INFO-659-900

Final Project

Predicting the Response to a Marketing Drip Campaign

Group members:
Alexandrea Morson
Lauren Tagliaferro
Maopin Yan
Nicole Buccigrossi

December 13, 2021

Table of Contents

1. Summary

1.1 Objective

The main objective of this project is to train a predictive model which will allow the company to maximize the profit of its next marketing campaign. A response model can provide a significant boost to the efficiency of a marketing campaign by increasing responses as well as reducing expenses. Therefore, this project is designed to predict who will respond to an offer for a product or service.

1.2 Data Source

Data Source: https://www.kaggle.com/rodsaldanha/arketing-campaign
Format: The data source is available in both .csv and .xlsx formats.

1.3 Features Description – Lauren/Ally

There are a total of 29 variables in this dataset, and "Response" will be regarded as the predicted variable. We omitted 3 variables: "ID", "Year_Birth", and "Dt_Custmer" while doing the visualization and modeling because these three variables will not influence the predicted results.

Following are the descriptions of the variables in this dataset.

Kidhome: Number of small children in customer's household.
Teenhome: Number of teenagers in customer's household.
Income: Customer's yearly household income.
MntFishProducts: Amount spent on fish products in the last 2 years.
MntMeatProducts: Amount spent on meat products in the last 2 years.
MntFruits: Amount spent on fruits products in the last 2 years.
MntSweetProducts: Amount spent on sweet products in the last 2 years.
MntWines: Amount spent on wine products in the last 2 years.
MntGoldProds: Amount spent on gold products in the last 2 years.
NumDealsPurchases: Number of purchases made with a discount.
NumCatalogPurchases: Number of purchases made using catalog.
NumStorePurchases: Number of purchases made directly in stores.
NumWebPurchases: Number of purchases made through the company's website.
NumWebVisitsMonth: Number of visits to the company's website in the last month.
Recency: Number of days since the last purchase.

2. Data Preparation

2.1 Importing Packages

```
library(ggplot2)
library(purrr)
library(tidyr)
library(caret)
library("e1071")
```

2.2 Read dataset and data clean

```
cc <- read.csv("marketing_campaign.csv")
head(cc)
```

Description: df [6 × 29]

	ID <int>	Year_Birth <int>	Education <chr>	Marital_Status <chr>	Income <int>	Kidhome <int>	Teenhome <int>	Dt_Customer <chr>	Recency <int>
1	0	1985	Graduation	Married	70951	0	0	2013-05-04	66
2	1	1961	Graduation	Single	57091	0	0	2014-06-15	0
3	9	1975	Master	Single	46098	1	1	2012-08-18	86
4	13	1947	PhD	Widow	25358	0	1	2013-07-22	57
5	17	1971	PhD	Married	60491	0	1	2013-09-06	81
6	20	1965	2n Cycle	Married	46891	0	1	2013-09-01	91

6 rows | 1-10 of 29 columns

There are a total of 29 variables and 2240 instances in this dataset. We found that there are 24 Nan values in the dataframe, we choose to drop those nan values as they are in a small amount. After we drop the nan values, there are 29 variables and 2216 rows.

```
cc <- na.omit(cc)
dim(cc)
```

```
[1] 2216   29
```

3. Data Transformation

Here we transformed the variables *AcceptedCmp1*, *AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Response, Complain* into a categorical (factor) variable (0 to “No”, 1 to “Yes”):

```
cc$AcceptedCmp1 <- factor(cc$AcceptedCmp1,levels=c(0,1), labels=c("No","Yes"))
cc$AcceptedCmp2 <- factor(cc$AcceptedCmp2,levels=c(0,1), labels=c("No","Yes"))
cc$AcceptedCmp3 <- factor(cc$AcceptedCmp3,levels=c(0,1), labels=c("No","Yes"))
cc$AcceptedCmp4 <- factor(cc$AcceptedCmp4,levels=c(0,1), labels=c("No","Yes"))
cc$AcceptedCmp5 <- factor(cc$AcceptedCmp5,levels=c(0,1), labels=c("No","Yes"))
cc$Response <- factor(cc$Response,levels=c(0,1), labels=c("No","Yes"))
cc$Complain <- factor(cc$Complain,levels=c(0,1), labels=c("No","Yes"))

head(cc)
```

