Online Video Data Analytics

Yaohui Ye
Wenxuan Cai
Benjamin Le
Jefferson Lai
Pierce Vollucci
George Necula, Ed.
Don Wroblewski, Ed.

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-112
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-112.html

May 14, 2015

Acknowledgement

We would like to thank George Necula for advising us throughout the entirety of this Capstone project. We would also like to thank Jibin Zhan from Conviva for introducing the problem space to us and Pat McDonough from Databricks for providing extensive technical support throughout our use of Databricks.

University of California, Berkeley College of Engineering

MASTER OF ENGINEERING - SPRING 2015

Electrical Engineering and Computer Sciences

Data Science and Systems

Online Video Data Analytics

Yaohui Ye

This **Masters Project Paper** fulfills the Master of Engineering degree requirement.

Approved by:

1. Capstone Project Advisor:

Signature: ____________________________ Date _____________

Print Name/Department: George Necula/Electrical Engineering and Computer Sciences

2. Faculty Committee Member #2:

Signature: ____________________________ Date _____________

Print Name/Department: Don Wroblewski/Fung Institute for Engineering Leadership

Abstract

This capstone project report covers the research and development of Smart Anomaly Detection and Subscriber Analysis in the domain of Online Video Data Analytics. In the co-written portions of this document, we discuss the projected commercialization success of our products by analyzing worldwide trends in online video, presenting a competitive business strategy, and describing several approaches towards the management of our intellectual property. In the individually written portion of this document, we discuss our implementation of two machine learning techniques, k-means clustering and logistic regression, and give detailed evaluation of these techniques on our dataset.

Contents

*Co-written with Benjamin Le, Jefferson Lai, Pierce Vollucci and Wenxuan Cai

I. Introduction

This report documents the Online Video Data Analytics capstone project completed in the course of the Data Science and Systems focus of the Master of Engineering degree at UC Berkeley. Through the collective efforts of Benjamin Le, Jefferson Lai, Pierce Vollucci, Wenxuan Cai, and Yaohui Ye, our team has not only characterized the need for effective data analysis tools in the domain of online video data, but has also developed analysis tools which attempt to address this need. As we will describe in detail in our Individual Technical Contributions, our work has produced many important findings and we have made significant strides towards a complete implementation of these tools. However, at the time of the writing of this report, additional work is required before our tools can be considered complete. That being said, our substantial progress has allowed us to form a very clear vision of what our finished tools will look like and how they will perform. Our vision leads us to believe that, once finished, our tools can be of great potential value to entities within the online data analytics industry. In order to understand how best to cultivate this value, we have extended our vision to depict tools to marketable products and we have evaluated the potential for our team to establish a business offering these products. In doing so, we have performed extensive research of the current market and industry which our potential business would be entering. The remainder of this report presents our findings and is divided into five sections. First we introduce our industry partner Conviva in the Our Partner section. Second, we present the objective of our work and the motivation behind the resulting products in the Products and Value section. Third, we introduce and describe the dataset leveraged by our products in the Our Dataset section. Fourth, our team characterizes our industry as well as our competitive strategy in the Trends, Market, and Industry section. Fifth, in our Intellectual Property section, we describe how we plan to protect the value of our work. Sixth, the Individual Technical Contributions section of this report details our specific contributions toward the goals of our project. Finally, the Conclusion section contains a retrospective analysis of the significance of this work and provides an outlook on the potential for continuation of our work in future endeavors.

II. Our Partner

This project is sponsored by Conviva, a leading online video quality analytics provider. Conviva works with video content providers, device manufacturers, and developers of video player libraries to gather video quality metrics from content consumers. Through our partnership with Conviva, we have access to an anonymized portion of their online video quality metric dataset for the development of our products. We also have access to Conviva engineers for collaboration purposes who provide domain knowledge and on site support. For the purpose of the business analysis forthcoming, the entity, “we”, will refer to our capstone team as a separate entity from Conviva. Furthermore, we consider Conviva to be a close partner to our capstone team on whom we can rely for continuous access to their dataset.

III. Products and Value

A vast and painfully prevalent gap exists between the amount of data being generated around the world and the global tech industry’s ability to utilize it. According to IBM, “every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone” (“Bringing Big Data”). While the already monumental quantity of data continues to grow, scientists and engineers alike are just beginning to tap into the power of this data. This is not to say that data does not already pervade nearly every imaginable aspect of life today; it does. Large amounts of data crunching and predictive analysis go on behind the scenes of numerous activities, from returning search queries, to recommending movies or restaurants, to predicting when and where the next earthquake will occur. However, there remains a massive body of questions and problems in both academia and industry that researchers have been unable to use data to answer. One domain in which better utilization of data could yield tremendous benefit is that of online media. Our team aims to serve this niche by building tools that address two critical challenges of online video data analysis: accurate real-time anomaly detection on large scale data and subscriber churn analysis.

Online video providers struggle to consistently serve a TV-like experience with high quality video free of buffering interruptions. Many factors within the "delivery ecosystem" affect the throughput of a video stream and, ultimately, the end user's viewing experience (Ganjam et.al 8). These factors include "multiple encoder formats and profiles, CDNs, ISPs, devices, and a plethora of streaming protocols and video players" (Ganjam et al. 8). An automatic anomaly detection and alert system is necessary in order to both inform a video content provider when their customers are experiencing low quality and, among the many possible factors, diagnose the primary cause of the problem. For example, if all customers experiencing frequent buffering belong to a certain ISP, then the alert system should flag that ISP as the root of the problem. The challenge that plagues many current solutions, however, is related to the aforementioned growth in the amount of collected data. While it is easy to detect when and why predictable measurements misbehave at small scale, it is hard to do so with high accuracy at large scale, across a range of system environments. To meet this challenge, our team has developed our Smart Anomaly Detection system to detect when and why truly anomalous and interesting behavior occurs in measured data. Such a system would greatly help content providers improve both their operational performance and efficiency. This value will be passed down to the viewers who benefit from higher service quality.

A second problem for subscription-based online video content providers is the ability to retain their subscribers. While the problem of diagnosing and eventually reducing subscriber churn has existed as long as the subscription service model, only recently has the tech industry developed the capacity and means to use big data to do so (Keaveny). Furthermore, largely due to the fact that online video hosting and distribution is a relatively new service, nearly all previous works in the area have focused on other domains such as telecommunications or television service subscriptions (Keaveny; Verbeke 2357-2358). Our team's Subscriber Analysis toolset aims to fulfill this unmet need by developing predictive models of viewer engagement and churn based on viewing activity and service quality data. Being able to predict churners and identify characteristic predictors from the data allows companies to focus on addressing the problems most critical to their viewers, thereby reducing churn rates. As proven by Zeithaml, there is a real, high cost associated with subscriber churn (Zeithaml). Thus by aiding in the reduction of churn, our Subscriber Analysis product can both help content providers increase revenues and result in higher overall satisfaction for those who purchase online video subscriptions.

For the reasons described above, our team is confident that our Smart Anomaly Detection and Subscriber Analysis products are important and valuable to both content providers and their customers, the viewers.

IV. Our Dataset

Conviva provided 4.5 months of session summary data from a single anonymous content provider for our research and development. 73,368,052 rows of session summaries are in this dataset. Each session summary represents a single instance of a viewer requesting a video object. In addition to service quality data, the type of device used by the viewer, the approximate location of the viewer, and metadata about the video content being accessed are collected into 45 columns. Subscription and demographic information about a viewer beyond their location are not available within this dataset. Fields that might otherwise help identify the anonymous content provider such as video content metadata were also anonymized by Conviva prior to data transfer to protect their customer.

Although the data was preformatted by Conviva before being transferred to our capstone team, we identified two important challenges implicitly encoded within this data through exploratory data analysis and follow-up communication with Conviva engineers. First, several of the fields in the session summaries are not as reliable as we initially believed. For example, fields such as `season` and `episodeName` are often empty. Second, our initial dataset included data generated by an artificial “viewer” that was used by Conviva for testing purposes and exhibited very strange, abnormal behavior. This was very important to keep in mind as we developed and evaluated our tools based on this data.

For our Smart Anomaly Detection product, Conviva informed us of the two most important metrics in assessing QoS that they wished to detect anomalies for. First is the number of attempts in watching online video over a time period. Low number of attempts indicates that users may be unable to access the content due to a datacenter failure. A high number of attempts signals the presence of a viral video. Second is the video start failure (VSF) rate. The VSF rate is the percentage of attempts that have failed to begin properly. VSFs may be caused by bugs in the video player software or by improper encoding/decoding of video content. Unlike attempts, low VSF rate is not a concern for video content providers. However, high VSF rate indicates major issues in the content delivery pipeline. To determine if an attempt has ended in VSF, we look at the `joinTimeMs` and `nrerrorsbeforejoin` columns in the data. The table below provides description about these 2 columns. An attempt ends in VSF if `joinTimeMs < 0 AND nrerrorsbeforejoin > 0`.

Column Name	**DataType**	**Description**
joinTimeMs	int	How long this attempt spent joined with the video stream. If this attempt has not yet joined, then this value will default to -1.
nrerrorsbeforejoin	int	How many fatal errors occurred before video join

For Subscriber Analysis, on the other hand, the nature of the problem is that we cannot know beforehand which fields within the session summary are useful in distinguishing viewers who are likely to churn. At the same time, as a consequence of the first of the challenges mentioned above, the Subscriber Analysis product should not indiscriminately use all fields of the session summary, including both reliable and unreliable fields. Thus, a central component of the work in Subscriber Analysis revolves around selecting a subset of these fields to use to form “features” to be used by the product.

V. Trends and Strategy

Having defined our team’s product and established both how they generate value and for whom they are valuable, we can focus on how we plan to bring these products to market from the standpoint of a new business. Amidst an era of rapid information and especially within the technology-abundant Silicon Valley, bringing such innovations to market requires understanding the market and having a well-formed competitive strategy. In this section, we describe the social and technological trends relevant to our product as well as the market and industry our business would be entering. We then describe the strategy we have developed that would allow our business to be successful in this competitive environment.

Why Us, Why Now

In the past five years, the number of broadband internet connections in the United States has grown from 124 million in 2009 to 306 million in 2014, leading to a compound annual growth rate of 19.8% per year (“Num. of Broadband Conns.”). This growth is indicative of the ever-growing role the Internet plays in daily life. Along with the growth of the Internet, as both a cause and effect, comes the spread of online services. In her article for Forbes, Erika Trautman, CEO of Rapt Media, states that “each year, more and more people are ditching cable and are opting for online services like Netflix and Hulu.”

The emergence of online video services has been so disruptive a shift in video distribution, that it incited a 2012 public hearing concerning public policies from the Senate Committee on Commerce, Science, and Transportation. In the hearing, leaders from technology juggernauts and state senators alike echoed the same viewpoint: online video services are the future of video distribution. Susan D. White, the Vice Chair for Nielsen, a leading global information and measurement company, reported that “the use of video on PCs continues to increase—up 80 percent in the last 4 years…Consumers are saying, unequivocally, that online video will continue to play an increasing role in their media choices” (U.S. Sen. Comm. on Commerce, Sci. & Trans. 9).

Of course, similar to other industries, a business seeking to enter today’s online video industry must meet a myriad of both business and engineering challenges. Unlike many of these industries, however, our industry is well-positioned to easily collect and analyze vast amounts of data to meet these challenges. Out of these conditions, the online video analytics (OVA) industry emerged, helping to translate and transform this data into useful insights that can be directly used by online video providers. A report by Frost & Sullivan summarized the rapid growth in the market:

> Still a largely nascent market, online video analytics (OVA) earned $174.7 million in revenue in 2013. It is projected to reach $472 million in 2020 as it observes a compound annual growth rate (CAGR) of 15.3%....The growth of OVA is largely attributed to the high demand for advanced analytics from online video consumption (Jasani).

