Pipeline Presentation

Sponsor - Amplero : Time Series
Team - Time Turners

Khyati Parekh
Niharika Sharma
Rajiv Veeraraghavan

Introduction

Problem Statement

- Predict user churn among prepaid users from multivariate time series
- Even a 1-3% decrease in user churn can tremendously increase in the revenue for the mobile network.

Data
Issues with the incoming data
and
Recommended solution

Data

The dataset consists of **usage statistics** and **social attributes time series** for approximately **12M** prepaid mobile phone users.

Each user is identified by a **unique entityId**. The user ids and other information have been **anonymized**. Each user is also known as an entity.

To each entityId correspond **many time series**. The period spanned by the time series is at least **120 days**, ending Sept. 27, 2017.

Data

Usage time series capture the voice, sms, data usage of each entity.
Examples: VoiceCallsPerDayTimeSeries,
CarrierReportedSubscriptionStateDeltaTimeSeries.

Social time series capture the social network structure of each entity.
Examples:
OutboundSMSNetworkPageRankLast7DaysTimeSeries,
InboundSMSNetworkOnNetFractionOfDegreeLast7DaysTimeSeries.

State and event time series capture the status of an entity and discrete events relevant to that entity.
Examples: RechargeTimeSeries, PlanDeltaTimeSeries, CallsToCareAgentTimeSeries.

File Structure

State_recharge_etc.jsonish.gz

"CallsToCareAgentPerDayTimeSeries": {

"VoiceCallsPerDayTimeSeries": { ** number of calls per day

"RichCarrierSubscriptionStateDeltaTimeSeries": {

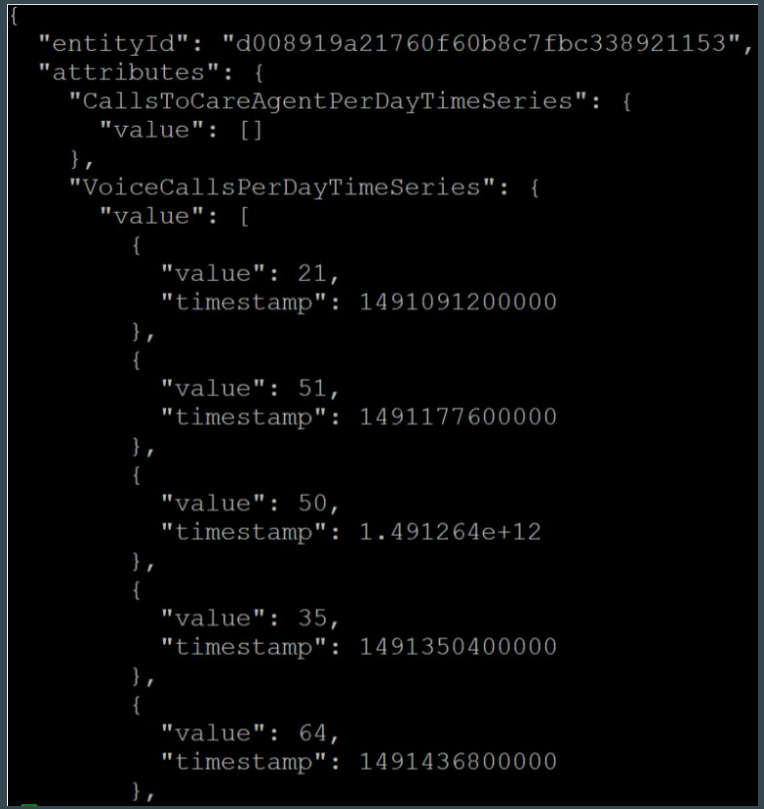
"CallsToCanadaPerDayTimeSeries": {

"CreditAppliedTimeSeries": {

"AutoPayDeltaTimeSeries": {

"RechargeTimeSeries": { ** amount and time of a pre-paid account recharge

"CarrierReportedSubscriptionStateDeltaTimeSeries": { ** is an entityID active or inactive? Needed for churn



The usage*.jsonish.gz files contain usage information for each entity. The usage is according to

SMS, Data, and Voice. These are the relevant time series:

"CompactSMSPerDayTimeSeries": {

"CompactDataKBPerDayTimeSeries": {

"CompactVoiceSecondsPerDayTimeSeries": {

However, these **time series are nested**. The idea is that SMS usage time series can be subdivided in component time series according to a finer-grained classification of usage: inbound SMS, outbound SMS, outbound SMS to international recipients, inbound SMS from international senders, etc. and similarly for other dimensions of usage. Here is an example:

"CompactSMSPerDayTimeSeries": {

"OutboundNotInNetworkInternationalCompactSMSPerDayTimeSeries": [

"OutboundInNetworkNotInternationalCompactSMSPerDayTimeSeries": [

Scaling

We have 51 gzipped files.

