2019

GroupAssigment
Task1

Saad Ahmed EL-15128
Maha Siraj EL-15100

Part 1: Business understanding

CRISP-DM standard process requires us to obtain understanding the business context (Larose, 2005). In our case, our objective is to build up model for the response to direct mail marketing. The model shall have balanced characteristics of both predictive accuracy and interpretability.

In our direct mail marketing response problem, our ultimate goal is to formula a model that assists the mailing decision, which leads to increase overall profit. The problem is for classification, so appropriate method such as logistic regression, discriminant analysis and support vector machines are considered as potential optimal model.

Cost & Benefits table

Cost benefits table for our clothing store requires management judgement and estimation. Firstly, the explanation for the positive and classification rate is demonstrated below:

Firstly, the estimated net profit per visit is the average revenue per visit times the Average net profit margin ratio. The formula is 113.89*0.1873 = 21.33 ($) per visit.

Then the cost of mailing per customer is $ 2, which is an approximate estimation including cost of material postage and handling cost.

This will give us a net profit per visit 21.33 - 2= 19.33 ($)

Finally, there are four different outcomes which are true negative, true positive, false negative and false positive. Their associated costs and explanations are demonstrated in the following table 1.

(Table 1 cost table)

Outcome	**Classification**	**Actual response**	**$ Cost**	**Rationale**
True negative	Nonresponse	Nonresponse	0	Since there is no mail sent and no response, there is no lost profit.
True positive	Response	Response	-19.33	The true customers are correctly classified and pointed out by mailing the marketing brochures, therefore the amount assigned to them is estimated profit minus cost of mailing.
False negative	Nonresponse	Response	21.33	The cost incurred is the opportunity costs of losing customers that are actually proved to be profitable.

False positive	Response	Nonresponse	2	Since the customer has been sent the marketing brochures and they did not respond, the loss is only the cost of mailing and other marketing per customer.

Part 2: Data understanding

2.1 Important Predictors Description

There are 49 predictors and one response in our dataset. We explain these predictors as table A1 shown in appendix. In the text blow, we list several predictors having potential contribution to estimating the classification of the response.

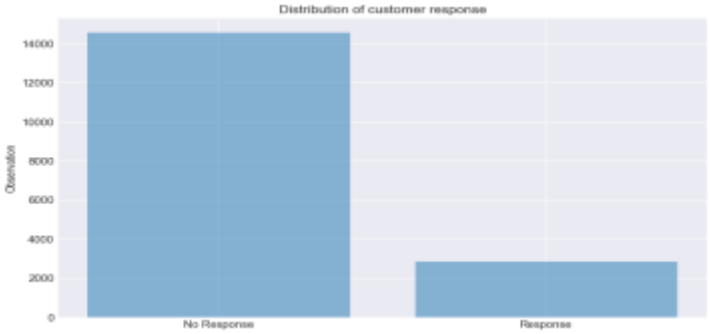
- *DAYS*: number of days customer has been on file
- *FREDAYS*: number of days between purchases
- *AVRG*: average amount per visit
- *MON*: total net sales
- *MAILED*: number of promotion mailed in the last year

2.2 Exploratory Data Analysis

Our exploratory data analysis has a focus on figuring out the individual distribution of the predictors and plotting out the relationship between each predictor and their classification.

Basic understanding of dataset

The dataset contains 48 covariates and 13044 observations. When the response variable coded as 1 indicates a response toward the marketing, and 0 for non-response towards the marketing. After the count for 0 and 1, 0 takes a large proportion of the whole dataset at 10910 while 1 only take a minority for 2134. Figure 1 and table 1 below shows that there is an imbalanced data distribution since most of the responses (83.6%) are nonresponse. This should be a concern for our model evaluation to account for the imbalanced issue, such as considering the precision rate.

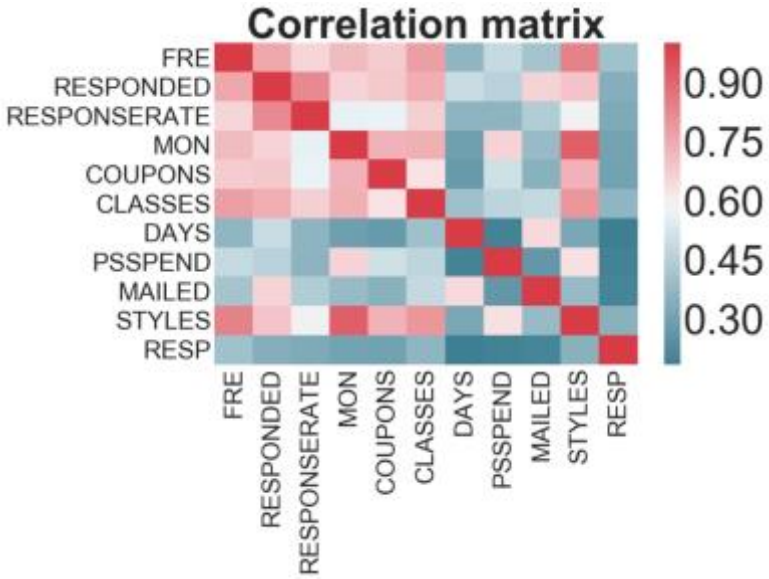


(Figure 1 Distribution of customer response)

(Table2: Response proportion)

