Business Understanding

Determine business objectives

Background

In a saturated apparel market, it is imperative for businesses to employ various tactics in order to increase sales and grow their customer base, effectively. The clothing store chain, Cotton On, recognises this as a core element to their success and thus has exhausted an immense amount of time and effort into trialling different marketing strategies to promote sales. Following from this, Cotton On requires analysis of the effectiveness of each marketing strategy for the corresponding demographic group.

The ultimate objective for Cotton On is to increase the profitability of its overall operations. To do so, the chain needs to ensure that only cost-efficient strategies are implemented. Hence, it is imperative to build a cost-benefit table to detail the impact of using direct mail marketing.

Outcome	**Classification**	**Actual response**	**Cost**
True negative	Nonresponse	Nonresponse	
True positive	Response	Response	
False negative	Nonresponse	Response	
False positive	Response	Nonresponse	

As this project is a relatively new area for Cotton On, an organisational structure and a key individual to oversee the operations is lacking. This presents an opportunity for the clothing store to procure an in-house analyst team that will have the responsibility of being the steering committee. It would further benefit from commissioning a data analyst. Since this project is still in the conceptualisation stage, there is a lack of hierarchy and participants.

Currently, Cotton On has been able to establish large databases for customer and product information as it is a well-known clothing store chain. However, there are no procedures in place to clean and prepare the information collected for analysis. Thereby increasing the unreliableness of the data. Hence, a classification system can facilitate this problem whereby a customer is classified into different categories upon encountering and “purchasing” a product from the store. More specifically, if given the attributes of a customer observed through their purchasing behaviour (*e.g. type of items purchased*), it is possible to the classify their reaction (*e.g. response or nonresponse*) for specific marketing strategies.

As discussed previously, this project is still in its initial stages, therefore there is no evidence of solutions or alternatives to be implemented in order to analyse their customer base. There are obvious benefits to the success of this project such as more effective targeting of different consumer groups dependent on tailored marketing strategies, whilst maintaining, or reducing, costs. This will meet the objective of increasing profitability, and thus will be welcomed.

Defining business objectives

The project is commissioned with the following objectives:

- For all customers, the program is able to classify based on their characteristics what their response to direct mail marketing will be
- Identify whether customers require an additional form of marketing strategy to respond
- Promote sales for Cotton On through tailored marketing strategies
- Ultimately, reduce operational costs and increase profitability

Business success criteria

Tentatively, the project will be judged on its success if:

- Accurately identify customers’ responses to direct mail marketing with an error margin of less than 10%
- Minimise costs associated with Nonresponse customers

Assessing the situation

Inventory of resources

Despite the presence of an existing in-house Database Management System that takes the form of operational databases, there are no database applications or hardware accessible to support the interaction between the data collected and Cotton On's marketing department. Moreover, a data warehouse has yet to be built to facilitate efficient and quality analysis, therefore it still requires data extraction and cleansing.

Personnel

Evidently, there is internal expertise in gathering and simple processing of this data for simple marketing purposes. However, more sophisticated analysis of these relationships for various marketing and sales aspects is lacking. Employing a data analyst that is also proficient in database management would be cost effective.

Data

Availability of large data will facilitate better insights into consumer behaviours, allowing for improvements to marketing strategies to promote sales. Because the project is in its introductory stages, it will be beneficial to limit this study to consumer responses for direct mail marketing. The project can be expanded later.

Risk

The monetary outlays for the consultants and marketing expenses are the greatest concerns for this project since the ultimate goal is to increase profitability. Therefore, it is imperative to stay within budget. Aside from this, there is nothing noteworthy.

Requirements, assumptions and constraints

Requirements

There are no legal and security requirements in regard to this project's results since information is collected anonymously and from basic marketing evaluation metrics that do not reveal consumers' identities. As this is a new project, all stakeholders involved are willing and have approved of the project scheduling requirements. Results shall be implemented into Cotton On's established databases and the created software to improve the classification of customers for enhanced marketing strategies.

Assumptions

Economic: The most prominent factor would be that Cotton On operates in a highly saturated retail market, denoting that it is subject to competitors' ability to attract and retain customers.

Data quality assumption: This becomes a threat because this will compromise the quality of the data collected. There is the possibility of systematic error occurring in the data collection stage. One of them being sample-selection error because Cotton On does not operate in a monopoly whereby its customers represent the whole population. But rather it only captures a segment of the market and thus this 'sample' might not accurately detail how consumers respond. This is followed by data processing error to which incorrect data entries will manipulate the overall insights. However, it is assumed that these threats have not occurred and data collection has been conducted in a consistent manner that ensures these errors have not occurred. Additionally, data utilised is an accurate representation of the population.

Upon completion, final results must be presented to the key stakeholders in a business report. It needs to include model evaluation based on at least two substantively different models for classifying Cotton On's customers. Along with including the project range (confidence interval) for the expected gross profit per customer contacted based on each model. The user would also require interpretation of the findings and final remarks on how to proceed with these insights.

Constraints

As discussed previously, there are no legal constraints of accessibility issues on the usage of this data since it is generic information that do not reveal consumers' identities. All funding will be covered by Cotton On.

Risks and contingencies

This outlines the possible risks that may present itself over the course of the project and contingency plans:

Risk	Impact & Plan

Scheduling	There are no issues with scheduling as there is no predetermined deadlines and this would be an ongoing project for Cotton On.
Financial	Again, no immediate issues present itself as the chain would be funding this project (assuming it remains profitable).
Data	There are existing procedures in place to prevent the likelihood of poor quality data. However, if this was an issue, Cotton On needs to be prepared, in regard to time and the financial costs, to gather more data that is relevant to the project.
Results	Even if the initial results are less dramatic than expected, this does not compromise the project as it signifies classification of customers is possible and fulfils the project objectives. It now provides scope for improvements on the classification algorithms and further iterations on other marketing strategies to be achieved.

This discussion reveals there are no obvious risks that require immediate attention, and necessary contingency plans have already been outlined. If Cotton On no longer finds the project feasible, it shall be halted until the steering committee can resample their customer base and re-evaluate.

