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. After prediction, for those customers who respond to this market strategy, direct mailing is sent. For those who don’t respond, we don’t use this promoting strategy. This helps use advertising budget more efficiently and enhance the customer relationship management.

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

From our existing data, we infer that the expected gross profit from a customer could be calculated by

,

And the final result is 471.01.

For advertisement cost, it mainly contains printing fee and postage. 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.

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 | |

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 47 numerical variables and 4 ordinal variables: HHKEY, CC\_CARD, VALPHON and WEB. Among them, CC\_CARD and WEB have already been dealt in 1-0 binary format so we only need get dummy variable for VALPHON. Additionally, HHKEY stands for Customer ID which is a random set which is of no use for prediction, so we delete it.

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.

**Exploratory data analysis**

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.

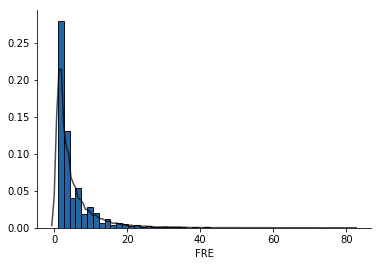
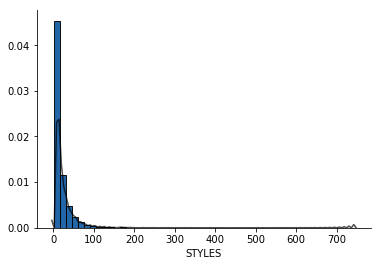
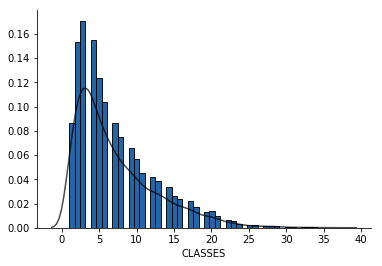
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FRE | CLASSES | STYLES | RESPONDED | RESPONSERATE |
| count | 17392 | 17392 | 17392 | 17392 | 17392 |
| mean | 5.094124 | 7.165823 | 17.38293 | 1.204462 | 17.36571 |
| std | 6.428738 | 5.375465 | 25.09014 | 1.848853 | 24.69021 |
| min | 1 | 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 |

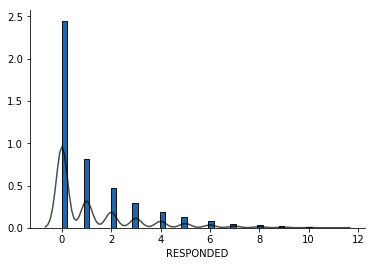
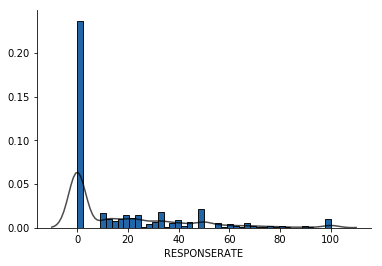
From this table, we could see mean varies, also, from our domain knowledge, they are measured in different units which imply further data standardization. In addition, skewness of all variables is far above 0 which means they are all positive skewed. Also, for kurtosis, they are all different from 3 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**

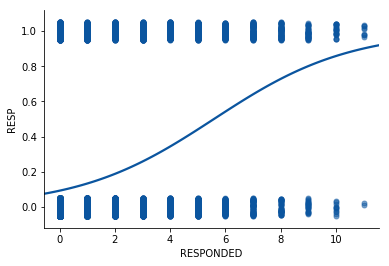
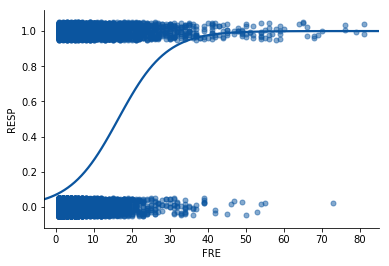
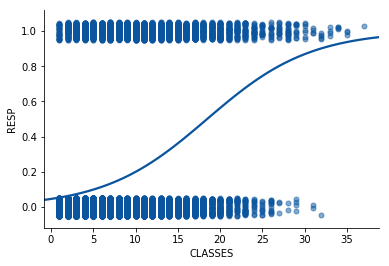
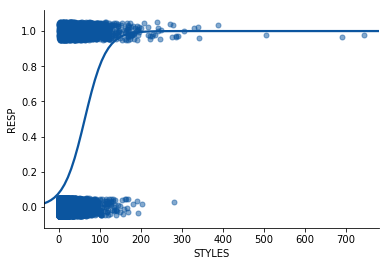
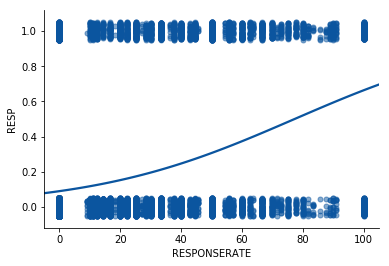
We then use histograms to analyze the distribution of the numerical predictors.





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.

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)

Model selection

1. GDA
2. Logistic Regression
3. Naïve Based Method
4. KNN

Model evaluation

ROC

Reference

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