Improving Customer Engagement at VMWARE through Analytics

(Harvard Business Case Study)

Group Members:

- Suhail Shaikh
- Vikalp Mehta

Improving Customer Engagement at VMWARE through Analytics (Harvard Business Case Study)

(1) How does B2B personalized marketing differ from B2C marketing?

It is more fundamental to understand the fact that Personalization of marketing is as important for B2B as it is for B2C. for decades personalization has been the foray for B2C marketing but now they are mixing up a bit and B2B is also focusing on crafting and tailoring their brand strategies to be more personalized and focused on end consumer.

- B2B marketing campaigns are adopting B2C strategies on different platforms. For example, Social media presence for B2B brands has been more mandatory now a days, as customers are seeking interaction with brands and in this fast-changing world a prompt reply with a great engagement strategy will be a win-win situation. B2B companies are not known for their fast replies and quick resolution, a prompt engagement structure online will improve customer sentiment about the brand. This also will help boosting up its personalized interaction with customer.
- We are living in an era where information is as easily available to make a first impression and perception about a brand. B2B brands can have a larger impact of this as one bad perception and a contract is lost. B2B brands needs to work upon their website and point of contacts with potential customers. A good experience with customer at those touch points will improve next steps in B2B process.
- B2B companies are now moving into the experiential marketing tactics as well. I have seen many tire manufacturers like JK tire in India indulging in experiential space with people. They have this one stop shop for all tire needs for every segment. Where they let consumer feel the brand and know about it in detail. This has helped them generate a great awareness regarding the brand.
- B2B companies must understand the decision-making process with great details as opposite to B2C it has more decision makers and more processes and systems in place. In order to get things done company needs to understand the complexity of decision making, Understand the customer journey and need of experience among the few.

(2) How is the propensity to respond model different from the traditional propensity to buy model?

In these changing times every B2B company is looking for some change in traditional personalized marketing techniques for its customer. VMW is no different as they want to use their data for some good reason.

- Changing strategy based on data helps predicting the right customer and personalizing experience for them, which in turns saves cost and reduces conversion time.
- This “Propensity to respond model” will use predictive analytics to predict further course of action for a set of customers and will help generate appropriate solution for them.
- This model will also segment its customers basis demographics, geographic and physiological factors. This segmentation will further help company to tailor its marketing strategy as per segmentation results.

Improving Customer Engagement at VMWARE through Analytics (Harvard Business Case Study)

(3) What kind of modeling problems can Kiran and Anit expect when the classes are not represented adequately? Which classification models should be used to handle these problems?

With advent of computing it has become easy now to compute and solve mathematical problems which require a lot of processing. An imbalance data might have given you nightmares earlier because of following reasons.

- Imbalance data can cause accuracy paradox, as it will lead to prediction of most repeated class in data.
- Imbalance data will lead to inability to apply statistical techniques which are more suitable for balanced data than imbalanced one.
- Using imbalance data introduces bias in analysis intentionally and it only few methods like logistic regression and SVM are not prone to it.

(4) VMW has identified more than 600 predictor variables. The data is also highly imbalanced (why?). Do you think that techniques such as logistic regression can be applied when the number of variables is large? Justify your answer. If your answer is no, what variable reduction techniques can be used?