Terminology

There are no particular terminology required for this project.

Cost/benefit analysis

Benefits	**Costs**
The classification algorithm shall reduce costs related to the execution of marketing strategies for Nonresponse customers	There are no costs required for data collection since this is already completed and no external databases are utilised.
More effective sales growth can be achieved by identifying which customers respond to which marketing strategy.	There will be minimal costs for result deployment as it can simply be installed into the hardware component
The advancement of knowledge regarding customers is highly valuable due to greater understanding of the behaviours of different customer groups.	There will be operational costs associated with the creation of the hardware for users to interact with the data

Determining data mining goals

Data mining goals

- Type of data mining problem: This is a Classification problem. We need to classify customers into their respective groups based on certain attributes
- Predictive capabilities: Predict and classify customers based on their response outcomes to specific marketing strategies following customer transactions
- Desired outcomes: Generating response predictions for all new customers

Data mining success criteria

The methods utilised for evaluation of these models will include:

- Confusion matrix
- Cost Matrix
- Metrics such as: Precision, Recall, and the F-measure
- Receiver Operating Characteristic Area Under Curve

Benchmarks for evaluating success are:

- 90% accuracy in identifying and classifying Cotton On customers
- Precision and Recall being each respectively at least above 80%
- ROC Area Under Curve being greater than at least 80%

Successful deployment of the model results is crucial to the success of the project.

Producing a product plan

A product plan is constructed:

Phase	Time	Resources	Risks
Business understanding	3 days	All analysts	Alignment of goals and objectives
Data exploration	1 week	All analysts	Data problems
Modelling & evaluation	2 weeks	Data mining consultant, business analyst	Data problems. Constructing feasible models
Deployment & presentation of results	1 week	Business analyst	Inability to implement results

Assessing tools and techniques

This project can be structured as a binary classification problem. Therefore, there are a plethora of potential tools to use:

- KNN
- Logistic Regression
- Naïve Bayes
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Decision Trees
- Random Forests
- Extremely Random Forests
- Adaptive Boosting
- Gradient Boosting
- Ensemble voting

All these tools will be carried out on the training data and then based on their cross-validation scores, we then select a final model.

Data Understanding

Data description report

Data has been collected from multiple existing sources including:

- **Purchase data:** Significant levels of information has been collected attaining to the purchase behaviours of customers over different time periods and store locations. It also specifies Cotton On's marketing response outcomes.
- **Product and marketing database:** Detailed description on different product attributes has been provided, along with outcomes of previous marketing efforts.
- **Customer database:** This database collects more insightful information pertaining to typical consumption behaviours and basic demographic attributes of Cotton On's customers.

At this moment, no other data will be required for the project as there is a sufficient amount (*21,740 observations*) available to draw valuable insights from, and assist in making accurate predictions. It will beneficial for analysis to class the product and customer database in accordance to the purchase data, to better classify the customer groups.

The dataset is presented in a tabular format, consisting of 50 variables and a response variable:

HHKEY	***ZIP_CODE***	***REC***	***FRE***	***MON***	***CC_CARD***	***RESP***
9955600066402	1001	208	2	368.46	0	0
9955600073501	1028	6	4	258	1	1

9955600076313	1056	327	2	77	0	0
9955600078045	1118	66	8	846.06	1	0
9955600078517	1107	49	1	87.44	0	0

Fig. 1 Snapshot of dataset including first 5 observations, 6 variables, and response variable

Appendix 1.1 provides a detailed explanation of each variable and their respective type. The binary variable of VALPHON will be converted into a numerical binary format (0 for N and 1 for Y) for ease of data analysis.

Data exploration report

Variables have been grouped into different categories based on their data type for ease of data exploration:

Type	***Variables***
y	RESP
drop	HHKEY, PC_CALC20
categorical	ZIP_CODE, CLUSTYPE
discrete	REC, FRE, PROMOS, DAYS, CLASSES, COUPONS, STYLES, STORES, MAILED, RESPONDED, STORELOY
binary	CC_CARD, VALPHON_Y, WEB
continuous	MON, AVRG, TMONSPEND, OMONSPEND, SMONSPEND, PREVPD, FREDAYS, LTFREDAY, AMSPEND, PSSPEND, CCSPEND, AXSPEND
fractions	GMP, MARKDOWN, RESPONSERATE, HI, PERCRET, PSWEATERS, PKNIT_TOPS, PKNIT_DRES, PBLOUSES, PJACKETS, PCAR_PNTS, PCAS_PNTS, PSHIRTS, PDRESSES, PSUITS, POUTERWEAR, PJEWELRY, PFASHION, PLEGWEAR, PCOLLSPND

Fig 2. Variable grouping

No missing data was detected in the dataset.

Preliminary exploration into the RESP variable indicates that 83.4% of Cotton On's customers did not respond to their promotional letters. This implies the existence of imbalanced classes. It appears that all variables, aside from zip code, appears to be statistically skewed at a 1% significance level. This is corroborated through following analysis.

Correlation matrix is constructed in order to identify any correlation between the variables.

