Machine learning project

CUSTOMER CHURN PREDICTION

Dataset Link :

https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction

https://raw.githubusercontent.com/AbhishekYadav-01/Encryptix/main/CUSTOMER%20CHURN%20PREDICTION/Churn_Modelling.csv
(Github Link)

Link of the google collab file :

https://colab.research.google.com/drive/1fufTbxiZaV17nsyfDb-eTwWX666vM50Q?usp=sharing

Github Repository Link :

https://github.com/AbhishekYadav-01/Encryptix/tree/main/CUSTOMER%20CHURN%20PREDICTION

Aim :

Develop a model to predict customer churn for a subscription based service or business. Use historical customer data, including features like usage behavior and customer demographics, and try algorithms Logistic Regression, Random Forests, or Gradient Boosting to predict churn.

Dataset Loading :

The dataset used in this project can be found on Kaggle and GitHub:

- Kaggle: Customer Churn Prediction Dataset
- GitHub: Customer Churn Prediction Dataset

The dataset includes the following features :

- **Customer Information:** RowNumber, CustomerId, Surname
- **Demographics:** Geography, Gender, Age
- **Account Information:** CreditScore, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary
- **Target Variable:** Exited (1 if the customer has churned, 0 otherwise)

Data Preprocessing :

- **Handling Missing Values:** No missing values were present in the dataset.
- **Encoding Categorical Variables:** 'Geography' and 'Gender' were encoded using LabelEncoder.
- **Feature Scaling:** Numerical features were scaled using StandardScaler.

This is how the dataset looks like :

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

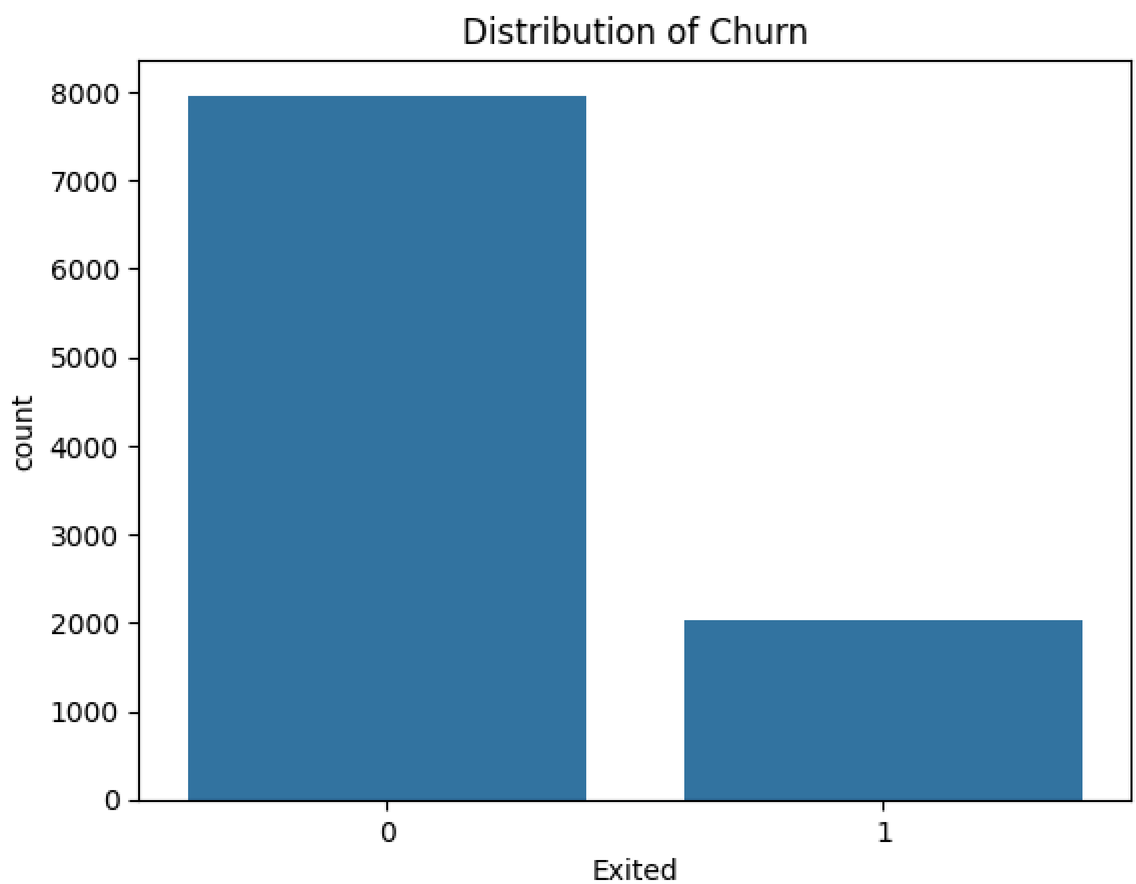
This is the mathematical information of the data :