Response proportion in training set	
0 (nonresponse)	1 (response)
0.836	0.164

2.2.1 Correlation

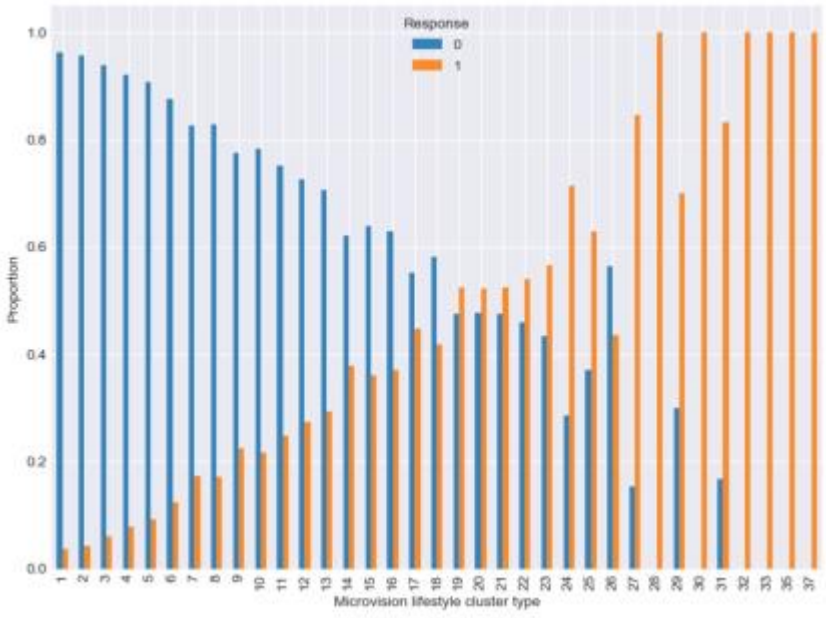


(Figure 2 correlation matrix)

We plot the correlation matrix as shown in figure A1 in appendix. Figure 2 above lists the most important features that are highly correlated with the target. FRE has the highest correlation with the response at 0.406, followed by *CLASSES* at 0.369. The rest of variables such as *STYLES*, *RESPONSED*, *RESPONSERATE*, *MON* and *COUPONS* all have a correlation that is above 0.3 in absolute term. It is reasonable to believe that these predictors have the best capability to our classification problem.

2.2.2 Individual predictors' distributions

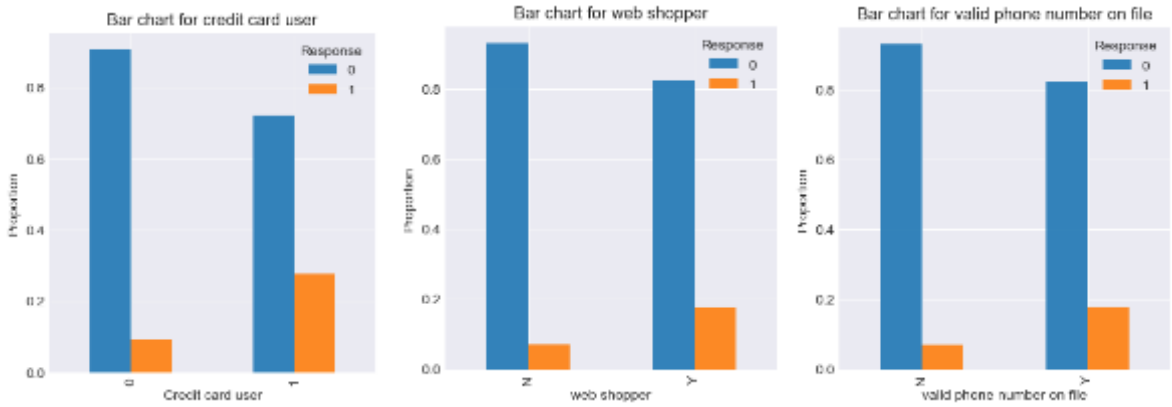
Categorical



(Figure 3 Bar chart for CLUST)

From figure 3 above, *CLUST* (micro vision lifestyle cluster type) may explain the classification of our response. For higher index of *CLUST*, customer is more likely to directly respond to marketing mail.

The categorical variables contain mainly 3 binary variables, including CC_CARD, WEB and VALPHON. The figures below illustrate the proportion of nonresponse and response within each category such as being credit card user or not.



(Figure 4 Bar chart for CC_CARD) (Figure 5 Bar chart for WEB) (Figure 6 Bar chart for VALPHON)

The bar charts give us the following ideas: firstly, the credit card users tend to be responsive to the marketing initiatives and on the other hand, non-credit card users tend not to respond to the marketing brochures. In addition, those customers who are web shoppers would be

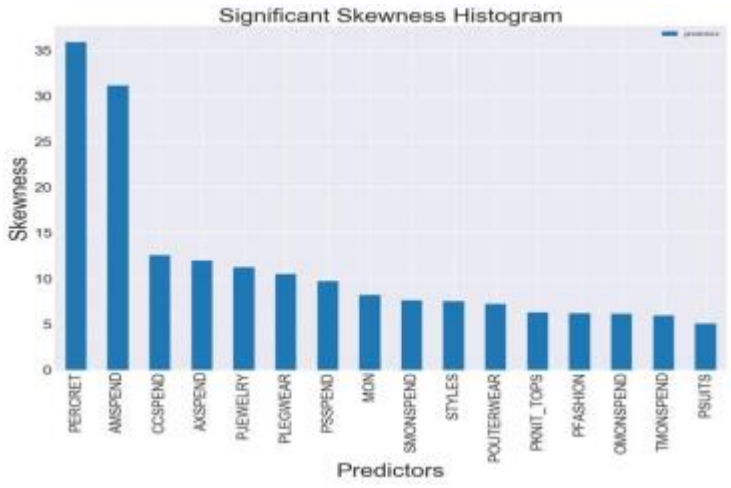
more responsive towards the marketing initiatives and therefore classified as response. Valid phone number on file does not indicate much difference for classifying the customers.

Numerical variables

ZIP-CODE and *HHKEY* (customer ID) are the numerical variables that provide less information for classification.

HHKEY is unique to each customer; hence, it contains no information related to predicting the customer response towards marketing promotions. *ZIP-CODE*, although provides information related to geographic location to customers, they should be treated as categorical variables. Since the *ZIP-CODE* will generate many dummy variables, it is not appropriate to include with the purpose of maintaining a reasonable dimension.

