Bank Churn Prediction and Analysis project

# Name:

Back Churn Predication and Analysis Project

# Business Problem:

Customer churn—the phenomenon of customers leaving the bank for competitors—is a significant challenge for financial institutions. Churn can result in lost revenue, increased customer acquisition costs, and an overall decline in customer loyalty. Understanding the reasons behind why customers leave (churn) and why they stay (non-churn) is essential for developing effective strategies to reduce churn rates and increase customer retention.

This project focuses on identifying the factors contributing to customer churn and using predictive modeling techniques to classify customers at risk of leaving the bank. By analyzing customer demographics, account details, and behaviors, we aim to gain valuable insights that will guide the bank in developing targeted retention strategies and improving customer satisfaction.

# Business Impact:

The analysis of customer churn and predicate can provide valuable insights that help the bank retain its customers. By identifying key factors influencing churn, the bank can implement targeted strategies to improve customer satisfaction and engagement. This could result in a lower churn rate, increased customer loyalty, and enhanced overall profitability. Additionally, understanding why customers leave can help the bank better allocate resources and optimize its service offerings.

# Dataset(s):

Bank Churn Prediction Dataset – Kaggle

The dataset for this project is available on Kaggle under the title Bank Churn Prediction. You can access the dataset here:   
 [Bank Churn Prediction Dataset](https://www.kaggle.com/datasets/bonginstates/bank-churn-prediction/data)

**Target Features:**

The columns that will be used to simplify and analyze the problem statement are:

- CustomerId: Unique identifier for each customer.  
 - Surname: Last name of the customer.  
 - CreditScore: A measure of the customer's credit history.  
 - Geography: The country or region of the customer.  
 - Gender: The gender of the customer.  
 - Age: Age of the customer.  
 - Tenure: Number of years the customer has been with the bank.  
 - NumOfProducts: The number of products the customer holds with the bank.  
 - Balance: Account balance of the customer.  
 - HasCrCard: Whether the customer has a credit card.  
 - EstimatedSalary: The estimated salary of the customer.  
 - isActiveMember: Whether the customer is actively using the bank's services.  
 - Exited: Whether the customer has churned (Exited = 1) or stayed (Exited = 0).

**List of Analysis**

Certainly! Here's how the insights and analysis would look, following the same structure as the first example:

**1. Credit Score Distribution**

• **Objective**: Analyze the distribution of customer credit scores and how they relate to overall customer characteristics.  
• **Analysis**:

* Use histograms and box plots to visualize the distribution of credit scores.
* Apply the viridis colormap to make the plots visually engaging.
* Calculate the skewness and check for any significant outliers.  
  • **Insight**: The majority of credit scores are above 500, with a near-normal distribution around 650, suggesting a healthy financial profile for most customers. This analysis helps to identify potential financial risks or areas for improvement.

**2. Age Distribution and Outliers**

• **Objective**: Understand the age distribution and potential outliers that may impact customer behavior.  
• **Analysis**:

* Use histograms and box plots to visualize the age distribution of customers.
* Apply the viridis colormap for consistency.
* Highlight any outliers, especially customers above 60 years old.  
  • **Insight**: The age distribution is right-skewed, with a peak around age 40, suggesting that the majority of customers are in their middle adulthood. The box plot shows that most ages are clustered near the median, with few outliers above 60, which might be important for understanding customer retention and preferences.

**3. Balance Distribution**

• **Objective**: Investigate the distribution of customer balances and their significance in customer behavior.  
• **Analysis**:

* Use histograms and box plots to analyze the balance distribution.
* Focus on how many customers have a zero balance and how the non-zero balances are distributed.
* Apply the viridis colormap to the plots for better visualization.  
  • **Insight**: The distribution reveals that most customers have balances close to zero, with a few high-value accounts. This suggests that the majority of customers may have low or no active financial products, which could influence their engagement with additional services or products.

**4. Estimated Salary Distribution**

• **Objective**: Explore the estimated salary distribution to understand customer income and purchasing power.  
• **Analysis**:

* Visualize the salary distribution with histograms and box plots.
* Apply the viridis colormap to enhance visual appeal.
* Evaluate the spread and presence of outliers in salary data.  
  • **Insight**: The salary distribution is quite even, with a variety of income levels spread across the range. This suggests a diverse customer base with varying purchasing power. Understanding salary distribution can inform marketing strategies and potential product offerings tailored to different income segments.

