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 direct mail marketing. 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, direct mail will not be sent. This helps the company to use advertising budget more efficiently, and finally achieve profit maximization.

The cost-benefit table is built as follows:

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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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.

In conclusion, the costs corresponding to each outcome are

|  |  |  |  |
| --- | --- | --- | --- |
| Outcome | Classification | Actual Response | Cost |
| True Negative | Nonresponse | Nonresponse | 0 |
| True positive | Response | Response | $45.56 |
| 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 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, which means we don’t need to deal with NaN.

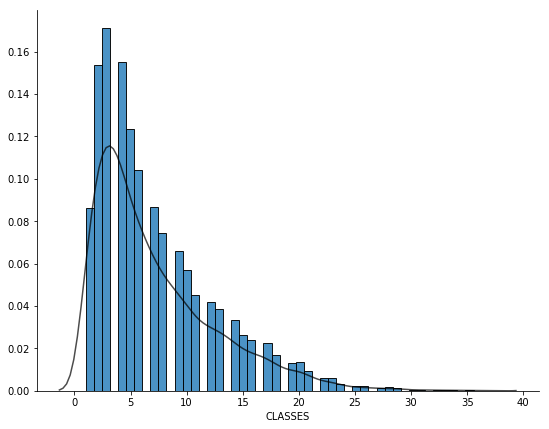
Firstly, we split our data (80/20) into training and test sets to evaluate performance of different models. The random\_state order is unique based on one of our student ID. Among our training data, 16.58% customers respond to direct mail marketing, which is not a large number.

Looking at the statistical summary**(Appendix)** the remaining variables, 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.

**Exploratory data analysis**

Regarding to categorical variables, we have some detailed analysis on CLUSTYPE. It is a variable containing 50 classes which describe find some interesting patterns in CLUSTYPE.

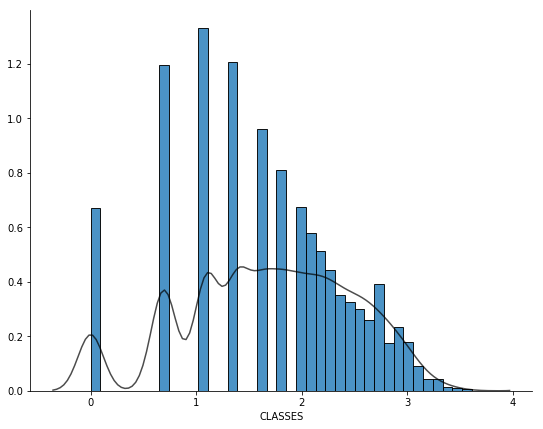
For numerical variables, from our previous analysis, they tend to be right-skewed. Here we plot “CLASSES” as an example. It explains the number of different product classes purchased for each customer. We should expect that customer with high number of different product classes purchased tend to have a higher response probability since they are more likely to try new products.



We could find more people tend to buy smaller classes of goods while less people prefer to purchase various classes of goods. The obvious unnormal patterns obviously call for data normalization.

Referring to transformation, log function is the first choice. However, some numerical variables 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.

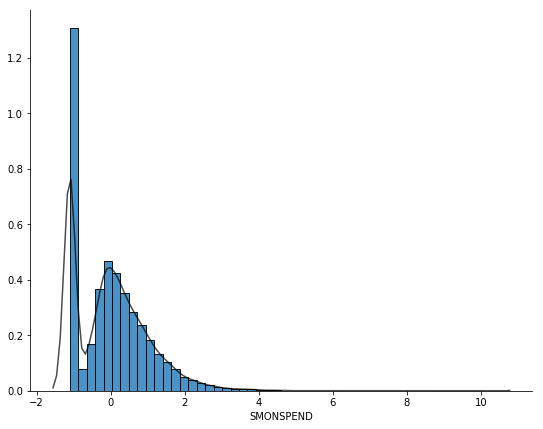
To test whether we have achieved normalization, we plot the distribution of “CLASSES” again. Although it is not perfectly gaussian distributed, it has been improved a lot in skewness and generally meet the assumptions of normality.



