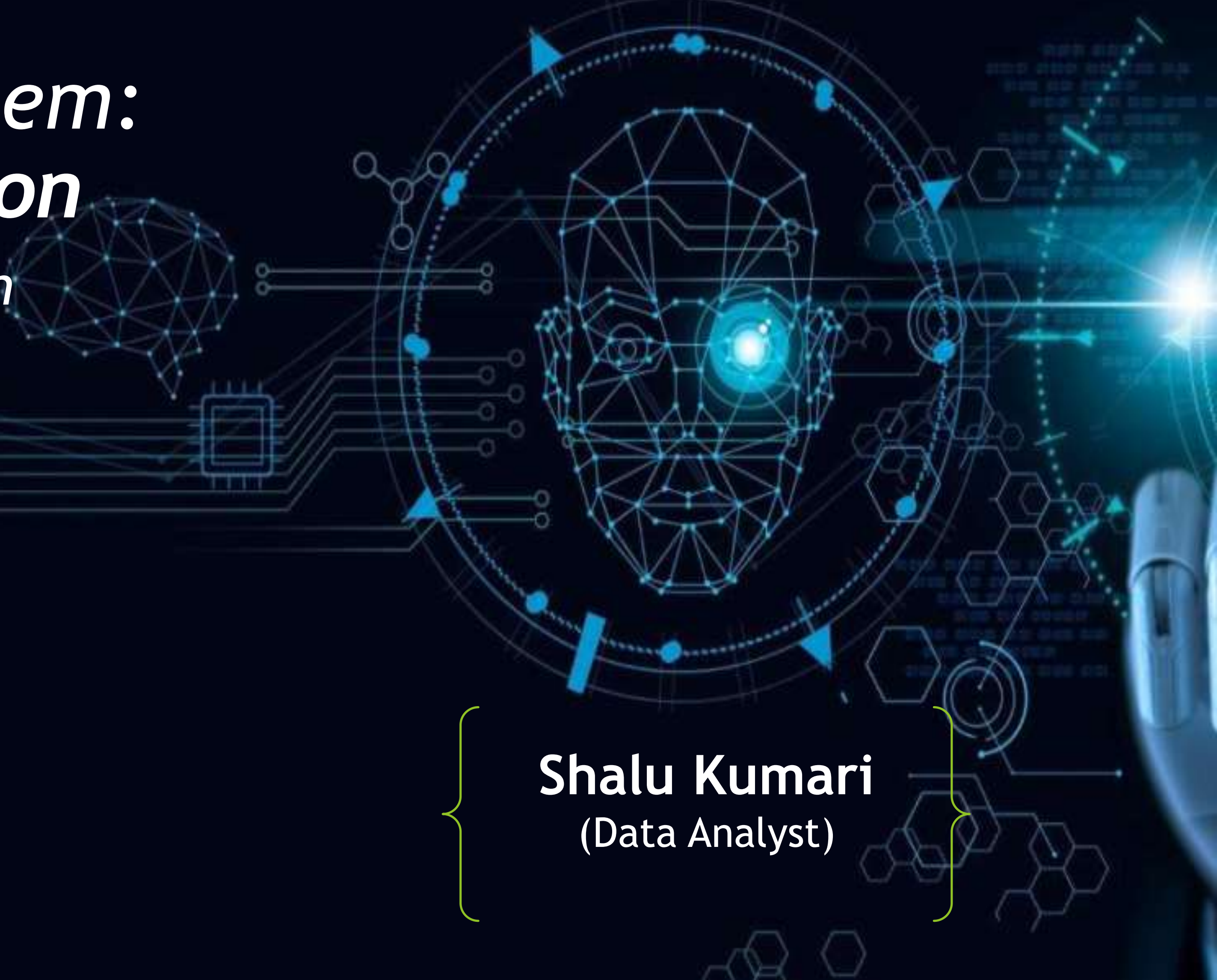
Solving the Business Problem: ***Customer Churn Prediction***

Improving Customer Retention Through Predictive Analytics



Shalu Kumari
(Data Analyst)

