Final Churn Prediction Report

Andres Perez

Executive Summary

This report documents a comprehensive data science workflow for predicting customer churn. The project covers data exploration, feature engineering, baseline and advanced modeling, and business-focused evaluation. The main goal was to maximize the identification of customers at risk of churning, providing actionable insights for retention strategies.

1. Business Context and Objective

Customer churn is a critical business problem for subscription-based companies, as retaining existing customers is often more cost-effective than acquiring new ones. The objective of this project was to build a predictive model to identify customers likely to churn, with a particular focus on maximizing sensitivity (recall) to ensure as many churners as possible are detected for targeted retention efforts.

2. Data Overview

- **Data sources:**
  - Raw churn data (customer activity, support, and engagement metrics)
  - Engineered features (created in feature_engineering.Rmd)
  - Evaluation set (for generating predictions on unseen data)
- **Data quality:**
  - Addressed missing values, duplicates, and outliers during data exploration
  - Created new features to capture customer behavior and engagement
  - Managed class imbalance using resampling techniques (ROSE)

3. Feature Engineering

- **Top 10 engineered features used in modeling:**
    1. Customer.Months
    2. Days.Since.Last.Login
    3. CHI.Score.Mon0
    4. Activity_Score
    5. CHI.Score
    6. Logins
    7. Views_log
    8. Logins_log
    9. Views
    10. Login_View_Interaction
- **Process:**
    - Applied log transformations, ratios, and interaction terms to capture non-linear relationships
    - Selected features based on importance from a random forest model
    - Focused on features with the most predictive power for churn

4. Modeling Workflow

- **Baseline models:** Logistic Regression, Random Forest, XGBoost
- **Class balancing:** Used ROSE to address class imbalance in the training set
- **Hyperparameter tuning:** Performed grid search and cross-validation (especially for XGBoost)
- **Threshold optimization:** Selected probability threshold to maximize F1 score
- **Advanced experiments:** Cost-sensitive XGBoost (class weights), ensembling with Random Forest
- **Modular workflow:** Each stage documented in a separate Rmd for clarity and reproducibility

5. Model Comparison

Comparison of Key Models on Test Set

Model	Sensitivity	Specificity	F1	AUC
Baseline XGBoost	0.50	0.89	0.25	0.76
Cost-Sensitive XGBoost	0.42	0.90	0.23	0.72
Random Forest	0.27	0.95	0.23	0.78
Ensemble	0.27	0.93	0.19	0.75

6. Best Model Selection

The best model for this project was the **Baseline XGBoost** model, which achieved the highest sensitivity (0.50) and a strong AUC (0.76) on the test set. This model was selected because the business priority is to maximize the identification of churners, even at the expense of some precision and specificity. The final model used the top 10 engineered features, class balancing with ROSE, and threshold optimization for F1 score.

7. Business Interpretation and Recommendations

- The selected model can identify 50% of churners, allowing the business to proactively target at-risk customers with retention offers.
- The trade-off is a lower precision (F1 = 0.25), meaning some non-churners may be incorrectly flagged, but this is acceptable given the business goal.
- The model's predictions can be used to prioritize outreach, design targeted campaigns, and allocate retention resources more efficiently.
- For best results, the business should monitor the impact of interventions and continue to refine the model as more data becomes available.

8. Lessons Learned and Next Steps

- **Challenges:**
  - Strong class imbalance and weak linear relationships between features and churn made prediction difficult.
  - Even advanced techniques (cost-sensitive learning, ensembling) did not substantially improve sensitivity or F1.
- **What worked:**
  - Feature engineering and class balancing improved model performance over the raw baseline.
  - XGBoost with threshold optimization provided the best trade-off for the business goal.
- **Next steps:**
  - Explore additional features (e.g., customer demographics, external data)
  - Try alternative algorithms or ensembling strategies
  - Work with business stakeholders to refine the definition of churn and intervention strategies
  - Monitor model performance over time and retrain as needed