	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

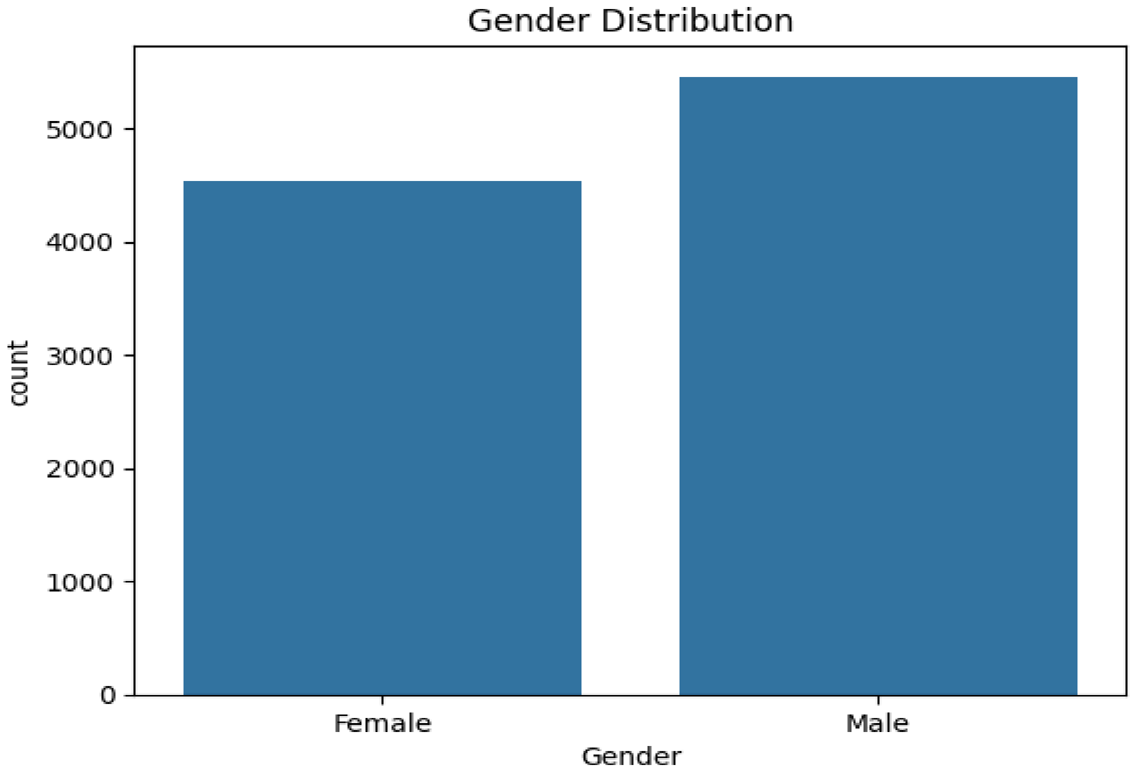
Checking for missing values :

```
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
```