Problem Statement :

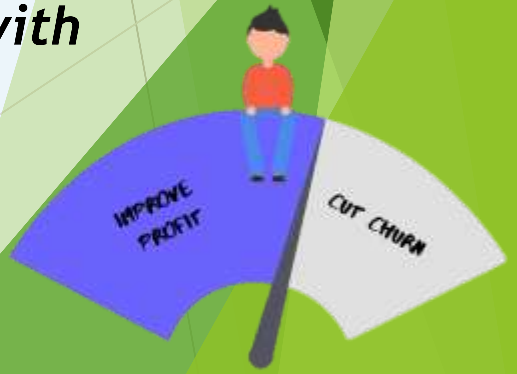
Customer churn, is a significant challenge for businesses in various industries.

Customer Churn Rate = $\left(\frac{\text{Lost Customers}}{\text{Total Customers at the Start of Time Period}} \right) \times 100$

Losing customers can lead to <u>reduced revenue and market share</u>, highlighting the importance of <u>accurately predicting</u> and mitigating churn

In the context of banks, customer churn specifically refers to the phenomenon where account holders close their accounts, discontinue using banking services.

The objective of this project is to predict whether a customer will continue with their account or close it (i.e., churn).



Dataset Description :

The dataset contains information on bank customers, including those who have churned (Exited = 1) and those who continue to be customers (Exited = 0).
Key dataset specifications: 165,034 rows & 14 columns
The dataset includes the following attributes:

Attribute	Description
id	A unique identifier for each customer.
Surname	The customer's surname or last name.
Credit Score	A numerical value representing the customer's credit score.
Geography	The country where the customer resides (France, Spain, or Germany).
Gender	The customer's gender (Male or Female).
Age	The customer's age.
Tenure	The number of years the customer has been with the bank.
Balance	The customer's account balance.
NumOfProducts	The number of bank products the customer uses (e.g., savings account, credit card).
HasCrCard	Whether the customer has a credit card (1 = yes, 0 = no).
IsActiveMember	Whether the customer is an active member (1 = yes, 0 = no).
EstimatedSalary	The estimated salary of the customer.
Exited	Whether the customer has churned (1 = yes, 0 = no).

Type of Machine Learning Task

- **Supervised Learning - Classification Task**

  As we haven to classify whether 1 or 0 (churned or not) on the basis of given feature (‘Exited’)

Overview :

In this project, I aimed to predict customer churn in a banking dataset using machine learning techniques. I performed extensive data analysis, including **exploratory data analysis** (EDA) to understand the characteristics of the dataset, and then proceeded with **data preprocessing**, **model building**, and **evaluation**.

i built three classification models: **Logistic Regression, Random Forest, and XGBoost,** and evaluated their performance using various techniques:

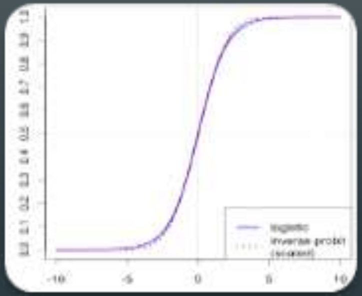
- **Normal**: without any balancing technique
- **Over-sampling**: using SMOTE (Synthetic Minority Over-sampling Technique.)
- **Under-sampling**: using RandomUnderSampler

I found that **over-sampling produced the best results**, particularly for identifying churners. Among the models, **XGBoost consistently performed the best**, offering a good balance of precision and recall for identifying churners.
Additionally, i visualized the performance of the models using **confusion matrices and classification report.**

Overall, this project demonstrates the importance of **addressing class imbalance in predictive modeling tasks**, especially in scenarios like customer churn prediction, where identifying minority classes (churners) accurately is crucial for **business decision-making** so to retain the customer as much as possible.

