Mobile Application User Segmentation and Churn Prediction

Capstone Project

Liana Mehrabyan
Supervisor: Arnak Dalalyan

Akian College of Science and Engineering
American University of Armenia

June, 2018

Outline

1. Problem Statement and Motivation
   - Mobile App User Segmentation
   - Dataset Description
2. Proposed Approaches
   - User Segmentation
   - Data Labeling
   - Churn Prediction
3. Experimental Results

Outline

1. Problem Statement and Motivation
   - Mobile App User Segmentation
   - Dataset Description
2. Proposed Approaches
   - User Segmentation
   - Data Labeling
   - Churn Prediction
3. Experimental Results

App Development Industry

- One of the leading business fields.
- 69.7 billion US dollars revenue in 2015.
- 188.9 billion US dollars revenue prediction for 2020.

App Development Industry

Main souces of App revenue:

- App Store purchases
- In-App purchases
- In-App advertisement

App Development Industry

Criterion for a successful app:

App Development Industry

Criterion for a successful app:

- High download rates?

App Development Industry

Criterion for a successful app:

- High download rates?
  - No. 21% of users use the app just once after downloading.

App Development Industry

Criterion for a successful app:

- High download rates?
  - No. 21% of users use the app just once after downloading.
- Alternative?

App Development Industry

Criterion for a successful app:

- High download rates?
  - No. 21% of users use the app just once after downloading.
- Alternative?
  - High number of active users.

Tracking App Usage Data

- Highlighting some app drawbacks that otherwise cannot be seen.
- Segmentation of users to meet customers' needs.
- Making the app more personalized.
- Identifying users that are at risk of churning.

Outline

Dataset Description: Session-Based Data

The Data

Dataset of 11.000.000 observations and 8 features. Each observation represents information about one session of a user.

The Features

- device id
- timestamp
- crashed
- duration
- screens count
- OS version
- app version
- country

Dataset Description: User-Based Data

Derivation of new dataframe

Such dataset does not meet the needs of this project.
Solution: derivation of a per-user dataframe of 16 features.

Features

device id, last session, total duration, average duration, average number of generated screens, average IAT: average time between two sessions, num ber of crashes, number of sessions, crash rate, max IAT, Recency, R score, F score, M score and RFM.

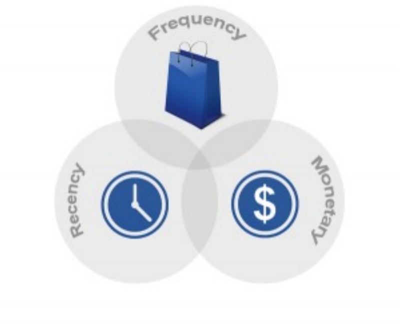
Outline

User Segmentation Approaches

The following approaches were used to analyze the data from user segmentation perspective:

- RFM Customer Analysis
- Gaussian Mixture Models
- Decision Tree Classifiers
- Random Forest Classifiers

RFM Customer Analysis



- Recency: the time between the present and the last product consumption.The higher the interval, the lower is the recency score.
- Frequency: number of times the customer uses the product over a certaintime interval. The higher the number of usage, the higher the F score.
- Monetary: the monetary value of the consumption.

RFM Customer Analysis

- Obtain R, F and M values.
- Assign scores on a scale from 1 to k based on what quantile interval the value falls into.
- Calculate RFM score by $100 * R + 10 * F + M$.

RFM Customer Analysis

- Recency: time interval between the last recorded session time in the dataframe and the last session of the user.
- Frequency: number of sessions per month.
- Monetary: total duration spent using the application.

Gaussian Mixture Models

Preliminary Analysis Conclusions:

- Many erroneous entries.
- Many outliers: univariate outlier detection tools are useless.

