# USABLE ARTIFICIAL INTELLIGENCE PROJECT REPORT INFO-I513

# BANK CUSTOMER CHURN PREDICTION

## TEAM

**SAI SRIKAR GANDHE**
**EMAIL: sgandhe@iu.edu**

**FHARIYA ASEEM FATHIMA**
**EMAIL: fha@iu.edu**

**Site for Data Set: https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers?resource=download**

# ABSTRACT

The purpose of this study is to investigate the factors that contribute to bank customer churn and to develop an accurate predictive model. The dataset contains information on customer demographics, account details, transaction history, and customer interactions with the bank. Several machine learning models, including Logistic Regression, Decision Trees, Random Forest, XGBoost and k-Nearest Neighbors (k-NN). The results suggest that the most important factors affecting churn, the best-performing model can be achieved by comparing accuracies.

# INTRODUCTION

The banking business has gotten increasingly competitive in recent years, and customer churn has become a serious concern. Customers can simply transfer banks, making it critical for banks to identify the causes of churn and implement effective retention strategies. Customer churn prediction is a challenging topic that requires a detailed understanding of customer behavior and interactions with the bank. We address the subject of bank customer churn prediction in this project and offer a machine learning-based strategy and to construct an accurate churn prediction model.

# DATA SET

**About the dataset:** This data set comprises information about bank customers, and the target variable is a binary variable that indicates whether the customer has left the bank (closed his account) or not.
It comprises of 10,000 records containing demographic and banking history information from customers in three countries: France, Germany, and Spain.

**Attributes**:

- **RowNumber-** represents the number of the record (row)
- **CustomerId-** contains random values and has no effect on the customer's decision to leave the bank.
- **Surname-** A customer's surname and it has no effect on their decision to leave the bank.
- **CreditScore-** A customer with a higher credit score is less likely to leave the bank, hence it can have an effect on customer churn.
- **Geography-** the location of a customer can influence their decision to leave the bank.
- **Gender -** worth it's investigating whether gender influences a customer's decision to leave a bank.
- **Age -** It is obviously crucial, as older clients are less likely than younger customers to leave their bank.
- **Tenure**- the number of years a customer has been a bank's client. Older customers are typically generally committed and much less probable to leave a bank.

- **Balance**-Another good indicator of customer churn, as clients with higher balances are less likely to leave the bank than those with smaller balances.
- **NumOfProducts**-refers to the quantity of products purchased by a customer through the bank.
- **HasCrCard**- indicates whether a consumer owns a credit card. This column is also crucial because credit card holders are less likely to leave the bank.
- **IsActiveMember**- active clients are more likely to stay with the bank.
- **EstimatedSalary**- as with balance, people with lower salaries are more likely than ones with greater salaries to quit the bank.
- **Exited**- whether or not the customer left the bank.

## DATA COLLECTION:

- The data was obtained from Kaggle.
- The Data was collected from a hypothetical bank's database.
- No, it doesn't come from a poll/survey, the data collected from a hypothetical bank's database.
- Yes, the data used from Kaggle (https://www.kaggle.com/mathchi/churn-for-bank-customers)
- It is possible to re-collect the data, but the source data is unavailable.
- It is difficult to say definitively whether any aspect of the data was collected manually. It is also possible that some of the data was cleaned or processed manually after it was collected. However, this would depend on the specific details of the data collection process and any subsequent data cleaning or processing steps taken.

# DATA MANAGEMENT:

## Libraries used:

```
In [20]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns
         import plotly.express as px
```

```
In [47]: from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(Feature, target, test_size=0.2, random_state=42)
```

### Logistic Regression

```
In [48]: from sklearn.linear_model import LogisticRegression
         from sklearn.metrics import accuracy_score
         from sklearn.metrics import confusion_matrix
```

```
In [50]: from sklearn.metrics import classification_report
         print(classification_report(y_test, y_prediction))
```

```
In [51]: from sklearn.tree import DecisionTreeClassifier
```

```
In [54]: from sklearn.ensemble import RandomForestClassifier
```

```
In [57]: from xgboost import XGBClassifier
```

```
In [60]: from sklearn.neighbors import KNeighborsClassifier
```

```
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
```

```
In [63]: # using GridSearchCV to find the best hyperparameters for Random Forest
         from sklearn.model_selection import GridSearchCV
```

- The data was preprocessed by detecting and eliminating missing values.

```
In [28]: data.isnull().sum()
```

```
Out[28]: RowNumber          0
         CustomerId         0
         Surname            0
         CreditScore        0
         Geography          0
         Gender             0
         Age                0
         Tenure             0
         Balance            0
         NumOfProducts      0
         HasCrCard          0
         IsActiveMember     0
         EstimatedSalary    0
         Exited             0
         dtype: int64
```

- Some variables were recoded to numeric values to allow for easier analysis.

```
In [41]: data['Geography'] = data['Geography'].astype("category")
         data['Gender'] = data['Gender'].astype("category")

In [42]: data = pd.get_dummies(data, columns=['Geography', 'Gender'])

In [43]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   CreditScore        10000 non-null  int64  
 1   Age                10000 non-null  int64  
 2   Tenure             10000 non-null  int64  
 3   Balance            10000 non-null  float64
 4   NumOfProducts      10000 non-null  int64  
 5   HasCrCard          10000 non-null  int64  
 6   IsActiveMember     10000 non-null  int64  
 7   EstimatedSalary    10000 non-null  float64
 8   Exited             10000 non-null  int64  
 9   Geography_France   10000 non-null  uint8  
 10  Geography_Germany  10000 non-null  uint8  
 11  Geography_Spain    10000 non-null  uint8  
 12  Gender_Female      10000 non-null  uint8  
 13  Gender_Male        10000 non-null  uint8  
dtypes: float64(2), int64(7), uint8(5)
memory usage: 752.1 KB
```

```
In [25]: data
```