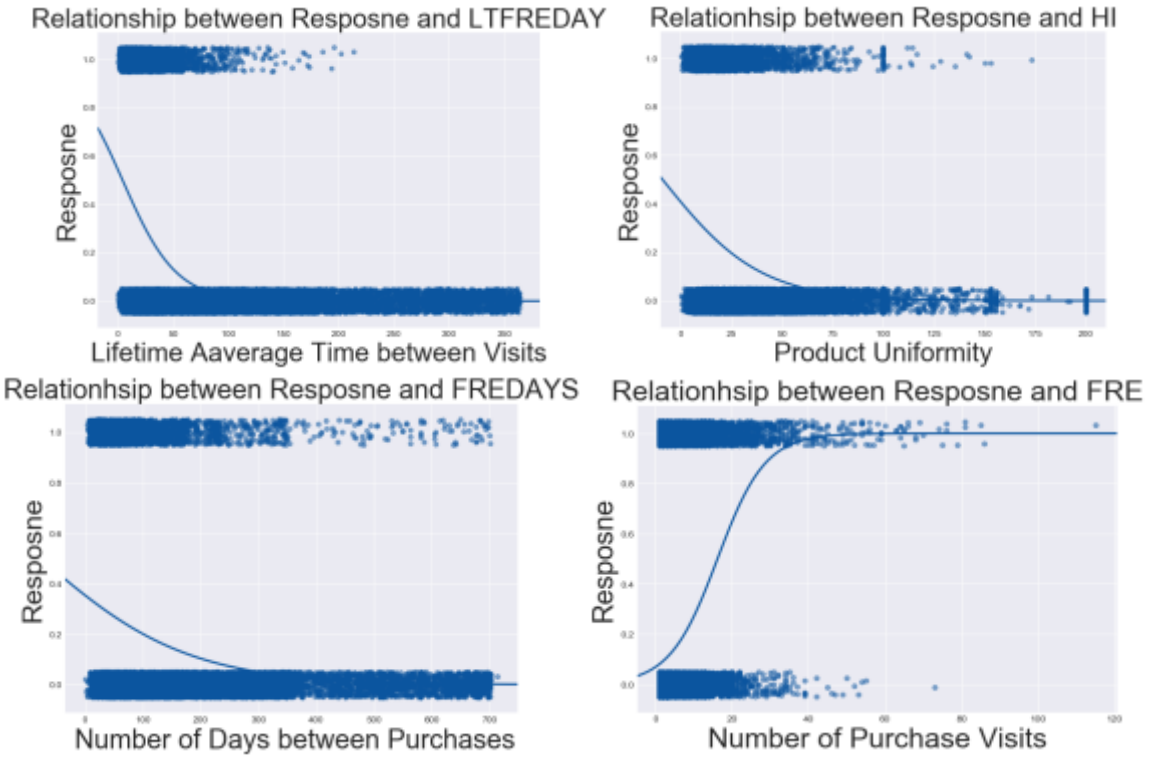
Skewness:



(Figure 7 Skewness for predictors (only include the predictors with an absolute value of skewness larger than 5))

As illustrated in figure 7 above, some variables such as *PERRCENT*, *AMSPEND*, *CCSPEND* and *AXSPEND* have skewness higher than 10. These variables should be a major concern. After performing their plots with transformation and considering the domain knowledge, we then decide to discard these variables. For example, *PERCENTAGE* is the ratio calculated from profit and revenue, which are the variables already included in the dataset. These kind of variables does not provide additional information to our model.

2.2.3 Bivariate relationship



(Figure 8 Plot of bivariate relationships for key continuous predictors)

Figure 8 for the bivariate relationships for the key continuous variables are demonstrated above. Firstly, FRE shows a positive correlation with the response. A higher *FRE* would associate with the customers who are responding to the market brochures. *LTFRIDAY* and *FRIDAYS* show the negative association with the response, since a higher figure in *LTFRIDAY* tends to be classified as nonresponse. Finally, the graph shows that a higher Hi tends to be classified as non-response to the marketing initiatives.

Part 3: Data Preparation with Feature Engineering

3.1 Removing irrelevant variables and creating dummy

To begin with, we check that there is no missing data. The dataset includes some variables that are not relevant to the classification, such as *HHKEY*. Also, covariates such as *ZIP-CODE* have many categories that do not add much information to classifying customers. So, we remove the *HHKEY* and *ZIP-CODE* in model building.

We prepare the data by creating dummy for categorical variables (*VALPHON*), clearly separating the training data with response 0 and response 1. This will assist us to figure out differences between non-response and response.

3.2 Box-cox transformation

To make the variables approximately normal distributed, we perform box-cox transformation for all continuous variables and decide to use the transformed variables as predictors.

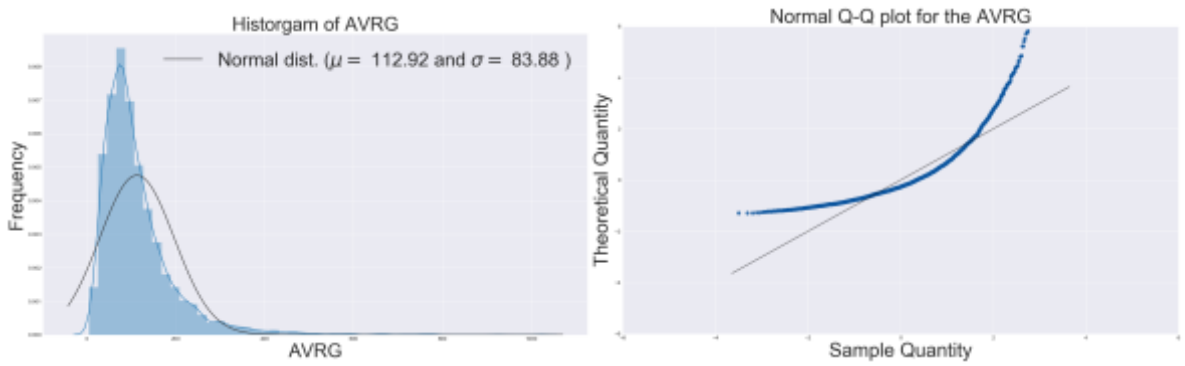
To begin with, we check that there is no missing data. The dataset includes some variables that are not relevant to the classification, such as *HHKEY* (customer ID). Also, covariates such as *ZIP-CODE* have many categories that do not add much information to classifying customers.

We prepare the data by creating dummy for categorical variables (*VALPHON*), clearly separating the training data with response 0 and response 1. This will assist us to figure out differences between non-response and response.

