<Page intentionally left blank>
Using Application Generated Data to Provide Personalized User Experience in Software Applications.

By Priya Kher

Submitted to the MIT System Design and Management Program on May 16\textsuperscript{th}, 2018 in partial fulfillment of the requirements for the Degree of Master of Science in Engineering and Management

Abstract

Delivering quality user experience is the most critical goal of any product development and marketing team in any organization. With the advancement of technologies in the fields of data science and data computation, it is now possible to know users more efficiently and create solutions that satisfy their needs to the fullest.

In this thesis, I explore how the digital ecommerce and online content provider companies are utilizing many different personalization methods which are helpful in increasing the rate of successful transactions, however, a similar trend is not visible in SaaS applications. Cloud computation has made software both easily accessible and replaceable, putting a lot of stress on both the value of the product as well as the user experience.

Many software companies still follow the traditional approach of creating static personas for product design and marketing purposes and create one fits all solution. Machine/application data, which is continuously generated by the software applications, tracking each and every user activity, can be extremely useful in understanding the user behavior and thus giving companies the ability to create more personalized and adaptive solutions.

I explore data generated about a pedagogical website at MIT which is used to support instruction in computation—open to students from all the departments. I applied machine learning algorithms to show that there are different clusters/classes of students in a class. By tracking student activity and performance on class website, it can be predicted which class they belong to. This information can be used to develop customized solutions for all the students.

Thesis Advisor: John R. Hauser

Title: Kirin Professor of Marketing, MIT Sloan School of Management
<Page intentionally left blank>
Acknowledgements

I would like to thank my Thesis advisor Professor John R Hauser for providing me an opportunity to work under his guidance. I am grateful for his valuable insights and thorough feedback for my thesis.

I would like to express my sincere gratitude and thanks to Prof Abel Sanchez for providing data of his class for my research. I am also thankful to him for providing valuable data computation and visualizations lessons in his class that immensely helped in my thesis research.

I would also like to thank System Design and Management department for guiding and supporting me patiently throughout my studies at MIT.

Finally, I would like to thank my partner, Prathmesh Phalke, for without his loving support, I would not be able to complete this educational journey.
# Table of Contents

1. Introduction ......................................................................................................................... 9

2. Background .......................................................................................................................... 10
   2.1 Personalization in digital Commerce ................................................................................. 10
   2.2 Software as a service ......................................................................................................... 13
   2.3 User Profiling ................................................................................................................... 14
   2.4 Adaptation ........................................................................................................................ 15
   2.5 Challenges ....................................................................................................................... 16

3. Literature .............................................................................................................................. 17
   3.1 User Need Analysis ......................................................................................................... 17
   3.2 Application Data Analysis ............................................................................................... 18
      3.2.1 K Means - Clustering of Unlabeled data ................................................................. 18
      3.2.2 Naive Bayes ........................................................................................................... 20

4. Case Study ............................................................................................................................ 21
   4.1 Overview of Test Application ........................................................................................... 21
   4.2 Data ................................................................................................................................... 23
   4.3 Data analysis ..................................................................................................................... 24
   4.4 Results .............................................................................................................................. 29

5. Conclusion and Recommendations ....................................................................................... 30

6. Future Work ......................................................................................................................... 31

7. References ............................................................................................................................ 32

8. Appendix ............................................................................................................................... 34
## Table of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The user modeling-adaptation loop (De Bra, 2017)</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>By region</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>Page views per user</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>Device Type</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>Number of page views per page in the website</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>Page view activity per day (in class time)</td>
<td>26</td>
</tr>
<tr>
<td>7</td>
<td>Page View activity per user (outside class time)</td>
<td>26</td>
</tr>
<tr>
<td>8</td>
<td>User Events data</td>
<td>26</td>
</tr>
<tr>
<td>9</td>
<td>Intermediate – Advance Content</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>No experience – Beginner Content</td>
<td>29</td>
</tr>
<tr>
<td>11</td>
<td>Post Survey Data</td>
<td>29</td>
</tr>
<tr>
<td>12</td>
<td>Personalization and adaptation framework based on user data</td>
<td>31</td>
</tr>
</tbody>
</table>
1 INTRODUCTION

In this digital transformation era, developing a web or mobile based application to fulfill user needs has become significantly easy and cheap over the last few years, making it difficult to compete on product price. For every use case there are several providers available in market ready with a solution. End users have an immense power to choose amongst thousands of online services available over the web. In this fierce environment, even technically advanced systems or products can fail if they are not providing the best experience to users. Creating a remarkable customer experience is one the most challenging tasks that many companies are currently facing to maintain competitive differentiation.

Over the last few years marketing teams have been using data to understand the right market segment for their products and to identify customer needs. The growing use of data driven product design has helped firms to build more useful and reliable products for their users. Though the use of data is focused on improving product development and increasing profits, it is also educating users that they can have more and more personalized products/services which will fulfill their needs and requirements to the fullest.

