Part B- Churn Prediction

Introduction

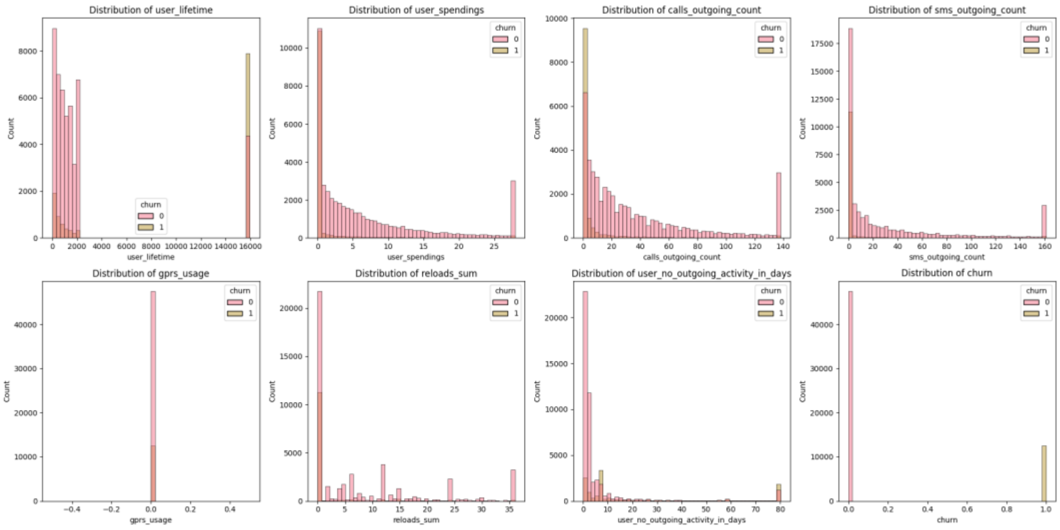
Our task is to develop a churn prediction model for a telecom company using customer behavior data to be able to identify customers at risk of churning to take actions. Our dataset contains 60,000 customer records with 66 features including usage, financial, and behavioral metrics.

Assumptions about data:

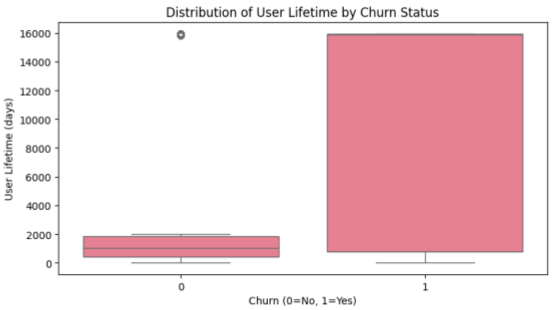
1. All features are collected over the same time period
2. Missing values indicate no activity
3. 'Last 100' metrics represent the most recent customer behavior
4. Customer IDs are unique

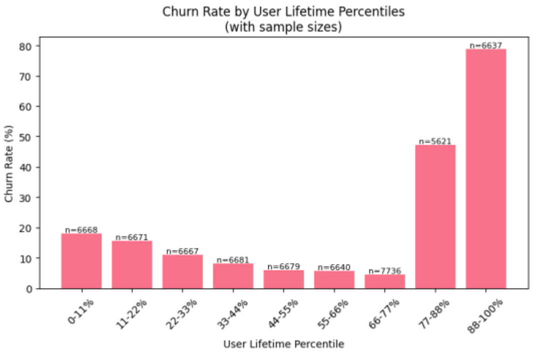
EDA

The dataset is imbalanced with 80% non-churned users and 20% churned users. Distribution of key usage metrics are as follows:

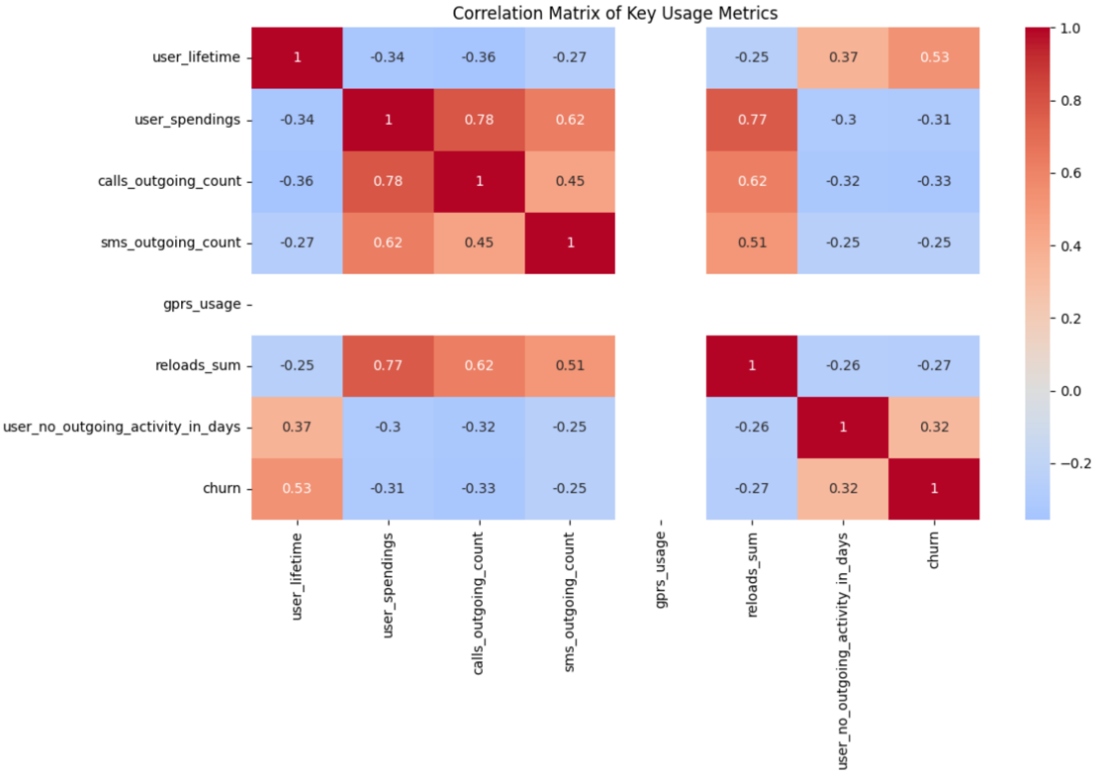


When we analyze user lifetime further, we see a U-shaped relationship. Highest churn risk in the longest lifetime percentile. Second highest risk in the newest customers. Most stable customers are in the middle range.





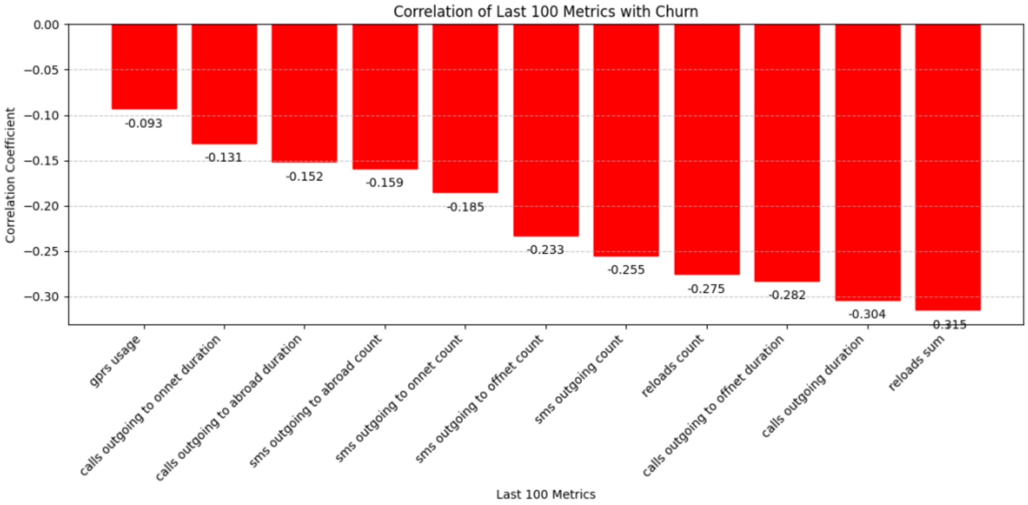
Looking at correlation of key usage metrics:



1. User lifetime shows the strongest positive correlation with churn
2. Usage metrics (calls, SMS, reloads) are negatively correlated with churn
3. Strong interconnection between spending and usage metrics
4. Inactivity days positively correlate with churn

Correlations of “last 100” metrics show:

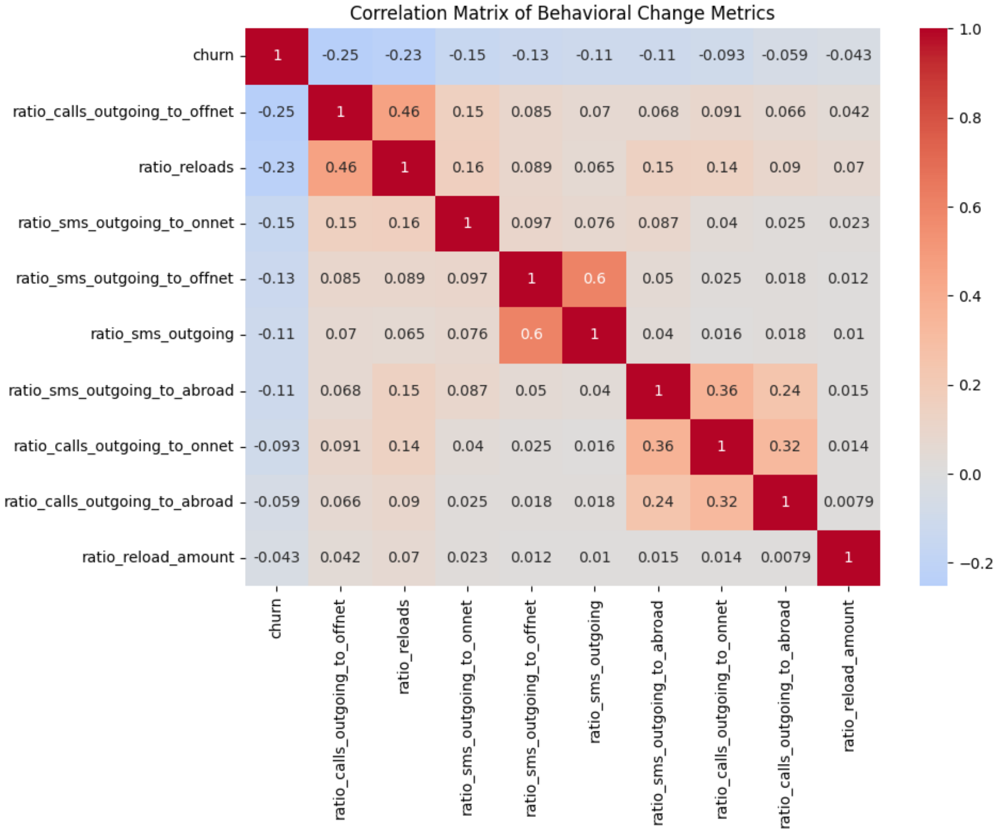
1. All "last 100" metrics show negative correlations with churn
2. Reload sum and outgoing calls are strongest predictors
3. Recent GPRS usage has the weakest correlation



