Part1 Business understanding

This is a dataset from a clothing store chain. It requires to build a model to predict whether a customer will respond to the direct mail marketing. The final mailing decision will be based on prediction model. For those customers who are projected to respond to this market strategy, direct mailing is sent. For those who are not, no mail will be sent. This project helps the company have a deep understanding of its customers, to use advertising budget more efficiently and finally achieve profit maximization.

From our previous inference, potential lost profit is likely to be higher than the useless advertisement cost, then we use the following terminology.

The cost-benefit table is built as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification** | | | | |
| **Nonresponse** | | | | **Response** |
| **Actual** | **Nonresponse** | Advertisement cost saved | Useless advertisement cost | |
| **Response** | Potential lost profit | Profit | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification** | | | | |
|  | | | |  |
| **Actual** |  | True negative | False Positive | |
|  | False negative | True Positive | |

Detailed explanation and specific costs for the four outcomes are as follows:

**TN (True Negative).** This is a correct prediction. The classification model predicts that customer will not respond to direct mailing, so no promotion booklets or coupons are sent. And since in reality, this type of customers will not respond, the advertisement cost is saved. No loss is under this outcome.

**FN (False Negative).** This is an incorrect prediction. The model predict that customer will not respond to direct mailing and no mail has been sent. However, this kind of customers will respond in reality, which means potential profit is lost.

From our existing data, the revenue per visit per person could be calculated by obtaining the mean of the variable “Average amount spent per visit”, which is $113.89. The profit margin for clothing retailer industry is 40% **(Reference),** which indication the lost profit is $45.56.

**FP (False Positive).** This is an incorrect prediction. The model predict that the customer will react to mailing and the advertisement cost has been incurred. In reality, this type of customer will not react to the marketing strategy. The total loss is advertisement cost mainly including postage and printing cost. Refer to the Australia Post Website, for direct mail marketing, it provides different types of charge. Since we know little information about this Clothes Store, so to be conservative, we choose the expensive type which is $1.550. Also, we assume other cost (such as printing fee) is $0.450. In total, advertisement cost per person is $2. Then the loss for this outcome is $2.

**TP (True Positive).** This is a correct prediction. The model predicts the customer will react to advertisement and sent a promotion mail. In reality, after receiving the mail, the customer will contribute to the store’s sales. This creates a profit of $45.56-$2=$43.56.

In conclusion, the costs corresponding to each outcome are

|  |  |  |  |
| --- | --- | --- | --- |
| **Outcome** | **Classification** | **Actual Response** | **Cost** |
| True Negative | Nonresponse | Nonresponse | 0 |
| True positive | Response | Response | -$43.56 |
| False negative | Nonresponse | Response | $45.56 |
| False positive | Response | Nonresponse | $2 |

Part2 Data understanding

This is a dataset containing 21740 customer records with 50 independent variables and 1 dependent variable. In the 50 independent variables, it contains 44 numerical variables and 6 ordinal variables: HHKEY, CC\_CARD, VALPHON and WEB, CLUSTERTYPE. Among them, we delete HHKEY and ZIP\_CODE. HHKEY stands for Customer ID which is a random number, it might be based on the customer acquisition time but have no use for prediction, so we delete it. For ZIP\_CODE, it is difficult for interpretation and might be irrelevant to our prediction.

Among them, CC\_CARD and WEB have already been dealt in 1-0 binary format in our dataset so we only need get dummy variable for VALPHON.

After checking, we find no missing value in both independent and dependent variables, then no efforts need to be made to deal with missing values.

Firstly, we check the full dataset and find among all 21470 records, 16.61% customers respond to our direct mailing promotion.

Since our aim is to understand the behaviors of respondent customers and predict response, which target at help build better customer relationship management and achieve profit maximization. Different from pure statistical model, we need to look deeper and further in our variables.

**Categorical variables**

We have 4 four categorical variables (CLUSTYPE, VALPHON, CC\_CARD and WEB), which can intuitively describes the basic characters and preference of these respondent customers.

**CLUSTYPE**

Firstly, we focus on CLUSTTYPE, which refer to Microvision lifestyle cluster type (market segmentation category defined by Claritas demographics), to give a general description of these customers. The top 5 customers are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **lifestyle** | **Type Name** | **probability** | **Brief introduction** |
| 10 | Home Sweet Home | 12.07 | married couples with one or no children, above average household income, own their home, concentrated in the suburbs |
| 1 | Upper Crust | 9.45 | families with older children located in the suburbs, very high levels of income and education and work in executive, managerial and other professional occupations |
| 4 | Mid-Life Success | 7.91 | households with very high incomes living in suburban areas, homeowners with very high property values, primarily working in white-collar occupations such as sales. |
| 16 | Great Beginnings | 6.58 | married families with children, located in rural areas, a household income very near the national average, own their home and work in blue-collar occupations |
| 8 | Movers and Shakers | 5.03 | households containing singles and couples, with two workers and no children, live in the suburbs and some urban areas and have high levels of education and income. |

Generally, our common customers are to some extent focus on high level income households. Especially for Upper Crust, which rank 2rd in our customer segments, has the highest income of all the segments with a median of three times the national average.

