Supervised learning- Classification(2)

Parcours Progis

Etudes, Medias, communication, Marketing

Bahareh Afshinpour.

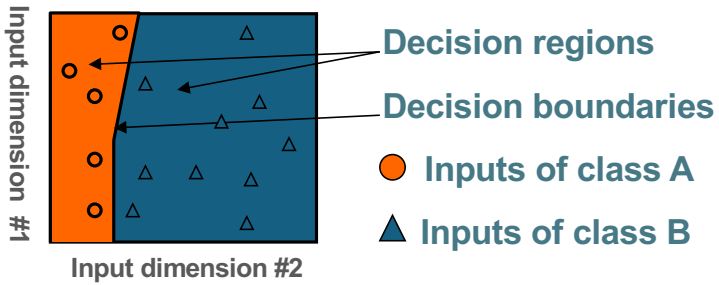
12.01.2026

Université Grenoble Alpes

Sciences Po Grenoble UGA

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

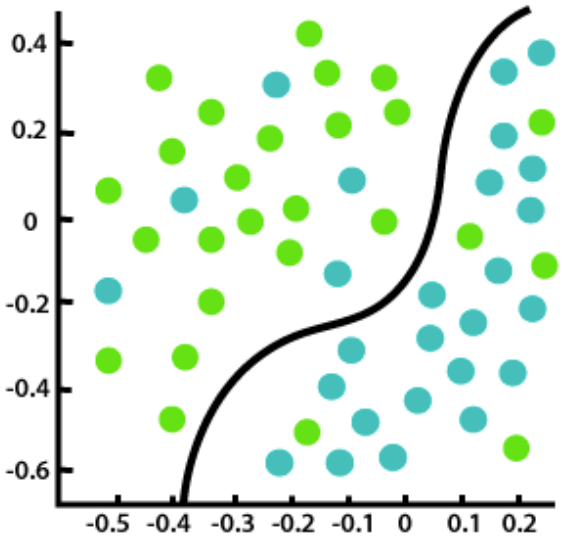
Classification: Terminology



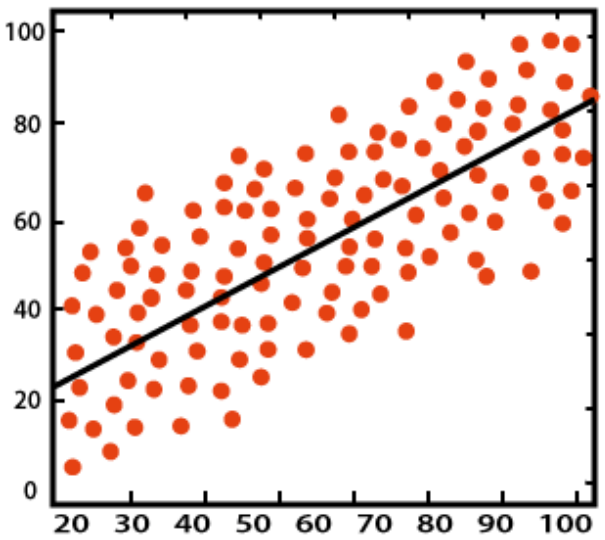
- The way a classifier classifies inputs is defined by its decision regions.
- The borderlines between decision regions are called *decision-region boundaries* or simply ***decision boundaries***.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Classification



Classification

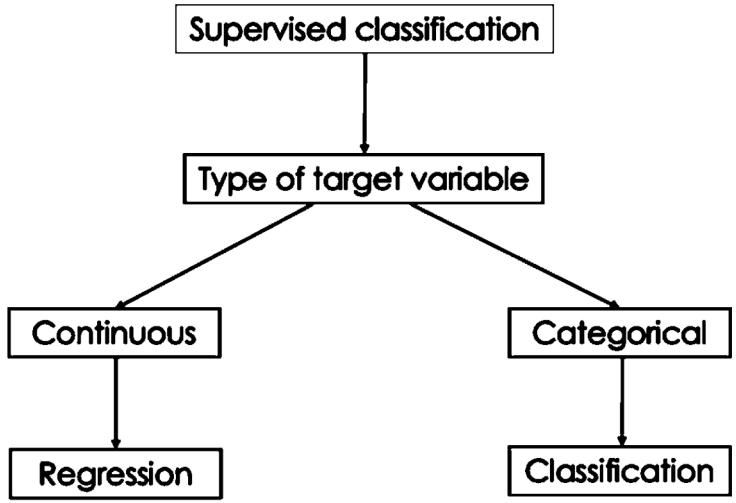


Regression

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

What is Classification in Machine Learning?

- Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data.



https://www.datacamp.com/blog/classification-machine-learning

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

k-Nearest Neighbor (k-NN)

To classify an example *d*:

- Find *k* *nearest* neighbors of *d*
- Choose as the label the majority label within the *k* nearest neighbors

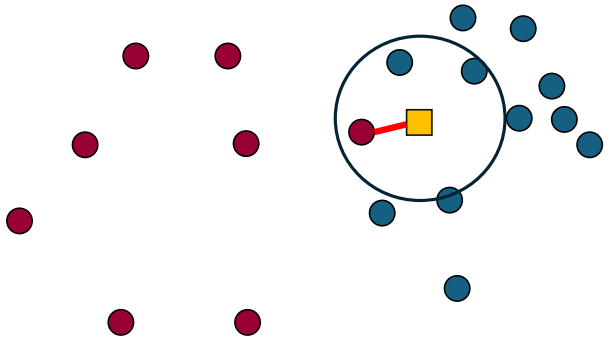
PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

What about this example?

K=3

3-NN

blue

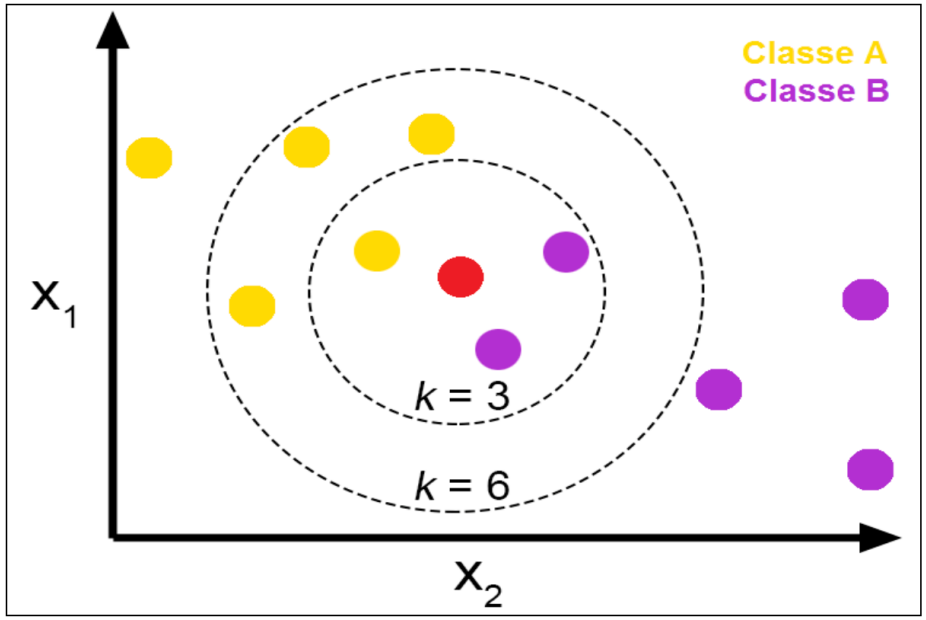


label 1

label 2

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

k-Nearest Neighbor (k-NN)



