MODELLING CONSUMER RESPONSE TO MARKETING

Rhys Kilian

November, 2017

Executive Summary

This report outlines the approaches undertaken to build a model to predict the response of customer to direct mail marketing. This is becoming increasingly important for classic brick-and-mortar stores which are facing headwinds for the profitability for their stores. Therefore, developing a classification model to predict a customer's response to a direct mail marketing campaign will be important for the efficiency and profitability of their efforts.

This investigation begins by first collating the data from previous marketing campaigns, ensuring that it is free from errors. This is the data which will be used to train the models, as well as evaluate them once the model selection process is complete. However, to select a model a criteria was developed which was the expected profitability of a customer from the campaign. This first involved creating a cost-benefit matrix which defined the losses (and gains) associated with incorrect and correct classifications of response and non-response. The two models which produced the highest expected profitability per customer using the training data were then selected for further model evaluation. However, these candidate models were first tuned using a cross-validation approach, to maximise the precision prior to the model validation phase.

The two selected models, which were the Quadratic Discriminant Analysis (QDA) and Logistic Regression models, were then evaluated against two benchmarks. These benchmarks were models representing the situation in which direct mail was send to all of the customers, and to none of the customers respectively. The best performing model at this phase was the QDA model which had an expected profit per customer of \$0.75. This represented a 4.2% improvement over the benchmark in which direct mail was sent to all customers. Over the course of a year, it was concluded that this would result in a substantial improvement for the retail store, thus justifying this analysis.

Contents

1 Introduction

Direct mail marketing is one of many potential marketing strategies which can be used to promote sales. However, before marketing strategies are decided upon, it is important to have an understanding of the different strategies potential impact on sales and thus the overall profitability of the store. Predictive analytics, specifically classification analysis is a useful method for using prior data from the store to predict the customers response.

For the clothing store chain in question, data collected from previous customers can be used to classify which customers will respond to direct mail marketing. By identifying potential customers the clothing store is able to best target their direct marketing efforts to customers which are predicted to respond. Furthermore the store can also predict the overall impact of the direct marketing. Using the test data in the model evaluation revealed that the QDA model was the highest performing model of all models considered at this stage. Such that, the expected profit from each customer using this model is $0.75

2 Business Understanding

In recent years, and especially given the growth of Amazon and other online retailers, physical stores have seen a significant decline in profitability (Hodson, Perrigo and Hardman, 2017). Given the tough industry conditions, targeted and effective marketing to consumers is of critical importance to ensuring the continued success for a company. Current marketing strategies not only include in-store promotions and direct mail marketing, but also newer forms of e-marketing such as e-mail and web advertisements (Chinta, 2006).

The goal of developing a classification model for predicting a customers response to direct mail marketing ensures the highest return for marketing efforts, ultimately increasing the profitability of the store. In order to predict this, it is beneficial to develop a cost-benefit table for different classifications, as seen in Table 1, which enables a comparison of different combinations of classifications (correct/incorrect) in terms of an expected profit or loss.

Table 1: Cost-Benefit Matrix

Outcome	**Classification**	**Actual Response**	**Cost**	**Rationale**
True Negative	Nonresponse	Nonresponse	$0	No marketing costs, No profits
True Positive	Response	Response	-$19.78	Profit - Marketing costs
False Negative	Nonresponse	Response	$22.78	Lost profits, No marketing costs
False Positive	Response	Nonresponse	$3.00	Marketing cost, No profit

The cost benefit table was made from several assumptions. Is is assumed that the promotional materials and postage costs $3.00 per item. Furthermore the profit margin of the store is assumed to be 20%. Looking at the average spend per visit for a customer given by the data, the mean average spend per visit is $113.90, which would be the amount expected for a customer to spend in response to a marketing promotion.

- **True Negative** - No sales are made, no costs are incurred. The customer would not have responded to promotion had any promotion been sent.
- **True Positive** - Promotion is mailed (cost of $3), customer responds and spends average of $113.90, store makes a profit of $22.78. Thus the benefit is $22.78 minus cost of $3, leading to a total benefit of $19.78.
- **False Negative** - Promotion is not mailed to a customer who would have responded. No cost is incurred, however the store loses a potential profit of $22.78, thus incurring a cost of $22.78.
- **False Positive** - Promotion is mailed to a customer who does not respond. Only the cost of sending promotion is incurred $3.

For direct mail marketing, it can be seen that a false positive is far less serious, with significantly less costs, than a false negative.

3 Data Understanding

The classification model is based off detailed customer data, which contains 50 different predictors as well as the response of the customer to direct mail marketing (the response under consideration). These variables can be classed into 6 broad categories customer demographics, amount spent, purchase history, location of purchase, purchase characteristics and marketing history as summarized in Table 2.

Table 2: Broad Categories of Predictors

Category	**Variables**
Customer Demographics	Customer ID, Zip code, Microvision lifestyle cluster type (50 categories), Valid phone number on file
Amount Spent	Total net sales, Average amount spent per visit, Gross margin percentage, Credit card user
Purchase History	Number of purchase visits, Amount spent in past month, Past three months, Past six months, Amount spent in same period last year, Number of days customer on file, Number of days between purchases, Lifetime average time between visits
Location of Purchase	Amount spent in each franchise (4 variables), Number of stores customer purchased at, Web shopper
Purchase Characteristics	Number of different product classes purchased, Percentages spent on different types of clothing (15 variables), Product uniformity (low score = diverse spending), Percent of returns, Total number of individual items purchased by the customer
Marketing History	Number of coupons used by the customer, Number of marketing promotions on file, Number of promotions mailed in last year, Number of promotions responded to in the past year, Promotion response rate for past year, Markdown percentage on customer purchases

3.1 Exploratory Data Analysis

Gaining an understanding of the data is a crucial element to building a suitable classification model. In particular, it is useful to understand the behavior of key variables, the relationships between variables, as well as any anomalies which exists in the data. Only the training data is used to perform exploratory data analysis, leaving the test data for model evaluation only.

A common issue with large datasets is the presence of missing data, which reduces the available sample size for a predictor, potentially causing issues with modeling and excluding key facts about data. Fortunately, there is no missing data present in the dataset. This can be expected given the data appears to have been collected using an electronic customer database, thus giving complete information.

3.1.1 Univariate Analysis - Response

The first stage to understanding the training data is to investigate the distribution of the response variable. Out of the 15,218 data points in the training set, 2,534 have a response of 1 (indicating a response to the direct mail marketing) whilst 12,684 have a response of 0 (indicating no response). This corresponds to a 16.7% response rate to the direct mail marketing promotion.

3.1.2 Correlation between Predictors and Response

With 50 different predictors in the dataset, investigating the correlation between different predictors, as well as the correlation between the predictors and the response will allow for greater understanding of the relevance of predictors. Predictors which are highly correlated with the response can be seen as containing significant information for the prediction of the response. However high correlation between predictors could potentially suggest multi-collinearity issues are present, and that similar information about the response is contained in multiple predictors. In cases where the correlation between predictors is significantly high it is worth considering dropping one or more predictors to prevent any collinearity issues.

The full correlation matrix for all 50 predictors can be seen in Figure 10 in the Appendix A. From this matrix it can be seen that there are a couple of predictors which are highly correlated with one another. These include the variables high correlation between 'classes', 'coupons', 'styles' and 'stores', 'mailed' and 'responded', as well as 'tomonspend', 'omonspend' and 'smonspend'. It can be seen that these variables are somewhat related, for instance the number of promotions mailed would have a significant correlation with the number of promotions responded to.

Figure 1 shows the correlation between predictors and the response for the nine predictors most highly correlated with the response. The predictors 'fre' (the number of purchase visits), 'classes' (number of different product classes purchased), 'styles' (number of individual items purchased by the customer), 'responded' (number of promotions responded to in the past year), 'mon' (total net sales), 'coupons' (number of coupons used by the customer), 'stores' (number of stores customer purchased), 'responserate' (promotion response rate for the past year) and 'smonspend' are all highly correlated with the response. Given the high correlations, the distributions of these variables will be interesting to investigate.

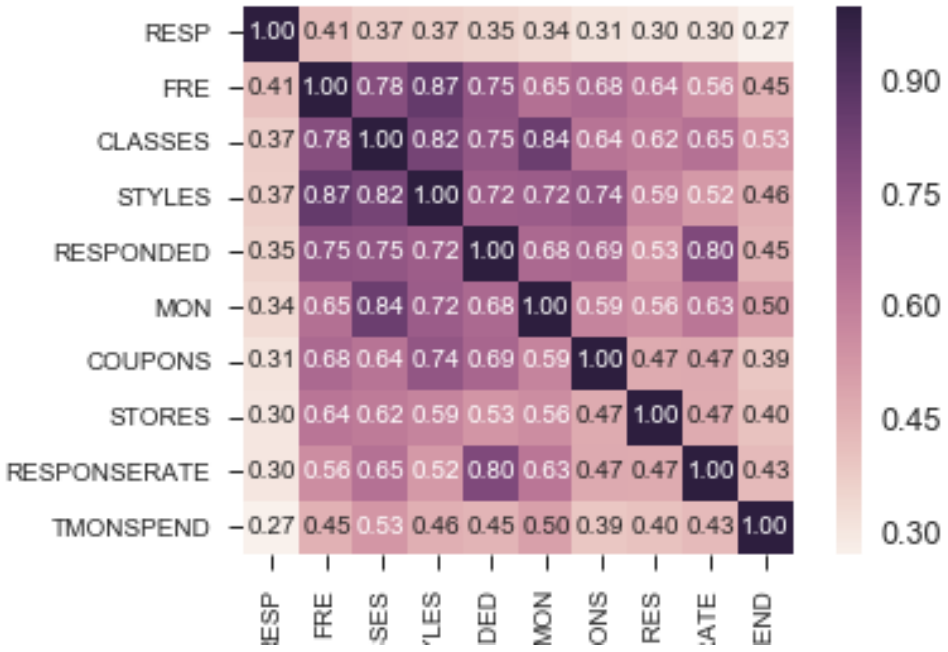

Figure 1: Correlation Matrix for Predictors Strongly Correlated with Response

Interestingly, the correlation matrix in Figure 1 shows that whilst some predictors are more highly correlated, there is no single important predictor, with the highest correlation being 0.41. In fact the most significantly correlated predictors have a correlation between 0.41 and 0.27, suggesting a similarly moderate correlation across many predictors. Thus it it likely that a number of predictors are needed to accurately classify the response.