We check this by calculate the mean and standard deviation of our previous largest variance variable SMONSPEND, the output is

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

Then we plot the distribution



Although there might be outliers, it is much less skewed and generally standard normally distributed.

**Derive new variables**

Looking through our variables, some predetermined relationship seems to already be in our dataset, which might create multicollinearity. We should look deeper in our variables.

(1)

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)

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 this three variables:

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

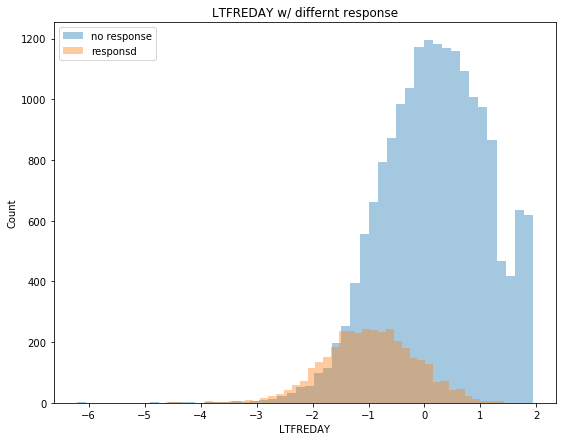
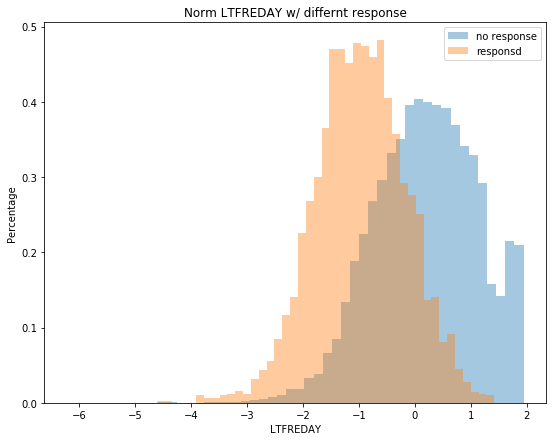
However, the correlation between total net sales and average amount spent per visit is just 0.395475, which is not as high as we expect, so we just keep the original variables.

**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:

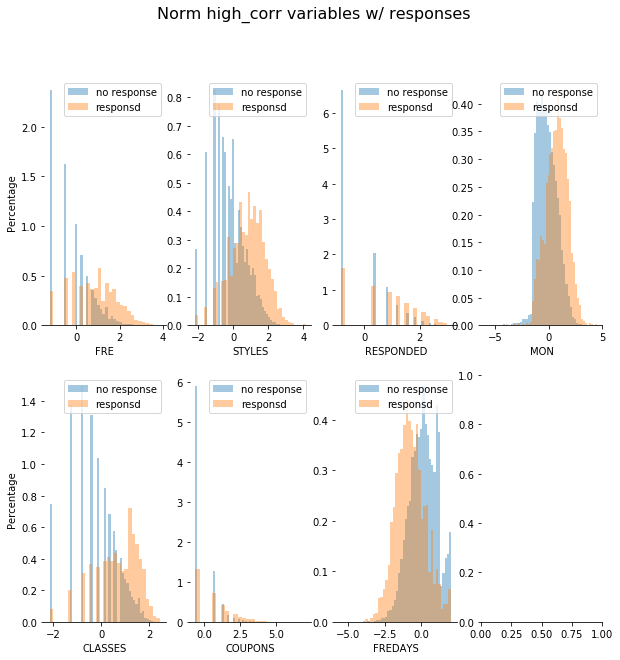
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LTFREDAY | FRE | STYLES | RESPONDED | MON | CLASSES | COUPONS | FREDAYS |
| 0.434007 | 0.400003 | 0.368682 | 0.337048 | 0.333518 | 0.328448 | 0.324806 | 0.32318 |

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.