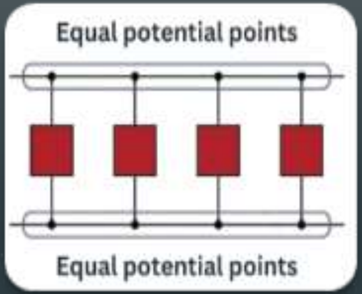
Algorithm Used :

Logistic Regression



- Logistic regression is like drawing a straight line through data points to separate them into two groups (churned and not churned)
- Why? : As it straightforward and easy to interpret hence due to its simplicity and interpretability is a good choice a for churn predictions.

Random Forest



- Random forest is like asking a bunch of friends for advice, then making a decision based on the most popular answer.
- Why ? : captures complex relationships in the data, is less prone to overfitting and requires minimal feature engineering

XG Boost (Extreme Gradient Boost)

- XGBoost is like a team of experts working together to solve a problem, with each expert focusing on a specific aspect.
- It builds a sequence of decision trees, where each tree corrects the errors made by the previous ones. It combines the predictions of all trees to make a final prediction.
- Why? : Due to it's speed and high performance

EDA (Exploratory Data Analysis) :

Data Cleaning

Dropped Unnecessary columns - "CustomerId" , "Surname"

No nulls & duplicates were found.

Handled outliers by capping extreme values at the 95th percentile - "CreditScore", "Age" & "Balance"

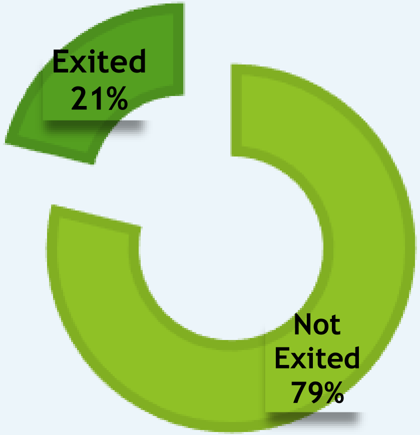
Univariate Analysis

Dataset is **Imbalanced** ("Exited") -

- 0 - 79% (approx.)
- 1 - 21% (approx.)

Insights into

- gender distribution - 2
- number of products - 4
- Tenure - 10
- Geography - 3



Bivariate Analysis

Explored the impact of **categorical variables** (e.g., credit card ownership, active membership) **& numerical variables** (e.g., credit score, balance, estimated salary) on **customer churn.**

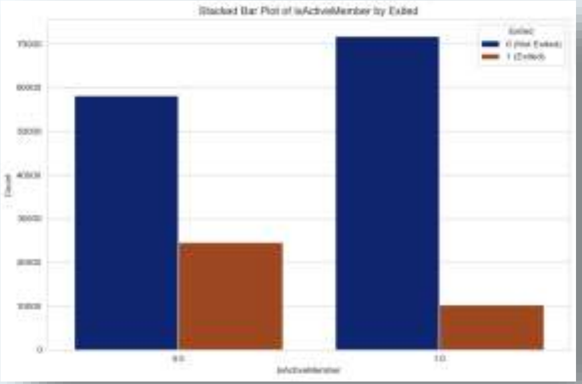
Found that **higher credit scores** & **larger A/c balances doesn't gaurantee loyalty**

Customer churn analysis highlighted that customers **aged 50-60**, **females**, and those with a **tenure** of **0-2 years** exhibited the **highest churn rates.**

Multivariate Analysis

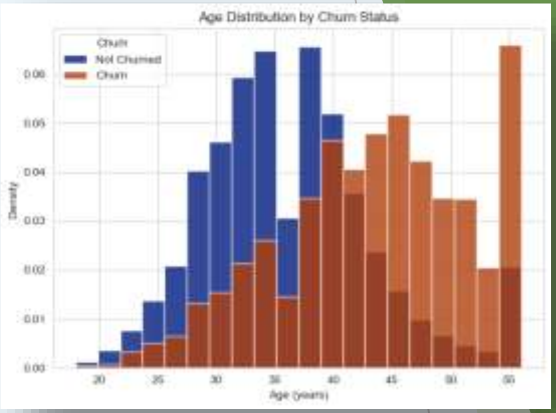
Collinearity among **variables** was checked using a correlation heatmap, **revealing no strong multicollinearity** issues.

