Classification models to optimize profitability in a direct mail marketing campaign

Business Understanding

Business and Data Mining Objectives

This study explores methods to increase the profitability of a clothing store chain through a direct mail marketing promotion. The business would like to know which customers are more likely to respond to direct mail marketing so they can decide which customers to mail promotions to and reduce unnecessary expenses. Hence, this is a classification problem. This study will use diverse classifiers to address the business problem of profitability. Given historical information about customer behaviors and store events, the models will classify which customers are likely to respond to the direct mail marketing promotion.

The model selection and evaluation criteria is aligned with a measure of business success, specifically the overall profit (or negative cost). Various models will be applied to training data, and well-performing models will be selected and evaluated on a validation and test data set. After applying the most effective classifier, the business can expect a higher promotion response rate and increased profit generated from direct mail marketing promotions.

This study will follow the CRISP-DM data mining methodology, which consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Sumathi, Sivanandam, & SpringerLink, 2006).

The Cost-Benefit Table

In classification, there are four decision outcomes: true/false positive and true/false negative. In order to select and evaluate classification models, we use measures such as error rates, sensitivity (true positive rates), specificity (true negative rates), and precision (one minus false discovery rate). In the direct mail marketing response problem, the cost of each decision is different. Specifically, the cost of a false negative is much higher than the cost of false positive, so it is not reasonable to simply use measures that treat the cost of all outcomes as equal in magnitude to evaluate models. Therefore, it is necessary to build a cost-benefits table to compare models. Below we will discuss the loss matrix (**Table1**) in detail and analyze the cost of the four possible outcomes.

		Classification	
		$\hat{Y}$=0 (Non-response)	$\hat{Y}$ =1 (Response)
Actual	Y=0 (Nonresponse)	True Negative (Loss Avoided) $0	False Positive (Mailing Cost) $1
	Y=1 (Response)	False Negative (Opportunity Cost) $14.8	True Positive (Expected gross profit minus mailing cost) -$13.8

Table1.The Cost-Benefit Table (loss matrix)

A true negative occurs when the model makes a correct prediction that the customer will not respond to direct mail marketing, so the company will not mail them. The loss is avoided and the cost is zero as there is no mailing or production costs. There is no potential profit loss, since those customers will not respond to the mail promotion even if they received it.

A false positive occurs when the model expects the customer to respond to the mail marketing promotion, but the customer does not. Hence, the business would incur a direct mailing cost without making any profit. In this analysis, we assume the cost per mailing is $1, which includes postage fee, printing cost, salary cost, and all other associated costs.

A true positive is when the model correctly predicts that a customer will respond to the mail promotion. In this case, the company will mail to a customer and the customer will make a purchase during the promotion period, generating profit. According to historical data, the average amount spent per visit is $113.90 (*Appendix Table1*). Given that the clothing store industry profit margins are generally within a range of 4 to 13 percent (Crane, 2007), the expected profit per visit is $14.80 ($113.90 × 13%). Assuming that the cost per mailing is $1, the cost of a true positive outcome will be -$13.8 ($1 - $14.8). The negative cost here means the business is generating profit.

A false negative is when the model predicts the customer as nonresponsive to the promotion, but the customer is in fact responsive. The company will not mail these customers who would actually respond to the mailing promotion. As a result, there will be a massive opportunity cost. Using the estimates above, the potential profit loss is $14.80 per customer.

Using the confusion matrix and cost-benefit table to calculate overall cost, we have a specific criterion for model selection and evaluation. As noted above, the cost of a false negative is much higher than a false positive. We are therefore interested in minimizing false negatives, while maximizing profit. A decision threshold parameter is calculated based on the cost-benefit table, using the following formula (Scharth, 2017): $r^* = (\mathrm{C_{FP}} - \mathrm{C_{TN}})/(\mathrm{C_{FP}} + \mathrm{C_{FN}} - \mathrm{C_{TP}} - \mathrm{C_{TN}}) = (1 - 0)/(1 + 14.8 - (-13.8) - 0) = 0.034$. This threshold will be the decision rule for classifying a customer as positive or negative. If the estimated probability of a positive outcome (the customer will respond) is higher than 0.034, the customer will be classified as positive. Note this threshold is very low due to the large cost of false negatives.

Data Understanding

Data Summary

The data is collected from a clothing store chain with 21,740 customers and 51 variables such as number of purchase visits, total net sales and average amount spent per visit (*Appendix Table 2*). The response variable is a binary variable named 'RESP', indicating whether or not the customer responded to previous direct mailing market promotions. There are no missing value in this dataset.

The meanings of some variables are unknown, such as 'PC_CAL20', 'STYLES', and 'STORELOY'. The variable 'HHKEY' is the customer ID, which is unique for every customer and does not contain any helpful information for our analysis. The variable 'ZIP_CODE' represents the geographic location of the customer and could be useful, however we do not consider it as it is too difficult to implement for this study. These variables are therefore dropped, leaving 46 variables, including 4 categorical variables, 9 discrete variables, 32 continuous variables and the target variable. *Appendix Table 3* shows the 25 variables whose correlation coefficients with the response is greater than 0.1. The following analysis will focus on these variables.

Imbalanced Data Problem

The data here is highly imbalanced, as there is an overwhelming proportion of negatives (nonresponses) in the data, shown in **Figure 1**. Not addressing this could results in misleading conclusions. Only 16.5% of all customers responded to past mailing promotions. If the model is trained on data with 83% nonresponsive customers and simply predicts all customers will respond to the mailing promotion, the error rate will be only 17%. One may conclude this model is reasonably good as the error rate is relatively low, but it is in reality meaningless and not increasing profit effectively. We will attempt to address this by oversampling the data to balance the classes.

