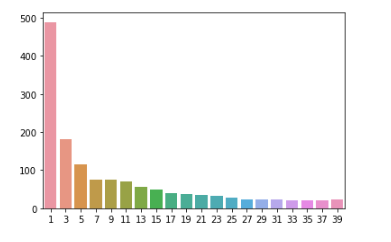
**Anirudh Vunnam**

**06/20/2022**

**Introduction**

This project will enable a magazine company to understand the cause of subscription decline by analyzing a dataset of its subscribers. The marketing dataset that was analyzed contains records of subscribers and their individual preferences for different products. By analyzing this dataset, the company questions what marketing strategies work to their benefit or against them based on the output of the analysis. To achieve this, the report focuses on investigating who will respond to an offer for a product or service. By doing this, the company could use the conclusions acquired as a reference in the marketing of its products.

**Data Cleaning**

The dataset contains 2240 records and 29 variables. We can see that the feature ID represents the customer IDs of the various customers over a while, but the data does not reflect a consecutive collection of data from all the customers, which is why the length of our data is not in line with the range of the customer IDs in this feature. Due to the fact mentioned above, the feature is distorting the data a bit, so we will be dropping the feature from our data as it is not adding any value to our data and hence not required for our analysis. From the dataset, we can observe the feature 'Year\_Birth' represents the different years of birth of the customers. The data collected contains some gaps in this feature; in terms of particular years in which customers were born. So our data does not contain all the years starting from 1893 to 1996. From the education feature, We can see that there are classes in this feature, but there are 2 among them which mean the same but are represented in two different ways; i.e. '2n Cycle' and 'Master'. So we will replace all occurrences of the class'2n Cycle' with 'Master', for a better value representation. The feature Dt\_Customer represents the dates of the customer’s enrolment with the company. The data in this feature is represented in type str. To improve the value representation of this feature, and to do meaningful feature engineering with this feature, we will be converting the values in this feature to data type DateTime.We also get to see that there are missing values in the variable income. The best way to impute the missing values in this feature is to utilize relative imputation strategies. We can use the model of the feature for imputation purposes of the missing values in this feature if we assume that the data is a high degree representation of the bigger population, which means that any new data point will be in line with the distribution of the data. 

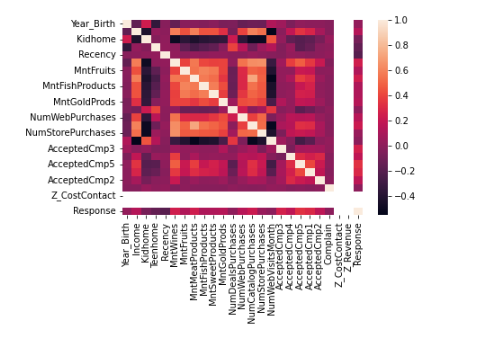
Because of the fact we cannot be sure of how representative our data is of the larger population, we will be using KNNImputer in this case outlier detection using Local Outlier Factor. The number of outliers detected by LOF is 34 and the dimension of the data after removing outliers is (2206, 26)

**EDA**

In this section, an exploration of what variables may drive the default will be stated. The exploration is divided into two parts: numerical variables and categorical variables.

**Understand numerical variables**

To understand numerical variables, correlations between numerical features and the response are focused on, Based on the heat map of correlations, preliminary conclusions relating to the business problem can be obtained.



From this, we can conclude that most of the numerical features correlate with the variable “Income “, and according to the color depth, it can be correctly assumed that the variable “mnt fish products” and “Mntmeatproducts” have a high correlation with “Numstorepurchases”. We can see from above that a lot of values in the heat map representing the correlations between the features are null values (excluding the diagonal values); this is due to the presence of imbalanced data in some of the features in our data which leads to 0 variance, resulting in a null value for correlation with those features. The features with imbalanced features will be handled in the later stage of the modeling.