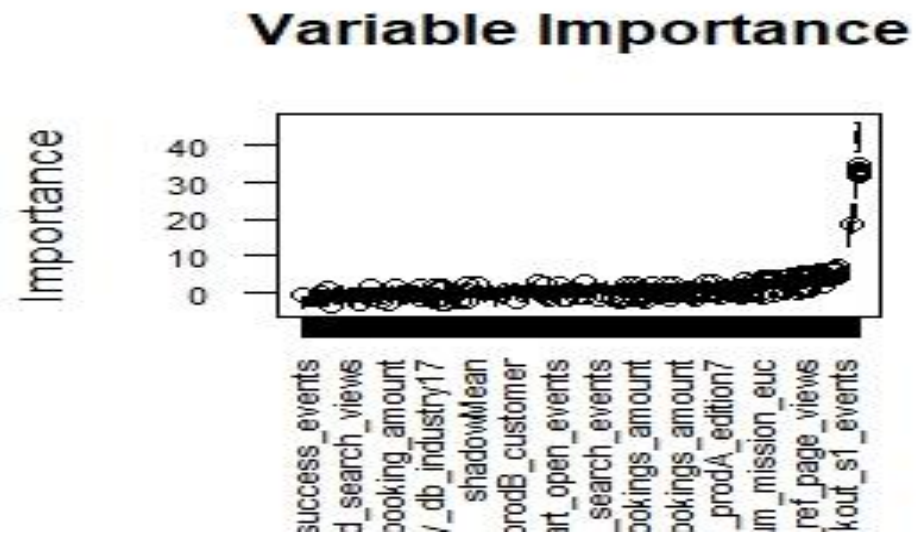
Data is highly imbalanced in VM ware case as more than 97% of the predicting class is "0" class. That means that out of 50000 instances 48600 instances are predicting one class and that is where the problem in modelling arises. There are 700 variable classes in the data. This scenario in Machine learning is known as "**curse of Dimension**". It makes data analysis very difficult and tiring process as computational power of different software's and tools are limited. It is optimum to do feature selection and identification to put important features in model to predict better.

While dealing with such cases it is always important to clean data as large data contains many anomalies. There are plenty of method for feature selection like subset, wrapper and embedded methodssubset methods: These are subset method for feature selection as they select a combination of variable basis their importance and keep trying different selection to come out with few. This method works effectively when there is limited number of feature present in Data. As it takes combination of 2 raise to power of number of feature present in data set and as the number of feature increases in data set to more than 10, combination increases to 1024.

- **Wrapper Method:** This method includes forward, backward and recursive feature selection methods. These method selects and drop variable one by one basis their p values and keep important ones in group. This performs best result while forming feature groups. These techniques also help to reduce collinearity among features as they check and add and remove variable.

Improving Customer Engagement at VMWARE through Analytics (Harvard Business Case Study)

- **Embedded Method**: This method checks for overfitting and assign penalty to large coefficient of features. This helps settling trade off between Bias and Variance. Lasso is the method frequently use for feature selection.
- We also have some of the algorithms like Random Forest and Logistic regression which helps determine importance of variable subsequently. These methods select variable with different measures like Gini impurity and entropy.
- This model improvises performance basis feature selection methods like Boruta, Ranger, Random Forest and Logistic Regression. This has reduced number of important features to 92 from 700. This improves predictive power of model and makes it more efficient. I am attaching sample output of Boruta applied in Feature selection.



(5) What are the efficient meta-algorithms that you can use to aggregate the model? Why do you think the model performs better than the individual models?

Meta algorithm takes different algorithm and combine those to predict output in a more efficient way. Meta Algorithm can combine different model together and uses their individual vote to predict output better.

Reason why meta algorithm performs better are as follows.

- Meta algorithm is like taking each vote into account. For example, if you like to invest in a stock and you gather advice from social media, investor, your friend, a stock broker and a

journalist. Individually they all may have different advice and suggestions to invest. Best way might be to give equal weightage to all the advice and combine them together to decide.

- This method makes our prediction better and effective as each model has its own way and experience to predict. Each model comes with different hypothesis, modelling technique, number of instances used to predict and seed. This helps in better prediction.
- Different learning algorithm helps making a model stable. As the final output is not dependent upon the single model but on the vote of different models. We may or may not use same model to predict.

(6) Use the sample data provided to develop a Random Forest model. Comment on the model development and accuracy of the model.

Random Forest is prone to imbalance in the Data-set and generally more biased towards the majority class. We did try to improve this by implementing Smote technique which balances out the data i.e. gives equal number of rows for all the target classes which prepares the data well for analysis.

Another important thing which is important to note in case of Random Forest is that, evaluation measures are not limited to accuracy as it is biased towards most frequent target class in Data. Hence precision is the most sought metric to evaluate performance of Random Forest.

