# Bank Customer Churn Prediction Final Report

## Executive Summary

This project addresses customer churn in the banking industry by developing a predictive model using the provided dataset. The model identifies key factors contributing to customer churn, providing actionable insights to reduce churn and improve customer retention.

We employed the DIDA framework (Data, Insights, Decisions, and Advantage) to structure our analysis and derive practical recommendations. Our final model offers both predictive power and interpretability, guiding the bank in strategic decision-making.

## DIDA Framework

### Data

We used daily customer data across multiple branches, including features such as:

- Credit score
- Balance
- Age
- Gender
- Geography
- Customer complaints

This data underwent preprocessing steps, including normalization, and one-hot encoding of categorical variables to prepare it for model training.

### Insights

Probability of customer churn based on relevant predictors such as credit score, bank balance, age, gender, etc.

- Logistic regression revealed that customer satisfaction scores and balances are the most critical predictors of churn.

- Classification trees highlighted interactions between age and geography, providing additional actionable insights.

- Customers who had previously complained were at significantly higher risk of churn, which initially dominated the predictions.

Key qualitative insight: Addressing low satisfaction scores and handling complaints proactively can significantly reduce churn risk.

### Decisions

Which customers to target with retention strategies, to reduce the chance of churn

The models support better decision-making compared to intuition alone by identifying:

- High-risk customers (low satisfaction and high balances) who need immediate intervention.

- High-churn geographies requiring localized retention strategies.

- Complaint management opportunities to mitigate churn early.

### Advantage

• Improved customer retention strategies through focused interventions.

• Enhanced targeted marketing efforts, reducing churn rates and increasing customer lifetime value.

• Resource optimization by prioritizing actionable segments.

## Managerial Challenges and Insights

### Challenges

1. The initial model’s overreliance on the “Complain” feature led to skewed predictions.

2. Balancing predictive power and interpretability in the final models.

3. Addressing class imbalance in the dataset as churn cases were underrepresented.

### Logistic Regression V1

In our original Logistic Regression model, we found that the ‘Complain’ predictor was too dominant in predicting customer churn; almost a 1:1 relation. Therefore, if a customer had complained in the past, they were extremely likely to churn/exit as well. This was giving us an accuracy and AUC of close to 99% (refer to MLGroupProject_Logistic_v1.ipynb). While this performance might appear impressive at first glance, it highlighted a significant issue: the model’s reliance on a single feature limited its ability to generalize and diminished the interpretability of other variables. This iteration underscored the need for a more balanced approach to understanding the predictors of customer churn.

### Logistic Regression V2

To address the limitation of the first model (Logistic Regression V1), we removed the ‘Complain’ predictor so that our model does not consider the highly predictive ‘Complain’ feature and provide a clearer view of the relative importance of other features. This gave us an AUC of around 0.85. This version revealed the nuanced influence of variables such as customer satisfaction, geography, and account balance, which were previously overshadowed. The revised model also provided actionable insights for designing targeted retention strategies (refer to MLGroupProject_Logistic_v2.ipynb)

## Python Implementation

### Logistic Regression

- **Rationale:** Chosen for its simplicity, interpretability, and effectiveness for binary classification.
    - **Procedures:**

1. Normalized numerical features: numerical features were normalized to ensure equal weight in the model, avoiding bias from variables with larger scales (e.g., balance vs. tenure).

2. One-hot encoded categorical variables (e.g., geography, gender): Categorical variables like geography were one-hot encoded to allow the model to interpret them without assuming ordinal relationships.

3. Applied **cross-validation** to improve model robustness: Logistic Regression models were evaluated using cross-validation to ensure robustness and avoid overfitting, especially given the class imbalance in the dataset.

- **Performance Metrics:** Accuracy and AUC guided model selection.

### Classification Tree

- **Rationale:** Useful for capturing non-linear relationships and providing interpretable decision rules.

- **Procedures:**

1. Split the dataset into training and testing subsets.
2. Limited tree depth to prevent overfitting.
3. Evaluated using AUC and accuracy, comparing with logistic regression.