Key Findings of EDA :



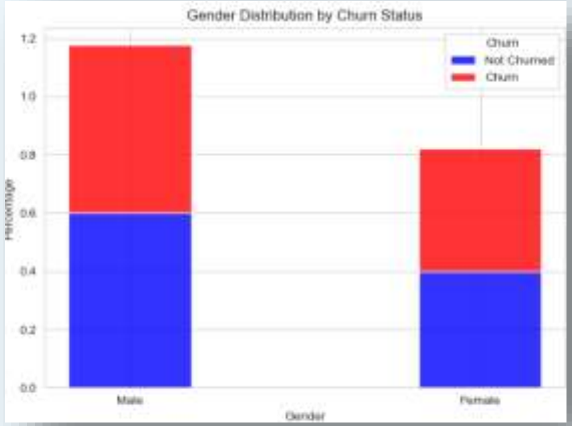
Important factors influencing customer churn:

Age, Gender, Tenure, Credit score, A/c balance, and active status

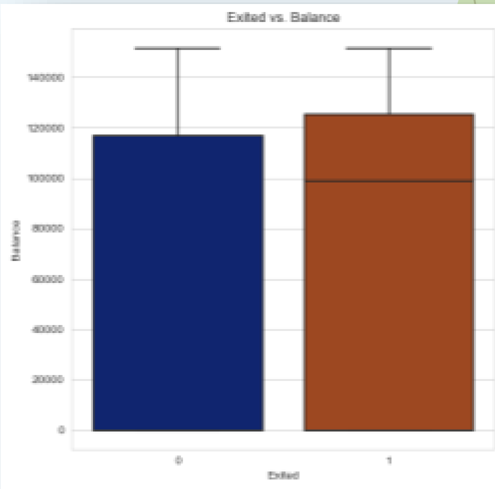


More likely to churn :

Females & customers with **shorter tenure**



Customer churn tends to occur more frequently among those with **higher account balances may be due to dissatisfaction or better offers** elsewhere.



Data Preprocessing :

Converted Categorical Variables to numerical:

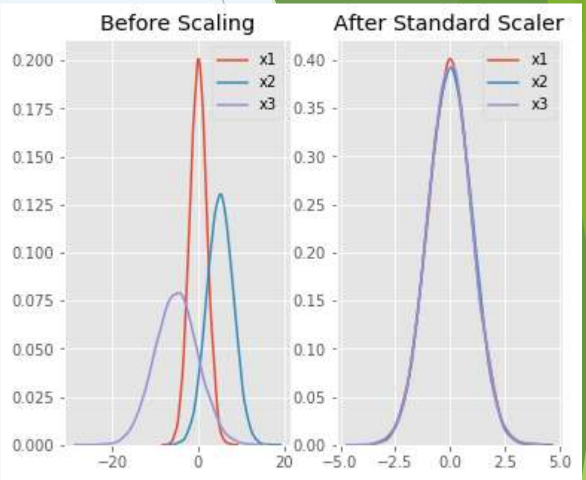
"Gender" : Transformed into numerical representation **by replacing** male as 1 and female as 0

"Geography": Utilized **one-hot encoding** to represent categorical values as binary.

Standard Scaling:

Performed standard scaling to **normalize numerical features.**

Ensures all variables are on a similar scale, preventing features with larger magnitudes from dominating the model.



id	Gender	Gender2
165029	Female	0
165030	Male	1
165031	Male	1
165032	Female	0
165033	Male	1

Geography_France	Geography_Germany	Geography_Spain
True	False	False
True	False	False
True	False	False
True	False	False
False	False	True

Geography_France	Geography_Germany	Geography_Spain
1	0	0
1	0	0
1	0	0
1	0	0
0	0	1

Model Building :

In model building , machine learning algorithms are **trained on historical data** to make predictions.
ML Algorithm that I have used is **Logistic Regression, Random Forest & XG Boost**

Three techniques were employed for building and evaluating predictive models for customer churn prediction:

Normally

Models built without any balancing technique

Over-Sampling

Adding **more copies of the minority class** to balance the dataset.