Gaussian Mixture Models

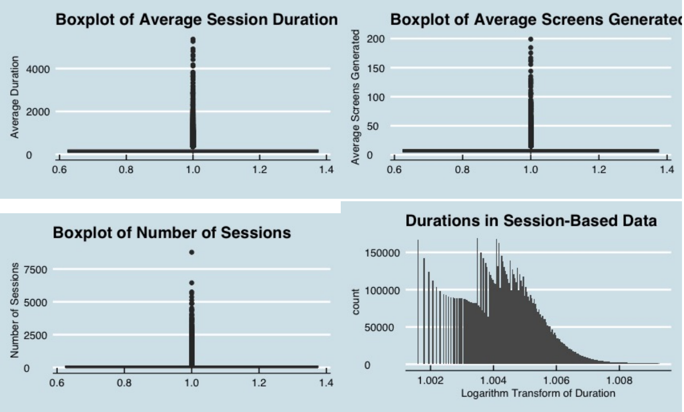
Preliminary Analysis Conclusions:

- Many erroneous entries.
- Many outliers: univariate outlier detection tools are useless.
- Possibility of having mixture data.

Gaussian Mixture Models

Preliminary Analysis Conclusions:

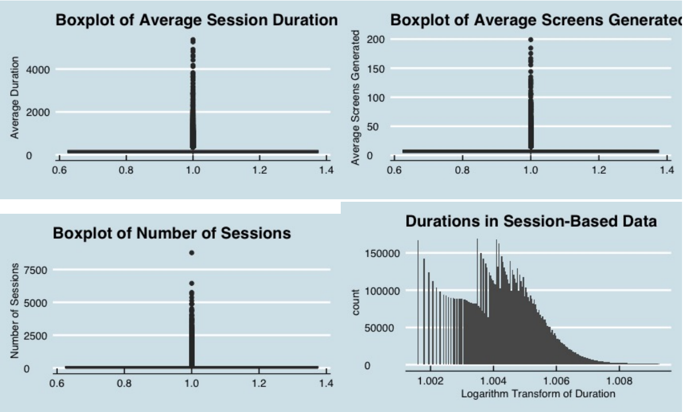
- Many erroneous entries.
- Many outliers: univariate outlier detection tools are useless.
- Possibility of having mixture data.



Gaussian Mixture Models

Preliminary Analysis Conclusions:

- Many erroneous entries.
- Many outliers: univariate outlier detection tools are useless.
- Possibility of having mixture data.



Gaussian Mixture Models

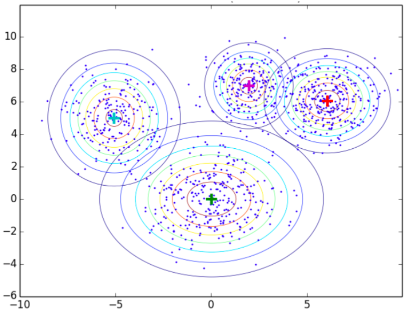
Definition

A K-component Gaussian mixture is a weighted sum of K Gaussian densities given by the form:

$$p(x) = \sum_{k=0}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{1}$$

where each Gaussian density $N(x|\mu_k, \Sigma_k)$ is called a **component** having its mean μ_k and covariance Σ_k and the parameters π_k are called **mixing coefficients**.

Gaussian Mixture Models



Parameters:

- $\pi = [\pi_1\ \pi_2,\ ..,\ \pi_K]$
- $\mu = [\mu_1,\ \mu_2,\ \ldots\ ,\ \mu_K]$
- $\Sigma = [\Sigma_1, \Sigma_2, \ldots, \Sigma_K]$

GMM: Parameter Estimation

- Models are typically learned by using maximum likelihood estimation techniques.
- Finding the maximum likelihood solution for mixture models is usually analytically impossible.
- Numerical methods used instead.

