Machine Learning Engineer Nanodegree

Capstone report

Basmah Alabdullatif

12/28/2021

Predict Responses to Marketing Campaign

Overview:

Marketing campaigns have become an important tool for companies to attract new customers. A well-planned campaign is essential to reduce expenses and get a good marketing ROI at the same time. Therefore, companies need a successful and cost-efficient campaign strategy to help companies stay competitive and gain more market shares.

Problem statement:

Spending the marketing budget on the wrong customers (customers less likely to purchase products) will waste the companies money.

Having previous campaigns data, and how customers acted to them, we trained a model to classify whether a customer is likely to engage in our current campaign or not, to head the marketing effort in the right direction.

Dataset:

This data-set from Kaggle contains information of 2240 customers with 29 columns including demographics, buying habits, and how they acted to previous campaigns, here are the features:

1- ID: Unique identification code for each customer
2- Year_Birth: The DOB Year of the customer
3- Education: Customer's level of education.
4- Marital_Status: Customer's status of Marriage
5- Income: Customer's annual Income
6- Kidhome: Number of children under 13 in Customer's house
7- Teenhome: Number of children between 13-19 in Customer's house
8- Dt_Customer: Date of customers enrollment
9- Recency: Number of days since last purchase
10- MntWines: the amount of money spent on Wines in the last 2 years.
11- MntFruits: the amount of money spent on Fruits in the last 2 years.
12- MntMeatProducts: the amount of money spent on Meat products in the last 2 years.
13- MntFishProducts: the amount of money spent on Fish products in the last 2 years.
14- MntSweetProducts: the amount of money spent on Sweet products in the last 2 years.
15- MntGoldProds: Dollar the amount of money spent on Gold products in the last 2 years.
16- NumDealsPurchases: Number of purchases made with a discount.
17- NumWebPurchases: Number of purchases made through the company's website.
18- NumCatalogPurchases: Number of purchases made using the catalog.
19- NumStorePurchases: Number of purchases made directly in-store.
20- NumWebVisitsMonth: Number of visits made through the company's website.
21- AcceptedCmp1: 1 if the customer accepted the offer in the 1st campaign, 0 otherwise.
22- AcceptedCmp2: 1 if the customer accepted the offer in the 2nd campaign, 0 otherwise.
23- AcceptedCmp3: 1 if the customer accepted the offer in the 3rd campaign, 0 otherwise.
24- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise.
25- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise.
26- Complain: 1 if the customer complained in the last 2 years, 0 otherwise.
27- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise.
28- Z_CostContact cost spent on this customer (for marketing).
29- Z_Revenue revenue gained from this customer.

Data-set link: https://www.kaggle.com/rodsaldanha/arketing-campaign

Solution statement:

We trained machine learning models using previous data to predict which customer is worth spending the marketing effort on. This is a classification problem that aims to classify each customer as (will respond, will not respond) based on their personal data, demographics, buying habits, and how they acted to previous campaigns.

Benchmark model:

To assess how our final model is performing, we compared the results of the final model with the following benchmark: Data Analysis Campaign project by Rodolfo Saldanha using deep learning (Keras) which gave a 0.867 accuracy score. More details can be found here:

https://www.kaggle.com/rodsaldanha/data-analysis-campaign

After testing different algorithms we decided to use a logistic regression algorithm that outperformed the benchmark score.

Evaluation metrics:

As this is a classification problem, we calculated the following metrics to evaluate the algorithms:

- **Accuracy:**

accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.

$$\frac{True\ predections}{Total} = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Precision:**

Precision evaluates the fraction of correct classified instances among the ones classified as positive.

$$\frac{True\ predection}{Classified\ as\ positive} = \frac{TP}{TP+FP}$$

- **Recall:**

quantifies the number of positive class predictions made out of all positive examples in the dataset

