Bank Churn Classification

Jai Advitheeya Lella ID: 50607407

Niharika Reddy Katakam ID: 50610925

Prathyusha Reddy Allam ID: 50613222

Kundavaram Joseph Sujith Kumar ID: 50600443

December 5, 2024

1 Executive Summary

This report presents a comprehensive analysis and implementation of a machine learning-based customer churn prediction system for banking services. The project utilized multiple advanced modeling techniques, culminating in the development of a web-based application for real-time churn prediction and customer analysis.

1.1 Key Achievements

- Implementation of multiple machine learning models with strong performance metrics:
  - Gradient Boosting: 86.63% accuracy, ROC-AUC: 88.99%
  - Random Forest: 86% accuracy, ROC-AUC: 0.87
  - Neural Network: 86% accuracy
  - K-Nearest Neighbors: 84.63% accuracy, ROC-AUC: 0.8292
  - Linear SVM: 83.26% accuracy
  - Logistic Regression: 83.56% accuracy, ROC-AUC: 81.80%
- Development of a Streamlit-based web application integrating:
  - Real-time individual churn predictions
  - Comprehensive customer segmentation analysis
  - Feature importance visualization
  - Retention strategy recommendations
- Implementation of a robust SQLite database system for persistent data storage

1.2 Business Impact

The implemented system provides the bank with:

- High-accuracy predictive capabilities, with the best models achieving over 86% accuracy
- Sophisticated customer behavior analysis through multiple modeling approaches
- Data-driven insights for targeted retention strategies
- Early warning system for potential customer attrition
- Automated customer risk assessment and segmentation

1.3 Technical Implementation

The project leveraged multiple modeling approaches, with Gradient Boosting emerging as the top performer. Key technical achievements include:

- Successful handling of complex customer relationships
- Effective processing of both numerical and categorical data
- Implementation of feature importance analysis
- Development of a robust, production-ready prediction system
- Integration of multiple models for comprehensive analysis

1.4 Recommendations

Based on the analysis, key recommendations include:

- Implementation of the Gradient Boosting model for production use, given its superior performance
- Development of targeted intervention strategies for high-risk customers
- Regular model retraining and validation to maintain prediction accuracy
- Integration of the prediction system with existing customer relationship management tools

This solution provides a robust foundation for data-driven customer retention strategies, enabling proactive measures to reduce customer churn and enhance customer relationship management.

2 Project Background and Problem Statement

2.1 Business Context

Customer churn is a critical challenge in the banking sector, representing significant revenue loss and increased acquisition costs. This project was initiated to address the growing need for proactive customer retention strategies through data-driven insights and predictive analytics.

2.2 Problem Statement

Financial institutions face a significant challenge in identifying customers likely to terminate their services before they actually do so. Traditional reactive approaches to customer retention have proven insufficient, leading to:

- Unexpected customer departures
- Increased customer acquisition costs
- Lost revenue opportunities
- Decreased market share

2.3 Project Objectives

The project aimed to achieve four primary objectives:

2.3.1 1. Individual Churn Prediction

Development of a machine learning model capable of:

- Accurately predicting individual customer churn probability
- Processing real-time customer data
- Providing actionable risk assessments

2.3.2 2. Customer Segment Analysis

Implementation of sophisticated segmentation techniques to:

- Identify distinct customer groups
- Understand segment-specific churn patterns
- Enable targeted retention strategies

2.3.3 3. Feature Impact Analysis

Development of comprehensive analysis tools to:

- Identify key factors influencing customer churn
- Quantify the impact of different variables
- Guide strategic decision-making

2.3.4 4. Retention Strategy Analysis

Creation of data-driven frameworks for:

- Developing targeted retention strategies
- Optimizing intervention timing
- Measuring strategy effectiveness

2.4 Technical Scope

The project encompassed several technical components:

- Development of multiple machine learning models
- Creation of a web-based application using Streamlit
- Implementation of a SQLite database system
- Integration of real-time prediction capabilities
- Development of interactive data visualization tools