Spurred by the massive opportunity in this market, our team has worked with Conviva to identify two of the most significant technical challenges faced by content providers: real-time detection of anomalies in a rapidly changing, unpredictable environment and efficiently reducing subscriber churn.

The challenge of retaining subscribers has existed as long as the subscription-based business model itself. As the competitive landscape of the online video market continues to evolve, the ability to diagnose and mitigate subscriber churn is a crucial component for business success. Sanford C. Bernstein estimated Netflix's average annual churn rate at 40-50%, which translates to 24-30 million subscribers (Gottfried). Reducing this churn rate by even a small fraction and keeping the business of these subscribers could mean significant increases in revenue. Just as critical for success is the ability to detect and respond to anomalies or important changes in metrics such as network usage and resource utilization. On July 24, 2007, 18 hours of Netflix downtime corresponded with a 7% plummet in the company's stock (Associated Press).

As previously described, our team provides solutions to these challenges through our Subscriber Analysis and Smart Anomaly Detection products. We believe that while these solutions, which use a combination of statistical and machine learning techniques, are powerful, our primary value and competitive advantage lies in our use of the unique dataset available to us through our partnership with Conviva. In the following sections, we discuss in detail how we plan to establish ourselves within the industry. In particular, we describe how we will position ourselves towards our buyers and suppliers as well as how we will respond to potential new entrants and existing competitors to the market.

Buyers and Suppliers

One of the most important components of a successful business strategy is a deep and accurate understanding the different players involved in the industry. In particular, an effective strategy must define the industry's buyers, to whom businesses sell their product, and its suppliers, from whom businesses purchase resources. In this section, we provide an overview of important entities related to our industry and present an analysis of our buyers and suppliers.

Potential customers for the global online video analytics market include content providers, who own video content, and service providers, who facilitate the sharing of user-generated video content (Jasani; Smith). Among the content providers are companies such as HBO, CCTV, and Disney, who all bring a variety of original video content to market every year. These businesses serve a huge user base and are able to accumulate large amounts of subscriber data. HBO alone was reported to have over 30 million users at the beginning of 2014 (Lawler). This abundance of data presents massive potential for improving these companies' product quality and, correspondingly, market share. Our Subscriber Analysis product can realize some of this potential by helping understand the experience and behavior of their users. Furthermore, with our Smart Anomaly Detection product, content providers can be made aware when significant changes occur in viewer behavior, system performance, or both. These tools can lead to a more valuable product, as seen from the content provider's viewers. While service providers such as Twitch or Vimeo differ from content providers in that they tend to offer free services, the success of these companies is still highly dependent on retaining a large number of active users. Thus, we target service providers in much the same way as we target content providers. Overall, we find that content and service providers, as buyers, are at an advantage in terms of business leverage over us, as sellers. This is primarily due to low switching costs, which arise from the fact that other businesses such as Akamai and Ooyala offer products for processing video data similar to ours (Roettgers). Because buyers ultimately make the choice choosing where to send their data on which both Smart Anomaly Detection and Subscriber Analysis depend, it can be difficult to deter customers from switching to our competitors. However, as we describe later in this paper, our unique approach towards churn analysis may differentiate us from our competitors and decrease buyer leverage over us.

On the other end of the supply chain, we also must consider who our suppliers will be and what type of business relationship we will have with them. Because our product exists exclusively as software, we require computing power and data storage capacity. Both of these can be obtained through the purchase of cloud services. Fortunately, the current trends indicate that cloud services are becoming commoditized, with many vendors such as Amazon, IBM, Google and Microsoft offering very similar products (Hanley). Though our buyers benefitted from low switching costs between us and our competitors, we face even lower switching costs between our suppliers. This is because while there is a considerable amount of effort involved with integrating a monitoring or analytical system with a new set of data, migrating the services between the machines which host them is almost trivial, involving only a transfer the data and minor machine configuration. In addition to cloud services, to a certain extent, we are dependent on device manufacturers and developers of video player libraries. We require them to provide an Application Program Interface (API) which we can use to gather online video analytics data from users. Fortunately, prior relationships with these device manufacturers and developers have been established through our partner Conviva. Conviva can help us open APIs for new devices and video players to maintain the flow of data required for our products.

As Porter argued, strategic positioning requires performing activities either differently or more efficiently than rivals (“Five Competitive Forces” 11). Our partnership with Conviva affords us a large quantity of high quality data for our algorithms to utilize, giving us a slight advantage compared to other services. In order to maintain and build upon this advantage, however, we must focus on developing our products to utilize this data and yield results in a superior manner. Thus, it is clear that our ability to differentiate from competing products and outperform them is key to our business strategy and the following sections describe how we can do so.

New Entrants

“Know yourself and know your enemy, and you will never be defeated” (Sun Tzu 18). This proverb can be applied to almost any competitive situation, from warfare to marketing. Interpreting this teaching in the context of business strategy, we identify that understanding the rivalry among existing and potential competitors is essential to a lasting competitive advantage. This interpretation fits well within the framework of Michael Porter’s five competitive forces. We now examine new entrants through the incumbent advantages and barriers to entry that work to keep this force as a low threat to both of our products. Porter recognized seven incumbent advantages (“Five

Competitive Forces" 4-6). The first is supply side economies of scale in which established incumbents have tremendous strength. The code behind a given analysis program is a fixed cost which scales well with an increased number of users, thus reducing the marginal cost of the code with each customer. The servers that receive and process the various users' data are linear, but scale with the number of customers acquired. The real advantage comes from the exponential power of the data supplied by these same customers, a theme we have come back to repeatedly in this paper. As the breadth and quantity of data increase with the combined user base of our customers, our algorithms become increasingly powerful and allow the incumbent product to outperform new entrants. This leads into our second advantage, demand side benefits of scale. As the authority in the field of providing content providers with analytics, incumbents can encourage customer demand by using their data on content quality improvements to provide hard evidence of the bottom line improvement new users can expect. "Increasingly powerful predictive analytics tools will unlock business insights [and drive revenue]" (Kahn 5). Demonstrating that our tools provide access to increases in revenue is key to nurturing demand.

Switching from an incumbent's service provides another barrier to entry, customer switching costs. While switching from one online service to another is not prohibitively expensive considering the benefits offered, the most impacting loss is in the past data the incumbent analysis provider's algorithms had of user's performance. "As we increase the training set size L we train on more and more patterns so the test error declines" (Cortes et al. 241). Via additional training examples, the incumbent's algorithm would consistently outperform the new entrant as the new entrant slowly acquires a pool of data comparable to that of the incumbent.

Just as it does not appear expensive for a customer to switch, it appears feasible for new entrant to join due to minimal physical capital requirements. With Platform as a Service (PaaS) providers, a new entrant merely needs a codified algorithm and a client or two to get started. Still, it is again the data that proves key to providing value to our customers. Importantly, new entrants cannot attain this data until they acquire clients, a classic catch-22 which serves as an inhibiting capital requirement for new entrants.

The global reach of our data partner, Conviva, provides both a size independent advantage as well as an unequal access to potential distribution channels in that it allows for direct international sales in the form of immediate integration of our tools with the systems of our partner's customers. The last relevant advantage as discussed by Porter, concerns restrictive government policy. Privacy concerns do arise when personal data is used, however there are standards for anonymization to be employed when using such data (Iyengar). While governments do allow the use of such data, it has to be acquired by legal means, which means a new entrant is restricted in its means of gathering new data for its algorithms. Thus, after a thorough analysis of the potential new entrants of our industry, the incumbents' advantages suggest that the threat of new entrants is a relatively weak force in our industry.

Existing Rivals

Another category of threats that a successful business strategy must address is that of existing rivals. As Porter described, the degree to which rivalry drives down an industry's profit potential depends firstly on the intensity with which companies compete and secondly on the basis on which they compete ("Five Competitive Forces" 10). We analyze these two parts for each of our products separately.

As machine learning grows in popularity, research into anomaly detection and other analyses of time series data is receiving greater attention both in academia and in industry. A survey of anomaly detection techniques shows a variety of techniques applied in a diverse range of domains (Chandola). Our strategy must take into account the threat of commercialization of technologies into industry competitors. For example, in 1994 Dipankar Dasgupta used a negative selection mechanism of the immune system to develop a "novelty" detection algorithm (Dasgupta). In addition to these potential competitors, there already exist several important industrial competitors working on anomaly detection. In January, 2015, Twitter open-sourced *AnomalyDetection*, a software package that automatically detects anomalies in big data in a practical and robust way (Kejariwal). Our Smart Anomaly Detection product is comparable to products from industry competitors such as Twitter; it is able to integrate with various sources of data, perform real-time processing, and incorporate smart thresholding with alerts. Although our competitors may try to research and develop a superior anomaly detection algorithm, we believe that our superior quantity and quality of data provided by Conviva gives us an edge over our competitors. Thus, we characterize competitive risk for Smart Anomaly Detection as weak. To a large extent, the competitors of Subscriber Analysis include the content providers themselves. Netflix spends $150 million on improving content recommendation each year, with the justification that improving recommendations and subscriber retention by even a small amount can lead to significant increases in revenue (Roettgers). These content providers have the advantage that they have complete access and control over the data they collect. If most companies were able to build an effective churn predictor in-house, the industry would be in trouble. However, we are confident that the quality of our Subscriber Analysis product will overwhelmingly convince content providers facing the classic “buy versus build” question, that building a product of similar quality would demand significantly more resources than simply purchasing from us (Cohn). This confidence is further supported by Porter in the context of the tradeoffs of strategic positioning (“What Is Strategy?” 4-11). In addition to content providers, there also exist competitors such as Akamai and Ooyala, who offer standalone analysis products to content and service providers. These competitors tend to focus on the monitoring and visualization of the data. In contrast, Subscriber Analysis focuses on performing the actual analysis to identify the characteristics and causes of subscriber churn.

Still, our most important advantage over these competitors remains our ability to perform in-depth churn analyses based on the abstraction of session summaries, which consist of a unique combination of metrics exclusively related to service quality. To the best of our knowledge, this is unique to previous and existing works in subscriber churn analyses. Our research has shown that the most prominent existing analysis approaches all incorporate a significant amount of information, often involving direct customer surveys or other self-reported data. Because service quality data is abundant and easy to obtain compared with demographic data, our Subscriber Analysis product can appear extremely appealing to potential customers. This easy to collect and consistent subset of video consumption data means our product has the potential to scale much better than existing approaches which require highly detailed, case-specific, and hard to obtain datasets. However, we cannot guarantee that this algorithmic advantage be sustained as our competitors continue their own research and development. Thus, we conclude that threat of competition to Subscriber Analysis is moderate.

Substitutes

The final element of our marketing strategy concerns the threat of new substitutes. Porter defined substitutes as products that serve the same purpose as the product in question but through different means (“Five Competitive Forces” 11). We first discuss potential substitutes for our Smart Anomaly Detection product.

The gold standard for most alert systems is human monitoring. Analogous to firms hiring security monitors to watch over buildings, video content providers can hire administrators to keep watch over network health. A more automated substitute is achieved through simple thresholding, in which hardcoded thresholds for metrics such as the rate of video failures trigger an alarm when exceeded. Content providers can also utilize third party network performance management software from leaders like CA, Inc. This type of software alerts IT departments of potential performance degradation within the companies' internal networks (CA Inc. 4). Similarly, content providers can pursue avenues besides Subscriber Analysis to reduce churn rates. Examples include utilizing feedback surveys and consulting expert market analysts. Feedback from unsubscribers is an extremely popular source of insight into why customers choose to leave and can go a long way in improving the product and reducing churn rate. These often take the form of questionnaires conducted on the company’s website or through email. In addition, content providers commonly devote many resources towards consulting individuals or even entire departments with the goal of identifying marketing approaches or market segments that generate lower churn rates.