The bivariate relationships between these significantly correlated variables shown in Figure 11 in Appendix A, reveals the different patterns of bivariate relationships between the response and predictors as well as predictors themselves.

3.1.3 Bivariate Analysis

In order to gain a greater understanding of both the dataset and the potential classification models, the bivariate relationship between key predictors and the response variable should be further investigated. However due to the number of predictors in the dataset, only a small selection will be considered as a generalization of trends found in the dataset. Further plots can be found in Appendix A.

One of the interesting variables to consider is the the microvision lifestyle cluster type, which segments the population into 50 categories based on family structure, income, education and profession. Figure 12 in Appendix A shows the relationship between these different cluster types and the response. From this plot, it would seem the store in question attracts prosperous clients with the wealthiest cluster, cluster 1, being the second most prevalent category.

A violin plot combines elements of a box plot and density plot, allowing the visualization of both the distribution of the data and its probability density. This is significant in revealing the modality of a distribution, as can be seen in Figure 2 and 3. For 'fre' Figure 2 shows that for both a response and no response, the common values are right skewed towards fewer purchase visits. However notably the mean and upper quartile range for a response is higher than no response, indicating a higher number and greater spread of

average visits. Figure 3 for 'classes' similarly shows that for no response, the number of different product classes purchased is right skewed, with a small spread of data. In comparison for a response, the number of product classes is more evenly distributed, with a higher mean and larger spread of data. Significantly it is noted that the majority of predictors are significantly right skewed, with a low mean and high upper quartile range.

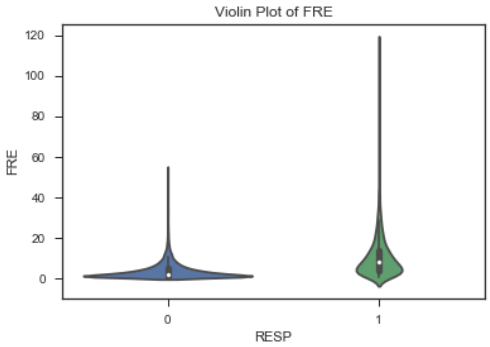

Figure 2: Violin Plot for 'fre'

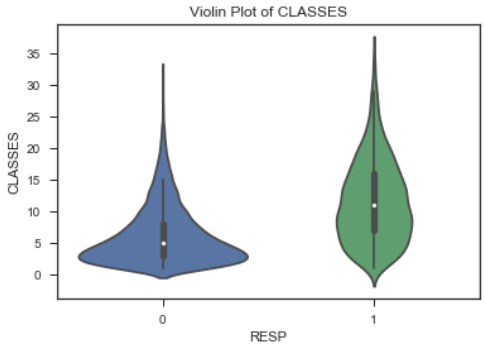

Figure 3: Violin Plot for 'classes'

A kernel density estimate (KDE) plot is another useful method to visualize the relationship between data, providing an estimate of the probability density function for a variable. For a binary classification problem, KDE is a useful tool as it allows for a direct comparison of the distributions between a response and no response. For some predictors such as 'avrg' (average amount spent per visit) shown in Figure 4, the distributions for response and no response are almost identical. In comparison for predictors including 'classes' as seen in Figure 5 the distributions are fairly unique, with different elements of peakedness (or spread) and different peaks (or maximum). These KDE plots likewise confirm the right skew present across most predictors.

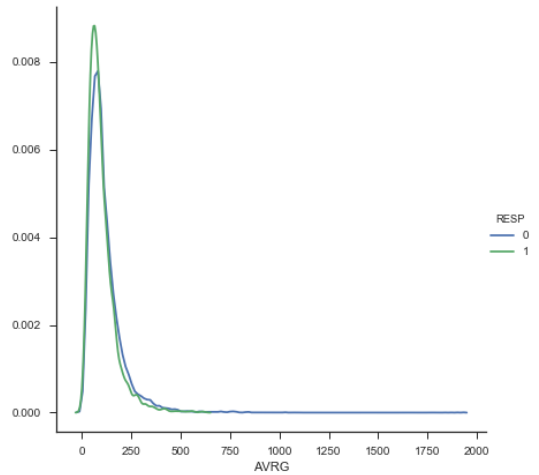

Figure 4: KDE Plot for 'avrg'

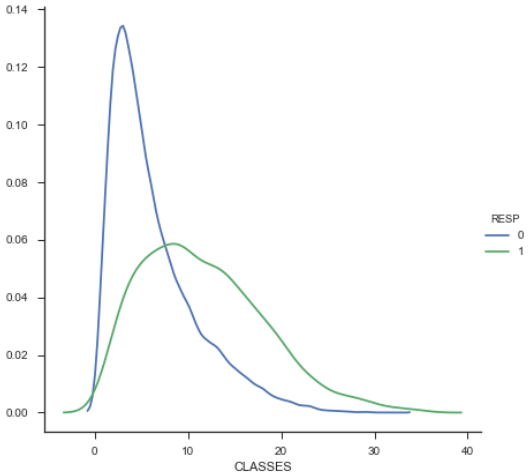

Figure 5: KDE Plot for 'classes'

The similarity between certain distributions and the skew of the data has the potential to violate particular model assumptions. Many classification models work best on symmetric data (if not data that is normally distributed). In particular, Gaussian discriminant analysis assumes the data to have a normal or Gaussian distribution and works on the principle of different classes having differing distributions. The potential violations of these assumptions should be considered in further data processing.

At this stage it is also interesting to plot a basic logistic regression onto the data, to further visualize the difference in the response between variables. Figure 6 shows a logistic regression on 'avrg', for which the logistic curve does not follow the typical s shape. As previously discussed, the difference between response and no response for this variable was minimal, potentially indicating that 'avrg' does not contain significant information for the prediction of a response. In contrast, the logistic regression of 'classes' as seen in Figure 7 shows a typical logistic curve, indicating a clear relationship between classes and the response.

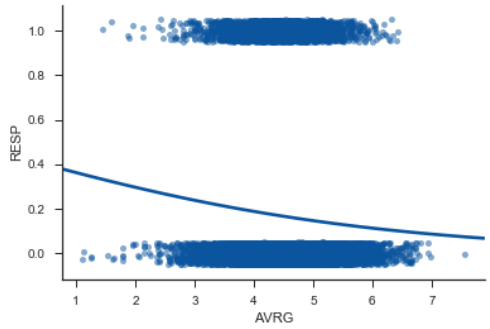

Figure 6: Logistic Plot for 'avrg'

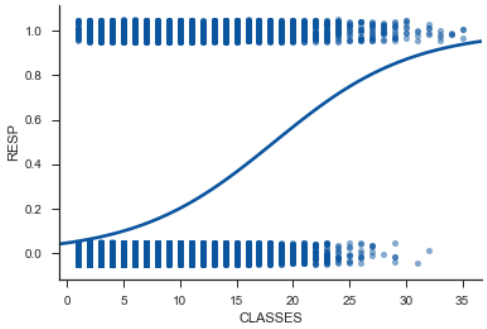

Figure 7: Logistic Plot for 'classes'

It can be hypothesized based off these plots, that the predictors with a higher correlation to the response will have clear difference between the distribution of the response and no response classes as well as a clear relationship when a basic bivariate logistic regression is applied.

4 Data Preparation

Given the large number of variables present in the dataset, it is useful to undertake a degree of data preparation, including selecting, cleaning and transforming variables to ensure these variables are suitable for constructing a classification model. Significantly, the same processing should be applied to both the training and test data. This ensures that the model can be used on the test set.

4.1 Data Selection

Firstly, certain variables can be excluded from further analysis. Customer ID is used to identify each customer, and is an encrypted value. Thus it contains no information which will be useful to predicting which customers are likely to respond to mail marketing. The zip code of the customer is seemingly numeric, however is a categorization of a geographic location of the customer. Whilst this variable might be of some use for prediction later, it is removed in the meantime due to its large variability.

4.2 Feature Engineering

To improve the classification performance of the two models, feature engineering was used to create additional predictors from the data provided. Feature engineering is also advantageous as it generally generates variables which yield more flexible and simple models through the transformation of data and the addition of predictors.

The creation of additional predictors was informed by the exploratory data analysis conducted at the beginning of this report. For instance, observing the bivariate relationship between the response and the predictor, and the distribution of the predictor, it is possible to determine whether non-linear terms are appropriate for variable selection. In addition, features were constructed by understanding what the variable represent. For example, two additional spending variables were created by a simple subtraction of cumulative sales values.

As expected, careful consideration was made to ensure that these additional predictors, and the feature engineering conducted, did in fact improve the classification performance. Techniques such as L1 logistic regression were used as a variable selection method, which were a subset of the larger model selection problem which used cross-validation to verify the classification rates did improve by such feature engineering.

4.2.1 Spending Variables

The data contains three cumulative spending variables which were of interest, which represented the amount spent by a customer in the last: month, three months and six months. As these are cumulative values, the amount spent by a customer in the previous month is also going to be contained within the last three and six months. Hence, this is redundant data, which is being triple counted within the model building process.

By subtracting the amount spent in the last month from the last three months, and the last three months from the last six months, two more appropriate variables were created as shown below. The cumulative variables for the past three and six months were also dropped.