Example of the use of kNN for classification. if k=1, the test point (in read) is assigned to class B. If k=3, it is assigned to class A.

https://www.cs.ucdavis.edu/~koehl/Teaching/AIX0008/Chapters/Chapter3/notes.html

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

K-Nearest Neighbor (k-NN)

- The k-Nearest Neighbor (kNN) model is an intuitive way to predict a quantitative response variable:

to predict a response for a set of observed predictor values, we use the responses of other observations most similar to it

- kNN is a non-parametric learning algorithm. When we say a technique is non parametric , it means that it does not make any assumptions on the underlying data distribution.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Optimal choice of K

- K should be chosen optimally:
  - If K is too small, sensitive to noise points
  - If K is too large, neighborhood may include points from other classes

- How can we optimize K?
  - Build multiple models with different k
  - Then, evaluate the model

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

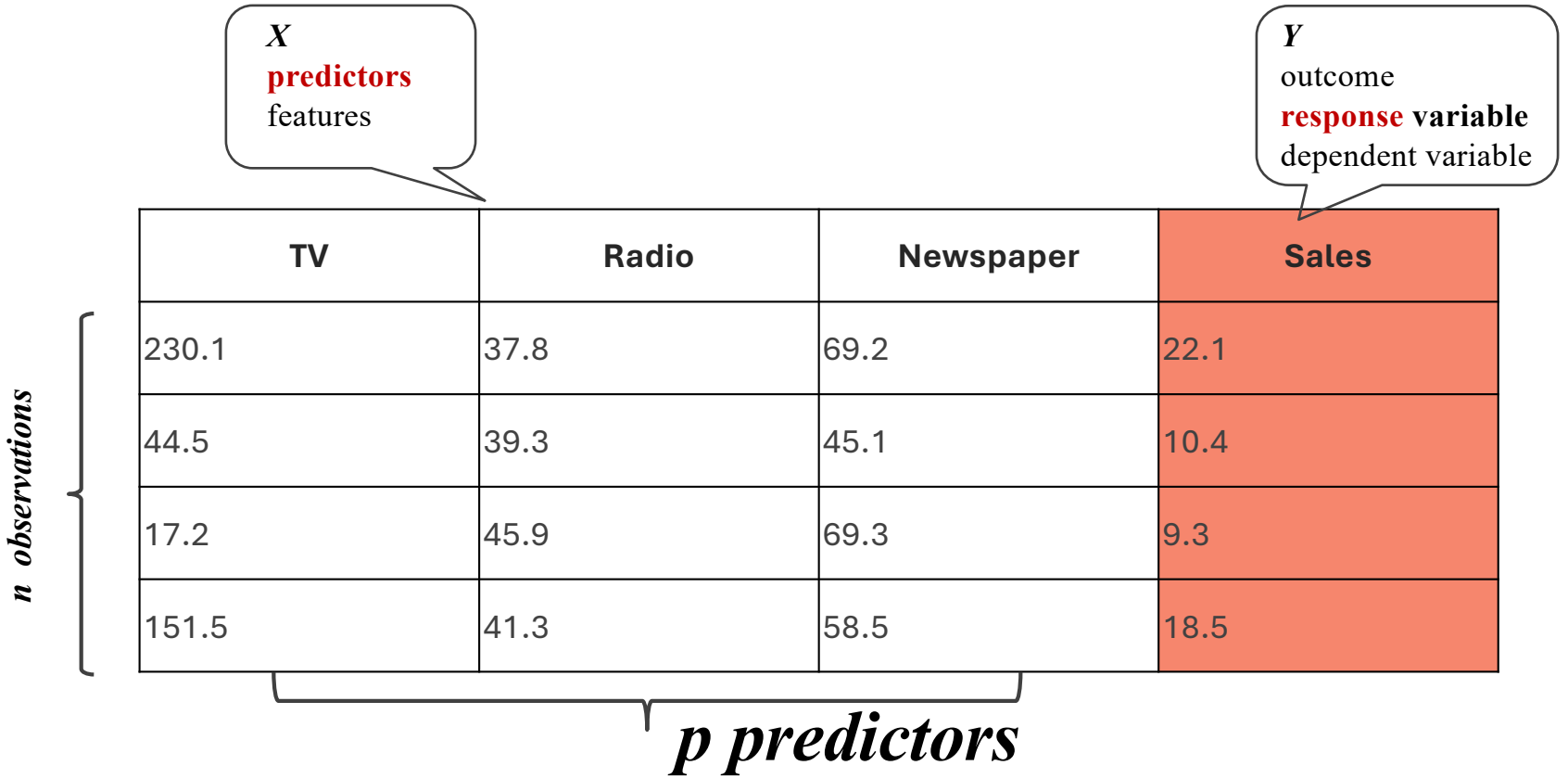
Example

The Advertising data set consists of the sales of a particular product in 200 different markets, and advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. Everything is given in units of $1000.

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5

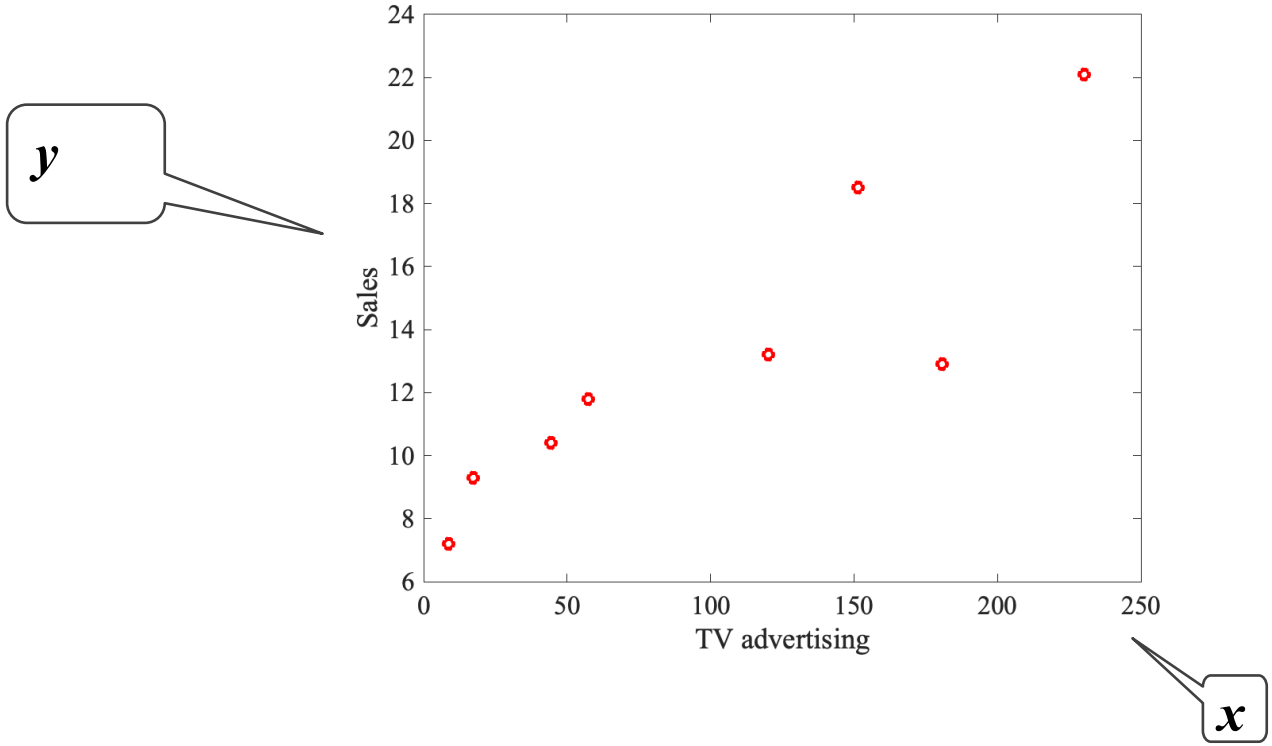
PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Response vs Predictor Variables



PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

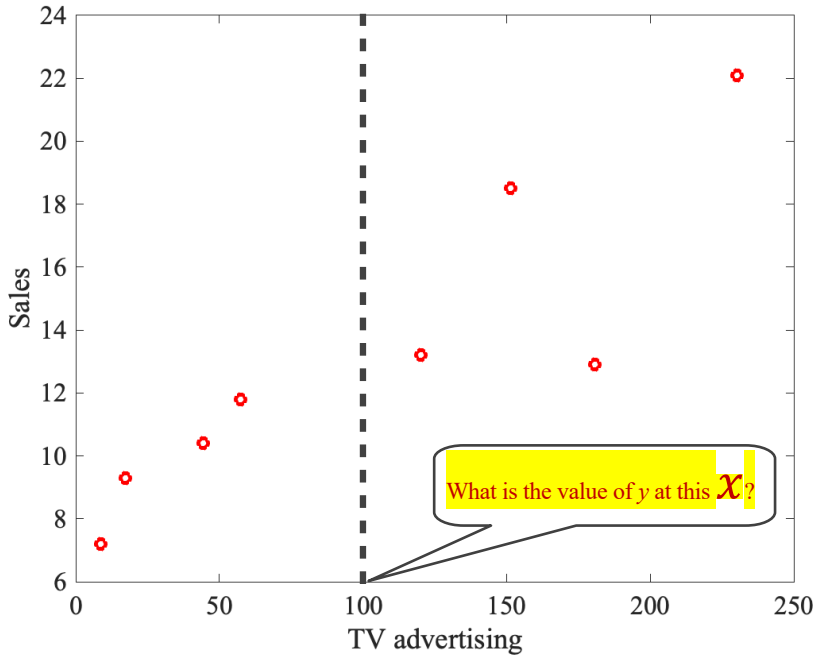
Prediction Model



PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

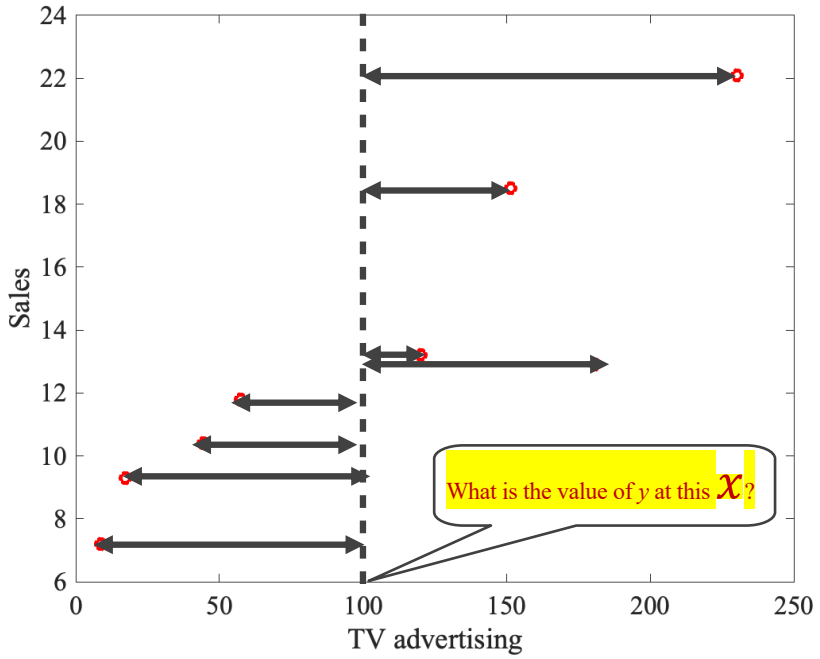
Prediction Model

How do we predict the sales value (y) for a given TV advertising value (x)?



PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

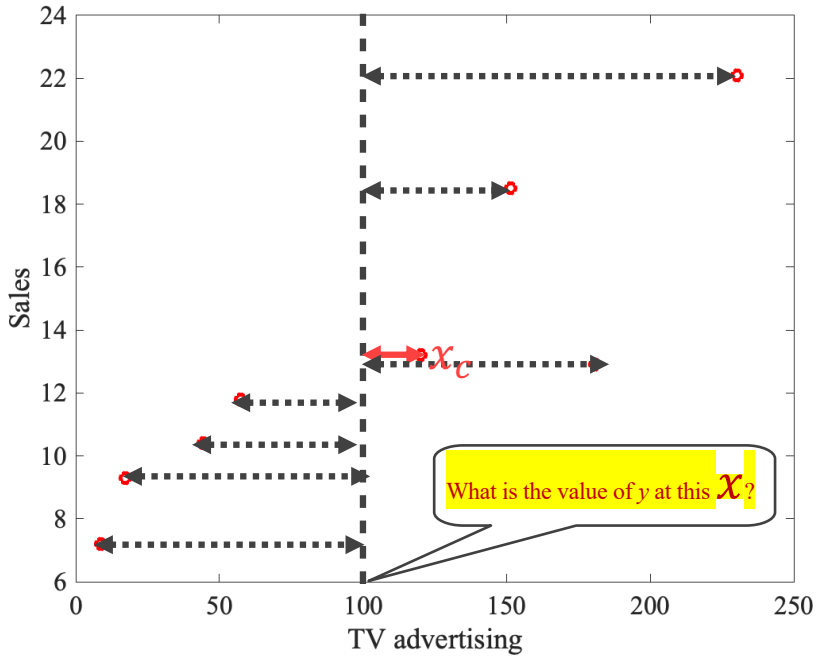
Prediction Model: 1-Neighbour



1. Find distance D(x,x_i) to all other points

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

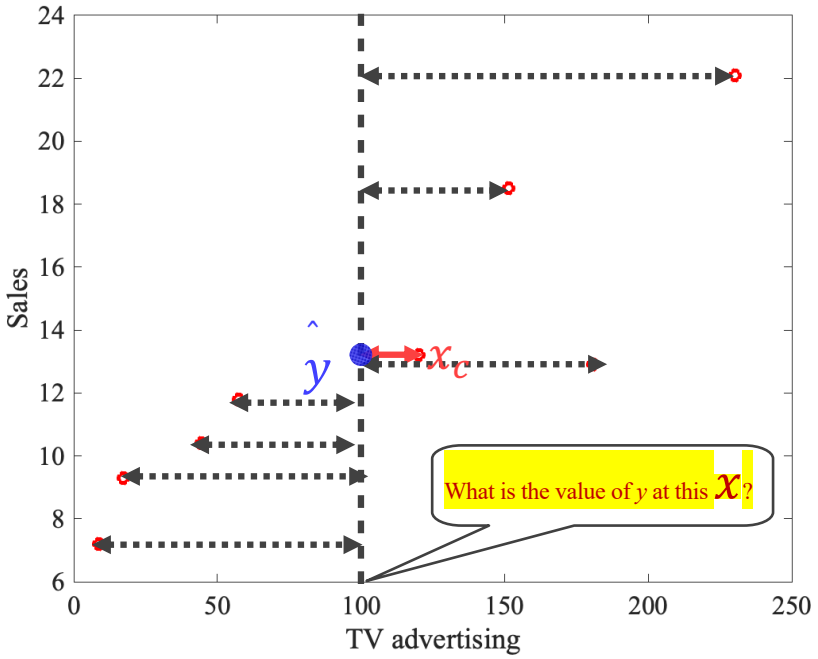
Prediction Model: 1-Neighbour



1. Find distance D(x,x_i) to all other points
2. Find closest x_c

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Prediction Model: 1-Neighbour



