Online Video Data Analytics

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

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2015-100
http://www.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-100.html

May 14, 2015

University of California, Berkeley College of Engineering

MASTER OF ENGINEERING - SPRING 2015

Electrical Engineering and Computer Science

Data Science and Systems

Online Video Data Analytics

Wenxuan Cai

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 Science

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 and evaluate a combination of using Weibull Distribution, pruning technique, and online learning to detect anomaly in video start failure data.

Contents

*Co-written with Benjamin Le, Jefferson Lai, Pierce Vollucci, and Yaohui Ye

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 seven 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 ConclusionConcluding Reflections 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 Contribution

Overview

As the development of online content distribution channels and the growth of online video providers, more and more people have joined the population of web streaming which brings the television experience onto their individual screens. The prevalence of Internet video service, as a public trend, takes on-demand video to people's daily life. Erika Trautman, CEO of Rapt Media, in her article for Forbes commended that "each year, more and more people are ditching cable and are opting for online services like Netflix and Hulu" (Trautman).

The rapid growth of online video services contributes to the explosion of video contents all over the Internet and the skyrocketing of video viewer info that is valuable to explore. At this moment, Youtube is reported to serve over 1 billion users and 300 hours of video are uploaded per minutes (Youtube). Over recent years, video suppliers have started to commercialize their products, during which we observe two industrial trends: large scale concurrent access from different parts of world has put a heavy burden on the technical infrusture of video hosting sites; content providers rely on subscription service to generate revenues. My capstone team, consisting of Benjamin Le, Pierce Vollucci, Jefferson Lai, Yaohui Ye, and me, works on online video analytics and attempts to address these two challenges.

Basically, the project is divided into two parts. On one hand, we explored the idea of Smart Anomaly Detection and used machine learning techniques to develop a real-time video platform anomaly monitor; on the other hand, we launched the Subscriber Analysis project, whose goal was to develop predictive models of viewer engagement and churn based on viewing activity and service quality data. The team members selected which sub-team to join based on their own interests and skills. While Ben and Pierce contributed consistently to the Smart Anomaly Detection, Jefferson and

Yaohui spent all of their time working on Subscriber Analysis. Since I expressed interests in both sides, I had affiliation with both teams over the year and contributed to different parts of both projects.

My work over the year can be divided into two chronological phases. In the fall semester, I devoted my efforts to working on Subscriber Analysis. Specifically, I researched different methods to discover a metric to quantify customer engagement, which would play an important role in predicting future subscriber churn. However in spring semester, I worked as a member of Anomaly Detection sub-team on the video start failure[1](VSF) data set and investigated Weibull analysis to detect outliers. Together with Pierce's work on MADe (Seo), we sought the way to characterize the VSF data pattern and to determine the threshold for classifying anomalous data points. At the same time, Ben used time series analysis method to detect anomaly on the seasonal video attempts data set. Eventually, we integrated our work together into the Smart Anomaly Detection, which combined a series of statistical and machine learning techniques into a product that intelligently detects true anomalies while being robust to noise and false alarm.

Literature Review

As a part of customer behavior analysis, customer engagement is a measurement of a user's experience and satisfaction on a brand or product (Chaffey). From web surfing to video watching, from online retailing to Internet service, customer engagement is becoming a critical factor of strategic marketing and drives the commercialization decision of a company. In order to discover an effective way to integrate customer engagement into marketing analysis, various international high-profile conferences and seminars have brought the topic of customer engagement as the primary theme onto the table (Campanelli). The broad spectrum of ongoing research on the topic has encouraged a variety of attempts to build a system to measure and improve customer satisfaction and loyalty in the context of customer

[1] Video start failure, or VSF, is defined as a failed attempts to play the video.

engagement. If we look at a finer granularity, customer engagement can further refer to customer behaviors, marketing practice, and metrics (Summerfield). While a global standard is still on its way, different parties have established for their products different types of metrics, including duration of visit, frequency, click-through rate, etc. Among them, we are particularly interested in the way to quantify a user's engagement in an online video. Although a lot of efforts are put into developing metrics for user engagement on webpages, there is very little publishing which sheds light on the subscriber churn in online video industry (Trautman), nor metrics for online video engagement. Nevertheless, my inspiration came from the engagement measurement for a user on web pages. Published in IMSA 07, Wadee, Miroslav, and Moiez argued that the time spent on a web page was sufficient to infer a user's interest (Wadee). Based on our data set, I was interested in transferring the same idea from webpages to online videos. Thus, in phase one, most of my work focused on inferring a user's engagement on a video from his or her watching time.