$$\frac{positive\ predection}{all\ actual\ positive} = \frac{TP}{TP+FN}$$

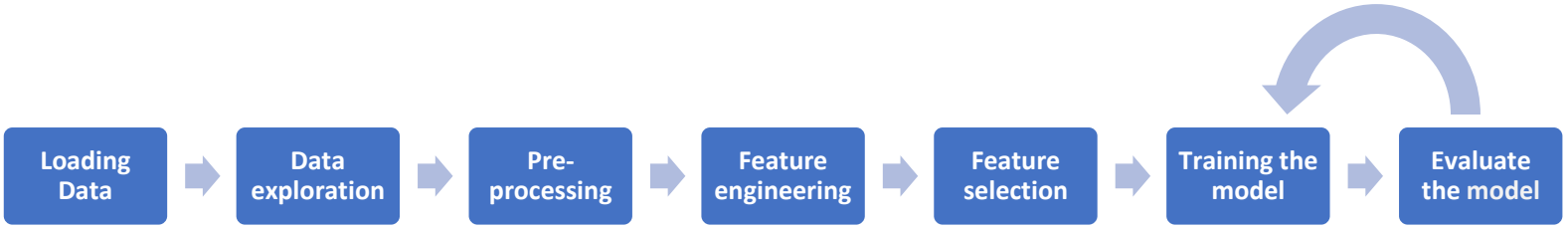
- **Confusion matrix:**

Confusion matrix summarize the performance of the classification algorithm.

		True class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

Project design:

We used Python 3.6 in this project. We followed the steps below to implement our machine learning model:



- **Libraries used:**
  - numpy
  - pandas
  - seaborn
  - matplotlib
  - collections
  - imblearn
  - sklearn
  - datetime

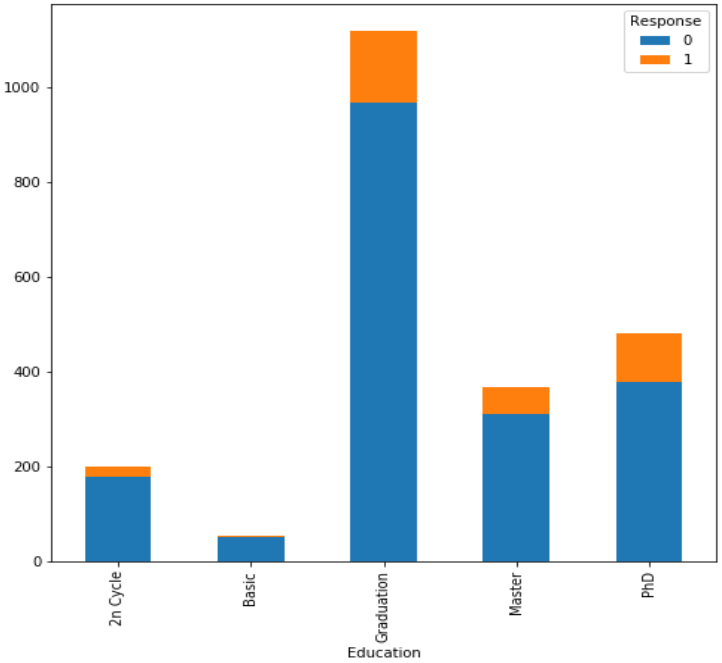
1. **<u>Loading Data:</u>**

This data is available on Kaggle as a .csv file, we simply load it to our workspace as pandas data frame.

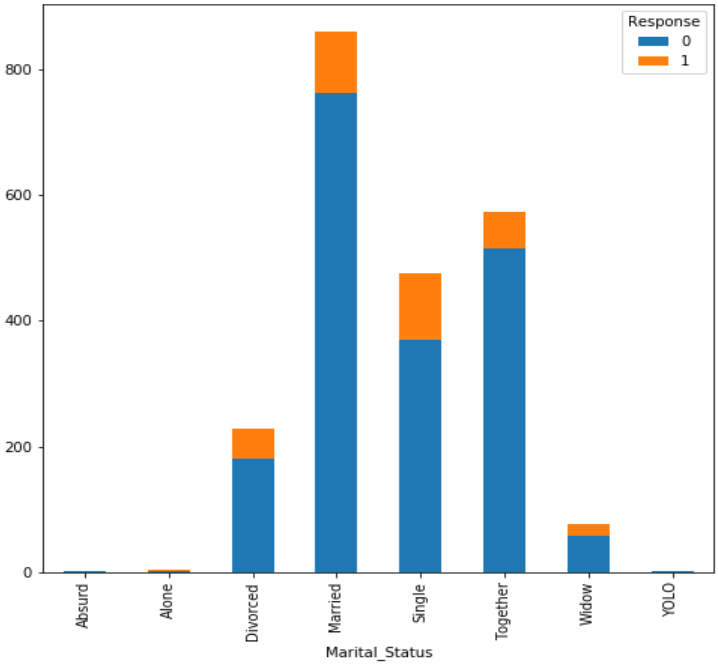
2. Data exploration:

After loading the data, EDA has been done to understand the data, find data patterns, trends and extract useful insights to select features. Here are some interesting observations we found:

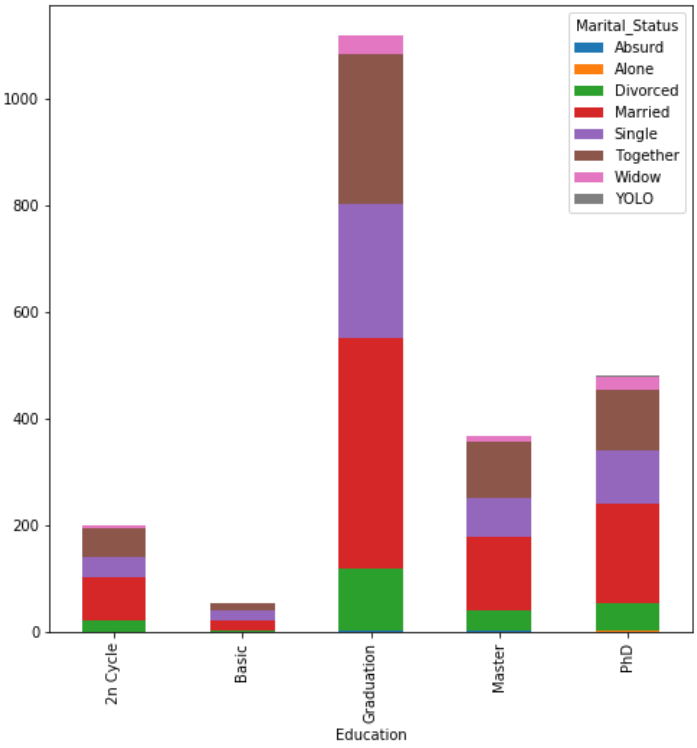
First, we plot the responds from each educational level. We can see that most of the people responding to the current campaign have graduation degree. However, greater percentage of positive responses came from people with phD.



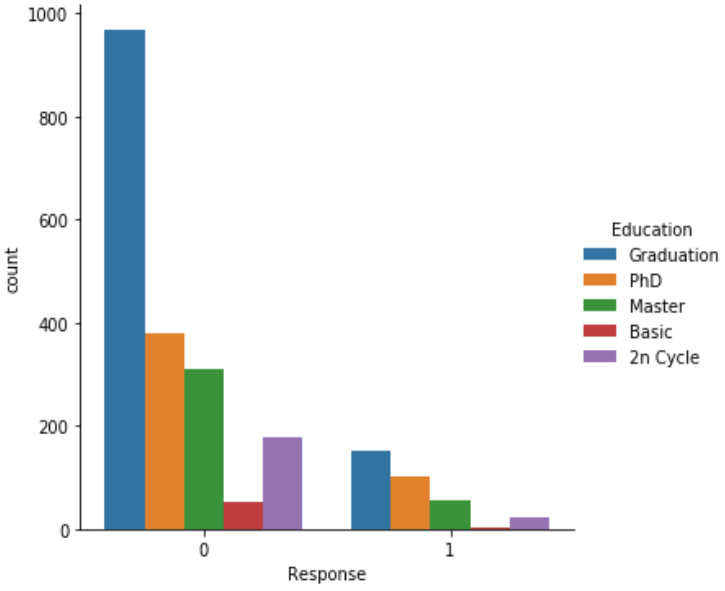
We also plot the responds from each material status. We can see that single people have the most positive responses.



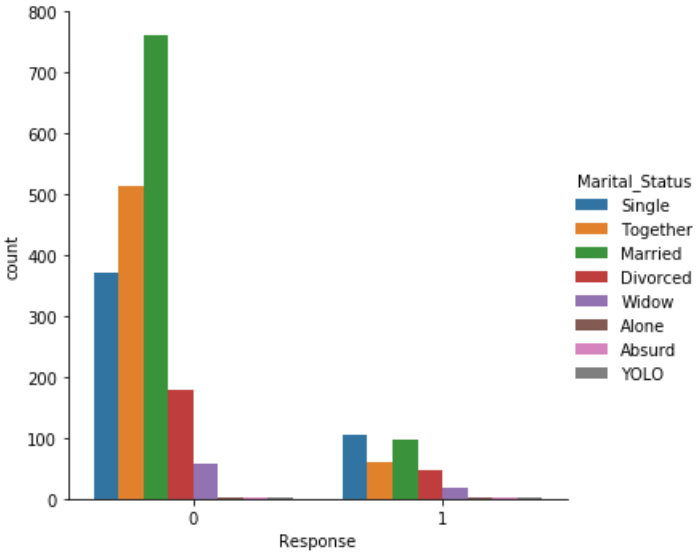
From the following bar chart, we can see that all educations levels have almost same proportion from all marital status.