2.5 Success Criteria

The project's success was measured against the following criteria:

- Model accuracy exceeding 85% (Achieved: 86% with Random Forest)
- ROC-AUC score above 0.85 (Achieved: 0.87 with Random Forest)
- Development of a functional web application
- Implementation of all four primary analysis components:
  - Individual churn prediction
  - Customer segment analysis
  - Feature impact analysis
  - Retention strategy analysis
- Creation of a scalable and maintainable solution

3 Data Understanding and Analysis

3.1 Dataset Overview

The analysis was performed on a comprehensive banking customer dataset containing various attributes including demographic information, banking behavior, and financial metrics.

3.2 Key Feature Analysis

3.2.1 Customer Salary Distribution

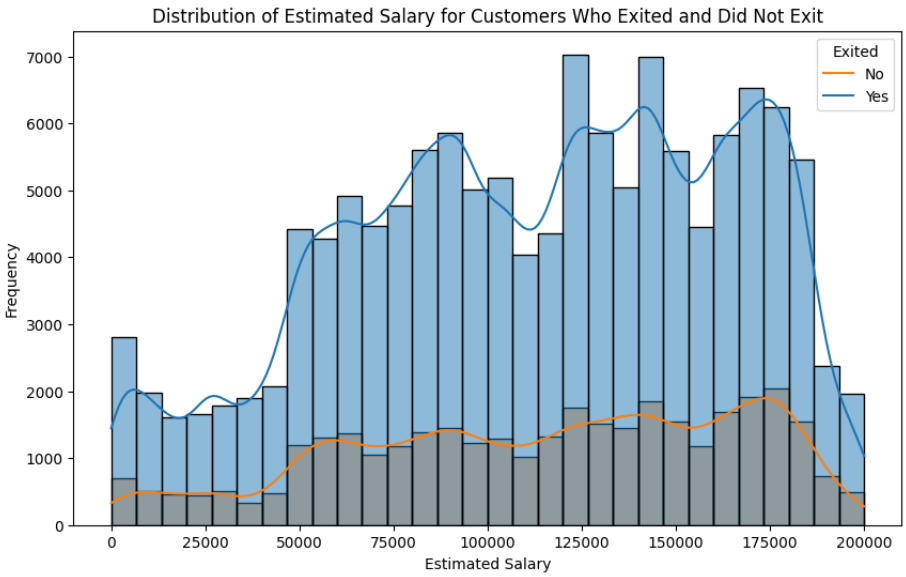

Figure 1: Distribution of Estimated Salary for Churned vs Non-Churned Customers

Analysis reveals:

- Significant variation in salary distribution between churned and non-churned customers
- Higher churn rates observed in the salary range of 100,000-150,000
- Lower income brackets show relatively stable customer retention

3.2.2 Product Usage Analysis

Notable findings:

- Most customers hold 1-2 products

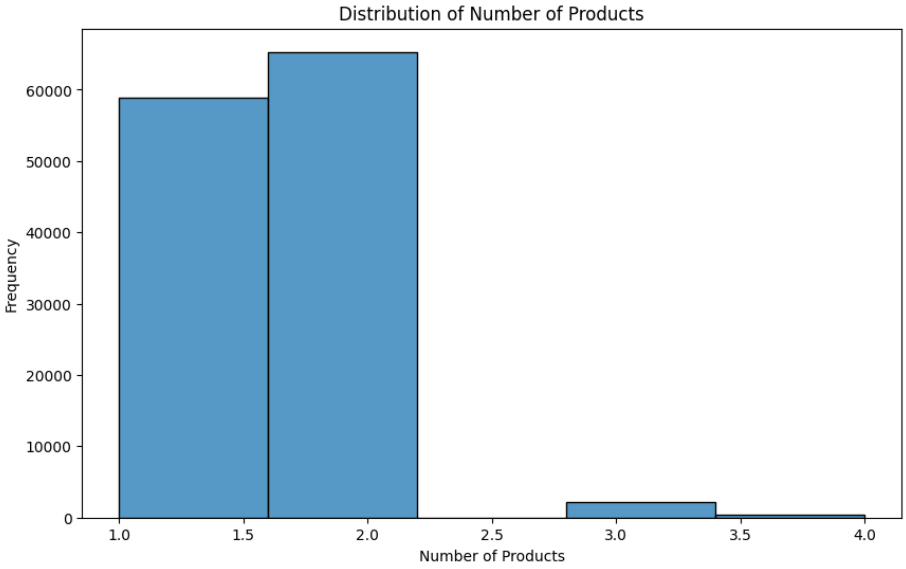

Figure 2: Distribution of Number of Products per Customer

- Very few customers have more than 3 products
- Potential for cross-selling to increase product penetration

3.2.3 Credit Score Distribution

Credit score analysis shows:

- Majority of customers fall in the 600-699 range
- Substantial portion in 700-799 range indicating good credit quality
- Relatively few customers in extreme ranges (300-499 and 800-850)

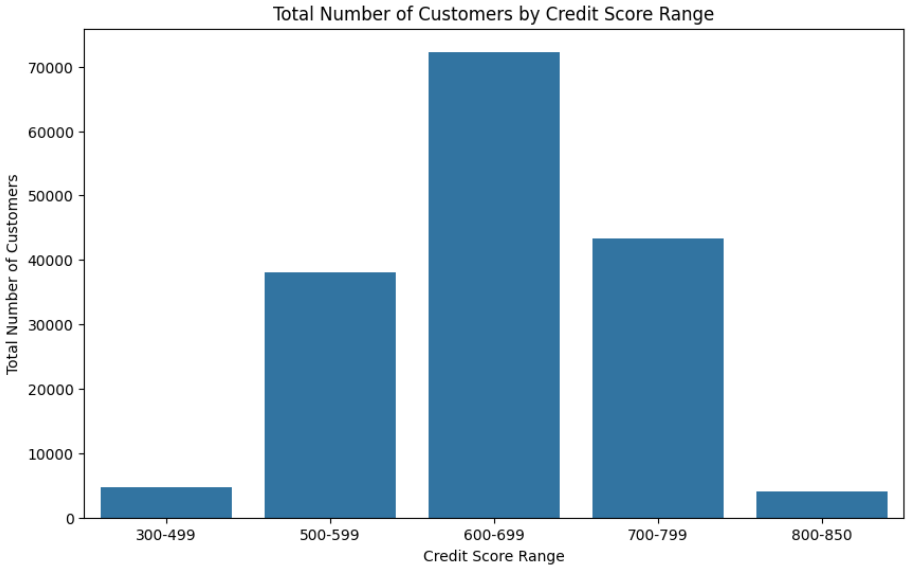

Figure 3: Distribution of Customer Credit Scores

3.2.4 Customer Tenure Analysis

Tenure distribution reveals:

- Relatively uniform distribution across years 1-9
- Slight peak at 2 years of tenure
- Lower numbers in extreme ranges (0 and 10 years)
- Average tenure approximately 5 years