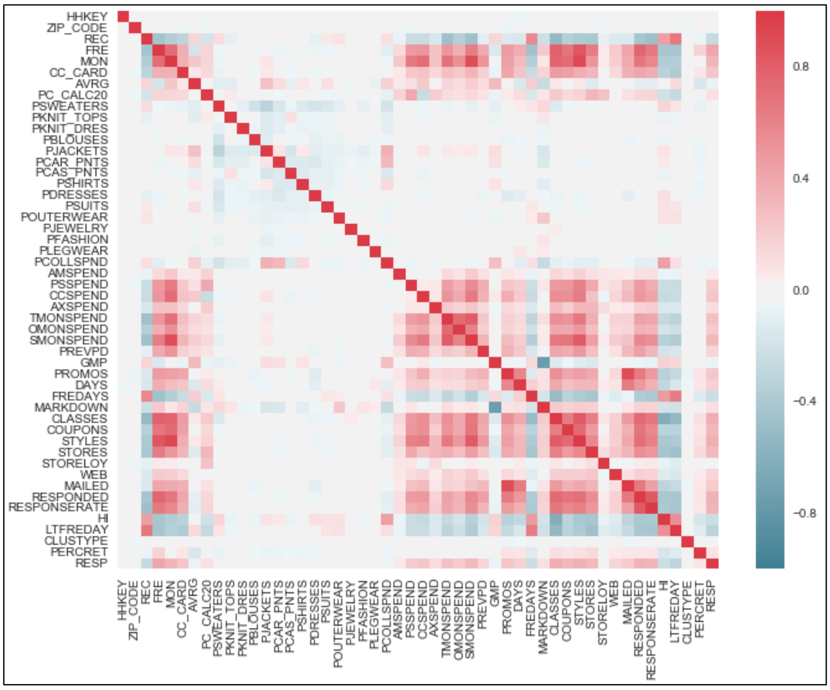


Fig. 3 Correlation matrix of all predictors

Although majority of the variables appear to be uncorrelated, there are a few that are significantly correlated based on a correlation value of 0.75 or above.

- FRE, CLASSES
- FRE, STYLES
- FRE, RESPONDED
- MON, SMONSPEND
- MON, CLASSES
- MON, STYLES
- TMONSPEND, SMONSPEND
- SMONSPEND, STYLES
- GMP, MARKDOWN
- PROMOS, MAILED
- CLASSES, STYLES
- RESPONDED, RESPONSERATE

This suggests that there are possible existences of multicollinearity between the variables. Further analysis of these correlations, through pair plots in Appendix 1.2, illustrates this existence. Majority of the bivariate relationships are positively geared, which has intuitive sense. For example, total net sales (MON) should increase with the number of product classes purchased (CLASSES), and this aligns with the plot. Whilst, markdown percentage on purchases (MARKDOWN) is negatively correlated with gross margin percentage (GMP). Despite the large observations clusters towards the bottom-end of the trend line, the bivariate relationships are still geared towards the trend lines.

Box plots and histograms have been utilised to analyse the distribution of discrete and continuous variables, respectively, as seen in Appendix 1.3 and 1.4. It can be visually extrapolated that the variables are not normally distributed with extreme skewness. This enables scope for feature engineering to occur, including transformation and interaction.

Pivot table for DAYS, FRE, RESPONDED classed into CLUSTYPE groups
Through this cross tabulation, it is able to be extrapolated that all market segmentation categories have been Cotton On's customers for roughly equal durations – 398.218 to 458.207 days on file. Additionally, it can be identified that lower digital literacy customer segments (based on Claritas ConneXions Lifestage Groups) make less frequent purchase visits and are less responsive to promotions. Majority of these lower digital literacy customer groups fall in the Mature Years life-stage, suggesting that this occurrence could possibly be due to their lower disposable incomes. Appendix 1.5 highlights the customer groups that were considered outliers or abnormal to the norm.

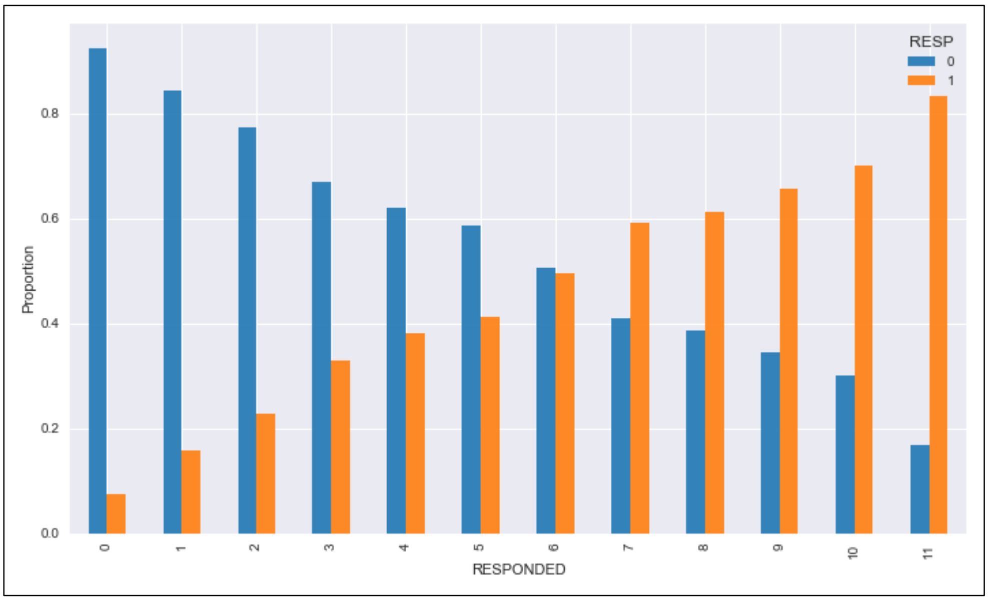


Fig. 4 Bar plot of cross-tabulation between RESPONDED and RESP

This graphical representation of the cross-tabulation between RESPONDED and RESP illustrates the proportion of Y (1) and N (0) responses to the most recent promotional material dependent on the number of promotion responses in the past year. This graph explains that for those customers who responded to 7 or more promotions, they are proportionally more likely to respond Yes to most recent promotional material. Whereas, customers who responded 6 or less times are more likely to respond No. This is intuitive to real-life occurrences. This emphasises Cotton On's need to classify their customer base in order to effectively tailor their marketing efforts based on different customer segments. Appendix 1.6 depicts the entire table of these relationships.

