

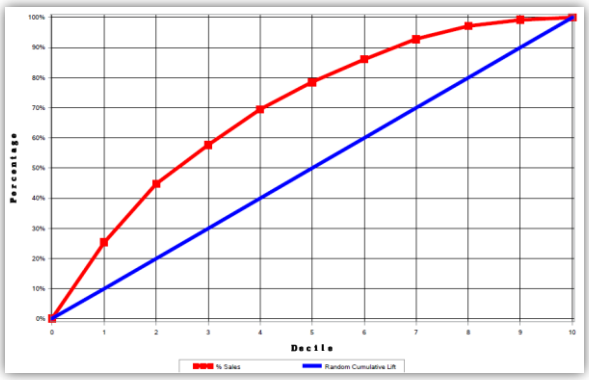

Figure5. Test Lift Chart

Scoring/Deployment process

Scoring new accounts involves creation of all of the variables that are in the model building process. SAS Enterprise Miner Score node outputs the score code which is used in future to score the new input dataset. Figure 6 shows deployment strategy followed at West. As part of the process, each account is assigned a decile based on its propensity to buy. After decile assignment is done, strategy is designed to notify customers for future PPV events. Notifications are sent out on the day of the event, one day before, two days before or maximum three days before the event. Accounts belonging to the top deciles are sent notifications on all four days to evaluate the best days for response.

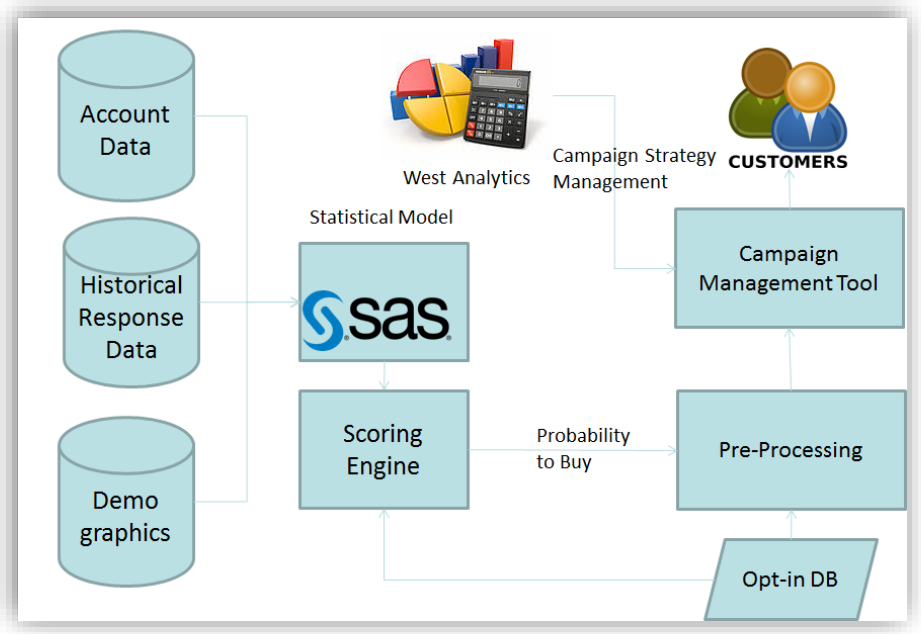

Figure6. West Corporation's Deployment Architecture

RESULTS

When comparing current PPV campaign results to past like campaigns, there is substantial overall lift - 40%+ better performance in both response rates and revenue. The best accounts generated 1.15% response rate, while the worst accounts generated just 0.15% response rate.

CONCLUSION

Center for Data Science team at West uses SAS Enterprise Miner and Predictive Analytics path to help their clients in improving business performance and generating additional revenue. Modeling exercise also helps operational teams at WEST to design proactive treatment strategies.

REFERENCES

Khots, Dmitriy. 2015. Unstructured Data Mining to Improve Customer Experience in Interactive Voice Response Systems. *Proceedings of the SAS Global Forum 2015*, Dallas, TX, SAS. Available at https://support.sas.com/resources/papers/proceedings15/3141-2015.pdf.

ACKNOWLEDGMENTS

The authors would like to thank Rhonda Gibler for her encouragement and support throughout the process. The author would like to thank Shruti Palasamudram for her guidance and suggestions throughout this study. The authors would also like to thank Dr. Goutam Chakraborty for inviting the team to present this topic at SAS Global Forum 2017.

CONTACT INFORMATION

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

Krutharth Peravalli, West Corporation

E-mail: kperaval@west.com

Sumit Sukhwani, West Corporation

E-mail: ssukhwani@west.com

Dr. Dmitriy Khots, West Corporation

E-mail: dkhots@west.com