Companies are also leveraging data to ask for personalized needs by shifting gears towards advanced technologies such as voice assistants and chatbots, which use machine learning, deep learning and NLP, taking personalization to a next level. Every company would like to build an application which pleases all its users, but it is not an easy task. In addition to the cost of building such a complex application which is personalized for each and every user, it is extremely difficult to both accurately understand each and every individual completely and to accurately take an action such as recommendation, adaptation etc. based on user needs. Though personalization with 100% accuracy many not be yet achievable, but some steps, which are implementable, should be taken now to achieve and maintain significant competitive differentiation.

For many years, the first step in any product design is to identify how your idea/product can serve a variety of potential end users. This process typically begins with identifying segments based on certain high level factors. Once the segment is finalized, marketers try to identify the user
 personas. User personas are created by doing extensive market research, interviewing or surveying users, observing the daily tasks of potential users, and are then categorized into different buckets. This is the most common approach that almost all tech companies use.

These personas are used by different departments such as marketing, sales, and product to improve their respective activities. In the next chapters, I cover how personalization of an application affects (what?). Despite creating different personas, users with different personalities and context are not able to efficiently utilize applications that have been designed to fulfill their exact needs. Especially in SaaS applications, where the intent of the user is to get a task done rather than purchasing something, it is even more important that users are able to extract value from the application and are able to navigate themselves in the application. End users of SaaS application are also end users of many retail and content provider websites and have been already exposed to wide ranges of personalized features and content. These ends users are getting used to these personalized features and are now demanding and expecting similar features in more complex SaaS applications. It is high time that SaaS companies adopt new intelligent technologies to improve customer experience.

2 BACKGROUND

2.1 PERSONALIZATION IN DIGITAL COMMERCE

According to Gartner (Gartner, n.d.), “Personalization is a process that creates a relevant, individualized interaction between two parties designed to enhance the experience of the recipient. It uses insight based on the recipient's personal data, as well as behavioral data about the actions of similar individuals to deliver an experience to meet specific needs and preferences”

For the last few years, personalization has been a one of the top priorities of many digital commerce companies to attract and retain more customers and thus increase their revenues. Many large scale companies have been successful by implementing these methods.
Many of these personalization efforts have been restricted either to advertising or to recommending new products. In advertising, ads across various channels are customized according to the preferences of the end users and are shown only to those users whose preferences best match to the product attributes. In product recommendation, new products or content is suggested to users to purchase or watch based on their browsing and purchase history. Overall, the main objective of all these methodologies is achieve a successful purchase.

In the digital commerce industry, many personalization methods of varied sophistication have been in use for many years. They have proven to be successful in increasing the revenues. Dhanalakshmi & Lakshmi, 2014 conducted a survey to discover what different methods of personalization were currently being used in the market for personalization.

Table 1: List of Personalization Methods

<table>
<thead>
<tr>
<th>No.</th>
<th>Source of Data</th>
<th>Technique</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Server log files+referrer+agent</td>
<td>multivariate K-means method</td>
<td>Personalized URL recommendation</td>
</tr>
<tr>
<td>2</td>
<td>Firefox add on Alter Ego</td>
<td>C/NC, TF, TF-IDF and personalized B25 weighting</td>
<td>Personalized search results from search Engine</td>
</tr>
<tr>
<td>3</td>
<td>Analyzing Documents</td>
<td>Conceptual Term Frequency</td>
<td>Document Clustering</td>
</tr>
<tr>
<td>4</td>
<td>Mobile phone with sensors</td>
<td>Graph based Ontological representation and Cosine similarity</td>
<td>Measure Personalization for Mobile Users</td>
</tr>
<tr>
<td>5</td>
<td>Client</td>
<td>Ontology representation and RSVM</td>
<td>Personalization for Mobile users</td>
</tr>
<tr>
<td>6</td>
<td>Mobile phone</td>
<td>XML Schema</td>
<td>Personalization for Mobile users</td>
</tr>
<tr>
<td>7</td>
<td>Documents analyzed from sentence, document and corpus level</td>
<td>Natural Language Processing, Verb argument structure, Conceptual term frequency</td>
<td>Document Clustering</td>
</tr>
<tr>
<td>8</td>
<td>Click through data(positive and negative preferences)</td>
<td>Bipartite graph, Personalized Agglomerative</td>
<td>Clustering Concept based user profiles</td>
</tr>
<tr>
<td></td>
<td>Community contributed Photos</td>
<td>Bayesian learning</td>
<td>Recommendation of next travel location</td>
</tr>
<tr>
<td>---</td>
<td>-------------------------------</td>
<td>-------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>10</td>
<td>User Domain Knowledge, Domain Knowledge</td>
<td>(Association Rule Interactive post-Processing using Schemas and Ontologies)</td>
<td>Interesting Association Rules</td>
</tr>
<tr>
<td>11</td>
<td>User’s browsing behavior from client Add-on in the Browser</td>
<td>Ontology, Personalized page view (PPV) graph</td>
<td>Personalization (Preferred information in the shortest path)</td>
</tr>
<tr>
<td>12</td>
<td>Overview of various personalization approaches with examples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>access log files of ISP cache proxy servers</td>
<td>Fuzzy-Temporal Association Rule Mining</td>
<td>Dynamic personalization</td>
</tr>
<tr>
<td>14</td>
<td>Unseen sites</td>
<td>Bayesian learning and expectation-maximization (EM) techniques</td>
<td>Extracting Information from unseen websites</td>
</tr>
<tr>
<td>15</td>
<td>Concepts from Web snippets</td>
<td>frequent patterns</td>
<td>Personalized search results from search Engine</td>
</tr>
<tr>
<td>16</td>
<td>Analyzing Documents</td>
<td>tfidf+ distributional features</td>
<td>Document Clustering</td>
</tr>
<tr>
<td>17</td>
<td>Website and individual web page</td>
<td>tree templates and schema</td>
<td>Extracting data from web pages</td>
</tr>
<tr>
<td>18</td>
<td>Browsing behavior from Mobile phones</td>
<td>ontology-based, multi-facet (OMF) profile, RSVM</td>
<td>Personalized search results from search Engine</td>
</tr>
<tr>
<td>19</td>
<td>User’s search query</td>
<td>Ontology, semantic mining</td>
<td>Knowledge based information</td>
</tr>
<tr>
<td>20</td>
<td>Domain Knowledge and user preferences</td>
<td>Ontology</td>
<td>Recommendation of web page</td>
</tr>
<tr>
<td>21</td>
<td>Overview of Semantic Web and Ontologies</td>
<td>Ontology</td>
<td>Recommendation of web page</td>
</tr>
<tr>
<td>22</td>
<td>KBData Mining Using Semantic Web</td>
<td>Ontology</td>
<td>Personalization</td>
</tr>
<tr>
<td>23</td>
<td>Weblog and SQL log from SDSS Sky server</td>
<td>Normalization</td>
<td>Web log Analysis</td>
</tr>
</tbody>
</table>
2.2 SOFTWARE AS A SERVICE