Description: df [6 × 29]

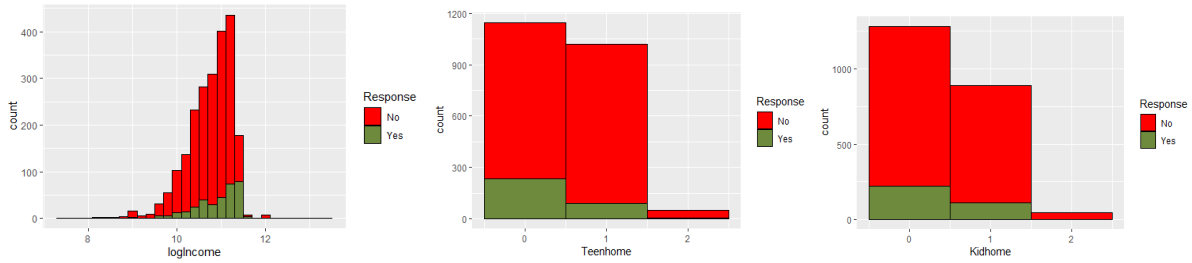
	ID <int>	Year_Birth <int>	Education <chr>	Marital_Status <chr>	Income <int>	Kidhome <int>	Teenhome <int>	Dt_Customer <chr>	Recency <int>
1	0	1985	Graduation	Married	70951	0	0	2013-05-04	66
2	1	1961	Graduation	Single	57091	0	0	2014-06-15	0
3	9	1975	Master	Single	46098	1	1	2012-08-18	86
4	13	1947	PhD	Widow	25358	0	1	2013-07-22	57
5	17	1971	PhD	Married	60491	0	1	2013-09-06	81
6	20	1965	2n Cycle	Married	46891	0	1	2013-09-01	91

6 rows | 1-10 of 29 columns

4. EDA and Data Visualization

In the Visualization part, we will not show the codes. We will only show the plots we made and give some explanations about the relationship between “Response” and the other variables. The codes are in the .html file in the attachments.