A Kruskal-Wallis ANOVA rank test was conducted to test:

- H_0: all the medians are equal
- H_1: one or more medians are not equal, between response being 0 or 1

	H Stat	*p values*	*Reject, α=0.01*
REC	2202.13	0	TRUE
FRE	4096.08	0	TRUE
PROMOS	1313.31	0	TRUE
DAYS	1600.12	0	TRUE
CLASSES	3077.12	0	TRUE
COUPONS	2395.66	0	TRUE
STYLES	3820.81	0	TRUE
STORES	2042.75	0	TRUE
MAILED	1093.56	0	TRUE
RESPONDED	2698.85	0	TRUE
STORELOY	1570.09	0	TRUE
CC_CARD	1265.24	0	TRUE
VALPHON_Y	236.33	0	TRUE
WEB	555.28	0	TRUE

Fig. 5 Predictors that are statistically significant from ANOVA test

Utilising this non-parametric ensures minimal assumptions are made about the distribution of the categorical predictors, which is necessary due to the skewness of most variables.

The table above presents all of the categories where there is enough statistical evidence, at a significance level of 0.01, to reject the null hypothesis. This specifies that these predictors do not necessarily have equal medians – i.e. it is possible to reject that the distribution of these variables are the same. Therefore, these variables are statistically different when classifying whether RESP is 0 or 1.

	SQRT(VIF)
MON	167.34
CCSPEND	110.65
PSSPEND	88.71
AMSPEND	31.66
AXSPEND	24.49

Fig 6. Variables that have a variance inflation factor (VIF) > 10

To evaluate the extent of the possible presence of multicollinearity in the dataset, VIF is calculated for each predictor. Root VIFs larger than 10 indicates high correlation between predictors which may violate the linear regression assumptions. Here it can be seen that total net sales (MON) is significantly correlated with the amount spent at the four different franchises (CC, PS, AM and AX). Therefore, these four variables should be removed.

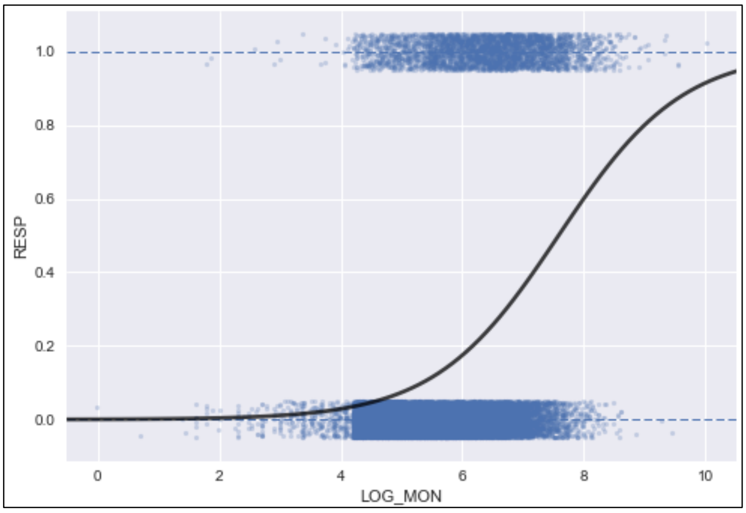


Fig 7. Logistic regression between LOG_MON and RESP

The variable MON has been log transformed for the purpose of this analysis since the variable is significantly skewed. A logistic regression plot has been conducted on LOG_MON and RESP in order to illustrate the relationship between total net sales and whether response is 0 or 1. The graph depicts there is no clear decision boundary between the two responses, but rather observations are clustered in the middle. Moreover, variance is too high and thus this warrants further analysis.

Further exploratory analysis could be considered for the continuous variables bounded in between 0 and 1. These variables take on a percentage or fractional value and therefore warrant further attention. However, the methods and algorithms used in the modelling phase are robust enough to account for these variable types.

Data verification report

It is important to verify the data to ensure quality insights and predictions can be achieved.

- Missing data: there is no evidence of missing data.
- Data errors: data collected are automatically generated therefore there is no concern.
- Measurement errors: since data is collected based of basic transaction metrics, there is no possibility for this type of error to occur.
- Bad metadata: this is concern since there is no official dataset documentation that details the description of each variable. However, the variable names are fairly intuitive and thus can be easily deduced.

Data preparation

Data selection

With credit to the EDA, it was determined that the zip code variable (‘ZIP_CODE’) was redundant due to the number of unique classes (7419). As a result of this, categorically encoding the thousands of classes would be meaningless for any model. Training the models with the zip codes numerically as ordered values would also be inaccurate. In light of this, the zip codes were removed. For future analysis, one could consider grouping the zip codes into bins if more geographical information was available.

Four more variables corresponding to the amount spent at each franchise (‘AMSPEND’, ‘PSSPEND’, ‘CCSPEND’, and ‘AXSPEND’) were removed following calculations of their variance inflation factors. This calculation assessed potential multicollinearity in the variables, and it was determined that these four variables were collinear with total amount spent (‘MON’). In order to not violate any multi-linear regression assumptions (with consideration of logistic regression), these variables were removed.

It is also worth noting that the customer ID (‘HHKEY’) was removed as a predictor variable, alongside ‘PC_CALC20’, which could not be identified or interpreted.

Data cleaning

As noted in the data exploration phase, there were no issues in regard to missing data, data errors, measurement data. Therefore, there is no need to clean the data. Bad metadata has been rectified through research on each individual variables and cross-verification from multiple sources.

Data construction

Transformation of variables

Following thorough exploratory data analysis, it is evident that several predictors are skewed and need addressing in order to train models which rely on assumptions of Gaussian data, such as logistic regression which uses maximum likelihood estimation. A skewness test which tested the null hypothesis that the sample data comes from a normally distributed population returned results indicating that all continuous and discrete variables were skewed. Further analysis could consider a Jarque-Bera test in order to compare kurtosis as well.

With these results, alongside illustrations of skewness with boxplots for discrete variables and histograms for continuous variables, transformations were considered. Both log- and Box-Cox transformations were assessed as can be seen below in the case of the predictor for total amount spent ('MON').