Software as a Service is a licensing and delivery model used to rent a software application over the web instead of purchasing the software (PC Magazine, n.d.). A subscription model provides users flexibility to pay for the software only when they need it and doesn’t bind them for long term commitments or contracts. Even though SaaS has been around since 1960 and was used earlier by IBM and other mainframe companies to provide data storage and computing facilities for top banks, it started gaining popularity in late 90s with the advent of internet as well as the shrinking cost of computers.

The Internet provided a means to host and access these applications globally. In the early years, SaaS applications were most popular with customer resource management (CRM) software, but now they have spread across all the industries and applications. The growing popularity of on-demand software has made SaaS a very successful delivery model. Almost all major on premise software applications have either already moved to the cloud or are in process of doing so. Though a SaaS type model brings a much value to end users by giving them flexibility, it also adds a risks of increasing churn rates compared to the previous on premise model.

The growth of platform as a service model such as AWS, Microsoft Azure, has given many small businesses incentives to enter the world of SaaS, creating many players in the market and making competition even more difficult. With the availability of software on web, end users don’t have to invest their time and money in infrastructure making it easier to switch providers.

To manage the churn rates, SaaS applications are continually trying to ensure that customers receive ongoing value and continue to subscribe to them. SaaS applications, like on premise applications, are aimed to solve complex problems and deliver high value to customers, making them complex web applications compared with other e-commerce applications over the web. SaaS application providers spend considerable time, effort, and money to conduct demos with the clients and to create online learning material to better coach their end users and help them adopt the application as early and smoothly as possible. Many large-to-small scale SaaS companies such as Microsoft, Cisco, Salesforce,
Tableau etc. have their own certifications to both market their application as well as create proficient users of their application.

2.3 USER PROFILING

For any user-profiling method, getting user information is the most crucial part. Two most popular methods for user profiling are:

1. Explicitly asking user information, e.g. collecting user demographic information while setting up user accounts. Other ways of getting explicit feedback could be by asking if he/she liked or disliked a recent purchase/product or by ask users to rate the products. Though explicit methods provide rich and useful data, these methods burden the user. Not many users are enthusiastic about providing so much information or even accurate information.

2. Implicitly extracting information: User information such as browsing history/activity, search history etc., can be extracted using browser cache/agents, proxy servers, web logs, search logs etc. ([Brusilovsky, Kobsa, & Nejdl, 2007])

As discussed earlier, creating Personas is one of the popular methods used by designers especially in early product design ([Pruitt & Grudin, 2003]) and marketers to selectively target end users

Concept of Personas was first introduced by software designer Alan Cooper in his book “The Inmates Are Running the Asylum” to serve in the UX design process. Gartner([Gartner, n.d.]) defines Personas as a means of putting a more-human face on a segment. Personas take the dry facts of segments and give them a more-human form, infusing them with personal features and emotions. Organizations may find it easier to create an experience for a persona versus a specific segment of the employee population.