For other three categorical variables, we draw the cross table.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| VALPHON | 0 | 1 |  | CC\_CARD | 0 | 1 |  | WEB | 0 | 1 |
| RESP |  |  |  | RESP |  |  |  | RESP |  |  |
| 0 | 0.932 | 0.818 |  | 0 | 0 | 1 |  | 0 | 0.847 | 0.546 |
| 1 | 0.068 | 0.182 |  | 1 | 0.905 | 0.721 |  | 1 | 0.153 | 0.454 |

We could find a higher proportion of the customers with valid phone is responded compared to customers without valid phone. Also, credit card user is less likely to respond than non\_credit card user. Web shopper is more likely to respond than non-web shopper.

In conclusion, for categorical variables, we could rationally expect our customers tend to wealth people who has valid phone number on file, do not use credit card and prefer to use web shopping.

**Numerical variables**

In order to concentrate on numerical variables, we first temporarily put categorically variables aside.

**Statistical summary**

We first have a good at statistical summary**(Appendix)** of the remaining 48 variables to get roughly statistical understanding. We pay attention on mean, variance, skewness and kurtosis to have a rough judgement of the unit scale and normality of our variables. It is obvious that all the variables have quite different means and variances. These inconsistent scales can make prediction biased and hence call for data standardization in the following data preparation steps. Normality is

another important character since it is a crucial assumption under quite a lot of predictive models such as LDA and QDA. To observe the normality, binary variables are excluded because of natural disobedience to normal distribution. For numerical variables, besides GMP and DAYS, the skewness of all the remaining variables have positive skewness with highest one up to 37.6707. This implies these variables might violate requirements of normal distributions and tend to be like lognormal. The kurtosis to variables are all different from 3 ranging from -1.78169 to 128.704, which is quite unstable and further confirm the unnormal characters.

*For details, tracking his/her purchasing behavior in our brand, we could find*

*Regarding to the money spent, which might be the most important part to us. It contains money spent on each clothing category (15 in total), money spent on each store (4 in total) and amount spent in the past one month, three months and six months as well as amount spent in the same period last year. Total net sales and Gross profit margin. average amount spent per visit.*

*By looking through all of our 48 variables, we could roughly divide them into 5 types.*

*Categorical variable types:*

*3 flag variables: • Flag: credit card user• Flag: valid phone number on file. • Flag: web shopper.*

*And CLUSTERTYPE which we have already addressed before.*

*Accounting figures: Sales type: Number of purchase visits, Total net sales*

*Specific types of goods*

*Amount spent for each of four different franchises (4 variables)*

*Amount spent in different periods*

*Respond indicators:*

**Normalization**

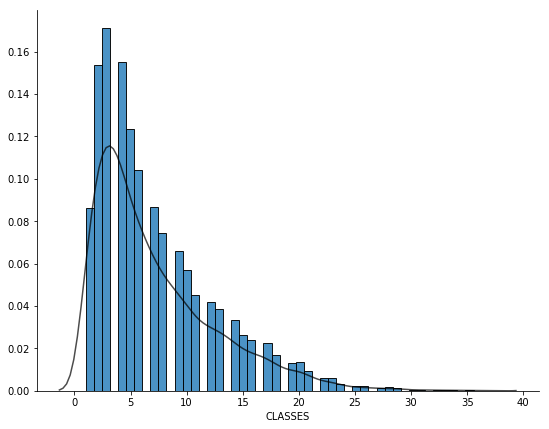
Firstly, we aim to use visualization to detect and solve unnormal situations in our variables.

Since we have a wide range of variables, and we could not illustrate them one by one, so we divided them into several types **according to customer purchasing behavior** and choose typical one in each type for illustration.

There is one type of variable which directly related to customers preference on **purchase product classes** and **purchase location.**

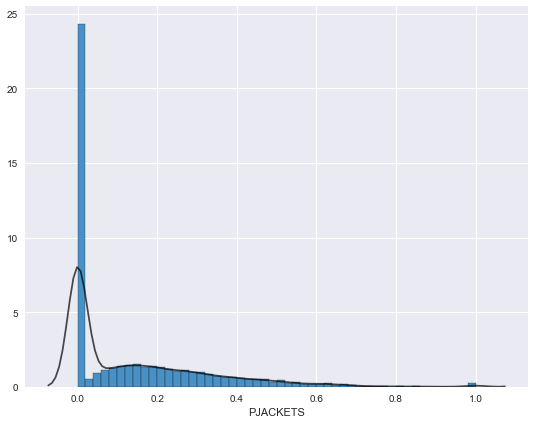
It contains: percentages spent by the customer on specific classes of clothing (15 variables), CLASSES: number of different product classes purchased, amount spent for each of four different franchises (4 variables), STORES: number of stores the customer purchased at.