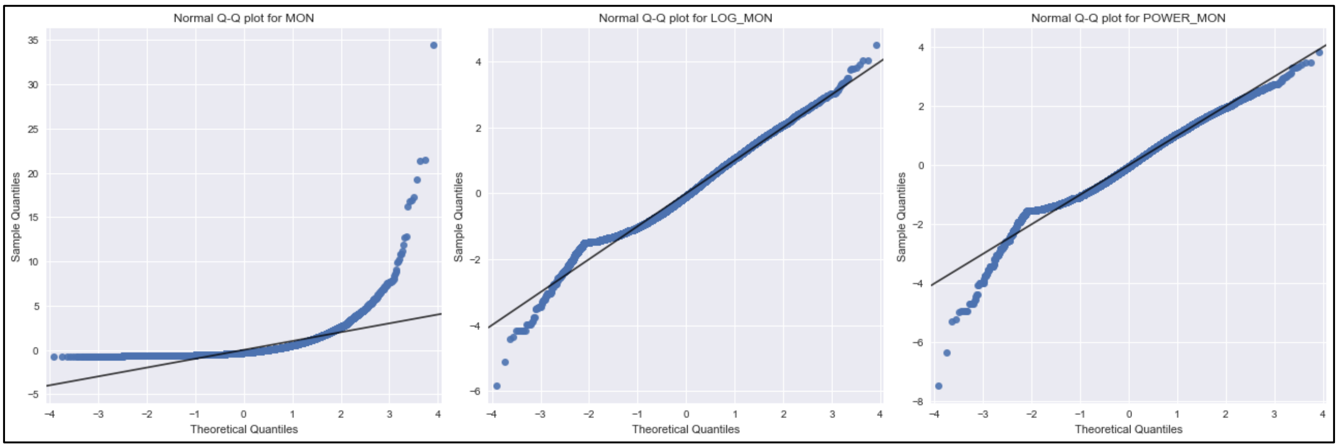


Fig. 8 Normality plots for MON, LOG_MON and POWER_MON

Here, it can be seen that both transformations reduce the skewness in the upper-tail, however struggle to address that in the lower-tail. These results were shared when transforming the other skewed predictors as well.

Encoding categorical variables

The cluster types ('CLUSTYPE') predictor has 51 unique classes, allowing for some flexibility in encoding. One method would be to one-hot encode all the cluster types, at the expense of computational costs. However following EDA, it was clear that several cluster types occurred far more frequently than others. With this information, the top nine cluster types were one-hot encoded, with the remaining all being grouped into a tenth category. Whilst a somewhat arbitrary decision, this improve some of the trained model's predictive capacities (namely K-Nearest Neighbours and Logistic Regression). This encoding could be further tuned if desired.

Also, worth noting is the one-hot encoding of the valid phone number predictor ('VALPHON') which was provided in the original data set with two possible observations, 'Y' and 'N'.

Constructing interactions

Interactions were created following EDA and intuitive domain knowledge assessment in order to consider relationships between binary/discrete and continuous variables within the dataset.

The seven interactions created were:

Interaction	***Variables***	***Description***
Interaction 1	CC_CARD; MON	Credit card owned (binary) : Total spent (continuous)
Interaction 2	WEB; MON	Web shopper (binary) : Total spent (continuous)
Interaction 3	PROMOS; MON	Promotions on file (discrete) : Total spent (continuous

Interaction 4	CLASSES; MON	Product classes purchased (discrete) : Total spent (continuous)
Interaction 5	COUPONS; MON	Coupons used (discrete) : Total spent (continuous)
Interaction 6	RESPONDED; MON	Responses to mail promos (discrete) : Total spent (continuous)
Interaction 7	MAILED; FREDAYS	Promos mailed (discrete) : Days between store visits (continuous)

Fig. 9 Description of interactions created

The rationale for each interaction is as such:

Interaction	*Rationale*
Interaction 1	Owning a credit card may change how much you spend
Interaction 2	Being a web shopper may change spending patterns in a physical store
Interaction 3	Number of promotions you have on file may change how much you spend
Interaction 4	Number of different product classes purchased may change how much you have spent (could highlight more diverse spending pattern too)
Interaction 5	Number of coupons may change spending pattern
Interaction 6	Number of mail promos responded to may change amount spent
Interaction 7	Number of promos mailed to a customer may change number of days between store visits

Fig. 10 Rationale for each interaction created

For further analysis, interactions with the variables for one month ('OMONSPEND'), three month ('TMONSPEND', and six month spending ('SMONSPEND') could be considered too.

Standardising the predictors

In the case that the models being trained apply regularisation and dimension reduction, the predictors in the cleaned and feature engineered dataset were standardised with *Scikit-Learn's* Standard Scaler method. Given the skewed nature of these predictors, and the subsequent likelihood of outliers, robust scaling could be considered too.

Data integration

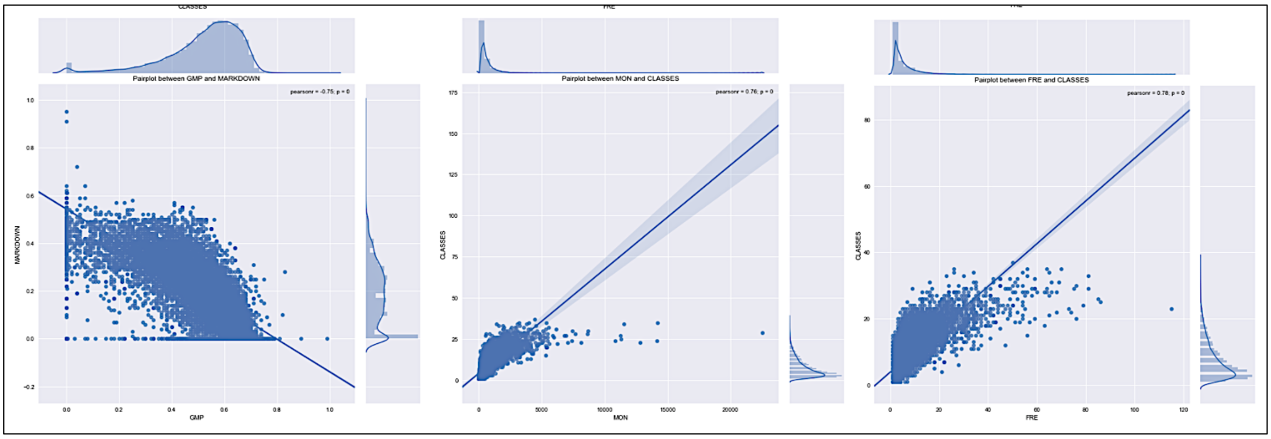
Since data is collected from one source that is internally warehoused, there is no need to obtain further data from other sources. Therefore, no effort is required for data integration.