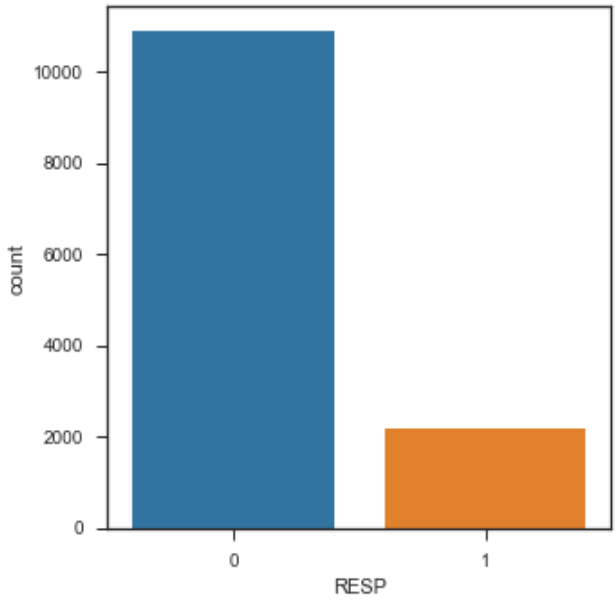

Figure 1. Count plot of response and nonresponse

Skewness in Numerical Variables

Descriptive statistics are used to analyze each variable (*Appendix Table 1*), which show that the majority of variables have high skewness and kurtosis. We visualize the distribution of these numerical variables with histograms and boxplots (*Appendix Figures 4* and *5*), with some indicating the presence of outliers. For example, **Figure 2** shows 'LTFREDAY' to be right-skewed with several outliers. Descriptive tables also show that some of the numerical variables have a minimum value of zero, with many entries equal zero. When doing data processing, an adequate transformation method is needed for these variables.

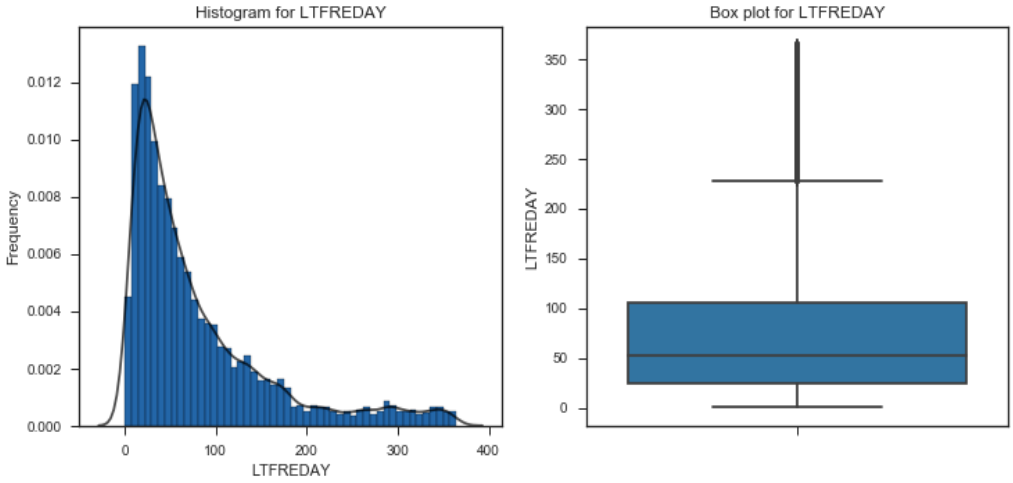

Figure 2. Histogram and Boxplot of 'LTFREDAY'

Univariate Relationship Between Predictors and Response

Univariate regression plots are drawn to explore the relationship between the numerical variables and the response. Plots with more significant relationships are shown in *Appendix Figure 1.1-1.13* and *2.1-2.9*. Nearly all numerical variables show positive relationships with the response. For instance, the regression plot of 'MON' (total net sales) in **Figure 3** shows that as 'MON' increases, the probability of that customer responding also increases. 'LTFREDAY', 'FREDAYS', 'HI', and 'REC' all have negative relationships with the response variable. These variables are concerned with the time between each store visit or purchase, and diversity among purchases. Customers with longer time periods between each purchase or visit and with less diversified spending patterns are less likely to respond to the promotion.

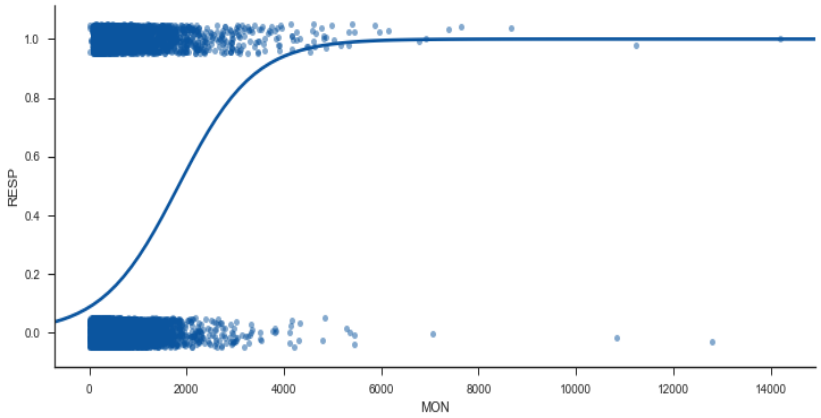

Figure 3. Regression plot of 'MON'

To analyze the relationship between binary variables and the response, we look at cross-tabulations (*Appendix Table 4*). These tables show that 27.4% of credit card users respond to mail, while only 9.6% non-credit card users respond to mail. The joint probabilities for customer response for each binary outcome are all different. We consider all binary variables in model building.