In contrast to the sparsity of published studies on online video engagement and subscriber retention, anomaly detection has much richer fields of research ongoing in both academia and industry. Twitter recently posted a blog about anomaly detection on seasonal data by employing time series decomposition (Kejariwal). Twitter also open-sourced its R package to the community. Rather than the autoregressive function model we used for the seasonal video attempts data, Twitter revealed a new way to do time series analysis in anomaly detection. Ben, who worked on Smart Anomaly Detection, successfully transformed Twitter's algorithm from R into Python and gained some positive results on identifying outliers in our video attempts data. However, as stated by Chandola that techniques in different domains of anomaly detection could rarely be used upon each other (Varun), the contextual nature of anomaly detection prevented us from applying the same technique on the non-seasonal VSF data set. For our VSF data, which was neither seasonal nor normal, a widely used method was Weibull analysis due to its ability to assume the characteristics of many different types of distributions. The flexibility of Weibull distribution contributes to its popularity among engineers and broad usage in modeling reliability data (Martz). For example, Weibull analysis is used to model the distribution of product life data, typically for time-to-failure of a product, to estimate manufacture metrics such as reliability time, warranty time, etc (Life). In my phase-two work about anomaly detection on VSF data, I adopted the idea of Weibull analysis to evaluate and validate a threshold for anomalous points by fitting a Weibull distribution.

Method and Materials

More and more services are gaining intelligence from knowledge over big data. Unlike traditional industries, online video providers are already well-positioned to easily collect vast amounts of data about the quality of their products. Although we didn’t have a direct relationship with those video suppliers, we obtained access to online video service data from our cooperation with Conviva Inc, an online video platform which gathers video session data and offers solutions for video analytics and optimization (Kishore). Conviva’s partnerships with different types of content providers including HBO, Disney, etc, gave it a rich data source (Summers). As a part of our contract with Conviva, we signed the NDA and anonymized some data fields for privacy reason. After that, most of our experiments were performed on a 4-month video session data of a particular customer of Conviva.

Regarding the specific individual technical contribution, my timeline in each phase can be generalized into three stages: learning and familiarization with data set and tools; related material research; implementation and result validation.

Phase 1. Tools familiarization and Exploratory Data Analysis

During the first stage of phase one, I spent most of my time exploring the Conviva’s data under the principle of Exploratory Data Analysis (EDA) and learning knowledge of Spark and Databricks. Hosted under Apache Open Source Licence, Spark is a fast and general-purpose clustering computer system which efficiently schedules parallelled jobs into distributed computation nodes and utilizes the in-memory cache to reduce IO communication overhead incurred in traditional map-reduce frameworks such as Hadoop (Zaharia). We chose Spark to be our primary data processing tool for a couple of reasons. On one hand, a lot of machine learning problems require iterative processing over training set to learn the best parameters, which typically applies the same function repeatedly. Traditional mapreduce framework saves the intermediate results onto disks and reads them back later if needed, which incurs unnecessary IO overhead. Spark, using fast in-memory cache for those intermediate values, avoids the performance penalty of data reload and speeds up the entire training process (Zaharia). On the other hand, besides Conviva, we had an industrial partnership with Databricks, whose primary product is a cloud based programmable Spark interface. The startup actively improved the functionality and stability of their tools, and we gained access to their beta version product, where we could program and run Spark code on our EC2 cluster. The online Spark workshop video gave us a quick introduction into the Spark world, and we obtained further domain knowledge by working on the project. Other big data platforms such as Tableau also provided data visualization and processing functionalities, however, unlike Databricks, it didn't have Spark support so we wouldn't be able to use existing Spark machine learning library. Besides that, Databricks offered a programmable interface called cloud notebook which were easy to use when doing customized data processing. We also created a team mailing list during that time so that anyone could easily ask help from the entire team when running into technical difficulty.