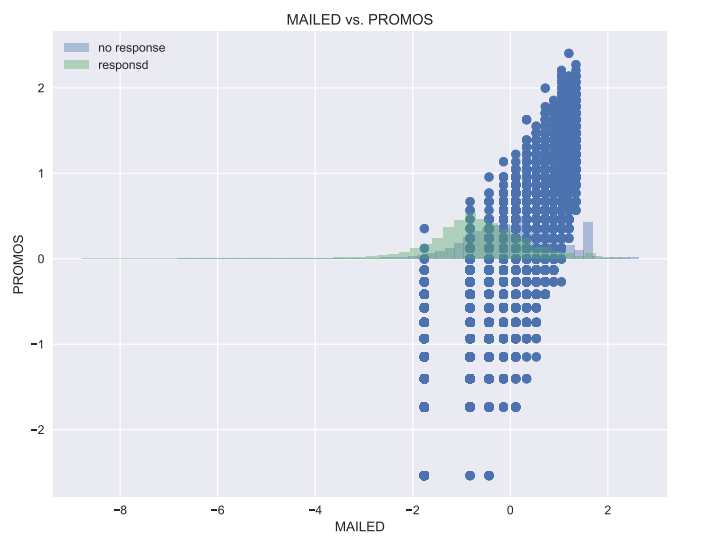
**Flag variables**

**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:

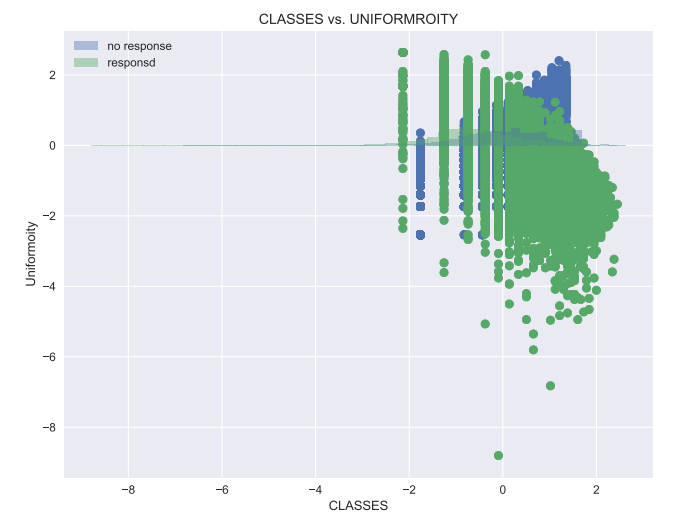
After analysis of correlations with response variables, we also find correlation matrix among independent variables and “absolute correlation>0.8” as multicollinearity. (For those absolute value lower than 0.8, we replace with “Low”)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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:



**下面的先不要看**

Interesting base models

1.

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:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Multicollinearity** | | | | | | | |
| CORRELATION | STYLES |  | CORRELATION | PROMOS |  | CORRELATION | MON |
| MON | 0.919288 |  | MAILED | 0.89619 |  | SMONSPEND | 0.889327 |
| FRE | 0.846314 |  |  |  |  |  |  |
| SMONSPEND | 0.816488 |  | CORRELATION | RESPONDED | |  |  |
| CLASSES | 0.812781 |  | RESPONSERATE | 0.826608 |  |  |  |

To be conservative, these columns haven’t been directly deleted, in variable selection part, PCA will be used to do further analysis.

We then do some exploratory data analysis on the training data. We conclude that 16.582% customers respond to direct mailing marketing.

By exploring predictors’ correlations with the response (Appendix correlation\_absolute), we find “FRE”, “CLASSES”, “STYLES”, “RESPONDED” and “RESPONSERATE” have top 5 correlation with response.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| RESP | FRE | CLASSES | STYLES | RESPONDED | RESPONSERATE |
| 1 | 0.4076 | 0.3728 | 0.3641 | 0.3506 | 0.3315 |

The following is the descriptive statistics for these 5 predictors (The descriptive statistics for all the predictors is in appendix). Sample skewness and kurtosis are added to verify the distributions of variables.