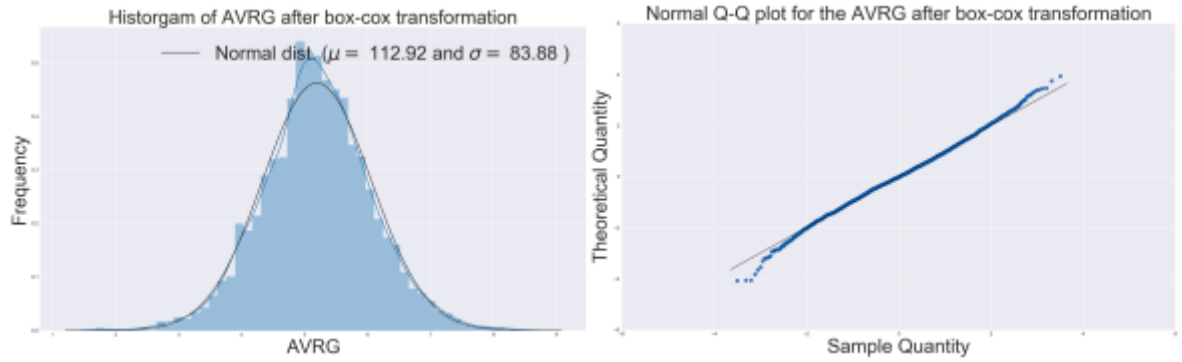
To make the variables normally distributed, we performed box-cox transformation for all variables and decide to use the transformed variables as predictors. The transformation results are impressive for some variables such as *AVRG* and *MARKDOWN*.

Box-cox successfully tackles the problem of skewness and abnormal tail behaviour. For example, prior to the box-cox transformation, *CCSPEND* has thinner tails and significant skewness. This has been improved after performing box-cox transformation.

For the rest of the numerical variables, we plot the histogram and perform the box-cox transformation, hoping to achieve normality and symmetry of the distribution. The results show that the normality improves for some variables such as *AVRG* and *SMONSPEND*. The achievement from box-cox transformation is shown below when comparing figure 9 and figure 10.



(Figure 9 Distribution of AVRG before transformation)

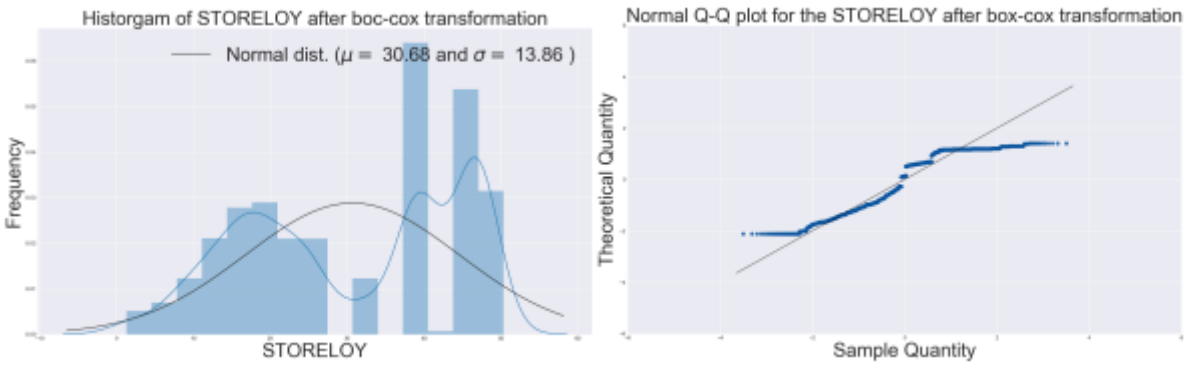


(Figure 10 Distribution of AVRG after box-cox transformation)

However, a set of predictors with a spike at 0 and remaining distribution above 0, such as spending at store. Our treatment towards this is to create 'flag variables', representing that there is not spending at store. The remaining predictors, despite the box-cox transformation, still have an asymmetric distribution, heavy tails or multimodal pattern. The selection towards these variables are based on our domain knowledge.

However, most the predictors after transformation still have an ugly shape distribution. One kind of distribution is multimodal.

Figure 11 illustrates that *STORYLOY* still has a distribution that contains multiple peaks. The QQ plot also shows that the behaviours on both tails are abnormal. The box-cox transformation may not be a good way to solve this multimodal issue.



(Figure 11 Distribution of STORYLOY after box-cox transformation)

3.3 Standardizing Variables and Split Dataset

One of the key step is to partition the data into training set and test set with 80% and 20% respectively. The training sets are used to fit different model, and the test sets perform independently for model evaluation.

After that, we standardised the numerical predictors of both training set and test set through minus the mean of training set and then dividing by standard deviation of training set, which will reduce the variability between different scale of predictors. And the logistic regression can also benefit from data transformation, which can be a way to account for nonlinearities and reduce the impact of predictor outliers.

The categorical variables are not necessary to standardised since the scale of variables are restricted to zero and one.

Part 4: Modelling methodology

4.1 Logistic regression

The logistic regression model is most widely used model for binary classification problem. It requires building up a linear combination of predictors. After that, the linear combination is plugged into a logit function, which transfers the response to a restricted interval from 0 to 1. Finally, classify the response to the one has the larger likelihood. As a generalised linear model, the input does not have any distributional assumption. In our case, our goal is to estimate the conditional probability and classify the customers into the response group or nonresponse group.

In our EDA discussion, we found that the values taken by numerical variables are not well separated between the response (1) and non-response (0). The possible values taken when response equal to 0 cover the values taken when responses equal to 1. This fact can be accounted for by the imbalance data, since most of the response variable take 0. Thus, we will focus more on results of precision than sensitivity.

The higher slope of the curve, the higher potential relationship between the predictors and classification.

In terms of the advantage of using logistic regression, it will work better if there is only one single decision boundary that is smooth and non-linear. Also, this model can be easily interpreted via odd ratio. However, it requires that every data point should be independent, otherwise the model will attempt to overweight the significance of the observations. Although this model predicts the results based on independent values, it might variate the accuracy of its predictions. To avoid overfitting and dimensional exposure, regularisation including ridge and lasso would be applied into logistic regression.

