DMW MINI PROJECT

Varun Gadde 9512063
Vedant Gaikwad 9512073

BANK CUSTOMERS EXITING PREDICTIVE MODEL

A bank is investigating a very high rate of customer leaving the bank. We investigate and predict which of the customers are more likely to leave it

- 1 Dropping unnecessary columns
- 2 Variable Analysis
  - 2.1 Categorical Variables
  - 2.2 Numerical Variables
- 3 Machine Learning
  - 3.1 Modelling
  - 3.2 Logistic Regression
  - 3.3 Classification with SVM (Support Vector Machine)
  - 3.4 Classification with Naive Bayes
  - 3.5 Classification with Decision Tree
  - 3.7 K-Fold Cross Validation (cross_val_score)

1. Loading Libraries

```
# Linear algebra
import numpy
```

ii. Importing data

```
# Input data files
!gdrive download --recursive 1-TMP-Ebu-Fc1xcrfbrcr-CiamcGcnfoc
```

```
Downloading 1TMP-Ebu-Fc1xcrfbrcr-CiamcGcnfoc with
Complete!
Downloaded 1TMP-Ebu-Fc1xcrfbrcr-CiamcGcnfoc
Downloaded 1TMP-Ebu-Fc1xcrfbrcr-CiamcGcnfoc
rate: 1.0MB/s, elapsed 1.0MB/s, total 1.0MB/s
Downloaded: 1 files, 1.0MB/s, elapsed 1.0MB/s, total 1.0MB/s
Downloaded: 0 files, 0B (0 B/s)
```

```
ls -ltr DMW-Mini-Project-Classification-2021
ls -ltr DMW-Mini-Project-Classification-2021
cd DMW-Mini-Project-Classification-2021
ls -ltr
cd DMW-Mini-Project-Classification-2021
```

```
total 0
drwxr-xr-x 2 root root 4096 Jan  6 14:12 DMW-Mini-Project-Classification-2021
drwxr-xr-x 2 root root 4096 Jan  6 14:12 DMW-Mini-Project-Classification-2021
/content/DMW-Mini-Project-Classification-2021
total 0
-rwxr-xr-x 1 root root 0 Jan  6 14:12 Churn_Modelling.csv
```

```
# Directories refer to data
data_dir=".
```

```
[0]  0    15634602  Hargrave  619  France  Female  42
[1]  1    15647311  Hill      608  Spain   Female  41
[2]  2    15619304  Onio      502  France  Female  42
[3]  3    15701354  Boni      699  France  Female  39
[4]  4    15737888  Mitchell  850  Spain   Female  43
[5]  5    15574012  Chu       645  Spain   Male    44
[6]  6    15592531  Bartlett  822  France  Male    50
[7]  7    15656148  Obinna    376  Germany Female  29
[8]  8    15792365  He        501  France  Male    44
[9]  10   15328328  Ali       654  France  Male    27
```

```
# Directories refer to data
data_dir="."
```

```
Churn_Modelling.csv  DMW-Mini-Project-Classification-2021
Churn_Modelling.csv  DMW-Mini-Project-Classification-2021
total 2000
-rwxr-xr-x 1 root root 1000000 Jan  6 14:12 Churn_Modelling.csv
drwxr-xr-x 2 root root 4096 Jan  6 14:12 DMW-Mini-Project-Classification-2021
```

```
# Input data files available in the read-only "../input/" directory
# For example, running this (by default) will print all files in the input directory
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
/kaggle/input/churn-modelling/Churn_Modelling.csv
```

```
# Input data files available in the read-only "../input/" directory
# For example, running this (by default) will print all files in the input directory
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
/kaggle/input/churn-modelling/Churn_Modelling.csv
```

```
# This data set consists of 3333 entries, 21 columns
# We can check missing values using the .isnull() method
data = pd.read_csv("churn.csv")
data.isnull().sum()
```

```
Customer ID      0
Surname          0
Credit Score     0
Geography        0
Gender           0
Age              0
Tenure           0
Balance          0
Num of Products  0
Has Credit Card  0
Is Active Member 0
Estimated Salary 0
Churn            0
dtype: int64
```

```
# From above, we understand that there is no missing value in column.
# We can also check the unique values in each column of the dataset.
# We can check missing values also using the isnull() method as below
data = pd.read_csv("churn.csv")
data.nunique()
```

```
Customer ID      10000
Surname           2932
Credit Score       460
Geography            3
Gender               2
Age                 70
Tenure              11
Balance           6382
Num of Products      4
Has Credit Card      2
Is Active Member     2
Estimated Salary  9999
Churn                2
dtype: int64
```

1. Dropping unnecessary columns

```
# Dropping unnecessary columns
data.drop(["Surname","Customer ID","Surname"], axis=1, inplace=True)
data.head()
```

	Credit Score	Geography	Gender	Age	Tenure	Balance	Num of Products	Has Credit Card	Is Active Member	Estimated Salary	Churn
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

```
# Dropping unnecessary columns
data.drop(["Surname","Customer ID","Surname"], axis=1, inplace=True)
data.head()
```

```
KeyboardInterrupt: Traceback (most recent call last)
/usr/local/lib/python3.7/dist-packages/IPykernel_launcher.py in <module>
      2 data.drop(["Surname","Customer ID","Surname"], axis=1, inplace=True)
      3 data.head()
      4 data.drop(["Surname","Customer ID","Surname"], axis=1, inplace=True)
      5 data.head()
/usr/local/lib/python3.7/dist-packages/pandas/core/frame.py in drop(self, labels, axis, index, columns, level, inplace, errors)
   4117 
   4118 
   4119 
   4120 
   4121 
   4122 
   4123 
   4124 
   4125 
   4126 
   4127 
   4128 
   4129 
   4130 
   4131 
   4132 
   4133 
   4134 
   4135 
   4136 
   4137 
   4138 
   4139 
   4140 
   4141 
   4142 
   4143 
   4144 
   4145 
   4146 
   4147 
   4148 
   4149 
   4150 
   4151 
   4152 
   4153 
   4154 
   4155 
   4156
```