When exploring through the dataset, I learned that each video session summary contained more than 150 features, including timestamps, video type, buffer ratio, which we were particularly interested in for measuring customer engagement. Since the data was clean and formatted, we didn't need to do much prepossessing. The only transformation we did was to aggregate 10-min video session summaries when doing anomaly detection.

EDA suggests to use visualization tool at the beginning to identify potential patterns in the dataset (Tukey), so I first made an extraction tool which constructed video watching history for each user from the 4-month data. Basically, the tool aggregated existing session summaries for each user and generated a list of video the users had watched over time. It was not difficult to create the watching history for a single user, but became challenging to construct all users' history concurrently given the huge amount of video sessions and distinct users. The solution here was to avoid constructing any global mapping or collecting data on a single node, but programed in a mapreduce manner that kept user records distributed across nodes and joined them in the end.

After that, I wrote a visualization tool which counted the unique viewers for each video, and plotted the video popularity chart. It not only helped me filter eligible videos to measure customer engagement, but also facilitated Subscriber Analysis to detect the common interests among the customers.

Result - Video popularity visualization

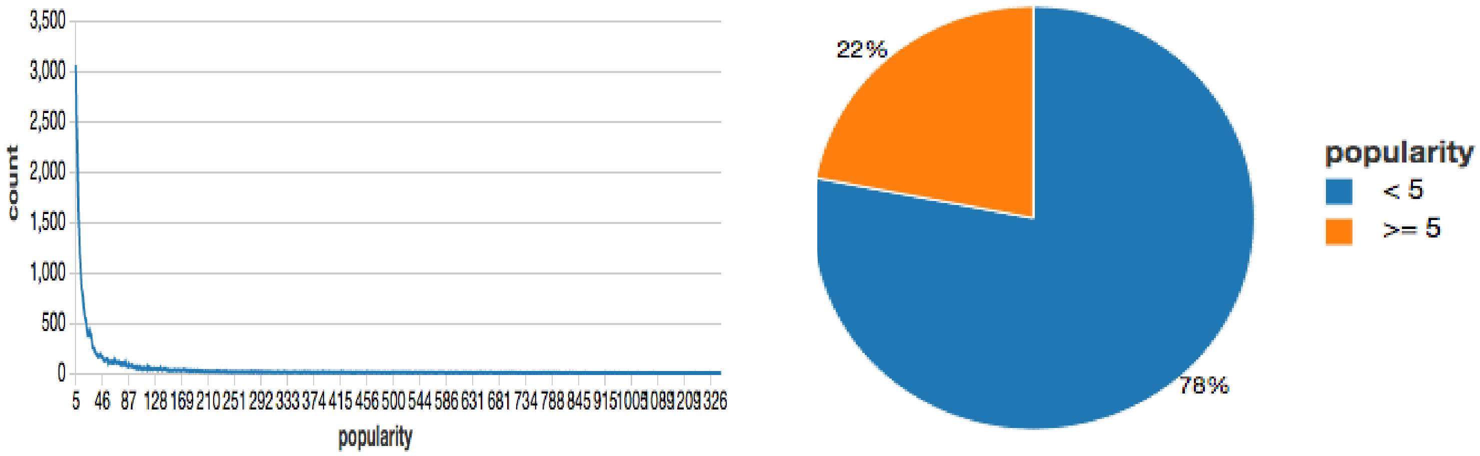

Figure 1. Video popularity plot from the visualization tool

The visualization tool was handy to use. It aggregated the distinct viewerships for each video and sorted videos according to popularity, and revealed useful statistics about the data set. A video session is defined as an instance of a view of a video asset by one device, and contains all the feature metrics that we worked with. From 4 months of video session summaries, we found 1,082,701 unique viewers, who watched

5,092,834 distinct videos in total. **Figure 1** shows the distribution of video popularities. From the plots we can see that around 78% videos have equal or fewer than 5 unique viewers. The long tail of the video popularity data indicates a skewed distribution. While most videos have around 2 or 3 distinct viewers, the popular contents attracted more than 3000 individuals to watch. This value drops sharply as we move to less popular videos. A further breakdown of video popularity statistics is shown below.