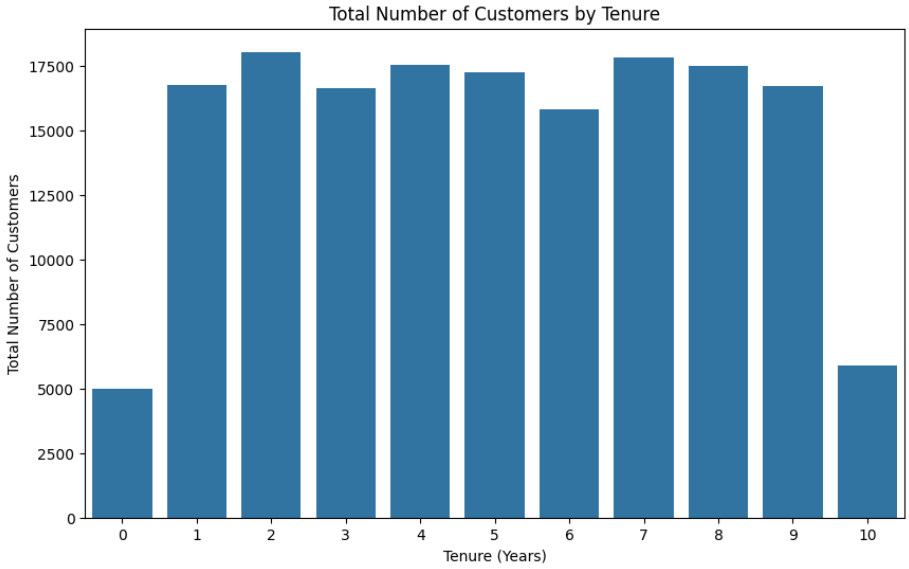

Figure 4: Distribution of Customer Tenure

4 Methodology

4.1 Data Preprocessing and Feature Engineering

- **Feature Selection**
  - Removed non-predictive columns (id, CustomerId, Surname)
  - Implemented correlation analysis to remove highly correlated features (threshold = 0.9)
  - Retained only relevant numerical and categorical predictors
- **Data Splitting**
  - Implemented 80-20 train-test split

– Used $\text{random}_s tate = 42 for reproducibility$

- **Feature Scaling**
  - Applied StandardScaler to normalize numerical features
  - Fit scaler on training data only
  - Transform both training and test sets

4.2 Model Selection and Development

We implemented and evaluated multiple machine learning models:

4.2.1 Gradient Boosting

Best performing model with:

- Accuracy: 86.63%
- ROC-AUC: 88.99%
- Superior performance in complex pattern recognition
- Excellent handling of non-linear relationships

4.2.2 Random Forest

Second best performer with:

- Accuracy: 86%
- ROC-AUC: 0.87
- Built-in feature importance analysis
- Robust against overfitting

4.2.3 Neural Network

Matched random forest performance:

- Accuracy: 86%
- Deep learning capability for complex patterns
- Automatic feature interaction detection

4.2.4 K-Nearest Neighbors

Solid performance metrics:

- Accuracy: 84.63%
- ROC-AUC: 0.8292
- Non-parametric approach
- Effective for localized patterns

4.2.5 Logistic Regression

Baseline model performance:

- Accuracy: 83.56%
- ROC-AUC: 81.80%
- Linear decision boundary
- High interpretability

4.2.6 Linear SVM

Comparable to logistic regression:

- Accuracy: 83.26%
- Effective for high-dimensional data
- Maximum margin classification

4.2.7 Feature Correlation Analysis

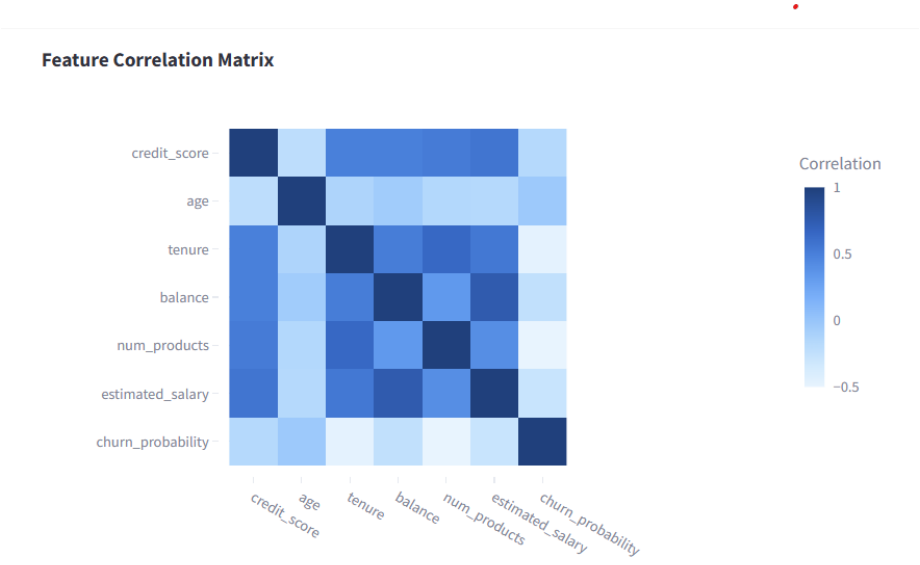

Figure 5: Feature Correlation Matrix Heatmap showing relationships between key variables