Analysis of Lifestyle Cluster Type Variable

The Microvision lifestyle cluster type variable is a categorical variable with 50 categories. Clustering techniques are considered for this variable to reduce the number of classes. The proportions of response and nonresponse for each cluster type are shown in *Appendix Figures 3.1* and *3.2*. However, the bar plots show insignificant differences between types and the actual meaning of each class is unknown, two issues which make clustering not feasible. As a result, we do not consider clustering for this lifestyle variable.

Correlation Between Predictors

Figure 4 is the correlation heat map, illustrating the correlations between the top 25 variables with the highest response variable correlation. Dark colors indicate a stronger correlation, either negative or positive, between two variables. For instance, the correlation between 'FREDAYS' and 'LTFREDAY', and the correlation between 'OMONSPEND' and 'TMONSPEND' are both strong positive correlations. Many of those correlated variables have a similar meaning. Hence, we will consider combining some variables as well as principal component analysis (PCA).

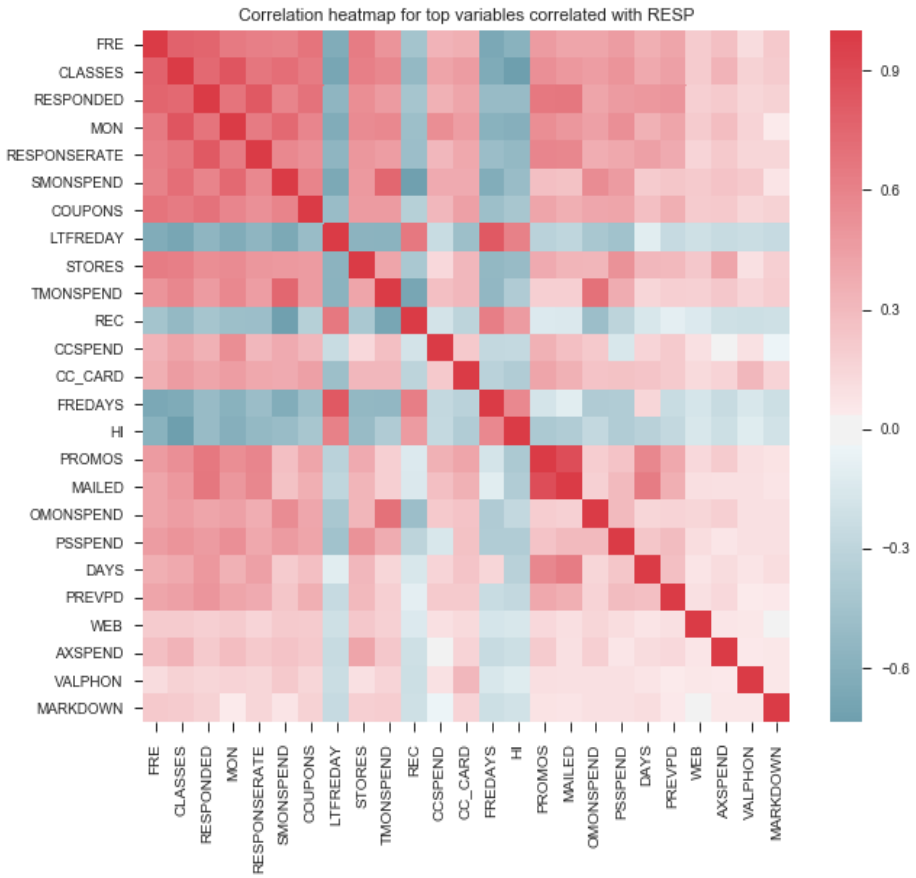

Figure 4. Correlation Heat map for top variables related to 'RESP'

Data preparation and Feature Engineering

Data Cleanse

The first step is to assess whether the data are complete, as we want to address variables with a large number of missing values. Missing values and other issues are analyzed in exploratory data analysis. It shows that this dataset does not have any missing values, nor a large number of nonsensical entries. There are two further minor issues to be addressed. 'VALPHON' is a binary variable, and the two classes for it are 'N' and 'Y'. In order to build models using 'VALPHON', 'N' and 'Y' are replaced by integers 0 and 1. Secondly, 'CLUSTYPE' (Microvision lifestyle cluster) is a discrete variable that has 51 unique values, considered a numerical variable when it is in fact categorical. 'CLUSTYPE' is converted into categorical variable, though we ultimately decide to not use it.

Data Transformation

When applying models such as Logistic Regression and Gaussian Discriminant Analysis, an assumption is made of normally distributed data, which in practice usually yield better results. Exploratory data analysis illustrates that the distributions of many numerical variables are highly skewed. However, some of those variables can have a minimum value at 0, which means that a log transformation for those variables is not valid.

Two types of data transformation are applied in this case. A log transformation is applied for variables that do not have any zeros, such as 'RESPONSERATE', 'SMONSPEND', 'TMONSPEND', and 'OMONSPEND'. For several variables with zero values, such as 'LTFREDAY', 'HI', 'FREDAYS', 'MON', 'CCSPEND', a square root transformation is applied. Data transformation can normalize variables' distributions. **Figure 5**, for example, shows the distribution for 'LTFREDAY' is closer to normal after a log transformation.