Researchers clearly define persona as a set of user behavior. Unfortunately, they are often used as a synonym for:

- Demographic — a subset of a population usually based on age, gender, ethnicity, language, education, wealth, etc.
- Market Segment — a group of consumers to whom your product/service serves a value
• Role — a description of a set of responsibilities and objectives for an individual function that is linked hierarchically in a broader organizational structure.

Due to this misrepresentation of personas or user profiles, designers often misunderstand their end users as well misunderstand the concept of personalization. People in same segment, role or demographic have individual needs, behaviors and knowledge which highly influence their interaction with the product.

Some companies do extensive research of user behavior or personalities to create good quality user personas and use it as a guide for system design, however, static personas aren’t very helpful and can’t provide long lasting user experience.

2.4 ADAPTATION

Customization and adaptation are two different processes that are essential to enabling personalization and satisfying user needs in any application. Customization is a process that is guided by requests made specifically by the user; whereas, adaptation is process that is guided by the system which interprets the responses from the end user.

An application should automatically adapt to several attributes, such as, the capabilities of the device at hand, network connectivity, and most importantly the user’s activities, location, and context. However, applications must provide end users with some control to customize the adaptation process any level.

Adaptations can be of two types: 1) Content adaptations and 2) User interfaces (UI) adaptations.

Content adaptation could range from personalized recommendations or information displayed within the application to personalized alerts (recommendations, offers) or emails being sent to the user to increase likelihood of successful transactions, e.g. recommendations provided Netflix, Amazon and other content provider or e-commerce applications.
**UI adaptations** are changes made to the application interface itself to better suit the needs of the user. These adaptations could be as simple, as automatically adjusting the layout of the application depending on type of device (platform adaptation) of the user, to as smart, as changing the question on a Google form based on the language of the question the user has provided or changing styles based on location.

Adaptations of both the content and UI can either be expert or data driven. Expert driven adaptations in applications are either based on the prior knowledge of a human being who has strong knowledge to guide end user especially in a less complex system or based on prior knowledge and discussions of group of experts who collectively create a semantic database of rules of adaptations.

In large, complex, and dynamic applications, adaptations are mostly data driven as it becomes impossible for humans to do so. Data driven adaptations focus on analyzing the content as well the user behavior. In large social networking, e-commerce, or online content provider applications, collaborative filtering has been the most popular tool of adaptation. To avoid risks of collective bias as well as attention bias, i.e., people are not able to see anything other than recommendation and hence buy only recommended products, hybrid approaches of both data and experts are gaining popularity.

### 2.5 Challenges

Many organizations struggle in developing and executing a digital personalization strategy successfully due to lack of knowledge and expertise. Personalization is still considered as a secondary value over the core value of the SaaS applications and hence doesn’t often receive attention or funding in many companies. As seen in Table 1 above, there are many methods currently being user for personalization/user profiling, and form application column of these methods, we can infer that many of these methods are specifically used in e commerce, content provider or search engines.

However, this trend is changing, and more and more companies are creating a data science team/department to focus on utilizing the data generated from their applications to improve application performance, adoption, and user experience.
Organizations, which are trying to develop ubiquitous applications that are smart enough to adapt itself to all possible scenarios and provide an effortless experience to all its users, face many difficulties. It is challenging to foresee, and provide support for, diverse circumstances, settings and technologies. It is also challenging to support and manage devices required to build applications (user interfaces) that are robust enough to support complex and dynamic interactions dependent on contexts of use.

To understand implicitly the user needs it is very difficult and important to get meaningful data from the application and to build models on the non-labeled data. Setting up code to generate every user activity is not a big issue but to filter out relevant features is.

Unsupervised learning is one of the challenging fields within Machine learning. It is very subjective, as there is no simple goal for the analysis, such as prediction of a response and can be biased towards intuitions.

One of the problems with many data mining methods is that they generate a large number of patterns, and most of the patterns, while being statistically acceptable, are trivial, or not relevant to the application at hand.

The deep integration of the adaptation rules with the content and engine result into very complex systems. In addition to the challenges of developing these systems, adaptive systems pose new challenges: diminished predictability and comprehensibility, diminished controllability, obtrusiveness, infringement of privacy, and diminished breadth of experience (Jameson, n.d.)

3 LITERATURE

3.1 USER NEED ANALYSIS

Implementing any new product or a new feature requires a good understanding of the needs and expectations of the end users. To improve user experience, it’s important to understand existing user behavior and identify user pain points. Three of the four phases described in Voice of customer [(Griffin & Hauser, 1993)] - customer needs, a hierarchy, priorities, and customer perceptions of performance were used in this research. Qualitative interviews and surveys were conducted of all end users and stakeholders
involved in the system to identify the needs. Needs were manually structured into categories and prioritized using anchored scale method by asking users to rate them on a scale of 1-5. Customer perceptions are derived from market research[(Popovici & Mironov, 2015)] about how customers perceive the performance of products that compete in the market being studied. If no product exists as yet, the perceptions indicate how customers now fulfill those needs.