**Understand categorical variables**

To understand the impacts of categorical features, visualization plots were drawn to give insight into judging whether category variables had a significant impact on the subscription to the product. Binary categorical features with imbalanced data are identified to prevent false outlier detection of the minority classes mentioned before, we will be excluding them from the outlier detection and removal step. And to prevent majority class prediction by the classification algorithms, we will be handling the above features to balance the frequency of the binary classes later in the process. After observing the table in appendix 1, we can get a preliminary conclusion: customers who have an education level of 'Graduation' show the highest rejection levels in the last campaign, where the effect of 'Income' was insignificant towards the kind of response. We can also see that the level of the customer's income is insignificant towards their levels of a rejection response, for the customers having a 'Master' in education.

For the customers with a 'Basic' educational background we can see that the rejection levels are relatively lower than for customers with other educational backgrounds; but within these classes of customers, the ones having a low level of income show significantly higher levels of a rejection response towards the last campaign. From assessing the customer in terms of their age, the length of the business partnership influences the response made to the campaigns. We can conclude so far that we can see a similar level of relatively higher rejection responses from customers who are old and former customers. We can also see a similar level of relatively lower but independently higher, rejection levels from new and current customers. A common pattern among all kinds of customers is that the highest levels of rejection responses are shown by the customers in the age bin of 3 to 7. Based on the above preliminary conclusions, we can think that among the categorical variables, the variable "Education", variable "Marital Status", variable “kid home" and variable "numDealsPurchase" have a significant impact on the preference for products and campaigns. These variables have to be focused on in the analysis part.

**Analysis**

**Modeling**

After completing the preliminary analysis of the data set, this report will complete further analysis by establishing models based on the preliminary conclusions. To accurately draw up a model, more thought is on how suitable the chosen method will affect the prediction of trends. This section will expound on the logistic regression model and decision tree model used. After establishing the model it can then be optimized through future engineering methods to increase the accuracy of the model. In turn, determining what approach they can take to increase their market share, splitting the data for hyper-parameter optimization and model selection.

**Logistic regression model**

The first model we will analyze is the logistic regression. It is used because the dependent variable is of binary classification achieves better results with logistic than with any other model .It also supports the categorizing of data into discrete classes through the study of relationships in the variables from a given set of labeled data. Secondly, this model is great at interpreting different features, and these separate features have a large influence on the final results. Finally, a logistic regression model is established for prediction. By interpreting the value and coefficients of each variable in the data, we can get some conclusions. Among them, the numerical variables, variable "customer age", and variable "income", have a significant role in the product being bought from the company. This shows that in case the household where there is we can see high levels of rejection responses from customers: having 0 teenagers or kids, having only kids and o teenagers, having equally kids and teenagers, having only teenagers. On the contrary, customers with 70% kids and 30% teenagers are showing lower rejection responses towards the last campaign. In order to summarize the above observations, customers with all kinds of combinations of kids and teenagers (except 30-70 ratio) in their family show similarly high rejection responses. The use of this model will enable a future analysis to be done on both binary and categorical data.

