Data Analysis and Feature Engineering Report

Insights for Customer Churn Prediction

Prepared by: RetentionX

May 2025

Abstract

This report presents a comprehensive analysis of customer churn data, including statistical insights and feature engineering efforts. Statistical tests and visualizations uncover key predictors of churn, while new features and transformations enhance predictive modeling. The findings provide actionable insights for improving customer retention strategies.

Contents

1 Statistical Analysis and Insights

This section summarizes the statistical tests and insights derived from advanced data analysis:

- **T-tests**: Conducted on numerical features to compare means between churned and non-churned customers. Features like `CreditScore` and `Balance` showed statistically significant differences, indicating their potential predictive power.
- **Chi-squared Tests**: Performed on categorical features such as `Geography` and `Gender` to evaluate their relationship with churn. Results highlighted that `Geography` had a significant association with churn rates.
- **Correlation Matrix**: Visualized relationships between numerical features to identify multicollinearity and feature importance. High correlations were observed between `Balance` and `EstimatedSalary`, suggesting potential redundancy in predictive modeling.
- **Recursive Feature Elimination (RFE)**: Identified the top 5 most relevant features for churn prediction, including `Tenure`, `BalanceToSalary`, and `ProductsPerTenure`. These features were selected based on their contribution to model performance.

1.1 Key Insights

- `CreditScore`, `Balance`, and `EstimatedSalary` showed significant differences between churned and non-churned groups, with churned customers generally having lower `CreditScore` and higher `Balance`.
- Strong correlations between `Balance` and `EstimatedSalary` suggest potential redundancy, necessitating dimensionality reduction techniques.
- RFE highlighted `Tenure`, `BalanceToSalary`, and `ProductsPerTenure` as critical predictors, emphasizing the importance of customer lifecycle and financial behavior in churn prediction.
- Customers with higher `ProductsPerTenure` ratios were less likely to churn, indicating that engagement over time plays a key role in retention.

2 Feature Engineering Summary

This section outlines the new features, transformations, and their expected impact on model performance.

2.1 New Features

- `Tenure` (in months): Derived from `last_active_date` and `signup_date`. Captures the duration of a customer's relationship with the company, expected to improve the model's ability to capture lifecycle patterns.
- `BalanceToSalary` Ratio: Measures financial stability by dividing `Balance` by `EstimatedSalary`. Anticipated to enhance predictive power for financial behavior.
- `ProductsPerTenure`: Calculated as the ratio of products owned to tenure. Designed to capture customer engagement over time and its impact on churn likelihood.
- `ChurnProbabilityScore`: A derived feature based on historical churn patterns, providing a probabilistic estimate of churn likelihood for each customer.

2.2 Transformations

- **Scaling Numerical Features**: Applied `StandardScaler` to normalize features like `CreditScore`, `Age`, and `Balance`, ensuring comparable scales and improving model convergence.
- **Encoding Categorical Features**: Used one-hot encoding for variables like `CardType` and `Geography` to convert them into numerical format, ensuring appropriate representation in the model.
- **Log Transformation**: Applied to skewed features such as `Balance` and `EstimatedSalary` to reduce skewness and improve normality.
- **Interaction Features**: Created by combining existing features, such as `Age * Tenure` and `BalanceToSalary * ProductsPerTenure`, to capture complex relationships between variables.

2.3 Expected Impact

- Improved model performance by reducing feature skewness and enhancing interpretability.
- Better handling of categorical variables and numerical feature scaling for dataset consistency.
- Enhanced predictive power through interaction features and derived metrics like `ChurnProbabilityScore`.

3 Visualization Summary

- **Churn Rate by Tenure**: Boxplots revealed that customers with shorter tenures are more likely to churn, underscoring the importance of early engagement strategies.
- **Feature Importance**: Bar charts of logistic regression coefficients highlighted the top 10 features influencing churn, with `Tenure` and `BalanceToSalary` being the most significant.
- **Customer Segmentation**: Cluster plots based on `Balance` and `EstimatedSalary` identified distinct customer groups, aiding in targeted marketing strategies.

4 Conclusions

This report provides a robust foundation for predictive modeling through comprehensive statistical analysis and feature engineering. Key predictors of churn, such as `Tenure`, `BalanceToSalary`, and `ProductsPerTenure`, were identified, and data transformations enhanced model readiness. The insights and engineered features offer actionable strategies for improving customer retention.