Out[25]:

| | CreditScore | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited | Geography_France | Geography_Germany | Geo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 619 | 42 | 2 | 0.00 | 1 | 1 | 1 | 101348.88 | 1 | 1 | 0 | |
| **1** | 608 | 41 | 1 | 83807.86 | 1 | 0 | 1 | 112542.58 | 0 | 0 | 0 | |
| **2** | 502 | 42 | 8 | 159660.80 | 3 | 1 | 0 | 113931.57 | 1 | 1 | 0 | |
| **3** | 699 | 39 | 1 | 0.00 | 2 | 0 | 0 | 93826.63 | 0 | 1 | 0 | |
| **4** | 850 | 43 | 2 | 125510.82 | 1 | 1 | 1 | 79084.10 | 0 | 0 | 0 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **9995** | 771 | 39 | 5 | 0.00 | 2 | 1 | 0 | 96270.64 | 0 | 1 | 0 | |
| **9996** | 516 | 35 | 10 | 57369.61 | 1 | 1 | 1 | 101699.77 | 0 | 1 | 0 | |
| **9997** | 709 | 36 | 7 | 0.00 | 1 | 0 | 1 | 42085.58 | 1 | 1 | 0 | |
| **9998** | 772 | 42 | 3 | 75075.31 | 2 | 1 | 0 | 92888.52 | 1 | 0 | 1 | |
| **9999** | 792 | 28 | 4 | 130142.79 | 1 | 1 | 0 | 38190.78 | 0 | 1 | 0 | |

10000 rows × 14 columns

```
]: data
```

]:

| NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited | Geography_France | Geography_Germany | Geography_Spain | Gender_Female | Gender_Male |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 101348.88 | 1 | 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 1 | 112542.58 | 0 | 0 | 0 | 1 | 1 | 0 |
| 3 | 1 | 0 | 113931.57 | 1 | 1 | 0 | 0 | 1 | 0 |
| 2 | 0 | 0 | 93826.63 | 0 | 1 | 0 | 0 | 1 | 0 |
| 1 | 1 | 1 | 79084.10 | 0 | 0 | 0 | 1 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2 | 1 | 0 | 96270.64 | 0 | 1 | 0 | 0 | 0 | 1 |
| 1 | 1 | 1 | 101699.77 | 0 | 1 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 42085.58 | 1 | 1 | 0 | 0 | 1 | 0 |
| 2 | 1 | 0 | 92888.52 | 1 | 0 | 1 | 0 | 0 | 1 |
| 1 | 1 | 0 | 38190.78 | 0 | 1 | 0 | 0 | 1 | 0 |

- No, other data is merged.
- There was no other data merged in or manually manipulated.

ANALYSIS:

- The analysis used in this project which involves the use of machine learning algorithms to predict customer churn in a bank. To evaluate the performance of the algorithm, we used cross-validation to estimate the accuracy of the model on unseen data. We also used metrics such as precision, recall, and F1-score to evaluate the performance of the model. Also, we have found ROC curves and AUC score for all models to compare the models.
- The analysis done here is to predict customer churn in a bank using machine learning algorithms.
- Data pre-processing: The data was cleaned and pre-processed by handling missing values, encoding categorical variables, and scaling numerical features.
- Feature selection: Features were selected based on their importance in predicting churn.
- Model selection: Five different machine learning models were evaluated: Logistic Regression, Random Forest, XGBoost, Decision

Tree, and KNN. The Random Forest model was chosen as it achieved the highest accuracy.

- Model evaluation: The accuracy of the final model was evaluated using confusion matrix and also found ROC curve and AUC score.
- There is no data sub setting.
- There is no subgroup analysis.

ARGUMENT:

- The core argument is that machine learning can be used to predict customer churn with high accuracy.
- We have visualized the customers who are exited and retained by using pie diagram from that visualization we got to know that 20.4% customers have exited, and 79.6% customers are retained as per the data.
- It is not a causal argument because while some of the features in the bank churn project, such as age, credit score, balance, and tenure, may have a causal relationship with the target variable, there may be other confounding factors that are not included in the dataset that could affect the customer's decision to leave the bank. Therefore, it's important to be cautious in making causal claims based on correlation analysis alone.

# Design

## INTERVENTION:

The goal of the design solution is to predict bank churn using machine learning algorithms and provide the best model using accuracy and AUC score.

- The target behavior of the intervention is to predict which customers are at a higher risk of leaving the bank, based on their behavior and characteristics. The machine learning-based churn prediction model can analyze customer data and identify patterns that indicate a customer is at a higher risk of churn. The goal is to reduce the customer churn rate, which can lead to higher profits for the bank.
- The expected behavior of my intervention is that it will accurately predict whether a bank customer will churn or not with a high degree of accuracy.
- The form of delivery of my intervention is a machine learning model that will be trained on historical data, which will then be used to predict whether a customer will churn or not based on their current data.
- The system can analyze the customer data and predict the likelihood of churn.
- To measure the effects of my intervention, I will use metrics such as accuracy, AUC score, and confusion matrix. These metrics will help me determine the effectiveness of my model in predicting bank churn accurately.

