# Introduction

In this paper, we will evaluate given dataset and select appropriate statistical model to fit the data. We will use fitted model to make prediction followed by testing the accuracy of model with new data.

The given dataset is customer data of a magazine company, data points include demographics of customer along with other internal parameter such as, customer’s date of registration with company, customer activities such as number of deals purchased, number of web purchases, catalog purchases, purchases in store and number of website visit each month. We have other information such as recency and customer interest in different products such as wines, fruits, meat, fish, sweet and gold.

The given set have 29 columns and 2240 records, few of the features in given dataset is difficult to understand without the data dictionary, however, we know that the target variable is “Response” column, which is indicator of user subscription of magazine. ‘Response’ in given dataset is a binary variable where 1 indicates user is a subscriber whereas 0 indicates that user choose not to subscribe magazine.

Since we are working with binary variable as target, we will be using Logistic regression to fit the model and predict the user’s likelihood of buying subscription as binary output. We will also fit the same data using KNN classifier to compare the difference of accuracy between two models.

## Logistics Regression

Logistics regression is a statistical predictive model used when dependent or target variable is binary or categorical. It is widely used in medical research to predict the medical/health status patients based on multiple independent variables.

For example: Research can fit a model with heart patient data, this model can help them to evaluate new patients’ likelihood of heart diseases based on historical trends. One can also calculate increase in probability of getting heart diseases by each increment in Body mass index of patients using the same model.



# Analysis and Implementation

For the analysis of the given dataset, we will use Python as data procession language. We will start by important all the libraries and package we intent to use for analysis and modelling.



The given data set is csv file, where data is separated by “;”, we will use pandas to import data indicating the semi-colon as delimiter.

Imported dataset includes, 2240 records with 29 columns. The first column “ID’ will be removed from data frame as it will not be useful but rather misleading in the model.

### **Exploratory data analysis and Data cleaning**

We start of by checking for missing values in dataset. 24 records in the dataset are missing Income data of customer. Since Income is a big range continuous variable, we will fill the missing value with average income of respective “Marital Status” category.

1. data['Income'] = data['Income'].fillna(data.groupby('Marital\_Status')['Income'].transform('mean'))

While evaluating descriptive statistics of numerical variables in the dataset, we found that two feature variables ‘Z\_CostContact’ and ‘Z\_Revenue’ are constant. Since the constant variable will add no value in model, we will drop these feature variables from data frame.

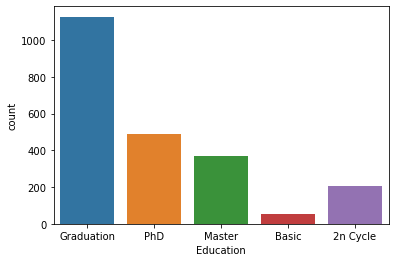
1. data.drop({'Z\_CostContact','Z\_Revenue'}, axis=1, inplace=True)

### **Transforming Categorical Variables**

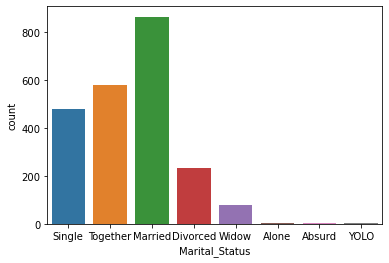
We have three categorical variables, 'Education' with 5 levels of education, 'Marital\_Status' with 8 categorical levels of customer’s marital status and 'Dt\_Customer' which is customer registration date.

Customer registration date is not formatted as date; hence we will start with formatting as date and transforming to Number of days since customer’ registration with the company, which as number, can be use in model as feature.



Moving to next categorical variable “Education”, we see 5 levels of education. However, few of the levels are synonyms and creates duplicate category. We will standardize this into 4 category, Basic, Bachelor, Master and PhD.

Since, the education level is an ordered category, i.e. level of education increase from Basic to PhD, we will map categories using **ordinal encoding.**

1. #changing categorical variable education to ordinal encoding
2. mapper = {'Basic': 0, 'Bachelor':1, 'Master': 2, 'PhD': 3}
3. data['Education'] = data['Education'].replace(mapper)

Now, we will investigate last categorical variable “Marital\_Status”. Based on frequency distribution bar graph, there are 5 major categories of marital status and 3 categories with very small distribution. Like education column, this also has multiple duplicate categories. We will first standardize the categories of marital status and change the categorical variable into dummy variable using “One-Hot-encoding” technique. We will remove one of the newly created dummy variables to avoid multicollinearity.