1. Find distance D(x,x_i) to all other points
2. Find closest x_c
3. Define $\hat{y}(x) = y(x_c)$

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Advantages of KNN Algorithm

- Very easy to understand and explain
- Non-parametric : no assumption about data distribution
- Naturally supports multi-class classification: no modification needed for more than two classes.
- Adaptable through K and distance metric

Weaknesses: computationally expensive - High memory usage – very sensitive to feature scaling (standardization)- sensitive to noise and outlier- Performance degrades in high dimensions- choise K is critical

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Performance Metrics

- Accuracy
- Sensivity
- Specificity
- Presision
- F1 score

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Accuracy

- It is defined as the rasio of correctly predicted instances to the total instances:

$$Accuracy = \frac{Correct\ prediction}{Total\ cases} * 100\%$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\%$$

Where TP = True Positives,
TN = True Negatives,
FP = False Positives, and FN = False Negatives.Read

False or True: Based on how accurate the prediction is.

Positive or Negative: are the output of the prediction.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Accuracy

200 data point

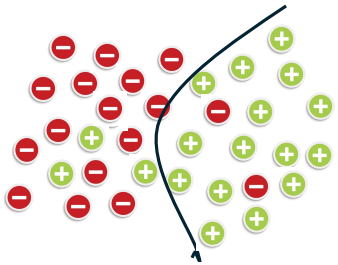
100 Healthy person

100 sick person (patient)

Our prediction

70 healthy person | 30 sick person

20 healthy person | 80 sick person



PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Accuracy

False or True: Based on how accurate the prediction is.

200 data point

100 Healthy person

100 sick person (patient)

Our prediction

70 healthy person

True 

30 sick person

False 

20 healthy person

False 

80 sick person

True 

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Accuracy

200 data point

100 Healthy person

100 sick person (patient)