There are three types of files:

a) 20 usage*.jsonish.gz files contain the usage data;

b) 20 social*.jsonish.gz files contain social network data;

c) 1 file called state_recharge_etc.jsonish.gz contains state, event, and some usage data.

Each file is 8 GB or more compressed. After uncompressing each file is 25 GB or more.

Scaling

Using the following scalable methods:

- D2.8xlarge instances
- Spark on EMR - 2 core node cluster
- S3

Missing data points

We have 25 million entity IDs in the dataset, from which we extracted the entity IDs with churn activity.

How? There are multiple subscription state types - Inactive, Active and cooling (CarrierReportedSubscriptionStateDeltaTimeSeries)

Outcome - Only ~60000 entity ID have “state = cooling” which means that they left a supplier, which is 0.25% of the whole dataset

Values missing for multiple entity ID in the Time Series and sub-Time Series

Challenges in data pipeline creation

S3 Data access - Multiple boto clients

Large Compressed files - Read line by line, can’t store all the data in one variable, cannot store single Time Series data to a dataframe because of memory issues

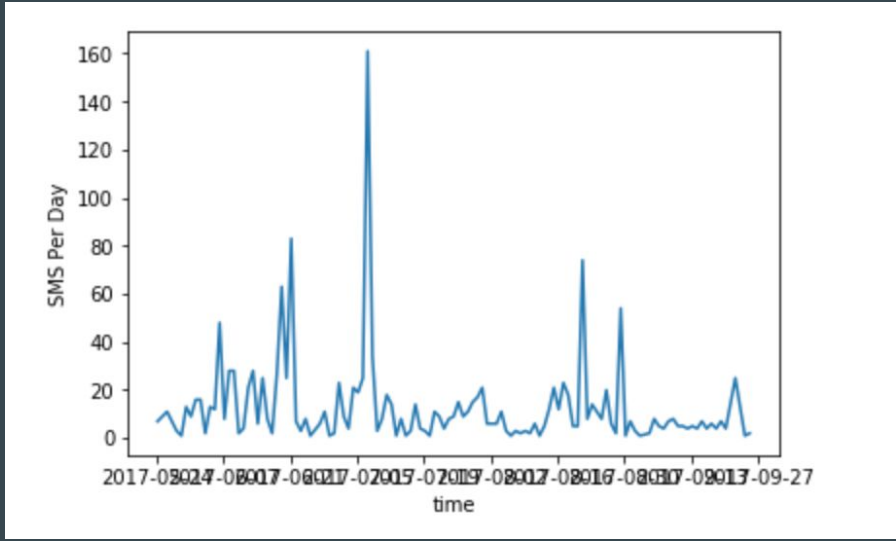
GZip documentation

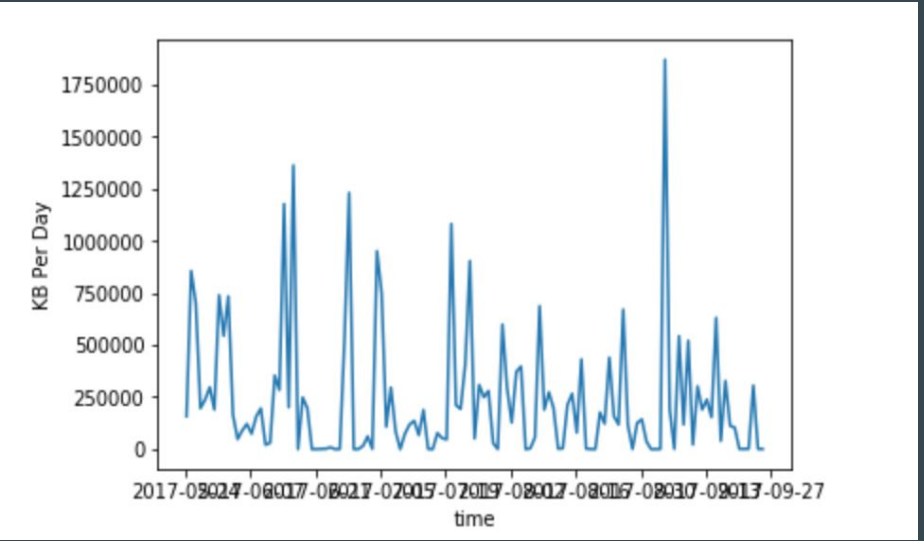
Missing values

Churner / Non-churner data extraction - EntityID with churn activity only 0.25%

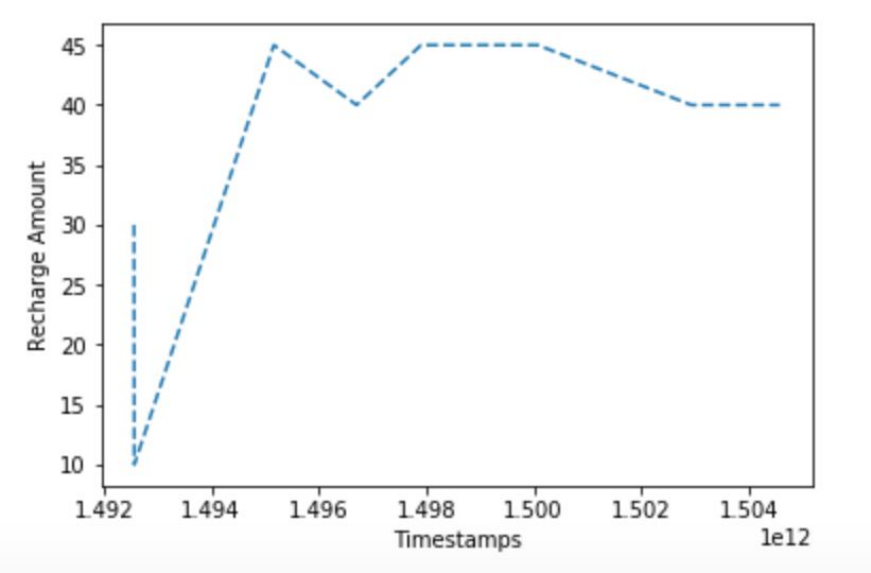
Data Analysis
and
Statistical Models

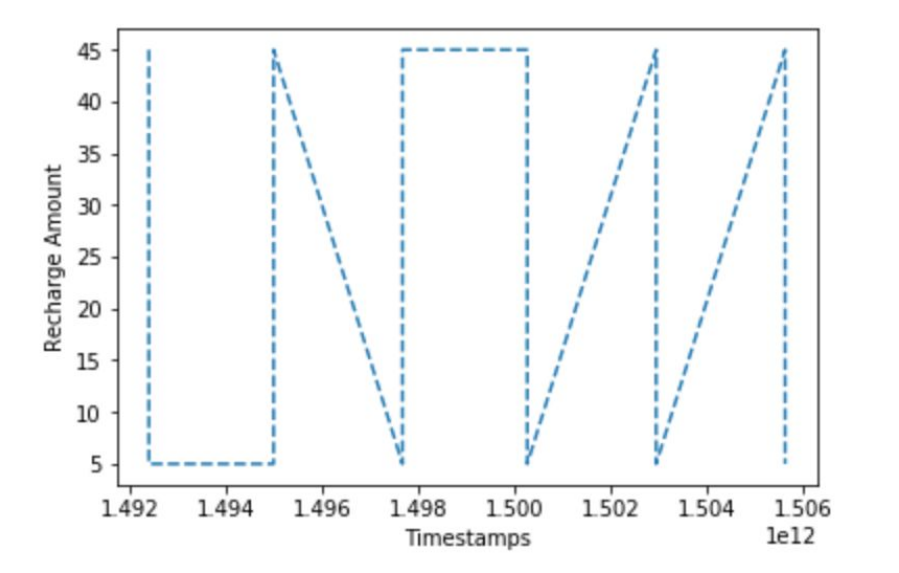
SMS/Data Usage Trends





Recharge Trends





Statistical Approaches

1. Time Series Similarity Techniques
    a. Time series for churned users might be completely different subspace compared to users who are non-churned
2. Formulating the problem as a markov chain to predict churn
    a. Hidden Markov Models, Long Short Term Memory Networks

THANK YOU!