Expectation Maximization Algorithm

Goal

Given a set of observations x_1,x_2,....x_N to model Gaussian Mixtures, one can represent this data as an NxD matrix X where the i^{th} row is the transpose of x_i. The goal is to maximize the log-likelihood given by:

$$log\ p(X;\pi,\mu,\Sigma) = \sum_{n=1}^{N} ln\Big\{\sum_{k=1}^{K}\pi_k N(x_b;\mu_k,\Sigma_k)\Big\} \tag{2}$$

Expectation Maximization Algorithm

The EM algorithm can be summarized as follows:

Expectation Maximization Algorithm

The EM algorithm can be summarized as follows:

1. Initialize the parameters π, μ, Σ and the initial value of the log likelihood.

Expectation Maximization Algorithm

The EM algorithm can be summarized as follows:

1. Initialize the parameters π, μ, Σ and the initial value of the log likelihood.
2. E-step: estimate the posterior probability $\gamma(z_{nk})$ of the i^{th} observation belonging to the k^{th} component:

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n; \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i N(x_n; \mu_i, \Sigma_i)} \tag{3}$$

Expectation Maximization Algorithm

The EM algorithm can be summarized as follows:

1. Initialize the parameters π, μ, Σ and the initial value of the log likelihood.
2. E-step: estimate the posterior probability $\gamma(z_{nk})$ of the i^{th} observation belonging to the k^{th} component:
$$\gamma(z_{nk}) = \frac{\pi_k N(x_n; \mu_k, \Sigma_k)}{\sum_{i=1}^{K} \pi_i N(x_n; \mu_i, \Sigma_i)} \tag{3}$$
3. M-step: update the parameters π, μ, Σ given the current posterior probabilities:
$$\mu_k^{t+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \tag{4}$$
$$\Sigma_k^{t+1} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{t+1})(x_n - \mu_k^{t+1})^T \tag{5}$$
$$\pi_k^{t+1} = \frac{N_k}{N} \tag{6}$$

Expectation Maximization Algorithm

- Evaluate the log likelihood

$$log\ p(X;\pi,\mu,\Sigma)=\sum_{n=1}^{N} ln\Big\{\sum_{k=1}^{K}\pi_k N(x_b;\mu_k,\Sigma_k)\Big\} \tag{7}$$

- Check for convergence of the parameters or the likelihood.

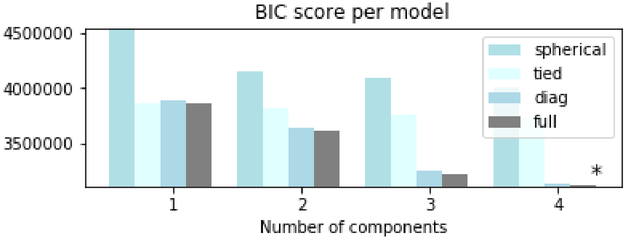
Model Selection

Definition

Given a finite set of models, let MLL_i be the maximum log likelihood of the i^{th} model. And let d_i be the dimension of the i^{th} model. Then, the penalty BIC_i for the model M_i is given by:

$$BIC_i = MLL_i - \frac{1}{2} d_i logn \tag{8}$$

Model Selection:



Gaussian Mixture Models

Resulting sub-populations:

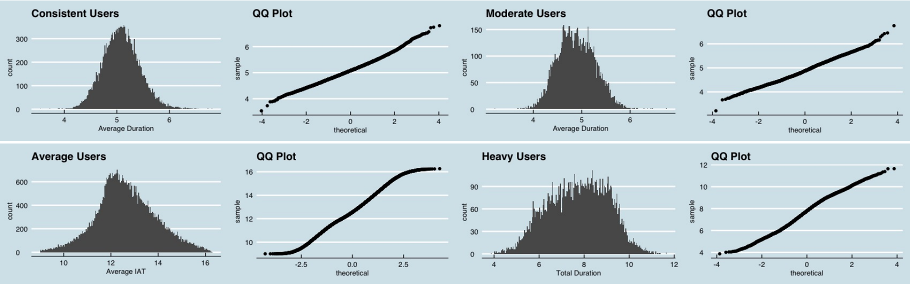
Average Characteristics of User Subgroups						
User Group	Duration	Num. of Screens	Num. of Sessions	RFM Score	IAT	Users in the group
Group 1	172.52	7.57	367.30	333.0	40372	17923
Group 2	142.15	6.71	182.78	251.91	62352.8	8313
Group 3	131.68	6.40	23.87	275.63	649615	54000
Group 4	425.15	15.05	12.03	277.43	869347	9188

Gaussian Mixture Models

User sub-populations:

- **Consistent Users:** users with many sessions of adequate duration and screen count as well as high RFM score and low session inter arrival times.
- **Moderate Users:** users having moderate amount of sessions with corresponding duration and a moderate RFM score.
- **Average Users:** users having less average records but, surprisingly, high RFM score.
- **Heavy Users:** users having less but sessions that are considerably longer with more screens generated but with used within larger time intervals.