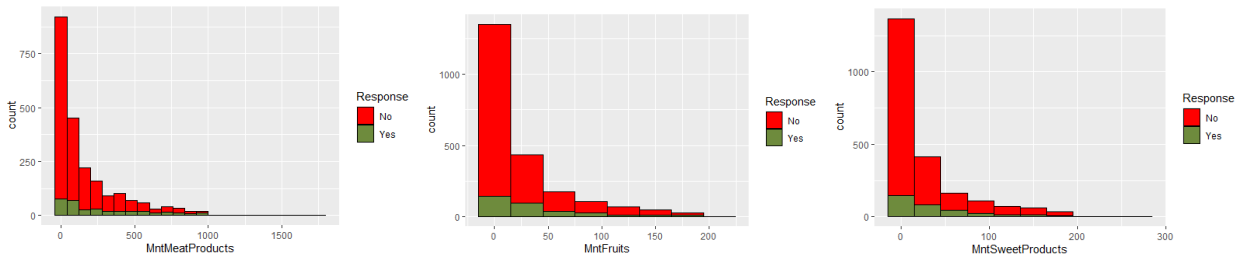
a) Response & LogIncome



b) Response & Teenhome

c) Response & Kidhome

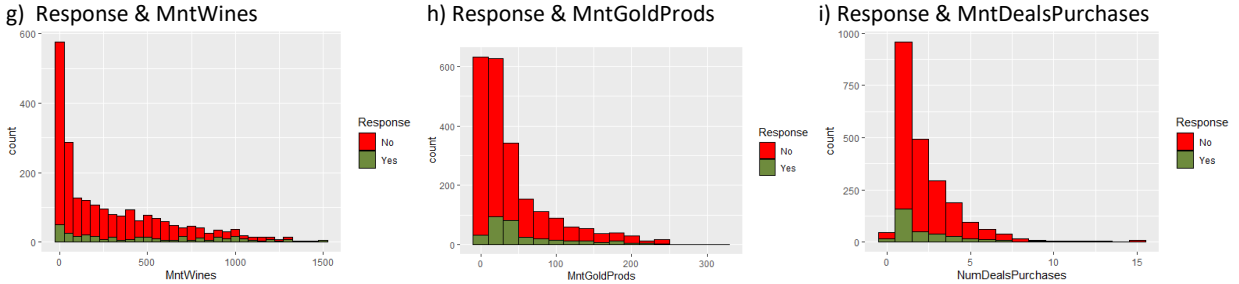
d) Response & MntMeetProducts



e) Response & MntFruits

f) Response & MntSweetProducts

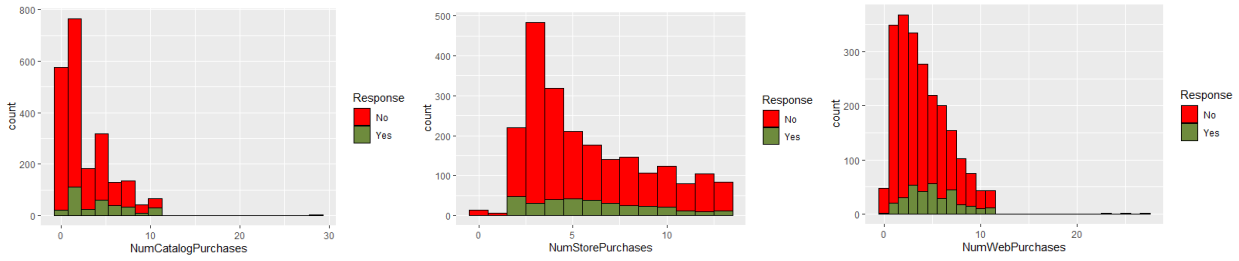
g) Response & MntWines



h) Response & MntGoldProds

i) Response & MntDealsPurchases

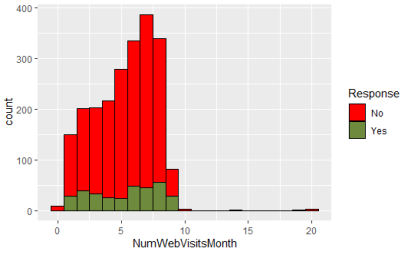
j) Response & NumCatalogPurchases



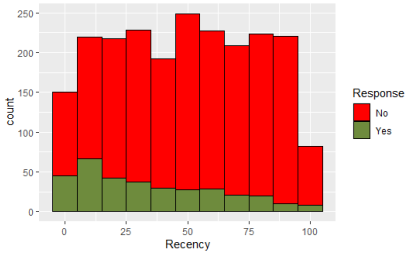
k) Response & NumStorePurchases

l) Response & NumWebpurchases

m) Response & NumWebVisitsMonth



n) Response & Recency



The following are the descriptions of the plots above.

a) We found that with the increase of income, there is a higher probability of response "yes".
b) Most customers in the dataset have either 0 or only 1 teen in their household. We found that having a teen in the household decreases the likelihood of the response yes.
c) Most customers in the dataset have either 0 or only 1 small child in their household. We found that having small children in the household decreases the likelihood of the response yes.
d) There is a large variance in the dataset between the amount spent on meat products in the last 2 years and response. Most customers in the dataset spent 0 on meat products in the last 2 years. Overall, the amount spent on fruit products is not a good predictor of response.
e) There is a large variance in the dataset between the amount spent on fruit products in the last 2 years and response. Most customers in the dataset spent 0 on fruit products in the last 2 years. Overall, the amount spent on fruit products is not a good predictor of response.
f) There is a large variance in the dataset between the amount spent on sweet products in the last 2 years. Most customers in the dataset spent 0 on sweet products in the last 2 years. Overall, the amount spent on sweet products is not a good predictor.
g) There is a slight variance in the dataset between the amount spent on wine products in the last 2 years. Most customers in the dataset made 5 purchases of wine products in the last 2 years.
h) There is not much variance in the dataset between the amount spent on gold products in the last 2 years. A large amount customers in the dataset made 0- 5 purchases of gold products in the last 2 years.
i) We found that of customers who made purchases with a discount, the most frequent number of purchases with a discount is just 1. There are only minor differences in response ratios of a different number of purchases made with a discount.
j) We found that customers who made 10 or more purchases using a catalog are likely to respond yes, however, the majority of customers made either 0 or 1 purchase using a catalog.
k) There are only minor differences in response ratios of a different number of purchases made directly in stores. However, it is worth noting that among customers who have made 1 or fewer purchases in-store, the response ratio is 0.
l) We found that are only minor differences in response ratios between a different number of purchases through the company's website. The majority of customers in the dataset made between 1 and 5 web purchases.
m) We found that customers who made 9 visits to the company's website in the last month were most likely to respond yes. Other than this group of customers, there are only minor differences in response ratios.
n) We found that the fewer number of days since the last purchase made customers more likely to respond yes, however, there is a large variance in the dataset of the number of days since the last purchase and response.