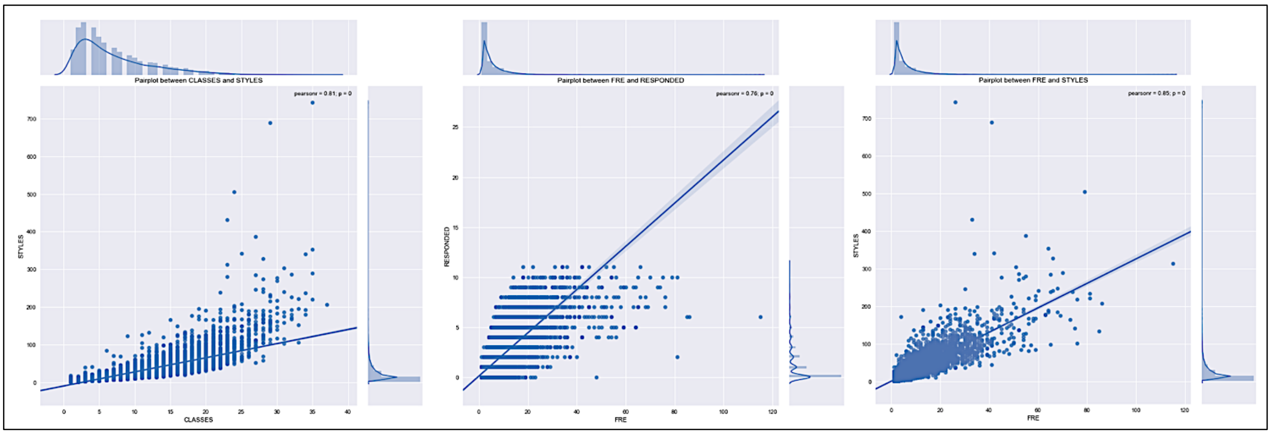
Appendix 1.1

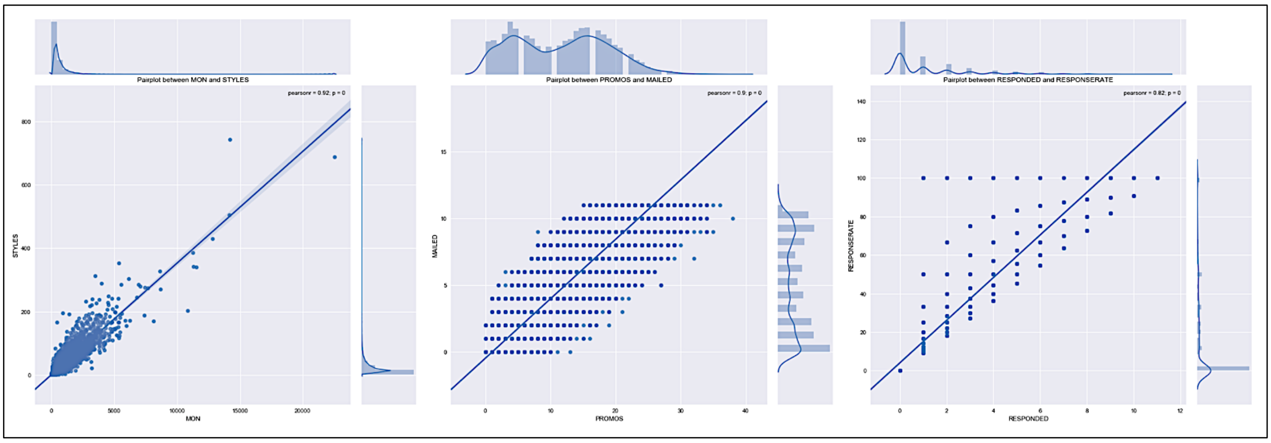
Variable	***Explanation***	***Type***
HHKEY	Customer ID	Numeric
ZIP_CODE	Zip code	Numeric
REC	Frequency of visits	Numeric
FRE	Number of purchase visits	Numeric
MON	Total net sales	Numeric
CC_CARD	Flag: credit card user	Numeric
AVRG	Average amount spent per visit	Numeric
PC_CALC20	*Unknown*	Numeric
PSWEATERS	Percent spent on sweaters	Numeric
PKNIT_TOPS	Percent spent on knit tops	Numeric
PKNIT_DRES	Percent spent on knit dresses	Numeric
PBLOUSES	Percent spent on blouses	Numeric
PJACKETS	Percent spent on jackets	Numeric
PCAR_PNTS	Percent spent on cargo pants	Numeric
PCAS_PNTS	Percent spent on casual pants	Numeric
PSHIRTS	Percent spent on shirts	Numeric
PDRESSES	Percent spent on dresses	Numeric
PSUITS	Precent spent on suit	Numeric
POUTERWEAR	Percent spent on outerwear	Numeric
PJEWELRY	Percent spent on jewellery	Numeric
PFASHION	Percent spent on fashion	Numeric
PLEGWEAR	Percent spent on leg wear	Numeric
PCOLLSPND	Percent spent on collectibles	Numeric
AMSPEND	Total amount spent at AM store	Numeric
PSSPEND	Total amount spent at PS store	Numeric
CCSPEND	Total amount spent at CC store	Numeric
AXSPEND	Total amount spent at AX store	Numeric
TMONSPEND	Total amount spent over the last 3 months	Numeric
OMONSPEND	Total amount spent over the last 1 month	Numeric
SMONSPEND	Total amount spent over the last 6 months	Numeric
PREVPD	Amount spent in the same period last year	Numeric
GMP	Gross margin percentage	Numeric
PROMOS	Number of marketing promotions on file	Numeric
DAYS	Number of days customer has been on file	Numeric
FREDAYS	Number of days between purchases	Numeric
MARKDOWN	Markdown percentage on customer purchases	Numeric
CLASSES	Number of different product classes purchased	Numeric
COUPONS	Number of coupons used by the customer	Numeric
STYLES	Number of different styles purchased	Numeric
STORES	Number of stores customer has shopped in	Numeric
STORELOY	*Unknown*	Numeric
VALPHON	Flag: valid phone number on file	Boolean
WEB	Flag: web shopper	Numeric
MAILED	Number of promotions mailed in the last year	Numeric
RESPONDED	Number of promotions responded to in the past year	Numeric
RESPONSERATE	Promotion response rate per year	Numeric
HI	Product uniformity	Numeric
LTFREDAY	Lifetime average time between purchases	Numeric