Popularity	<10	10 ~ 20	20 ~ 50	50 ~ 100	> 100
Percentage	92%	2%	2%	1%	3%
Count	4,685,407	101,856	101,858	50,928	152,785

Table 1. Video popularity

Although simple, this tool turned out to be really useful to discover interesting videos to work on and to understand the nature of our video session data. For example, if we hoped to generate user targeted recommendations or to catch the ongoing content trend, it could point us to the relevant subset of videos to take a closer look into. Since all the fields it operated on were standard features in the heartbeat messages sent from customers to Conviva, this tool could be used on other customers' data directly. It was also easy to extend to visualize the aggregation of other video session fields such as buffer ratio distribution.

Phase 1. User engagement quantization

The visualization tool helped us to locate the subset of videos of interest to work on, and the next step was to find a proper way to quantify customer engagement from video session data. However, this was still a challenging problem given the circumstance we were in. On one hand, as mentioned in the earlier section, there was very little publishing on quantifying customer engagement from video session data either on an established standard which we could follow; on the other hand, although we had more than 150 features to work with, relatively few of them could be taken to measure customer engagements. Moreover, useful features such as total playing time or total session numbers were highly correlated. Important feature like video length was also unreliable to use. For a detailed explanation of correlation between these features, please refer to Jefferson's and Yaohui's reports.

In this case, video playing time was the most significant feature which we could rely on, and thus our research focused on defining a numerical relationship between the video playing time and a score of user engagement. The methods we hypothesized were largely based on the limited materials online on related items such as webpage engagement. In order to gain some suggestions from the expert who worked on the real-world version of the problem, in October we got in touch with one Conviva engineer, Jim, over skype to exchange ideas on measuring customer engagement and discussed possible video content analysis that we could perform over the data set. Although we didn't formalize any complete algorithm nor collected any results, we list our learning and discussed ideas here in hope that engineers who might work on this topic in the future could consider the pros and cons of these approaches and gain some useful insights from our analysis.

1. Using plain video playing time. This is the simplest method, and is able to characterize the engagement score via a step function. It ensures the monotonicity of the engagement score, but is in risk of oversimplification to assume a linear relationship between viewing time and customer engagement.
2. Using polynomial or exponential function. This is a more reasonable assumption because it magnifies the rewarding rate as the increase of staying time (Wadee). The problem with this and last approaches is that they do not put an upper bound on the engagement score and thus are not a not a fair measurement for videos of different lengths.

3. Using sigmoid function. This is the ubiquitous S-curve that heavily used by economists, technologists and scientists in predictive model and trend mapping (Frederick). The sigmoid function is defined as

$$S(t) = \frac{1}{1+e^{-t}}.$$

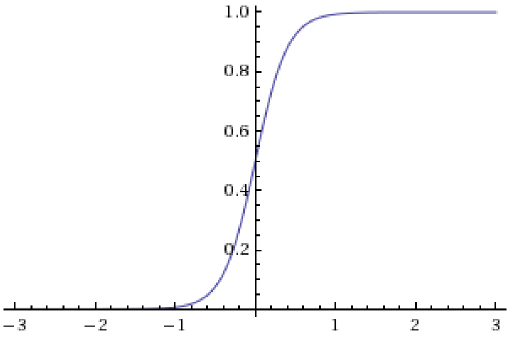

Figure 2. Sigmoid Function Plot

It exhibits a progression from small beginnings that accelerates and approaches a climax over time. The assumption here is that user engagement is supposed to increase at the biggest rate around the mean viewing time. Netflix also utilizes the sigmoid function when computing the quality of experience metric in its patent *Measuring User Quality of Experience for a Streaming Media Service* (John).