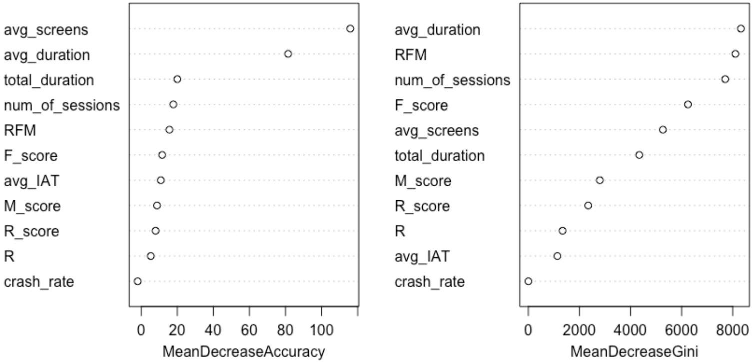
Customized Approach For Each Subgroup: Outlier Detection



Values being more than 3 standard deviations away from the mean were replaced with the mean value of the feature.

Understanding User Behavior

Decision Trees: understanding what leads the user to a certain sub-population.
Random Forest: understanding the importance of the features.



Outline

Labeling The Data

Motivation

In practice, developers are unable to retrieve information whether a certain user has deleted the app or not. Hence, there is a need for defining customer churn in a way that it will be statistically meaningful and have profound theoretical basis.

Alternating Renewal Processes

Definition

Suppose that (Ω, F, P) is a probability space and $I \subset \mathbb{R}$ has finite cardinality. Suppose further that for each $\alpha \in I$, there is a random variable $X_\alpha : \Omega \rightarrow \mathbb{R}$ defined on (Ω, F, P). The function $X : Ix\Omega \rightarrow \mathbb{R}$ defined by $X(\alpha, \omega) = X_\alpha(\omega)$ is called a stochastic process with indexing set I, and is written $\{X_\alpha, \alpha \in I\}$.

Alternating Renewal Processes

- An alternating renewal process alternates between two states up and down.
- Define $\{U_n, n \geq 1\}$ and times system being down $\{D_n, n \geq 1\}$.
- Consider a state variable $Z(t)$ that is 1 is the system is up at time t and 0 if the system is down at time t.

Definition

A renewal process $N(t), t \in T$ with a state variable $Z(t)$ and duration sequences D_n and U_n is called an alternating renewal process.

Renewal Reward Theory Results

Limiting Proportions

long-run proportion up $= \lim_{x \to \infty} \frac{1}{t} \int_0^t Z(s)ds = \frac{E[U]}{E[U]+E[D]}$

$\lim_{x \to \infty} P(Z(t) = 1) = \frac{E[U]}{E[U]+E[D]}$

$\lim_{x \to \infty} P(Z(t) = 0) = \frac{E[D]}{E[U]+E[D]}$

Churn Definition

Session Inactivity

From Alternating Renewal Process theory, the probability of the session being inactive in the long run can be calculated by:

$$\frac{averageIAT}{averageIAT + averageduration}$$

Definition

A user is assumed to be at risk of churning if:

$$\frac{Recency}{Recency + Lastsessionduration} \geq 1.1 \frac{averageIAT}{averageIAT + averageduration}$$