## DESIGN:

- The design solution is based on the use of machine learning algorithms such as decision tree, KNN, logistic regression, random forest, and XGBoost to predict bank churn. My design rationales are based on the assumption that past behavior and demographic information are critical indicators of whether a bank customer will churn or not.
- One alternative design idea that was considered was to use a simple rule-based system that would identify at-risk customers based on a set of predetermined rules. However, this approach was ruled out because it would not be able to capture the complex patterns in customer behavior that can lead to churn. Another alternative design idea was to use a neural network-based model, but this was ruled out because it would require a larger amount of data and more computing resources than the machine learning-based models used in the project. Additionally, these algorithms can be more challenging to interpret than traditional machine learning algorithms.
- The design includes the physical design of a machine learning model that will be trained on historical data to predict bank churn. The interaction model involves inputting the customer's past behavior and demographic information into the model, which will then output a prediction on whether the customer will churn or not. The interface design will be a user-friendly dashboard that displays the results of the model, including the accuracy, AUC score, and confusion matrix.
- The physical design of this solution involves implementing the machine learning models in a programming environment such as Python, with the necessary libraries for data preprocessing, feature selection, model training, and evaluation. The models will be optimized using hyperparameter tuning techniques to improve their

accuracy, and the best performing model will be selected based on the evaluation metrics such as accuracy and AUC score.

- The interaction model of this design involves training the machine learning models on the historical data of bank customers and using the trained model to predict the likelihood of future customers to churn.
- The bank can integrate this machine learning model into its existing customer management system to predict the likelihood of customers churning and take proactive measures to retain them.
- The bank can use the insights gained from the analysis of customer behavior patterns to improve its customer service and product offerings, thereby reducing the likelihood of customers churning.
- The bank can use this design solution to identify the most influential factors contributing to customer churn and take steps to address those factors. This could involve providing personalized offers to customers, improving the quality of service, or addressing specific pain points that are leading to customer dissatisfaction.

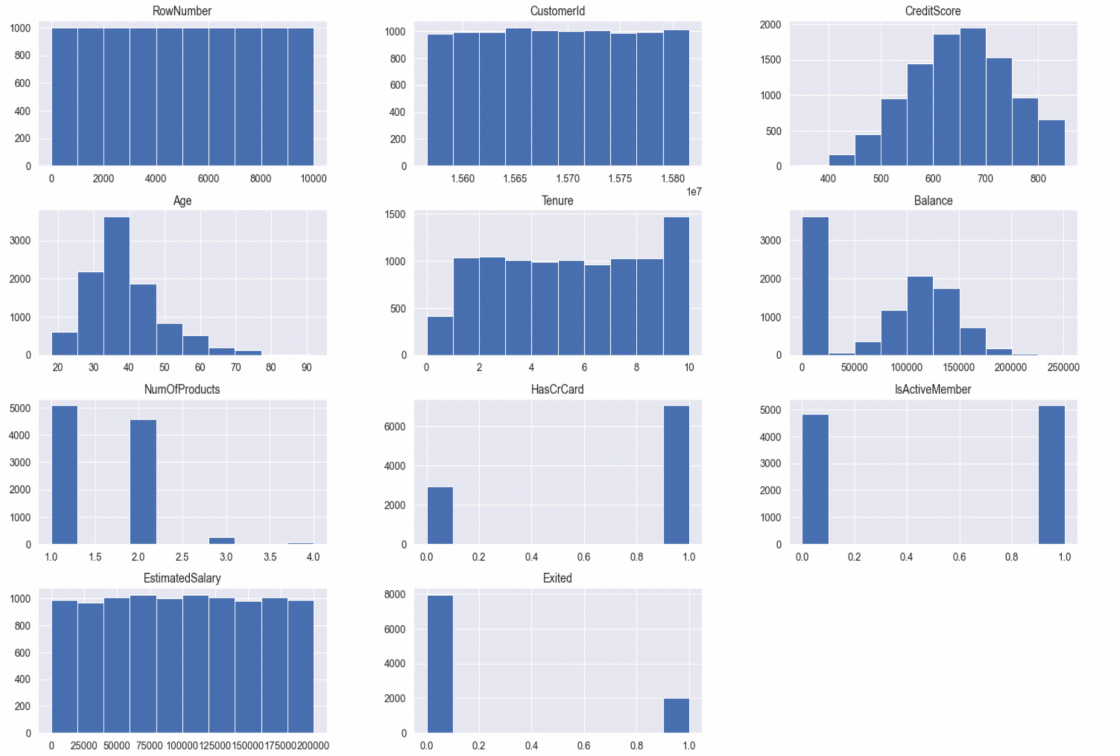
## Data Visualization

Visualizations that are used in this project are,

**Histogram:**

Plotting histograms for all attributes.