Porter classified a substitute as a high threat when the substitute offers superior price/performance (“Five Competitive Forces” 12). With this in mind, we found that the overall threat of substitutes for Smart Anomaly Detection product is low. In contrast to human monitoring, our product offers a superior value proposition to our buyer. According to Ganjam et. al, many factors, including "multiple encoder formats and profiles, CDNs, ISPs, devices, and a plethora of streaming protocols and video players," affect the end user’s viewing experience (Ganjam 8). The complexity of this delivery ecosystem requires equally complex monitoring with filters to isolate a specific ISP, for example, and to determine if its behavior is anomalous. Such large scale monitoring does not scale efficiently when using just human monitoring. Similarly, simple thresholding poses little threat as a substitute because fine tuning proper thresholds over multiple data streams is difficult and time consuming. Many false positives and negatives still occur, despite such fine tuning (Numenta 11). Network performance management software, on the other hand, poses a considerable threat to us. However, while they are excellent at detecting problems within a content provider’s internal network, they alone cannot increase the quality of service. Xi Liu et al. argue that an optimal viewing experience requires a coordinated video control plane with a “global view of client and network conditions” (Liu 1). Fortunately, thanks to our partnership with Conviva, we have the data necessary to obtain this global view.

Just as with Smart Anomaly Detection, the threat of substitutes for Subscriber Analysis is also low. Although feedback surveys are direct and easy to implement, there are several inherent issues associated with them. Perhaps most prominently, any analysis that uses this data format must make a large number of assumptions in order to deal with uncontrollable factors such as non-response bias and self-report bias

(Keaveny). Expert opinion, whether gathered from a department with the company or through external consult, is the traditional and most common approach towards combating subscriber churn. This method, while very effective, tends to be extremely expensive. Still, as demonstrated by Mcgovern's Virgin Mobile case study, expert opinion can lead to identifying the right market segment, lower churn rates, and ultimately a successful business (McGovern 9).

To mitigate the threat of substitutes, Porter suggests offering "better value through new features or wider product accessibility" ("Five Competitive Forces" 16). For Smart Anomaly Detection, there are several avenues to pursue to provide a better value proposition to our buyers. For example, we can develop more accurate predictors with additional data from Conviva and explore new machine learning algorithms. For Subscriber Analysis, the threat of substitutes continues to be low because, unlike the examples given above, our product can perform effective analyses and generate valuable insights in an automated, efficient fashion. Data obtained through direct customer surveys, while potentially cheap, come bundled numerous disclaimers and can lead to a certain stigma from the subscriber's perspective. Furthermore, although data obtained through surveys, such as demographic information, might be more helpful in characterizing churners, by focusing on providing churn analysis based only on service quality data, our Subscriber Analysis product has at least one significant advantage. Service quality data from content consumers can be more easily gathered compared to data such as demographic information. Consequently, our product can be more appealing and accessible to content providers, especially those who do not have access to, or would like to avoid the cost of obtaining, personal data about their users. We also point out that both Subscriber Analysis and the substitutes such as those described above can be used in combination with each other. In such a case, our Subscriber Analysis product becomes even more appealing. This is because it can use the data from customer feedback to yield further improved performance. Our product would also make tasks such as identifying appropriate market segments much easier and cheaper to accomplish for content providers.

Strategy Summary

In summary, there are several social and technological trends which make now the right time for commercializing our Subscriber Analysis and Smart Anomaly Detection products. The most prominent among these are the rapid growth in internet connectivity and the spread of online services. In order to evaluate how well positioned we are to capitalize on the opportunity created by these trends, we developed a business strategy through competitive industry and market analysis from several different perspectives. From the perspective of buyers and suppliers, though we find that buyer power is significant, over time we expect to differentiate ourselves from our competitors by leveraging both the superior size of our dataset and our more efficient overall use of the data. We find that supplier power is low for our industry because the only significant resource we require is available through cloud services, an industry in which we have high buyer power and which is quickly becoming commoditized. From the perspective of rivals, the threat of new entrants is low due in large part to the superior quantity and quality of our data as well as the benefits of scale we would stand to benefit from as incumbents. Similarly, while existing competitors do present a threat, we find that our use of superior data and unique approach gives us a significant competitive advantage over them. Finally, we see a weak threat from the perspective of substitutes because we offer superior value at a cheaper price to our customers that only improves in combination with other techniques. Taken together, our evaluations lead us to believe that there is significant potential for a sustained competitive advantage over competitors, and that now is an opportune time to pursue it.

VI. Intellectual Property

Equally important to a team’s ability to build a valuable product and bring it to market is its ability to protect that value. In this section, we explain how we, as a business pursuing the strategy above to bring Subscriber Analysis and Smart Anomaly Detection to market, intend to sustain and protect the value of our work.

The traditional method for protecting the value of a new technology or innovation is obtaining a legal statement regarding ownership of intellectual property, IP, in the form of a patent. Indeed, patents have performed well enough to remain a primary mechanism for IP protection in the US for more than 200 years (Fisher). Unfortunately, when it comes to software, the rules and regulations regarding patents become dangerously ambiguous. The recent influx of lawsuits involving software patents has been attributed to the issuance of patents that are unclear, overly broad, or both (Bessen). Despite software patent laws being an active and controversial topic, these discussions have simply left more questions unanswered. The *Alice Corporation v. CLS Bank* Supreme Court case in 2013 is oft cited as the first source of information about software patentability, and even this case has been criticized for the court's vagueness (*Alice Corporation v. CLS Bank*). As noted by patent attorney and founder of IPWatchDog.com Gene Quinn, a definitive line should be drawn by the courts: a patent describing only an abstract idea, without specific implementation details, is invalid and cannot be acted upon (Quinn).

Thus, faced with the question of patentability, our team must examine the novelty of our Subscriber Analysis and Smart Anomaly Detection products. The goals of Subscriber Analysis and Smart Anomaly Detection are to diagnose the causes of subscriber churn and intelligently detect important changes in measured data respectively. Because these goals are rather broad, there exist a number of existing implementations, both old and new, with similar objectives. As a team considering patentability, we look towards the novelty of our specific approach and implementation. In the course of this introspection, we note that our implementation amalgamates open source machine learning libraries such as SciKit-Learn, published research from both industry and academia, programming tools such as those offered by Databricks, and finally the unique data afforded to us through our partnership with Conviva. With this in mind, we conclude that current patenting processes are flexible enough such that by defining our implementations at an extremely fine granularity, we would likely be able to obtain a patent on our software. However, we strongly believe that there exist several significant and compelling reasons against attempting to obtain a patent for our work. In this section, we elaborate on these reasons and describe an alternative method for protecting our IP which better suits our situation and business goals.

There is an abundance of existing anomaly detection patents of which we must be wary. Several of these patents are held by some of the largest companies in the technology sector, including Amazon and IBM. For example, *Detecting anomalies in Time Series Data*, owned by Amazon, states that it covers “The detected one anomaly, the assigned magnitude, and the correlated at least one external event are reported to a client device” (U.S. Patent 8,949,677). One patent owned by IBM, *Detecting anomalies in real-time in multiple time series data with automated thresholding*, states that in the submitted algorithm, a “comparison score” is calculated by comparing “the first series of [observed] normalized values” with “the second series of [predicted] normalized values” (U.S. Patent 8,924,333). In observance of these patents, we must be wary of litigation, especially when it concerns large technology companies. Recently, many companies in the tech industry, both small and large, have come under fire with a disproportionate number of patent infringement lawsuits (Byrd and Howard 8). Some optimists argue that most companies need not worry, because large technology companies are likely filing patents defensively. However, these companies are often the ones who play prosecutor in these patent infringement cases as well. For example, IBM, a holder of one of these anomaly detection patents, has a history of suing startups prior to their initial public offerings (Etherington). More recently, Twitter settled a patent infringement lawsuit with IBM by purchasing 900 of IBM’s patents (Etherington). In a calculated move by IBM, Twitter felt pressured to settle to protect their stock price in preparation for their IPO. Thus, we must be extremely careful in how we choose to protect our intellectual property. If this means filing a patent, then we must be prepared to use it defensively. This is likely to require a very large amount of financial resources. As we do not currently have these resources to spare and cannot guarantee that the protection offered would be long lasting or enforceable, we seek an alternative to patenting.

The goal of our Subscriber Analysis product is to predict the future subscription status of users based on past viewing behavior. Despite our research on existing patents, our team has been unable to find many patents which pose a legal threat to Subscriber Analysis. Most active patents on video analytics focus on video performance and forecast, such as Blue Kai Inc's *Real time audience forecasting* (US Patent App. 20120047005). In contrast, the patent field of quantization and prediction of subscriber behavior remains largely unexplored. Despite several commercial solutions on the market, there has not been a corresponding number of patents. Thus, Subscriber Analysis does not face the same level of risk of litigation compared to Smart Anomaly Detection. However, there are a handful of patents in other domains that we need to be wary of. *System and method for measuring television audience engagement*, owned by Rentrak corporation, describes a system that measures audience engagement based on the time he or she spends on the program (US Patent 8,904,419). In short, it constructs a viewership regression curve for different video content and measures the average viewing length. For a new video, the algorithm infers the level of viewer engagement based on the video content and the duration the viewer watched. While viewer engagement is a critical component for predicting behavior in Subscriber Analysis, we also incorporate additional data. These include viewing frequency, content type, and video quality. Under such circumstances, we do not see it as necessary to license patents such as the one above for two reasons. First, and perhaps most importantly, we apply churn analysis in the domain of online video, whereas most relevant patents apply to other older domains. Second, our algorithm incorporates a unique set of features corresponding to the data provided by Conviva.

The decision to pursue and rely on a patent in the software is an expensive one in both time and financial resources as well as a risky one due to the tumultuous software patent environment. As such, while we may pursue a patent, it will not be relied upon for our business model. As such, we have two additional IP strategies to investigate, open sourcing and copyrighting.

Open source software is software that can be freely used, changed, and shared (in modified and unmodified form) by anyone, subject to some moderation (Open Source Initiative). Open sourcing has become increasingly popular; both the total amount of open source code and the number of open source projects are growing at an exponential rate (Deshpande, Amit et al). For the purposes of our endeavor, it is not the novelty of our approach but our dataset and partner provided distribution network that distinguishes us. As the algorithms used are already publicly available, open sourcing our code does not cost us anything but provides us the shield of using open source software for our business and the badge having our code publically exposed and subject to peer review. Our business model would entail providing a value-added service company, dedicated to helping customers integrate their existing systems with our anomaly detection library. Through our partnership with Conviva, we have an established distribution network to our potential customers who we can offer immediate integration with Conviva's existing platform. This is a significant advantage as while open source is openly available to all users, they are primarily for experienced users. Users have to perform a significant amount of configuration before they begin using the code, which can pose quite a deterrent. While we will use the open source codebase as a foundation for our service, we will additionally provide full technical support in designing a customized solution that meets the customer's needs. By pivoting towards this direction, we add additional monetary value to the product that we can sell and bridge the technical gap for unexperienced users, relying on a SAAS implementation style for our business model instead of on a patent.