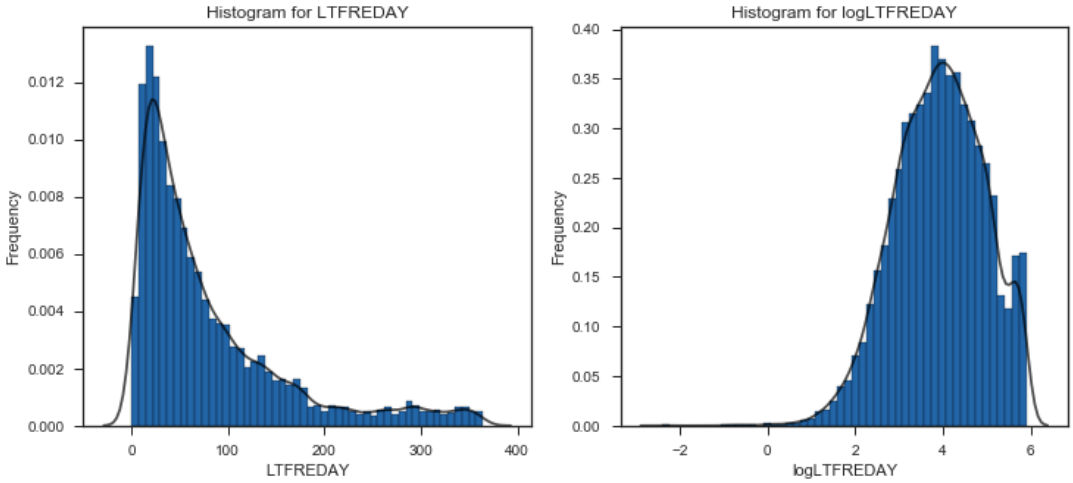

Figure 5. 'LTFREDAY' before and after a square root transformation

Variable Creation

Analysis of the correlation between response and predictors show that, 'OMONSPEND', 'TMONSPEND', and 'SMONSPEND' can be three useful predictors in model building. However, after checking the documentation of these variables, it is apparent the information they contain overlap. For example, amount spent in the last month ('OMONSPEND') is obviously included in both amount spent in the last three months ('TMONSPEND') and in the last six months ('SMONSPEND'). Thus, new variables are created to capture each of these separately.

The three variables 'OMONSPEND', 'TOMONSPEND' and 'STMONSPEND' are kept, and two new variables are derived from them. Amount spent in previous months 2 and 3 ('TOMONSPEND') is calculated by subtracting 'OMONSPEND' from 'TMONSPEND'. Amount spent in previous months 4, 5 and 6 ('STMONSPEND') is calculated from 'SMONSPEND' - 'TMONSPEND'.

Standardization

Data standardization is needed for this analysis. Although applying data transformations have already reduced the variability, there are still substantial differences in variability between numerical variables. For example, the standard deviation for 'REC' is 104.56 while for 'STORES' it is 1.62. Clearly, there is a large variability between these two variables and their fields should be standardized.

The data is standardized by subtracting the mean of each variables from each observation, and then dividing by the standard deviation. After standardization, all the numerical variables have a standard deviation of 1 and a mean of zero.

It is necessary to standardize the data since many classification methods rely on the assumption of a Gaussian distribution. By satisfying this assumptions, we expect the models to perform better. Additionally, in order to conduct principal component analysis, data needs to be standardized. If the variables have different variabilities, PCA will disproportionately favor variables with larger magnitude and therefore larger absolute variances, as it is a variance maximizing method.

Principle Component Analysis

PCA is a procedure used to transform the correlated variables into several uncorrelated components. Analysis of the correlation heat map in **Figure 4** suggests that a collinearity problem exists between some predictors. For example, the correlation between 'FREDAYS' and 'LTFREDAY' is very high at 0.862. The existence of collinearity can influence the performance of some classifiers that assume non-collinearity, like logistic regression. Principal components are created, and both PCA models and non-PCA models will be explored in modeling.

Table 2 shows the four principal components created for variables that contains similar information. More specifically, the combined predictors have high correlations with each other and have similar meanings. 'PCA_1' is concerned with the customer purchasing pattern, 'PCA_2' incorporates lifetime average time between visits and the number of days between purchases and 'PCA_3' and 'PCA_4' represent the likelihood of response to past promotions and the number of previous promotions on file, respectively.

Principal Components	Variables Used
PCA_1	'CLASSES', 'HI'
PCA_2	'FREDAYS','LTFREDAY'
PCA_3	'RESPONDED','RESPONSERATE’
PCA_4	'MAILED','PROMOS'

Table 2. Principal Components Created

Data Resampling

Exploratory data analysis shows that this dataset is highly imbalanced, as only 16.5% of customers responded to previous promotions. Most classification methods perform better when the number of observations in each response class are roughly equal. This study used a resampling technique, specifically the Synthetic Minority Over-sampling technique (SMOTE) and built two extra training datasets, which have different proportions of responses. One of the training sets is balanced with a 50/50 ratio, while the other set is oversampled so that 70% respond to the promotion and 30% do not.

Final Datasets

After the data preparation and feature engineering, the final data sets are the original dataset, the original dataset with principal components, 50/50 balanced data and 70/30 oversampled data, both without principal components. All the variables included in these datasets are cleaned and processed using data transformation and standardization as described above. New variables created from the time between visits or purchases are also included.

Modeling

The models here will be evaluated with respect to the decision threshold calculated from the cost-benefit analysis. The data used are the training data after data transformation, creation of new variables, and standardization. All the models will be applied under the four scenarios mentioned above. The performance of models are then compared before and after applying SMOTE and PCA techniques on a validation set, which is an unseen subset of the training data for model selection.

Nine different modeling methods are applied: naive Bayes, logistic regression, regularized logistic regression (ℓ_1 regularized and ℓ_2 regularized), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), regularized QDA, random forest, and neural networks. Model validation suggests that logistic regression and regularized logistic regressions, LDA, random forest, and neural networks, all perform better than other methods. The complete validation scores are shown in *Appendix Table 5*. This study uses profit per customer as the criterion for model validation, shown in **Table 3**. Note this is expressed as negative cost.