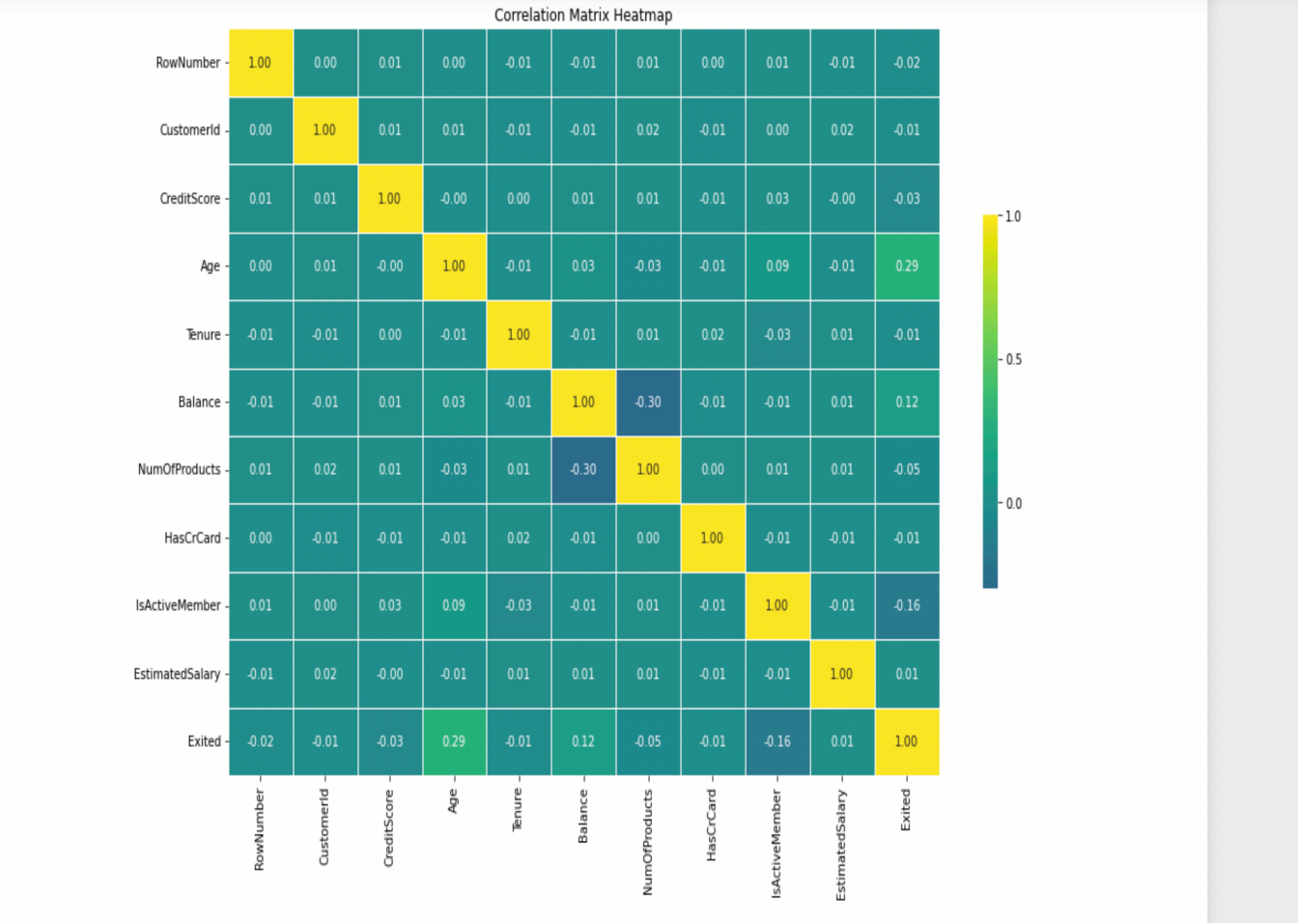
```
In [56]: data.hist(figsize=(25,15))
         plt.show()
```




**Heat Map:** Heatmaps are frequently used to represent the relationship between different variables in a dataset. The color of each cell in a correlation matrix represents the degree of correlation between the corresponding pair of variables.

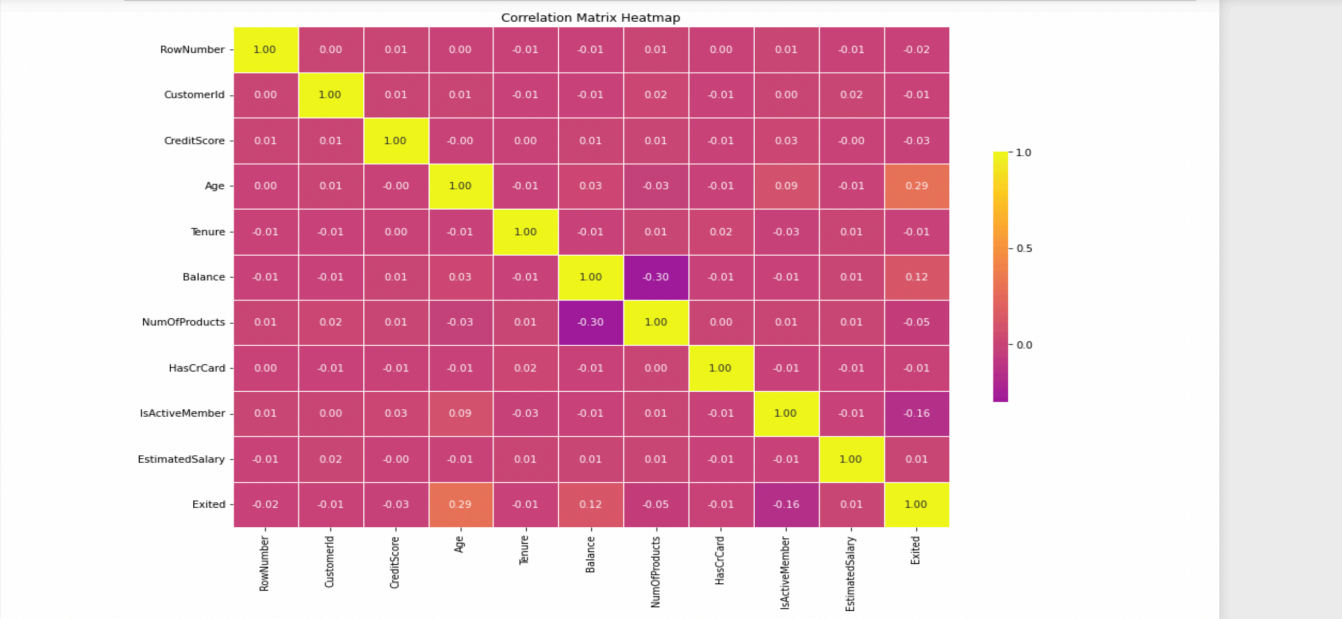
Viridis:

```
plt.figure(figsize=(15, 10))

sns.heatmap(data.corr(), cmap='viridis',center=0,annot=True,fmt='.2f',linewidths=0.5,
            cbar_kws={'shrink': 0.5, 'ticks': [-1, -0.5, 0, 0.5, 1]} )
plt.title('Correlation Matrix Heatmap')
plt.show()
```