Copyright for software provides another IP Strategy option. While debate continues to surround software patents, copyrights are heavily applied in software. As expressed by Forbes's Tim Worstall, "there's no doubt that code is copyright anyway. It's a specific expression of an idea and so is copyright." There are several differences in the protection offered by copyrights compared to that of patents. While a patent may expose a very specific invention or process to the public and protect for 20 years, a copyright offers much broader protection while still providing the threat of lawsuit for enforcement. The copyright lasts 90 years past the death of the author and offers statutory damages (Copyright.gov). In addition, the scope of what it encompasses proves more relevant to our endeavor. “Multiple aspects of software can qualify for copyright protection: the source code, the compiled code, the visual layout, the documentation, possibly even the aggregation of menu commands” (Goldman). By protecting the numerous aspects of our project, copyright provides us adequate security. Besides the advantages of the protection offered, the process is affordable and efficient. Copyright is automatic as soon as a work is completed, though to file for statutory damages, one must formally register for a fee of less than $100 and an application turnaround time of under a year (Copyright.gov). In addition, even prior to completion of the work, we can preregister with a detailed explanation of the work in progress.

All IP strategies come with risks and copyright is no different. While pursuing a strategy of trade secrets would make our code more private, we would risk losing our protection should the secret be compromised. Also, as a general security principle in the computer science field, only the bare minimum should be relied upon to be kept secret to minimize risk of loss. However, completely publicizing our code for our copyright can be equally dangerous as the competition could copy our code with only slight rewrites. To remedy this, we can limit access to the raw code and only publish the required first and last 25 pages of code needed to attain a copyright on the entire work. In addition to this measure, it is our unique dataset that is the source of our code’s advantage over our competitors, and this is already protected by our partner, Conviva, in its aggregated form as a trade secret,

We believe that the novelty of our code and the application of our techniques to our unique dataset would allow us to obtain a software patent. However, while a patent may be most effective at reducing our risk of litigation, we look to alternatives due to the current complexity of filing a software patent and the immense amount of financial resources required to do so. Our research has led us to two very appealing alternatives: open sourcing and copyrighting. For the reasons stated above, we believe that while each of these alternatives have their own risks, their respective merits make them more appropriate for our use than patenting. Moving forward, we plan to employ open sourcing, as we expect that building a large, open community of support will encourage adoption and most benefit our products.

VII. Technical Contributions

Overview

Online video data analytics, as the name suggests, refers to analysing performance and content data generated by online media players of viewers from all over the world. According to Conviva, our partner who provides us with large amount of valuable video data, they gather data of over 4 billion streams per month over 1.6 billion unique devices from all its customers (Conviva Inc.). Data of such volume has great potential to reveal unordinary behaviors over the content delivery network and the users, which deserve the attention of content providers. The main purpose of our capstone project is to generate useful insights by applying some machine learning techniques to the large dataset and to provide a possible solution to the problems we discovered, including how to accurately find anomalies in real time and how to predict users' engagement and subscription status. Our capstone project is divided into two subprojects, Smart Anomaly Detection and Subscriber Analysis, and in order to increase overall productivity, our team is also divided into two subteams, accordingly. The Smart Anomaly Detection team has been working on building toolsets to discover anomalies from streaming data that are meaningful for content providers and the Subscriber Analysis team, which I'm in, has been focusing on analysing users' past behavior data to figure out the reasons for churn or unsubscription.

Within the Subscriber Analysis subteam, Jefferson and I worked together towards the same goal while focusing on different aspects, so that we can get more comprehensive and conclusive results. To be specific, I worked on some exploratory data analyses, related-work research, feature selection experiments, implementation data-specified clustering and classification algorithms. In the following parts, I'll describe these tasks I've been working on in detail and how these tasks contribute to the goal of our subproject.

Related works

The main focus of our subproject is the history of users' behavior and the churn management. On the one hand, there exists quite little work about user behaviors in the context of online videos. A user behavior model was built using the probabilities and the distributions of different features observed in the dataset, making simulating user behavior possible (Vilas). This model can help understand and simulate each single video session but it's not yet able to provide customized prediction for each user based on their viewing history and lacks in-depth exploration of features. Afterwards, the same research group further extends their research mentioned above by including analysing the characterizations of server workload, from which content providers could benefit a lot. The type of data they possess resembles ours but as we share different visions of the end product, we took different paths and our project focused mainly on the subscriber side.

On the other hand, the study of subscriber churn management has a much longer history and more people's attention. Subscriber churn has always been an important issue in service industry and it's vital for the development of companies. The first model of customer switching in service industries was presented in 1995 (Keaveney) and was improved later on in 2001 (Keaveney and Parthasarathy). They analysed the factors that might be effective in discriminating between switchers and continuers of online services and this was treated as a marketing problem when there wasn't so much data. Later on, as data mining became popular and stronger computation resources became available, techniques other than empirical findings are applied to churn management in the Telecom industry, such as Decision Tree and back propagation Neural Network (Hung). Hung et al started their experiment by developing possible variables from other research and interviews with telecom experts, like customer demography, bill and payment analysis, customer care/service analysis, and then segment the customers based on empirical findings of these features. Such segmentation helped improve the performance. Some other techniques were also implemented and experimented, including Hybrid Neural Networks (Tsai 2009), Support Vector Machine (Coussement) and Logistic Regression (Burez). However, an issue with churn management is that customer churn is often a rare event thus the classes in the training set are imbalanced (Burez). The majority of real-world customers are non-churners, resulting in unbalanced weighting for those who are actually leaving. There have been numerous works proposing possible solutions to such a problem in the context of churn prediction. Algorithms like Weighted Random Forest, Sampling, Gradient Stochastic Boosting and Logistic Regression were implemented and compared on various datasets using different evaluation metrics (Burez, Chen). The experiment results show that Weighted Random Forest and Sampling are improving the accuracy the minority class and have better performance than most of existing algorithms (Chen), but Logistic Regression can sometimes outperforms others thus always need to be considered (Burez). However, we weren't able to find any existing research which addresses customer churn using only data about usage type and quality, but instead, they mainly user demographic information, interactions between subscribers and service providers and subscription types. They obtained data from either direct survey (Keaveney) or companies' databases, which greatly differ from the data we have access to. For example, in Keaveney's paper (Keaveney, 2001), he experimented with different features, including external influence, interpersonal influence, experiential influence, usage frequency, usage intensity, etc. and he found that customers' prior experience, satisfaction with information provided, as well as demographic data like income and education level are particularly useful when identifying switchers and continuers, while what we have is user usage and performance information and a very important part of our project is to infer and construct the definitions of critical subscriber information such as viewer engagement, user churn or subscription status from our limited data set.

As shown above, most of previous works are only focusing on one aspect of our project and have access to data with great detail. For our project with specific dataset, we used some statistical analyses to infer critical information from the data and then integrated our ideas with a number of the existing techniques. In the next part, I'll describe my work about the inference and integration, along with my other tasks.

Methods, Materials and Results

My individual technical contributions within the Subscriber Analysis team include tool familiarization and exploratory data analysis, user information inference, implementation and analyses of a K-Means clustering algorithm, and implementation and analyses of Logistic Regression classification algorithm. In this section I'll describe the methods, experiments and results of these tasks.

To begin with, we'll list the feature that we used for the following experiments. The features, which were extracted by Jefferson Lai, are shown in Table 1 and for more detailed descriptions, please refer to Jefferson's technical contribution (2015). The indices in the table would be used in some graphs to represent corresponding features.

Feature Index	Feature Name	Feature Description	Hypothesis/Justification
0	`nsessions`	Number of sessions initiated.	The number of sessions reflects how often a viewer watches
1	`totalplayms`	Total playing time.	Engagement Metric. Included for autoregressive purposes
2	`avgplayms`	Average playing time per session.	Indirect measure of content type and quality of service
3	`avgpausedms`	Average paused time per session.	Pause time could represent content type or viewing quality
4	`avgjoinms`	Average join time per session.	How long viewer has to wait for video start affects viewer experience
5	`avgstopms`	Average stop time per session.	Stop time could represent content type or viewing quality

6	`avgsleepms`	Average sleep time per session.	Sleep time could represent content type or viewing quality
7	`ndevices`	Number of unique devices.	More engaged viewers may use multiple devices
8	`nconntypes`	Number of unique connection types.	More connection types could mean more engagement
9	`ncountries`	Number of unique countries.	Could show interesting viewer behavior and/or signal loyalty
10	`flives`	Fraction of livestream sessions.	Distribution of content types correlates with tendency to churn
11	`fvsfs`	Fraction of video start failures.	Proportion of videos which failed to start reflects service quality
12	`febvs`	Fraction of exits before video starts.	Proportion of videos closed by the viewer before the video starts reflects service quality
13	`fjoined`	Fraction of successful video starts.	Proportion of videos which started successfully reflects service quality

Table 1. Feature used in our experiments

Tool familiarization and Exploratory Data Analysis

The first task of this initial stage is to familiarize myself with the tools and dataset that would be used throughout the whole project. The computation tool we are using to analyse data is Apache Spark, a fast and general engine for large-scale data processing (Spark), which was originally developed by graduate students and professors from UC Berkeley (Zaharia). Spark introduces an abstraction called resilient distributed datasets (i.e., RDD), which is a collection of objects partitioned across different machines that enables Spark to perform efficient in-memory computation for iterative and interactive algorithms. Databricks Cloud is the primary product of another major partner of ours, Databricks. Databricks Cloud was developed based on Spark and aims at making big data easier by providing a zero-management cloud platform with interactive workspace for exploration and visualization. By collaborating with Databricks, we got access to the powerful computation platform without worrying about the configuration of worker machines. Spark is one of the most popular state-of-the-art techniques for big data analyses and it was a great opportunity for us to have access to it beta version of such tools. To get started with Spark, we followed the intro videos posted by Databricks and I was able to attend one of the Spark workshops hosted at the headquarter of Databricks and discussed the issues we encountered during the learning process with the engineers.

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques, to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set (NIST). Such techniques were also used in previous works (Hung). The data that we have access to contains millions of session summaries of every video that every user watched within a certain period of time, fields of a summary include `buffering time`, `error count`, `location`, `type of device`, `session length`, etc. From the initial EDA, we found some issues with the dataset, for example, a field named ‘`Content Type`’ only has two possible values, while we expect this field contains richer information so we can actually use it as a feature. We also found that there were a lot of missing weeks of users, i.e. users were inactive in certain weeks and such records wouldn’t show up in any aggregation. We need to found these entries and add them manually, which could help avoid some unnecessary problems during actual implementation. By conducting EDA, we gained these insights about how to perform analysis on this dataset and discovered some ‘traps’ that should be avoided to make our result more valid.

User Information Inference

As we mentioned above, the data we have only contains usage and performance information but if we want to predict a user’s subscription status, we first need to give the user a label of whether he/she has left. For example, if we train on the first 10 weeks of data, we will label the churners’ status only based on these data and do the same thing for the last 10 weeks while testing. In addition to subscription status, we also need to decide how to define a user’s engagement.

In order to measure a user's engagement and integrate that measurement into the labelling algorithm of subscription status, we first calculated the playing time of each user in each week on record. Using `playing time` as the primary measurement is reasonable since this is the more direct indication of a user engaging in the service, while other metrics might have some bias. For instance, the count of sessions within in each week only takes the count into consideration but not the length of each session, which is unfair for those who watch several long movies and has bias towards those who glimpse through a lot of video but only stay for a very short time. After the calculation, we classified users into different groups based on their playing time and gave each of them a label of their engagement level, including high, medium, low and none. To infer the subscription status, we tried to find that if a user has dropped more than one level and kept low engagement for sometime, what is the possibility that the level may ever go up. To explain the measurement better, we can look at the example graph in Figure 1.