**5. Exited (Churn) Distribution**

• **Objective**: Analyze the proportion of customers who exited versus those who stayed.  
• **Analysis**:

* Use a bar plot to visualize the distribution of customers who stayed (Exited = 0) and those who exited (Exited = 1).
* Calculate the percentage of customers in each group to understand the overall churn rate.  
  • **Insight**: 79.6% of customers stayed, while 20.4% exited. This indicates a significant class imbalance, which is important when building churn prediction models and applying retention strategies. The higher proportion of non-exited customers could help prioritize retention for those at risk of exiting.

**6. Geography Distribution**

• **Objective**: Explore the geographic distribution of customers and potential churn across different countries.  
• **Analysis**:

* Create a bar plot showing the percentage of customers from each country (France, Germany, Spain).
* Calculate the percentage of customers from each region and highlight any regional patterns.  
  • **Insight**: 50.1% of customers are from France, while the rest are divided between Germany (25.1%) and Spain (24.8%). This indicates that the majority of the customer base is from France, which may impact regional customer service and marketing strategies.

**7. Gender Distribution**

• **Objective**: Analyze the gender distribution of customers and its potential impact on churn.  
• **Analysis**:

* Use a bar plot to visualize the percentage distribution of male and female customers.
* Calculate the percentage of each gender in the dataset.  
  • **Insight**: 54.6% of customers are male, while 45.4% are female. This slight gender imbalance may not significantly affect churn behavior but could be considered when designing targeted marketing campaigns or retention strategies.

**8. Tenure Distribution**

• **Objective**: Investigate the customer tenure and its relationship with customer retention.  
• **Analysis**:

* Use a bar plot to show the tenure distribution.
* Focus on customers with tenure at the extremes (0 and 10 years) to see if there are any specific patterns.  
  • **Insight**: Tenure is fairly evenly distributed, with a few customers at the extremes. This shows that customers are spread across different lengths of time with the company, and customers with very short or long tenures are less frequent, possibly indicating churn risk at both ends.

**9. Number of Products**

• **Objective**: Analyze the distribution of the number of products owned by customers and its potential correlation with churn.  
• **Analysis**:

* Use a bar plot to display the distribution of customers with one, two, three, and four products.
* Focus on customers who have one or two products, as they represent the majority.  
  • **Insight**: 50.8% of customers have one product, and 45.9% have two products, with only a small proportion owning three or four products. This suggests that customers are typically sticking to one or two products, which could influence cross-selling or upselling strategies.

**10. Credit Card Ownership**

• **Objective**: Understand the relationship between credit card ownership and customer retention.  
• **Analysis**:

* Use a bar plot to visualize the percentage of customers who own a credit card.
* Calculate the percentage of customers with and without a credit card.  
  • **Insight**: 70.5% of customers own a credit card, which suggests that a significant portion of the customer base is engaged with financial products. This may also be an indicator of customer loyalty and purchasing power.

**11. Active Membership Status**

• **Objective**: Analyze the proportion of customers who are active members and their potential churn behavior.  
• **Analysis**:

* Use a bar plot to visualize the distribution of active vs. inactive members.
* Calculate the percentage of active members in the dataset.  
  • **Insight**: The dataset shows a fairly equal distribution between active and inactive members, with 51.5% of customers being active. Understanding the behavior of active vs. inactive members can inform targeted retention strategies, particularly for those who are inactive and may be at risk of churning.

**12. Correlation Plot**

* **Objective**: Explore relationships between different numerical features.
* **Analysis**: A heatmap with the 'viridis' color map is used to visualize the correlation between features. The values range from -1 to 1, indicating weak correlations (values close to 0). This suggests that the features are largely independent, with no significant linear relationships between them.
* **Insight**: Weak correlations imply that no feature stands out as strongly predictive of others, meaning that interactions between features might be minimal.

**13. Customer Churn by Geography (Stacked Bar Plot)**

* **Objective**: Compare customer churn across different countries.
* **Analysis**: A grouped bar chart is plotted to compare churn counts across countries (France, Germany, and Spain). The plot reveals the highest churn rate in Germany, followed by Spain and France. The stacked bars also show the relative churn percentages in each region.
* **Insight**: Germany exhibits the highest churn rate, indicating potential regional issues such as service quality or market competition, despite having a smaller customer base than France.

**14. Customer Churn by Gender (Stacked Bar Plot)**

* **Objective**: Analyze the impact of gender on customer churn.
* **Analysis**: A grouped bar chart displays churn counts for male and female customers. The exit rate for female customers is slightly higher than for males.
* **Insight**: Gender does appear to influence churn, with females showing a higher likelihood of exiting, suggesting a potential area for targeted retention efforts.