Our prediction

70 healthy person — 30 sick person

20 healthy person — 80 sick person

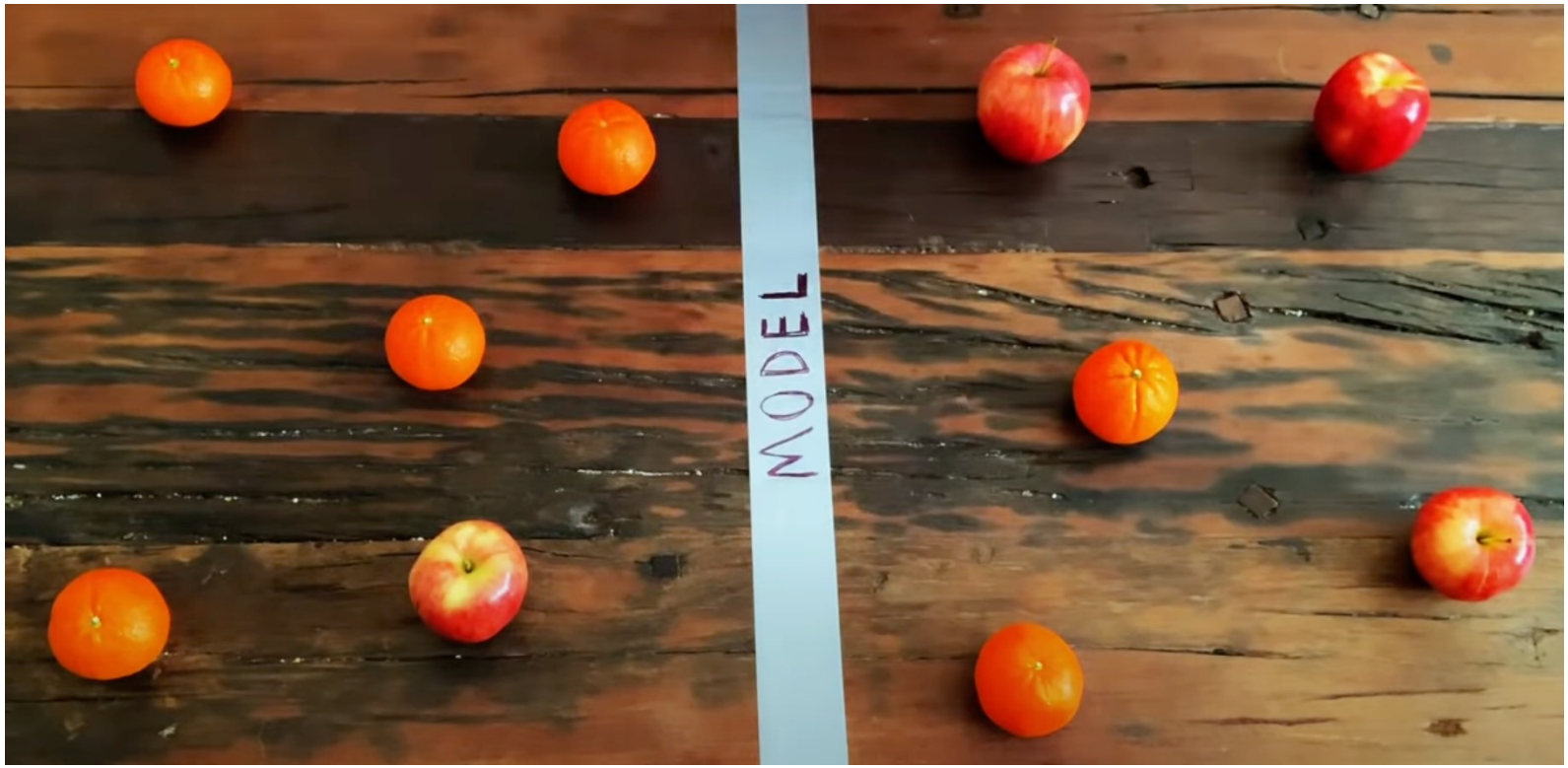
True 

False 

False 

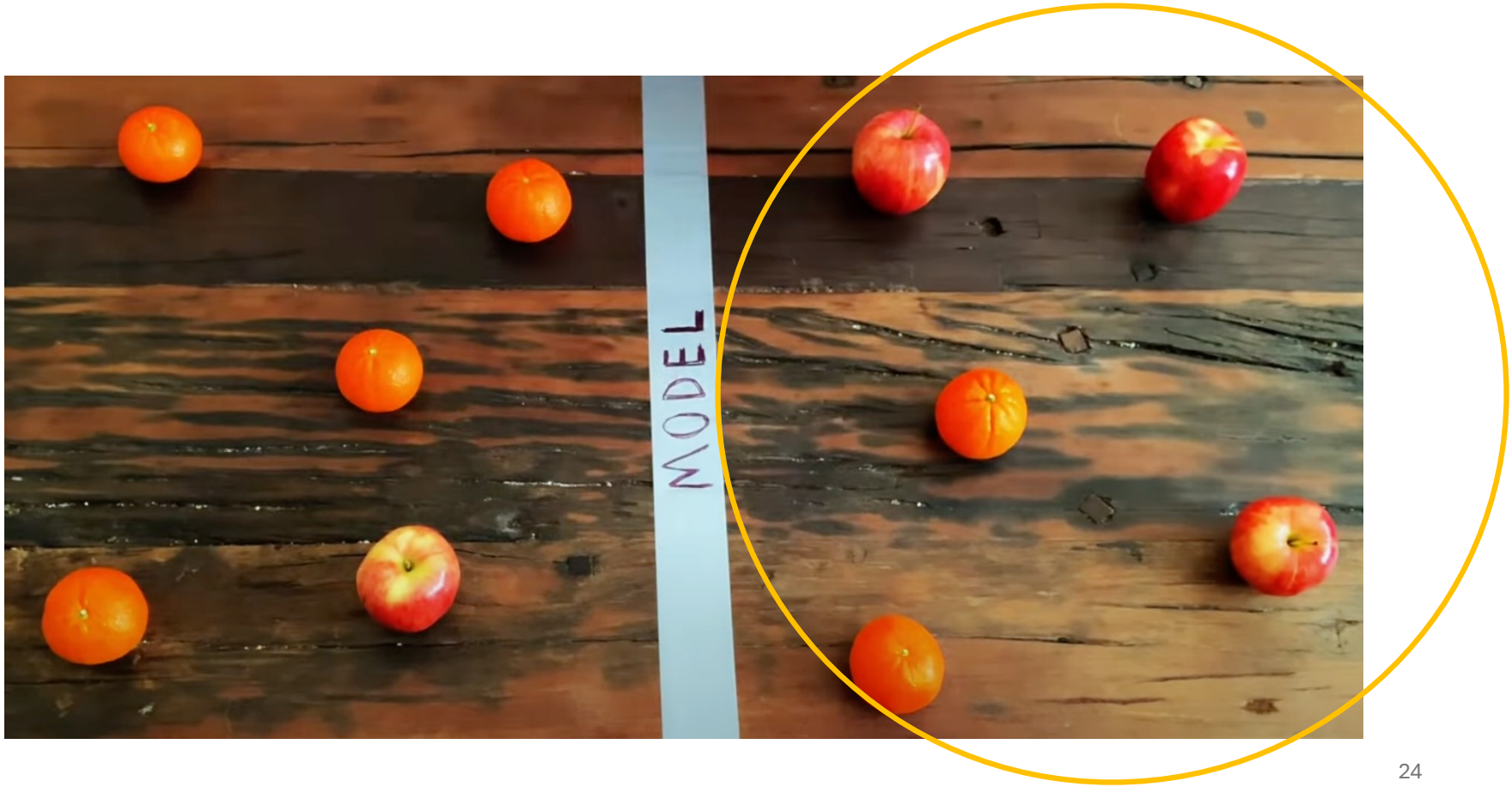
True 

Accuracy=(80+70)/200= 150/200=0.75*100=75%


MODEL

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

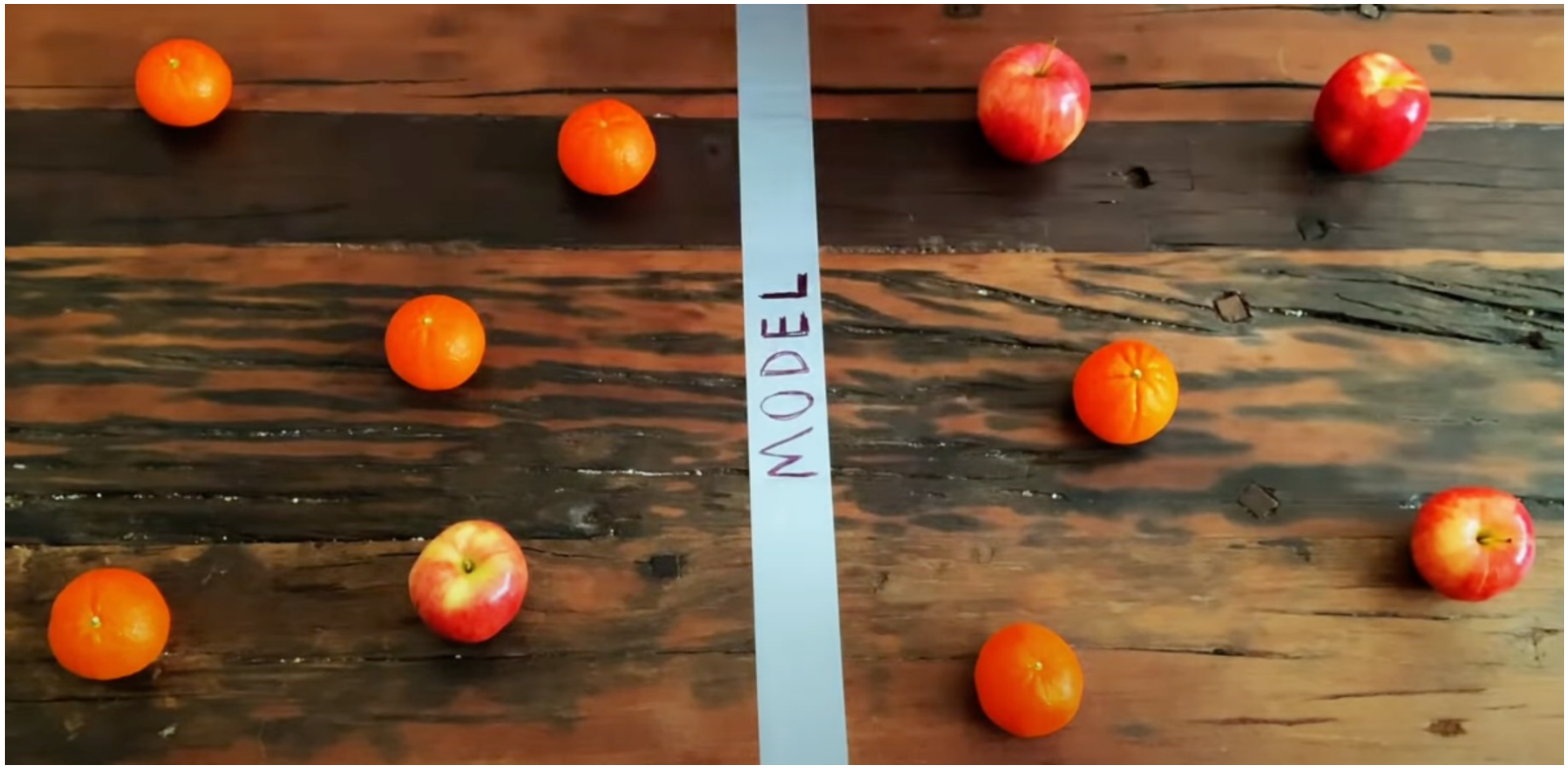
Anything on this side of the model will be classified as an appeles.



PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Oranges

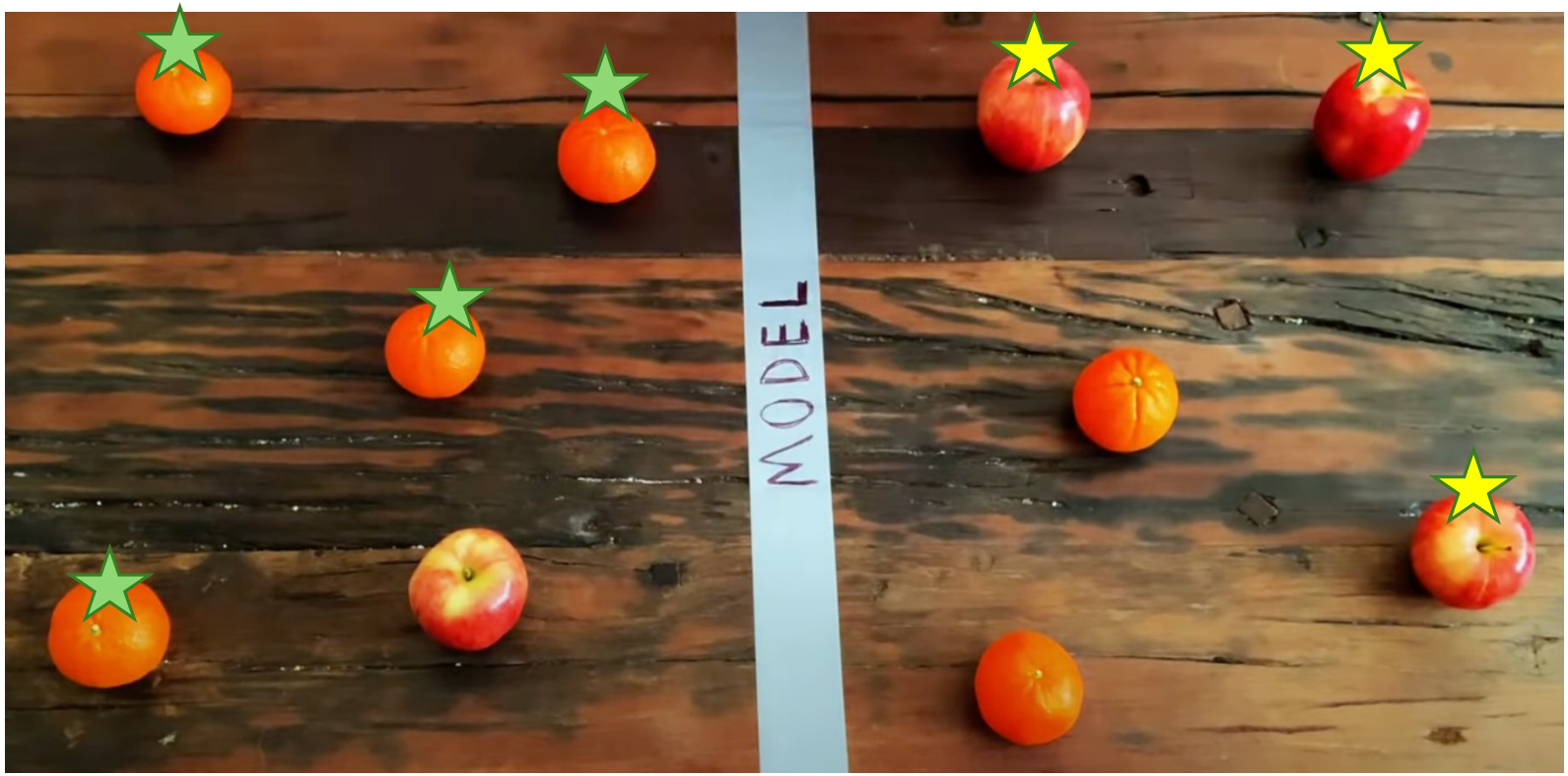
Apples



PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Accueracy is 7/10=70%

Oranges **Apples**



TP

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Customer Churn

- Customer churn refers to the situation where a customer stops using a company's services or products.
- In the banking sector, churn prediction is crucial because retaining customers is generally cheaper than acquiring new ones.

Based on customer data, predict whether a customer will leave the bank (churn = 1) or stay (churn = 0).

- This is a binary supervised classification problem.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Customer Churn

- Load the file
- Display the dataset shape and the first 5 rows

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv('./Banking_churn_prediction.csv')
```

```
df.shape
df.head()
```

	customer_id	vintage	age	gender	dependents	occupation	customer_nw_category	cu
0	1	2101	66	Male	0.0	self_employed	2	
1	2	2348	35	Male	0.0	self_employed	2	
2	4	2194	31	Male	0.0	salaried	2	
3	5	2329	90	NaN	NaN	self_employed	2	
4	6	1579	42	Male	2.0	self_employed	3	

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Customer Churn

- **what do the rows and columns represent?**

	customer_id	vintage	age	gender	dependents	occupation	customer_nw_category	cu
0	1	2101	66	Male	0.0	self_employed	2	
1	2	2348	35	Male	0.0	self_employed	2	
2	4	2194	31	Male	0.0	salaried	2	
3	5	2329	90	NaN	NaN	self_employed	2	
4	6	1579	42	Male	2.0	self_employed	3	

Each column represents a customer attribute

Each row represents one bank customer

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Dropping the ID column

- The id column is only a unique identifier and does not carry predictive information.

```
df=df.drop(['customer_id'],axis=1)
df.head()
```

	vintage	age	gender	dependents	occupation
0	2101	66	Male	0.0	self_employed
1	2348	35	Male	0.0	self_employed
2	2194	31	Male	0.0	salaried
3	2329	90	NaN	NaN	self_employed
4	1579	42	Male	2.0	self_employed

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Data Types Inspection

- Check the data types of all columns

‘customer_nw_category’ and ‘dependents’ should be treated as categorical variables, not numerical ones.

```
df.dtypes
```

```
vintage                          int64
age                              int64
gender                          object
dependents                     float64
occupation                      object
customer_nw_category             int64
current_balance                float64
previous_month_end_balance     float64
average_monthly_balance_prevQ  float64
average_monthly_balance_prevQ2 float64
current_month_credit           float64
previous_month_credit          float64
current_month_debit            float64
previous_month_debit           float64
current_month_balance          float64
previous_month_balance         float64
churn                            int64
```

Object

Object

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Defining Features and Target

- Target variable is : churn
- Features: all remaining column

```
Y=df[['churn']]
X=df.drop(['churn'],axis=1)
```

ebit	previous_month_debit	current_month_balance	previous_month_balance	churn
).20	0.20	1458.71	1458.71	0
6.27	100.56	6496.78	8787.61	0
6.73	259.23	5006.28	5070.14	0
).47	2143.33	2291.91	1669.79	1
3.62	1538.06	1157.15	1677.16	1

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Splitting numerical and categorical features

```
num=X.select_dtypes(include="number")
char=X.select_dtypes(include="object")
```

```
num.head()
```

	vintage	age	current_balance	previous_month_end_balance	a
0	2101	66	1458.71	1458.71	
1	2348	35	5390.37	8704.66	
2	2194	31	3913.16	5815.29	
3	2329	90	2291.91	2291.91	
4	1579	42	927.72	1401.72	

```
char.head()
```

	gender	dependents	occupation	customer_nw_category
0	Male	0.0	self_employed	2
1	Male	0.0	self_employed	2
2	Male	0.0	salaried	2
3	NaN	NaN	self_employed	2
4	Male	2.0	self_employed	3

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Outlier Analysis

- K-NN relies on distance calculations. Extreme values can distort distances and **negatively affect predictions**.