	Cost	Lower Bound 95% CI	Upper Bound 95% CI
Random Forest	-1.7	-1.688	-1.712
Naïve Bayes	-0.88	-0.869	-0.891
Logistic	-1.71	-1.698	-1.722
ℓ_1 regularized	-1.71	-1.698	-1.722
ℓ_2 regularized	-1.71	-1.698	-1.722
LDA	-1.71	-1.698	-1.722
QDA	-0.52	-0.509	-0.531
Regularized QDA	-0.62	-0.609	-0.631
Neural Network	-1.72	-1.708	-1.732

Table 3. The Validation Score of All Models

Analysis of Key Models

Baseline Model

A baseline model is established as a benchmark to see if the models are effectively increasing the overall profit. Since the cost of false negatives is much higher than the cost of false positive, we could simply mail to all customers to result in no false negatives. However, this may not be cost effective. The baseline model is therefore this scenario where all customers are mailed. The cost per customer is shown in **Table 4**. Although the error rate is very high due to all the false positives, this strategy still amounts to a profit of $1.49 per customer. The models we develop are compared to this baseline profit.

	TN (0)	TP (-13.8)	FN (14.8)	FP (1)	Error rate	Total cost per customer
Mail all customers	0	731	0	3617	0.832	-1.49

Table 4. The Validation Score of The Baseline Model (without oversampling and PCA)

Logistic Regression and Regularized Logistic Regressions

The logistic regression model is a discriminative algorithm that classifies each observation based on the conditional probability of an observation belonging to a certain class given the input data, and a decision threshold. Exploratory data analysis has identified that many variables have a clear relationship with the response. This implies the change in the probability for a customer to respond is associated with the increase in the value of numeric variables or the change in categorical variables. Given the observed input data, logistic regression will estimate the probability of a customer

responding by logistic function, given as $p(x) = exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)/(1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_j))$. The logistic regression coefficients will be estimated by using the maximum likelihood method, meaning the coefficients will be chosen such that they maximize the probability of obtaining the observed data.

Forward selection is used to find the best subset of predictors. This technique will compare validation scores of models using increasingly larger possible subsets of predictors until additional predictors worsen model performance. This finds the smallest subset with the lowest prediction error. The output from forward selection suggests that all the existing predictors should be kept.

Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient
Constant	-2.38	VALPHON	0.07	LTFREDAY	-1.95
MAILED	-0.19	PSSPEND	-0.17	STMONSPEND	-0.003
FRE	0.45	CC_CARD	-0.11	MARKDOWN	-0.01
STORES	-0.08	CCSPEND	-0.06	OMONSPEND	-0.01
DAYS	0.32	PREVPD	0.02	RESPONDED	-0.13
REC	-0.07	CLASSES	-0.06	TOMONSPEND	-0.02
MON	0.24	FREDAYS	0.76	RESPONSERATE	0.17
HI	-0.03	PROMOS	0.13	COUPONS	-0.08
WEB	0.06	AXSPEND	-0.02		

Table 5. The Coefficients of Predictors in Logistic Regression

The coefficients of logistic regression, shown in **Table 5**, tell the exact relationship between the response and the predictors. However, some relationships could be misleading due to confounding. For example, 'FREDAYS' and 'LTFREDAY' are two variables concerned with the frequency between purchases or visits, respectively. Obviously, these two variables should have either both positively or negatively associated with the probability of a customer responding, since they have very similar meanings. However, in the multiple logistic regression model, they have very different coefficient values, 0.76 for 'FREDAYS' and -1.95 for 'LTFREDAY'. This problem is caused by the high correlation between these variable. The association between predictors and the response analyzed by univariate regression plots (*Appendix 1.1-1.13* and *2.1-2.9*), as well as simple logistic regression using one predictor, is given in **Table 6**.

Predictors regarding spending	
MON	Positive
TOMONSPEND	Positive
OMONSPEND	Positive
STMONSPEND	Positive
PREVPD	Positive
Predictors regarding past promotions	
PROMOS	Positive
MAILED	Positive
RESPONDED	Positive
RESPONSERATE	Positive
Predictors regarding product diversity	
CLASSES	Positive
HI (low score means diverse spending patterns)	Negative
Predictors regarding frequency between purchases or visits	
FREDAYS	Negative
LTFREDAY	Negative

Table 6. The general association between response and each predictor given by simple logistic regression

To attempt to remedy this, regularized logistic regressions are also applied to shrink the coefficients and reduce the effect of correlation between predictors. Both ℓ_1 and ℓ_2 regularised logistic regressions are applied. The validation score shows that these regularized models perform equally as well as unregularized logistic regression. The cost is -$1.71 for all three models, meaning a profit of $1.71 per customer. Hence, regularization does not improve the model.

Linear Discriminant Analysis

The linear discriminant analysis classifier (LDA) used a different subset of predictors in order to satisfy the assumption of discriminant analysis. The LDA classifier assumes that the observations within each class are all from a multivariate Gaussian distribution. That is, they come from a normal distribution with common variance and a class-specific mean vector. As a rule, binary variables are excluded from the LDA model.

The validation output also suggests that LDA performs as well as regularized and unregularized logistic regression. The overall cost for LDA is the same as it is for logistic regression at -$1.71. Further evaluation of these models with the same overall cost is presented in a later section when we introduce the test dataset.

Random Forest

The random forest is a supervised classification algorithm which creates a “forest” with a number of decision trees. Each decision tree is trained separately on the training data, and a classification is made by first feeding the data point into each tree and recording each outcome. The classification is determined from a majority vote i.e. the category most classified among all the trees. Using this method is advantageous in that we do not require standardization or variable selection beforehand. With only two hyperparameters involved, namely the number of variables in the random subset and the number of trees in the forest, tuning this algorithm is fairly computation-friendly.