SMOTE (Synthetic Minority Over-sampling Technique) -

a smarter way to oversample, it creates synthetic samples that are similar to the existing minority class samples.

Under-Sampling

trimming down the majority class to match the minority class

Random Under Sampler do under-sampling by randomly removing samples from the majority class.

Evaluation Matrix of Classification Task -

Confusion Matrix

(summary of correct and incorrect predictions)

- **True Positives (TP):** Instances that were correctly classified as positive.
- **True Negatives (TN):** Instances that were correctly classified as negative.
- **False Positives (FP):** Instances that were incorrectly classified as positive.
- **False Negatives (FN):** Instances that were incorrectly classified as negative.

Classification Report

- **Precision:** The proportion of correctly predicted instances of a class out of all instances predicted as that class
- **Recall :** The proportion of correctly predicted instances of a class out of all instances that truly belong to that class.
- **F1- score :** It is a combination of precision and recall into a single value. It gives you a balanced measure of how well model is performing.
- **Accuracy :** the proportion of correctly classified instances out of the total instances.

Evaluation of Models (without balancing) :

Model: Logistic Regression

	precision	recall	f1-score	support
0	0.86	0.95	0.90	26023
1	0.71	0.41	0.52	6984
accuracy			0.84	33007
macro avg	0.78	0.68	0.71	33007
weighted avg	0.83	0.84	0.82	33007

Model: Random Forest

	precision	recall	f1-score	support
0	0.88	0.94	0.91	26023
1	0.72	0.53	0.61	6984
accuracy			0.86	33007
macro avg	0.80	0.74	0.76	33007
weighted avg	0.85	0.86	0.85	33007

Model: XGBoost

	precision	recall	f1-score	support
0	0.89	0.95	0.92	26023
1	0.73	0.55	0.63	6984
accuracy			0.86	33007
macro avg	0.81	0.75	0.77	33007
weighted avg	0.86	0.86	0.86	33007

Insights :

Logistic Regression: Performs slightly lower than other models.

Random Forest: Offers a balanced precision and recall.

XG Boost: Excels at identifying recall, crucial for targeting customer retention efforts.

Overall Insights:

All models achieve good accuracy (84% to 86%), but accuracy alone isn't sufficient due to data imbalance.

XG Boost demonstrates the highest average performance across precision, recall, and F1-score.

XG Boost notably stands out for its higher recall in identifying churning customers (55%).

Evaluation of Models (with Over- Sampling) :

Over-Sampling Model: Logistic Regression

	precision	recall	f1-score	support
0	0.91	0.76	0.83	26023
1	0.45	0.74	0.56	6984
accuracy			0.75	33007
macro avg	0.68	0.75	0.69	33007
weighted avg	0.82	0.75	0.77	33007

Over-Sampling Model: Random Forest

	precision	recall	f1-score	support
0	0.90	0.90	0.90	26023
1	0.64	0.63	0.63	6984
accuracy			0.84	33007
macro avg	0.77	0.77	0.77	33007
weighted avg	0.84	0.84	0.84	33007

Over-Sampling Model: XGBoost

	precision	recall	f1-score	support
0	0.90	0.92	0.91	26023
1	0.68	0.62	0.65	6984
accuracy			0.86	33007
macro avg	0.79	0.77	0.78	33007
weighted avg	0.85	0.86	0.86	33007

Insights:

•**Logistic Regression**: Lowest F1-score, highest precision for non-churners, lower recall for churners.
•**Random Forest**: Good recall but slightly lower precision compared to others.
•**XGBoost**: Maintains lead, excels in both precision and recall for churners, balanced performance.

Overall Performance:

•**Effect of Over-sampling**: Improved performance for all models, especially in recalling churners.
•**XG Boost Dominance**: Best performer overall, highest F1-score, balanced precision and recall.
•**Random Forest**: Follows closely behind XG Boost in F1-score.
•**Logistic Regression**: Lowest F1-score but highest precision for non-churners.