**Support vector machine**

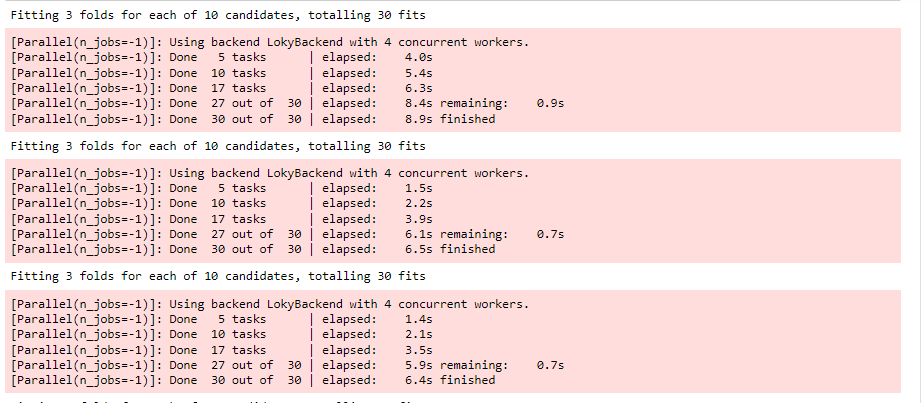
Use of this model ensures that one produces vital accuracy of the predictions with less computation power needed. This model can be applicable when it comes to both the regression and classification tasks but is largely used on classification purposes. The model’s objective is to find the most appropriate plane with a maximum margin. In order to classify future data more appropriately, the margin distance provides stability for the data points. Support vectors in other words are data points that influence the orientation and position of the hyperplane. Here we take the output of a linear function and check if the output is greater or less than 1, if it is greater than 1, it is identified with one class, and if -1 it is identified with a different class. By doing this, the reinforcement range of the selected values [-1,1] now can act as the margin. From our marketing dataset, we will implement the SVM algorithm by maintaining the number of classes as two. Selecting the feature to use is done by selecting the variable “total spent on the money” and variable income. We then take these features and plot them to visualize the relationship among the values. ” total spent on the money” is vital to acquiring data about the company subscription as it differs across the whole graph, to assess this is crucial as the number of sales a customer does is greatly influenced by the income they get. Clearly placing the learning rate and configuring the regularization parameter ensures that the number of epochs is maintained at appropriate levels.

**Decision tree model**

The second model used is the decision tree model. This model has advantages such dividing of each subscription market on each of these variables to get groups that are well defined in terms of their behavior. From the case, we can assume that dividing customers based on their income at an average amount is a major classifier because from this values can we analyze the majority of the data. At this point, one identifies one group of customer prospects who are more likely to respond favorably to the marketing campaign. In case this group may be more than the target, the decision tree moves on to the next step of building the decision tree. The advantage of Decision trees over the outdated method of decision-making is that it takes up several variables to predict the most accurate outcome. Combining the original data set, we can think that among the dependent variables, the “age” variable should be used as the highest priority judgment indicator to infer whether the customer will be for the campaign or against it. Lastly, the secondary judgment indicator is “days since enrollment”, and then from the visual decision tree we can use “income” “complain” and “customer age” as further judgment because the above variables have a significant impact on the segmentation of the products

**Optimize**

Based on the conclusions of the above two models, the model will be optimized next, optimization to enhance the prediction of variables. The optimizer is tuned to produce the most beneficial model for the scheme.



We already know that there is a high correlation between the variable “days since enrollment” and the variable “income". To preserve the characteristics of the different variables at the same time without making the accuracy of the model affected by the correlation between the dependent variables, a new variable is inserted into the data set and delete variable “age”. At the same time, as a criterion for judging, the percentage of kids in households is more efficient than a specific age criterion, and it is also easier to classify when used as a criterion for judging.

On the optimization of the data set, the decision tree model is reestablished as the accuracy of the decision tree is improved and the rank of variables according to their importance has changed. Based on the decision tree, we can find that the variable “income” is still the most important factor influencing the liking of a marketing campaign. The variable “age” is ranked second, followed by the “kid number” in a household. Their importance is vital and can be considered as being close. This data combined with the newly obtained decision tree visualization we can conclude that the variable “income”, variable “customer age”, variable “number of days”, variable “education”, the variable “complain” and variable “kids number” should be considered when predicting the success rate of a campaign. The specific judgment process can follow the decision tree as shown in the figure.