(Keahey, n.d.) It is important to see the complete picture of customers, i.e., to have a complete understanding of not only how each customer is transacting with your company, but how each customer is finding out about offerings, comparing alternatives, discussing products and services in their social networks and interacting with related products and services. Visualization can be very helpful in making the individual analytic components understandable and by tying them together into a comprehensible “big picture.” In addition, visualization can help you can start with a simple customer data set and add more views of the customer.

3.2 APPLICATION DATA ANALYSIS

Data generated by a course website, which will be discussed in detail in Chapter 4, was used in this research. The data was analyzed using K means clustering method. The data is further classified using Naïve Bayes classifier.

3.2.1 K Means - Clustering of Unlabeled data

In unsupervised learning, clustering is an important first step to understand the data and identify the hidden patterns in an unknown dataset. The aim of the K-means algorithm is to divide M points in N dimensions into K clusters so that the within-cluster sum of squares is minimize [(Hartigan & Wong, 1979)]. K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results.
The algorithm consists of two separate phases. The first phase selects K centers randomly, where the value K is fixed in advance. The next phase is to take each data object to the nearest center. Though many other metrics can be used, Euclidean distance is generally considered to determine the distance between each data object and the cluster centers.

When all the data objects are included in some clusters, the first step is completed and an early grouping is done. Recalculating the average of the early formed clusters. This iterative process continues repeatedly until the criterion function is minimized.

The Euclidean distance is defined as between one vector, \( x = (x_1, x_2, ..., x_n) \), and another vector, \( y = (y_1, y_2, ..., y_n) \). The Euclidean distance \( d(x_i, y_i) \) can be obtained as follow:

\[
d(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

The process of K-means algorithm is as follows:

Input: Number of desired clusters, K, and a database \( D = \{d_1, d_2, ..., d_n\} \) containing n data objects.

Output: A set of K clusters

Steps:

1) Randomly select K data objects from dataset D as initial cluster centers
2) Calculate the distance between each data object \( d_i (1 \leq i \leq n) \) and all K cluster centers \( c_j (1 \leq j \leq K) \)
3) Assign data object \( d_i \) to the nearest cluster.
4) For each cluster \( j (1 \leq j \leq K) \), recalculate the cluster center.
5) Return to step 2, loop over the dataset D until the centroid values are fixed and the cluster member does not move to another cluster.

To calculate optimum number of clusters in K mean clustering, elbow method can be used.

The idea of the elbow method is to run K means clustering on the dataset for a range of values of K (e.g. k from 1 to 10), and for each value of k calculate the sum of squared errors (SSE). Plot a line chart of the SSE for each value of K. With the addition of each new cluster the SSE will reduce and will eventually reduce to 0 when K is equal to number of data points as each data point will be a cluster. Goal is to select a value of K which has low SSE and adding a new cluster doesn’t reduce SSE significantly. Elbow represents
where we start to have diminishing returns by increasing $K$ and hence is called as "elbow criterion". 
[(KETCHEN Jr. & SHOOK, 1996)]

3.2.2 Naïve Bayes

[(Webb, Keogh, Miikkulainen, Miikkulainen, & Sebag, 2011)] The Naïve Bayes Classifier technique is based on Bayesian Theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naïve Bayes can often outperform more sophisticated classification methods and has been employed in many applications. It assumes that all attributes (i.e., features) of the examples are independent of each other given the context of the class, i.e., an independence assumption. While this independence assumption is often violated in practice, naïve Bayes nonetheless often delivers competitive classification accuracy.

Computational efficiency: Training time is linear with respect to both the number of training examples and the number of attributes, and classification time is linear with respect to the number of attributes and unaffected by the number of training examples.

Incremental learning: Naïve Bayes operates from estimates of low order probabilities that are derived from the training data. These can readily be updated as new training data are acquired. Because naïve Bayes always uses all attributes for all predictions, if one variable is missing, go with the prior, resulting in graceful degradation in performances. It is also relatively insensitive to missing attribute values in the training data due to its probabilistic framework.

Bayes rule provides a decomposition of a conditional probability that is frequently used in a family of learning techniques collectively called Bayesian Learning with an assumption that the attributes are conditionally independent given the class. (Munro et al., 2011):

$$P(y|x) = P(x|y) \frac{P(y)}{P(x)}$$

$P(x)$ is called the prior probability, $P(x|y)$ is called the posterior probability, and $P(y|x)$ is called the likelihood.

For attribute value data:
\[
P(x|y) = \prod_{i=1}^{n} P(x_i|y)
\]

where \(x_i\) is the value of the \(i\)th attribute in \(x\), and \(n\) is the number of attributes.

\[
P(x) = \prod_{i=1}^{k} P(c_i)P(x|c_i)
\]

where \(k\) is the number of classes and \(c_i\) is the \(i\)th class. Thus, \(P(y|x)\) can be calculated by normalizing the numerators of the right-hand-side of the equation.