Methodologies

4.1 Methods

Logistic Regression

Logistic regression is a "supervised machine learning classifier that extracts real-valued features from the input, multiplies each by weight, sums them, and passes the sum through a sigmoid function to generate a probability. A threshold is used to make a decision" [1]. The reason why we selected this methodology is this dataset is going the predict "Yes" or "No", and the probability of getting "Yes" or "No".

Support Vector machines (SVMs)

SVMs is an unsupervised learning approach, it attempts to find natural clustering of the data to groups, and map new data to these formed groups [4]. In this dataset, as the data are not linearly separable, we choose kernel = "radial" for the separation on the non-linear dataset. The reason why we selected this methodology is this method works well on classification and can group different kinds of data.

P-values

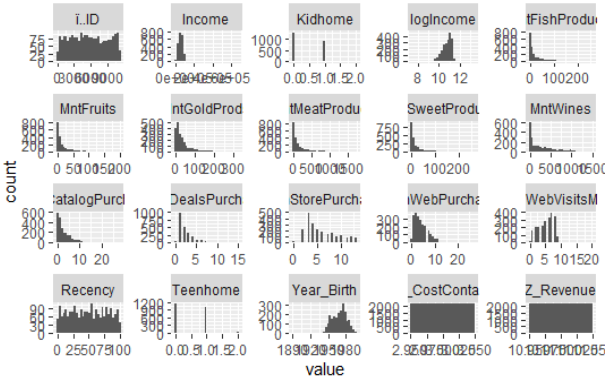
In this project, we used p-values under the logistic regression model to select features. Features with small p-values mean the results are significant [2].

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.493e+00	4.129e-01	-6.038	1.56e-09	***
Income	2.182e-06	2.954e-06	0.739	0.46014	
Kidhome	1.706e-01	2.081e-01	0.820	0.41247	
Teenhome	-1.039e+00	1.939e-01	-5.357	8.47e-08	***
Recency	-2.548e-02	2.826e-03	-9.015	< 2e-16	***
MntWines	1.490e-03	2.993e-04	4.978	6.43e-07	***
MntFruits	-4.874e-04	2.244e-03	-0.217	0.82801	
MntMeatProducts	1.969e-03	4.934e-04	3.991	6.58e-05	***
MntFishProducts	-1.610e-03	1.697e-03	-0.949	0.34262	
MntSweetProducts	1.774e-03	2.124e-03	0.835	0.40346	
MntGoldProds	3.140e-03	1.478e-03	2.124	0.03365	*
NumDealsPurchases	5.981e-02	4.783e-02	1.250	0.21117	
NumWebPurchases	1.028e-01	3.356e-02	3.062	0.00220	**
NumCatalogPurchases	1.086e-01	3.933e-02	2.763	0.00573	**
NumStorePurchases	-1.680e-01	3.278e-02	-5.124	3.00e-07	***
NumWebVisitsMonth	1.940e-01	4.520e-02	4.292	1.77e-05	***

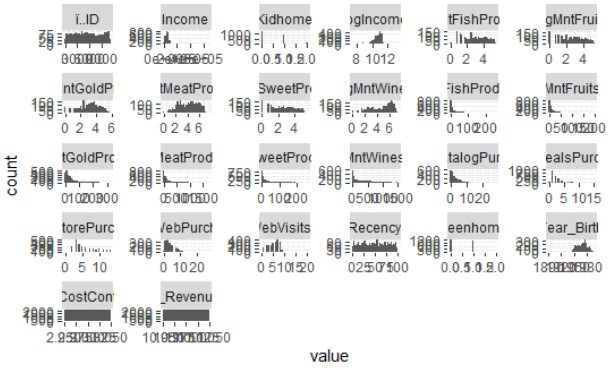
P-values feature selection

Logarithm Transformation

Log transformation reduces or removes the skewness of the original data, it works on those original data following a log-normal distribution [3]. In this project, some features are in the right skewness distribution, after the logarithm transformation, they turned into a relatively normal distribution, see the graphs below, left is before log transformation and right is after log transformation.