Random forests do not return regression coefficients like logistic regression; instead, it outputs a score showing the importance of each feature in determining the classification (Albon, 2016., "View Feature Importance", Para. 1). The feature importance is an estimate of what fraction of the classification a feature contributes to. The list of the features and their importance scores is provided in **Table 7**.

Predictors Feature Importance			
FRE	0.0936	HI	0.0423
CLASSES	0.0421	PROMOS	0.0240
RESPONDED	0.0136	MAILED	0.0159
MON	0.0485	OMONSPEND	0.0118
RESPONSERATE	0.0283	PSSPEND	0.0270
COUPONS	0.0081	DAYS	0.0595
LTFREDAY	0.3175	PREVPD	0.0119
STORES	0.0129	AXSPEND	0.0065
REC	0.0363	MARKDOWN	0.0315
CCSPEND	0.0369	TOMONSPEND	0.0205
FREDAYS	0.0815	STMONSPEND	0.0298

Table 7. Predictors' feature importance in Random Forest method

Features on top of the tree contribute a larger fraction to the final prediction than features deeper along the tree. The most important feature here is 'LTFREDAY' with an importance score of 0.3175, followed by 'FRE' (0.0936) and 'FREDAYS' (0.0815). This result is similar to what is obtained from logistic regression in that it highlights the same explanatory variables.

Neural Networks

On a high level, neural networks are a method in which the covariates are input into the model in the input layer, then various weighted transformations are applied to the covariates through hidden layers and activation functions. Each of these layers has a certain number of units, which are simply inputs into the layer either externally or from the previous layer. The activation functions are used to define the output of a unit, given its input, mapping the input between 0 and 1. The penultimate output is a set of predictors that are weighted linear combinations of the transformed covariates. A regression is then run on these predictors to produce classifications. The specifications of the neural network model used here are relatively simple. More complex models were tested with more hidden layers and units, but a simpler model ended up performing better. The model used had one input layer with 25 units and one output layer. The activation for the input layer is the rectified function, which takes an input value and outputs the same value if it is positive, and zero otherwise. The output layer uses a sigmoid function which condenses the output to be between 0 and 1 since this is a classification problem.

The output is not 0 or 1 however, it is simply condensed between those two numbers. The decision threshold is then applied to this output to determine the classification. A neural network model can be very useful for predictions since they allow for very complex relationships between the variables and response, but they are very difficult to interpret because they are a black box in the sense it is unknown what exact values and transformations are being used to create the model. Nevertheless, we implement it here for its predictive performance.

Alternative Methods and Other Techniques Included

Alternative models considered are naïve Bayes, quadratic discriminant analysis (QDA), and regularized QDA. The naïve Bayes classifier is a generative model based on the assumption that all predictors are conditionally independent given the class model label. QDA is a discriminant method, with the assumption that observations from each class are all drawn from a Gaussian distribution, but have different covariance matrices. Although the simplicity of the naïve Bayes classifier can potentially help prevent the model from overfitting, it may lead to highly biased probabilities. In this study, the validation output shows that the performance of these models is not exceeding the baseline model, so we do not consider them further.

A principle component analysis (PCA) is also conducted following the method stated under feature engineering. Although PCA can combine correlated predictors while preserving information as well as reduce the number of predictors for simplicity, the non-PCA models perform better than the PCA ones. Comparing the validation scores before and after applying PCA, we find that including principal components does not help increase profit. Hence for simplicity, we choose to proceed with non-PCA models.

The SMOTE resampling method is also applied to balance the data, however, models built on the resampled data are ultimately not selected. This study applied both balancing and oversampling, using the ratios 50/50 and 70/30. Model validation shows that neither method helps increase profit. This is likely because of the way the cost-benefit matrix is defined. In practice, resampling the data when the cost-benefit matrix is very asymmetrical only sometimes leads to better classification results, but in this case, it does not.

Assumptions
Gaussian discriminant analysis methods require that the observations within each class are from a multivariate Gaussian distribution. In addition, QDA classifiers result from assuming that each class has its own covariance matrix, while LDA assumes that observations from each class have a common variance. Histograms are used to check the distribution of the observations in exploratory data analysis. Although some variables have skewed distributions, data transformation have been applied to reduce the skewness. By standardizing the data, the variance of each variable is also regularized.

The performance of logistic regression can be affected if the data contains too many outliers or if high multicollinearity exists. Exploratory data analysis has revealed some unusually large values for some variables, however, it is difficult to tell whether these observations are outliers or simply skewed observations. This analysis assumes that there are few outliers after data transformation and standardization. The multicollinearity problem is ultimately left as is, since it does not appear to affect model performance.

Evaluation

Up to this point, different models have been built and validated on the transformed training dataset. Model selection suggests that there are six approximately equally well-performing models: logistic regression, regularized logistic regression (ℓ_1 and ℓ_2), LDA, random forest, and neural networks. All selected models are non-PCA models build on unbalanced data, since these techniques did not help improve the model performance. The selected models will now be evaluated and discussed in terms of their practicality in being applied to the business problem.