Table 3: Feature Engineering: Spending Variables

New Variable	**Previous Variable**	**Feature Engineering**
Month23	TMONSPEND - OMONSPEND	Subtracting last months spend from the previous three to create spending variable in month two and three
Month456	SMONSPEND - TMONSPEND	Subtracting previous three months spend from the past six to create spending variable in months four, five and six.

However, it should also be noted that this technique produced values which were negative. This is not intuitive as it is expected that the cumulative values of a greater number of months should be larger than or equal to a lesser number of months. It was therefore concluded that this was due to errors in the entry of data, and any observation with values less than zero were removed to prevent NaN issues after taking a log or square root transformation.

4.2.2 Polynomials

It was also possible to estimate whether a linear relationship between the response and the predictor was appropriate using the EDA. In cases where the relationship appeared to be non-linear, higher order polynomial terms were constructed to capture this relationship. Furthermore, square root variables were also created to model this possible relationship too. A non-exhaustive list of the polynomial terms created as shown in table 4.

Table 4: Feature Engineering: Polynomials

New Variable	Previous Variable	Feature Engineering
RESPONDED-s2	RESPONDED	Squaring the number of promotions responded to in the past year
RESPONDED-3	RESPONDED	Cubing the number of promotions responded to in the past year
RESPONDED-Sq	RESPONDED	Square rooting the number of promotions responded to in the past year

4.2.3 Dummy Variables

Dummy variables were also used to represent categorical variables. In this data set only one predictor needed to be engineered into dummy variables and was the VALPHON. This therefore became a variable which took the value of 1 when the customer had a valid phone number of file.

4.3 Data Formatting

4.3.1 Skewed Data Transformation

Furthermore, three data transformations were performed to reduce the skewness of the data. This included the box-cox, natural logarithm one plus, and the square root. As several models used in the investigation depend on the distribution of data, namely the LDA and QDA methods, it is expected that making the data more normal, or at least more symmetrical, will improve classification performance. As Doane and Seward (2011) proposed, it is appropriate to transform data which has a skewness of greater than 0.5.

While there is much debate as to most appropriate data transformation method, it was concluded that using a cross-validated approach would be optimal. As such, each of the three transformation methods were applied to the data, and the method which yielded the lowest classification error rate selected. It was found that natural logarithm one plus transformation had the lowest classification error rate, followed by the box-cox and finally the square root. Hence, the natural logarithm one plus transformation was used.

5 Modeling

Several different classification models were initially considered, including k nearest neighbors, logistic, L1 regularized, L2 regularized, linear discriminate analysis, quadratic discriminate analysis and regularized quadratic discriminate analysis. Ultimately, using the cost-benefit matrix established in Section 2 it is possible to determine the generalization performance of the models and thus select the best performing models. However hyper-parameters will be selected using cross-validation precision as discussed in Section 8.

5.1 Modeling Techniques and Assumptions

Whilst there are many different models which could be considered, the following models were chosen due to their relatively small computational intensity, ability to be interpreted and characteristics of the data. With classification models, no one method dominates in every situation, as each result in differing decision boundaries and rest on different assumptions.

5.1.1 Justification for Models Considered

There are many different classification models which could be considered and implemented on this data. The models considered below have been selected based on computational intensity, interpretability and domain knowledge.

Notably both parametric (logistic, Gaussian discriminant analysis) and non-parametric (KNN) methods have been used. These models are considered standard for classification problems, and often yield reliable results with minimal assumptions. These three broad model classifications further consider different approaches to classification.

5.1.2 Logistic Regression - Baseline

The logistic regression is a linear regression model which has a categorical response variable. In essence the logistic regression models the conditional distribution of the response given the predictors. In order to ensure that the probability of X lies between 0 and 1, the logistic function as seen in Equation 1, is used.

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}} \tag{1}$$

Using the maximum likelihood method, to fit the model, and manipulating the result, it is possible to obtain an equation which is linear in X, as shown by Equation 2. The maximum likelihood method estimates the coefficients such that the predicted probability corresponds as closely as possible to the observation. The left hand side of this equation is called the logit. An increase of one unit in X, results in a β_1 change in the log odds.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{2}$$

The logistic regression is most similar to a linear regression. The model output can be used as in Equation 1 to predict the probability of being in a certain category. Crucially, logistic regression does not assume a distribution of the observations.

Regularization methods, which shrink the coefficients of a model towards zero can also be applied to the logistic regression. These methods help to improve prediction accuracy, by reducing the variance of the model. No one regularization method outperforms the other leading to relying on cross-validation to determine the best approach.

- **L1 Regularization** - L1 penalty regularization on the logistic regression works in a similar method to the lasso regularization for a linear regression, using the penalty term $\lambda\Sigma|\beta_j|$. The L1 penalty shrinks the coefficients of the classifier towards zero, and for certain λ values forces the coefficients of some variables to be zero, allowing for both regularization and variable selection. Cross-validation is used to determine the optimal value of λ. L1 regularization is suited to models with a small subset of significant predictors.

- **L2 Regularization** - L2 penalty regularization for the logistic regression is comparable to the ridge regression, with a penalty term $\lambda\Sigma\beta_j^2$ which shrinks the coefficients of the classifier towards zero. In L2 regularization, coefficients of similar importance are shrunk by the same amount, reducing the impact of correlation between predictors. However L2 regularization does not perform any variable selection, leading to potential interpretability problems. Again λ is selected using cross-validation error.

5.1.3 Linear Discriminate Analysis (LDA)

In contrast to previous methods discussed, LDA models for each response class the distribution of predictors X ($Pr(X = x|Y = k)$, the uses Bayes' theorem as shown in Equation 3 to determine $Pr(Y = k|X = x)$. When the distributions of the predictors are close to normal, LDA and logistic regression returns similar results.

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_l^K = 1\pi_l f_l(x)} \tag{3}$$

If the terms are specified correctly, the Bayes classifier will have the lowest error rate for all potential classifiers. LDA is useful for situations where classes are well-separated, the number of samples (n) is small and X is approximately normally distributed in each class, or there are more than two response classes. LDA makes the assumption that the observations are taken from a Gaussian distribution, which has a mean vector and covariance matrix common across the two classes.

However LDA does somewhat rely on the poster probability threshold, which for a two-class case is normally set to 50%, which has the lowest overall error rate. Lowering this threshold will result into more observations being classed into the default class. Given the distribution of the response variables under consideration, and the relatively small cost of a false positive, the threshold of 0.5 was used.

As the covariance matrix is assumed to be the same for all classes, LDA has a comparatively smaller number of parameters. Whilst reducing the flexibility of the model, this also results in a lower variance. However if the assumption that the covariance matrix is the same is wrong, LDA can have high bias. LDA is thus recommended for cases where a reduction in variance is required, such as having few training observations.

5.1.4 Quadratic Discriminate Analysis (QDA)

QDA like LDA uses Bayes' theorem to predict the class of the response, assuming observations follow a Gaussian distribution. However under QDA, a different covariance matrix is calculated for each class. According to the Bayes classifier, under QDA the observation is assigned to class k for which Equation 4 is maximized.

$$\delta_k(x) = -\frac{1}{2}(x - \mu_k)^T\Sigma_k^{-1}(x - \mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k \tag{4}$$

QDA, having variable covariance matrices, results in a flexible classifier, which has quite high variance. Thus QDA is preferable where variance is not a major concern, such as larger training sets, or where the classes cannot be assumed to have a common covariance matrix. Like KNN, QDA can be applied to problems with a non-linear decision boundary, as QDA results in a quadratic decision boundary.

- **Regularized Quadratic Discriminate Analysis** - Regularization methods can also be applied to QDA, to achieve a compromise between the covariance matrix

properties of QDA and LDA. Equation 5 shows how the covariance matrix can be calculated, for an α selected by cross-validation.

$$\Sigma_k(\alpha) = \alpha\Sigma_k + (1-\alpha)\Sigma \tag{5}$$

5.1.5 K Nearest Neighbors (KNN)

KNN is a non-parametric approach to classification, which uses K observations closest to X, specifically those belonging to class k, to calculate the conditional probability. The KNN approach results in an approximation of the Bayes classifier, which is nonparametric.

$$P(Y = k|X = x) = \frac{1}{K}\Sigma I(y_i = k) \tag{6}$$

K in the KNN classifier is a hyper-parameter, which should be determined using a cross-validation approach. A lower K results in a more flexible decision boundary. As a non-parametric approach, KNN is suitable for problems with a highly non-linear decision boundary. However a draw-back to this method is no information can be gained about the importance of various predictors.

5.2 Model Building

5.2.1 Scoring Metric

The goal is to find a model which maximizes the generalization performance. This is equivalent to finding the model which yields the largest expected profit based on the validation data. Firstly, the training data which is 80% of the entire data set is further split into a validation and a training set. The models described in the previous section are built using the training data, with consideration of the hyper-parameters which affect the generalization performance. This generalization performance is then estimated using the validation set. Due to the relatively large size of the data a training and validation set was deemed to be appropriate.

Moreover, it was also inappropriate to use the accuracy of prediction alone as the measure of model performance. This is because the penalties for misclassification are not equal, as defined by the cost-benefit matrix in Section 2. Therefore, the scoring measure used in this analysis is the estimated profitability per customer, is given by the following formula:

$$\hat{R} = \frac{TN \times L_{TN} + FP \times L_{FP} + FN \times L_{FN} + TP \times L_{TP}}{N} \tag{7}$$

The values for the loss function are provided in the cost-benefit matrix in Section 7. Therefore, it is intuitive that the best performing model is the one that maximizes the number of true positives and minimizes the number of false negatives. This is because the coefficients in the loss function have the largest impact on the expected profitability from a customer. Hence, the model which yields the largest expected profitability per customer will be the model selected based on the training data.