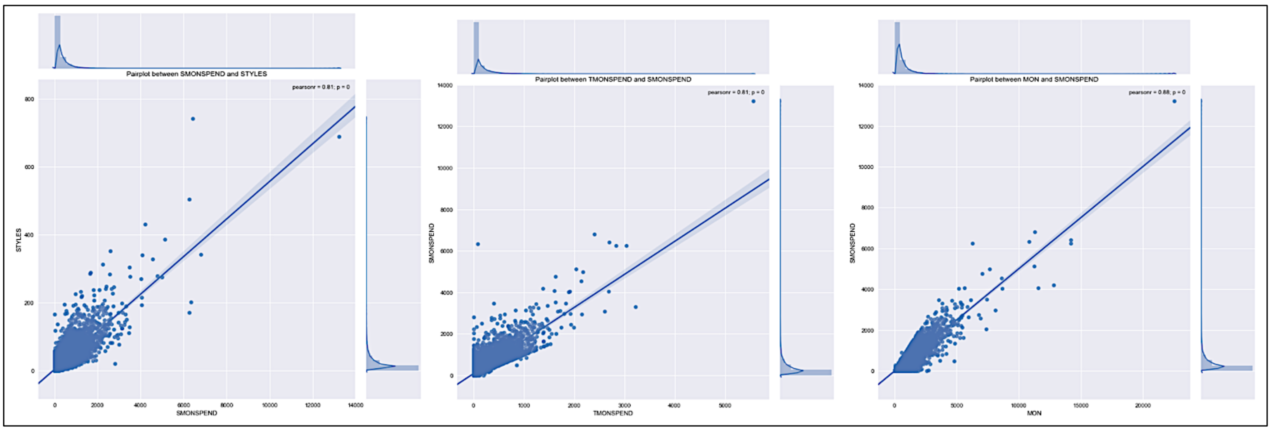
CLUSTYPE	Market segmentation categories	Numeric
PERCRET	Precent of returns	Numeric
RESP	Response to a promotion letter	Numeric