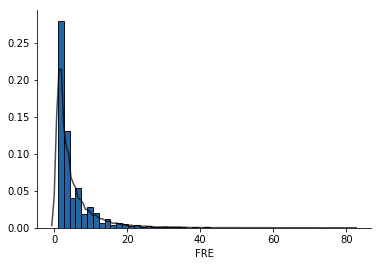
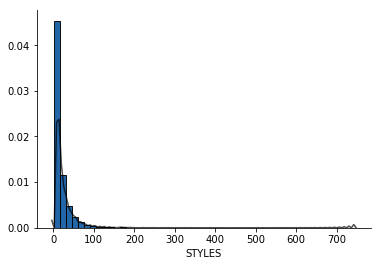
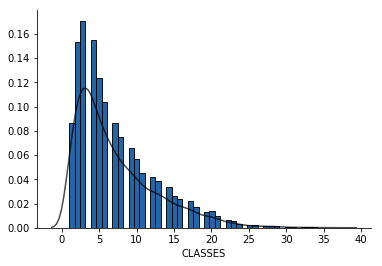
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FRE | CLASSES | STYLES | RESPONDED | RESPONSERATE |
| count | 17392 | 17392 | 17392 | 17392 | 17392 |
| mean | 5.094124 | 7.165823 | 17.38293 | 1.204462 | 17.36571 |
| std | 6.428738 | 5.375465 | 25.09014 | 1.848853 | 24.69021 |
| min | 1 | 1 | 1 | 0 | 0 |
| 25% | 1 | 3 | 5 | 0 | 0 |
| 50% | 3 | 6 | 9 | 0 | 0 |
| 75% | 6 | 10 | 20 | 2 | 30 |
| max | 81 | 37 | 743 | 11 | 100 |
| skew | 3.8394 | 1.30176 | 6.503903 | 1.971095 | 1.530816 |
| kurt | 22.72923 | 1.654053 | 101.5332 | 4.006621 | 1.775795 |

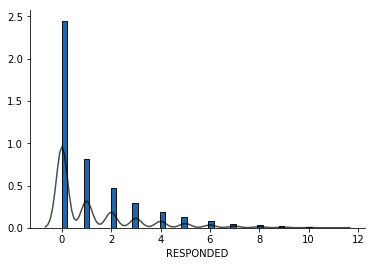
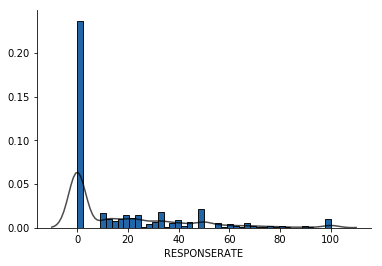
From this table, we could see mean varies, also, from our domain knowledge, they are measured in different units which imply further data standardization. In addition, skewness of all variables is far above 0 which means they are all positive skewed. Also, for kurtosis, they are all different from 3 ranging from 1.65 to 101.53 which express leptokurtic and extremely platykurtic. Since a majority methods and models require assumption of normality, further data normalization might be considered.

Feature engineering (data visualization)