5.2.2 Hyper-parameters

The models presented in Section 5.1 are dependent on the selection of hyper-parameters. To determine these values a ten-fold (k=10) cross validation method was used in which the candidate hyper-parameter was fit to the k-1 folds and tested on the kth fold. For each of the k folds the precision was calculated. Precision is given by Equation 8.

$$P(Y = 1|\hat{Y} = 1) = \frac{TP}{TP + FP} = \frac{TruePositives}{PositiveClassifications} \tag{8}$$

Precision is one of the many metrics which could be used to measure classification performance. Generally, when classifiers are considered, accuracy is one of the most popular metrics and can be defined as how often the classifier is correct. However, for the purposes of this business problem, accuracy is not an appropriate measure of success due to the imbalances in correct and incorrect classifications as outlined in the cost-benefit matrix.

When the formula of precision is considered, it is intuitive to see that it is a measure of when the prediction is responded, how often is it actually correct. Since the ultimate goal of this analysis is to develop a classifier which maximizes profits generated from direct mail marketing, it is optimal to maximize the number of true positives and minimize the number of false positives, hence precision is an appropriate metric to be used.

Finally, the value for the hyper-parameter was selected based on the average precision on each of the k folds. The values for the hyper-parameters are provided in Table 5 and are shown in terms of a precision vs hyper-parameter plot for L1 logistic, L2 logistic, QDA regularized and KNN in Figures 16 to 19 respectively in Appendix C.

Table 5: Summary of Hyper-parameter Results

Model	**Hyper-parameter**	**CV = 10 Precision**
L1 Logistic	$\lambda = 0.0001$	0.318
L2 Logistic	$\lambda = 1292$	0.800
QDA Regularized	$\alpha = 0.75$	0.421
KNN	$k = 2$	0.480

5.3 Model Assessment Results

Using the validation set results the model which would be put forward to the company is selected. The candidate models are compared based on the expected profit per customer involved with the direct marketing campaign. The model which resulted in the highest expected profit would be the one which was selected and further evaluated using the test data. As expected, attention was made to ensure that the proposed model had not been over-fit on the training data. This was ensured by using a randomized sampling process when selecting the validation data, and via consideration of the bias-variance trade-off.

The expected profit per customer is shown in Table 6. The values shown in this table are the negative value (i.e. higher is better) of the loss function shown in Equation 7.

Table 6: Summary of Validation Set Results

Model	**Expected Profit per Customer ($)**
Logistic Regression	-1.558
L1 Logistic Regression	-2.073
L2 Logistic Regression	-2.977
LDA	-1.310
QDA	0.732
QDA Regularized	0.128
KNN	-3.073

Using Table 6 it is clear which models will not be appropriate for the client. Since the overall goal is to increase the profitability of the direct mail marketing campaign we can immediately discard those models which have a negative expected profit per customer.

Firstly, the logistic regression model performed poorly on the validation set. This could be due to a number of reasons, however, it is theorized that this classification technique does not perform well as it does not penalize for model complexity. As a result, the variance of the predictions may have increased which therefore results in predictions which are not precise as reflected in the low expected profit per customer. Additionally, multicollinearity issues may have been present, as shown in the high areas of correlation in Figure 10.

Surprisingly however, the L1 and L2 logistic regression methods performed the worst of all the candidate models. When analyzing the coefficients in the L1 logistic model, as shown in Figure 14 in Appendix B we observe than only 5 of the predictors have been selected. This is due to the penalty term of the model penalizing for complexity, resulting in a sparse model. While this reduces model complexity, it does have the effect of increasing bias under the bias-variance trade-off. As a result, this overly simplistic model that has a poor generalization error. Furthermore, the L2 logistic model selects coefficient values for the predictors which are similar to the logistic regression model. The logistic regression and L2 logistic models are shown in Figures 13 and 15 in Appendix B respectively. It is proposed that the L2 model suffers from the same variance and potential multicollinearity issues as the logistic model.

The models which performed the best from the candidate models are based on discriminant analysis. The LDA model performs the worst of these discriminant models, yielding a negative expected profit, suggesting that the assumption of a common covariance matrix across classes is not appropriate. However, the QDA and QDA regularized models both exhibit a positive expected profit per customer based on the validation data. This is promising as the end result is to find a model to increase the profitability of the direct mail marketing campaign. Of these two models, the QDA model without regularization is the best performing. This implies that the covariance matrix between classes is distinct, and therefore should be considered when completing further analysis.

Finally, the KNN model was the worst performing of all the models considered. This is likely due to the number of neighbours not being optimally selected under the bias-variance trade-off. In this instance, only values of $k = 1, 2, 3$ were tested due to the large computational expense. For this reason, the model is likely to be too complex, resulting in high variance but low bias.

However, the model evaluation phase will consider the two best models based on the results shown in Table 6. These are the QDA model and the logistic regression models.

6 Evaluation

For the model evaluation phase the test data will be used which is randomly sampled from 20% of the original dataset. During the evaluation stage the performance of the QDA model selected in the previous section will be compared against two baseline models. As such, the accuracy and generality of the model can be assessed. However, particular attention will be made to determining whether the model meets the business objective of increase the direct mail marketing profitability.

6.1 Baseline Model Performance

To determine the success of the proposed QDA model two benchmark models were established. These are simply models based on the following premises:

1. Send to everyone - sending promotional material to all customers
2. Send to no one - not sending any promotional material at all

These are considered to be the absolute minimum levels of performance at which the QDA model needs to exceed. As in the model selection stage, the metric used to compare the models is the expected profitability per customer. Therefore, the costs and benefits associated with correctly classifying customers will be considered which is therefore important to the clients profitability goal. The results of the two baseline models are shown in Table 7 noting that negative values imply a profit is generated as outlined in Table 1 in Section 2.

Remembering that a negative value indicates a profit it can be observed that the send to everyone model is the best performing of the benchmarks. This statistic per customer is the minimum level at which any candidate model needs to outperform on the test data.

It is interesting to note that despite the high error rate of 83.7% this model results in a profit to the company. However, when the type of error is taken into account this result makes sense. Clearly, the company would like to maximize the true positive result and can afford to mis-classify customers provided that it is a false positive due to its low cost. Nonetheless, the QDA model is evaluated against the send to everyone model.

Table 7: Baseline Model Evaluation

Model	**TN** ($0)	**TP** ($-19.78)	**FN** ($22.78)	**FP** ($3.00)	**Error Rate**	**Overall Cost**
Send to everyone	-	700	-	3604	83.74%	-$3034.00 (-$0.70 /customer)
Send to no one	3604	-	700	-	16.26%	$15946.00 ($3.70 /customer)

6.2 Model Evaluation Proposed Models

Using the test data, a similar performance table has been produced for the QDA model as well as the baseline logistic regression model. This is shown in Table 8.

Table 8: QDA and Logistic Regression Model Evaluation

Model	**TN** ($0)	**TP** ($-19.78)	**FN** ($22.78)	**FP** ($3.00)	**Error Rate**	**Overall Cost**
QDA	320	682	18	3284	76.7%	-3227.92 (-$0.75 /customer)
Logistic Regression	3459	221	479	145	15.5%	$6975.24 ($1.62 /customer)

Immediately, the evaluation of the QDA model shows that the expected profit from the direct marketing mail campaign for the company will increase using this technique. The $0.75 per customer represents a 4.2% improvement over the baseline model of send to everyone. On the other hand, the logistic regression model clearly under-performs. Despite

its lower error rate, the values contained within the loss matrix penalize misclassification heavily for false negatives of which the logistic regression model contains substantially more than the QDA model. Therefore, of these two models, the QDA model which has an expected profit of $0.75 per customer is the more preferred model.

While this amount may seem insignificant, it is a solid improvement when considering the scale at which the direct mail campaign may be operating at. For instance, if the company is assumed to be a large conglomerate company in Australia in which 1 million direct mail letters are sent, the higher performing QDA model will generate an expected $30,000 more profit per campaign than the send to everyone model, assuming the test data is representative of the population. Over several direct marketing mail campaigns, the expected improved performance of the QDA model will result in a larger bottom line profit value than the company would have otherwise expected.

Finally, it is also interesting to note the lower error rate of the QDA model when compared to the baseline model. While this value is still relatively large, the low misclassification rate for false negatives ensures that the overall cost to the expected profit is minimal.

6.2.1 Expected Profit Confidence Intervals

Nonetheless, it is important to consider the uncertainty associated with these estimates. For this reason the confidence intervals for the estimated profit have been calculated.

The confidence interval are based on the z-scores under the normal distribution, where the full calculation is provided in Appendix D. The results are shown in Table 9 for the QDA and logistic regression models. Again, here a negative value represent a profit, whilst a positive value represents a loss.

Table 9: Expected Profit Confidence Intervals QDA and Logistic Regression

	90	**95**	**99**
QDA	(-2345.12, -4138.47)	(-2177.66, -4314.54)	(-1852.26, -4661.10)
Logitic Regression	(7189.59, 6740.59)	(7229.77, 6694.573)	(7307.02, 6602.92)

Clearly, the QDA model is again the best performing model using the test data for model evaluation. Again all confidence intervals for the QDA model give an expected profit just below, if not exceeding the expected profit from the baseline, send to everyone model. Meanwhile the logistic regression confidence intervals show not even the lower bounds of the confidence intervals come close to matching the performance of the benchmark.

6.2.2 Assumption Checking

Furthermore, it is also important to check that the assumption inherent to the model design are met. The success of the QDA model is dependent on whether the classes are Gaussian distributed. This can be checked by analyzing the distribution for each of the classes, responded and not responded, for the predictors in the model. This process was completed with example plots shown in Figures 8 and 9, with further examples in the Appendix E, Figures 20 and 23.