Features before log transformation



Features after log transformation

5.2 Evaluation Metrics

In the following equations, TP stands for True Positive, FP stands for False Positive, TN stands for True Negative, FN stands for False Negative in the confusion matrix.

Accuracy

Accuracy represents the number of correctly classified data instances over the total number of data instances. Accuracy may not be a good measure if the dataset is not balanced (both negative and positive classes have a different number of data instances) [5]. Hence, we calculated the other metrics as well in order to evaluate the performance of the classifiers. Accuracy can be represented as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

F1-score

F1 is the harmonic mean of precision and recall. F1 is useful in situations where the analysis views precision and recall as equally valuable. The formula for calculating f1 is as follows:

$$\text{F} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}}$$

Precision

How close measured values are to each other. Measurements can be accurate but not precise and vice versa. Precision can be calculated by dividing true positive by the sum of the true positive and false negative as below:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

The recall is the number of values that were classified correctly. The recall values can be calculated by dividing the true positives by the sum of the true positives and false negatives as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Cohen's Kappa

Cohen's kappa statistic is a very good measure that can handle very well both multi-class and imbalanced class problems. As the dataset in our project is an imbalanced dataset, Cohen's Kappa would be a good evaluation indicator for the prediction results. The Cohen's Kappa equation is as follows:

$$K = \frac{2 * (TP * TN - FN * FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)}$$

5. Modeling

5.1 SVMs on the original dataset

We are going to try how SVM classifier trained on the dataset without features selection. We are going to try cost = 1, and gamma = 0.5 on this model.

```
svm_model <- svm(Response ~ Income+Kidhome+ Teenhome+ Recency+ MntWines+ MntFruits+ MntMeatProducts+ MntFishProducts+ MntSweetProducts+ MntGoldProds+ NumDealsPurchases+ NumWebPurchases+ NumCatalogPurchases+ NumStorePurchases+ NumWebVisitsMonth+ Z_CostContact+ Z_Revenue, Train_new, kernel="radial", cost=1, gamma=0.5)
```

```
confusionMatrix(prediction, Test_new$Response, dnn = c("Prediction", "Truth"))
```

Accuracy of Model Prediction: 0.862

The results are very good, except the Kappa value is low. The Accuracy is 0.862, the precision is 0.992, the recall is 0.8654 from the confusion matrix.

5.2 Logistic Regression Model on the original dataset.

We will train the Train_new dataset use all the features, and comparing the confusion matrix out with the new dataset after features selection.

```
default_lr_model_all <- glm(Response ~ Income+Kidhome+ Teenhome+ Recency+ MntWines+ MntFruits+ MntMeatProducts+ MntFishProducts+ MntSweetProducts+ MntGoldProds+ NumDealsPurchases+ NumWebPurchases+ NumCatalogPurchases+ NumStorePurchases+ NumWebVisitsMonth+ Z_CostContact+ Z_Revenue, data = Train_new, family = "binomial")

pred_lr_all <- predict(default_lr_model_all, newdata = Test_new, type = "response")
pred_lr_class_all <- ifelse(pred_lr_all >= 0.3, 1, 0)

results <- confusionMatrix(data=factor(pred_lr_class_all, levels=0:1),reference = factor(Test_new$Response, levels=0:1), positive = "1")
print(results)
```