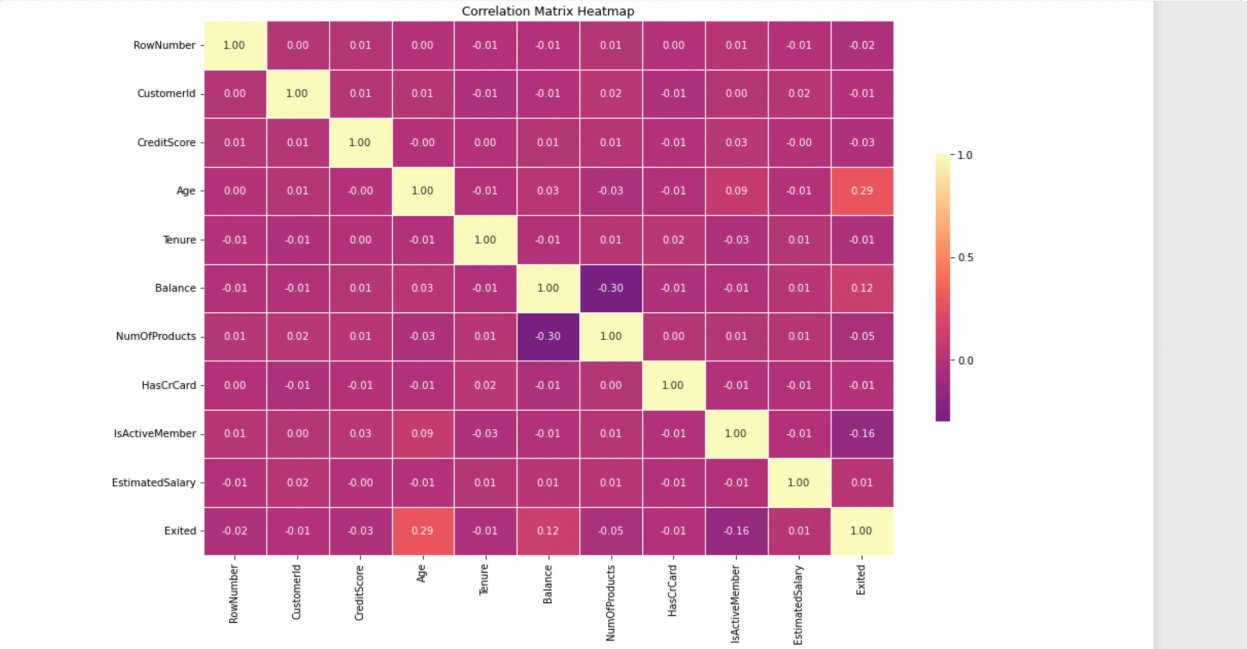

# Plasma:

```
plt.figure(figsize=(15, 10))

sns.heatmap(data.corr(), cmap='plasma',center=0,annot=True,fmt='.2f',linewidths=0.5,
            cbar_kws={'shrink': 0.5, 'ticks': [-1, -0.5, 0, 0.5, 1]} )
plt.title('Correlation Matrix Heatmap')
plt.show()
```



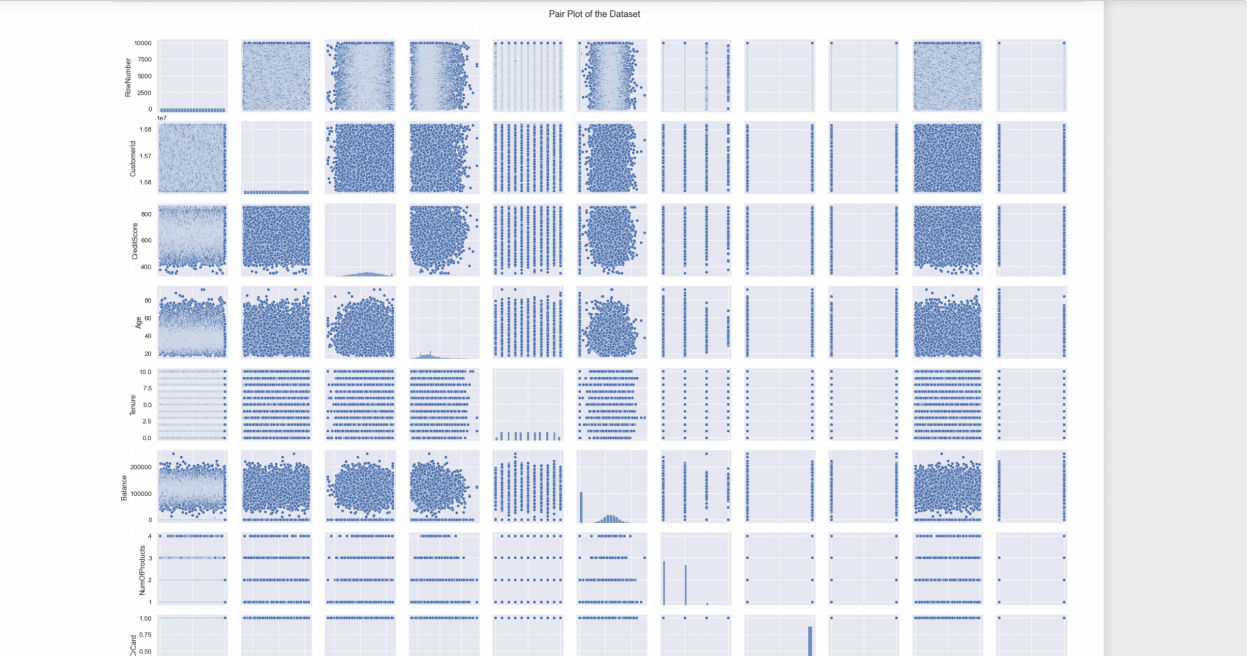

Magma:

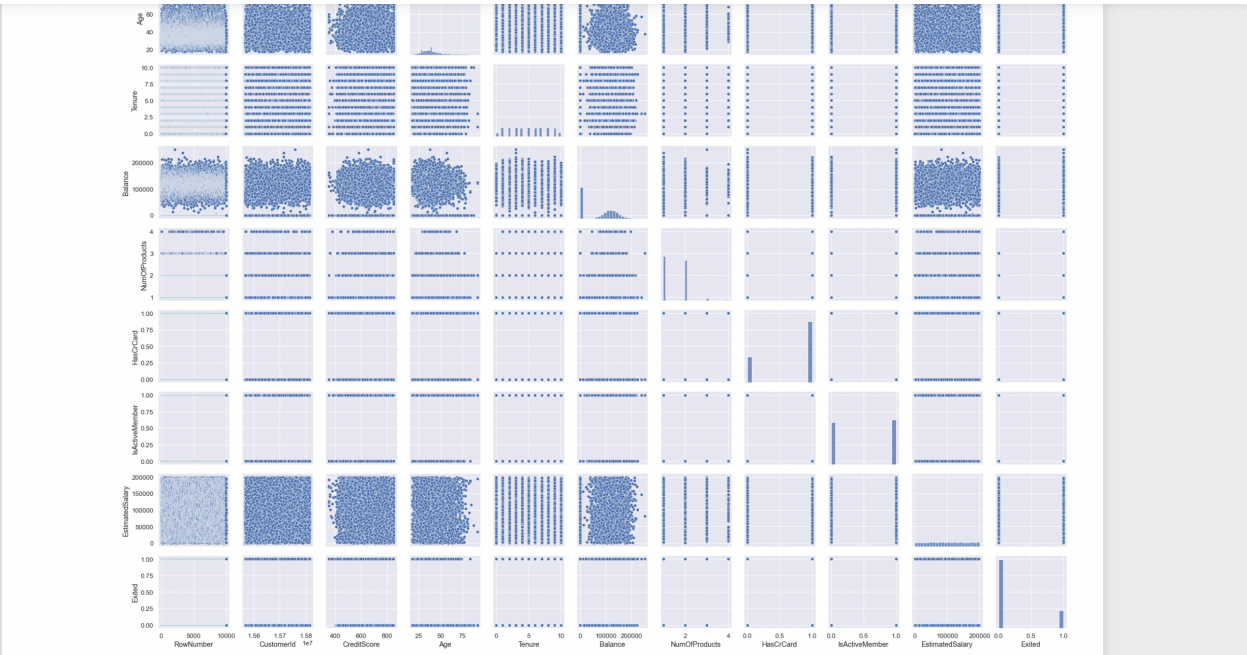
```
plt.figure(figsize=(15, 10))

sns.heatmap(data.corr(), cmap='magma',center=0,annot=True,fmt='.2f',linewidths=0.5,
            cbar_kws={'shrink': 0.5, 'ticks': [-1, -0.5, 0, 0.5, 1]} )
plt.title('Correlation Matrix Heatmap')
plt.show()
```




**Pair plot:** Pair plots allow us to visualize the relationships between multiple variables in a dataset. By plotting each variable against every other variable, we can quickly identify any correlations or patterns in the data.

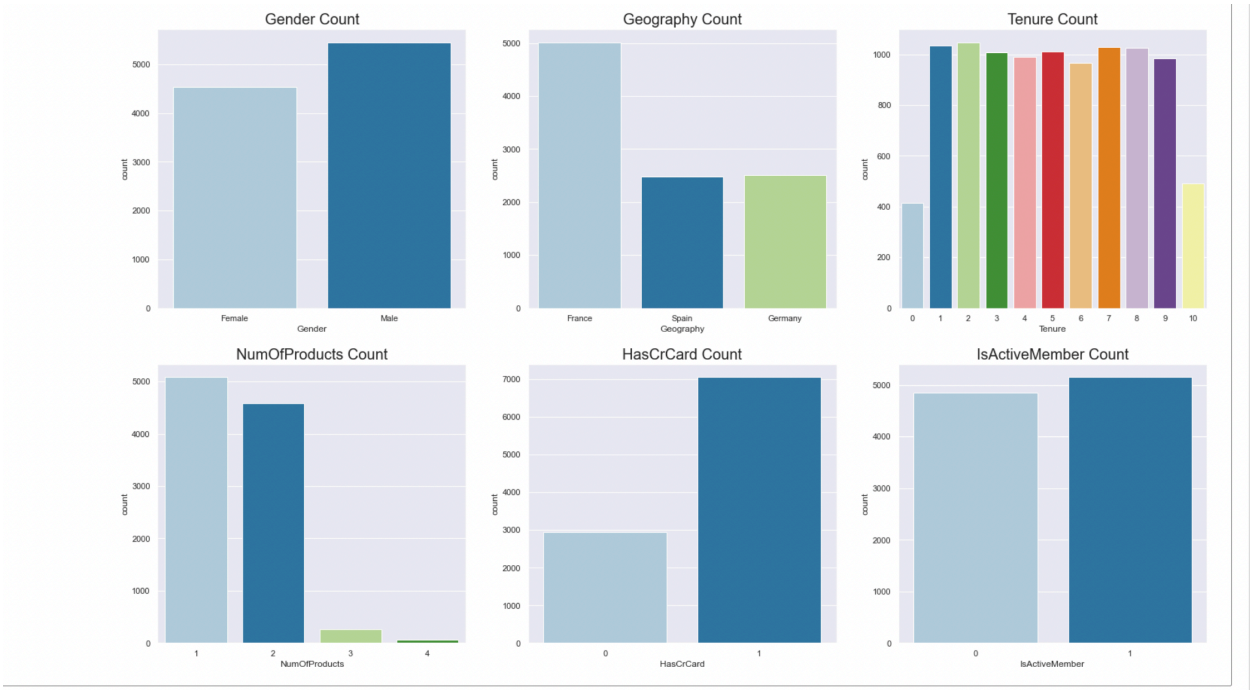
```
sns.pairplot(data)

# Adjust the plot aesthetics
sns.set(font_scale=1.2)
plt.subplots_adjust(top=0.95)

# Add a title to the plot
plt.suptitle('Pair Plot of the Dataset', fontsize=18)

# Show the plot
plt.show()
```