By optimizing the logistic regression model, decision tree, and support vector machine we can find from the summary in the appendix that the six variables derived from the optimized decision tree model have a significant impact on the product subscriptions. Finally, we can compare the F1 score which is the weighted average of precision and recall, and accuracy in the table below to get the conclusion that the optimization for the decision tree model was a success.

|  |  |  |
| --- | --- | --- |
| ***model*** | ***accuracy*** | ***F1 score*** |
| Optimized Logistic regression | 0.858 | 0.61 |
| Decision tree | 0.883 | 0.53 |
| Support vector machine | 0.906 | 0.58 |

**Conclusion**

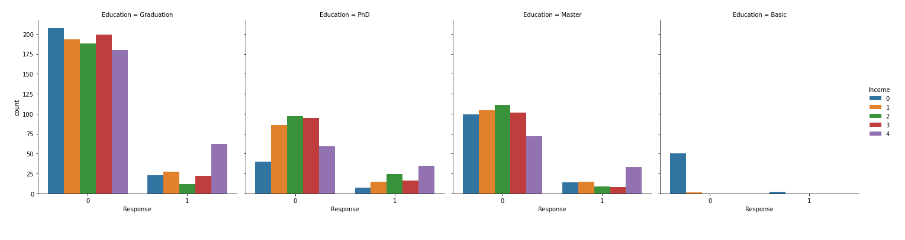
Based on the above analysis, we can find that it is not simple to accurately predict whether a customer will promote product campaigns and the optimized model is just about 80%. Moreover, by the series of analyses on the dataset, we can make an educated conclusion on factors that influence the response of customers to the marketing of products. We have found out that the most beneficial customers to the company do not complain largely compared to the average customers. From the conclusion of the model, we find out that the total days of customer enrollment with the company has a vital role if the customer will respond positively to a campaign. Finally, from this report, we can give the company insight into a set of procedures for judging whether the customer will like a product or not. If a product is chosen, it must have been influenced by different factors.

**Reference**

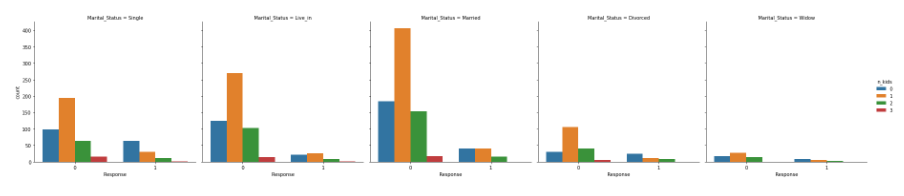
Nath, P. (2020). Metrics and marketing--model. *PsycTESTS Dataset*. <https://doi.org/10.1037/t81775-000>

**Appendix 1**

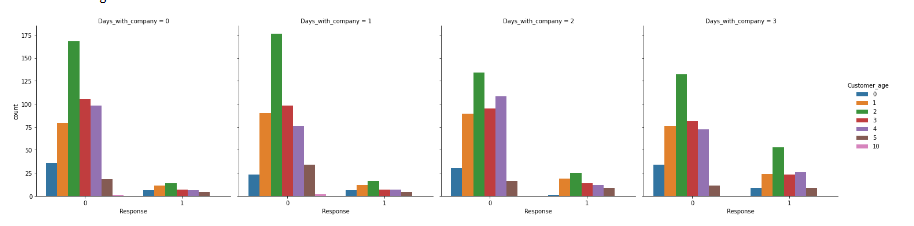
Visualization of Education and Income with Response of the customers



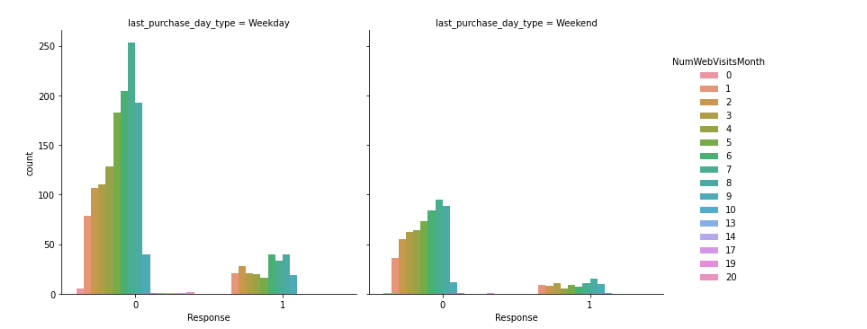
Visualization of Marital\_Status and n\_kids with Response.