# Results

**Logistic Regression**

- **Accuracy:** 84%
- **AUC:** 0.85

**Classification Tree**

- **Depth of the best pruned tree:** 6
- **AUC:** 0.85

## Managerial Recommendations & Key Insights

**Logistic Regression**

1. **Inactive Customers (1.112184):** High churn likelihood. Re-engage with loyalty programs.

2. **German Customers (0.920304):** Region-specific churn. Target with localized strategies.

3. **Female Customers (0.514588):** Higher churn risk. Develop tailored retention offers.

4. **Customers with 3+ Products (2.455100, 6.388572):** Overextension risk. Simplify product offerings.

5. **Low Satisfaction (0.203601):** Address complaints and improve service quality.

**Classification Tree**

1. **Older Inactive Customers in Germany (Leaf 31):** High churn risk. Re-engage older customers regionally.

2. **Customers with More than 2 Products (Leaf 5):** High churn. Bundle products to enhance value.

3. **Spanish Customers with Low Credit Scores (Leaf 53):** Address credit challenges with targeted programs.

4. **Inactive Members with High Balances (Leaf 9):** Retain with incentives and rewards.

5. **High-Earning Young Customers with Low Balances (Leaf 8):** Offer premium services to improve retention.

**Rules for Customer Churn Prediction**

1. **Rule 1: High churn probability**

If **Age ≤ 41.5**, AND **Number of Products > 2**, AND **Balance > $23,913.37**, AND **Is Not an Active Member**, AND **Geography = Germany**, Then the customer is **likely to churn**.

2. **Rule 2: Low churn probability**

If **Age ≤ 41.5**, AND **Number of Products ≤ 2**, AND **Balance ≤ $72,715.53**, AND **Geography ≠ Germany**, Then the customer is **unlikely to churn**.

3. **Rule 3: High churn probability**

If **Age > 41.5**, AND **Number of Products = 3 or 4**, AND **Is Not an Active Member**, AND **Geography = Germany**, AND **Balance > $87,266.75**, Then the customer is **very likely to churn**.

4. **Rule 4: Low churn probability**

If **Age > 41.5**, AND **Number of Products = 1 or 2**, AND **Is an Active Member**, AND **Estimated Salary ≤ $86,457.33**, AND **Balance > $42,398.04**, Then the customer is **unlikely to churn**.

5. **Rule 5: High churn probability**

If **Age ≤ 41.5**, AND **Number of Products > 2**, AND **Balance ≤ $23,913.37**, AND **Credit Score ≤ 460**, Then the customer is **likely to churn**.

6. **Rule 6: Very high churn probability**

If **Age > 41.5**, AND **Number of Products = 3**, AND **Is Not an Active Member**, AND **Geography = Spain**, AND **Point Earned ≤ 244**, Then the customer is **very likely to churn**.

7. **Rule 7: Low churn probability**

If **Age ≤ 41.5**, AND **Number of Products = 1**, AND **Geography ≠ Germany**, AND **Balance ≤ $23,913.37**, Then the customer is **unlikely to churn**.

## Summary and Final Recommendations

1. **Inactivity and Engagement:** Inactive customers are the most at-risk segment. Retention strategies like loyalty programs and personalized incentives are essential to re-engage them.

2. **Geographic and Demographic Factors:** German customers and older individuals show higher churn likelihood, requiring region-specific and age-tailored retention efforts. Female customers also exhibit higher churn risk, suggesting the need for personalized offerings.

3. **Product Overextension:** Customers with three or more products are highly likely to churn, indicating dissatisfaction or complexity. Simplifying offerings or bundling products can mitigate this risk.

4. **Satisfaction and Credit Impact:** Low satisfaction scores and poor credit ratings are strong churn indicators. Proactive service improvements and financial support programs are critical.

5. **High-Value Customers:** Younger, high-earning customers with low balances and inactive members with high balances need targeted outreach and premium services to retain their loyalty.

By addressing these insights, the bank can focus on reducing churn while enhancing customer satisfaction and lifetime value.