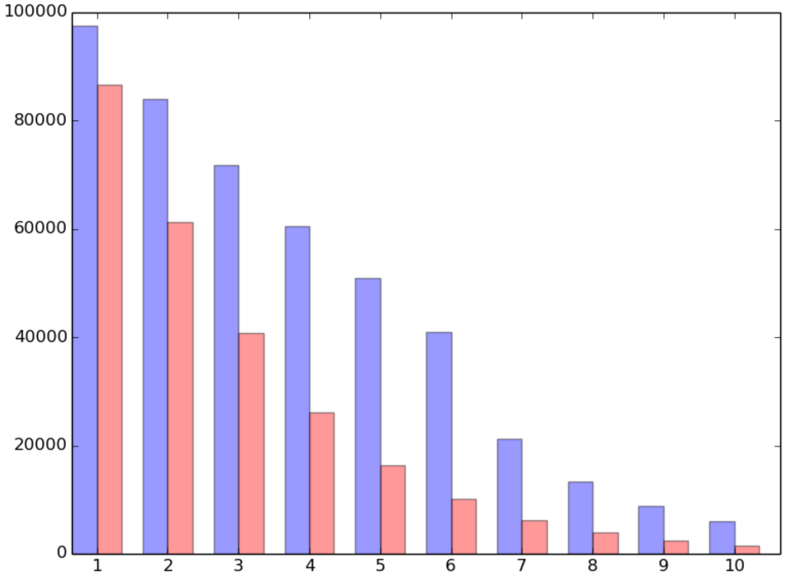

Figure 1. User number vs weeks

In this graph, the blue bars indicate the number of users whose engagement level dropped for x weeks, and the red bars indicate the number of users whose engagement level dropped for x weeks but an increasing engagement level of the user was observed after the x weeks in the dataset. To define churners, we are supposed to find the smallest x, so that for any other length of inactive weeks that's larger than x, the possibility of a user ever coming back would be almost the same. The x we got in the example graph is 5 since for all $x \geq 5$, the possibility of the engagement level going up remains almost the same, which means at the end of our dataset, if a user has been inactive for 5 continuous weeks, he will be labelled as a 'churner'. After such labelling, we have around 15% percent of all users were labelled as a 'churner' and the possibility that they might come back is around 30%.

This model was later further modified as described in Jefferson Lai's paper, to better work with our experiments and fit the goals of our subteam.

K-Means Clustering Algorithm

Materials and Method

After the first steps and with the features Jefferson extracted from the full dataset, I started my next task, implementing a clustering algorithm. The motive for clustering is that after exploratory data analysis and implementing the Linear Regression prototype, the team found the complexity of selecting the best features and got confused what features we should use and whether the dataset we have is enough for predicting customer churn and even predicting a user's engagement. Through clustering, we can group people who share similar features and analyse some statistics within each cluster. By analysing the quality of the clustering, we can gain deeper insight of our dataset.

I chose the most common clustering algorithm, K-Means, which was first used by James MacQueen in 1967 (MacQueen). K-Means clustering algorithm aims at partitioning n observations into k clusters. The first step of this algorithm is to choose k points (randomly, but none of these should be too close) as our initial cluster centers. The second step is to take each point of the given dataset and associate it to its nearest centroid. When there are no pending points, we continue to the third step, recalculating the centroids for each cluster that we get from the second step by using the means of all points within that cluster. After calculating the k new centroids, we calculate the new binding of each point and its nearest new centroid and start a new iteration from the second step. The centroids change after each iteration and the algorithm stops when no more centroids continue moving.

Initially, the algorithm aims to minimize a squared error function, which is

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} ||x_j^{(i)} - c_i||$$

where $||x_j^{(i)} - c_i||$ is any chosen distance measure between a data point $x_j^{(i)}$ and the cluster centroid c_i, is an indicator of the quality of clustering, measuring the similarity of points within a certain cluster. Instead of directly computing and minimizing the error function, in each iteration of the algorithm, we change the cluster center based on the grouping of data points, so that the distance between each point and its corresponding center can be shortened thus minimizing the error function.

In addition to this objective function, there are also several external evaluation metrics for clustering, such as silhouette score (Rousseeuw), purity, normalized mutual information, rank index and F measure (Manning). For our project, we want to know about what features can help distinguish churners from others and also the quality of the clustering. To learn about the ratio of churners within each cluster, i.e. its ability to separate churners from other, we chose purity as our secondary metric and to measure the tightness of each cluster, we chose silhouette score as the third.

To compute purity, each cluster is assigned to the class that is the most frequent in the cluster and the accuracy of such assignment is measured by counting the number of correctly assigned documents and dividing by n, which the number of points in this cluster(Manning). To observe the ratio of churners in a clearer way, we used that ratio as the metric, instead of using the majority. By using the new purity, we can find out if any of the clusters has a high percentage of churners and if they share many similarities with respect to the features we chose.

As for silhouette score, it has become an important metric for analysing and evaluating clustering, and it can visualize how well each data point lies within its cluster and the average of all points' silhouette scores within one cluster tells about the tightness of the grouping of these points. A simplified silhouette is

$$S(i) = \frac{b(i)-a(i)}{max\{a(i), b(i)\}}$$

where $a(i)$ is the distance between the ith point and its cluster center, and $b(i)$ is the distance between the ith point and its second closest cluster center. This is a simplified version of silhouette score, since for the original definition, we need to calculate n^2 distances for each iteration, where $n = 10^6$ in our dataset. We are testing with different cluster centers and feature sets, thus such amount of computation would be too expensive and such approximation would allow us to explore more parameters and perform more experiments. In the experiments, I calculated the silhouette score of each data point and take the average of all points within each cluster as the score for the cluster. We use these scores to measure the quality of the clustering.

Before actually implementing the main parts of the algorithms, we still need to consider some details. The most important issue is feature selection, and by following one of the most common feature searching approaches, greedy forward feature selection, we are able to compare the performance between different feature combinations. During the feature selection process, we use a greedy approach, which adds the local optimum at each step to the best feature set from last step, where optimum means the feature set gives the highest performance which will be described later, and uses the fulfilled feature set as the final feature set. At the beginning, we set the selected feature set as φ. At each step, we add each feature temporarily to the feature set and evaluate the result. Here, we define the best performance as the clustering gives the highest or the lowest purity while its average silhouette score is above a certain threshold, which means it can give a good separation of churners and non-churners and have a high quality of clustering at the same time. We choose the feature that gives the feature set the bestes performance and add it permanently to the selected feature set. Iterate the steps until no feature can increase the performance of the algorithm.

Another issue is data normalization. Since each feature has different ranges of values and it would create great bias if we don't normalize the features. Features with larger range include total playing time, total session time, total session count. These features should be scaled to 0 and 1 as others do and actually we care more about the users with less playing time as the majority of churners tend to be in this range. A simple solution is to shrink those with extremely high playing time to a reasonable range and then take a logarithm to amplify the lower part, which can be further optimized by analysing the distribution of each feature.

By implementing this clustering algorithm, we're hoping to find out if there would be a good clustering of the data with these certain features and use the clusters with a high/low purity to build a classifier which is pretty similar to random forest, that is we cluster all users using different feature sets and different numbers of clusters and select the top n clustering spaces with the highest/lowest purity as our classifier. Each clustering space contains several cluster centers and some of these centers contain the label of good clustering. For each user, we calculate the aggregated data from his viewing history and use each clustering space to place the aggregated data into a cluster and treat the purity of this cluster as the possibility of this user being a churner. All spaces vote to classify the user.

Implementation and Results

In actual implementation, I used the feature set in Table 1 and iterate over all features with different number of clusters. For a given feature set and number of clusters, I took the mean of each feature over all weeks, which represents the overall engagement of the user and quality of service he got, and calculated the squared error over the all clusters to represent the error of the clustering. For each iteration, we chose the feature resulting in the highest maximum or lowest minimum purity and silhouette scores higher than a threshold (which was 0.5 in the experiments) and added it to the selected feature set. We measure the quality of a clustering by considering all metrics, while mainly focusing on purity and silhouette score since the squared error is highly dependent on the distribution of the data.

First, we'd like to show the distribution of all purities we get from all the experiments we conducted, including those with different cluster numbers and different feature sets during the feature selection process, which was not as expected. This histogram of purity in Figure 2 shows that the highest purity is around 30% and the lowest is around 2.5%. The highest isn't high enough to distinguish churners and according to the raw data, there are always quite few points in the clusters with low purities, and these clusters aren't enough for our classification purpose. If we use a voting mechanism which is similar to random forest, all clusters will vote 'non-churner' since all churners seem to be evenly distributed to all clusters and clustering doesn't help separate churners from other users. The histogram shows the overall performance of all my experiments for clustering with regard to purity, before we dig into the details of the clusters.

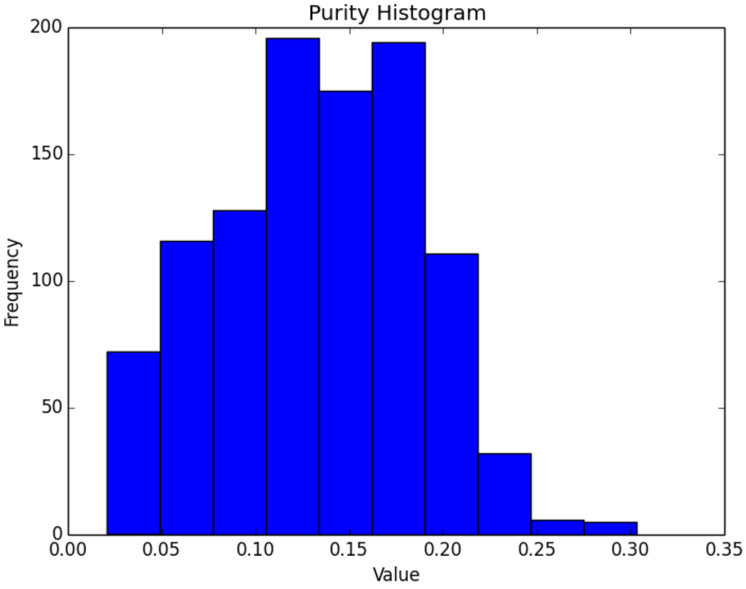

Figure 2. Histogram of All Purities

To verify the assumption, we looked into the details of one iteration of the clustering algorithm. We plotted most of the results from one run in Figure 3, with only feature 1 (`total playing time`) and 11 clusters, since this one gives the highest purity among all.

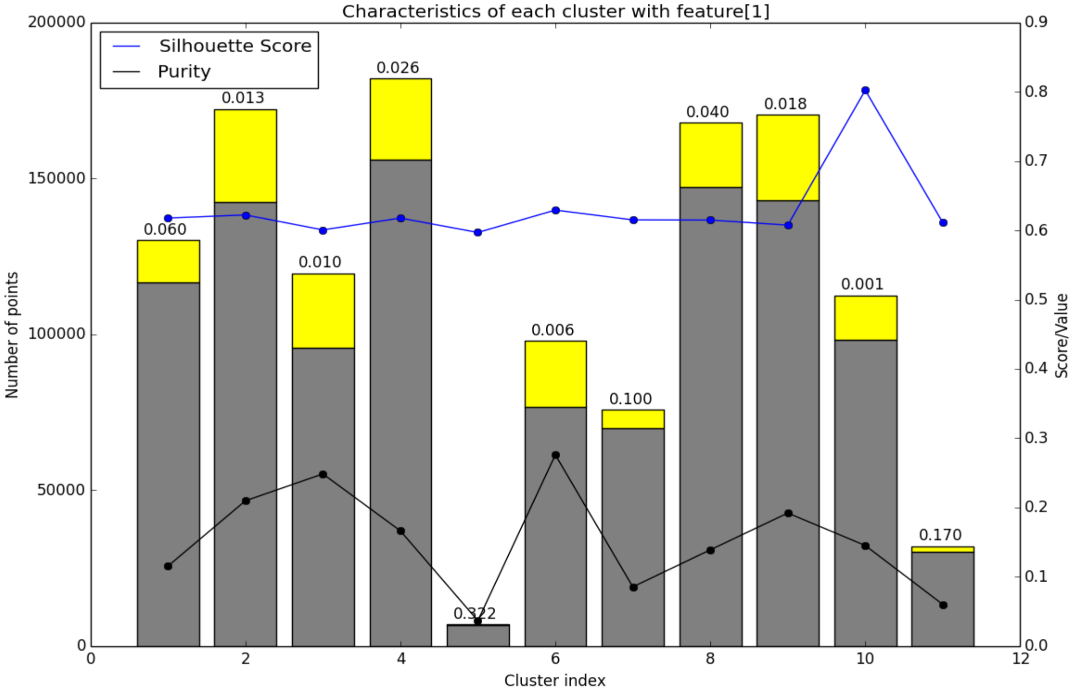

Figure 3. Characteristics of each cluster with total playing time (feature 1)

The yellow bars represent the churners and grey bars represent non-churners within this cluster. The number on top of the bars is the corresponding cluster center. The blue line connects the different silhouette scores of each cluster and the black line connects the different purity values of each cluster. Even though this clustering result contains the highest purity value, most clusters have approximately the same purity, which means data points are just evenly distributed in these groups and this feature doesn't give us a good separation of the churners. On the other hand, if we take a look at the silhouette scores, all clusters have a score higher than 0.55, which means the distance between one point and its second closest cluster center is more than twice that between the point and its closest center. We can say this is an good clustering, so the scores can evaluate the quality of clustering regarding its tightness. By observing other iterations, we can find that for some features, we can get very good clustering (with a silhouette score of 0.9998) but the purities remain low. We can conclude that the purity of each cluster isn't influenced by the quality of clustering and even though we might have a pretty good clustering, the purities are still too low to use.