Evaluation of Selected Models

Appendix Table 6 shows the complete model evaluation output generated on the test data. The test scores in terms of the business objectives are shown in **Table 8**. It can be observed that the neural network model performs best, as the cost of this model is the lowest at -$1.74. Applying the neural networks model can increase the overall profit by approximately $1.74 per customer. The 95% confidence interval indicates that the profit per customer when applying neural network is between $1.728 and $1.752. The performances of logistic regression, regularized logistic regression and LDA models are similar. Applying these models can increase the profit per customer by approximately $1.71. For these models, the 95% confidence interval of profit is between $1.698 and $1.722. The random forest in this scenario does not performs as well as the other models, but it can still increase the profit by approximately $1.69., with the 95% confidence interval predicting profit to be between $1.678 and $1.702. These models are separated by a very minor amount on a customer level, those these differences are far more apparent when multiplying out by the number of customers. Profits then differ by thousands of dollars.

	Cost	Lower Bound 95% CI	Upper Bound 95% CI
Random Forest	-1.69	-1.678	-1.702
Logistic	-1.71	-1.698	-1.722
L1 regularized	-1.71	-1.698	-1.722
L2 regularized	-1.71	-1.698	-1.722
LDA	-1.71	-1.698	-1.722
Neural Network	-1.74	-1.728	-1.752

Table 8. Evaluation score for selected models

Compared to the baseline model, all the selected models perform better. The baseline model is based on the strategy of sending the promotion to all customers, resulting in an overall profit of $1.49 per customer. The incremental profit per customer generated by using logistic regression (or regularized logistic regression or LDA), neural networks, and random forest, are $0.22, $0.25, and $0.20, respectively.

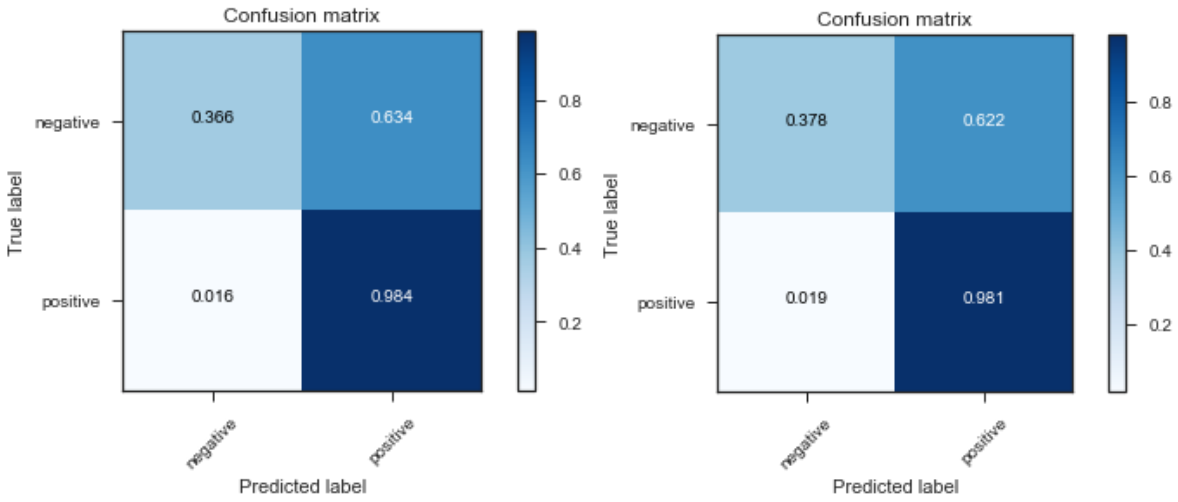

Figure 6. Confusion matrices of logistic regression (left) and LDA (right)

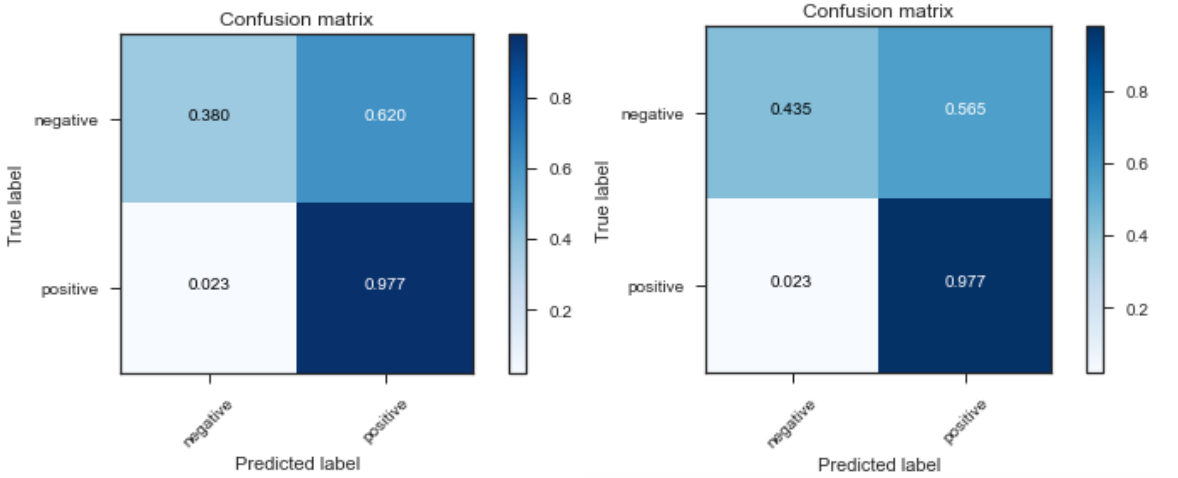

Figure 7. Confusion matrices of random forest (left) and neural networks (right)

Confusion matrices for each classifier are plotted (shown in **Figure 6** and **7**). The confusion matrices of logistic regression and regularized logistic regressions have no significant differences. The logistic regression classifier has a 98.4% sensitivity and a 36.6% specificity, which means that it has 98.4% probability of correctly classifying a responding customer and 36.6% probability of correctly classifying a customer that will not respond. The confusion matrix for LDA shows that it has a lower sensitivity at 98.1% but a higher specificity at 37.8%. For random forest and neural networks, the sensitivity rate is the same at 97.7%, while the specificity rates are 38% and 43.5% respectively. All four confusion matrices show very little false negatives i.e. minimal costs of missed opportunities.