4.2 Support vector machines

Besides of logistic regression, support vector machine can be an effective method for our classification problem. The basic idea for SVM is based on maximal margin classifier, which is to identify a hyperplane that has the farthest distance to the training observations (James et al, 2013).

In our case, the plot for support vector machines in EDA part shows that the two classes might not be separable by a hyperplane, hence there is no maximal margin classifier that could not be applied here. Then we incorporate the support vector classifier with a soft margin to train a linear SVM model, which allows for some observations to be on the incorrect side of margin, or even on the incorrect side of the hyperplane. The parameters C can be set up, which determines the amounts and the seriousness of margin's violation that we can accept. C is of great importance for the model selection, since it plays an essential role to determine optimal bias-variance trade-off. A higher C indicates a higher width for margin and higher tolerance for the violation of margin, which is associated with high bias and lower variance. On the other hand, however, A lower C gives a lower width for margin and lower tolerance for margin violation, which therefore results in a model that has low bias but high variance. We select the tuning parameter C, the budget for tolerance of misclassification, by using cross-validation method.

Support vector classifier has the following advantages. Firstly, it is more robust method that only observations that either lie on the margin or that violate the margin would affect the classification boundary. For the rest of the observations that are quite far, the decision hyperplane is not susceptible to the changes of them. Secondly, it has a better classification of most of the training observations and the choices of degree of violation are very flexible and at our discretionary that yields a reasonable classification results in real application.

Furthermore, we still attempt the polynomial kernel with a degree of 3 to incorporate the nonlinear relationship boundary with enlarged feature sets. This leads to a more flexible decision boundary and is equivalent to constructing a support vector classifier in a high-dimensional space with the degree of polynomials at 3.

Because the large number of predictors may lead to the overfitting issue, we also apply regularisation to this model. In this dataset, there is a high proportion nonresponse (0) among the response, we balance the weight of the two classes of our basic SVM and regularised SVM, then we use cross validation to select a best weight of two classes and apply this weight into the basic SVM model.

4.3 Gaussian discriminant analysis

We also apply Gaussian discriminant analysis including LDA and QDA to our dataset. The Gaussian discriminant analysis is to categorize the response into the class with the largest computing discriminant score by assuming the distribution of subpopulation.

The LDA model is relatively similar to logistic regression, because both of them produce a linear decision boundary in Bayes classification rule. The difference between them is the

algorithm in computing the decision boundary. LDA assumes a normal distribution for the class population which is assumed to be identical among classes and then computes the mean and variance of the distribution. Because of the same parameter of distribution, LDA generates a linear combination of predictors in discriminant function. Comparing to LDA, QDA assumes that the distribution for each class is not exactly same, hence QDA computes different mean and variance for different class and generates a quadratic combination as discriminant function. Finally, after plugging in these variables into a discriminant score function, the class could be determined as the one with the larger score. Because we could not understand the data before we construct the data, we perform both QDA and LDA.

In addition, we incorporate regularization to QDA because it may suffer high variance compared to LDA.

Part 5: Model evaluation

5.1 Baseline model for model evaluation

Decision threshold calculation:

$$\tau^* L_{TP} + (1-\tau^*) L_{FP} = \tau^* L_{FN} + (1-\tau^*) L_{TN}$$

The optimal threshold is $\tau^* = \frac{L_{FP}-L_{TN}}{L_{FP}+L_{FN}-L_{TP}-L_{TN}} = \frac{2-0}{2+21.33-(-19.33)-0} = 0.047$

If the probability of being 1 conditional on all predictors is greater than threshold 0.047, then it shall be classified as positive response. On the other hand, it shall be classified as negative response if the probability of being 1 conditional on all predictors is less than 0.047.

In this section, the benchmark model is constructed based on the following calculation:

We need to consider a single prediction rule: predicting that all customers will respond, or predicting all will not respond. In the strategy one, our decision is to send the customer brochures to all customers. This is associated with an expected profit of $1.77 per customer. However, if we do not send to the customers any brochures, this will yield a cost $3.77 per customer.

Therefore, we define ‘sending to everyone’ as our baseline model, since this is the best expected result that does not rely on model prediction.

5.2 Evaluation based on standard classification metrics

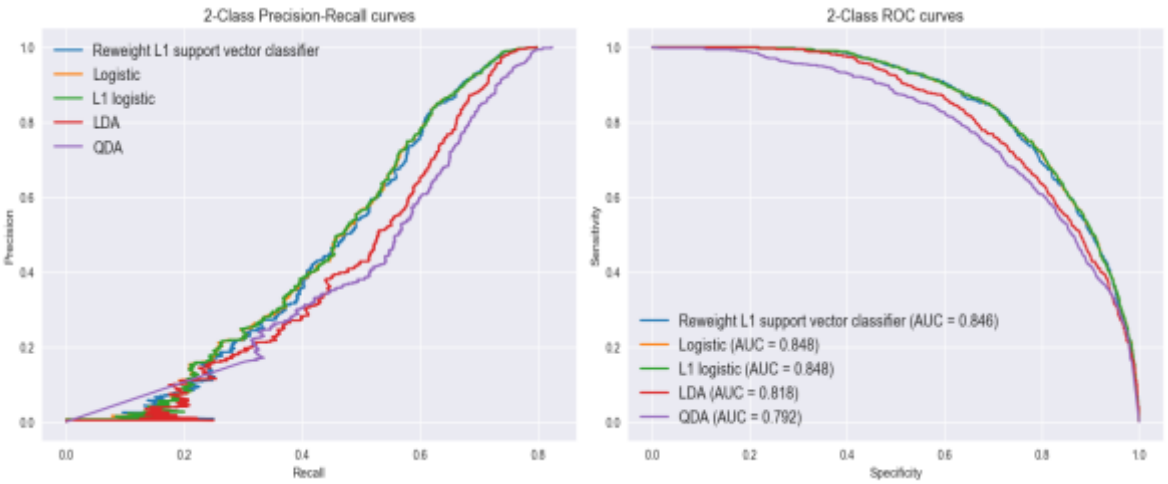
The table below shows the conventional measures for quantifying the classification accuracy, including sensitivity, specificity, AUC, precision, F score and average precision. Since in this dataset, the training data may have an imbalance pattern with a majority of

customers responding to the marketing. Therefore, we may have higher reliance on measures such as AUC and average precision, since they are more useful and relevant classification accuracy measures in the context of imbalanced data.