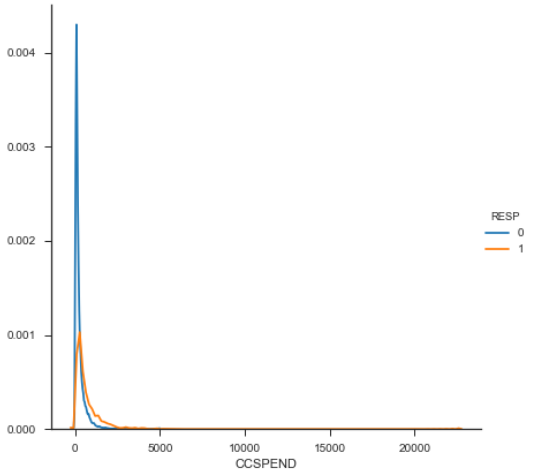

Figure 8: KDE, Before Transformation

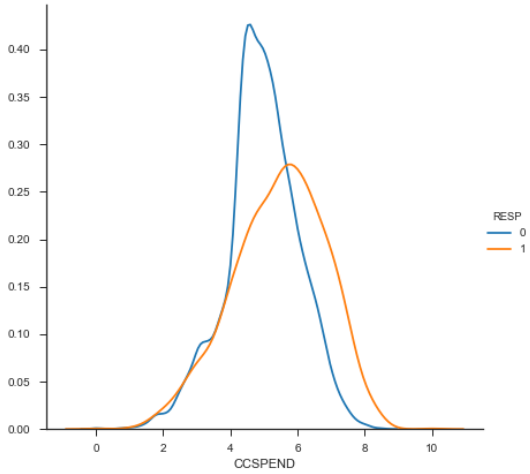

Figure 9: KDE, After Transformation

As can be seen from these plots, for the non-transformed variables, there is significant negative skew and high kurtosis. With this skew and kurtosis, the assumption of Gaussian distributed data would not hold, and thus QDA would not be appropriate. However for the log transformed variables, the data more closely resembles are normal distribution. For these plots, the data is exhibits only slight skew and kurtosis. Thus it can be seen the assumption of normality holds, and this transformed data meets the conditions required for QDA.

6.3 Next Steps

Under the CRISP-DM process a sub-activity of the model evaluation stage is determining the next steps. It is possible that the QDA model analyzed throughout the model selection phases could be improved upon, independent of the test data. For example, the following activities could be completed which could potentially result in a greater generalization performance on the training data for this particular model:

- *Data transformation*: this report considered three data transformation techniques. However, since the QDA model is heavily dependent on Gaussian distribution data, several other data transformation methods could be used to generate more normal predictor distributions
- *Feature creation*: additional features could be engineered to generate more insightful and simplistic approaches to modeling the data
- *Re-balancing the data*: the training dataset is heavily weighted towards customers who did not respond to direct mail marketing. Therefore, re-balancing would ensure that the model selected is not biased in this sense

Nonetheless, the decision has been made to proceed with the deployment phase. This is because the business objective of determining a model to generate a model with an expected profit per customer under direct mail marketing has been met.

7 Deployment

Since the QDA model was the most promising model selected based on the training and test data, yielding a positive profit per customer, it is the appropriate model to be deployed. Using this classification model the company should only send to customer in which the model predicts that they will purchase after receiving direct mail. However, it is important that the company record the information and the results of this further direct mail marketing campaign. Continued data collection should be considered, in a similar fashion to the data collected for building this model (which appears to originate from a customer loyalty program). The QDA model proposed can be further improved and developed using similar classification techniques outlined in this report. Training the model with additional data should lead to classifications which are more precise, and should be expected to yield a higher profit per customer.

8 Conclusion

In summary, the model put forward to the client for their direct mail marketing campaign was the Quadratic Discriminant Analysis (QDA) model. This classifier was estimated to have an expected profitability of \$0.75 per customer which represents a 4.2% improvement over a "send to everyone" strategy. Over the course of a business year, this is expected to result in increased profitability for the client and greater transparency for their decision making processes.

This model was selected on the bases of the expected profit per customer of which a number of different binary classification models were tested. Each of these models were trained using data which had been transformed in such a way to maximize the results for the client. However, the QDA model which resulted in a profitable score based on the training data, and the logistic regression model which did not, were selected for further evaluation. Again, the QDA model performed well under the unseen test data which confirmed the notion that this model would be the most appropriate based on the direct mail marketing needs. In fact, this model was compared against two potential baseline strategies and again was found to outperform these. Therefore, the QDA model is the model which should be deployed for the client's next direct mail marketing campaign to help increase expected profitability.

9 References

Chinta, R. 2006, *"Retail Marketing Trends in USA and Their Effects on Consumers and The Global Workforce"*, Business Renaissance Quarterly, vol. 1, no. 2, pp. 65-79.

Hodson, N., Perrigo, C. and Hardman, D. (2017). *2017 Retail Trends*. [online] Strategy&. Available at: https://www.strategyand.pwc.com/trend/2017-retail-trends [Accessed 22 Oct. 2017].

Doane, D. and Seward, L. (2011). Measuring Skewness: A Forgotten Statistic?. *Journal of Statistics Education*, 19(2), pp.1-18.