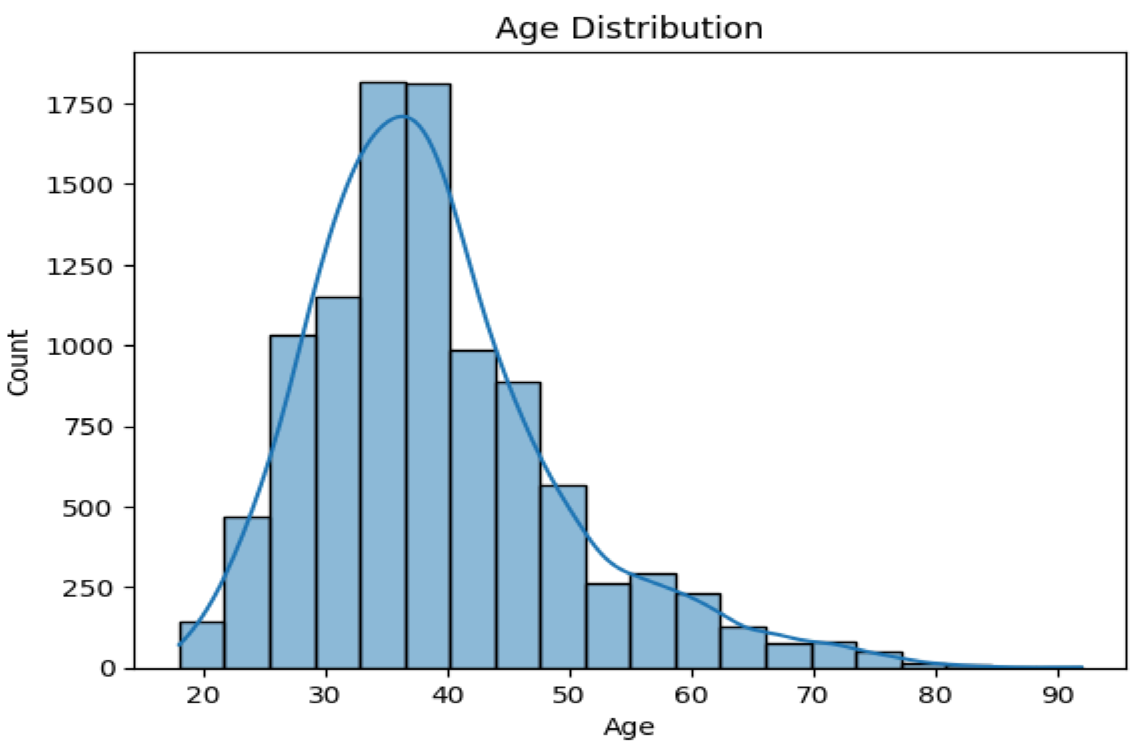
This is the distribution of the Churm (Count vs Excited) :



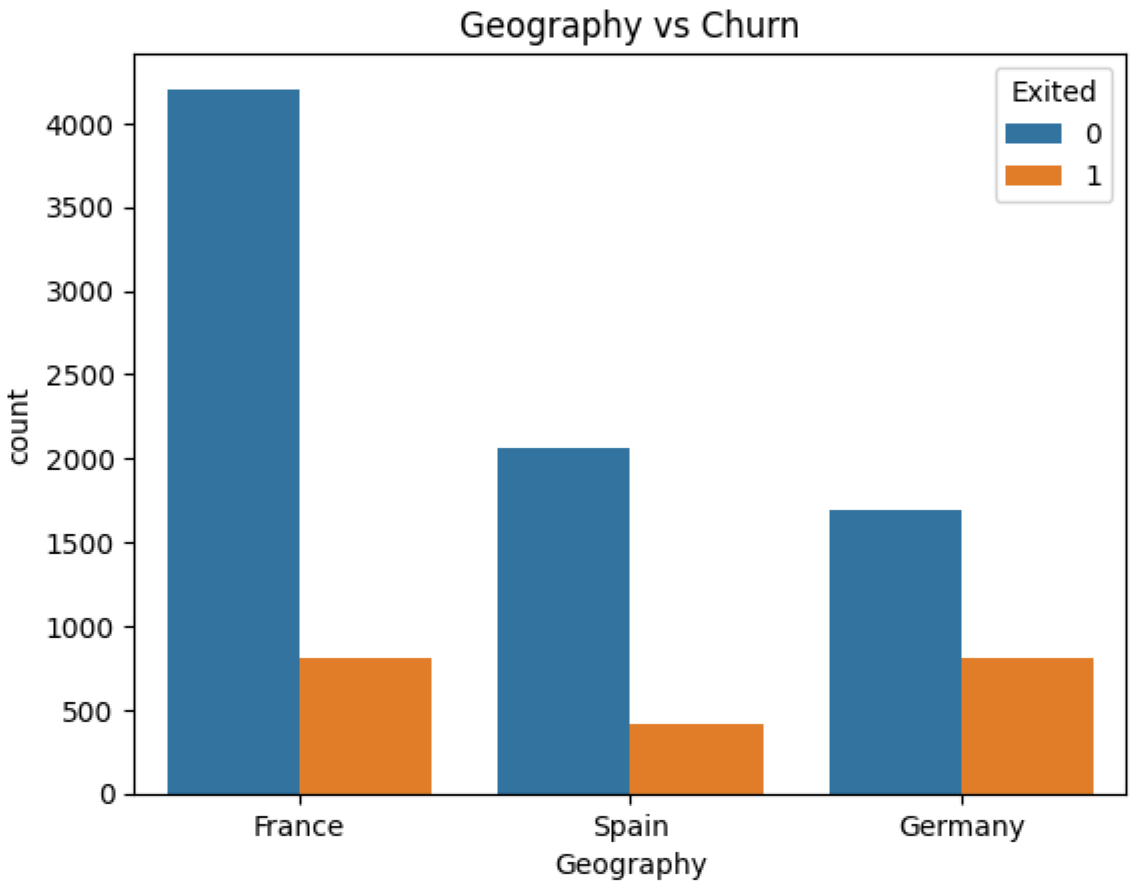
This is the distribution of the Gender (Count vs Gender) :



This is the distribution of the Age(Count vs Age) :



This is the graph of the Geography vs Churm :



Models Used :

These three models were trained using data:

- Logistic Regression
- Random Forest Classifier
- Gradient Boosting Classifier

Data Splitting :

The dataset was split into training (75%) and test (25%) sets to validate model performance.

