🚀 Customer Churn Prediction & Retention Strategy 🚀

Project Goal: To predict which bank customers are likely to churn and recommend personalized strategies to retain them. 🏦

Key Achievements & Deliverables

- **Data Preparation:** Processed & engineered 10+ features from the Bank Customer Churn dataset (`churn.csv`). 📊
  - *Examples:* `BalanceSalaryRatio`, `TenureGroup`, `AgeGroup`.
- **Model Building & Optimization:** Trained multiple ML models (Logistic Regression, Random Forest, XGBoost).
  - **Optimized XGBoost** emerged as the best performer after hyperparameter tuning. 🧠
- **Churn Risk Identification:** Identified high-risk customers with their predicted churn probability. ✨
- **Personalized Retention Strategies:** Developed a rule-based system to suggest tailored actions for at-risk customers. 🤝

Optimized XGBoost Model Performance

Our best model shows strong capabilities in identifying churners:

- **Accuracy: 87.15%** (Overall correct predictions) ✅
- **Precision: 80.80%** (8 out of 10 predicted churners are *actually* churners – **highly efficient targeting!**) 🎯
- **Recall: 49.14%** (Identifies nearly half of all *actual* churners – **significant improvement, but room to grow!**) 🎣
- **ROC-AUC: 0.8649** (Excellent ability to distinguish between churners and non-churners) 📈

Top Churn Drivers (Feature Importance)

The model highlighted the most influential factors contributing to churn:

- **Age** 👴
- **Number of Products** 📦
- **Estimated Salary** 💵
- **Credit Score** 💯
- **Balance** 💰
- *(Engineered features like `BalanceSalaryRatio` also played a key role!)* 💡

Retention Strategy Approach

Strategies are generated based on customer profile and predicted churn probability:

- **High Balance Customers:** "Provide a premium customer service contact." 📞
- **Older Customers (> 50):** "Offer retirement-friendly account options." 👵
- **Inactive Members:** "Send reactivation offer with 5% cashback." 📧
- **Very High Churn Probability (> 0.7):** "Immediate personal call from retention team." 🚨

Future Vision & Next Steps

To make this project production-ready and continuously valuable:

- **Production Deployment:** Containerize (Docker) & deploy to cloud (AWS/Azure/GCP). ☁️
- **MLOps:** Implement automated logging, monitoring, model versioning, and retraining pipelines. 🛠️
- **Enhanced UX:** Develop a dashboard or integrate with CRM systems. 📊
- **A/B Testing:** Empirically test different retention strategies. ✅

📊 Customer Churn Project: Dataset Snapshot 📊

Dataset Overview: Bank Customer Churn Data

- **Size:** 10,000 customer records 👥
- **Features:** 14 distinct attributes per customer 📋
- **Target:** `Exited` (1 = Churned, 0 = Stayed) 💔➡️✅

Key Feature Categories & Insights

- **Demographics:**
  - `Age`: Sample range 39-43 (diverse in full data) 🎂
  - `Geography`: Primarily France & Spain (Germany also present) 🌍
  - `Gender`: Sample shows more Females 👩
- **Financial Indicators:**
  - `CreditScore`: Healthy range (582-850) 📈
  - `Balance`: Wide variation (€0 to €125K+) 💰
  - `EstimatedSalary`: Sample range €79K-€113K 💸
- **Behavioral Metrics:**
  - `Tenure`: Customer longevity (1-8 years) ⏳
  - `NumOfProducts`: 1-3 banking products held 🛍️
  - `IsActiveMember`: Indicates customer engagement (Binary) ✔️

Data Quality & Preprocessing Notes

- **Completeness:** No missing values found (100% complete) 🎉
- **Potential Issues Identified:**
  - `Balance` = €0.00 for some customers (requires verification) 🤔
  - `RowNumber`, `CustomerId`, `Surname` are likely redundant and can be dropped 🗑️

🌍 Churn by Country: Key Insights 🌍

Customer Distribution & Churn Patterns

- **France:** Dominates customer base (~60% of total) 🇫🇷, but has the **lowest churn rate** (a benchmark for retention! ✅)
- **Germany:** ~20% of customers 🇩🇪, but shows the **highest churn proportion (~30-40% of its customers)** 🚨
  - *Germany's churn rate is 2-3x higher than France's!* 📈
- **Spain:** ~20% of customers 🇪🇸, with a moderate churn rate (midway between France & Germany) ⚖️