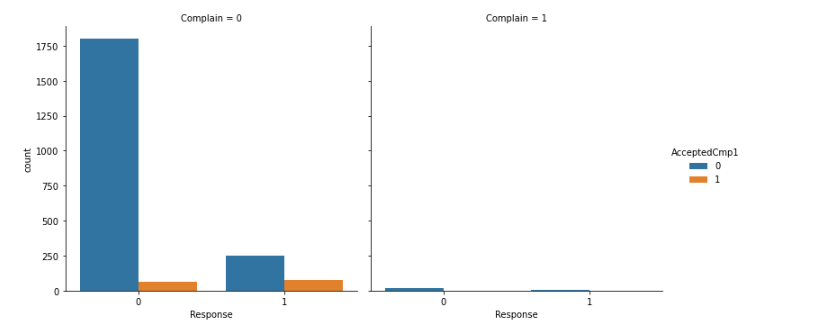
Visualization of Customer\_age, Days\_with\_company, and Response.



Visualization of NumWebVisitsMonth, last\_purchase\_day\_type with Response.

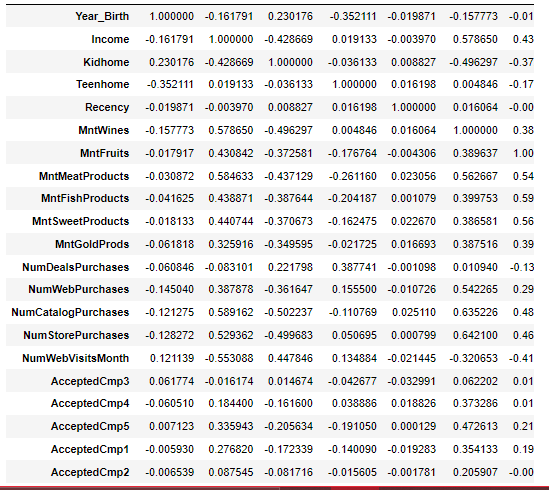


Visualization of AcceptedCmp1, Complain with Response.



**Appendix 2**

The summary of the optimized regression model



**Appendix 3**

Decision tree

Trial=0.351

Samples=18

Value=[51,23]

Class=0

subscription\_enrollment<=0.5

trial=0.42

Samples=500

Value=[500,125]

Class=0

true

false

Income<=46.0

Trial=0.461

Samples=460

Value=[270, 152]

Class=0

complain<=0.5

trial=0.459

Sample =170

Values=[230,140]

Class=0

Income<=20.0

Trial=0.23

Samples=120

Value=[60, 52]

Class=0

Store\_purchase<=0.5

Trial=0.461

Samples=320

Value=[170, 136]

Class=0

Web\_purchase<=0.3

Trial=0.31

Samples=120

Value=[34,20]

Class=0

complain<=0.5

Sample =170

Values=[230,140]

Class=0

Trial=0.488

Samples=12

Value=[4,2]

Class=0

Trial=0.0

Samples=134

Value=[43,27]

Class=0

Trial=0.432

Samples=32

Value=[14,6]

Class=0

Trial=0.351

Samples=18

Value=[51,23]

Class=0

Trial=0.476

Samples=21

Value=[74,52]

Class=0

Trial=0.23

Samples=56

Value=[48,7]

Class=0

Trial=0.432

Samples=32

Value=[14,6]

Class=0

Trial=0.248

Samples=16

Value=[23,2]

Class=0

Trial=0.563

Samples=8

Value=[45,2]

Class=0