Model Training :

Each model was trained on the training data to learn patterns and relationships in the features that predict churn.

1.Logistic Regression

- **Training:** This model i have trained using the fit method on the training data.

2.Random Forest Classifier :

- **Training:** This model i have trained using the fit method on the training data.

3.Gradient Boosting Classifier :

- **Training:** This model i have trained using the fit method on the training data.

Model Evaluation :

This is the some Actual vs Predicted for some data :

Logistic Regression :

```
status :
      Actual  Predicted
3555       1          0
4078       0          0
8445       0          0
5939       0          0
5583       0          0
1656       0          0
5550       0          0
1736       0          0
6297       0          0
```

```
6364       0          0
```

Random Forest:

```
status :
      Actual  Predicted
3555       1          0
4078       0          0
8445       0          0
5939       0          0
5583       0          0
1656       0          0
5550       0          0
1736       0          0
6297       0          0
6364       0          0
```

Gradient Boosting :

```
status :
      Actual  Predicted
3555       1          0
4078       0          0
8445       0          0
5939       0          0
5583       0          0
1656       0          0
5550       0          0
1736       0          0
6297       0          0
6364       0          0
```

Accuracy Scores :

After testing on 25 % of the data we got this much of the accuracies :

```
We got Logistic Regression Accuracy: 0.8124
We got Random Forest Accuracy: 0.8644
We got Gradient Boosting Accuracy: 0.8712
```

Results of the each models :

1. Logistic Regression Classification Report :

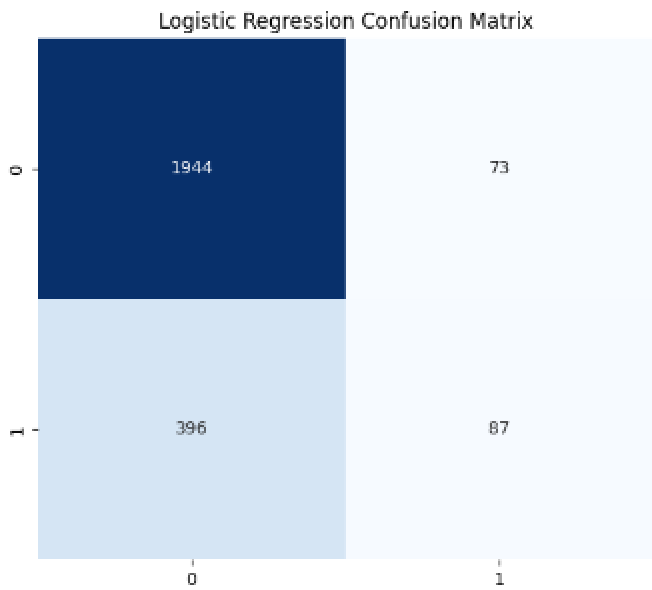
```
              precision    recall  f1-score   support

           0       0.83      0.96      0.89      2017
           1       0.54      0.18      0.27       483

    accuracy                           0.81      2500
   macro avg       0.69      0.57      0.58      2500
weighted avg       0.78      0.81      0.77      2500
```

The Logistic Regression model has a high precision (0.83) and recall (0.96) for predicting non-churners (class 0), but it struggles with predicting churners (class 1), with a low recall of 0.18. This indicates the model is more likely to predict non-churn than churn, leading to more false negatives.