**15. Customer Churn by Credit Card Status (Stacked Bar Plot)**

* **Objective**: Investigate the effect of credit card ownership on customer churn.
* **Analysis**: A grouped bar chart compares churn rates for customers with and without credit cards. The exit rates for both groups are similar, with a minimal difference in churn percentages.
* **Insight**: Having a credit card doesn’t have a strong impact on customer churn behavior, suggesting that factors other than credit card ownership may drive exits.

**16. Customer Churn by Active Member Status (Stacked Bar Plot)**

* **Objective**: Assess the relationship between being an active member and churn.
* **Analysis**: The plot compares the churn rates between active and inactive members. The exit rate is slightly higher for inactive members.
* **Insight**: Inactive membership is associated with a higher likelihood of exiting, which could indicate that engagement plays a key role in retention.

**17. Credit Score Distribution by Churn Status (Box Plot)**

* **Objective**: Examine the relationship between credit score and customer churn.
* **Analysis**: A box plot displays the distribution of credit scores between churned and non-churned customers. The distribution doesn't show any clear trend linking credit score to churn, with some higher credit score customers still exiting.
* **Insight**: Credit score does not appear to be a strong predictor of churn, suggesting that other factors might be more influential.

**18. Age Distribution by Churn Status (Box Plot)**

* **Objective**: Investigate whether age affects customer churn.
* **Analysis**: The box plot shows that older customers (especially those in their 70s and above) tend to have a higher likelihood of exiting, whereas younger customers have a lower exit rate.
* **Insight**: Age could be a potential factor in churn, with older customers more likely to leave, indicating that retention strategies may need to focus on this demographic.

**19. Tenure Distribution by Churn Status (Box Plot)**

* **Objective**: Analyze the role of customer tenure in churn behavior.
* **Analysis**: A box plot comparing tenure with churn shows that tenure has minimal impact on whether a customer exits or not.
* **Insight**: Tenure doesn’t appear to be a strong indicator of churn in this dataset, suggesting that other factors like engagement or product usage might play a bigger role.

**20. Balance Distribution by Churn Status (Box Plot)**

* **Objective**: Determine if customer balance correlates with churn.
* **Analysis**: The plot shows that many customers have a balance of zero, regardless of whether they exited or not, with no significant difference in balance distributions for exited versus non-exited customers.
* **Insight**: Balance does not seem to be a critical factor in customer churn decisions, as a large proportion of customers maintain a zero balance.

**21. Number of Products Owned by Churn Status (Box Plot)**

* **Objective**: Explore the relationship between the number of products a customer owns and their likelihood to churn.
* **Analysis**: The box plot reveals that customers with more products, especially those with three or four, have a higher likelihood of exiting. This could suggest that customers with multiple products might be more likely to leave.
* **Insight**: Holding multiple products is associated with a higher risk of churn, potentially indicating that customers with more complex relationships with the service are more likely to exit.

**22. Estimated Salary by Churn Status (Box Plot)**

* **Objective**: Assess whether estimated salary influences customer churn.
* **Analysis**: The distribution of estimated salary is similar for both exited and non-exited customers, showing no significant differences between the two groups.
* **Insight**: Estimated salary does not appear to be a significant factor influencing customer churn in this dataset.