We compared the precision before and after running Smote and found that precision is increased to 78% from 59% after correcting the imbalance.

```
confusionMatrix_RF


          Reference
Prediction     0     1     2     3     4     5
         0 48565     0     0     9     0     0
         1    36   800     9     3     0     0
         2     0     0     0     0     0     0
         3     0     0     0     4     0     0
         4     0     0     0     0    86     0
         5     0     0     0     0     4   490

Overall Statistics

               Accuracy : 0.9988
                 95% CI : (0.9984, 0.9991)
    No Information Rate : 0.9719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.978

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2  Class: 3 Class: 4 Class: 5
```

Improving Customer Engagement at VMWARE through Analytics (Harvard Business Case Study)

```
                     Class: 0  Class: 1  Class: 2  Class: 3  Class: 4  Class: 5
Sensitivity            0.9993   1.00000   0.00000  2.500e-01   0.95556  1.000000
Specificity            0.9936   0.99902   1.00000  1.000e+00   1.00000  0.999919
Pos Pred Value         0.9998   0.94340       NaN  1.000e+00   1.00000  0.991903
Neg Pred Value         0.9749   1.00000   0.99982  9.998e-01   0.99992  1.000000
Prevalence             0.9719   0.01600   0.00018  3.200e-04   0.00180  0.009799
Detection Rate         0.9712   0.01600   0.00000  7.999e-05   0.00172  0.009799
Detection Prevalence   0.9714   0.01696   0.00000  7.999e-05   0.00172  0.009879
Balanced Accuracy      0.9964   0.99951   0.50000  6.250e-01   0.97778  0.999960
```

RF after handling Imbalance

Average Prec= (0.9434 + 1 + 1 + 0.991903)/5 =78.70

```
onfusionMatrix_RF1
Confusion Matrix and Statistics

          Reference
Prediction     0     1     2     3     4     5
         0 48601     0     0    13     0     0
         1     0   800     9     3     0     0
         2     0     0     0     0     0     0
         3     0     0     0     0     0     0
         4     0     0     0     0    86     0
         5     0     0     0     0     4   490

Overall Statistics

               Accuracy : 0.9994
                 95% CI : (0.9992, 0.9996)
    No Information Rate : 0.9719
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9894

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000  1.00000  0.00000  0.00000  0.95556 1.000000
Specificity            0.9907  0.99976  1.00000  1.00000  1.00000 0.999919
Pos Pred Value         0.9997  0.98522      NaN      NaN  1.00000 0.991903
Neg Pred Value         1.0000  1.00000  0.99982  0.99968  0.99992 1.000000
Prevalence             0.9719  0.01600  0.00018  0.00032  0.00180 0.009799
Detection Rate         0.9719  0.01600  0.00000  0.00000  0.00172 0.009799
Detection Prevalence   0.9722  0.01624  0.00000  0.00000  0.00172 0.009879
Balanced Accuracy      0.9954  0.99988  0.50000  0.50000  0.97778 0.999960
```

