Amplero User Churn Prediction

Khyati Parekh, Niharika Sharma, Rajiv Veeraraghavan | Megan Hazen | University of Washington

Objective

- Performed a multivariate time series analysis on the data to find the probability of a user churn out.
- Motivation of the sponsor - Being able to predict a user's propensity to churn out of a subscription, and the timing of the churn event, makes marketing more efficient.

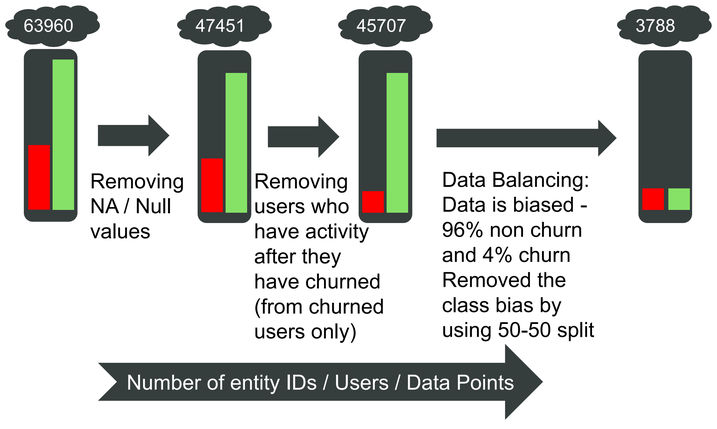
Input

- **Voice Call Time Series** - Number of calls per day
- **SMS Time Series** – Number of SMS per day
- **Data Time Series** – Data used in KB per day
- **Recharge Time Series** - Amount and time of a pre-paid account recharge
- **Carrier Reported Subscription State Delta Time Series** – Tells whether an entity ID active or inactive. Needed for churn

Feature Selection

Voting Classifier

- Features considered for each time series / Entity ID
  - Mean, Variance
  - Min, Max
  - Number of Inactive days



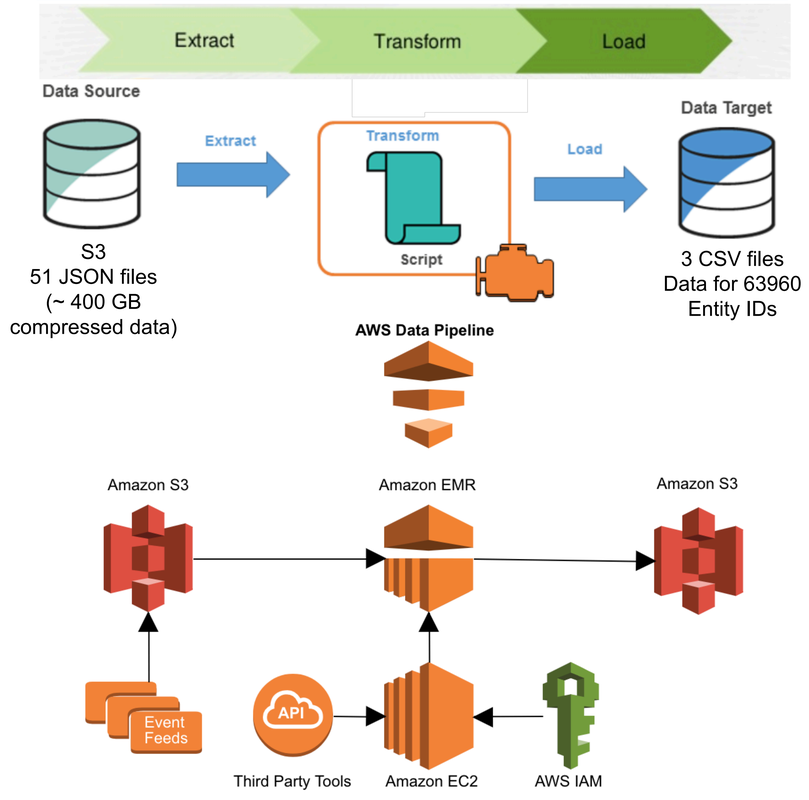
Neural Network

- Each time series is broken down into 100 chunks
  - The average value in each chunk is a feature
  - Therefore, 100 features are generated from each time series.

Models

Model name	Description
Voting Classifier	Consists of three different models - Logistic Regression, SVC [kernel="rbf], KNN [n=5] • Experimented with the feature vector (with & without inactive days) • Selected different K size for KNN • Experimented with RBF, linear and poly kernel for SVC
Neural Net	A multilayer perceptron network with 1 hidden layer and about 1000 hidden nodes to make the prediction

Method Pipelines (ETL)



Challenges

- Large Compressed files
  - Read line by line (parallel process also time consuming),
  - Cannot store Time Series data to a data frame because of memory issues.
- Not enough/ satisfactory GZip library documentation

Assessment

Voting Classifier (Logistic Regression, SVC [kernel="rbf], KNN [n=5]) including inactive days feature vector				
AUC	ACCURACY	RECALL	F1	PRECISION
0.812	0.810	0.761	0.806	0.857

Voting Classifier (Logistic Regression, SVC [kernel="rbf], KNN [n=5]) excluding inactive days feature vector				
AUC	ACCURACY	RECALL	F1	PRECISION
0.637	0.635	0.598	0.631	0.667

Neural Network				
AUC	ACCURACY	RECALL	F1	PRECISION
0.7489	0.75	0.7321	0.7446	0.7575

Neural Network (excluding 14 days of timeseries data just before churn)				
AUC	ACCURACY	RECALL	F1	PRECISION
0.658	0.658	0.644	0.653	0.663

- The best result was using Voting Classifier with Best Models: Logistic Regression, RBF SVC and KNN with soft voting [2, 3, 1] having ~0.80 as the final score.
- Neural Network model gave an accuracy of 75% when we consider the time series data till the date of churn
  - If we ignore 2 weeks of data just before churn, the accuracy of the model drops to 65%.

Summary

- Number of inactive days seems to be a very good predictor of performance. If we ignore number of inactive days as a feature, the predictive power of the model drops.
- Voting Classifier and the neural network models can enable mobile networks to identify users that may churn in the future. This information can be used to provide tailor made promotions to users.

Future Work

- We believe we can obtain improved performance if we consider convolution neural networks in addition to the current multilayer perceptron.
- Considering a much larger dataset to avoid issues with overfitting in a real time environment.

UNIVERSITY of WASHINGTON
eScience Institute