To discover the characters of this type of variables, recall our right-skewness inference from statistical summary, we plot “CLASSES” as an illustration here. It stands for the number of different product classes purchased for each customer. We should expect that customer with high number of different product classes purchased tend to have a higher response probability since they are more likely to try new products.



We could find most people concentrate on buying 0-10 goods, with this figure decreasing as the number of classes increase.

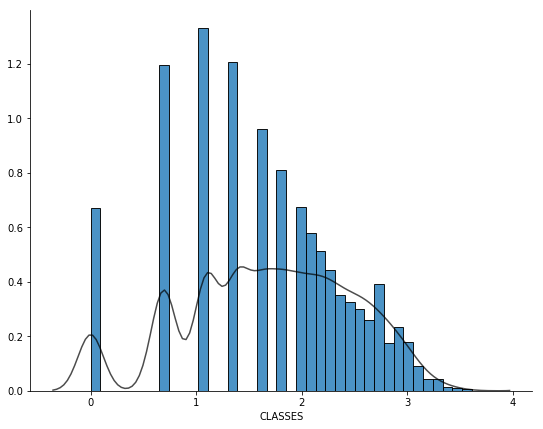
To look detailly on customers’ preference on clothing classes, we plot the histogram of variable “PJACKETS” as an example, which represent percentage spent on Jacket, as one of the 15 variables providing the percentages spent by the customer on specific classes of clothing.



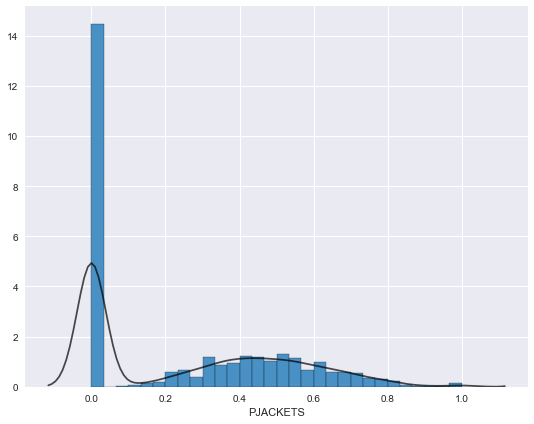
We can detect that almost 25% of customers do not buy jackets at all, while the percentage spent on jackets for the rest customers are right skewed. We could expect other 14 variables for percentage spent on specific goods have similar situations. And checking the dataset confirm our expectation.

Referring to transformation, log function is the first choice. However, corresponding to two conditions we have mentioned before, numerical variables like percentage of amount spent on specific goods contain 0 values, which could be inconvenient for us to apply log transformation. Therefore, for those variables only with positive values, we apply log transformation. For those containing 0 values, we choose square root transformation.

To test whether we have achieved normalization, we plot the distribution of “CLASSES” again. Although it is not perfectly gaussian distributed, it has been improved a lot in skewness and generally meet the assumptions of normality.

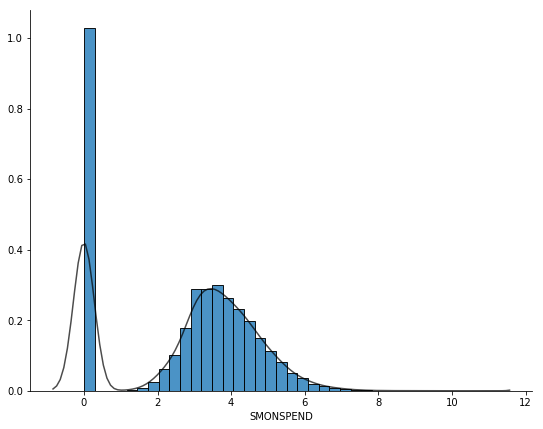


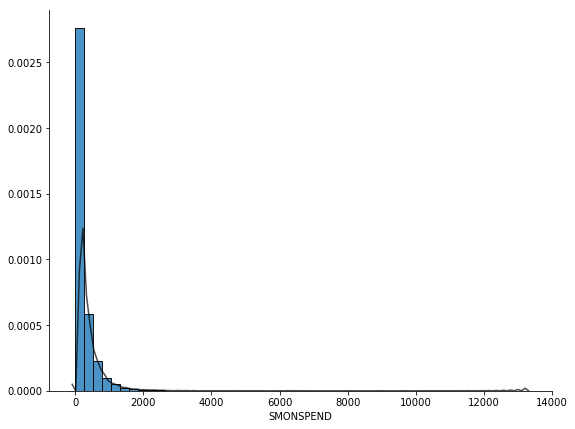
Also, after normalization, for “PJACKETS”, the plot is as follows:



We could find 0 spike still exist while the right part is generally normally distributed. We will address this problem later in our report. **For this condition, we consider to apply flag variable to this type of variables.( HAVE WE BUILT FLAG VARIABLES FOR ALL VARIABLES? IT SHOULD INCLUDE 4 STORES AND 15 CHOTHING TYPE)**