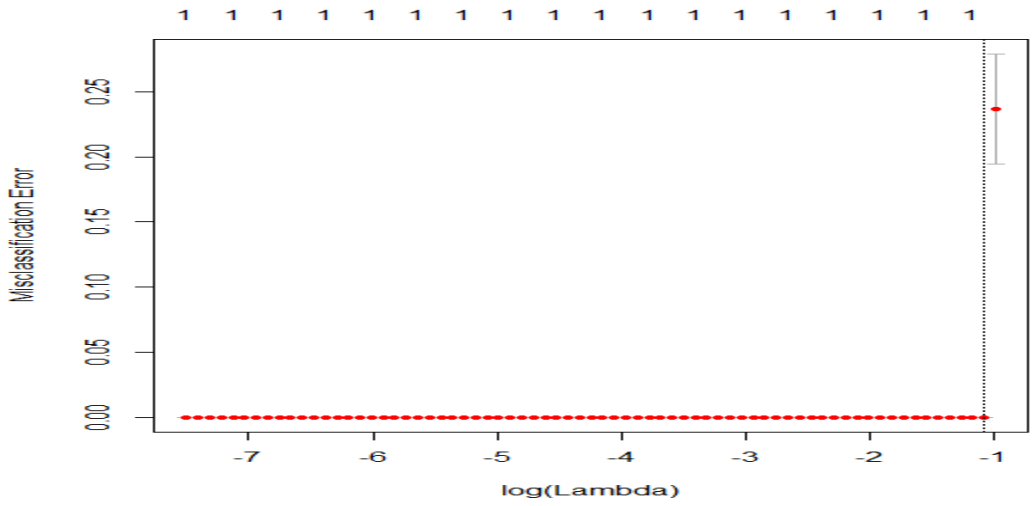
RF without handling Imbalance

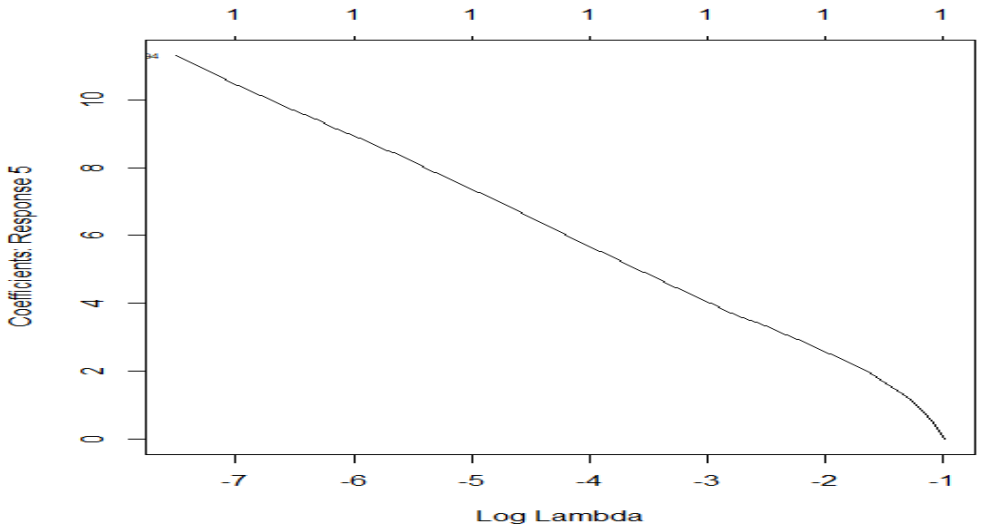
Prec =(0.98522 + 1 + 0.991903)/5 =59.54

Improving Customer Engagement at VMWARE through Analytics (Harvard Business Case Study)

(7) How different are regularized logistic regression models from standard logistic regression models? When should L1, L2 regularization be used to model the data? Develop a regularized logistic regression model on the given sample data. What insights do you obtain from this model?

We used cross validation with ridge and lasso algorithm to find optimal value of regularization parameter corresponding to minimum value of cross validation error. Plotted values are shown below for the case. Corresponding value of lambda for minimum misclassification error is .34.





We have to find out the optimal value of lambda at which error is minimum and coefficients are scaled to optimum value.

Improving Customer Engagement at VMWARE through Analytics (Harvard Business Case Study)

(8) Develop a couple of extreme gradient boosting models with different values for parameters (depth, eta, etc.) Discuss how the models differ from each other.

We made two xgboost model with different eta parameters .01 and .001. We found that precision increased for smaller value of eta. It reached to 98% from 95%. For imbalanced Data set it is always necessary to do resampling and we have done it using Smote technique. It has balanced out the predicting classes.

In case of imbalanced data set it is necessary to observe another evaluation matrix like precision because of accuracy paradox. Precision state the model accuracy and does not relate to real values.