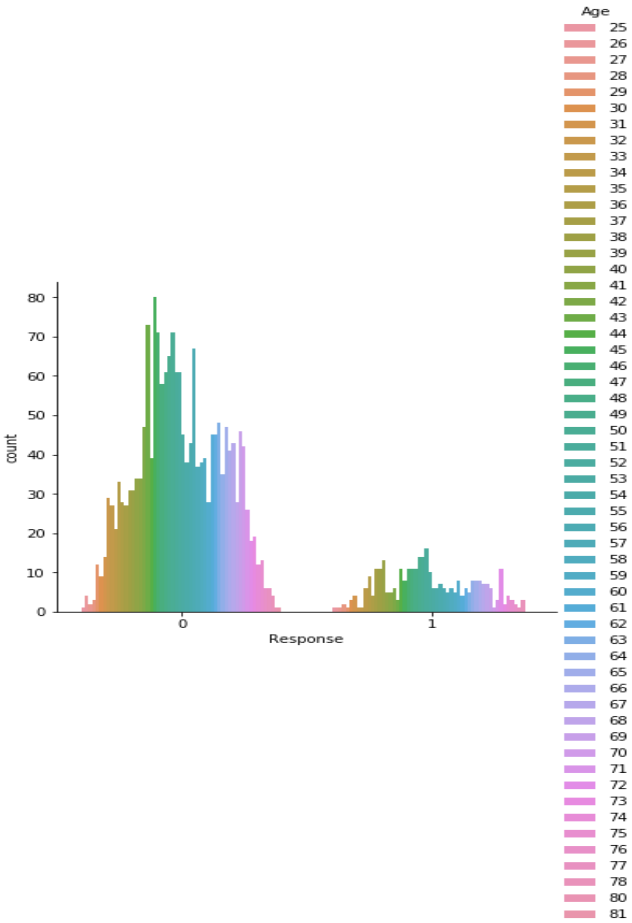
Then, we investigate the responses counts among different features. We can conclude from the following chart that the count of positive responses from master's and PhD holders is relatively high since it didn't drop sharply as graduation counts did. However, graduation holders are the majority so they have the highest count in both responding and non-responding.



We can also see from the following chart that the count of positive responses from single people is the highest. Although marred and together are the majority, their positive responses are lower than those from single people.



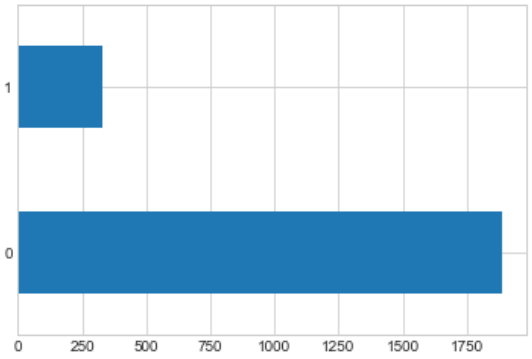
When it comes to the top age range to respond to the campaign, the following chart shows that people in their 50's have the highest responding counts.



3. Pre-processing:

Over sampling:

When we plot the count of each class (as shown below), we can see that our dataset is highly unbalanced. To overcome this issue, we apply oversampling using SMOTE. The majority class had 1888 samples while the other class had only 332 samples, after over sampling, both classes have 1888 sample for each.



Encoding categorical data:

We only have 2 categorical data: educational levels, and marital status. To encode marital status, we used one hot encoding. Since educational levels are ordinal, we map them as follows:

Basic	1	2n Cycle	2	Graduation	3	Master	4	PhD	5

Imputation:

There is only missing data in the income column (24 null values) that we decided to replace with the mean of the same education level.

Outliers:

we plot boxplots for each numerical column to find the outliers. We got some outliers in the following columns:

Income, Age, Spending, AverageCheck, ShareDealsPurchases

We delete all nonsense datapoints.

Scaling:

We normalize our data using MinMaxScaler.