Outline

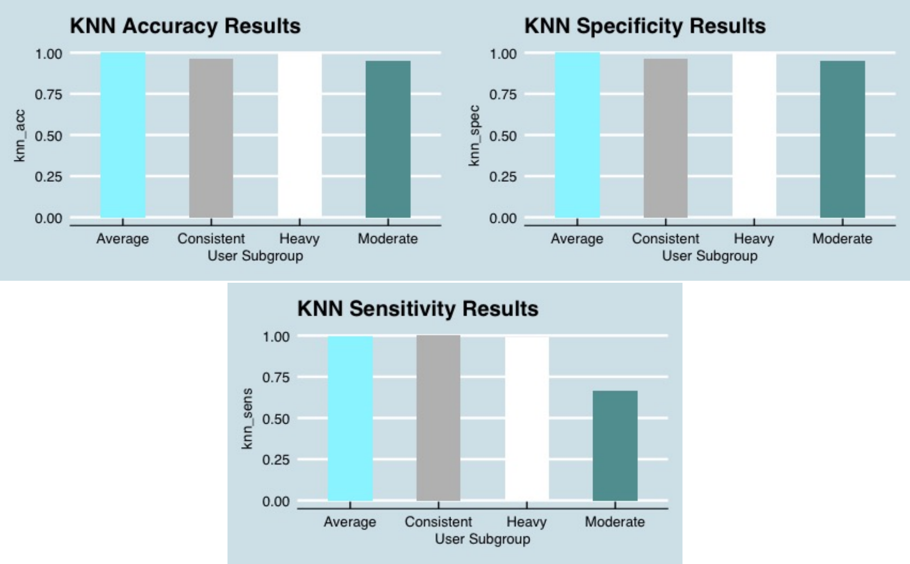
Classification Algorithms

- Naive Bayes Classifiers
- K Nearest Neighbours
- Support Vector Machines

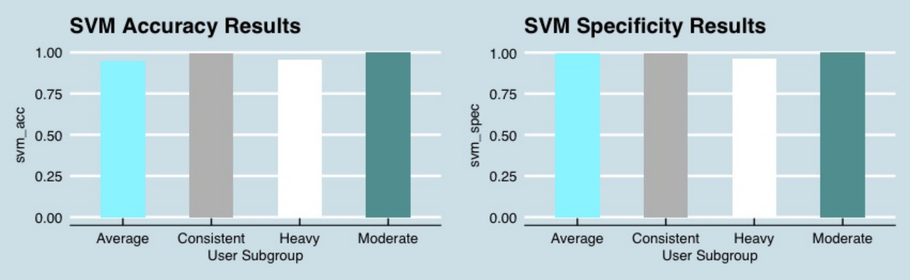
Supervised Learning

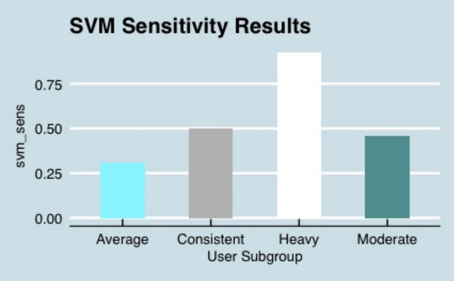
- Models based on individual characteristics of sub-populations.
- KNN, SVM Naive Bayes classifiers used.
- 70% of the data used for training and validation purposes. Other 30 used for testing. (Maintaining the proportion of target variable classes.)
- High class imbalance.
- SMOTE (Synthetic Minority Over-sampling Technique) used to handle class imbalance.