A Additional EDA

Covariance of Predictors

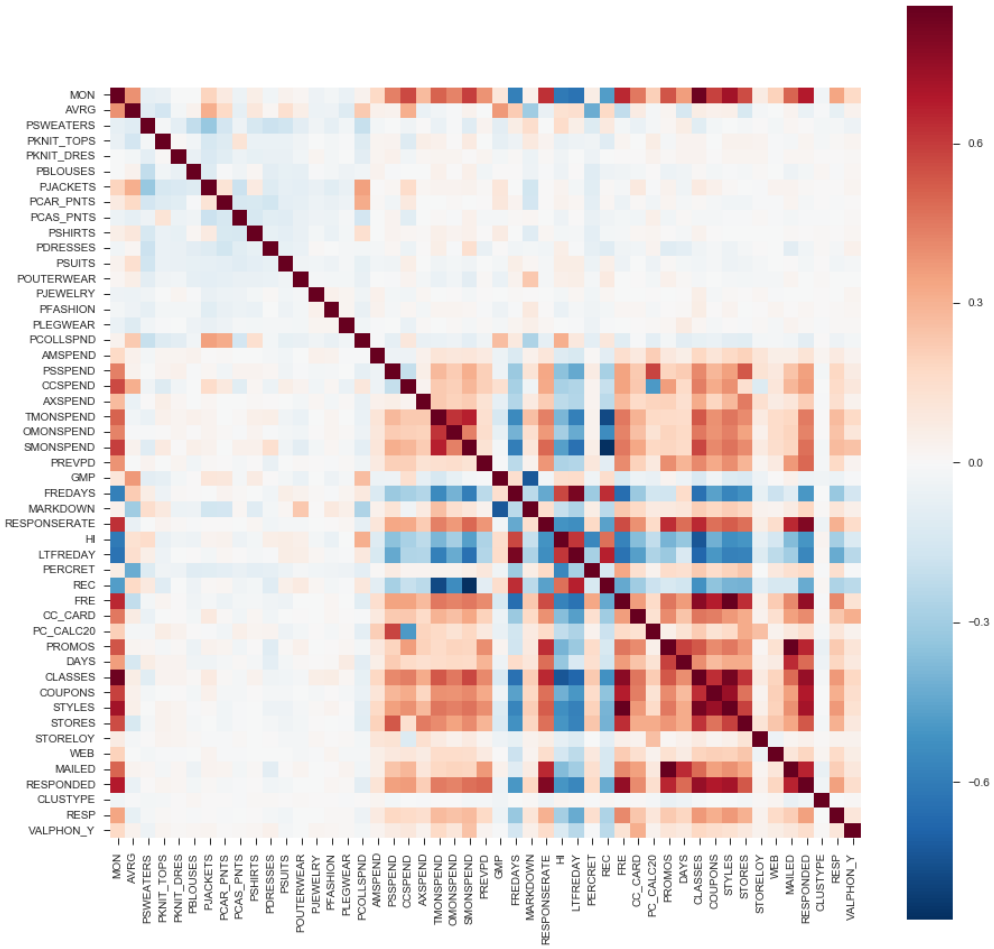

Figure 10: Correlation Matrix for All Predictors

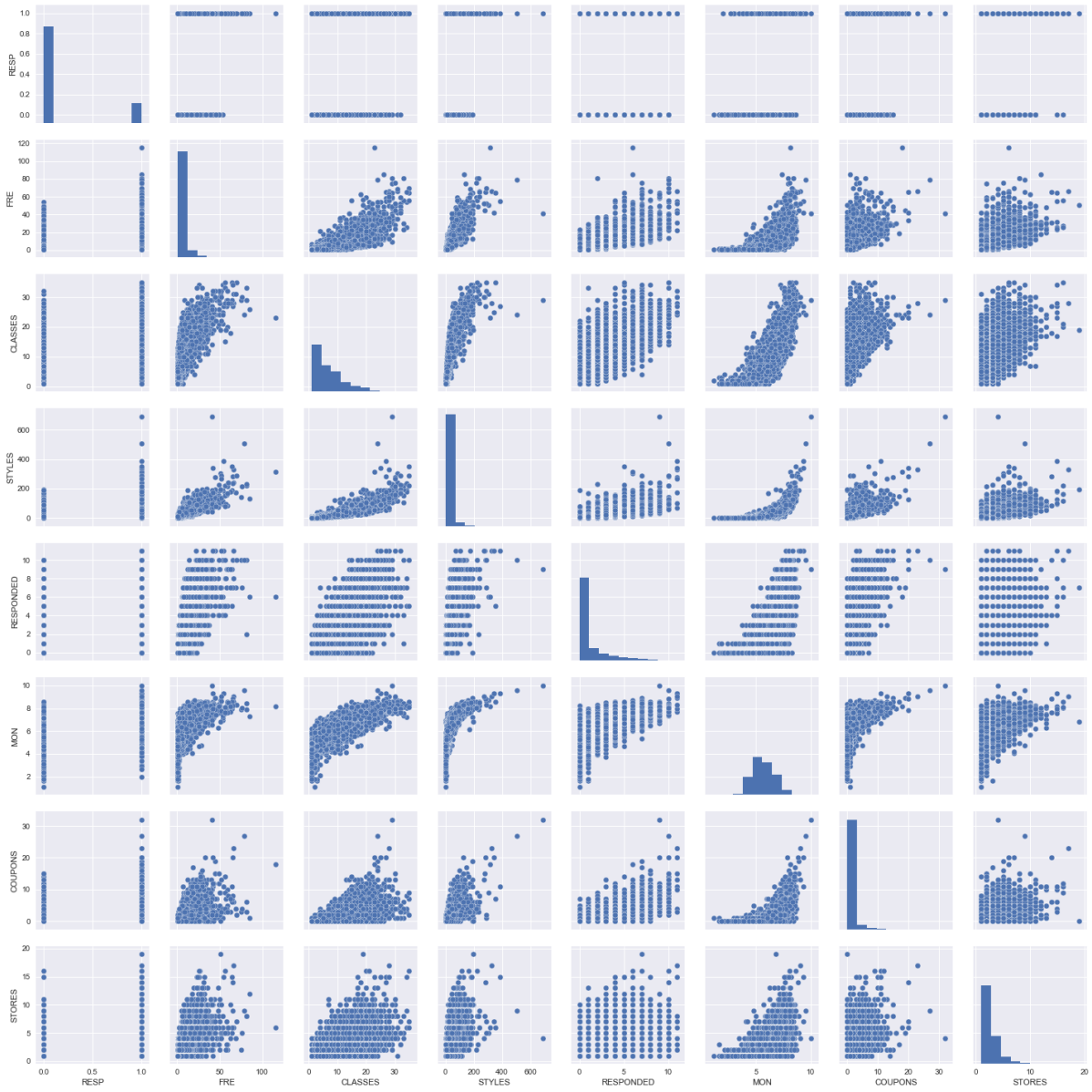

Figure 11: Correlation Matrix with Scatter Plots for Strongly Correlated Predictors