It is clear that neural network is the best at predicting which customers will not respond. In that case, the cost of mailing to non-responsive customers is saved. It does, however, have the highest amount of false negatives. Logistic regression is better at discovering which customer are likely to respond to the direct mailing marketing promotion. By applying this classifier, more purchases will be made in the promotion period, though the cost of mailing will be higher.

Data Mining Results and Limitations

Model evaluation has suggested that the neural network model outperforms the other models. The neural network model has a 97.7% probability of correctly classifying a responding customer and 43.5% probability of correctly classifying a customer that will not respond. The business can apply this model to decide which customers should be involved in the direct mail marketing promotion, however, the associations between predictors and the response discovered by the neural network unknown and not interpretable.

Considering interpretability, the logistic regression output shows the coefficients of each predictor. The general associations between predictors and response has been analyzed in **Table 5**. The performance of logistic regression is good, since it has the highest sensitivity rate and the second highest profit of $1.71 per customer. Logistic regression performs well when the data has less outliers and the distribution of predictors are approximately normal, which to an extent has been achieved by data transformed and standardized.

As for limitations, the dataset studied in this analysis is not complete. The actual meaning of some predictors are absent, e.g., ‘PC_CAL20’, ‘STYLES’, and ‘STORELOY’. In future data collection, the company should aim to avoid this kind of error and try to get more detailed information of their customers. For example, they can keep a record of the sex and age of customers, when in the year they usually purchase, and customers’ comments on products they bought. More complete information would help to build better models and improve performance.

Although the decision threshold has been adjusted based on cost-benefit analysis, it is not guaranteed to be the optimal choice. Future statistical modeling can study the choice of threshold and how it can affect model performance. By getting a more optimal decision threshold, the performance of classification models can be further improved. Additionally, the threshold can be adjusted based on new knowledge of the business.

Deployment

We have applied several effective classification models which has provided insights to potential responders among thousands of customers. However, the statistical results alone are not enough to explain the business problem. Data interpretation and key information regarding deployment follows.

If the company does not apply the mail promotion strategy, that is, does not send mail to any customer, the company is expected to have a cost of $2.49 per customer. If the company sends mail to every customer, with the cost of mailing at $1, the company gains $1.49 profit per customer. If the company sends mails to specific customers based on the logistic regression model, $1.71 profit per customer is expected. In this case, using logistic regressio increases the profitability by 14.77%. This increased profit primarily comes from decreasing the cost of sending promotions to nonresponsive customers.

More specifically, classification models for predicting whether customers would respond or not have the following commercial uses:

1. Logistic regression can help infer characteristics of possible responsive customers. As shown in **Tables 4** and **5**, the coefficients of the predictors in logistic regression can present the relationship between the response and those predictors. To some extent, it reflects which kind of customers that the clothing store is more likely to get a response from. Generally, a customer with more frequent purchase visits and longer time on file is more likely to respond to the direct mail marketing promotion. Also, the number of promotions that a customer received has a positive association with the probability of responding. Upon identifying these common characteristics of responders, the company would benefit from devising a strategy to specifically target those kinds of customers.

2. Furthermore, customer relationship management would be more effective. The Forrester Report and Nail (2000) indicates mail promotions have no significant financial advantage in obtaining new customers compared to other promotion strategies. The advantage of mail promotions is to maintain customer relationships. By keeping records and analyzing customer information, regulars are identified and the company can inform them of the latest promotions on time. Effective feedback are also obtained after each mail marketing promotion, which helps keep customers' information up to date. This leads to cost-efficient sales and cross-sellin. From this, the company can establish long-term customer loyalty for further profit.

3. For all these models, if the estimated probability of a positive outcome is higher than 0.034, the customer will be expected to respond. In order to reduce substantial mailing costs, a suggested strategy can be only mailing to the potential responders rather than all customers on file. Moreover, if the business really wants to focus on the most likely customers, they can mail a promotion only if all the models classify the customer as responsive. However, these stricter classification standard could lead greater opportunity loss.

References list

Albon, C. (2016). *Random Forest Classifier Example*. Retrieved November 2, 2017, from https://chrisalbon.com/machine-learning/random_for est_classifier_example_scikit.html

Crane, M. (2007). *How To Set Up A Clothing Retailer: Important Performance Metrics.* Retrieved October 20, 2017, from https://www.forbes.com/2007/01/09/retailer-financial-metrics-ent_cx_mc_0109fundamentalsmetrics.html

Nail, J., & Forrester (Firm). (2000). *The email marketing dialogue*. Cambridge, MA: Forrester Research.

Scharth, M. (2017). *QBUS2820 Predictive Analytics, Module 9, week 9: Classification I* [PDF]. Retrieved from https://elearning.sydney.edu.au/bbcswebdav/pid-4900899-dt-content-rid-20739523_1/courses/2017_S2C_QBUS2820_ND/QBUS2820-09.pdf

Scharth, M. (2017). QBUS2820 Predictive Analytics, Module 10, week 9: Classification II [PDF]. Retrieved from https://elearning.sydney.edu.au/bbcswebdav/pid-4933852-dt-content-rid-20860358_1/courses/2017_S2C_QBUS2820_ND/QBUS2820-10.pdf

Sumathi, S., Sivanandam, S. N., & SpringerLink (Online service). (2006). *Introduction to data mining and its applications*. Berlin, Heidelberg: Springer Berlin Heidelberg.