Project

Data Science

Agenda

- TEAM MEMBERS
- INTRODUCTION
- DATASET DESCRIPTION
- PREPROCESSING
- MODELS
- RESULTS
- DATASET LINK

يوسف يسري رضا حليم	2021170655
فيلوباتير جورج زكريا بسطوروس	2021170397
أنس سعد فتحي سعد	2021170100
اندرو عماد سامي	2021170098
بيتر ماجد جرجس باخوم	2021170128
يوسف رامي يوسف مسيحة	2021170631

TEAM MEMBERS

CUSTOMER CHURN, OR THE ATTRITION OF CUSTOMERS, IS A CRITICAL CONCERN FOR BUSINESSES ACROSS VARIOUS INDUSTRIES. IDENTIFYING FACTORS THAT CONTRIBUTE TO CUSTOMER CHURN AND BUILDING PREDICTIVE MODELS TO ANTICIPATE AND MITIGATE THIS PHENOMENON ARE ESSENTIAL FOR MAINTAINING A STABLE AND THRIVING CUSTOMER BASE

Our dataset consisting of :

- **ROWNUMBER**: AN IDENTIFIER FOR EACH ROW IN THE DA TASTE.
- **CUSTOMERID**: A UNIQUE IDENTIFIER FOR EACH CUSTOMER.
- **SURNAME**: THE SURNAME OR LAST NAME OF THE CUSTOMER.
- **CREDITSCORE**: A NUMERICAL REPRESENTATION OF THE CREDITWORTHINESS OF THE CUSTOMER.
- **CREDITSCORE**: A NUMERICAL REPRESENTATION OF THE CREDITWORTHINESS OF THE CUSTOMER.
- **GEOGRAPHY**: THE GEOGRAPHICAL LOCATION OR COUNTRY WHERE THE CUSTOMER RESIDES.
- **GENDER**: THE GENDER OF THE CUSTOMER (MALE/FEMALE).
- **AGE**: THE AGE OF THE CUSTOMER.

Our dataset consisting of :

- **TENURE:** THE NUMBER OF YEARS THE CUSTOMER HAS BEEN WITH THE BANK OR SERVICE.
- **BALANCE:** THE ACCOUNT BALANCE OF THE CUSTOMER.
- **NUMOFPRODUCTS:** THE NUMBER OF BANK PRODUCTS THE CUSTOMER IS USING.
- **HASCRCARD:** BINARY INDICA TOR (0 OR 1) WHETHER THE CUSTOMER HAS A CREDIT CARD.
- **ISACTIVEMEMBER:** BINARY INDICA TOR (0 OR 1) WHETHER THE CUSTOMER IS AN ACTIVE MEMBER.
- **ESTIMATEDSALARY:** THE ESTIMA TED SALARY OF THE CUSTOMER.
- **EXITED:** THE TARGET VARIABLE INDICA TING WHETHER THE CUSTOMER HAS CHURNED (1) OR NOT (0).

- we use functions like (.info) &(.head) to describe our dataset
- Check Nulls & duplicate and remove it
- we dropped three unnecessary columns
- Making one-hot encode for Geography column
- Encode Gender column
- Calculate z-score for specific columns then detect the outliers and delete it

Each algorithm has strengths depending on the dataset's complexity and feature interactions.

RANDOMFOREST

- Handles mixed data types
- Works well with non-linear relationships

SVM

- Effective in high-dimensional spaces
- Works well for smaller datasets

KNN

- Simple and effective for smaller feature spaces

REASON : Because our data is a numerical data

RANDOMFOREST:

3

```
RF Accuracy: 90.62%
f1 score =  90.61775699785439
recall score =  90.62306653539282
precision score =  90.61370204480262
confusion matrix:
[[1059  110]
 [ 104 1009]]
```

SVM:

1

```
SVM Accuracy: 85.19%
f1 score =  85.18131712745225
recall score =  85.2935638157647
precision score =  85.43791967414649
confusion matrix:
[[947 222]
 [116 997]]
```

KNN:

4

```
KNN Accuracy: 82.87%
f1 score =  82.78441984419844
recall score =  83.08477384852935
precision score =  83.95755630865156
confusion matrix:
[[ 867  302]
 [  89 1024]]
```

DEEP LEARNING MODEL :

2

```
Enhanced Neural Network Model Evaluation:
Accuracy: 0.8584574934268185
F1 Score: 0.8626116546150574
Recall Score: 0.9110512129380054
Precision Score: 0.8190630048465266
Confusion Matrix:
 [[ 945  224]
 [  99 1014]]
```

HERE IS THE LINK :

https://www.kaggle.com/datasets/filippoo/deep-learning-az-ann


thank you so much