A more comprehensive engagement estimation model would consider and assign different weights to other features such as video quality, startup latency, and the likelihood of interruptions in streaming playback. In this sense, a lot of work could be extended and experimented in the future to build a more robust and solid engagement model which allows the service providers to improve subscriber retention and engagement (John).

Phase 2. VSF data exploration

In phase two, I worked closely with Ben and Pierce on Smart Anomaly Detection. They had attained some positive results on the seasonal video attempts data set in fall. This time, the team switched the gears to work on VSF data. While Ben focused on anomaly detection on the subset of aggregated features, I worked closely with Pierce on detecting anomalous VSF data points.

Currently, we aggregated our VSF data set into hourly entries. The fields which we were interested in was the hourly VSF rate -- the percentage of VSF over the total video attempts within an hour. Here is a histogram plot of the aggregated VSF data set.

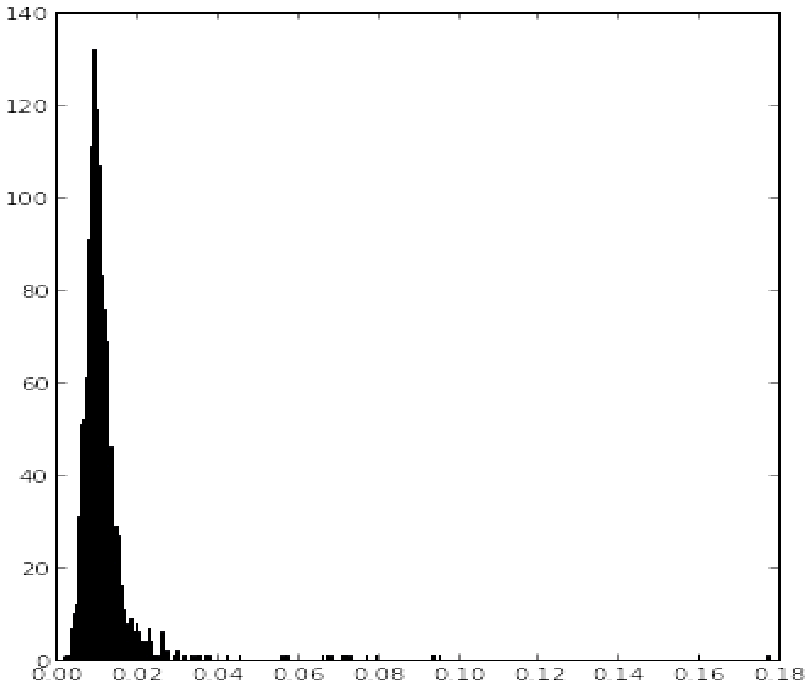

Figure 3. VSF Data Plot

The x axis is the VSF rate, and y axis represents corresponding frequency. Unlike the seasonal video attempts data, the VSF data was non-seasonal. From the graph we could observe that the distribution of VSF data had a spike in the middle and a long tail on the high end.

This different underlying patterns meant that our autoregressive method, which was used earlier to detect anomalies in seasonal video attempts data, would not work here. For the details of the autoregressive method, please refer to Ben’s report. This saved my time from reviewing all the previous work of Smart Anomaly Detection, and I could get started quickly. After going through some of the time series analysis paper the team had touched last semester, I focused on seeking the best model to catch the underlying pattern of VSF.

Although trying to solve the same problem, Pierce and I approached it in different ways. While Pierce worked on MADe method (Seo) which used medium value to estimate a more robust version of “standard deviation” and varied MADe sensitivities to dynamically catch outliers in the VSF data during different time periods, I looked for an approach to fit a distribution parametrically and used probability to compute the likelihood that a point occurred under the parameterized distribution. It would be classified as an anomaly if the VSF rate was larger than mean and the probability was smaller a certain threshold. That’s saying, we only care about suspicious outliers on the right hand side of the distribution. Although a VSF point with value 0 might also have a small probability and might be an anomaly due to the error in the recording system, we could not tell if it was an error or a rare phenomenon, so we chose to ignore this here.

