AIMS | African Institute for Mathematical Sciences
SENEGAL

Real-Time Customer Churn Prediction System Documentation

African Master's in Machine Intelligence (AMMI)-2025
MLOPS Project

Tiana, Abenezer, Ejah

Contents

1. Real-Time Customer Churn Prediction System Documentation

1.1 Introduction to the System

This document describes the complete implementation of the **Real-Time Customer Churn Prediction System**, designed to identify customers with high probability of churn using advanced machine learning techniques. The system processes telecommunications data to predict churn behaviors and provide actionable insights for retention strategies.

1.2 Feature Engineering

Feature engineering was performed to transform raw data into meaningful inputs for the ML models. This involved preprocessing the provided features and creating new ones to improve model performance.

1.2.1 Data Source **Source:** Kaggle Dataset (dataset, 2023)

1.2.2 Engineering Steps

Preprocessing

a. Handle missing values (e.g., impute with mean/median for numerical features).

b. Normalize/scale numerical features (e.g., using StandardScaler).

Categorical Encoding

- **Label Encoding**
- **One-Hot Encoding**

Feature Selection

- Use correlation analysis or feature importance from models (e.g., Random Forest) to select top features.
- Remove redundant features like CustomerID (non-predictive).
- Add new feature: **TotalCalls** (inbound and outbound calls.).

1.2.3 Final Features

Table 1.1: *Numerical Features used in the model*

Feature	**Description**
MonthlyRevenue	Average monthly revenue generated by the customer (in dollars).
MonthlyMinutes	Total number of minutes used by the customer per month.
TotalRecurringCharge	Total monthly recurring charges for the customer's plan.
OverageMinutes	Number of minutes used beyond the customer's plan allowance.
UnansweredCalls	Number of incoming calls not answered by the customer.
CustomerCareCalls	Number of calls made by the customer to the service department.
PercChangeMinutes	Percentage change in minutes used compared to the previous month.
PercChangeRevenues	Percentage change in revenue generated compared to the previous month.
ReceivedCalls	Number of incoming calls received by the customer.
DroppedBlockedCalls	Total number of calls dropped or blocked due to network issues.
MonthsInService	Number of months the customer has been with the service provider.
ActiveSubs	Number of active subscriptions associated with the customer's account.
RetentionCalls	Number of calls made to the retention team to address churn risk.
RetentionOffersAccepted	Number of retention offers accepted by the customer.
ReferralsMadeBySubscriber	Number of referrals made by the customer.
CurrentEquipmentDays	Number of days since acquisition of the customer's current device.
IncomeGroup	Income group of the customer.
CreditRating	Customer's credit rating score.
TotalCalls	*New feature:* Total number of calls combining inbound and outbound calls.

Table 1.2: *Categorical Features used in the model*

Feature	Description
RespondsToMailOffers	Indicates whether the customer responds to marketing offers sent by mail (Yes/No).
MadeCallToRetentionTeam	Indicates whether the customer contacted the retention team (Yes/No).
Occupation	Customer's occupation (e.g., Professional, Student).
PrizmCode	Demographic segmentation code based on Claritas PRIZM (e.g., U38 for urban, R15 for rural).

1.3 Models Used

Three machine learning algorithms were implemented and evaluated for churn prediction:

a. **Random Forest**

b. **XGBoost**

c. **Neural Network**

1.4 Model Evaluation

The models were evaluated using cross-validation and the following metrics: accuracy, precision, F1-score, and AUC-ROC.

Table 1.3: *Model evaluation results*

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	0.6610	0.4101	0.4021	0.4060	0.6452
XGBoost	0.7110	0.4965	0.2148	0.2999	0.6464
LightGBM	0.7192	0.5428	0.1618	0.2493	0.6636
Neural Network	0.6216	0.3859	0.5296	0.4465	0.6341
Stacking Ensemble	0.7037	0.4771	0.2937	0.3636	0.6617

1.4.1 Optimal Hyperparameters and Metrics Results

1.4.2 Observations

- **LightGBM** achieved the highest **accuracy (0.7192)** and **AUC (0.6636)**, indicating it provided the best overall predictive performance among the tested models.
- **XGBoost** showed competitive accuracy (0.7110) but suffered from low recall (0.2148), meaning it missed many true churn cases.
- **Random Forest** had moderate performance across all metrics, achieving a balanced trade-off between precision and recall.
- **Neural Network** exhibited the best **recall (0.5296)**, capturing more churn cases, but at the cost of lower precision and overall accuracy.
- The **Stacking Ensemble** model did not outperform the best individual model (LightGBM), suggesting that the combination of base learners did not lead to significant performance gains.

1.5 Conclusions

Final Conclusions will be there

References

dataset, K. (2023). Customer churn prediction dataset. https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom/data