Strategic Implications for Retention

- **Germany (High-Risk Market):**
  - **Urgent Need** for targeted interventions. 🎯
  - *Possible Causes:* Cultural preferences, local competition, service issues. 🤔
- **France (Retention Opportunity):**
  - Leverage low churn to **refine successful strategies** & replicate elsewhere. 🏆
- **Spain (Learning Opportunity):**
  - Benchmark against France's approaches to **improve retention**. 💡

💯 Credit Score vs. Churn: A Key Indicator 💯

Key Observations: Credit Score & Churn Behavior

- **Churned Customers (💔):**
  - Show a **wider spread** of scores, with more outliers. 📉
  - Have a **lower median score** (≈650) compared to retained customers.
  - More frequently have scores **below 600** (a high-risk zone). 🚨
- **Retained Customers (✅):**
  - Scores **cluster more tightly**, indicating consistent creditworthiness.
  - Exhibit a **higher median score** (≈700). ⬆️

Critical Thresholds & Strategic Implications

- **Scores < 600:** Show **2-3x higher churn likelihood**.
  - **Action:** Implement enhanced monitoring, consider credit counseling, or targeted support. 🛡️
- **Scores > 750:** Demonstrate **strong customer retention**.
  - **Action:** Reward loyalty, offer premium services, or relationship pricing. 🌟
- **Overall:** Credit score is a **strong predictor of churn**, enabling tiered, proactive retention strategies. 🎯

📉 Logistic Regression: Initial Model Performance 📉

Performance Breakdown

- **Accuracy: 80.9%** (Model correctly predicts 81% of cases overall) ✅
- **Precision: 59.7%** (When predicting churn, 60% are correct) 🎯
- **Recall: 18.9%** (Only detects **19% of actual churn cases** – a critical weakness!) 🚨
- **F1-Score: 28.7%** (Poor balance between precision and recall) ⚖️

Key Insights & Root Causes

- **Accuracy is Misleading:** High accuracy (81%) hides the fact that the model **misses 81% of actual churners** (Recall = 19%). This is due to **class imbalance** (more non-churners). 😬
- **Unacceptable Recall:** A recall of 0.19 is **unacceptable for business use**; it means we'd lose most at-risk customers. 💸
- **Model Limitations:** Logistic Regression struggles with complex feature interactions and non-linear relationships, contributing to its poor performance on churn prediction. 🚧

Conclusion:

This baseline model is **insufficient for effective churn retention**. Urgent improvements are needed, particularly in boosting recall, by addressing class imbalance and exploring more powerful models. 💪

🌳 Random Forest: Improved Model Performance 🌳

Significant Leap from Logistic Regression!

Our Random Forest model shows a strong step forward in predicting churn, offering a much better balance of performance.

Key Performance Metrics

- **Accuracy: 86.25%** ✅
  - Good overall correctness, but still mindful of class imbalance.
- **Precision: 77.5%** 🎯
  - **Strong!** Nearly 4 out of 5 predicted churners are *actually* churners. Efficient targeting!
- **Recall: 45.7%** 🎣
  - **Improved, but still the main challenge.** Over half of actual churners are *still being missed*.
- **F1-Score: 57.5%** ⚖️
  - Reflects better balance than before, but highlights the precision-recall gap.

In Essence:

The Random Forest model is **much better at precisely identifying churners**, but we're **still missing a substantial number** of at-risk customers. For effective churn prevention, **maximizing Recall remains paramount!** 💪

📊 Random Forest: What Drives Churn? (Feature Importance)

Top Influential Factors (Most Impactful on Churn Prediction)

- **Age:** By far the **most critical factor** – different age groups have distinct churn patterns. 👵➡️🧑
- **Number of Products (`NumOfProducts`):** Highly significant; fewer products mean less "stickiness." 📦
- **Estimated Salary:** Plays a substantial role in influencing financial behavior. 💵
- **Credit Score & Balance:** Both are highly important indicators of financial health & stability. 💯💰

Mid-Range & Less Influential Predictors

- **Mid-Range:** `BalanceSalaryRatio`, `Tenure`, `IsActiveMember` show good predictive power. 📈
- **Least Influential:** `Geography_Germany`, `Gender_Male`, `HasCrCard`, `Geography_Spain` have minimal direct impact on churn likelihood. 📉

