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.

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

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

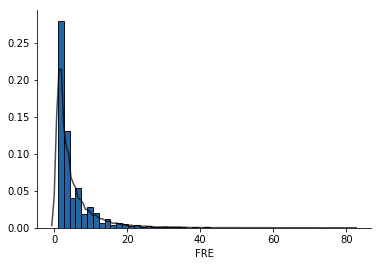
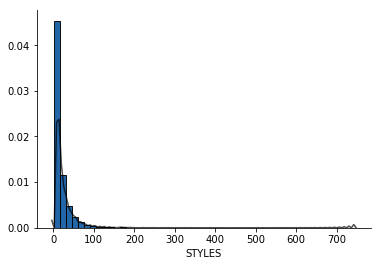
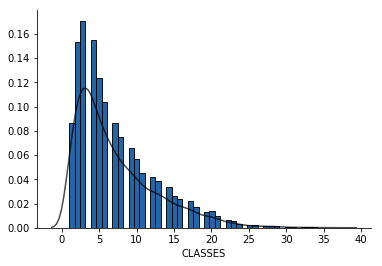
By exploring predictors’ correlations with the response (Appendix correlation\_absolute), we find “FRE”, “CLASSES”, “STYLES”, “RESPONDED” and “RESPONSERATE” have top 5 correlation with response.

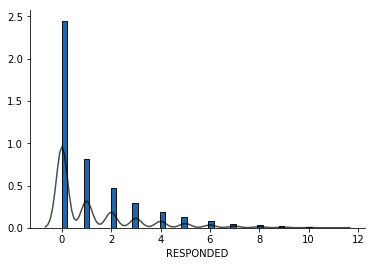
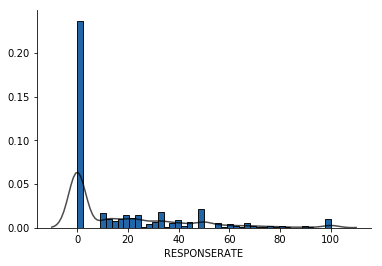
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| RESP | FRE | CLASSES | STYLES | RESPONDED | RESPONSERATE |
| 1 | 0.4076 | 0.3728 | 0.3641 | 0.3506 | 0.3315 |

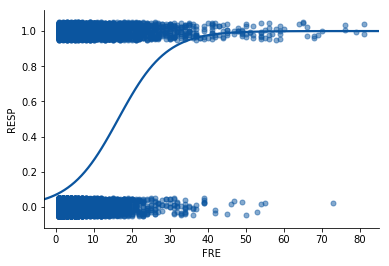
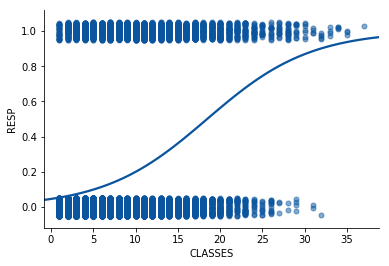
The following is the descriptive statistics for these 5 predictors (The descriptive statistics for all the predictors is in appendix). Sample skewness and kurtosis are added to verify the distributions of variables.

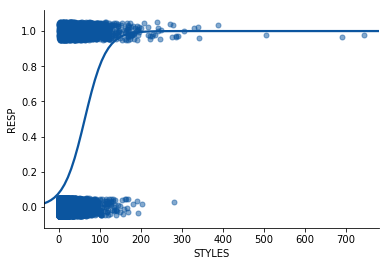
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FRE | CLASSES | STYLES | RESPONDED | RESPONSERATE |
| count | 17392 | 17392 | 17392 | 17392 | 17392 |
| mean | 5.094124 | 7.165823 | 17.38293 | 1.204462 | 17.36571 |
| std | 6.428738 | 5.375465 | 25.09014 | 1.848853 | 24.69021 |
| min | 1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

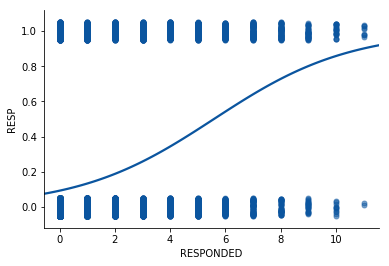
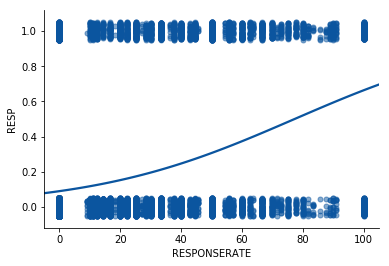
We then use histograms to analyze the distribution of the numerical predictors.





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.



Part3 Data preparation

Data transformation

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

1.PCA