**23**. **Churn Prediction (Machine Learning)**

* **Objective**: Build predictive models to identify customers who are likely to exit.
* **Analysis**:
  + Use machine learning algorithms (e.g., **Logistic Regression**, **Random Forest**, **XGBoost**) to train a churn prediction model.
  + Evaluate the model's performance using metrics like **accuracy**, **precision**, **recall**, and **F1-score**.
  + **Insight**: This model can help predict which customers are likely to churn, allowing for targeted retention efforts.
  + Here's the analysis of your model evaluation results in list format:
  + **Best Performing Model:**
  + **Gradient Boosting Classifier (gbc)**
  + **Accuracy:** 81.49%
  + **AUC:** 0.7313
  + **Recall:** 36.89%
  + **Precision:** 57.19%
  + **F1-Score:** 44.71%
  + **Kappa:** 34.24%
  + **MCC:** 35.49%
  + **Time Taken:** 0.445 seconds
  + **Analysis:** This model provides the best balance between accuracy and other key metrics, making it the optimal choice.
  + **Other Strong Models:**
  + **Extra Trees Classifier (et)**
  + **Accuracy:** 81.43%
  + **AUC:** 0.7783
  + **Recall:** 30.08%
  + **Precision:** 58.67%
  + **F1-Score:** 39.69%
  + **Analysis:** Very close to Gradient Boosting but with slightly lower recall. The AUC is higher, indicating better classification performance.
  + **CatBoost Classifier (catboost)**
  + **Accuracy:** 81.30%
  + **AUC:** 0.7472
  + **Recall:** 38.36%
  + **Precision:** 56.12%
  + **F1-Score:** 45.40%
  + **Analysis:** Similar to Gradient Boosting in performance, but slightly slower (1.6880 seconds). Has high recall and F1-score.
  + **Random Forest Classifier (rf)**
  + **Accuracy:** 80.76%
  + **AUC:** 0.7519
  + **Recall:** 31.35%
  + **Precision:** 54.67%
  + **F1-Score:** 39.72%
  + **Analysis:** Performs similarly to the top models but with a lower accuracy and recall rate.
  + **Other Models:**
  + **Light Gradient Boosting Machine (lightgbm)**
  + **Accuracy:** 80.66%
  + **AUC:** 0.7022
  + **Recall:** 37.03%
  + **Precision:** 53.62%
  + **F1-Score:** 43.64%
  + **Analysis:** Slightly lower performance in comparison to the best models but still a solid choice with acceptable recall and F1-score.
  + **AdaBoost Classifier (ada)**
  + **Accuracy:** 80.20%
  + **AUC:** 0.7064
  + **Recall:** 35.91%
  + **Precision:** 52.13%
  + **F1-Score:** 42.43%
  + **Analysis:** Good performance but with slightly lower accuracy and precision.
  + **Extreme Gradient Boosting (xgboost)**
  + **Accuracy:** 80.07%
  + **AUC:** 0.7085
  + **Recall:** 36.96%
  + **Precision:** 51.51%
  + **F1-Score:** 42.95%
  + **Analysis:** Similar to AdaBoost, with slightly better recall and precision, but still not as strong as Gradient Boosting.
  + **Weak Performing Models:**
  + **Logistic Regression (lr)**
  + **Accuracy:** 79.63%
  + **AUC:** 0.5809
  + **Recall:** 0%
  + **Precision:** 0%
  + **F1-Score:** 0%
  + **Analysis:** Extremely poor performance with no useful predictive power for the dataset.
  + **Dummy Classifier (dummy)**
  + **Accuracy:** 79.63%
  + **AUC:** 0.5000
  + **Recall:** 0%
  + **Precision:** 0%
  + **F1-Score:** 0%
  + **Analysis:** Just predicts random values, yielding no useful information.
  + **Naive Bayes (nb)**
  + **Accuracy:** 78.59%
  + **AUC:** 0.7482
  + **Recall:** 9.54%
  + **Precision:** 39.78%
  + **F1-Score:** 15.30%
  + **Analysis:** Despite having a decent AUC, it underperforms in recall and F1-score.
  + **Ridge Classifier (ridge)**
  + **Accuracy:** 76.74%
  + **AUC:** 0.0000
  + **Recall:** 13.61%
  + **Precision:** 32.95%
  + **F1-Score:** 19.22%
  + **Analysis:** Very poor performance, particularly in AUC and recall.
  + **K Neighbors Classifier (knn)**
  + **Accuracy:** 76.10%
  + **AUC:** 0.5246
  + **Recall:** 9.12%
  + **Precision:** 25.66%
  + **F1-Score:** 13.43%
  + **Analysis:** Poor performance overall with weak recall and F1-scores.
  + **Linear Discriminant Analysis (lda)**
  + **Accuracy:** 75.66%
  + **AUC:** 0.6571
  + **Recall:** 19.50%
  + **Precision:** 33.19%
  + **F1-Score:** 24.50%
  + **Analysis:** Moderate performance but still falls short in comparison to top models.
  + **Decision Tree Classifier (dt)**
  + **Accuracy:** 74.84%
  + **AUC:** 0.6156
  + **Recall:** 39.14%
  + **Precision:** 38.35%
  + **F1-Score:** 38.65%
  + **Analysis:** Performs slightly worse than some other classifiers with moderate recall.
  + **SVM - Linear Kernel (svm)**
  + **Accuracy:** 70.51%
  + **AUC:** 0.0000
  + **Recall:** 17.76%
  + **Precision:** 4.53%
  + **F1-Score:** 7.15%
  + **Analysis:** Poor predictive power with low recall and precision.
  + **Quadratic Discriminant Analysis (qda)**
  + **Accuracy:** 65.86%
  + **AUC:** 0.4950
  + **Recall:** 24.19%
  + **Precision:** 19.12%
  + **F1-Score:** 15.57%
  + **Analysis:** Very poor performance with weak recall and precision.

Here is the link for the code that covered the list of analysis and modeling : <https://github.com/Rama-Marhlh/Bank-Churn-Prediction-and-Analysis-Project/blob/main/Bank%20Churn%20Prediction%20Data%20EDA%20and%20Modeling.ipynb>