Confusion Matrix of Logistic Regression model :

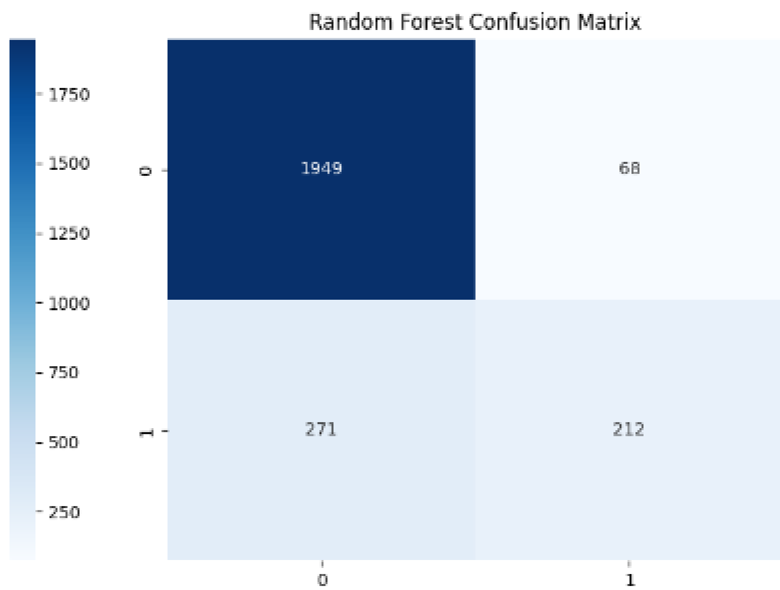


2.Random Forest Classification Report :

	precision	recall	f1-score	support
0	0.88	0.97	0.92	2017
1	0.76	0.44	0.56	483
accuracy			0.86	2500
macro avg	0.82	0.70	0.74	2500
weighted avg	0.85	0.86	0.85	2500

The Random Forest model performs better than Logistic Regression, with a higher overall accuracy (0.8644). It has a good balance between precision and recall for both classes. The recall for class 1 (churners) is 0.44, indicating better detection of churners compared to Logistic Regression.

Confusion Matrix of Random Forest model:



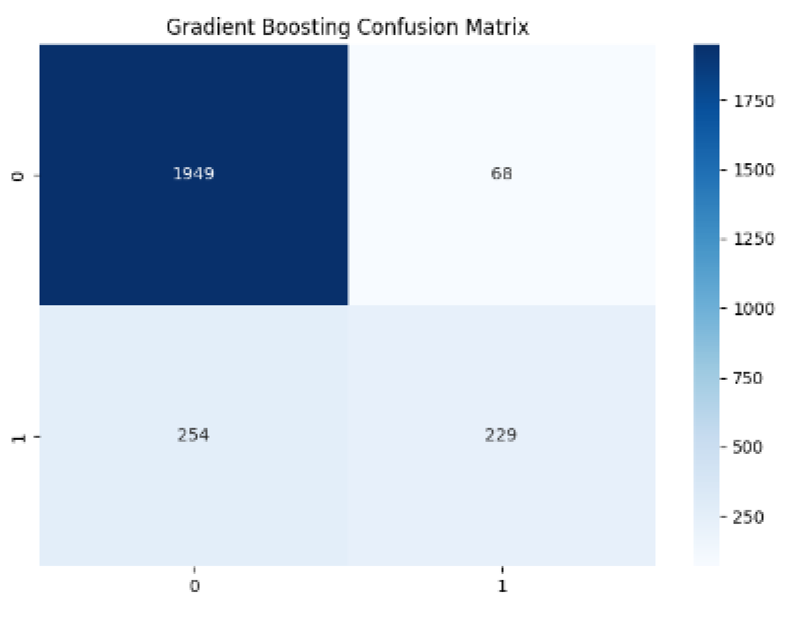
3.Gradient Boosting Classification Report :

	precision	recall	f1-score	support
0	0.88	0.97	0.92	2017
1	0.77	0.47	0.59	483
accuracy			0.87	2500
macro avg	0.83	0.72	0.76	2500

```
weighted avg       0.86      0.87      0.86      2500
```

The Gradient Boosting model achieved the highest accuracy (0.8712). It also has a high precision (0.88) and recall (0.97) for class 0, and improved performance for class 1 (precision of 0.77 and recall of 0.47), indicating a better balance in predicting both churners and non-churners.

Confusion Matrix of Gradient Boosting model:



Conclusion :

The project successfully developed and evaluated models for predicting customer churn, with Gradient Boosting emerging as the most accurate model, offering valuable insights for customer retention strategies. Gradient Boosting achieved the highest accuracy of 87.12%, indicating its effectiveness in predicting customer churn. It showed high precision (0.88) and recall (0.97) for non-churners and improved performance for churners (precision of 0.77 and recall of 0.47), demonstrating a balanced prediction for both classes. Random Forest also performed well with an accuracy of 86.44%, achieving a good balance between precision and recall for both classes and improving the detection of churners (recall of 0.44). Logistic Regression provided baseline performance with an accuracy of 81.24%, showing high precision (0.83) and recall (0.96) for predicting non-churners but struggling with predicting churners, leading to more false negatives due to a low recall of 0.18.