Actionable Insights for Retention Strategy

- **Prioritize Age-Based Strategies:** Tailor offers & communication by age segment. 🎯
- **Deepen Product Relationships:** Focus on upsell/cross-sell for customers with fewer products. 🤝
- **Monitor Financial Health:** Proactively engage customers based on changes in `CreditScore`, `Balance`, `EstimatedSalary`. ⚠️
- **Re-engage Inactive Members:** Target campaigns to reactivate dormant accounts. 📧
- **Segment by Key Drivers:** Craft personalized actions based on Age, Products, and Financial Profile. 🧑‍💻

🎯 High-Risk Customers & Retention Strategies 🎯

Identifying At-Risk Customers

- Our model identifies customers with **high churn probabilities** (e.g., 0.73 to 0.94). 📈
- **Key Profile:** Often high-balance customers (€105K - €137K+), making their retention critical. 💰
- **Diverse Product Holdings:** High balance customers can still be at risk, even with varying numbers of products (1, 3, 4). 📦

Current Retention Strategy

- **Primary Strategy:** "Provide a premium customer service contact." 📞
- **Rationale:** Aligns with high-value customers, offering dedicated support to address potential dissatisfaction. 🤝

Enhancements & Considerations

- **Strategy Diversity:** Explore more varied strategies beyond just premium contact (e.g., upsell for low `NumOfProducts`). 💡
- **Understand Trigger Points:** Investigate *why* customers are at risk (inactivity, service changes) for more targeted interventions. 🤔
- **Leverage All Feature Importance:** Ensure `Age` and other key drivers fully influence personalized strategies. 📊

🚀 XGBoost: Optimized Model Performance 🚀

Hyperparameter Tuning & Best Parameters

- **Tuning Process:** 3-fold cross-validation with 5 candidates. ✅
- **Best Parameters Found:**
  - `subsample`: 0.9
  - `n_estimators`: 100
  - `max_depth`: 5
  - `learning_rate`: 0.1
  - *(Also optimized `scale_pos_weight` for class imbalance!)* ⚖️

Key Performance Metrics (Optimized)

- **Accuracy: 87.15%** (Overall correct predictions) 📈
  - *Slightly higher than Random Forest (0.8625).*
- **Precision: 80.80%** (Over 80% of predicted churners are *actual* churners – **highly efficient targeting!**) 🎯
- **Recall: 49.14%** (Identifies nearly half of *actual* churners – **crucial improvement!**) 🎣
  - *Significant jump from Random Forest (0.4570).*
- **F1-Score: 60.88%** (Better balance between precision & recall) ✨
  - *Improved from Random Forest (0.5749).*
- **ROC-AUC: 0.8649** (Excellent ability to distinguish churners from non-churners) 📊

Conclusion:

The Optimized XGBoost model delivers **superior performance** across all key metrics, making it a robust choice for identifying churn risk with high precision and significantly improved recall! 💪

🚀 Customer Churn Prediction & Retention: Project Summary

Our Mission: Predict customer churn & deliver personalized retention strategies. 🏦

Key Project Highlights

- **Data Mastery:** Processed 10,000 customer records, engineered powerful new features (e.g., `BalanceSalaryRatio`, `TenureGroup`). 📊
- **Model Excellence:** Optimized **XGBoost** as our champion model. 🧠
  - **Performance:**
    - **Accuracy: 87.15%** ✅
    - **Precision: 80.80%** (Highly efficient targeting! 🎯)
    - **Recall: 49.14%** (Identifies nearly half of actual churners! 🎣)
    - **ROC-AUC: 0.8649** 📈
- **Top Churn Drivers:** `Age`, `NumOfProducts`, `EstimatedSalary`, `CreditScore`, `Balance` are key. 👴📦💵💯💰
- **Actionable Strategies:** Personalized recommendations (e.g., "Premium contact" for high-value clients, "Reactivation offers" for inactive members). 🤝

Impact & Next Steps

- **Impact:** Proactively identify and engage high-risk customers, minimizing revenue loss. 🛡️
- **Future:** Focus on **Production Deployment** (Docker, Cloud), **MLOps** (Monitoring, Retraining), and **UI Integration** for continuous business value. ☁️🛠️