Table 3 below gives information related to the statistics for model evaluation.

(Table 3 Model evaluation)

Model Evaluation Table						
Model	**Sensitivity**	**Specificity**	**AUC**	**Precision**	**F score**	**average precision**
L1 SVC	0.981	0.447	0.859	0.276	0.430	0.630
L1 SVC CV (0:15%, 1:85%)	0.981	0.448	0.859	0.276	0.431	0.630
L2 SVC	0.981	0.447	0.859	0.276	0.431	0.630
SVC	0.979	0.445	0.859	0.275	0.429	0.629
Logistic	0.977	0.450	0.859	0.276	0.431	0.629
L1 logistic	0.986	0.417	0.859	0.266	0.420	0.627
L2 logistic	0.981	0.420	0.858	0.266	0.419	0.625
Forward logistic	0.977	0.445	0.860	0.275	0.429	0.628
LDA	0.973	0.448	0.851	0.275	0.428	0.626
QDA	0.671	0.760	0.801	0.375	0.481	0.552
Regularised QDA	0.719	0.734	0.795	0.367	0.486	0.568



(Figure 12 Graphs of precision and sensitivity of models)

We can see that using the support vector classifier with incorporating L1 regularization can have the best performance in terms of average precision. Average precision is an area under the precision recall curve, which is a useful measure of success of prediction when the data between each class are not very balance. In this case, using support vector classifier

also gives us the best trade-off between precision and recall. Besides of this measure, it also has the second-best AUC, which represents a promising trade-off between sensitivity and specificity. Another model that is based on a rebalanced data with 15% of 0 and 85% of 1 also indicates promising results, demonstrating that the support vector classifier with L1 norm regularization is relative stable and not sensitive to the changes in proportion of training data.

Besides of this, logistic regression also performs reasonably when we incorporate L1, L2 norms regularization, or using forward selection. However, the performances based on the AUC and precision, on average, do not change too much between the logistic regression.

Finally, discriminant analysis yields the worst performance based on our evaluation results. The figure for AUC and Average precision is significantly lower than the logistic model and support vector classifier. This may be attribute to the following reasons: firstly, the categorical variables cannot be incorporated to the LDA and EDA model, therefore decreases the overall accuracy for classification. Secondly, the distribution for the errors is not normal, therefore the assumption for discriminant analysis is incorrect.

5.3 Evaluation based on business goal and objective

Our business problem here requires us to obtain a precise and accurate classification results for customers to minimize the economic cost per customer (equivalent maximizing profit). With this purpose on mind, our model evaluation should be primarily focus on the model that aims to achieve profit maximization objective.

It is essential to compute the expected economic cost per customer based on each model.

The formula is: $Expected\ economic\ cost: TP \times (-19.33) + TN \times 0 + FN \times 21.33 + FP \times 2$.

The expected economic cost per persons for each model is shown in table 4.

(Table 4 Expected economic cost for each model)

Expected Economic cost	
Model	**Average cost ($)**
L1 SVC	-2.368
L1 SVC CV(0:15%, 1:85%)	-2.371
L2 SVC	-2.357
SVC	-2.349
PCR Logistic	-2.346
Logistic	-2.346
L1 logistic	-2.357
L2 logistic	-2.324

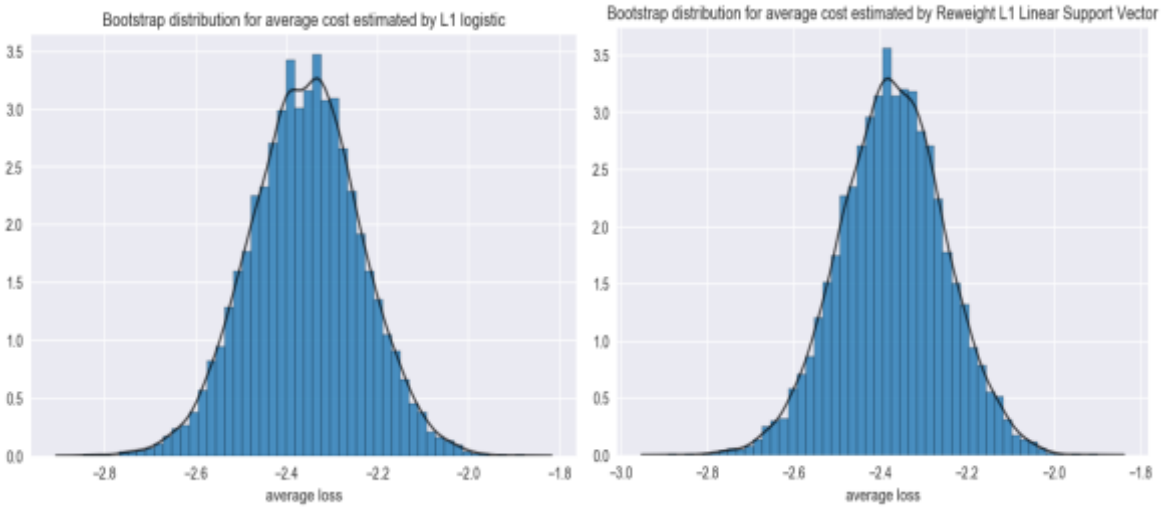
Forward logistic	-2.338
LDA	-2.314
QDA	-0.658
Regularised QDA	-0.961

The table 4 shows that SVC model with L1 norm regularization yields the minimized average costs associated with each customer at $ -2.371. Another way to interpret this figure is that based on the classification accuracy of using support vector classifier to predict whether a customer will respond to the marketing initiatives, we will expect to achieve a $2.371 profit per customer on average, which considers four different possibility such as true positive, true negative, false positive and false negative.

On the other hand, QDA gives us the poorest results for achieving the profit maximization goal, since it only gives less than $ 1 economic profit per person, much less than the benchmark result. Finally, the rest of the models yield similar performance, which are all above $ 2.3. The logistic regression with regularization gives better results, especially when the L1 norm regularization are incorporated.