To get an understanding of recency in usage metrics, I compared usage in the last 100 transactions with monthly averages for calls, SMS, and reloads. I found that:

- Churn declines in offnet calls, reload patterns and onnet SMS strongly signal churn risk.
- Offnet calls correlate with reload pattern; SMS types show moderate correlation.

Moderate correlation values highlight these as useful but complementary metrics for churn prediction models.



Feature Engineering

New features are added such as:

1. Account age, new user flag
2. Daily spending, reload patterns, balance management
3. Service distribution across onnet/offnet/abroad
4. Relative usage of calls/SMS/data
5. Recent vs. lifetime usage changes

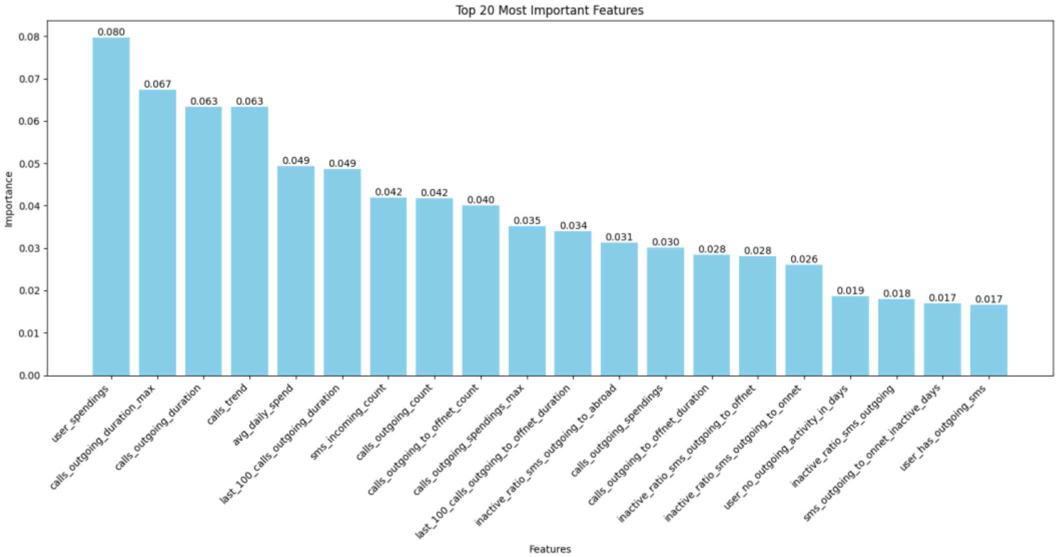
Modeling

Random Forest model is used with hyperparameter tuning and oversampling (SMOTE).
Results show that model shows strong ability to identify loyal customers and g ood balance between precision and recall for churn prediction

1. 86% overall accuracy
2. 94% precision for retained customers
3. 78% recall for potential churners
4. 0.902 ROC-AUC score

Feature Importance

User spending, calls outgoing duration, calls trend, average daily spend and SMS incoming count have the highest feature importance.


Top 20 Most Important Features
Importance
user_spendings
calls_outgoing_duration_max
calls_outgoing_duration
calls_trend
avg_daily_spend
last_100_calls_outgoing_duration
sms_incoming_count
calls_outgoing_count
calls_outgoing_to_offnet_count
calls_outgoing_spendings_max
last_100_calls_outgoing_to_offnet_duration
inactive_ratio_sms_outgoing_to_abroad
calls_outgoing_spendings
calls_outgoing_to_offnet_duration
inactive_ratio_sms_outgoing_to_offnet
inactive_ratio_sms_outgoing_to_onnet
user_no_outgoing_activity_in_days
inactive_ratio_sms_outgoing
sms_outgoing_to_onnet_inactive_days
user_has_outgoing_sms
Features