**Introduction**

The goal of this project is to predict whether a customer will subscribe to a magazine or not and then to see what factors are influencing this decision the most. The company has seen a decline in the number of subscribers when they expected a raise in this number as a lot of people are spending more time at home. We look at various features and see what the reason may be for this and what are the groups of people who have subscribed and who aren’t so the company can improve their marketing strategies. We aim to help the company understand the reasons for the decline and what could improve the subscriptions.

**Data Cleaning**

Data cleaning is one of the most important steps in the data analysis process and to get accurate results. For the purposes of this project, we have been given magazine company’s marketing dataset. We have 2,240 rows and 29 columns. We have various features like the age, income, kids, marital status, how much money they spent on wine, fruits, meat, fish, sweet and gold; how many purchases they made during discounts, at the store, on websites and how many times they visited the website and whether the customer accepted their offers during previous campaigns or not etc. We also have 2 columns z\_CostContact and z\_Revenue. These columns have constant values i.e., the values are not same in every row, and they are not contributing to the model in any way. Refer Appendix A table 1 and 2. So, we remove these values. We now consider the year of birth column. We now look at the birth year, refer Appendix A graph 1. We can see that after the year 1940, the graph looks like it has normal distribution. We have a few values before 1900 which means the customer is at least 100 years old. So, we consider these values outliers and remove values before 1940. Now, we look at the income column, refer Appendix A graph 2, we see a normal distribution till 200k but there is an outlier after 600k. So, we remove this value. We can also observe that this column has 1.1% i.e., 24 missing values. Since income is a very important feature to design the model, removing all the rows with missing values will have a huge impact on the model. So, we try to replace them with mean, median or mode. We now must decide which imputation method we will be using to fill the missing values. The goal is to find out which is a better measure of the central tendency of data and use that value for replacing missing values appropriately. For symmetric data distribution, one can use the **mean value** for imputing missing values. For skewed data distribution, we use **mode value or median value** for imputing the missing values [1]. So, we now use box plots and distribution plots to see the distribution of the data (Refer Appendix A graph 3 and graph 4). We see that the distribution of the data is fairly normal. Outliers data points will have a significant impact on the mean. The number of outliers is less than 15 which is less than 1% of the data. Hence, we use mean to replace the missing values. We now convert the date column from object to DateTime type and then we split it into day, month and year since our model only accepts numerical values. We also convert the marriage and education columns into integer columns using one hot encoding for the same purpose as stated above. We observe that the data for all other columns have normal distribution and have no missing values. So, we move on to the next step.

**Data Loading**

Initially we load the marketing data using scikit learn which is a machine learning library.

We then manually divide the data into training and testing sets. We use the training dataset to train the model and then we use the same model on our testing dataset so we can compare the results from training data and testing data to check the accuracy of the model.

**Logistic Regression model**

Chart, scatter chart

Description automatically generated It is a supervised machine learning algorithm. It is used to predict the probability of the dependant/ target variable. Binomial logistic regression model is one of the most popularly used regression models for Logistic Regression. It helps categorize data into two classes and predict the value of a new input as belonging to either of the two classes. It works by using a Sigmoid function to map the output probabilities. We put 0 and 1 on x and y axis and the graph looks as follows. In our example, we are considering personal loan feature as our dependent variable and all the other features are considered as independent variables. We then create a confusion matrix for the results to check the accuracy and results. We use sklearn library to perform these functions.

**Support Vector Machine**

Chart, scatter chart

Description automatically generated The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Hyperplanes are decision boundaries that help classify the data points. We use sklearn svm model in our example.

**Confusion matrix, Classification Matrix and Accuracy**

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. **The confusion matrix shows the ways in which your classification model** **is confused when it makes predictions. The matrix predicts true positive(TP) in x[0][0] position, false positive(FP) in x[0][1] position, true negative(TN) in x[1][0] position and false negative(FN) in x[1][1] position.**

Accuracy is used to find the percentage of how accurately our model has predicted the dependent variable. To find accuracy we use logistic regression score/ svm model score from sklearn. The formula we use to find the accuracy is true positive + true negative/ (Total of all the values i.e., TP,TN,FP and FN).