XGBOOST- maxdepth=10 and eta=0.01

Precision=(0.9999+0.9947+1+1+1)/5=99.89

```
ConfusionMatrix_XGB_maxdepth_10
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5
         0 2096    0    0    0    0    0
         1    0 2129    1    0    0    0
         2    0   11 2069    0    0    0
         3    0    0    0 2043    0    0
         4    0    0    0    0 2126    0
         5    0    0    0    0    0 2026

Overall Statistics

               Accuracy : 0.999
                 95% CI : (0.9983, 0.9995)
    No Information Rate : 0.1712
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9988

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   0.9949   0.9995   1.0000   1.0000   1.0000
Specificity            1.0000   0.9999   0.9989   1.0000   1.0000   1.0000
Pos Pred Value         1.0000   0.9995   0.9947   1.0000   1.0000   1.0000
Neg Pred Value         1.0000   0.9989   0.9999   1.0000   1.0000   1.0000
Prevalence             0.1677   0.1712   0.1656   0.1634   0.1701   0.1621
Detection Rate         0.1677   0.1703   0.1655   0.1634   0.1701   0.1621
Detection Prevalence   0.1677   0.1704   0.1664   0.1634   0.1701   0.1621
Balanced Accuracy      1.0000   0.9974   0.9992   1.0000   1.0000   1.0000
```

Improving Customer Engagement at VMWARE through Analytics (Harvard Business Case Study)

(9) Based on the different models results, what would be your final recommendation to create the propensity to respond model?

We have created a list of models to determine the effectiveness in the given situation and found out that models like Logistic regression and naïve byes were quite low on precision score. Random Forest did try to improve it significantly but xgboost came out to be a clear winner as its precision was highest among all other models. I would strongly recommend implementation of XGboost to prepare a propensity to respond model.

Model Name	Precision
Logistic Regression	54%
Naïve Byes	58%
Random Forest	79%
Xgboost	98%

(10) Discuss the possible deployment strategies for the model results so obtained.

As we can see that we have a great precision while predicting all the classes from XG boost. It would be great to strategies as per following.

- This model needs to be monitored over period to fetch maximum result. After successful runs and re-runs on real time data it may be used to chart out strategies.
- This case has precision of class 5 as one. This is the most desired class for VMW. If a consumer gets identified in this class sales department should not leave any stone unturned for conversion to potential sale.
- Some part of the model result can be implemented on a pilot basis in some department and their observations and results can be implemented across organization.

```
ConfusionMatrix_XGB_maxdepth_10
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5
         0 2096    0    0    0    0    0
         1    0 2129    1    0    0    0
         2    0   11 2069    0    0    0
         3    0    0    0 2043    0    0
         4    0    0    0    0 2126    0
         5    0    0    0    0    0 2026

Overall Statistics
```

Improving Customer Engagement at VMWARE through Analytics (Harvard Business Case Study)

```
               Accuracy : 0.999
                 95% CI : (0.9983, 0.9995)
    No Information Rate : 0.1712
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9988

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            1.0000   0.9949   0.9995   1.0000   1.0000   1.0000
Specificity            1.0000   0.9999   0.9989   1.0000   1.0000   1.0000
Pos Pred Value         1.0000   0.9995   0.9947   1.0000   1.0000   1.0000
Neg Pred Value         1.0000   0.9989   0.9999   1.0000   1.0000   1.0000
Prevalence             0.1677   0.1712   0.1656   0.1634   0.1701   0.1621
Detection Rate         0.1677   0.1703   0.1655   0.1634   0.1701   0.1621
Detection Prevalence   0.1677   0.1704   0.1664   0.1634   0.1701   0.1621
Balanced Accuracy      1.0000   0.9974   0.9992   1.0000   1.0000   1.0000
```

(11) Discuss the business implications of the models developed here. What will be the value addition to the marketing department and the sales department of VMW because of this exercise?

- Given some set of conditions basis location, product groups, download frequency and recency and target_hol category, we can always predict the next stage of customer decision cycle and personalize our marketing as per requirement.
- This model can be validated on the daily stream of data which has been flowing in as to monitor the performance of it. This will enhance performance of model in longer time.
- We can use this model to segment end consumer and use tailor made personalize marketing strategy for them. This will save some cost for the company as it is sometimes quite a costly affair to do email or cookie marketing.
- This model will help Marketing Department to tailor made its personalization marketing for its customer and the stage at which they are.
- Marketing department will now be able to decide upon consumer behavior observed from online website and social media. This will help them build great engagement strategy for their target segment.
- Depending upon the class identified Sales team can catch up with customer and make them aware about the next stage parameters and processes. This will increase consumer experience and potential success of conversion.