```
num.describe(percentiles=[0.01,0.05,0.10,0.25,0.50,0.75,0.85,0.9,0.99])
```

	vintage	age	current_balance	previous_month_end_balance
count	28382.000000	28382.000000	2.838200e+04	2.838200e+04
mean	2091.144105	48.208336	7.380552e+03	7.495771e+03
std	272.676775	17.807163	4.259871e+04	4.252935e+04
min	73.000000	1.000000	-5.503960e+03	-3.149570e+03
1%	1197.810000	9.000000	4.768920e+01	7.417820e+01
5%	1558.000000	23.000000	3.892485e+02	5.862515e+02
10%	1726.000000	28.000000	9.105650e+02	1.208590e+03

there is some drop between the 1 % and the minimum value (73-119) and there is a mild increase between the 99% and the maximum

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Outlier Analysis

Outlier Capping: instead of deleting extreme values we limit them to a reasonable range.

```
def outlier_cap(x):
    x=x.clip(lower=x.quantile(0.01))
    x=x.clip(upper=x.quantile(0.99))
    return(x)
```

When we say «cap at 1% » : the lowest 1% of values are considered too small

```
num=num.apply(lambda x : outlier_cap(x))
```

So, we replace values below 1st percentile with the 1st persentile value

```
num.describe(percentiles=[0.01,0.05,0.10,0.25,0.50,0.75,0.85,0.9,0.99])
```

	vintage	age	current_balance	previous_month_end_balance
count	28382.000000	28382.000000	28382.000000	28382.000000
mean	2093.031994	48.242865	6265.041971	6363.576965
std	264.421711	17.726483	9381.737713	9447.567053
min	1197.810000	9.000000	47.689200	74.178200
1%	1197.963900	9.000000	47.738448	74.212058
5%	1558.000000	23.000000	389.248500	586.251500

This keeps all data points but reduces the impact of extreme values.

Distances become more reasonable.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Missing value

```
num.isnull().mean()
```

```
vintage                          0.0
age                              0.0
current_balance                  0.0
previous_month_end_balance       0.0
average_monthly_balance_prevQ    0.0
average_monthly_balance_prevQ2   0.0
current_month_credit             0.0
previous_month_credit            0.0
current_month_debit              0.0
```

```
char.isnull().mean()
```

```
gender                0.018498
dependents            0.086780
occupation            0.002819
customer_nw_category  0.000000
dtype: float64
```

Missing value mainly appear in categorical variables.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Removing columns with more than 25% missing value

```
char=char.loc[:,char.isnull().mean()<=0.25]
```

Imputation of missing values

```
from sklearn.impute import SimpleImputer
imputer=SimpleImputer(missing_values=np.nan, strategy='most_frequent')
char_1=pd.DataFrame(imputer.fit_transform(char),index=char.index,columns=char.columns)
```

```
char_1.isnull().mean()
```

```
gender                  0.0
dependents              0.0
occupation              0.0
customer_nw_category    0.0
dtype: float64
```

by using this code, we want to replace the Nan value (missing value) in categorical features with the most frequent value for each feature.

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Encode Categorical Features

- **Why must categorical variables be encoded?**

  Machine learning algorithms require numerical inputs

```
# Create dummy features with n-1 levels
X_char_dum = pd.get_dummies(char_1, drop_first = True)
X_char_dum.shape
```

```
(28382, 21)
```

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Creating the final feature Set

```
X_all=pd.concat([X_char_dum,num],axis=1,join="inner")
X_all.head()
```

	gender_Male	dependents_1.0	dependents_2.0	dependents_3.0	dependents_4.0
0	True	False	False	False	False
1	True	False	False	False	False
2	True	False	False	False	False
3	True	False	False	False	False
4	True	False	True	False	False

5 rows × 33 columns

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Feature Scaling

Feature scaling: It is very important step, if you want to have a K-NN model. we will try to standardize each feature. After that, they become comparable.

Since K-NN is distanced based, feature must be on comparable.

```
from sklearn.preprocessing import StandardScaler
stdsc=StandardScaler()
X_std=pd.DataFrame(stdsc.fit_transform(X_all), index=X_all.index, columns=X_all.columns)
```

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Train Test Split

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test=train_test_split(X_std, Y, test_size=0.3, random_state=42)
```

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Building the K-NN classification model

- **K-NN classifies a data poit based on the majority class among its k nearest neighbors.**

```
from sklearn.neighbors import KNeighborsClassifier
KNN=KNeighborsClassifier(n_neighbors = 5)
KNN.fit(X_train,y_train)
```

Small K : sensitive to noise

Large K: smoother decision boundary

PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Model Evaluation

```
from sklearn import metrics
from sklearn.metrics import classification_report
y_pred_KNN=KNN.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred_KNN))
print(classification_report(y_test, y_pred_KNN))
```

```
Accuracy: 0.8193775689958896
              precision    recall  f1-score   support

           0       0.84      0.96      0.90      6967
           1       0.51      0.18      0.27      1548

    accuracy                           0.82      8515
   macro avg       0.67      0.57      0.58      8515
weighted avg       0.78      0.82      0.78      8515
```

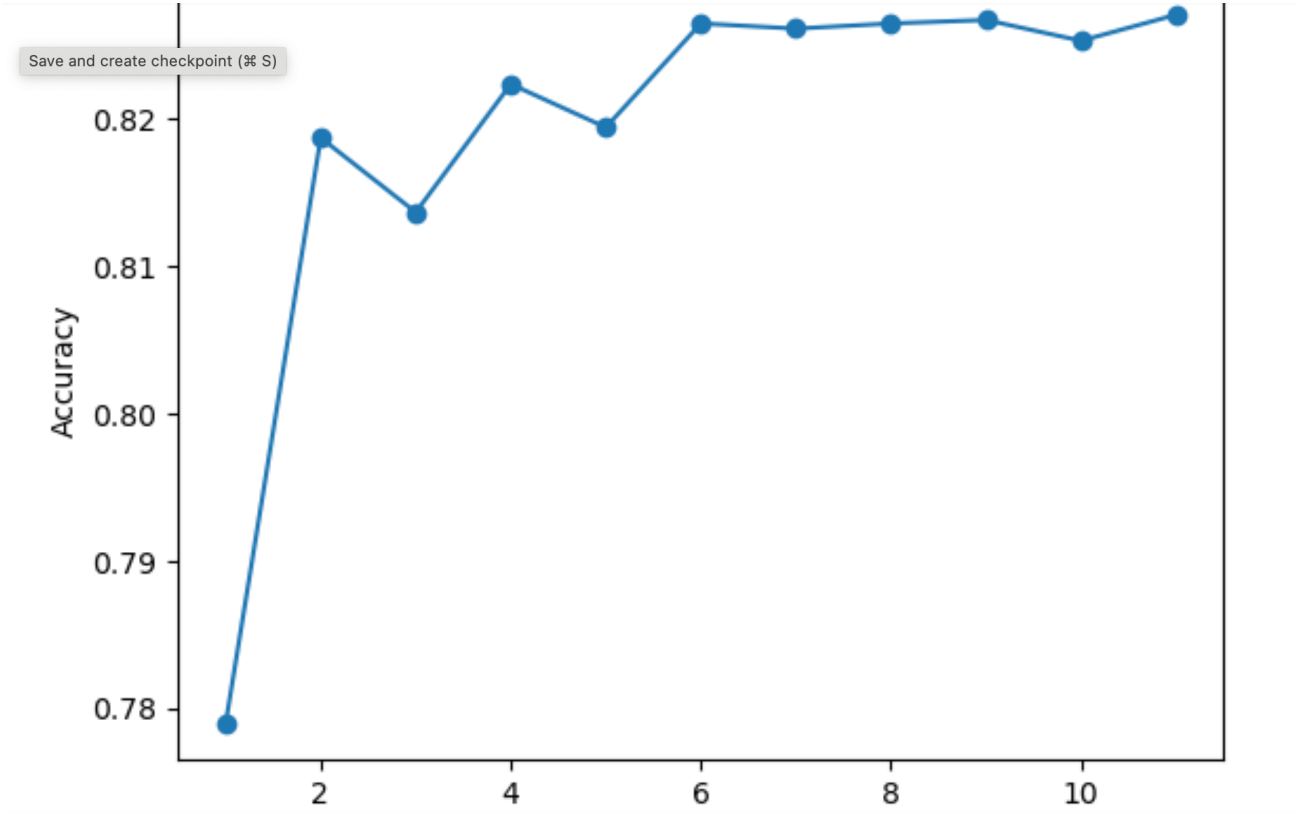
PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