Bivariate Analysis

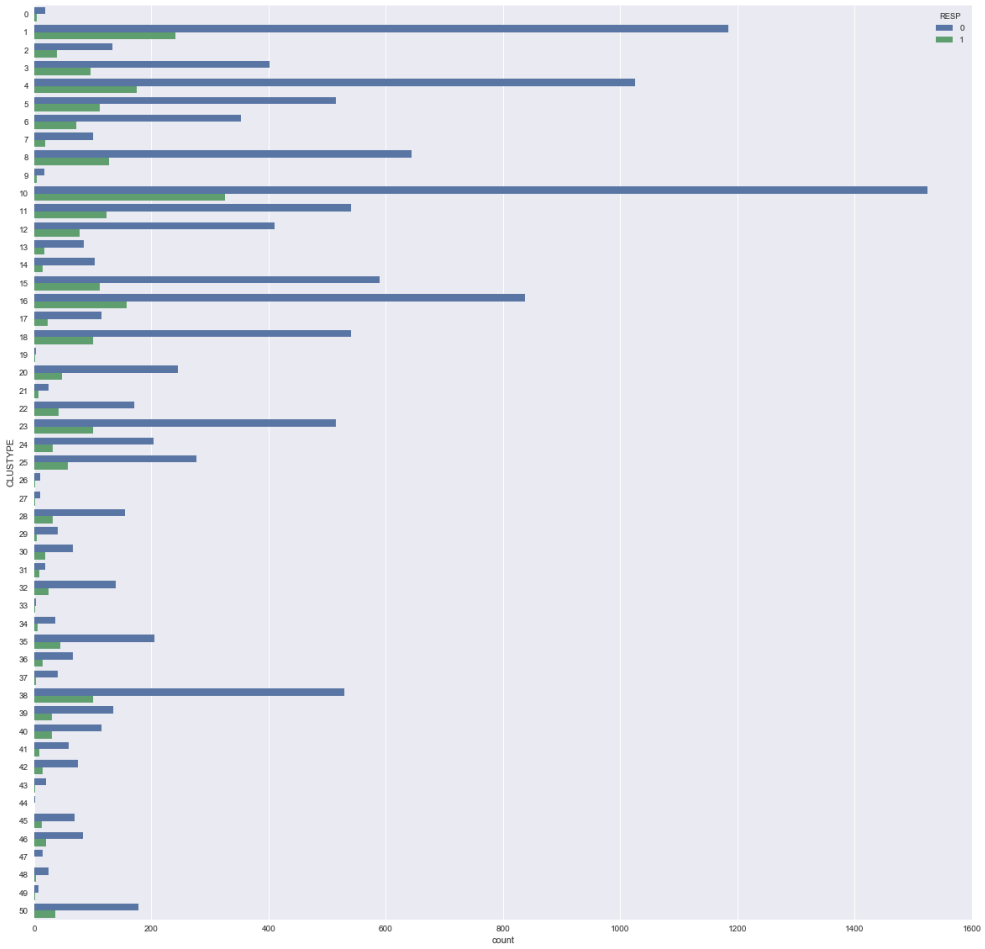

Figure 12: Microvision Count

B Model Coefficients

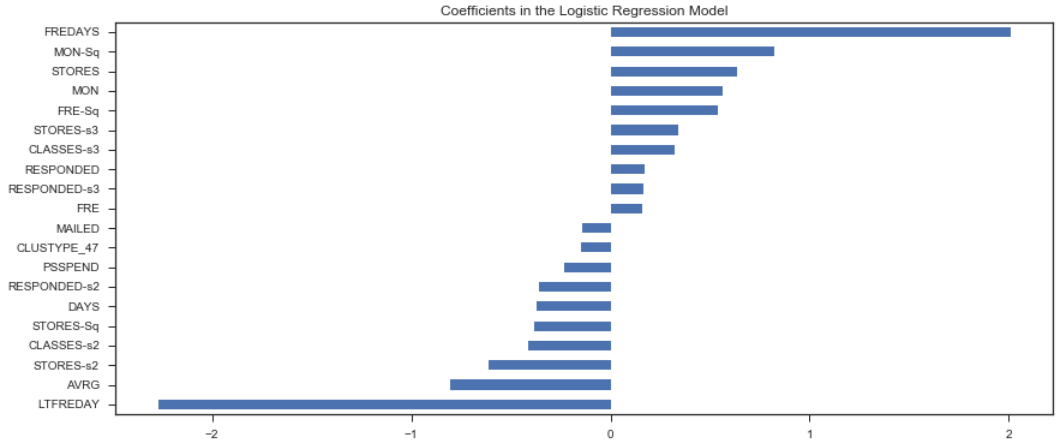

Figure 13: Logistic Coefficients

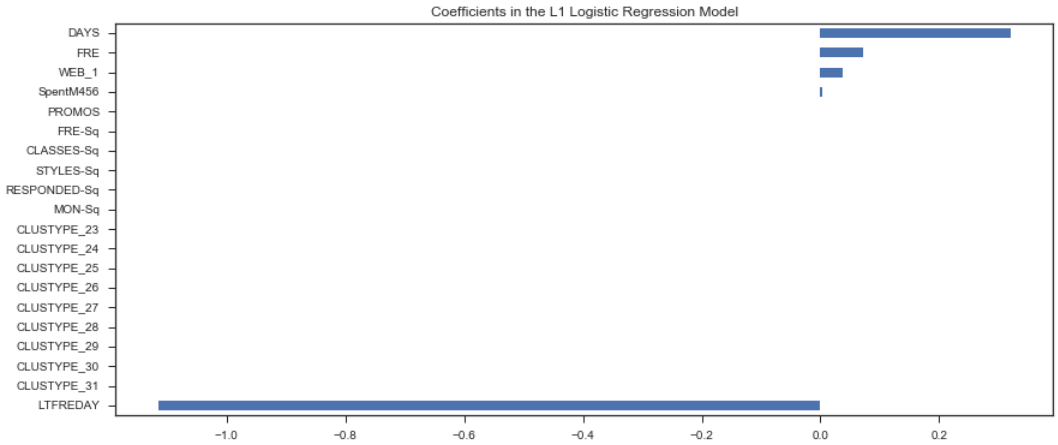

Figure 14: L1 Logistic Coefficients

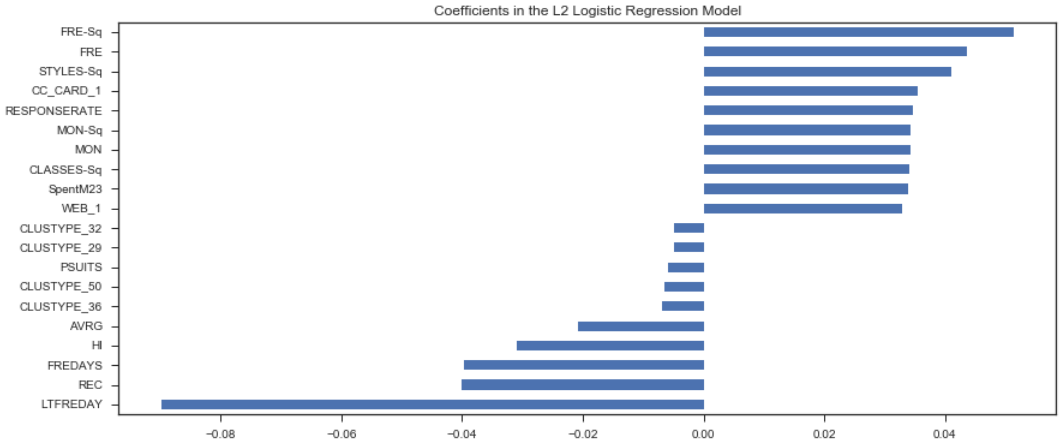

Figure 15: L2 Logistic Coefficients

C Model Hyperparameter Selection

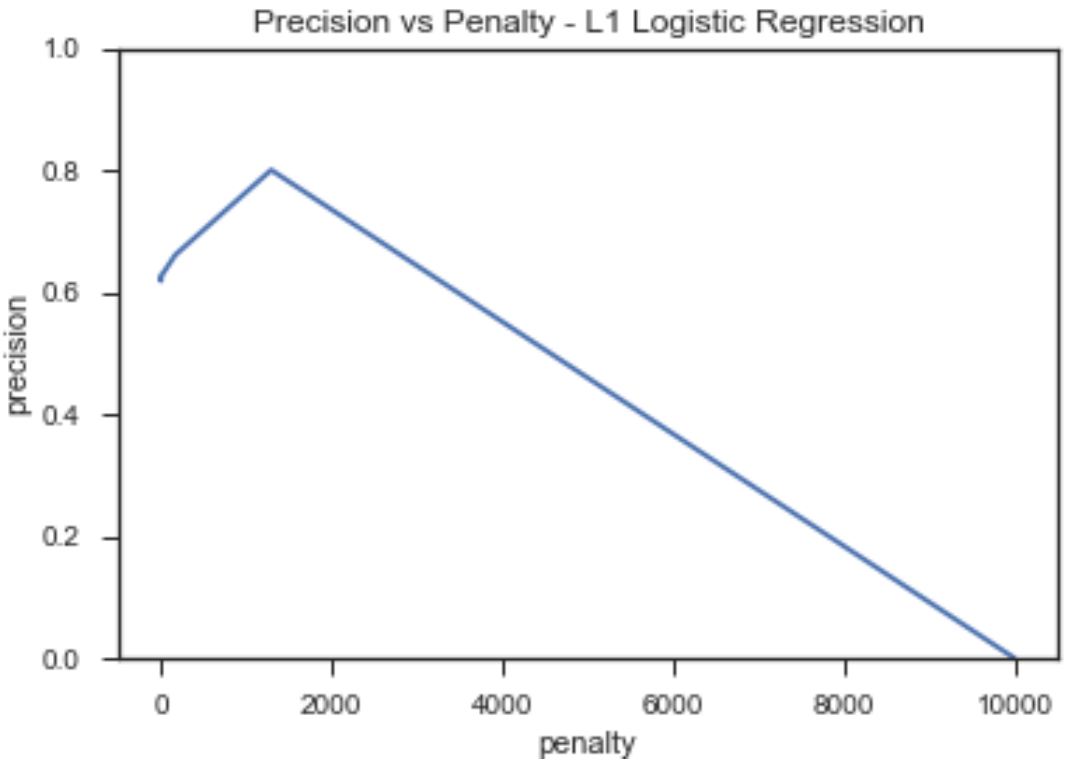

Figure 16: L1 Precision vs Penalty

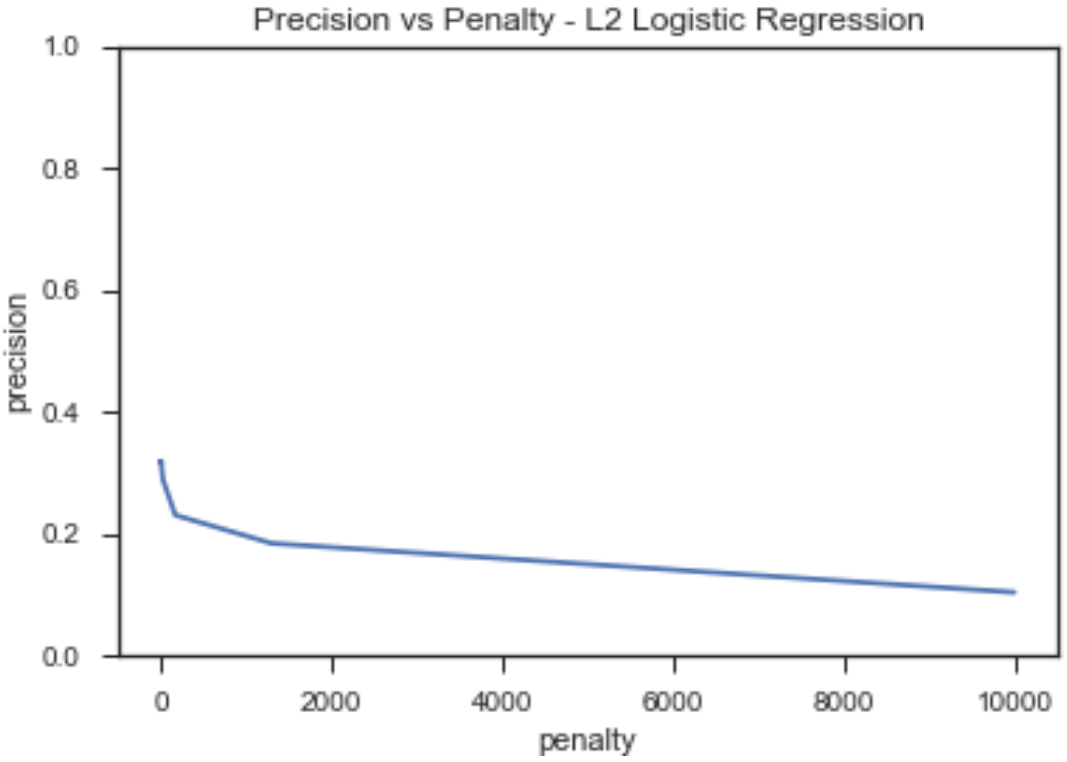

Figure 17: L2 Precision vs Penalty

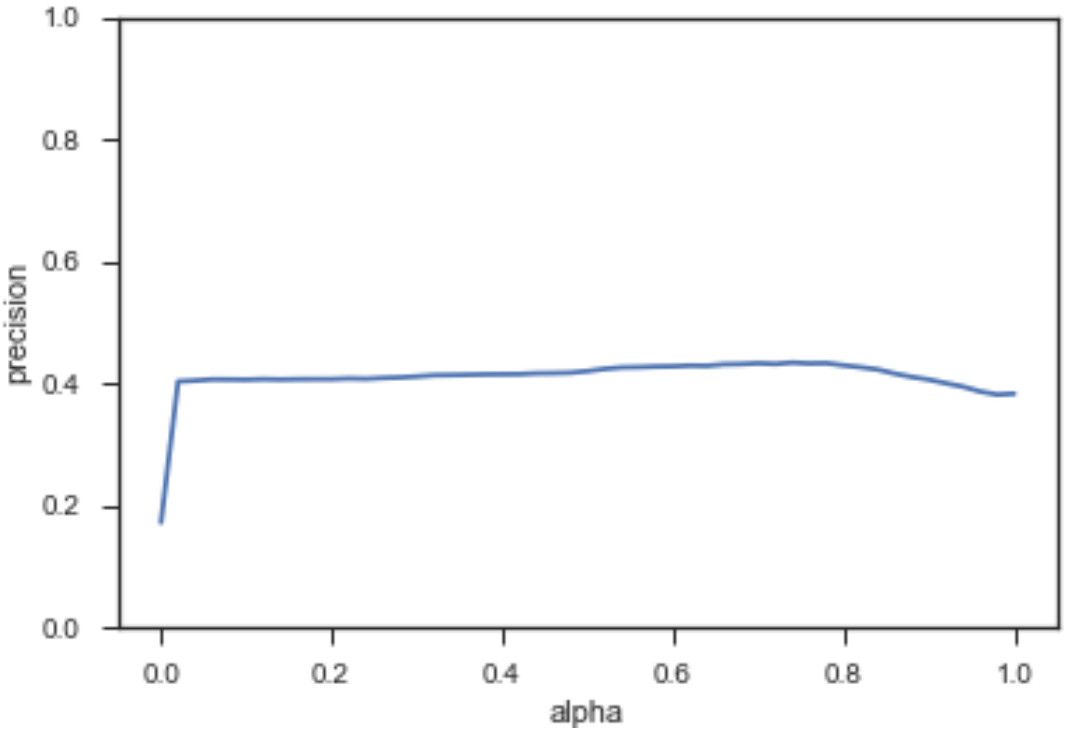

Figure 18: QDA Regularisation Precision vs Penalty

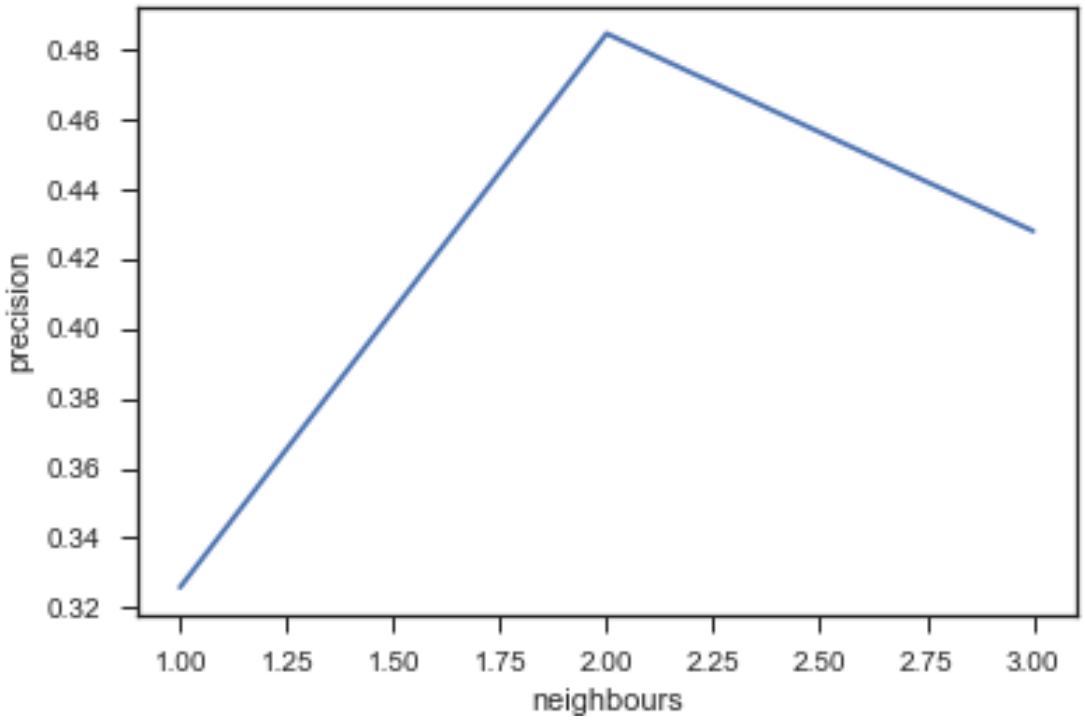

Figure 19: QDA Regularisation Precision vs Number of Neighbours

D Confidence Interval Expected Profit

The various steps involved in the calculation of the confidence intervals are as follows.