For categorical attributes, the required probabilities \(P(B)\) and \(P(A|B)\) are normally derived from frequency counts stored in arrays whose values are calculated by a single pass through the training data at training time. These arrays can be updated as new data are acquired, supporting incremental learning. Probability estimates are usually derived from the frequency counts using smoothing functions such as the Laplace estimate or an \(m\)-estimate.

For numeric attributes, either the data are discretized or probability density estimation is employed.

4 CASE STUDY

4.1 OVERVIEW OF TEST APPLICATION

I used data from the website “onexi.org” [(Onexi.org, n.d.)] that was made available for MIT subject 1.001: “Engineering Computation.” The subject focuses on teaching students’ fundamentals of computational thinking and data science in engineering computation. Because the subject is open to all the MIT students, students in the subject come from many different programs and backgrounds ranging from engineering courses such as Course 2 (Mechanical Engineering), Course 1 (Civil and Environmental Engineering), Institute for Data, Systems and Society, course 6 (Electrical Engineering and Computer Science) to management courses such as Leaders in Global Operations (LGO-MBA), System Design and Management.

Unlike other MIT classes which use MIT course management sites such as Canvas, Stellar, or LMOD, 1.001 uses its own website to host lectures, assignments, learning materials, and schedule information. Class
instructors have created a lot of learning data in form of videos, slides, and sample code files for students to learn about data science and computing in their own time.

A class learning website is very similar to an enterprise, SaaS application. Students enroll in the class, class instructors provide lessons in the classes and develop learning material and assignments, instructors and TAs together with the help of class website and communication tools such as piazza ensure that students are able to learn and realize the value, and have a good experience. If they are not able to enjoy the class, they would either drop the class in middle of the semester or not provide good year-end evaluations which can result in class cancellation.

Enrolling in a class in an educational institution is similar to purchasing a SaaS Software. Like SaaS software, it is purchased based on value and commitment is contingent on experience. Like SaaS freemium model (30-day trial period), students shop classes in first 1-2 weeks to determine for which classes they should register. Usually in the shopping period, students analyze the syllabus/course content, schedule, class expectations, and class experience to decide whether they can commit to the class. In the first trial month of SaaS applications, some users use the software, analyze if it is adding any value to their company for the given price and if it seems fair they may opt for monthly or yearly subscription. Typically, most SaaS providers provide limited user/data access in the trial period, therefore, restricting the decision to purchase the software in hands of very few people. After a company purchases the software subscription, more users start using it and explore various features in the application, and if they are not satisfied they may not renew their subscription. Keeping end users continually satisfied is a challenging task and identifying the reasons of dissatisfaction is even more challenging. Even high-value-delivering SaaS applications struggle to keep their end users engaged and provide them high quality experience. Professors face similar challenges in classes when users have different backgrounds. This problem is highly visible in online classes where anyone in the world can enroll in the class. In on premise classes, it can be challenging for the professor to adapt to the needs of every student in the classroom, however, a class website, which has the coaching content, can be very well adapted to analyze the class as well as adapt itself for different type of students in the class.
4.2 DATA

**Qualitative**

- **Interviews:** Based on interviews of some of the current students in the class, I identified following positive and negative feedback/sentiments for the class.

  (+) One of the instructors explained some really difficult concepts very well.

  (+) Videos posted on the website were very helpful in understanding the principles

  (-) Class website is very confusing, especially when it comes to tracking the schedule

  (-) It is not clear from the lecture site which classes have been taught and which is the latest class.

  (-) Assignments were too easy.

  (-) Assignment was way too difficult.

  (-) Finding information is really difficult and confusing.

  (-) It is difficult to identify on the website which link is a video vs slide vs zip file

  If someone is little late in class, the student has no idea which files to download to start the class work

**Quantitative**

1. Data was accessed from heap analytics (Heap Analytics, n.d.), which is a plugin used in the website to collect the clickstream data generated by the user. Data collected is from **1/1/2018 – 3/18/2018**

The data generated from the application was in 3 types:

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>User Data</td>
<td>A list of all users who were active in the time range of this export. Each line contains one JSON object which represents a single user, including all properties associated with the user.</td>
</tr>
<tr>
<td>Raw Page Data</td>
<td>A list of raw page views captured by Heap, including all properties</td>
</tr>
</tbody>
</table>
### Raw Event Data

A list of defined or custom events. Each line contains one JSON object which represents a single occurrence of an event, with all its properties, including which user performed the event and when it happened. Raw events are the underlying representation of event data, e.g. clicks with CSS selectors for web apps or touch / gesture events for mobile apps. Each line contains one JSON object representing a single event.

Refer to the [Appendix] for detailed data structure.

### 4.3 Data Analysis

**Part: I - Data Properties and Visualization**

As there was no training data available, my first step was to analyze the data by plotting multiple visualizations.