**A classification report has columns like precision, recall, f1 score and support. Precision is the percentage of predictions that were correct. For example, for true positives the formula would be TP/(TP+FP).** **Recall is the percentage of positive cases that were caught. For example: TP/(TP+FN). F1 score is the percentage of positive predictions that were correct. Formula for F1 score would be 2\*(Recall\*Precision)/(Recall\* precision)[4] Support is the number of samples of the true response that lie in the class.**

Confusion matrix results

1. Logistic regression array([[380, 12],

[ 52, 4]], )

1. SVM array array([[386, 6],

[ 47, 9]], )

**Conclusion**

From the analysis in the code attached we can conclude the following:

1. Refer Appendix A Result 1, we see the different variables and what is the significance and impact of each of them after we apply logistic regression model. We see the P value to determine which variable is most significant. We see the P value is 0 or very low for the variables number of teenagers at home, the number of days since the customer’s last purchase, number of store purchases, whether they accepted the offer in the previous campaign or not and the date of enrollment with the company. We see the coef value to determine the impact of these variables. We see the month and year they were enrolled with the company has a negative impact on the subscriptions. It means newer customers are not accepting the offer. We can observe that the people who has accepted the offers in the previous campaigns are likely to accept the offer again since all the coefficient values are positive. As the number of people purchasing in the stores are increasing, the subscriptions are also reducing. If the customer has purchased the magazine recently, the subscriptions rate is reducing which is the same case with the number of teenagers at home. We can see kids at home has a positive impact and year of birth has a negative impact, so the magazine is popular among younger people. The company can target people from that age group or try to include content for older people. The company can also try to advertise or offer discounts for subscriptions so the people purchasing magazines from the stores may feel more inclined to subscribe. The newer customers are not accepting the offer, so the company needs to look at the new content and what changes they have made and how to make the newer customers subscribe.
2. Refer the above confusion matrix results and Appendix A result 2 and 3 which shows the classification report for these 2 methods. We also found the accuracy of the 2 models which was found to be 85 for logistic regression and 88 for SVM. Both of these models seem to be accurately predicting the response but picking SVM would be good choice here because of logistic model's inability to find a maximum of the loglikelihood function and subsequently to find the solution for your data and the sets of dependent and independent variables you are using. Looking at our data, we have a lot of 0's and identical values in the marriage and education columns, which may be problematic for finding a solution. We can also observe in the results that the model has exceeded maximum number of iterations so it may not yield accurate results.
3. We see in Appendix A result 2 is the result for logistic regression and result 3 is the result for SVM. The precision value is higher for SVM and we can also refer the above confusion matrix results and see that the false positive values are less for SVM than when compared to logistic regression which is important for the company. If the false negatives are more the company is not losing any money but if the false positives are higher the cost to the company is high. The Recall is the percentage of positive cases that were caught. F1 score is the percentage of positive predictions that were correct. We can see that these values are higher for SVM which is beneficial for the company. As stated above in the 1st point, the company must focus on why the newer customers are not subscribing, what changes they have made recently that could have brought about this change. The company can focus on adding content for older people and target their advertising towards making or getting people to subscribe rather than buying from stores.

**References**

1. Python – Replace Missing Values with Mean, Median & Mode [October 3, 2021](https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/) by [Ajitesh Kumar](https://vitalflux.com/author/vitalflux/" \o "View all posts by Ajitesh Kumar) [Python - Replace Missing Values with Mean, Median & Mode - Data Analytics (vitalflux.com)](https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/#:~:text=1%20You%20can%20use%20central%20tendency%20measures%20such,be%20used%20on%20Dataframe%20for%20finding%20their%20values.)

**Appendix A**

***Table 1:***

***Graphical user interface, text, application

Description automatically generated***

***Table 2:***

***Graphical user interface, application

Description automatically generated***

***Graph 1:***

***Chart, histogram

Description automatically generated***

***Graph 2:***

***Chart

Description automatically generated with medium confidence***

***Graph 3:***

***Chart, box and whisker chart

Description automatically generated***

***Graph 4:***

***Chart, histogram

Description automatically generated***

***Result 1:***

***Table

Description automatically generated***

**Table

Description automatically generated**

***Result 2:***

***A screenshot of a computer

Description automatically generated with medium confidence***

***Result 3:***

***A screenshot of a computer

Description automatically generated with medium confidence***