**Numerical predictors**

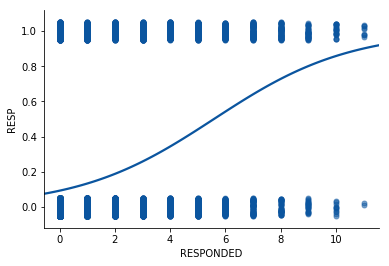
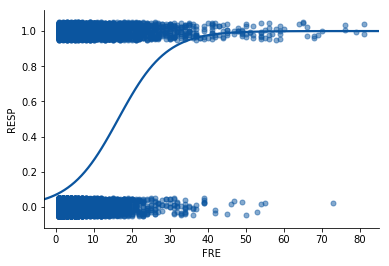
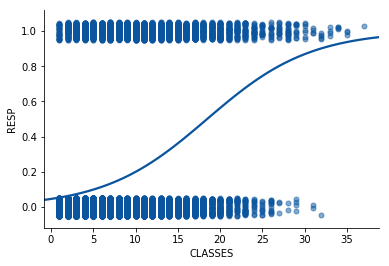
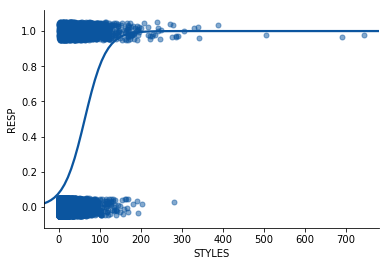
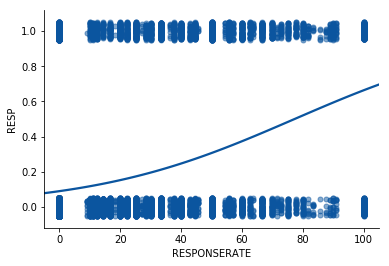
We then use histograms to analyze the distribution of the numerical predictors.





These are all numerical variables and their distribution confirm the conclusion of positively skewed distribution and look similar to lognormal distribution.

To explore the relationship between the numerical predictors and the response, we use univariate logistic regression.



FRE, CLASSES and STYLES tend to be strongly associated with a higher probability of response.

**Categorical variable**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

For categorical variables, we draw the cross table.

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.

Part3 Data preparation

Data transformation

Normalization

In order to meet assumptions of models, we do data normalization to transform data. We use log function to normalize previous data

From previous feature engineering pictures, the main variable

Normalization Standardization

Since different variables have different formats. For example, in order to tackle this incommensurable criteria problem, normalization is applied to transform original data with various scales into same scale, which bring standards and convenience to further analysis.

Data standardization is the critical process of bringing data into a common format that allows for collaborative research, large-scale analytics, and sharing of sophisticated tools and methodologies. (找一篇有类似定义的论文做reference)

Variable selection

PCA(normalization)

For a binary response with a 0/1 coding as above, if we use linear regression, some of our estimates might be outside the [0,1] interval, which violate the rules of probability. Because of this, classification methods are more preferred to qualitative response in our task.

Therefore, we use classification models including logistic regression, Quadratic Discriptive Analysis, KNN and tree method.

In particular the key differences between these two models can be seen in the following two features of logistic regression. First, the conditional distribution {\displaystyle y\mid x} is a [Bernoulli distribution](https://en.wikipedia.org/wiki/Bernoulli_distribution) rather than a [Gaussian distribution](https://en.wikipedia.org/wiki/Gaussian_distribution), because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0,1) through the [logistic distribution function](https://en.wikipedia.org/wiki/Logistic_function) because logistic regression predicts the **probability** of particular outcomes.

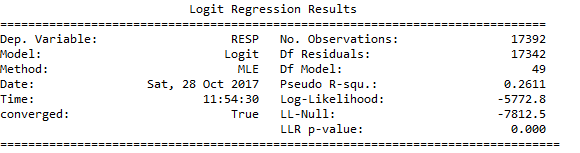
Methodology

Your analysis should take both predictive power and interpretability into consideration  
(possibly through different methods).

1. Logistic Regression

Logistic regression does not hold specific distribution assumptions. However, since coefficients in logistics regression are estimated by maximum likelihood method. Different from OLS, For maximum likelihood estimation, it need a large quantity of training set, which is just the case for our dataset.

Which is

Firstly, we generate an output of summary for logistics regression. 

From the selected variable, we sort variables based on coefficients, which can help in further explanation.