Appendix 1.2

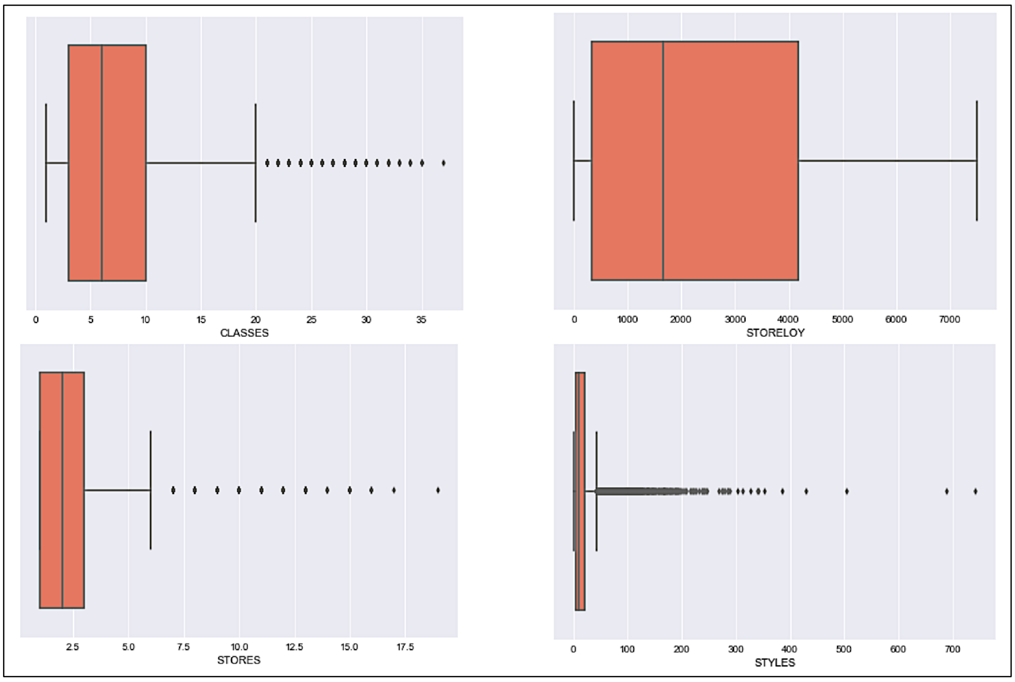


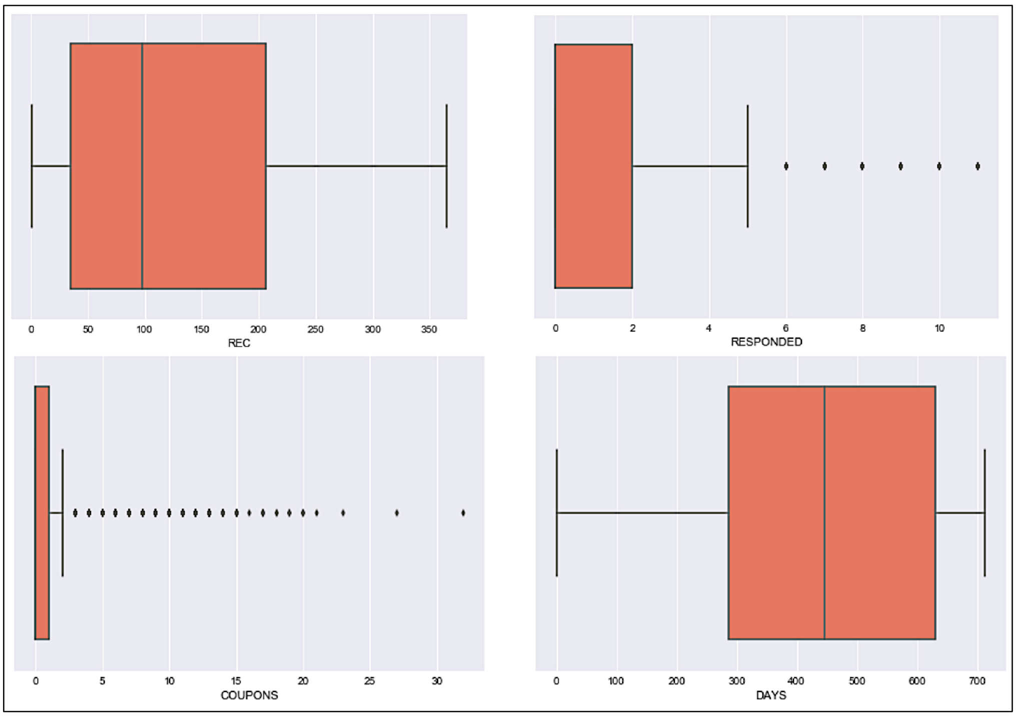


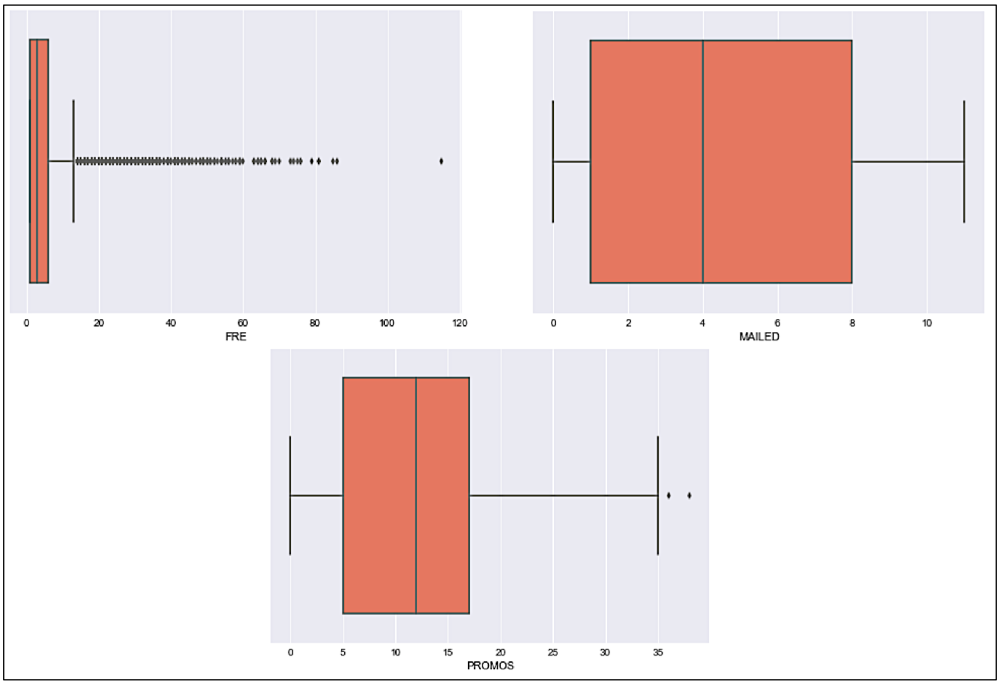




Appendix 1.3


0
5
10
15
20
25
30
35
CLASSES
0
1000
2000
3000
4000
5000
6000
7000
STORELOY
2.5
5.0
7.5
10.0
12.5
15.0
17.5
STORES
0
100
200
300
400
500
600
700
STYLES


0
50
100
150
200
250
300
350
REC
0
2
4
6
8
10
RESPONDED
0
5
10
15
20
25
30
COUPONS
0
100
200
300
400
500
600
700
DAYS



Appendix 1.4

HISTOGRAMS

Appendix 1.5

DAYS	Below lower limit	Above upper limit	
CLUSTYPE	**17**	**31**	**33**
DAYS	389.26	473.15	485.333
FRE	4.745	4.3	4.333
RESPONDED	1	1.075	1.667

FRE	Below lower limit					
CLUSTYPE	**44**	**47**	**49**	**43**	**29**	
DAYS	426	410.24	400.25	404.419	405.859	
FRE	1.25	2.52	2.812	3.71	3.766	
RESPONDED	0	0.52	0.375	0.677	0.781	
	Below lower limit					
CLUSTYPE	**9**	**48**	**14**	**34**	**42**	**26**
DAYS	458.207	421.32	420.468	408.382	402.405	433.615
FRE	3.862	3.92	4.064	4.091	4.091	4.154
RESPONDED	1.103	0.96	0.737	1	0.95	0.846
	Above upper limit					
CLUSTYPE	**27**	**2**	**39**	**36**		
DAYS	432.429	451.59	447.908	432.548		
FRE	5.857	5.888	5.956	6.161		
RESPONDED	1.095	1.478	1.48	1.161		

RESPONDED	Below lower limit			Above upper limit
CLUSTYPE	**44**	**49**	**47**	**33**
DAYS	426	400.25	410.24	485.333
FRE	1.25	2.812	2.52	4.333

RESPONDED	0	0.375	0.52	1.667

Appendix 1.6

RESPONDED	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***
RESP												
0	10816	3300	1754	925	560	373	204	96	58	29	12	2
1	894	613	518	455	345	262	200	139	92	55	28	10