To dig further, we also looked into the characteristics when using different sizes of feature sets. We want to compare the evaluation metrics among these feature sets, including error, average and maximum purity as well as average silhouette score.

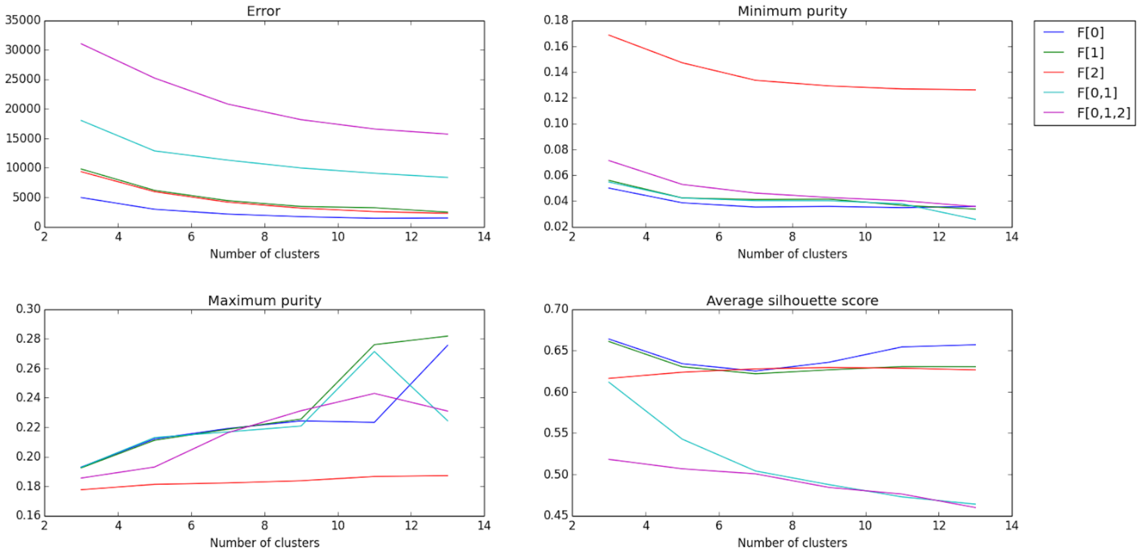

Figure 4. Metrics comparison when using different number of features

The results are shown in Figure 4, where we examined the performance for different features set, including `number of sessions` (Feature 0), `total playing time` (Feature 1), `average playing time` (Feature 2) and their combinations (Feature 0, 1 and Feature 0, 1, 2) as shown in the legend. The performance is determined by four metrics, error, minimum purity, maximum purity and average silhouette score. According to the result, If we try to use more than one feature, there won't be any improvement on these metrics. Since the dimension increases, error keeps increasing. The extra features make minimum purities comparable to single features while maximum purities outperformed by single features. Lower silhouette scores indicate a loosened clustering. However, when we continue the experiments and try to increase the number of clusters, the purities are becoming comparable with single features and the silhouette scores keep dropping. The results are reasonable, since more dimensions are making the clustering more complicated and more difficult to converge thus increasing errors and decreasing tightness of the clusters, and in the meantime, more features cannot change the inability of the dataset to separate churners.

By looking into the details of the clustering results, we can conclude that our purpose of clustering, i.e. separating churners and using the results to build a classifier, is not achievable by only using our current data. We can have a pretty good clustering, but the purity remains low even when we increase the number of clusters and the number of features.

Logistic Regression

Materials and Method

After Jefferson completed the k-NN regression and I completed the K-Means clustering, we decided to use different methods mentioned in the research papers we read to explore more options that could help improve our results. My last task was to implement logistic regression on our dataset. As mentioned in several research papers, logistic regression has been a common technique used in subscriber analysis and it sometimes outperforms other more sophisticated methods like artificial neural network (Burez). Comparing with neural network, logistic regression is a much simpler classifier and it is analogous to linear regression and the mechanism of this classification method is easy to comprehend. Logistic regression measures the relationship between a categorical dependent variable and some independent variables by estimating the probability of an event happening.

Suppose Y represents whether a user is a churner or not and $Y = 1$ represents that the user is a churner. Let p be the possibility of a user being a churner, thus the possibility of him being a non-churner is $1-p$. We use the definition of the odds of a user being a churner as $\frac{p}{1-p}$, i.e. the possibility of being a churner over the possibility of not being one. It's difficult to model a variable with restricted range, such as possibility. Thus in logistic regression, we model the logit-transformed probability as a linear relationship with predictor variables (UCLA: Statistical Consulting Group), which is

$$\text{logit}(p) = \log\frac{p}{1-p} = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n = \beta x$$

where x is the feature vector of each user and β is a matrix containing the parameters we are trying to learn from the training data. Transforming the equation to a representation of the possibility, we can get

$$p = \frac{1}{1+\exp(-\beta x)}$$

What we want to predict from the knowledge of user viewing history is not a precise numerical value of whether a user is a churner or not, but rather the probability he is a 'churner' rather than 'non-churner' (Educational & Professional Studies). Since Y can only be either 1 or 0, when given a user viewing history x, we can calculate the possibility of $Y = 1$, i.e. the possibility of the user being a churner. If this possibility is larger than 0.5, we'll predict the user being a churner, otherwise predict as a non-churner.

We wanted to do more with clustering, but we had to stop due to the unsatisfying results. By applying this classification algorithm to our dataset, we are trying to explore if classification can actually work well on such dataset and continue what we have left in clustering. In the meantime, we can further validate our assumption about whether the existing data is enough for us to predict customer churn.

Implementation and Results

In the implementation of Logistic Regression, to explore the different influences of different features for such a classification problem, we are also using the greedy forward feature selection. For each feature set we select, we extract the selected features from each week of users' viewing history and concatenate these features in chronological order. For example, when the feature set is [0, 1], we extract <feature0_week0, feature1_week0>, …, <feature0_weekn, feature1_weekn> and put all of them together as a vector, <feature0_week0, feature1_week0, …, feature0_weekn, feature1_weekn>. We use this vector to represent a user's history in this feature space. Our dataset is separated into two parts, the data of the first 10 weeks is for training and the last 10 weeks is for testing. Our labelling of churners are only based on users' viewing history of corresponding weeks.

As we have mentioned in previous sections, many users have inactive weeks in the dataset, so at this point, we only use users with a full history over the training set for training and those over the testing set for testing. Figure 5 shows the error rates for corresponding feature set when using only one feature. In our test set, we have a churner rate of 8%, after filtering out those without a full history.

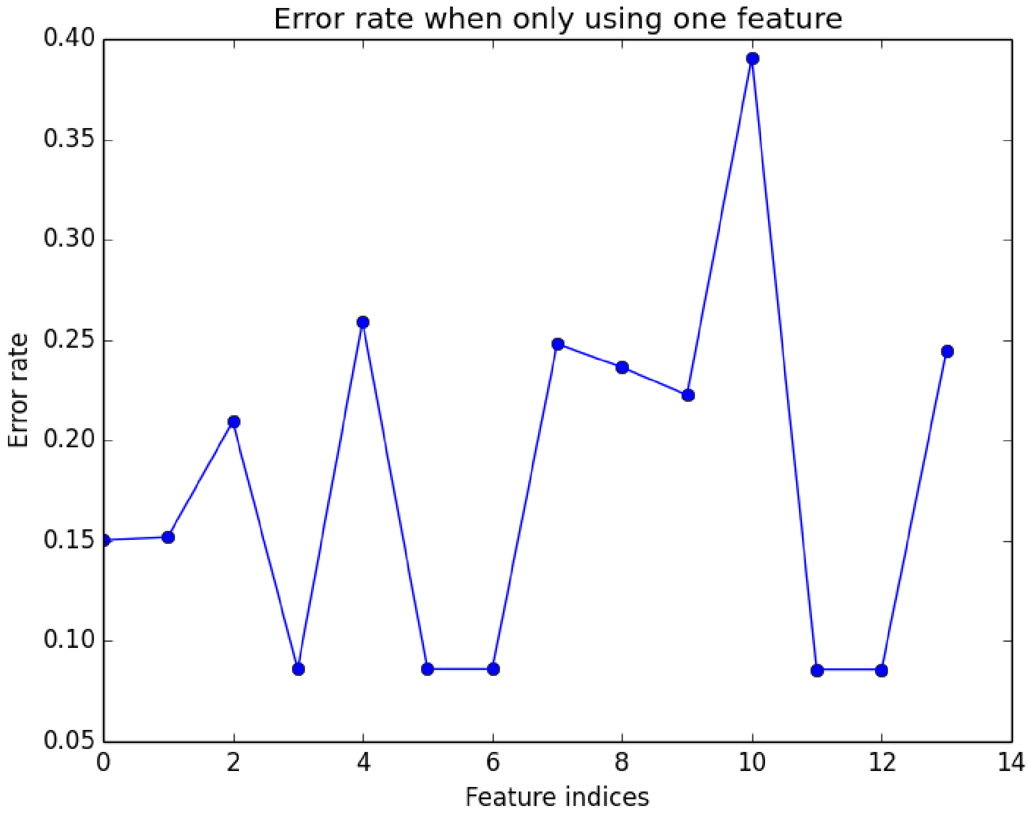

Figure 5. Error rate when only using one feature

From this graph, we can tell that different features have quite different error rates, and the lowest error rate that the features achieved is around 8%, which is equal to the ratio of churners in the test set. After digging deeper into the results, we found that when using these features, the classifier simply predicts every user as a non-churner and thus has a error rate that's equal to the churner ratio. Such results indicate these features, including `average paused time`, `average stop time`, `average sleep time`, `fraction of video start failures` and `fraction of exits before video starts`, are highly uninformative and the number of positive cases aren't enough for the predicting process. To confirm the assumptions, we looked into our data and found that the average of feature `average sleep time` over all users is zero, and the average of other features are also quite small, thus providing very little information.

Except for these features, we can see that other features have worse performance than predicting all users as churners, and `number of sessions` (Feature 0) and `total playing time` (Feature 1) have the highest accuracy. But in most cases, the classifier classifies most users to 'non-churners' while also misclassifies some 'non-churners' to 'churners', making the results even worse than classifying all users to 'non-churners'.