**Facts**

- The class website doesn’t explicitly ask for a user login, however, heap-analytics uses IP, browser and other user property information to create user ids
- The class website is open and is also used in some other distance schools as well.
- The class officially began on February 7th.
- On the first day there were X users. Current students: Y
- The class lecture schedule is in appendix []

For the purpose of relevancy to class, I excluded all the data prior to **02/01/2018**. I also filtered all the data which was accessed from outside “**United States**” as the course website was also used by a group of students in Saudi Arabia

1. Number of unique users since **02/01/2018**: 490
2. Number of page views per user: 29.4
3. Standard Deviation: **75.24**
Figure 6: Page view activity per day (in class time)

Figure 7: Page View activity per user (outside class time)

Figure 8: User Events data
Event data is more detailed as it provides information about every activity a user did in a given session. Data visualizations provided following insights:

1. List of areas/cities from where the website was accessed.
2. There was a wide range of distribution in:
   a. Page views per user
   b. Video views per user
   c. Slides views per user
3. Certain videos/lectures were more/less viewed than others.
4. User activities varied around different lecture and homework days.

Part: II - Machine learning

**K Means Clustering**

1. Clickstream data from heap analytics was exported in JSON format.
2. Dataset was imported in JavaScript files to filter the data.
3. Attributes for K means were calculated using JavaScript functions:
   a. Number of times videos watched by the student.
   b. Number of times lecture slides were viewed by the student.
   c. Number of time code material was downloaded by the student.
4. JSON Data was converted to csv format using the NPM package
5. Open the filtered csv file in R.
6. Missing values were replaced by 0 because missing value means that student never views/downloaded the material
7. Normalize the data for all the attributes.
8. Used elbow method to calculate optimum number of clusters
9. Initialize with 3 clusters.
11. Plot the clusters using package PLOTLY.
12. Analyze clusters.
Learning videos

- o: Medium Engagement, ◻: Low Engagement, ■: High Engagement

**Naïve Bayes classification.**

K means clustering provided an intuition that the class had 3 different groups of students who engaged with the application (Class website) differently. The purpose of Naïve Bayes is to predict the level of student knowledge or performance based on the application usage patterns.

Following steps were taken to create features to run naïve Bayes model

1. Extract transaction in JSON format from the click stream tracking plugin and convert it to the CSV file
2. Discretize continuous variables to factors based on expert advice.
3. Ideally, knowledge/performance level of students should be categorized based on their class scores or a pre-test, however, as per the class rules that data was not accessible. Students were categorized based on type of content they had watched. Students that watched videos such as home-work guides were marked as “no experience or beginners content”.
4. Using all these variables create a csv file
5. Split data into train and test data in 70%:30% ratio.
6. Run naïve Bayes in R studio
7. Measure results:
   a. Sensitivity = True Positive / Positive
   b. Specificity = True Negative / Negative
   c. Accuracy = (True Positive + True Negative) / (Positive + Negative)
4.4 RESULTS

K means clustering showed that there are the different clusters in a same class and Naïve Bayes shows that user usage data can be used to predict user category.

As per K means analysis: 5% of the class is very highly engaged, 20% high-medium engaged and 75% low-medium.

Naïve Bayes: Accuracy: 86%, Sensitivity: 0.9080, Specificity: 0.7333.

It was interesting to observe that, within both knowledge categories, there was different engagement behavior:

A post survey was conducted to explicitly ask students their experience in coding to confirm the results.
5 CONCLUSION AND RECOMMENDATIONS

As observed in the results, there clearly are clusters of students with different level of engagement, knowledge and performance level in the same class. Similarly, in software applications, users are from different backgrounds, knowledge and skill levels and preferences. With the availability of data and state of the art technologies, it is possible to develop adaptive solutions to give a more personalized experience to end users and help them maximize their value.

In the coming years, if companies do not deploy personalized solutions and continue providing one fit all solution, they will struggle to maintain a competitive advantage. In this class example, the data shows that many students, who watch only advance content, are not very engaged. My hypothesis is that these students are not very well challenged. This can be further tested by using actual performance scores. In software, if advanced end users are not highly engaged, is an indication that the application needs an upgrade or otherwise the end user will move to another application which provides more features and value.

On the other hand, there are students who watch help guides but are not overall engaging enough, such students are not able to get the complete value out of this class; they need attention. It can be further researched if performance of these students can be improved by sending them automated emails or notifications which would recommend them about topics that they haven’t viewed or about topics they should revise more to clear their concepts. Similarly, in software, users who are not engaged and are not efficiently utilizing the application should be recommended of features and be guided.

One of the challenges in performing this exercise was to get good quality data and set benchmarks for performance levels. For data collection, software designers and architects along with the data and product team must decide early on in the design phase what data points and events should be tracked.

Many software companies have already developed their own certification programs at different knowledge and performance levels to encourage their users to adopt and market their products. They could use similar criteria to judge the performance of their users. However, these performance benchmarks must be continually updated by tracking their user data.
The visibility of different clusters in engagement and content type in a simple class website application shows how differently users use a same application. In a more complex software application this problem is even more prominent and needs to be addressed.