5.4 Statistical Variability of the model evaluation result

We acknowledge that fitting the model based on the dataset is associated with the statistical variation for the estimated parameters and therefore the following results such as the cost per person in our case. Therefore, we quantify the statistical variation by using the bootstrap method. Through resampling the test dataset, we want to construct the interval regarding the cost per person to measure the uncertainty with our model evaluation results.



(Figure 13 Bootstrap Distribution for Average Cost Estimated by L1 Logistic & L1 Linear Support Vector)

The bootstrap distribution shown in figure 13 & 14 for the support vector classifier and logistic regression with L1 norm are shown below and the rest of the methods are displayed in appendix.

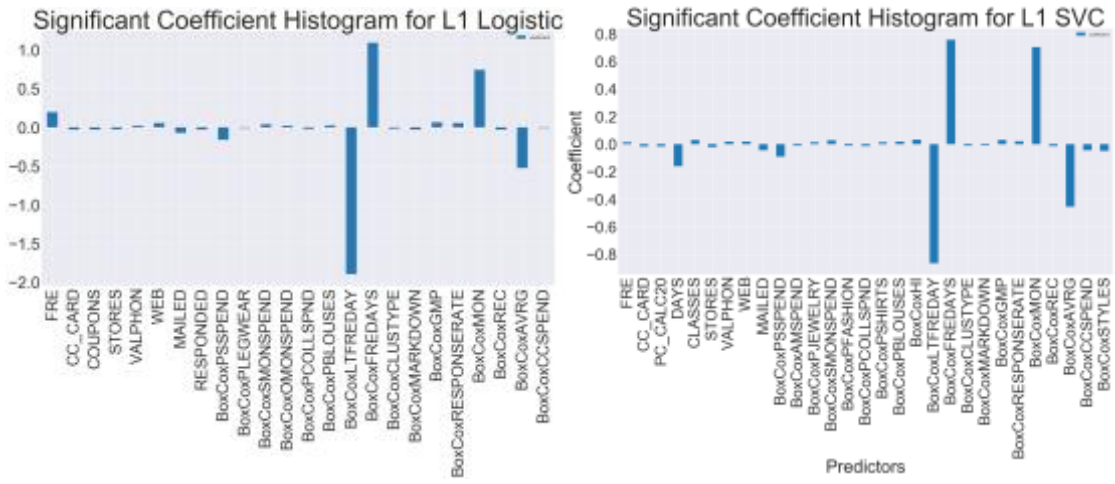
The table below gives information regarding to the confidence interval for the expected cost per unit using bootstrap samples for 10000 iterations. The results show that SVC has the best average performance that its upper limit and lower limit are better than the logistic and LDA. For the logistic regression, the confidence interval ranges from -2.67 to -2.04 with a standard error around 0.3. The result is relatively reasonable and it is not much worse than the support vector classifier method. In the end, LDA has the overall lowest performance.
In addition, using the L1 regularization for the logistic and support vector classifiers improve the estimation since the standard error is smaller and confidence level is relatively narrower. One explanation towards this is that L1 norm regularization performs variable selection, which is associated with a lower variance in coefficient estimation.

(Table 5 Confidence interval for expected costs per person)

Confidence Interval for Expected cost per person (95% CI)		
Model	**$ lower**	**$ upper**
Logistic	-2.669	-2.043
Logistic L1	-2.675	-2.044
SVC L1 (0: 15%, 1: 85%)	-2.691	-2.065
SVC L1 (balanced weight)	-2.688	-2.062
LDA	-2.637	-2.008

Therefore, in conjunction with previous traditional measure of model accuracy and our results for expected cost per person, we stipulate that Support vector classifier with L1 norm regularization and logistic regression with L2 regularization are the candidate model for our deployment for business application.

5.5 Model interpretation and business application



(Figure 14 Significant coefficient of L1 Logistic and L1 SVC)

Figure 14 above illustrates the coefficients of lasso logistic and lasso SVC whose absolute value exceeds 0.01. In our case, we found there are coefficients that may have large impact on classifying the response.

- Lifetime average time between visits (*LTFREDAY*)

This predictor has the largest negative coefficient in the model. In real life, the larger the numbers of average day between purchases in the lifetime, the less frequent the customer would make a purchase and therefore the less likely to respond to the marketing mails. Therefore, the customer is more likely to be classified as non-response when they have a larger number of day between purchase because they are less enthusiastic to purchase the item or response to the company.

- Number of days between purchases (*FREDAYS*)

This predictor has the largest positive coefficient in the model. In real life, people are less likely to purchase when they just shop recently. The larger the day between purchase, the more need they have to meet. Therefore, customers are more likely to be classified as response when they haven't purchase items for a longer period.

- Total net sales (*MON*)

This predictor also contributes significantly to both lasso logistic and lasso SVC. If a customer yields a high total gross profit to the company, this particular customer may have meaningful interest in the products of our clothing store. Therefore, when customers have a higher total net sales, they are more likely to be classified as response.

- Average amount per visit (*AVRG*)

In both L1 logistic and L1 SVC, average amount per visit is the second largest negative coefficient among all predictors. Although it may be considered that if customer has larger

average amount per visit, they may be more likely interest in products of clothing store. However, if customer purchases more items on average, they may be less frequently to make a purchase. Therefore, customers are less likely to respond to the direct mail if they have higher average amount per visit.

- Number of days customer has been on file (*DAYS*)

The fifth most significant predictors in L1 SVC are the number of days customers that has been on file. It can be seen that if the number of the days the customers has been on file is large, the customers will receive the marketing brochures more frequently than those who have less days on file. This will lead to negative attitudes towards marketing, therefore more incentive to be classified as non-response.

5.6 The effect of using model on actual accounting profit

One of the key problem for the business is to concern about the actual profit per customer that could be generated and reflected in the financial statement. Therefore, in this case we consider the actual profit or loss that arise during business. In the actual scenarios, as long as we send the brochure, the customers would then be determined as responding or not responding. Therefore, two cases, true positive and false positive should be considered in our calculation for actual profit.