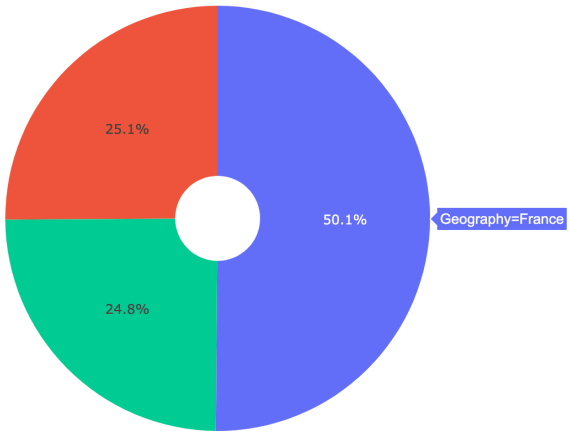
# Visualizing categorical variables:

```
# Visualizing categorical variables
fig, axs = plt.subplots(2, 3, figsize=(25, 15))
data_counts = [
    ('Gender', axs[0, 0]),
    ('Geography', axs[0, 1]),
    ('Tenure', axs[0, 2]),
    ('NumOfProducts', axs[1, 0]),
    ('HasCrCard', axs[1, 1]),
    ('IsActiveMember', axs[1, 2]),
]

# Create countplots
sns.set_theme(palette="Paired")
for data_count in data_counts:
    sns.countplot(data=data, x=data_count[0], ax=data_count[1])

# Setting the titles
axs[0, 0].set_title('Gender Count', size=20)
axs[0, 1].set_title('Geography Count', size=20)
axs[0, 2].set_title('Tenure Count', size=20)
axs[1, 0].set_title('NumOfProducts Count', size=20)
axs[1, 1].set_title('HasCrCard Count', size=20)
axs[1, 2].set_title('IsActiveMember Count', size=20)
plt.show()
```



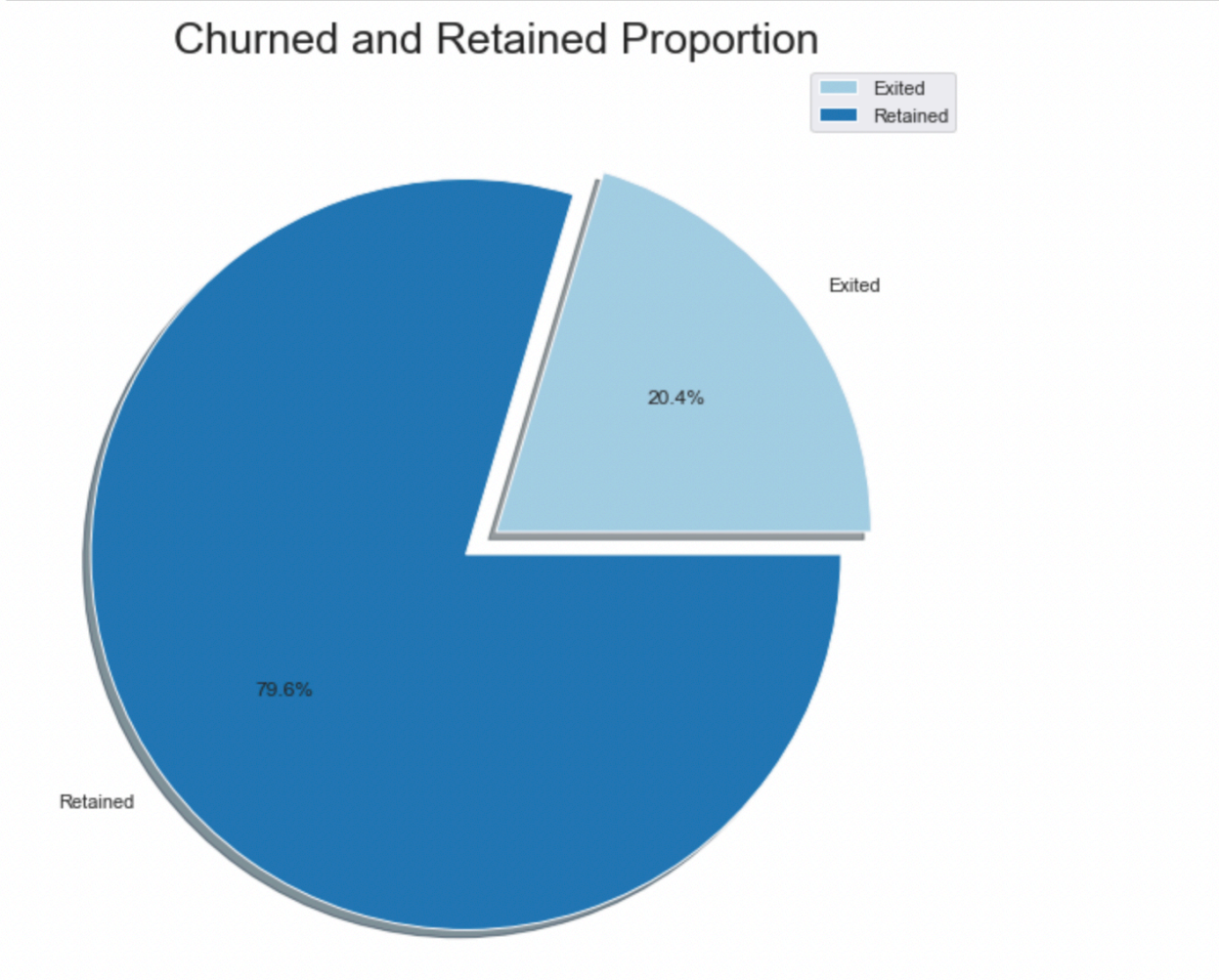
# Geography:

```
In [18]: px.pie(data, names='Geography', hole=0.2)
```