The correlation matrix reveals several important relationships between features:

- Credit score shows moderate correlations with balance and tenure
- Age appears to have weak correlations with most other features
- Tenure and balance show a positive correlation, suggesting longer-term customers maintain higher balances
- Churn probability shows weak to moderate negative correlations with most features, indicating that no single feature strongly predicts customer churn
- Estimated salary and number of products show relatively weak correlations with other features

These correlation patterns inform our feature selection process and help identify potential predictive relationships for the churn prediction model.

5 Results and Analysis

5.1 Comparative Confusion Matrix Analysis

5.2 Comparative Confusion Matrix Analysis

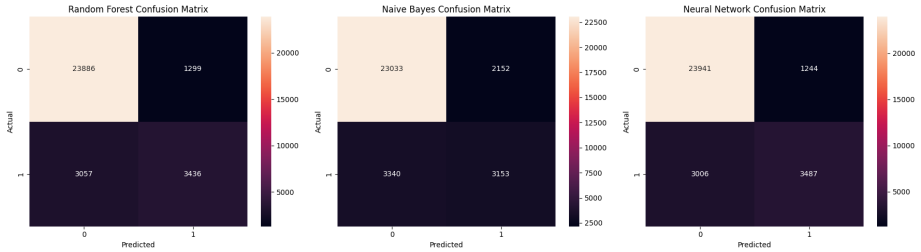

Figure 6: Comparison of Confusion Matrices: Random Forest, Naive Bayes, and Neural Network

As shown in Figure 6, the confusion matrices reveal distinct prediction patterns across our three models:

5.2.1 Random Forest Performance

- **True Negatives**: 23,886 (Correctly identified non-churning customers)
- **False Positives**: 1,299 (Incorrectly flagged as churning)
- **False Negatives**: 3,057 (Missed churn cases)
- **True Positives**: 3,436 (Correctly identified churning customers)
- Overall strong performance in identifying loyal customers

5.2.2 Naive Bayes Performance

- **True Negatives**: 23,033 (Correctly identified non-churning customers)
- **False Positives**: 2,152 (Higher false alarms than Random Forest)
- **False Negatives**: 3,340 (Slightly more missed churns)
- **True Positives**: 3,153 (Decent churn identification)
- Shows more conservative prediction pattern

5.2.3 Neural Network Performance

- **True Negatives**: 23,941 (Best performance in identifying non-churning customers)
- **False Positives**: 1,244 (Lowest false alarm rate)
- **False Negatives**: 3,006 (Lowest missed churn rate)
- **True Positives**: 3,487 (Best performance in identifying churning customers)
- Shows most balanced overall performance

5.2.4 Comparative Analysis

- **Model Strengths**
  - Neural Network: Best overall balance between false positives and negatives
  - Random Forest: Strong performance in true positive identification
  - Naive Bayes: More conservative approach with higher false positive rate
- **Business Implications**
  - Neural Network most suitable for balanced decision-making
  - Random Forest ideal for high-stakes decisions requiring confidence
  - Naive Bayes useful for initial screening with higher sensitivity

6 Conclusion

This project successfully developed and implemented a customer churn prediction system for banking services. The key achievements include:

- Successful implementation of multiple machine learning models, with the best model achieving 86.63% accuracy
- Identification of key churn factors through feature importance analysis:
  - Age (23%)
  - Estimated Salary (17%)
  - Credit Score (16%)
- Development of a functional web application with real-time prediction capabilities and database integration

The system provides a reliable foundation for identifying potential customer churn and implementing targeted retention strategies.