The confidence interval calculation is shown in the following: for those who are classified as response, the probability of truly response is P and the probability of negatively response is 1-P. Therefore, the event follows Bernoulli distribution. The confidence interval for this can be derived from $\hat{P} \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p}\times(1-\hat{p})}{n}}$ where $\hat{P}$ is the estimated probability of positive response given positive response classification(precision) and α is the significance level.

$Expected\ profit\ per\ person$:

$$\$\ (True\ Positive\ \times 19.33 - False\ Positive \times 2) \div Total\ positive\ classification$$

$$= (\hat{P}\ ) \times\ 19.33 - (1 - \hat{P}\ )\ \times\ 2$$

Table 6 shows the actual profit or loss given that the predicted classification is positively response.

(Table 6 Actual profit)

Outcome	Classification	Actual response	cost $
True positive	Response	Response	-19.33
False positive	Response	Nonresponse	2

The figures in this table is quite different from our profit-loss matrix defined in the previous section. The table here represents the actual profit that can generate per person. When the result is true positive, the profit generated is $19.33 per person since the profit margin per person is $21.33 and the $2 deduction is associated with cost of marketing. On the other hand, if the marketing brochures are mailed to the customers but there is no response regarding the mailing, this will incur an actual loss of $2.

The difference here is that in making economics decisions, we need quantify the opportunity cost, which is the benefits that we foregone when considering the second-best alternative. The profit and loss matrix in the section 1 should be used for us to make the model selection decisions.

Table 7 below demonstrates the effect of profit per customer on our financial statements.

(Table 7 Estimated profit for each model)

Actual Expected Loss Table			
Model	$ Actual expected loss	$ lower limit	$ upper limit
L1 SVC	-3.881	-4.164	-3.597
L1 SVC CV (0:15%, 1:85%)	-3.896	-4.179	-3.612
SVC	-3.868	-4.151	-3.585
Logistic	-3.891	-4.175	-3.608
L1 logistic	-3.682	-3.963	-3.402
LDA	-3.857	-4.140	-3.574

Here, we want to interpret the model that Support vector classifier still yields the highest profit per person at $ 3.90 per person, and the accuracy classification results would give a better profitability performance on the financial statements. The 95% confidence interval for the profit per person in support vector classifier is from $3.61 to $4.18. The statistical variation for the expected actual profit is not very large since the difference between middle point and upper or lower value is only $0.18. Also, Logistic regression with L1 regularization has an expected actual profit $3.68 per person, with the lower bound ($3.40) and upper bound ($3.96). The statistical variation is relatively larger with a 0.28 difference between the middle point and boundary limit. In terms of actual profit per person, LDA outperforms the logistic regression model with L1 regularization due to its high precision in positive classification. However, the profit improvement is not very significant between these two methods.

For example, if marketing department sends direct mail to 100,000 customers, then L1 SVC (reweighted) could increase actual profit by $(\$3.896 - \$1.77) * 100{,}000 = \$212{,}600$.

Summary and Conclusion

In summary, our optimal model is support vector classifier with L1 regularization. However, other method such as logistic regression with L1 regularization also has better performance in making decision related to direct marketing. These models can be deployed to the clothing stores to make the business decisions towards targeting potential customers that are with higher net sales and profit margin, lower average time between each purchase, larger number of days between each purchase and other customer characteristics. By pointing towards these potential characteristics, the customers are more likely to give a direct response to our marketing mail, and therefore the overall long-term profits can be maximised and good customer relationship and profitability management can be achieved to sustain the value creation within the clothing store business.

Reference

James, G., Witten, D., Hastie, T., Tibshirani, R., & SpringerLink (2013). *An introduction to statistical learning: With applications in R*. New York, NY: Springer New York.

Larose, D. T. (2005). Case Study: Modeling Response to Direct Mail Marketing Data Mining Methods and Models (pp. 265-316): John Wiley & Sons, Inc.

Appendix

Table 1 Predictors descriptions

Predictors	Description	Predictors	Description	Predictors	Description
HHKEY	customer ID	PSUITS	percentage spent by the customer on suits	FREDAYS	number of days between purchases
ZIP_CODE	zip code	POUTERWEAR	percentage spent by the customer on outer wear	MARKDOWN	markdown percentage on customer purchase
REC	number of days since last purchase	PJEWELRY	percentage spent by the customer on jewellery	CLASSES	number of different product classes purchased by the customer
FRE	number of purchase visits	PFASHION	percentage spent by the customer on fashion	COUPONS	number of coupons used by the customer
MON	Total net sales	PLEGWEAR	percentage spent by the customer on leg wear	STYLES	
CC_CARD	flag—credit card user	PCOLLSPND	percentage spent by the customer on collection	STORES	number of stores customer purchased at
AVRG	Average amount per visit	AMSPEND	amount spent for each of four different franchisers 1	STORELOY	number of stores customer shopped at
PC_CALC20		PSSPEND	amount spent for each of four different franchisers 2	VALPHON	valid phone number on file
PSWEATERS	percentage spent by the customer on sweater	CCSPEND	amount spent for each of four different franchisers 3	WEB	web shopper
PKNIT_TOPS	percentage spent by the customer on knit tops	AXSPEND	amount spent for each of four different franchisers 4	MAILED	number of promotion mailed in the last year
PKNIT_DRES	percentage spent by the customer on knit dress	SMONSPEND	amount spent in the past six months	RESPONDED	number of promotion responded to in the last year
PBLOUSES	percentage spent by the customer on blouses	TMONSPEND	amount spent in past three months	RESPONSERATE	promotion response rate for the past year
PJACKETS	percentage spent by the customer on jackets	OMONSPEND	amount spent in the past month	HI	product uniformity (low score = diverse spending patterns)
PCAR_PNTS	percentage spent by the customer on pants	PREVPD	amount spent the same period last year	LTFREDAY	lifetime average time between visits
PCAS	percentage spent by	GMP	gross margin	CLUST	Microvision lifestyle cluster

_PNTS	the customer on tops		percentage [(sales-COGS)/sales]	YPE	type
PSHIRTS	percentage spent by the customer on shirts	PROMOS	number of marketing promotions on file	PERCRET	percent of return
PDRESSES	percentage spent by the customer on dresses	DAYS	number of days customer has been on file	RESP	response

Figure A1 Correlation matrix containing all predictors