Evaluation of Models (with Under-Sampling) :

Under-Sampling Model: Logistic Regression

	precision	recall	f1-score	support
0	0.92	0.76	0.83	26023
1	0.45	0.74	0.56	6984
accuracy			0.75	33007
macro avg	0.68	0.75	0.69	33007
weighted avg	0.82	0.75	0.77	33007

Under-Sampling Model: Random Forest

	precision	recall	f1-score	support
0	0.93	0.81	0.86	26023
1	0.52	0.78	0.62	6984
accuracy			0.80	33007
macro avg	0.73	0.79	0.74	33007
weighted avg	0.84	0.80	0.81	33007

Under-Sampling Model: XGBoost

	precision	recall	f1-score	support
0	0.94	0.81	0.87	26023
1	0.53	0.79	0.63	6984
accuracy			0.81	33007
macro avg	0.73	0.80	0.75	33007
weighted avg	0.85	0.81	0.82	33007

Insights :

- **Logistic Regression**: Performance remains consistent with over-sampled model.
- **Random Forest**: Precision for churned class improved slightly, accompanied by increased accuracy but decreased recall.
- **XGBoost**: Similar to Random Forest with slightly improved precision and accuracy but decreased recall.

Overall Analysis:

- **Effectiveness of Under-Sampling**: Less beneficial compared to over-sampling due to decreased recall, crucial for retention efforts.
- **Considerations for Model Selection**: Under-sampling may not offer the desired balance between precision and recall for identifying churning customers.

Final Outcome :

Final Decision: **Over-Sampling Technique**

- Improved recall for churners, crucial for customer retention.
- Maintained overall accuracy.
- Balanced performance demonstrated by XG Boost's F1 score.

Selecting the Best Model:
Considering objective- **Maximize retention and prioritize catching every potential churner**.

XG Boost exhibits high accuracy, good precision for non-churners, and reasonable recall for churners.

- Missing a churner is costlier than reaching out to a non-churner who stays, making high recall a priority.

Comparison Table

	Logistic Regression	Random Forest	XGBoost
True Positives	5154	5449	**5527**
True Negatives	19742	20992	**21096**
False Positives	6281	5031	**4927**
False Negatives	1830	1535	**1457**
Class 0 -			
Precision	**0.91**	0.90	0.90
Recall	**0.76**	0.90	0.92
F1-Score	0.83	0.90	**0.91**
Class 1 -			
Precision	**0.45**	0.64	0.68
Recall	**0.74**	0.63	0.62
F1-Score	0.56	0.63	**0.65**
Accuracy	0.75	0.84	**0.86**
Strengths	High precision & recall for non-churners	High accuracy & recall for non-churners	Balanced performance, good recall for churners
Weaknesses	Lower recall for churners	Lower recall for churners	Lower recall for churners
Best for	Minimizing false positives, high precision	Balanced approach, high accuracy & recall	Balanced approach, good recall for churners

Further Improvement :

- To further improve our model, we can explore additional feature engineering techniques, such as creating new features or incorporating external data sources.
- We can also experiment with different model architectures and hyperparameter tuning techniques to optimize the model's performance.
- Additionally, continuous monitoring and updating of the model with new data can help ensure its effectiveness over time.

CONCLUSION

This project demonstrates the importance of predictive modeling in identifying and mitigating customer churn in the banking sector. By leveraging machine learning techniques, banks can proactively retain customers and maximize revenue.