Classification Results: KNN

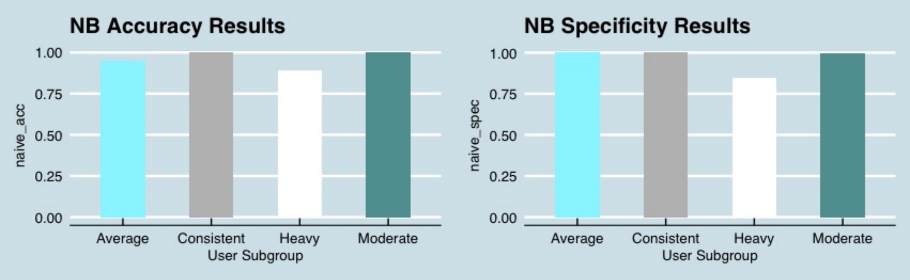


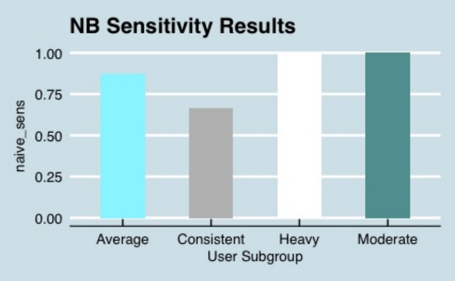
Classification Results: SVM





Classification Results: Naive Bayes





Choosing The Best Models

Definition

Consider Youden's J statistic given as:

$$J = \text{sensitivity} + \text{specificity} - 1$$

Best models per subgroup:

Consistent Users	KNN	0.964800
Moderate Users	Naive Bayes	0.998711
Average Users	KNN	0.993610
Heavy Users	KNN	0.988640

Summary

- Integration of customer segmentation in this field can considerably improve revenues.
- Analysis of user behavior data tracked while customers use the application is a useful tool for improving retention rates.
- Further work:
  - Extending the dataset with features considering user touches, buttons, app design and user demographics.
  - Developing real life business tool for updating model parameters from constantly flowing data.

Q & A

Questions?

Thank you for your attention!

For Further Reading I

- Worldwide Mobile App Revenues. *The Statistical Portal*. Retrieved from: `https://www.statista.com/statistics/269025/worldwide-mobile-app-revenue-forecast/`

- Divya D. Nimbalkar, Asst Prof. Paulami Shah, *Data mining using RFM Analysis*. International Journal of Scientific and Engineering Research, Volume 4, Issue 12, December, 2013, ISSN 2229-5518. Retrieved from: `https://pdfs.semanticscholar.org/5aa6/bcb19728998ff6f97cb68ce9e9670293be97.pdf`

For Further Reading II

- Khajvand M., Zolfaghar K., Ashoori S., Alizadeh S., *Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study*, Procedia Computer Science, Volume 3, 2011, pp. 57-63.
  Retrieved from:
  `https://doi.org/10.1016/j.procs.2010.12.011`
- Bishop C.M., *Pattern Recognition and Machine Learning*, 2006, pp. 430-435, ISBN-13: 978-0387-31073-2.
- Author NA, *The Bayes Information Criterion*, Massachusetts Institute of Technology, Dec. 2015.
  Retrieved from:
  `http://www-math.mit.edu/~rmd/650/bic.pdf`

For Further Reading III

- Erar B., *Mixture model cluster analysis under different covariance structures using information complexity*, University of Tennessee, Knoxville, 2011.
  Retieved from:
  `http://trace.tennessee.edu/cgi/viewcontent.cgi?article=2096&context=utk_gradthes`
- Russel S., Norvig P., *Artificial Intelligence, a Modern Approach, third edition.*, ISBN-13: 978-0-13-604259-4
- Fristedt B., Gray L., *A Modern Approach to Probability Theory.*, Birkhauser, Boston, MA, 1997.

For Further Reading IV

- Sigman K., *Introduction to Renewal Theory*, Columbia University, 2009.
  Retrieved from:
  http://www.columbia.edu/~ks20/stochastic-I/stochastic-I-RRT.pdf

- Whitt A., *Alternating Renewal Processes and The Renewal Equation*, Columbia University, 2013.
  Retrieved from:
  http://www.columbia.edu/~ww2040/3106F13/lect1119.pdf

- Chawla N.,Bowyer K.,HallL.,Kegelmeyer P. *SMOTE:SyntheticMinority Over-sampling Technique*,Journal of Artificial Intelligence Research 16 (2002) 321357.
  Retrieved from:
  https://arxiv.org/pdf/1106.1813.pdf

For Further Reading V

Dullaghan C., Rozaki E., *Integration of Machine Learning Techniques To Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers*, International Journal of Data Mining Knowledge Management Process (IJDKP) Vol.7, No.1, January 2017. Retrieved from: https://arxiv.org/pdf/1702.02215.pdf