1. data['Marital\_Status'] = data['Marital\_Status'].replace({'Alone': 'Single', 'Absurd':'Single', 'YOLO':'Single'})
2. data['Marital\_Status'] = data['Marital\_Status'].replace({'Together': 'Married'})
3. data['Marital\_Status'] = data['Marital\_Status'].replace({'Widow': 'Divorced'})
4. data = pd.get\_dummies(data, prefix= ['Marital\_status'], columns = ['Marital\_Status'])
5. data = data.drop(['Marital\_status\_Divorced'], axis = 1)

Now that we have transformed all the categorical variable into numerical variables, we can use all the features in data frame “data” to feed into the model.

### **Feature selection**

As part of feature selection, we will try to eliminate feature which are very weekly related with target variable ‘Response’. To do so, we will take reference of correlation coefficient matrix which give us the Pearson correlation coefficient between feature variable and target variable.

1. #corelation evaluation for feature importance
2. data.corr()["Response"]

Since, there are more than 20 feature variables, a heat map of correlation matrix is not very informative. We will rather use mathematical operator to filter feature with least correlation coefficient with target variable. Here, I have chosen to eliminate features which has less than 0.05 correlation with target variable.

1. data.drop({'Year\_Birth','NumDealsPurchases','NumStorePurchases','NumWebVisitsMonth','Complain'}, axis=1, inplace=True)

### **Training and Testing**

We will first divide the data frame into independent variable and target variable, we will then split the overall records into training and testing set in 7:3 ratio.

1. #dividing target variable from feature variables
2. X = data.drop(columns='Response')
3. y = data.Response
4. #spliting dataset to testing and trainning set
5. X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state= 21)

Most of the variables in dataset are small range number, but 'Dt\_Customer' and 'Income' has wider and different range. If not scaled, one might dominate other features. We will scale the variable to avoid one feature dominance over other.

1. #feature scaling as Income and Dt\_Customer value lies in different ranges.
2. scale = StandardScaler()
3. X\_train = scale.fit\_transform(X\_train)
4. X\_test = scale.fit\_transform(X\_test)

Since, the target variable is binary, we will implement a Logistic regression model, predicting whether a user will subscribe the magazine or not.

1. model1 = LogisticRegression(random\_state = 21)
2. model1.fit(X\_train, y\_train)
3. #testing model1 against x\_test dataset
4. y\_pred = model1.predict(X\_test)
5. cfm = confusion\_matrix(y\_test,y\_pred)
6. **print**("Confusion Matrix: \n", cfm)

Confusion Matrix:

[[547 16]

[ 67 42]]

The above model, when tested with unfamiliar testing set yield the Accuracy of 87.64%

Similarly, we will use the same data to train and test a KNN classifier model.

1. #Knn model
2. model2= KNeighborsClassifier(n\_neighbors = 9)
3. model2.fit(X\_train, y\_train)
4. y\_pred2 = model2.predict(X\_test)
5. cfm2 = confusion\_matrix(y\_test,y\_pred2)
6. **print**("Confusion Matrix: \n", cfm2)

Confusion Matrix:

[[549 14]

[ 80 29]]

When tested with same testing set, KNN model yield the Accuracy of 86.01%.

# Conclusion

In the above models, we were able to predict whether customer would buy subscription of magazine with the accuracy of 87%, this indicates that the feature selected to train model directly or indirectly influences the customer decision. We will closely look into the feature variables selected for the model and draw out report to stakeholder to understand what drives user to subscribe to the magazine.

Both education and income have good positive correlation with magazine subscription. Since, education is encoded as ordinal variable Basic as 0(lowest) and PhD as highest numerical value in series, this indicate customer with higher level of education (Masters, PhD) are more likely to subscribe compared to lower level education (Basic and Bachelors). Also, the customer with high income are quick to subscribe than customer with relatively low income.

Number of kids, Number of Teen and Recency have negative correlation with magazine subscribe. This indicates that, higher the number of kids, lower the chance of subscription. Moreover, targeted demographic should be unmarried, with no children. Comparison between kids and teen shows that parents with kids are more likely a possible subscriber than parents with teen in their home. Also, low the recency, higher the chances of subscription. If customer have not used the service since many days, there likeliness of dropping out of subscription increases.

Although, “MntMeatProducts” and “MntWines” distribution have highest standard deviation, they have strongest correlation with subscription among other mnt product features. This reflects that the popularity of magazine is more among Meat Product and Wines customer than Fish, Fruits, Sweet and Gold products.

With respect to customer activities, marketing should be focused on customers with multiple instances of web purchase and catalog purchases, as expected, customer who purchase catalog are very likely to buy subscription. Also, customer’s days since registration has good positive correlation as expected, stakeholder should invest on holding veteran customers.

Due to absence of data dictionary ‘Acceptedcmp’ variables are not understandable, however based on correlation coefficient, we can still advise that, ‘AcceptedCmp5’ is most favorable for subscription, followed by AcceptedCmp1, AcceptedCmp3, AcceptedCmp4 and AcceptedCmp2.