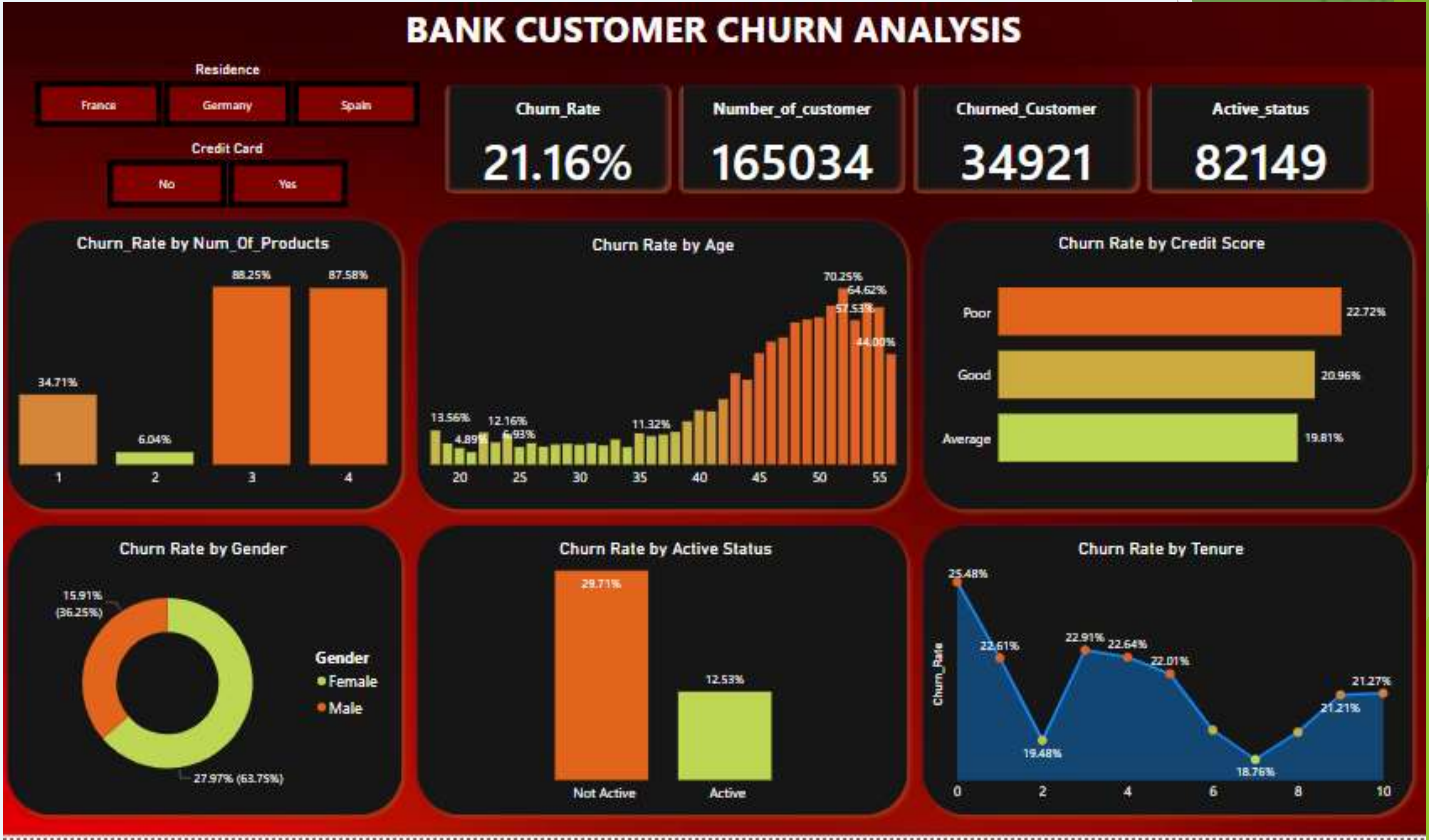
And chosen model, XGBoost, offers a balanced approach with high accuracy and recall, making it a valuable tool for banks in their customer retention efforts.

Application of this model in Other Industries :

- It has diverse industry applications beyond banking, including telecommunications, e-commerce, and subscription-based services.

- Model can be adapted to to predict customer churn in various sectors by customizing features and refining algorithms based on industry-specific attributes..

Power Bi Dashboard



Key insights :

•The churn rate for credit card customers is 21.16%.

•There is a positive correlation between the number of products a customer has and their churn rate.

•Customers with a poor credit score are much more likely to churn than customers with an average credit score.

•Active customers are less likely to churn than inactive customers.

•Customers with a tenure of less than 1 year are more likely to churn than customers with a longer tenure and many more.....

Thank You!