Accuracy of Model Prediction: 0.8416

5.3 Logistic Regression on selected features based on p-value selection

We selected 7 features with the lowest p-values: Teenhome, Recency, MntWines, MntMeatProducts, NumStorePurchases, NumWebVisitsMonth, NumWebPurchases. And use the logistic model train on the new dataset.

```
default_lr_model <- glm(Response ~ Teenhome+ Recency+ MntWines+ MntMeatProducts+ NumWebPurchases+ N
umStorePurchases+ NumWebVisitsMonth, data = Train_new, family = "binomial")

pred_lr <- predict(default_lr_model, newdata = Test_new, type = "response")
red_lr_class <- ifelse(pred_lr >= 0.3, 1, 0)

presults <- confusionMatrix(data=factor(pred_lr_class, levels=0:1),reference = factor(Test_new$Response, levels=
0:1), positive = "1")
```

Accuracy of Model Prediction: 0.8349

5.4 SVMs on the new dataset (after log transformation)

We transformed from Inf to 1, and -Inf to 0 (as 0 is the smallest value in the columns).

```
cc_new[cc_new == 'Inf'] <- 1
cc_new[cc_new == '-Inf'] <- 0
cc_new[is.na(cc_new)] <- 1
```

```
set.seed(0)
index_log<-createDataPartition(factor(cc_new$Response), p=0.8, list=F)
Train_log<-cc_new[index_log,]
nrow(Train_log)

svm_model_log <- svm(Response ~ Kidhome + Teenhome + Recency+ NumDealsPurchases +NumWebPurchases +
NumCatalogPurchases + NumStorePurchases + NumWebVisitsMonth +
logMntFishProducts+logMntMeatProducts+logMntWines + logMntFruits + logMntSweetProducts +
logMntGoldProds, Train_log, kernel="radial", cost=1, gamma=0.5)
```

```
prediction_log <- predict (svm_model_log, Test_log)
confusionMatrix(prediction_log, Test_log$Response, dnn = c("Prediction", "Truth"))
```

Accuracy of Model Prediction: 0.8597

6. Results

	Methods	Evaluation Metrics				
		Correct %	Precision	Recall	F	Kappa
1	**SVM**	**0.862**	**0.992**	0.865	**0.9242**	0.1725
2	**Logistic Regression** Variables: all	0.8416	0.50	0.4714	0.4853	0.3918
3	**Logistic Regression** Variables: p-values feature selection	0.8349	0.4091	0.4576	0.432	0.3388
4	**SVM** Variables: After log transformation	0.8597	0.9814	**0.8703**	0.9225	0.2114

7. Major Challenges and Solutions

a) For the SVMs model, the predicted variable "Response" should be transformed into categorical values, e.g. "Yes" or "No". However, for Logistic Regression Model, the predicted variable "Response" should keep in numerical value 0 or 1 in the evaluation matrix.
b) After the log transformation, there are some Inf and -Inf values in the dataset. Based on the characteristics of these data and the features, we must transform those Inf to 1 and -Inf to 0. Without transformations, the SVMs model does not work well.
c) This dataset is imbalanced, as most of them are "No". Thus, we choose the logistic regression model, which is not sensitive to the imbalanced dataset.
d) The 0 and 1 in the confusion matrix under Prediction and Truth are inversed, so we have to take care of the output scores. The sensitivity is the precision, and the Pos Pred Value is the Recall.

8. Conclusions and Future Work

In conclusion, the predicted results are better than what we thought, especially SVMs work very well on our dataset. What we did not expect is the original dataset, which without features selections, got the higher predictive results. In this case, SVMs works better than the logistic regression model. In the future, we could try cross-validation to get the average accuracies under these two models and to see how the accuracies change over time.

References:

[1] V. Singh and S. P. Lal. Digit recognition using single-layer neural network with principal component analysis., 2014. URL 10.1109/APWCCSE.2014.7053842.

[2] Stephanie Glen. F statistic / f value: Simple definition and interpretation, 2021.

URL https://www.statisticshowto.com/probability-andstatistics/f-statistic-value-test/.

[3] Kyaw Saw Htoon, Feb 29, 2020. Log Transformation: Purpose and Interpretation.
URL https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9

[4] Ben-Hur, Asa; Horn, David; Siegelmann, Hava; Vapnik, Vladimir N. ""Support vector clustering" (2001);". *Journal of Machine Learning Research*. **2**: 125–137

[5] Harikrishnan N B. Confusion matrix, accuracy, precision, recall, f1 score.
URL https://medium.com/analytics-vidhya/confusion-matrixaccuracy-precision-recall-f1-score-ade299cf63cd.