# Churned and Retained Proportion:

```
labels = ['Exited', 'Retained']
sizes = [data.Exited[data['Exited']==1].count(), data.Exited[data['Exited']==0].count()]
explode = [0, 0.1]
fig, ax = plt.subplots(figsize=(20, 10))
ax.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%', shadow=True)
ax.set_title("Churned and Retained Proportion", size=25)
handles, _ = ax.get_legend_handles_labels()
ax.legend(handles, labels, loc='best')
plt.show()
```




- The goal of using visualizations in this project is to help stakeholders understand the patterns and relationships in the data more easily and quickly than they would through raw data or tables. For example, the pair plot and heatmap could help identify which features are most correlated with churn, and the histogram could show the distribution of a particular feature among churned vs. non-churned customers.
- The effective data representations that could captivate your audience are pie graph that explains how much percentage of customers retained the bank and how much percentage of customers exited the bank. For finding the geography that is in this data I have used pie graph which shows the geography name and percentage of the data.

- I have chosen appropriate colors, fonts, and labels to make the visualizations easy to read and understand. Used consistent styles and formats across different visualizations to help users compare and contrast them easily. I have provided clear titles, captions, and legends to help users interpret the visualizations and understand what they are showing.

## Ethics:

- Some relevant values for designing with ethics in mind include fairness, transparency, privacy, and responsibility.
- The design should be transparent in how it uses the data and clear about any potential biases or limitations. It should prioritize privacy by not sharing personally identifiable information and being mindful of any potential harms to individuals or groups. It should be responsible by ensuring that the data is used in ways that align with ethical principles and by being willing to adjust the design or approach if needed.
- Ethical ways of using the data include using it for its intended purpose, ensuring that any analysis or conclusions are valid and reliable, and protecting the privacy of individuals. The ethical ways which we have used in our project is we have analyzed our data in a valid, reliable way and concluded. Un-ethical ways of using the data include using it for unintended purposes or in ways that harm individuals or groups, using faulty analysis or conclusions, or not protecting individuals' privacy. In our project we haven’t used unethical ways.
- It is possible that the design could disproportionately affect underserved, marginalized, low-resourced, and underrepresented populations, particularly if the factors used in the analysis are biased or discriminatory in some way but in this project, we have analyzed our model carefully without being bias on any data. To mitigate the risk of disproportionate impact, it is important to carefully consider the factors used in the analysis and to regularly review and adjust

the model as needed to ensure that it is fair and unbiased. It may also be beneficial to consult with experts in ethics and social justice to identify potential biases and develop strategies for addressing them.

- The design of data analysis and modeling is unlikely to have a significant impact on the world's environment, resources, and climate. However, it's worth noting that the collection and storage of data can have an impact on the environment, particularly in terms of energy usage and carbon emissions. Therefore, it's important to consider the environmental impact of data storage and processing infrastructure used in the project. Additionally, if the insights gained from the project lead to changes in the bank's business practices or customer behavior, there may be indirect environmental impacts to consider.
- The way to promote positive change in society would be to use the findings from the data analysis to improve customer service and overall customer experience. This could involve identifying pain points or areas of frustration for customers and working to address those issues. By improving customer satisfaction, the bank can not only reduce churn but also promote positive word-of-mouth and attract new customers.

## RESULTS:

| Models | Accuracy | AUC Score | Accuracy with Hyperparameter Tuning | AUC Score |
|---|---|---|---|---|
| Logistic Regression | 0.8125 | 0.78 | 0.811 | 0.78 |
| Decision Tree | 0.7905 | 0.69 | 0.836 | 0.77 |
| Random Forest | 0.8665 | 0.42 | 0.863 | 0.87 |
| XGBOOST | 0.861 | 0.85 | 0.8625 | 0.87 |
| KNN | 0.8235 | 0.74 | 0.823 | 0.74 |

Based on the results, the Random Forest and XGBoost models provided the highest accuracy and AUC scores, with both models showing an accuracy of around 86% and an AUC score of 0.87 after hyperparameter tuning. Overall, this project shows that machine learning algorithms can be effectively used to predict customer churn in the banking industry. Such models can help financial institutions to identify potential churners and take proactive steps to retain customers, thus improving customer satisfaction and reducing customer acquisition costs.