Another type of variables we are addressing is the amount spend according to the time, which includes TMONSPEND: Amount spent in the past 3 months, OMONSPEND: Amount spent in the past month, SMONSPEND: Amount spent in the past 6 months, PREVPD: Amount spent in the same period last year. Here we plot the histogram before and after normalization:



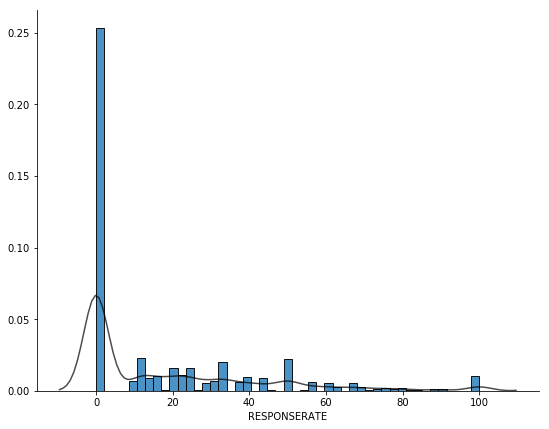
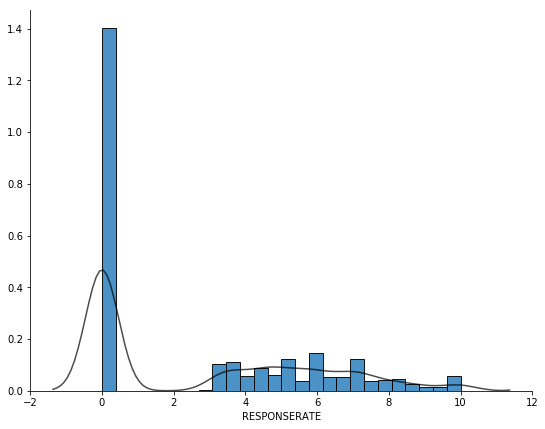


|  |  |  |
| --- | --- | --- |
|  | Mean | Standard deviation |
| SMONSPEND | 1 | 0 |

Then we plot the distribution

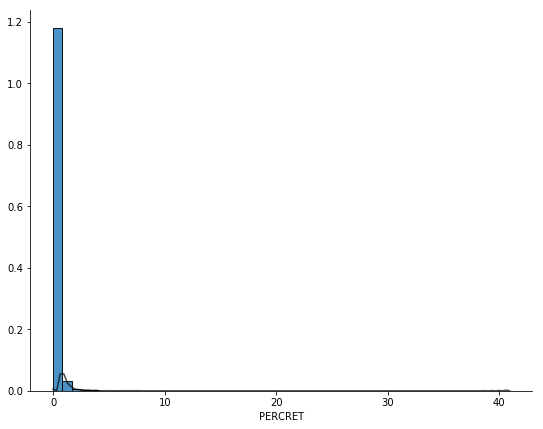
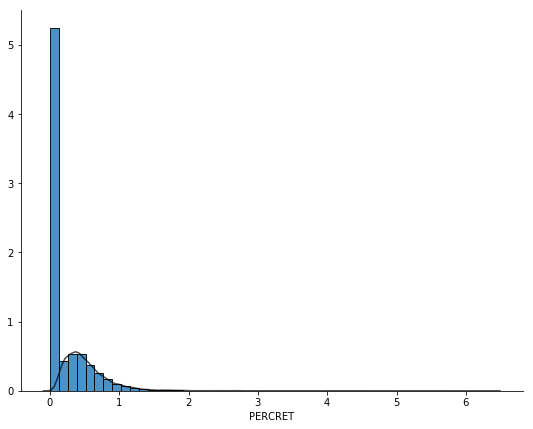
It is similar to what we have discussed before, it generally has a spike and the right part is inclined to be normal distribution. **Which also need to create flag variables for this type of variables.**

Then we move to the third type of variables. This is a group of variables related to marketing promotion. Which generally contain PROMOS: Number of marketing promotions on file, MAILED: Number of promotions mailed in the last year, RESPONDED: Number of promotions responded to in the past year, RESPONSERATE: Promotion response rate for the past year. From our previous analysis experience, we could rationally expect that distribution of RESPNSERATE is a spike at left corner while right part is a general normal distribution. By plotting the distributions before normalization and after normalization,



This confirm our expectation.

Another similar example is the PERCRET which is from our **forth type**: Percent of returns, since from our domain knowledge, only a small portion of customers will return the goods in clothes industry.



These plots also confirm our thoughts.

For this similar pattern, we will discuss the solution later.

(**Appendix**)We haven’t discussed variables related to time and number of buying and also coupon or markdown of buying, because we consider this are less important variable and tend to have a normal pattern. Because of limitation of report, we put this issue on appendix.