6 FUTURE WORK

For the simplicity of both this paper and the subject, simple data mining methods were employed for classification of students. However, for complex software applications, more advanced and sophisticated machine learning methods such as neural networks, Natural language processing, speech recognition and adaptation methods such as semantic rules should be explored. Also, further analysis should be done with clear performance metrics.

Additionally, personalization and adaptation methods should be researched in mobile applications, especially because SaaS applications haven't yet transitioned into this domain. Mobile applications are more complex as there is much less surface area for interaction and average user attention is lower than on desktop, but can provide rich data such as location information which can help better understand user context.


http://www.jmlr.org/papers/volume6/ando05a/ando05a.pdf

https://www.gartner.com/document/3839665?ref=solrAll&refval=201517962&qid=a40e4bbd47eeca2d499d55f4


https://www.gartner.com/document/3869863?ref=solrResearch&refval=201507615&qid=91e786c11eafb9481af696e7b4
<table>
<thead>
<tr>
<th>Data Source</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>User Data</strong></td>
<td><strong>Example:</strong> { &quot;user_id&quot;: 19866170676131, &quot;initial_referrer&quot;: &quot;direct&quot;, &quot;initial_landing_page&quot;: &quot;onexi.org/&quot;, &quot;initial_browser&quot;: &quot;Chrome 64.0.3282&quot;, &quot;initial_device_type&quot;: &quot;Desktop&quot;, &quot;initial_platform&quot;: &quot;Windows 10&quot;, &quot;initial_ip&quot;: &quot;182.73.1.130&quot;, &quot;initial_country&quot;: &quot;India&quot;, &quot;initial_region&quot;: &quot;National Capital Territory of Delhi&quot;, &quot;initial_city&quot;: &quot;Delhi&quot;, &quot;joindate&quot;: 1519303087007, &quot;properties&quot;: {} },</td>
</tr>
<tr>
<td><strong>Page Data</strong></td>
<td><strong>Example:</strong> { &quot;ip&quot;: &quot;96.40.154.63&quot;, &quot;city&quot;: &quot;Alhambra&quot;, &quot;path&quot;: &quot;/lectures.html&quot;, &quot;title&quot;: &quot;Lectures - Engineering Computation&quot;, &quot;domain&quot;: &quot;onexi.org&quot;, &quot;object&quot;: &quot;pageview&quot;, &quot;region&quot;: &quot;California&quot;, &quot;browser&quot;: &quot;Chrome 63.0.3239&quot;, &quot;country&quot;: &quot;United States&quot;, &quot;library&quot;: &quot;web&quot;, &quot;platform&quot;: &quot;Mac OS X 10.13.2&quot;, &quot;device_type&quot;: &quot;Desktop&quot;, &quot;landing_page&quot;: &quot;onexi.org/&quot;, &quot;user_id&quot;: 470707468338919, &quot;time&quot;: &quot;1514857429043&quot;, &quot;session_time&quot;: &quot;1514857412729&quot;, &quot;pageview_time&quot;: &quot;1514857429043&quot;, &quot;session_id&quot;: &quot;6741896447900808&quot;, &quot;pageview_id&quot;: &quot;4069969414821422&quot; },</td>
</tr>
<tr>
<td><strong>Event Data</strong></td>
<td><strong>Example:</strong> { &quot;ip&quot;: &quot;96.40.154.63&quot;, &quot;city&quot;: &quot;Alhambra&quot;, &quot;href&quot;: &quot;lectures.html&quot;, &quot;path&quot;: &quot;/&quot;, &quot;type&quot;: &quot;click&quot;, &quot;title&quot;: &quot;Engineering Computation&quot;, &quot;domain&quot;: &quot;onexi.org&quot;, &quot;region&quot;: &quot;California&quot;, &quot;browser&quot;: &quot;Chrome 63.0.3239&quot;, &quot;country&quot;: &quot;United States&quot;, &quot;library&quot;: &quot;web&quot;, &quot;platform&quot;: &quot;Mac OS X 10.13.2&quot;, &quot;hierarchy&quot;: @div.; navbar; navbar-fixed-top; navbar-inverse; {role=navigation}; @div.; container; {@div.; collapse; navbar-collapse; } @ul.; nav; navbar-nav; {li; } @a; {href=lectures.html}; &quot;, &quot;target_tag&quot;: &quot;a&quot;, &quot;device_type&quot;: &quot;Desktop&quot;, &quot;target_text&quot;: &quot;Lectures&quot;, &quot;landing_page&quot;: &quot;onexi.org/&quot;, &quot;user_id&quot;: 470707468338919, &quot;time&quot;: &quot;1514857428226&quot;, &quot;session_time&quot;: &quot;1514857412726&quot;, &quot;pageview_time&quot;: &quot;1514857412726&quot;, &quot;session_id&quot;: &quot;6741896447900808&quot;, &quot;pageview_id&quot;: &quot;5751005051518137&quot; },</td>
</tr>
</tbody>
</table>