However, we only used users with a full history during the experiments, which rarely happens in actual world, and to see if using all users' data would help improve the performance, we need to deal with the data for the missing weeks. One naive solution is to filling in zeros for those missing points with a key of <user_id, week> and use the stuffed dataset for training and testing. While this works for some features, features like fraction of successful video starts should also consider factors like geographical information, connection types and number of devices. This would require a more sophisticated method.

In order to better measure the quality of classification and avoid the fake low error rate, we added three more metrics, precision, recall and F1 Score. The definition of these metrics are as follows.

$$precision = \frac{True\ positive}{True\ positive + False\ positive}$$

$$recall = \frac{True\ positive}{True\ positive + False\ Negative}$$

$$F1 = 2 \times \frac{precision \cdot recall}{precision + recall}$$

In our specific problem, precision is the ratio of correctly predicted churners to the total number of predicted churners and recall is the ratio of correctly predicted churners to the total number of churners in our dataset. By measuring these metrics, we can have a much clearer idea of the performance of our classification. First, we only train the model with one feature and the results are shown in Figure 6.

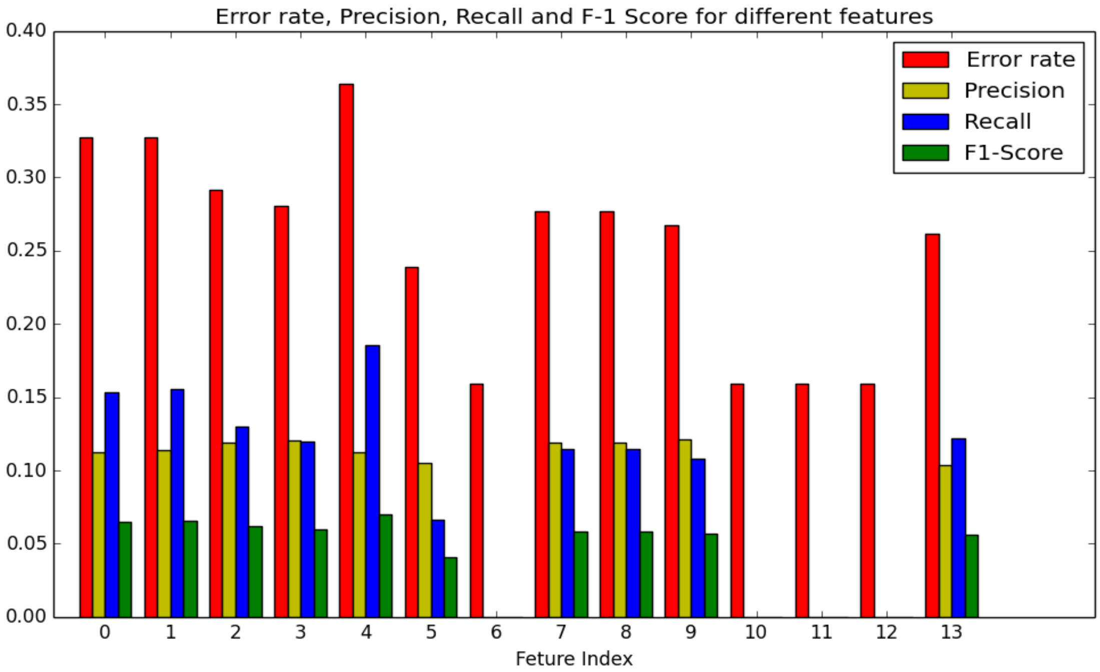

Figure 6. Error rate, precision, recall and F1-Score when using one feature with stuffed data

As shown in the graph, features that had an error rate that was equal to the churner ratio still have the same problem, as their recall is always 0 and precision is always infinity (it's difficult to represent infinity in the graph so I just used zero). As for other features, `average stop time` (Feature 5) has the lowest error rate as well as the lowest recall, which means this feature isn't good at finding churners, despite its low error rate. In the meantime, `average join time` (Feature 4) has the highest recall and highest F1-score. In our subscriber analysis, we care about recall more than error rate or precision, since recall measures the ability to find churners from all the users and a lower precision merely means we choose too many false positives, which isn't a large cost in our case. However, in general, most features have pretty high error rate and low recall and precision.

In order to figure out if adding more features for training can actually improve the performance, we added more features for `average join time` (Feature 4), since it had the highest recall and F1-score in our last experiment. The results for using different feature sets are shown in Figure 7. We can see that after adding to `average join time` (Feature 4), most combinations have a much better performance than the single feature regarding recall. But the best performance is just comparable with that of average join time alone and they also have a pretty high error rate.

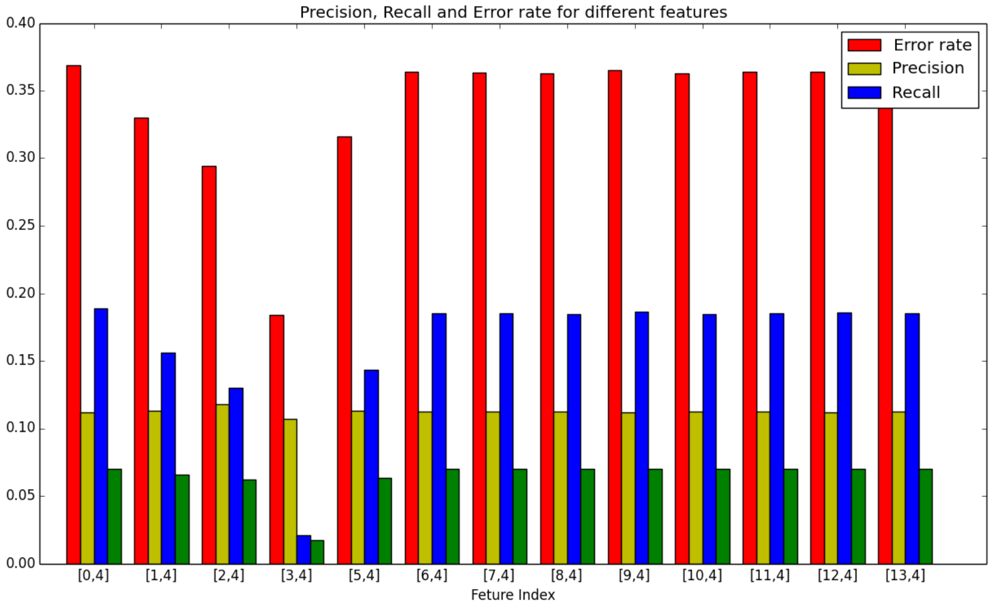

Figure 7. Error rate, precision, recall and F1-Score when using two features with stuffed data

The experiments with stuffed data achieved the highest recall of around 18% and the highest precision of around 12%, indicating `average join time` (Feature 4) plays a pretty important role in identifying churners and adding more features doesn't help improve the performance. We also used up to 4 features at the same time during the feature selection process, but there isn't any improvement in performance.

For logistic regression, we didn’t get any positive results, neither on the filtered data set nor on the stuffed data set and the classifier can identify at most 18% of all churners, which is pretty low and adding more features doesn’t help improve the classification performance.

Technical Summaries

Besides familiarization with the tools and the data and preprocess of the dataset, my work also includes clustering and classifying users. The first is building a clustering model and further using clustering results to build a classifier. We use different evaluation metrics to measure the quality of clustering and try to find clusters with highest or lowest purities values. Our initial purpose is to use these clusters to build a classifier using a similar mechanism to random forest, but our unsatisfying results showed the infeasibility of such methods. Most clusters in our results have purities within range (0.05, 0.25), and some of them have a high of around 0.3 while their high silhouette scores indicate that the quality of clustering is pretty good. We can conclude that for our dataset, we can have a good clustering but this clustering cannot separate churners from non-churners, thus stopping us from continuing with our initial plan. The second is using logistic regression to classify the users. We trained the model with different features sets. At the beginning, we only used users with full histories. The performance of the algorithm seems pretty good, but the problem is that some feature sets classify all users as non-churners and others have even worse performance than predicting everyone as a non-churner. However, this isn’t a real world case and we fulfill zeros to the data with missing weeks and use the new data set as the training set. Performance of this experiment is still pretty bad and the best the classifier can do is to identify 18% of all churners. Even though Jefferson’s results for k-Nearest Neighbor classification and random forest classification turned out to be pretty good, we used different approaches of data pre-processing and training, which worked best for our own algorithms. A valuable future work would be to use each others approaches and redo all the experiments, to see if there would be any different results.

The results for my part of work turn out to be negative, but there is still much room for improvement and optimization. Our results indicate the service quality data we have cannot be used to accurately predict customer churn by applying k-means clustering and logistic regression classification. In order to perform high quality subscriber analysis, we also need to get more demographic data as mentioned in many research papers.

VIII. Conclusions

The initial purpose of our capstone project was to develop toolsets for extracting meaningful insights from the online video data we obtained from our partner, Conviva. Considering the productivity of the team and also our interests, we divided the project into two parts, Smart Anomaly Detection and Subscriber Analysis and the team was also divided into two subteams accordingly.

Within the Subscriber Analysis subteam, Jefferson Lai and I spent the whole year trying to build a system to accurately predict users' engagement and their subscription status based on their viewing history. We used user behavior data as well as service quality metrics to infer critical user information and used all the data we had access to to train our various models. My contributions towards our final goal include related-work searching, critical information inference from existing data and implementing different machine learning algorithms. During the implementation, I've also applied many useful techniques to increase the credibility of our result, such as forward feature selection, evaluation metrics of clustering and classification, etc. In this section, I'll talk about our project status, project management insights and potential future work.

Project Status

At the beginning, we thought that predicting users' behavior based on his viewing history and the quality of service would be pretty interesting. We made a long list of things we thought we could do, including sentimental analysis of users' comments, content and genre analysis, etc. However, when we actually got the data, we found that we cannot perform most of these analyses since many fields aren't as expected and the data for quality of service was actually more useful for Anomaly Detection team. So as soon as we got to know the data, we started picturing the end product of our subteam, which was a predictor of users near future viewing time and engagement time.

After exploring the data, we found basic regression algorithms cannot yield a good performance on our dataset and users' behaviors are too unstable to predict. Then we turned to exploring the statistical correlation between user engagement and the features we have. At the beginning of the second semester, we changed our goal to proving that we don't have enough data to accurately predicting user engagement, based on the experiments we tried. Jefferson and I took a different path to this goal and tried different techniques towards our final goal. What we got is quite different from what we imagined at the beginning as we understand the problem more and more clearly.

Project Management Insights

Since our five-people team has been divided into two teams, it's actually easier to manage the progress within a smaller team. We used PivotalTracker as our project management tool and we can clearly see the users stories for our project at any point and edit the details of the stories. It worked pretty well at first, but as we got stuck on some problems, we tended to forget about the project management stuff and focused on the problem itself. Besides, there are so many factors that can influence our progress that it's difficult to keep track of everything. For example, at the beginning of this semester, we started to work locally due to the unavailability of the Amazon clusters. This process took longer than we thought but it didn't reflect on PivotalTracker and step by step, our schedule changed greatly from the original plan. This tool didn't really work well on our project, but instead, our team had one to two meetings every week to keep everything on track and everyone has a good idea of what other teammates have been working on. I do think the best way for project management is for the team to sit down and talk about their own progress every several days.