Also, for the accounting figures, we will discuss later in the functional variable part.

**Standardization**

Recall from our analysis in statistical(Appendix) summary before, our variances and means among variables vary largely. Although normalization help us to reduce these scales, by checking on the statistical summary (Appendix) on the normalized data, the difference still exists, with mean ranging from 0.0466976 with POUTERWEAR to 2.69896 with SMONSPEND

and variance ranging from 0.0696601 with PC\_CALC20 to 2.008351068 with PSSPEND.

Here we apply z-score to do standardization on our numerical variables, to achieve common mean 0 and variance 1.

For checking, we output distribution and statistical summary of SMONSPEND which originally has highest mean for illustration.

Mean of PC\_CALC20: 0.00

Mean of PSSPEND: -0.00

Std of SMONSPEND: 1.00

Std of POUTERWEAR: 1.00

It has achieved mean 0 and variance 1.

**Derive new variables**

Looking through our variables, some repetitive counts or predetermined internal relationship might exist in our variables, which might create sort of multicollinearity and harm the accuracy of our following models. We should look deeper in our variables.

**(1) Eliminate Repetition**

For example, OMONSPEND, TMONSPEND and SMONSPEND, which correspondingly represent amount spent in the past month, amount spent in the past 3 months and amount spent in the past 6 months. It is obvious that past 3 months includes information in the past month while amount spent in the past 6 months contains information both past month and past 3 months. To avoid these kind of multiple counts, two new variables 2-3MONSPEND and 4-6MONSPEND are derived to replace both TMONSPEND and SMONSPEND by using following formulas:

2-3MONSPEND = amount spent in the past 3 months - amount spent in the past month

4-6MONSPEND = amount spent in the past 6 months - amount spent in the past 3 months.

**(2) Functional variable**

Additionally, MON, FRE and AVRG, which separately represent total net sales, number of purchase visits and average amount spent per visit, could form an equation as follows:

Since average amount spent per visit might be linearly correlated with total net sales and need further analysis, so we build a correlation matrix between these three variables:

|  |  |  |  |
| --- | --- | --- | --- |
|  | FRE | MON | AVRG |
| FRE | 1 | 0.755201 | -0.30339 |
| MON | 0.755201 | 1 | 0.395475 |
| AVRG | -0.30339 | 0.395475 | 1 |

However, the correlation between total net sales and average amount spent per visit is just 0.395475, which is not as high as we expect, so we just keep the original variables.

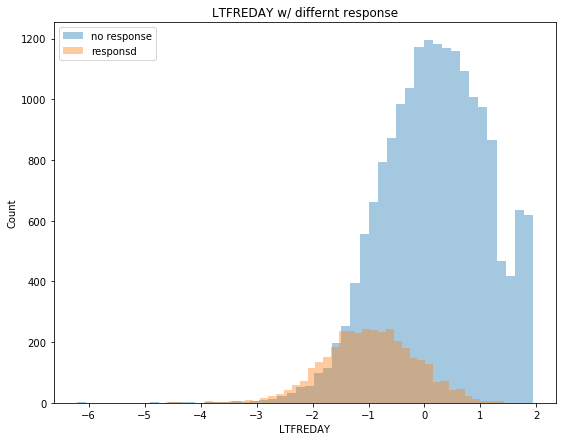
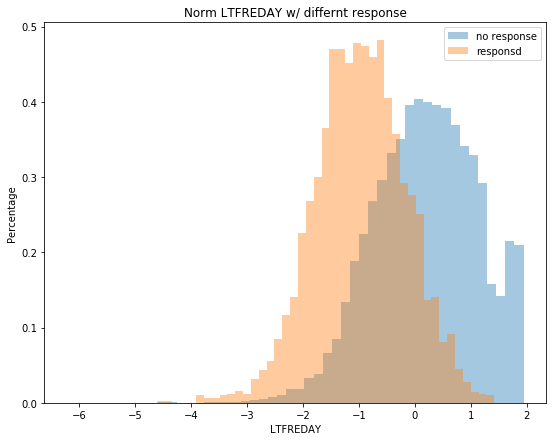
Here, we add categorical variables back to our dataset.

**Correlation**

Accomplishing derivation of variables, correlations between dependent variables and independent variables are calculated to find the most useful variables. The output for the top 8 absolute correlations are as follows:

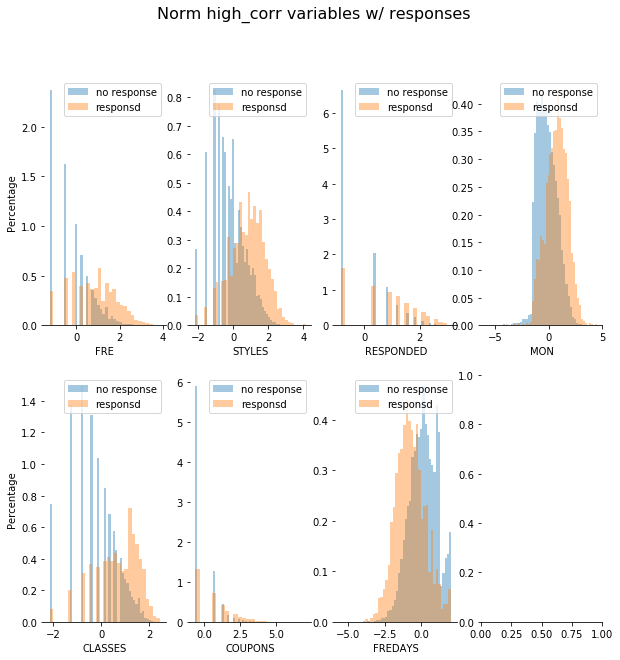
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LTFREDAY | FRE | STYLES | RESPONDED | MON | CLASSES | COUPONS | FREDAYS |
| 0.434007 | 0.400003 | 0.368682 | 0.337048 | 0.333518 | 0.328448 | 0.324806 | 0.32318 |

In order to understand the relationship between these variables and response, we draw a histogram of LTFREDAY with 2 different outcomes of response overlaying. Then, for better comparison and interpretation, we normalized the previous histogram by transferring vertical axis from count into percentage.



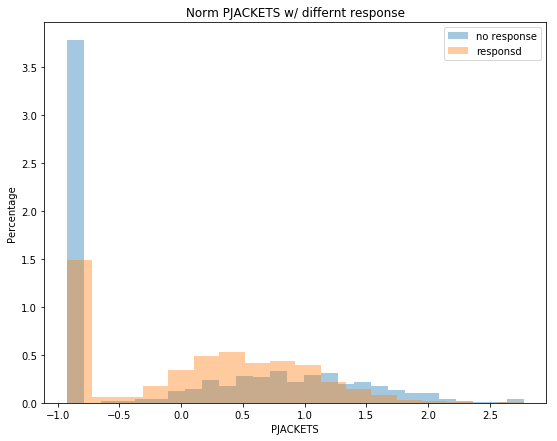
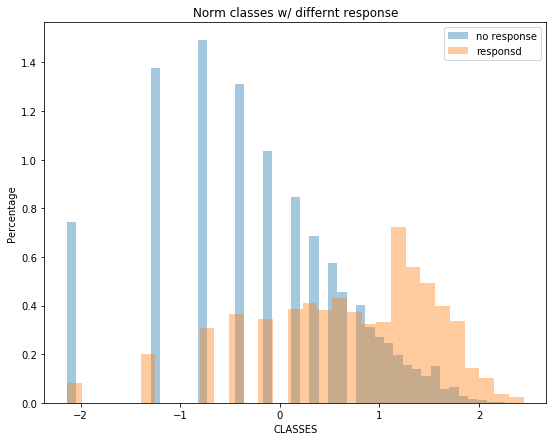
From the chart, we could conclude that with lower lifetime average time between visits, clients are more likely to response to marketing strategy. When lifetime increases, the percentage of response will decrease gradually. It makes sense because when the lifetime average time between visits are longer, it means customers have fewer interest in this store, resulting in lower probability response to marketing mail.

Similarly, we plot the remaining 7 normalized histograms.



The common trend is that except for FREDAYS, as the value of variables increases, they tend to have higher probability to respond. Recall these 6 variables, total number of individual items purchased by the customer, number of promotions responded to in the past year, total net sales, number of different product classes purchased, number of coupons used by the customer, when they increase, it implies that customers are more interested in this store, leading to high level respond. Conversely, when number of days between purchases increases, customers are less likely to respond.

In our previous analysis, we plot the histogram distribution of CLASSES. The following chart is consistent with our previous expectation that customer with high number of different product classes purchased tend to have a higher response probability.

Also, we have plotted jackets as well. Apart from people who do not buy jackets, we could see as customers spend more on jackets, they are less likely to respond to our market promotion. This is consistent with our previous conclusion with CLASSES. Since people concentrate on specific types, they are less interested in new products from promotion mail. This could also be confirmed by the plot of CLASSES.

**Multicollinearity**

We firstly find the correlation matrix for the entire training data, and set “correlation>0.8” as multicollinearity. The variables express multicollinearity are as follows:

After analysis of correlations with response variables, we also find correlation matrix among independent variables and “absolute correlation>0.8” as multicollinearity. (For those absolute value lower than 0.8, we replace with “Low”)

high corr exist(+ve): CLASSES,FRE 0.8

*high corr exist(+ve): CLASSES,MON 0.84*

*high corr exist(+ve): STYLES,FRE 0.86*

*high corr exist(+ve): STYLES,MON 0.88*

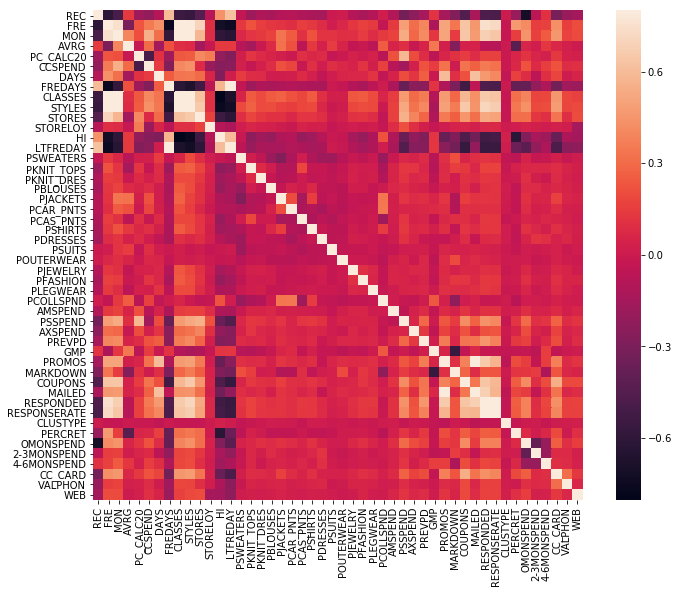
*high corr exist(+ve): STYLES,CLASSES 0.92*

*high corr exist(-ve): HI,CLASSES -0.8*

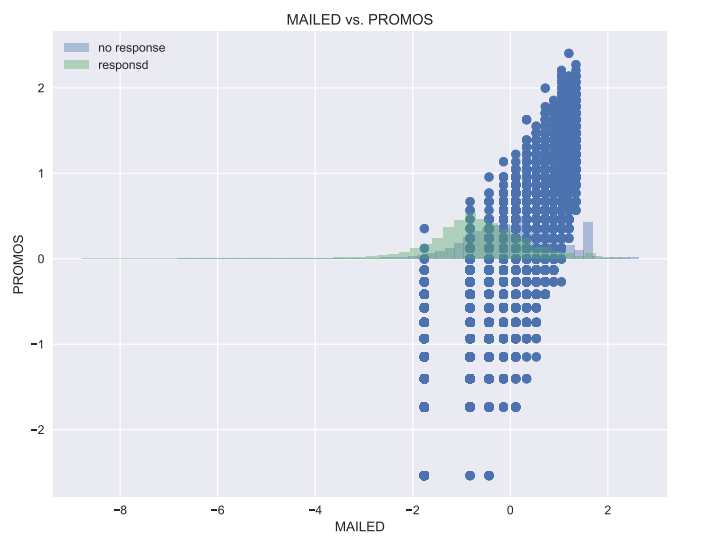
*high corr exist(+ve): LTFREDAY,FREDAYS 0.82*

*high corr exist(+ve): MAILED,PROMOS 0.9*

*high corr exist(+ve): RESPONSERATE,RESPONDED 0.94*

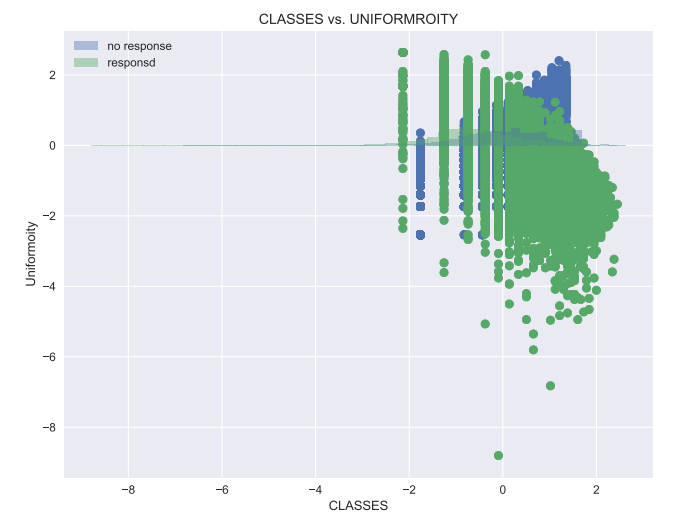


|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | LTFREDAY | MAILED | CLASSES | STYLES | RESPONSERATE | HI | |
| FRE | Low | Low | 0.804477 | 0.859882 | Low | Low |
| MON | Low | Low | 0.839951 | 0.8841 | Low | Low |
| FREDAYS | 0.821804 | Low | Low | Low | Low | Low |
| CLASSES | Low | Low | 1 | 0.923728 | Low | -0.80159 |
| STYLES | Low | Low | 0.923728 | 1 | Low | Low |
| HI | Low | Low | -0.80159 | Low | Low | 1 |
| PROMOS | Low | 0.89884 | Low | Low | Low | Low |
| RESPONDED | Low | Low | Low | Low | 0.939744 | Low |

For the positive relationship, we choose MAILED VS. PROMOS as an example for illustration

It is obvious there is positive relationship between mailed and promotion. This make sense because number of marketing promotion on file including number of promotions mailed in the last year.