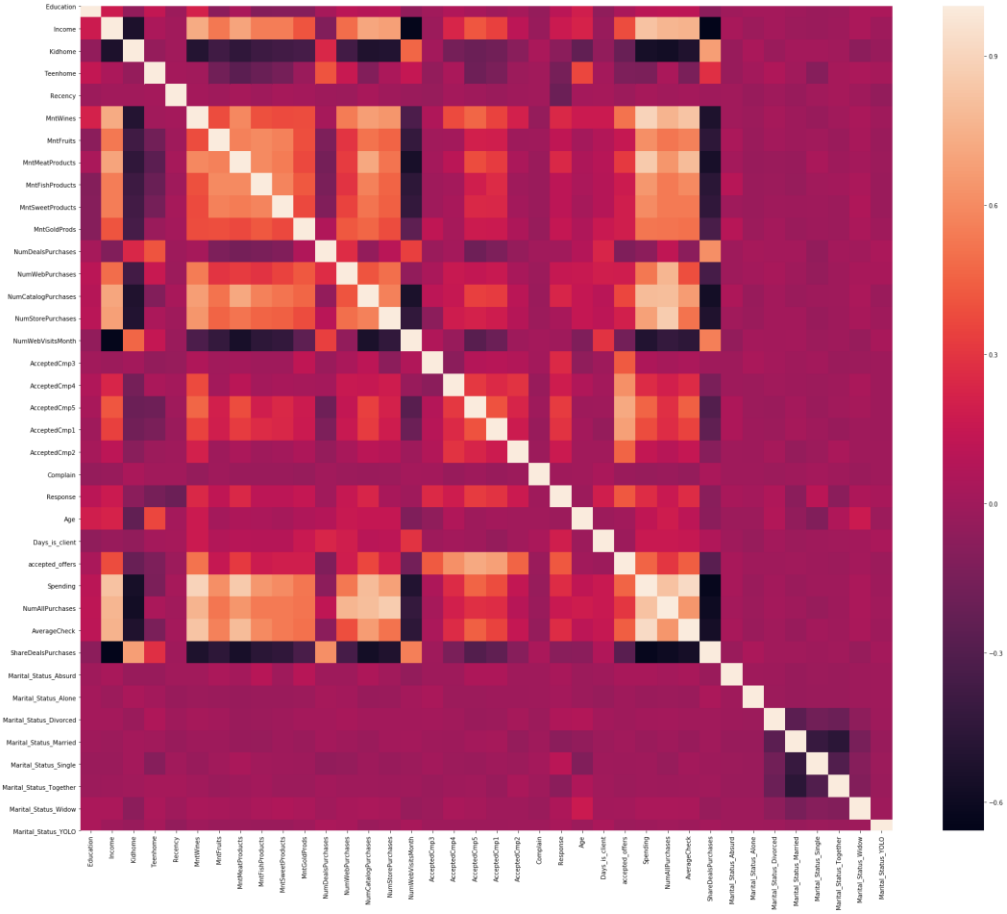
4. Feature engineering

we added new useful columns extracted from available ones. This include:

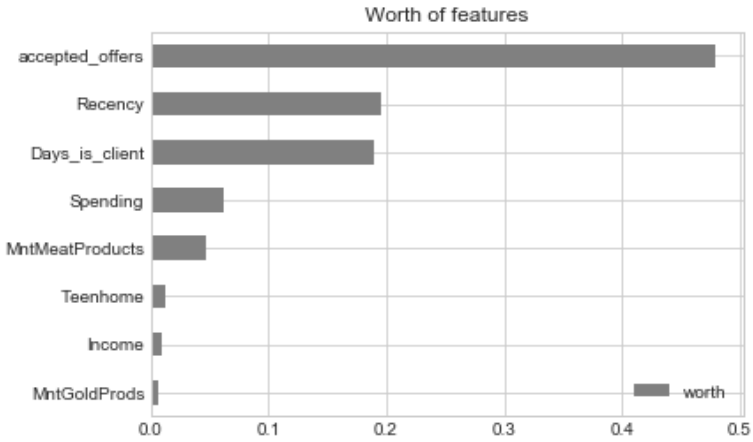
- Age from Year_Birth
- number of days the customer is enrolled with the company from Dt_Customer
- Number of offers accepted from the past campaigns.
- total number of purchases from all channels
- total Spending in all products
- total accepted offer
- Average Check
- ratio of DealsPurchases out of all purchases

5. Feature selection:

To select the most relevant features that will be useful for training the model we plot the correlation heat map :



Then we plot the top features that correlated with our target:



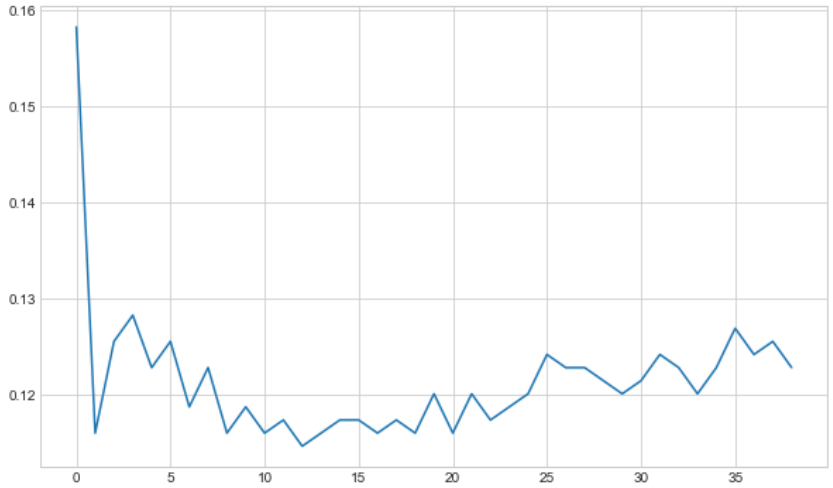
We select these features but exclude MntMeatProducts, and MntGoldProds because they are highly corelated with Spending.

6. Training the model

After splitting the data to train and test, we trained the model using several algorithms from sklearn library (after tuning their parameters):

a. KNN

Tuning hyperparameters:



Using elbow method, best **K** has chosen = 8
Training cross validation score **0.8729012722301313**

b. SVM:

After gridsearch we chose: 'C' = 10, 'gamma' = 1.438449888287663, 'kernel' = 'rbf'
Training cross validation score **0.881630623912503**

c. Decision trees

After gridsearch, we chose max_depth': 4
Training cross validation score **0.8634917406729488**

d. AdaBoost:

Parameters: max_depth=3 ,n_estimators=200
Training cross validation score **0.8487063023975775**

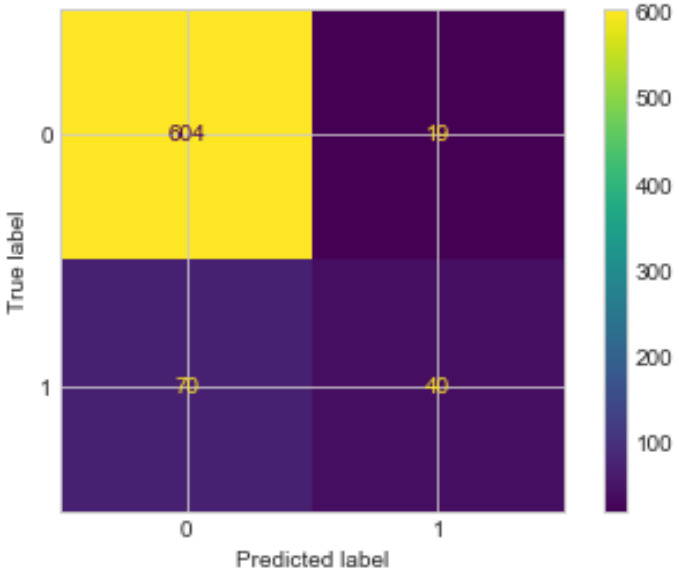
e. Logistic regression

After gridsearch we chose:'C' = 0.615848211066026, 'penalty' = 'l2', 'solver' ='liblinear'
Training cross validation score **0.8802951212347185**

7. Evaluate the model

The best algorithm in terms of training accuracy was **Logistic Regression**. We evaluated this algorithm using different measures:

	precision	recall	f1-score
0	0.89	0.98	0.93
1	0.75	0.35	0.47



Our algorithm has a poor recall score, that is, the ability to correctly predict positive responses is very low. Out of 110, it can only predict 40. However, it has fair precision and high accuracy.

Conclusion:

In this project, we analyzed customers' data set that includes demographic information, buying habits, and how they acted to previous campaigns. Our goal is to train a machine learning model to predict which customer is worth spending the marketing effort on. After evaluating many algorithms, we chose logistic regression with accuracy of 88%. This model also has High precision, recall and f1 score in terms of target value 0. which means we can exclude a lot of people that the model predicts as not worth the marketing effort. However, low recall and f1 score in terms of target value 1 means that the model misses a lot of positive responses.