From the absolute coefficients, the top 15 coefficents are

4 Coefficients which are above 10: MON LTFREDAY CCSPEND PSSPEND

Then model is fit with balanced weight.

logit = LogisticRegression(class\_weight='balanced')

precision recall f1-score support

0 0.95 0.71 0.81 3621

1 0.36 0.80 0.49 727

avg / total 0.85 0.73 0.76 4348

then based on the test data, we get confusion matrix.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 2577 | 1044 |
|  | 149 | 578 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification** | | | | |
| **Nonresponse** | | | | **Response** |
| **Actual** | **Nonresponse** | 0.592686 | 0.24011 | |
| **Response** | 0.034269 | 0.132935 | |

For other performance measures,

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Error rate | SE | Sensitivity | Specificity | AUC | Precision | | 0.274 | 0.007 | 0.795048143 | 0.711681856 | 0.834 | 0.356 |   For confidence interval,  After adding l1 regression,  When adding l2 regression |  |  |  |  |  |
|  |  |  |  |  |  |

1. QDA

QDA and LDA are two models require that observations from each class are drawn from a Gaussian distribution. The differences are LDA assume that K classes share a common covariance matrix while QDA hold assumption that each class has its own covariance matrix.

From our picture,

According to bias-variance trade-off, since our training set is very large,

Gaussian Naive Bayes Method

1. KNN

Model evaluation

The model evaluation should be based on at least three substantivelydifferentmodels.  
• The model evaluation should include a confidence interval for the expected gross  
profit from a customer for each model.

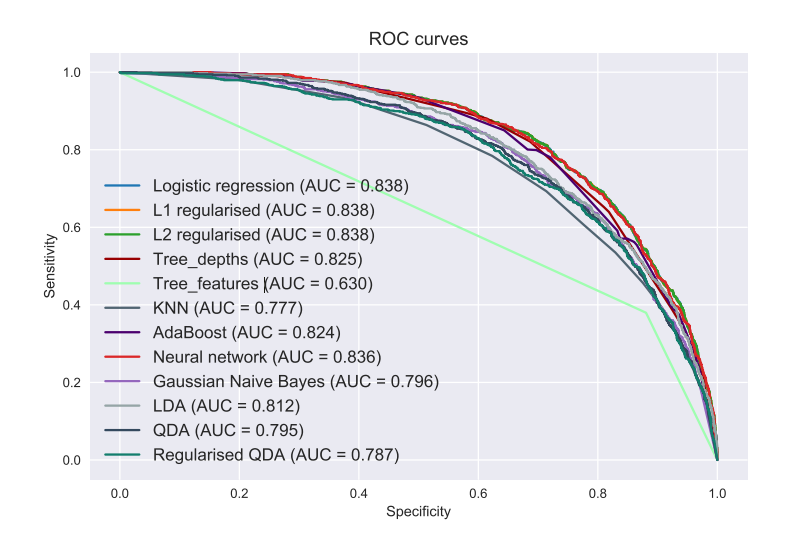
We select six indicators Error rate SE, Sensitivity, Specificity, AUC, Precision as our evaluation criteria. Based on the situation that the company want to maximize its profit and advertisement cost per person is greatly lower than the profit that we could get from the customer response, which is almost 200 times. Therefore, we consider more about the sensitivity rather than precision.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Logistic |  |  | Decision Tree\_depths | Gaussian Naive Bayes | QDA |
| Error rate | 0.149 |  |  | 0.292 | 0.242 | 0.239 |
| SE | 0.005 |  |  | 0.007 | 0.007 | 0.006 |
| Sensitivity | 0.22696 |  |  | 0.753782669 | 0.63823934 | 0.621733 |
| Specificity | 0.975973 |  |  | 0.69925435 | 0.782104391 | 0.788732 |
| AUC | 0.832 |  |  | 0.808 | 0.798 | 0.798 |
| Precision | 0.655 |  |  | 0.335 | 0.37 | 0.371 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

From previous output, it seems Decision Tree\_depths method performs well than other method with highest sensitivity 0.821183,

ROC



Reference

https://auspost.com.au/business/marketing-and-communications/bulk-mailouts/bulk-mail-options/acquisition-mail#tab2