First, find the z-scores in a normal distribution table. For example, for a 95% confidence interval:

$$\alpha = (1 - 0.95)/2 = 0.025 \tag{9}$$

Hence, the value to look up in the table is $1 - 0.025 = 0.975$. This gives a z-score of 1.906.

The confidence interval for the precision is given by:

$$\hat{p} = z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \tag{10}$$

From the definition of precision:

$$\hat{p} = \frac{TP}{TP + FP} \tag{11}$$

We arrive at the following equation:

$$\frac{TP}{FP} = \frac{-\hat{p}}{\hat{p} - 1} \tag{12}$$

Hence, using the confidence interval for precision and equation 12 a confidence interval can be found for $\frac{TP}{FP}$.

Now, taking the definition of the expected profit from the direct mail marketing campaign from the confusion and loss matricies:

$$ExpectedProfit = TN \times L_{TN} + FP \times L_{FP} + FN \times L_{FN} + TP \times L_{TP} \tag{13}$$

Rearranging this equation it is able to obtain the following expression:

$$ExpectedProfit = FP \times (L_{FP} + \frac{FN}{FP} \times L_{FN} + \frac{TP}{FP} \times L_{TP}) \tag{14}$$

Where $\frac{TP}{FP}$ is a confidence interval. Hence, a confidence interval can be obtained for the expected profit of that model.

The results for this analysis are shown in tables 10 and 11.

Table 10: QDA Expected Profit Confidence Interval

	90	**95**	**99**
z score	1.645	1.96	2.576
Precision	0.172	0.172	0.172
Precision LI	0.162537	0.160725	0.157182
Precision UI	0.181463	0.183275	0.186818
TP/FP Lower	0.194083	0.191505	0.186496
TP/FP Upper	0.221691	0.224402	0.229737
Profit Lower	-2345.12	-2177.66	-1852.26
Profit Upper	-4138.47	-4314.54	-4661.1

Table 11: Logistic Regression Expected Profit Confidence Interval

	90	**95**	**99**
z score	1.645	1.96	2.576
Precision	0.604	0.604	0.604
Precision LI	0.591737	0.589389	0.584797
Precision UI	0.616263	0.618611	0.623203
TP/FP Lower	1.449402	1.435394	1.408459
TP/FP Upper	1.605951	1.621996	1.653951
Profit Lower	7189.591	7229.767	7307.02
Profit Upper	6740.591	6694.573	6602.923

E Checking Assumptions QDA

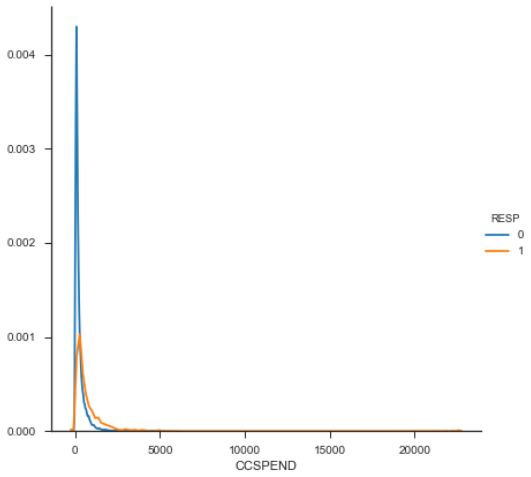

Figure 20: KDE Plot for 'CCSPEND' Before Transformation

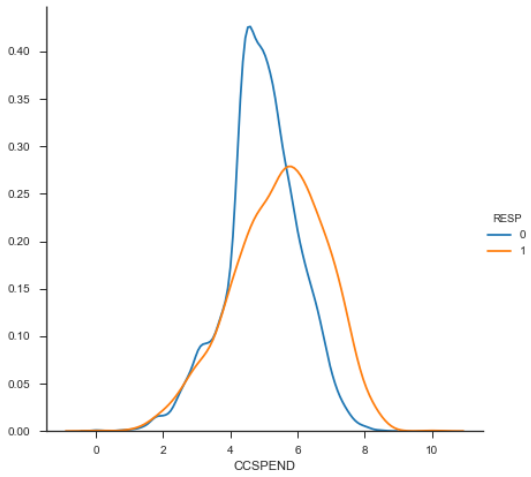

Figure 21: KDE Plot for 'CCSPEND' After Transformation

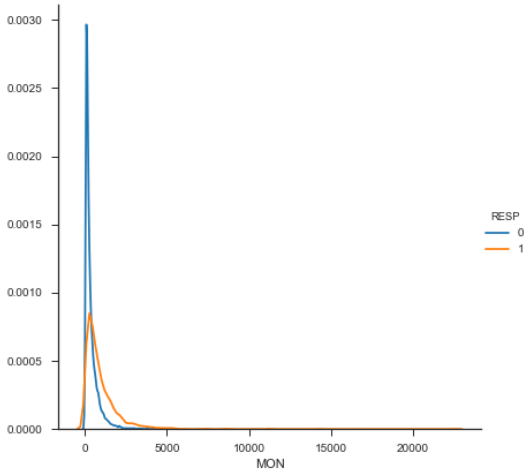

Figure 22: KDE Plot for 'MON' Before Transformation

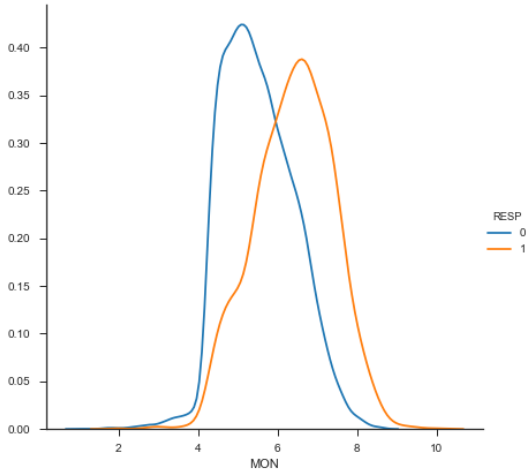

Figure 23: KDE Plot for 'MON' After Transformation

F Model Evaluation Results

Table 12: Model Evaluation Results Different Metrics

	Error rate	**SE**	**Sensitivity**	**Specificity**	**AUC**	**Precision**
Logistic	0.145	0.005	0.316	0.96	0.847	0.604
L1 regularised	0.142	0.005	0.264	0.973	0.842	0.658
L2 regularised	0.149	0.005	0.149	0.987	0.791	0.693
LDA	0.149	0.005	0.341	0.95	0.842	0.568
QDA	0.767	0.006	0.974	0.089	0.66	0.172
Regularised QDA	0.221	0.006	0.583	0.817	0.782	0.382

G ROC Curve

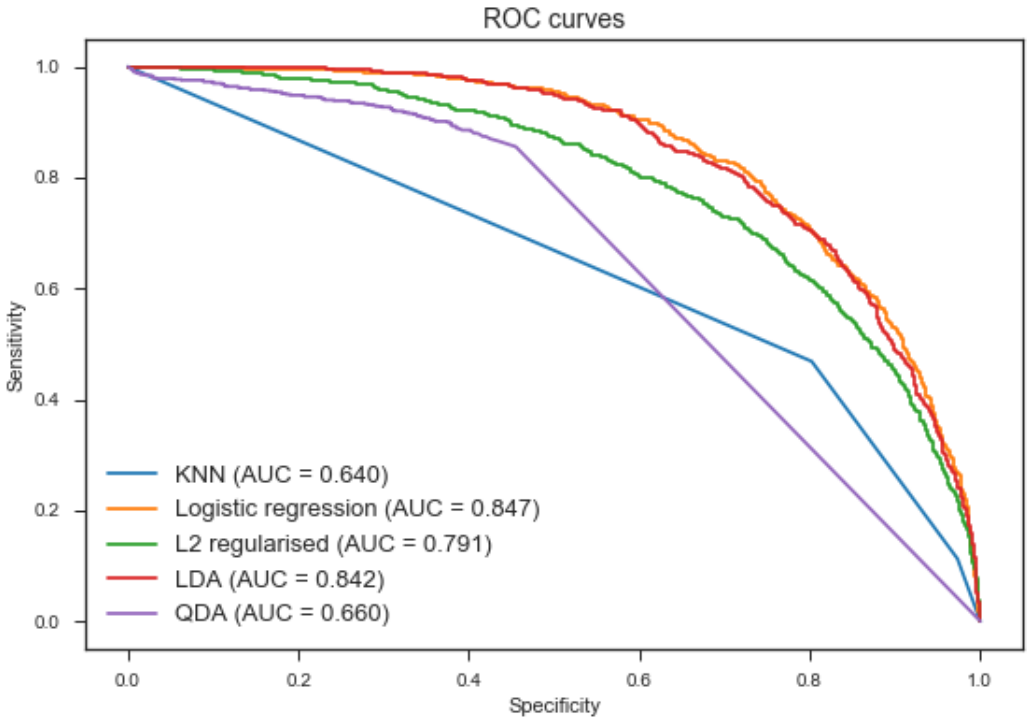

Figure 24: ROC Curve for Various Models