Phase 2. Normality test

Our first attempt was to use the Gaussian distribution, one of the most common continuous probability distributions used for real-valued random variable whose distributions are unknown in the natural and social sciences. (normal). There are a couple of advantages for choosing Gaussian distribution. First, a closed form solution existed for estimating the Gaussian parameters from the data (Anderson). Second, for a Gaussian distribution, the area between given standard units includes a determined percent area. This means we might be able to use z-score, defined as the count of standard deviation away from the mean, to output the confidence that whether a point is an anomaly. Although the **VSF Data Plot** we show earlier in the section reveals an uncommon right skewness in the distribution, the bell shape and clear spike in the middle drove us to fit a Gaussian distribution on VSF data. As we had 4-months VSF data, which was represented as 20,304 entries of 10-min aggregated sessions, we expected to learn the Gaussian parameters empirically. Maximum likelihood estimator is one of the most common methods used to compute the parameters of an assumptive distribution. Take our VSF data as an example, in order to use maximum likelihood estimator for our VSF data, we need to an assumptive distribution first. Given the distribution, we are able to write the probability of each data point, and the equation of the probability that we observe the entire data set. After that, what we need to do is to find the distribution parameters which give the highest probability of seeing the data set. It can be done by deriving the derivative of the equation and solving for the root. We are able to show that the equation's Hessian matrix, the matrix which filled with second derivatives of the equation, is positive semidefinite (Anderson-Darling). This indicates that the root we have calculated gives the global maximum value -- the maximum likelihood. The maximum likelihood estimator served the primary way to measure distribution parameters in our project, and was proved to be a useful technique in our Weibull Analysis. We will provide more details on this in the later sections.

Because 20,304 data points were not a small set to work on, before we ran the maximum likelihood trainer, we decided to perform an Anderson-Darling test on the VSF data first to verify our Gaussian assumption. Anderson-Darling is a modified version of Kolmogorov-Smirnov (K-S) test to measure whether a sample of data is drawn from a given distribution (Anderson). Unlike K-S test which typically assumes no parameters to be estimated in the tested distribution, Anderson-Darling test can be used to test normality based on the existing knowledge of the sample data (Anderson). Particularly, Anderson-Darling test achieves better accuracy than K-S test on normality test by giving more weight to the tail. With the usage of a specific distribution, Anderson-Darling test gains the advantages of allowing a more sensitive test.

Using this kind of descriptive statistics provided a fast way to understand what we worked on. Basically, statisticians call the assumption to verify as null hypothesis and the assumption which contradicts the null hypothesis as alternative hypothesis. During the test, significance level and p-value are two important concepts that statisticians particularly care about. Briefly speaking in context of fitting a distribution to a data set, the significance level is the probability of rejecting the null hypothesis if it is true. Intuitively, this is the tolerance of extreme observations, and is usually set to 1%, 5% or 10%. The p-value is the probability of observing the the result given that the null hypothesis is true, which gives the level of extremity of data. Theoretically, we accept the null hypothesis if p-value is greater than the significance level, and accept the alternative hypothesis otherwise.

Result - Normality test

We can observe the bell shape distribution from the graph. However, the Anderson-Darling test returned a p-value around 2.91e-06. Typically, statisticians use p-value of 0.05 as the threshold. (Anderson-Darling). Since our result was significantly below the threshold, we concluded that our VSF data was not likely to follow a Gaussian distribution. Thus, it seemed that it was not a good idea to learn Gaussian parameters from this VSF data, probably because of the right skewness of the VSF points. A potential alternative here was to use more complex Gaussians to fit the VSF data, which was called the Gaussian Mixture Model. Basically, Gaussian Mixture Model assumes that data points are generated from a mixture of finite number of Gaussians distributions with unknown parameters, and the training process will infer the best parameters to give maximum likelihood on the data (Wu). We were inspired by the Twitter time series data anomaly detection mentioned earlier, which suggested a potential to decompose the VSF data distribution into multiple Gaussians: one for standard VSF, one for anomalous VSF which mimicked the long tail, and optionally one for the noise. By assigning a bigger variance to the second Gaussian to make it flat and superimposing one upon another, it would simulate the right-skewed VSF curve.