For the negative relationship, there is only one pair CLASSES vs. UNIFORMROITY that correlation is below -0.8. The scatterplot is as follows:



For multicollinearity, we are really prudent towards them. We haven’t deleted them directly. On the contrary, since our models are only used for prediction rather than others, we keep this multicollinearity to increase our models’ accuracy.

**Flag variables**

Here we first use jacket as an example**. (为了说明什么)**

Among those customers who buy jackets and also use credit card, their response rate is 10.74%.

For those who buy jackets and have a valid phone number, their response rate is 15.63%.

For those who buy jackets and are also web shopper, their response rate is 1.88%.

We could conclude for these jacket buyers, if they have a valid phone number, they

In our previous analysis, we have mention a pattern with a spike at zero and the right part has followed a normal distribution. For this condition, flag variables are created to describe and solve this problem.

We have created 3 groups of flag variables in total. (**Appendix**)

*The first group is “type”, which contains 'PSWEATERS', 'PKNIT\_TOPS', 'PKNIT\_DRES', 'PBLOUSES', 'PJACKETS', 'PCAR\_PNTS', 'PCAS\_PNTS', 'PSHIRTS', 'PDRESSES', 'PSUITS', 'POUTERWEAR', 'PJEWELRY', 'PFASHION', 'PLEGWEAR', 'PCOLLSPND'.*

*The second group is “spent”, which contains 'AMSPEND', 'PSSPEND', 'CCSPEND', 'AXSPEND', 'OMONSPEND', 'TMONSPEND', 'SMONSPEND', 'PREVPD'.*

*The third group is “others”, which contains 'RESPONSERATE','PERCRET'.*

**Methodologies**

In order to build and evaluate our models, we split our data (80/20) into training and test sets to evaluate performance of different models. The random\_state order is unique based on one of our student ID. Among our training data, 16.58% customers respond to direct mail marketing, which is not a large number.

**Interesting Base models**

Different from regression problem, which usually use Naïve method as benchmark for model evaluation, classification problem does not have this kind of basic model. Therefore, based on the task background, we construct two extreme models which are classifying all customers into response category and classify all into nonresponse category.

In our test set, there are 4348 sets. Among them, 727 respond to our promotion while 3621 do not respond. And the actual total respond respondent rate is 16.72%.

For the first one which assume all customers respond, which means we send mails to everyone, the cost benefit table,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Outcome | Classification | Actual Response | Unit Cost | counts | Total cost |
| True Negative | Nonresponse | Nonresponse | 0 | 0 | 0 |
| True Positive | Response | Response | -$43.56 | 727 | -$31668.12 |
| False Negative | Nonresponse | Response | $45.56 | 0 | 0 |
| False Positive | Response | Nonresponse | $2 | 3621 | $7242 |
|  |  |  |  | Overall cost | -$24426.12 |
|  |  |  |  | Average cost | -$5.62 |

For the second one, since the prediction is that no one respond, so no mail should be sent:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Outcome | Classification | Actual Response | Unit Cost | counts | Total cost |
| True Negative | Nonresponse | Nonresponse | 0 | 3621 | 0 |
| True Positive | Response | Response | -$43.56 | 0 | 0 |
| False Negative | Nonresponse | Response | $45.56 | 727 | $33122.12 |
| False Positive | Response | Nonresponse | $2 | 0 | 0 |
|  |  |  |  | Overall cost | $33122.12 |
|  |  |  |  | Average cost | $7.62 |

**Methodology**

Since this is a classification problem, which require outcome constrain in [0,1], classification models are used here instead of linear regression methods.

**Choosing balancing method**

Since in our test set, response account for 16.72% while nonresponse account for 83.28%, which are imbalanced classes. This might cause problem because it tends to have high accuracy prediction towards nonresponse and low accuracy prediction with response type, resulting in bias in our prediction. Therefore, we choose both upsampling method and downsampling method to cope with our training set. **(Reference)**To be consistent and ensure comparison, we need to choose the best balance method and apply on all of our models.

Refer to **Appendix (1)**, we evaluate the performance of balance methods by comparing output from unbalanced model group and 2 different types of balancing model groups on Combined Bayes, Logistic, L1, L2 and Trees, LDA, QDA, QDA regularized.

Generally, performances of 2 balancing methods on all models are similar except for tree method, in which sensitivity of upsampling method is largely higher than downsampling method indicating **upsampling** is the best balance method in our case and we will use this in all the following predictive models except baseline models.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Decision Tree\_depths | |  |
|  | Upsampling |  | Downsampling |
| Error rate | 0.269 |  | 0.217 |
| SE | 0.007 |  | 0.006 |
| Sensitivity | 0.7290234 |  | 0.3672627 |
| Specificity | 0.731842 |  | 0.8668876 |
| AUC | 0.812 |  | 0.617 |
| Precision | 0.353 |  | 0.356 |

We assume that predictor variables are continuous, although decision trees can deal with qualitative variables.