Potential Future Work

The results for my part of work turned out to be pretty bad. But some more optimizations can be done with respect of clustering and classification:

- Use more sophisticated techniques for normalization: since we are mainly focusing on users with relatively low viewing time, we need to normalize the data so that data points are evenly distributed in range 0 to 1 to make the clustering more accurate.
- Find a more accurate way of labelling ‘churners’: currently we define a ‘churner’ by comparing the total playing time of the first half of the dataset and the second half. But churners could have different patterns and our program needs to be more general.
- Run the algorithms with more weeks of data: currently we have 20 weeks of data and the first 10 is used for training and the last 10 is used for testing, which seems a little insufficient for labeling a churner.
- Use statistical tests for result evaluation: we did a lot of experiments but due to the unsatisfying results, we only used some statistical metrics to evaluate our result. However, if we did the optimizations above, we should try more advanced methods.
- Exchange data preprocessing methods with Jefferson: since we both worked on classification, we can actually use each other’s method and redo all the experiments and see if there would be an improvement of the performance.

IX. Acknowledgements

We would like to thank George Necula for advising us throughout the entirety of this Capstone project. We would also like to thank Jibin Zhan from Conviva for introducing the problem space to us and Pat McDonough from Databricks for providing extensive technical support throughout our use of Databricks.

References

Alice Corporation v. CLS Bank. 573 U.S. Supreme Court. 2014. Print.

Associated Press. “Netflix reeling from customer losses, site outage.” *MSNBC*. MSNBC. 24 July 2007. Web. 15 Feb. 2015.

Bessen, James. “The patent troll crisis is really a software patent crisis.” *Washington Post*. The Washington Post. 3 Sept. 2013. Web. 27 Feb. 2015.

Biem, Alain E. “Detecting Anomalies in Real-time in Multiple Time Series Data with Automated Thresholding.” International Business Machines Corporation. US Patent 8,924,333. 30 Dec. 2014.

"Bringing Big Data to the Enterprise." IBM. N.p., n.d. Web. 13 Apr. 2015.

Burez, Jonathan, and Dirk Van den Poel. "Handling class imbalance in customer churn prediction." *Expert Systems with Applications* 36.3 (2009): 4626-4636.

Brundage, Michael L., and Brent Robert Mills. “Detecting Anomalies in Time Series Data”. Amazon Technologies, Inc., assignee. U.S. Patent 8,949,677. 3 Feb. 2015.

Byrd, Owen, and Brian Howard. 2013 Patent Litigation Year in Review. Rep. Menlo Park: Lex Machina, 2014. Print.

CA Inc. “Manage Your Network Infrastructure for Optimal Application Performance.” *CA Technologies*. n.p. n.d. 13 Feb. 2015.

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM Computing Surveys (CSUR)* 41.3 (2009): 15.

Chen, Chao, Andy Liaw, and Leo Breiman. "Using random forest to learn imbalanced data." *University of California, Berkeley* (2004).

Cohn, Chuck. “Build vs. Buy: How to Know When You Should Build Custom Software Over Canned Solutions.” *Forbes*. Forbes Magazine, 15 Sep. 2014. Web. 7 Apr. 2015.

Connelly, J.P., L.V. Lita, M. Bigby, and C. Yang. “Real time audience forecasting.” US Patent App. 20120047005. 23 Feb. 2012.

Conviva. "About Us." *Conviva*. n.p., n.d. Web. 28 Feb. 2015.

Cortes, Corinna, Lawrence D. Jackel, and Wan-Ping Chiang. "Limits on learning machine accuracy imposed by data quality." *KDD*. Vol. 95. 1995.

Coussement, Kristof, and Dirk Van den Poel. "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques." *Expert systems with applications* 34.1 (2008): 313-327.

Dasgupta, Dipankar, and Stephanie Forrest. "Novelty detection in time series data using ideas from immunology." *Proceedings of the international conference on intelligent systems*. 1996.

Deshpande, Amit and Riehle, Dirk. "The total growth of open source." *Open Source Development, Communities and Quality*. Springer US, 2008. 197-209.

Educational & Professional Studies. "What Is Logistic Regression?" University of Strathclyde. N.p., n.d. Web. Accessed on April 13, 2015. <http://www.strath.ac.uk/aer/materials/5furtherquantitativeresearchdesignandanalysis/unit6/whatislogisticregression/>.

Etherington, Darrell. "Twitter Acquires Over 900 IBM Patents Following Infringement Claim, Enters Cross-Licensing Agreement." TechCrunch. N.p., 31 Jan. 2014. Web. 25 Feb. 2015.

Fisher, William W. “Patent.” *Encyclopaedia Britannica Online*. Encyclopaedia Britannica Inc.

Ganjam, Aditya, et al. "Impact of delivery eco-system variability and diversity on internet video quality." IET Journals 4 (2012): 36-42.

García, Roberto, et al. "Statistical characterization of a real video on demand service: User behaviour and streaming-media workload analysis." *Simulation Modelling Practice and Theory* 15.6 (2007): 672-689.

Goldman, Eric. "The Problems With Software Patents (Part 1 of 3)." *Forbes*. Forbes Magazine, 28 Nov. 2012. Web. 01 Mar. 2015.

Gottfriend, Miriam. “Bullish Investors See New Hope for Netflix Profit Stream.” *The Wall Street Journal*. The Wall Street Journal. n.d. Web 14 Feb. 2015.

Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." *The Journal of Machine Learning Research* 3 (2003): 1157-1182.

Hanley Frank, Blair. "Amazon Web Services Dominates Cloud Survey, but Microsoft Azure Gains Traction - GeekWire." *GeekWire*. Geekwire, 18 Feb. 2015. Web. 02 Mar. 2015.

Harvey, Cynthia. "100 Open Source Apps To Replace Everyday Software." *Datamation*. N.p., 21 Jan. 2014. Web. 28 Feb. 2015.

Hung, Shin-Yuan, David C. Yen, and Hsiu-Yu Wang. "Applying data mining to telecom churn management." *Expert Systems with Applications* 31.3 (2006): 515-524.

Iyengar, Vijay S. 2002. “Transforming data to satisfy privacy constraints.” *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '02). ACM, New York, NY, USA, 279-288. Web. 12 Feb. 2015.

Jasani, Hiral. “Global Online Video Analytics Market.” *Frost & Sullivan*. n.p. 5 Dec. 2014. Web. 12 Feb. 2015.

2015 Lai, Jefferson. Technical Contributions to 'Online Video Data Analysis’.

Kahn, Sarah. "Business Analytics & Enterprise Software Publishing in the US." IBISWorld (2014): 5. Web. 11 Feb. 2015.

Keaveney, Susan M. "Customer switching behavior in service industries: An exploratory study." The Journal of Marketing (1995): 71-82.

Keaveney, Susan M., and Parthasarathy, Madhavan. "Customer switching behavior in online services: An exploratory study of the role of selected attitudinal, behavioral, and demographic factors." *Journal of the Academy of Marketing Science* 29.4 (2001): 374-390.

Kejariwal, Arun. "Introducing Practical and Robust Anomaly Detection in a Time Series." Twitter Engineering Blog. Web. 15 Feb. 2015.

Lawler, Richard. "Netflix Tops 40 Million Customers Total, More Paid US Subscribers than HBO." *Engadget*. N.p., 21 Oct. 2013. Web. 15 Feb. 2015.

Liu, Xi, et al. "A case for a coordinated internet video control plane." Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication. ACM, 2012.

MacQueen, James. "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. No. 14. 1967.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. "Evaluation of Clustering." Introduction to Information Retrieval. Cambridge, England: Cambridge UP, 2009. 356-60. Print.

McDonald, John H. "Handbook of Biological Statistics." Simple Logistic Regression. Sparky House Publishing, Dec. 4, 2014. Web. Accessed on Apr. 8 2015. <http://www.biostathandbook.com/simplelogistic.html>.

Mcgovern, Gale. Virgin Mobile USA: Pricing for the Very First Time. Case Study. Boston. Harvard Business Publishing, 2003. Print. 9 Jan. 2010.

NIST. "1. Exploratory Data Analysis." 1. Exploratory Data Analysis. N.p., n.d. Web. Accessed on Mar. 2015. <http://www.itl.nist.gov/div898/handbook/eda/eda.htm>.

"Number of Broadband Connections." *IBISWorld*. IBISWorld. 3. Web. 12 Feb. 2015.

Numenta. "The Science of Anomaly Detection." *Numenta*. n.p. n.d. 13 Feb. 2015.

Open Source Initiative. "Welcome to The Open Source Initiative." *Open Source Initiative*. N.p., n.d. Web. 28 Feb. 2015.

Porter, Michael. "The Five Competitive Forces That Shape Strategy." *Harvard Business Review Case Studies, Articles, Books*. N.p., Jan. 2008. Web. 12 Feb. 2015.

Porter, Michael. "What is Strategy?." *Harvard Business Review Case Studies, Articles, Books*. N.p., Jan. 2008. Web. 12 Feb. 2015.

Quinn, Gene. "A Software Patent Setback: Alice v. CLS Bank." *IP Watch Dog*. n.p. 9 Jan. 2015. Web. 27 Feb. 2015.

Roettgers, Janko. "Netflix Spends $150 Million on Content Recommendations Every Year." *Gigaom*. N.p., 09 Oct. 2014. Web. 15 Feb. 2015.

Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): 53-65.

Shelby County v. Holder. 570 U.S. Supreme Court. 2013. Rpt. in Dimensions of Culture 2: Justice. Ed. Jeff Gagnon, Mark Hendrickson, and Michael Parrish. San Diego: University Readers, 2012. 109-112. Print.

Smith, Sarah. "Analysis of the Global Online Video Platforms Market." *-- LONDON, Jan. 5, 2015 /PRNewswire/* --. Reportbuyer, n.d. Web. 02 Mar. 2015.

Spark. "Apache Spark™ - Lightning-Fast Cluster Computing." Apache Spark™ - Lightning-Fast Cluster Computing. N.p., n.d. Web. Accessed on Mar. 15, 2015. <https://spark.apache.org/>.

Sun Tzu, and James Clavell. *The Art of War*. New York: Delacorte, 1983. Print. 17-18.

Trautman, Erika. “5 Online Video Trends To Look For In 2015.” *Forbes*. Forbes Magazine, 08 Dec. 2014. Web. 14 Feb. 2015.

Tsai, Chih-Fong, and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks." *Expert Systems with Applications* 36.10 (2009): 12547-12553.

UCLA: Statistical Consulting Group. Introduction to SAS. N.p., n.d. Web. Accessed on Mar. 2015. <http://www.ats.ucla.edu/stat/sas/notes2/>.

United States. Cong. Senate. Committee on Commerce, Science, and Transportation. *The Emergence of Online Video : Is It the Future? : Hearing Before the Committee on Commerce, Science, and Transportation*. 112th Cong., 2nd sess. Washington: GPO, 2014. Web. 15 Feb. 2015

Verbeke, Wouter, et al. "Building comprehensible customer churn prediction models with advanced rule induction techniques." Expert Systems with Applications 38.3 (2011): 2354-2364.

Vilas, Manuel, et al. "User behavior analysis of a video-on-demand service with a wide variety of subjects and lengths." *Software Engineering and Advanced Applications, 2005. 31st EUROMICRO Conference on*. IEEE, 2005.

Vinson, Michael, B. Goerlich, M. Loper, M. Martin, and A. Yazdani. “System and method for measuring television audience engagement.” US Patent. 8,904,419. 26 Sep. 2013.

“What Does Copyright Protect? (FAQ) | U.S. Copyright Office.” *What Does Copyright Protect? (FAQ) | U.S. Copyright Office*. N.p., n.d. Web. 01 Mar. 2015.

Worstall, Tom. "The Supreme Court Should Just Abolish Software Patents In Alice v. CLS Bank." *Forbes*. Forbes Magazine, 29 Mar. 2014. Web. 01 Mar. 2015.

Zaharia, Matei, et al. "Spark: cluster computing with working sets." *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*. 2010.

Zeithaml, Valarie A. "Service quality, profitability, and the economic worth of customers: what we know and what we need to learn." Journal of the academy of marketing science 28.1 (2000): 67-85.