KENT STATE UNIVERSITY®

Customer Churn in Bank – Classification By Use Of Machine Learning

Presented by Vaishnavi Donde

Course - Capstone Project in Business Analytics

Group - 11

KSU ID – 811300129

Instructor - Mostafa Arkdani

Table of Contents

Project Overview

- Research Objective – To develop robust models for customer churn classification using machine-learning
- Methodology – To compare existing classification models, incorporate real-time data, and consider domain-specific features.
- Goals - As we know, it is much more expensive to sign in a new client than keeping an existing one. It is advantageous for banks to know what leads a client towards the decision to leave the bank. Churn prevention allows banks to develop loyalty programs and retention campaigns to keep as many customers as possible.
- Purpose - Contribute insights and methodologies to revolutionize customer churn in the banking sector.
- Expected benefits - Improved efficiency, better decision-making, and overall benefits for the bank.
- Target Audience - Both academic community and industry practitioners.

Data Description

- Number of Observations - 10000
- Number of variables: 14
- Features-

- Numerical – RowNumber, CustomerId, CreditScore, Age, Tenure, Balance, NumOfProducts, EstimatedSalary
- Categorical - Surname, Geography, Gender, HasCrCard, IsActiveMember,
- Target variable - Exited

Data Preparation and Exploration

- The dataset was sourced from Kaggle, a website that provides easy access to freely downloadable open datasets. The dataset encompasses several columns, each contributing differently to the prediction of customer churn.
- Data Cleaning: Removing or fixing errors and inconsistencies in data to ensure it's accurate and useful.
- Type Conversion: Changing data from one format to another, like turning text into numbers.
- Handling Missing Values: Finding and dealing with empty spots in our data, either by filling them in or removing them.
- Feature Engineering: Creating new useful features or variables from existing data to help improve analysis or models.
- Standardization: Making sure data follows a consistent format or scale.
- Exploratory Data Analysis (EDA): Using graphs and statistics to understand the main characteristics of our data and spot any patterns or anomalies.

Methodology

Here is the step-by-step methods we followed to get to the conclusion

1. Selected 7 essential features out of the original 14 attributes by employing techniques like MRMR, Chi2, ANOVA, and the Kruskal-Wallis test.
2. Created 24 unique machine-learning classification models.
3. Performed 3 runs using a set of 7 features along with three different holdout and test splits (15%|10%, 15%|15%, 15%|20%).

Variable	MRMR	Chi2	ANOVA	Kruskal	Median
CreditScore	8	7	7	7	**7**
Geography	3	3	6	6	**5**
Gender	4	6	4	5	**5**
Age	5	2	1	1	**2**
Tenure	9	8	8	8	**8**
Balance	6	5	3	4	**5**
NumOfProducts	1	1	5	3	**2**
HasCrCard	7	9	10	10	**10**
IsActiveMember	2	4	2	2	**2**
EstimatedSalary	10	10	9	9	**10**

Methodology

4. Assessed model performance using confusion matrices to measure true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
5. Tracked CPU time for each model to evaluate computational efficiency along with predictive accuracy.
6. Utilized the Ohio Supercomputer Center (OSC) for high-performance computing to manage our large dataset.
7. Evaluated models based on a combination of predictive accuracy and computational efficiency to determine the best model for practical use.

Set 1 – 7 Features			
Runs	Run 1	Run 2	Run 3
Holdout	15%	15%	15%
Test	10%	15%	20%

Performance Assessment

- The comprehensive evaluation of various machine learning models, based on CPU time, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), offered detailed insights into their efficiency and accuracy for customer churn classification in a bank.
- The Gaussian Naïve Bayes model was notable for its excellent computational efficiency, with a CPU time of just 0.05718 seconds, while RUSBoosted Trees provided an ideal balance with CPU time of 0.23678 seconds
- The RUSBoosted Trees model provided the best balance of accuracy and efficiency according to the weighted metrics. It achieved a TP rate of 65.8%, TN rate of 13.5%, FP rate of 6.9%, and FN rate of 13.85%, resulting in a total weight of 0.44166.
- Utilized OSC's high-performance computing resources to manage the intensive computations needed for model training and evaluation, ensuring timely and efficient processing.

Conclusion

- The comprehensive evaluation of various machine learning models, based on CPU time, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), offered detailed insights into their efficiency and accuracy for customer churn classification in a bank.
- Key factors were identified as significant predictors, offering valuable insights into the drivers of churn.
- The Gaussian Naïve Bayes model stood out for its excellent computational efficiency, with a CPU time of just 0.05718 seconds, while the RUSBoosted Trees model provided an optimal balance with a CPU time of 0.23678 seconds.
- The RUSBoosted Trees model was recognized as the best in terms of accuracy and efficiency based on weighted metrics.
- Leveraged the Ohio Supercomputer Center (OSC) for high-performance computing, ensuring efficient handling of extensive data and complex algorithms needed for model training and evaluation.
- The model and insights from the analysis can guide data-driven decisions in customer churn classification, enabling banks to proactively address issues and enhance customer satisfaction.

KENT STATE UNIVERSITY

Thank You.