Visualizing the effect of K

```
import matplotlib.pyplot as plt
from sklearn.metrics import  accuracy_score

k_values = range(1, 12)
accuracies = []


for k in k_values:
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)
    y_pred_k = knn.predict(X_test)
    accuracies.append(accuracy_score(y_test, y_pred_k))


plt.figure()
plt.plot(k_values, accuracies, marker='o')
plt.xlabel('Number of Neighbors (K)')
plt.ylabel('Accuracy')
plt.title('K-NN Accuracy vs K')
plt.show()
```

To understand how the choice of K influences model performance

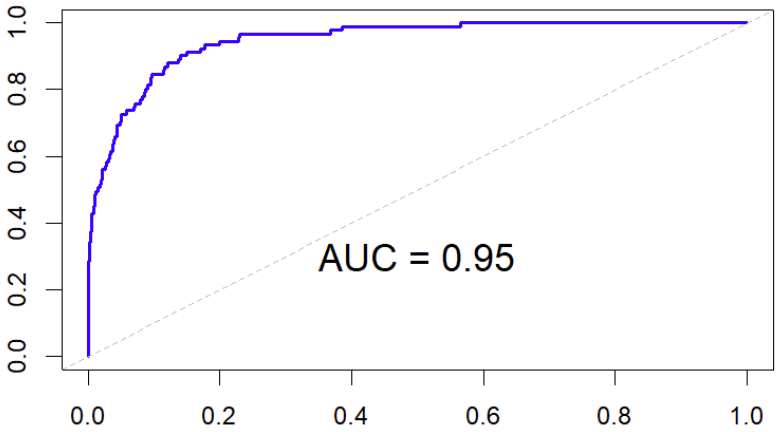
PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING



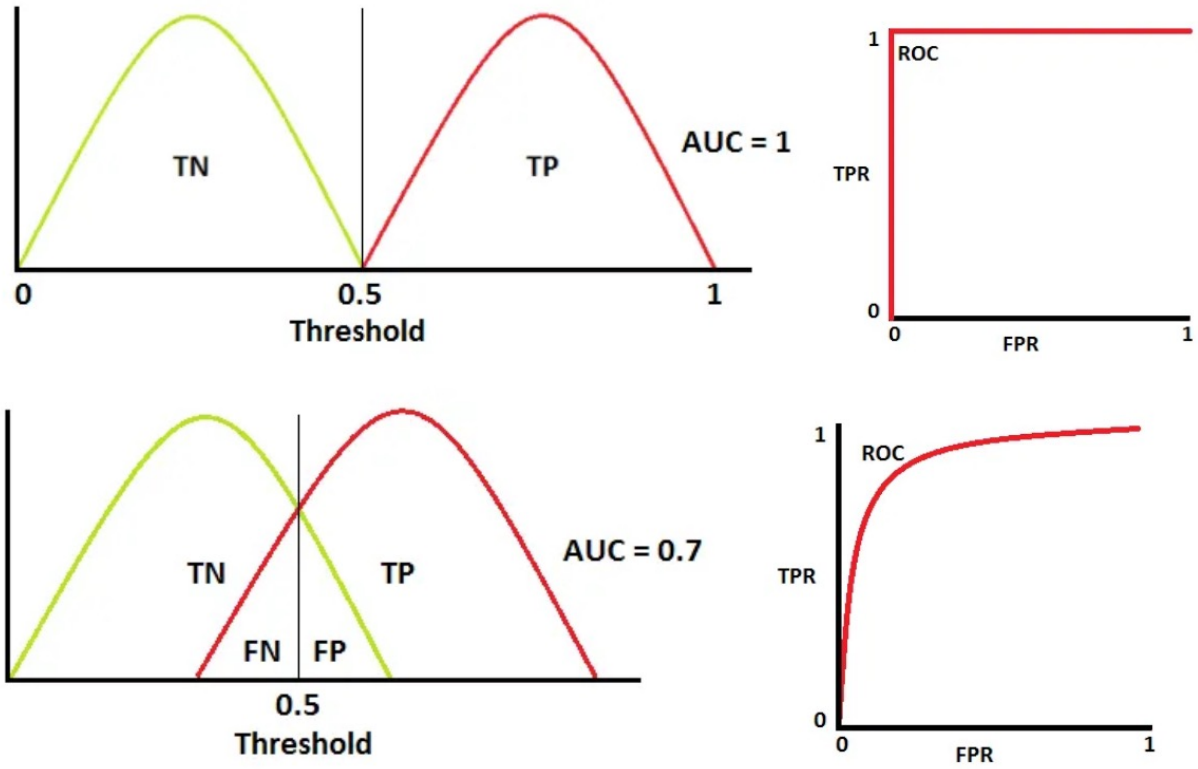
PROGIS
ETUDES, MEDIAS, COMMUNICATION, MARKETING

AUC (Area Under The Curve)

- It tells how much the model is capable of distinguishing between classes.
- Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.
- An excellent model has AUC near to the 1 which means it has a good measure of separability.



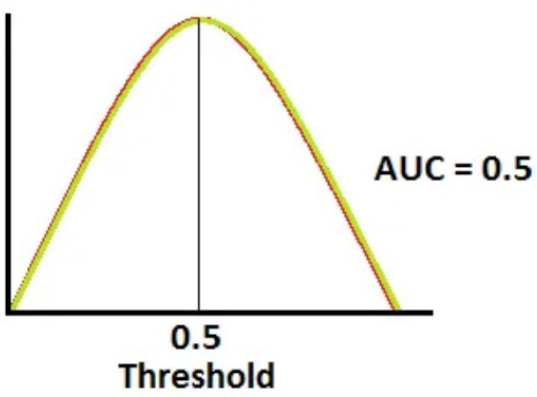
https://www.listendata.com/2023/08/calculate-auc-in-r.html

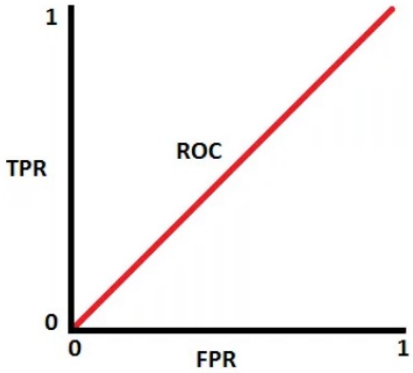


https://medium.com/towards-data-science/understanding-auc-roc-curve-68b2303cc9c5

The model perfectly able to distinguish between positive class and negative class.

When AUC is 0.7, it means there is a 70% chance that the model will be able to distinguish between positive class and negative class.





when AUC is 0.5, it means the model has no class separation capacity whatsoever.