Phase 2. Weibull analysis

Before putting all the efforts on the complicated Gaussian Mixture Model, we still expected to seek a simple distribution which could catch the right skewness of the data, so we tried Weibull. Same as Gaussian distribution, Weibull distribution is a member of exponential family, and is defined as

$$P(x, a, c) = ac(1 - \exp(-x^c))^{a-1} \exp(-x^c) x^{c-1}$$

where **x** is the input value, **a** and **c** are the scale parameters. Weibull distribution doesn’t have the symmetric shape as Gaussian. Different shape and scale parameters give it the flexibility to model the unilateral skewness of the data, either left or right. We have introduced the maximum likelihood estimator in the section Normality test, but unfortunately Weibull distribution didn’t have a closed-form solution for the maximum likelihood estimator. In this case,we used gradient descent to approach the maximum likelihood estimator of the Weibull parameters. Basically, gradient is the generalization of function derivatives and points the direction to the maximum or minimum point of the function. What gradient descend does is to follow the direction of gradient and move step by step to the max/min point (Hugo). **Figure 4** shows an example plot of gradient descend on the isocontour of the function.

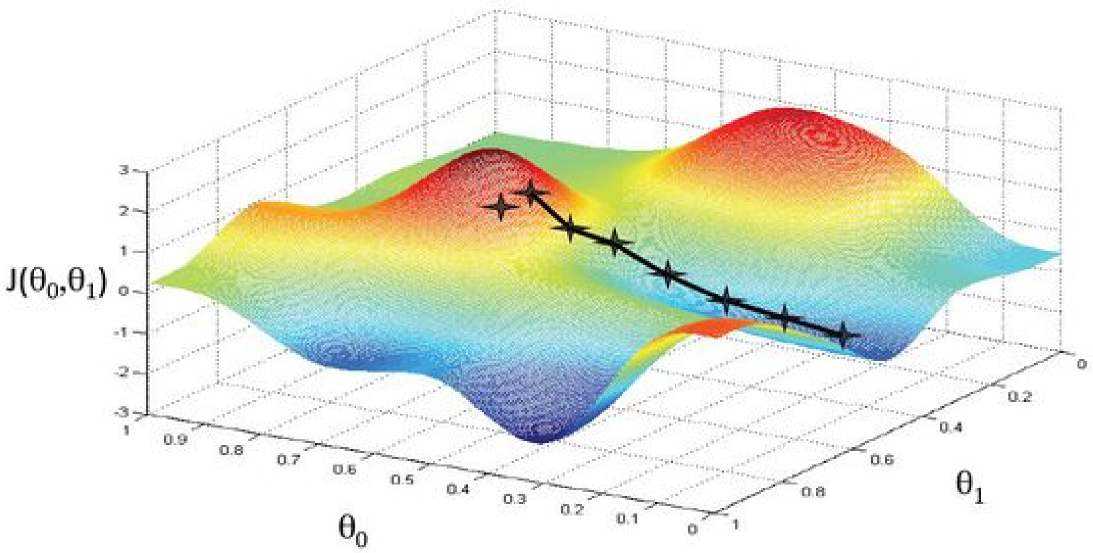

Figure 4. Gradient descend plot (source)

In this graph, the red peak on the left is the maximum point, and the black line shows the path of gradient descend. It computes the gradient at each place with cross, and move along the direction of the gradient. In our case, the maximum likelihood Weibull parameters were generated from iterative training on the VSF data set. During each iteration, it computed the gradient with respect to the two parameters separately and did a batch update to climb on the isocontour towards the global maximum. Since any function from the exponential family are guaranteed to be convex (Lauritzen), with an appropriate learning rate the gradient descent would converge at the maximum likelihood value after iterations.

After we trained the scale and shape parameters, the next step was to validate the goodness of fit. There were two aspects of fit that we cared about.

1. The goodness of fit on the entire VSF data, including the peak and skew. This indicated the overall fit between the distribution and data set.
2. How likely the trained Weibull distribution simulated the long tail of our VSF data, the right skew which corresponded to the anomalous subset of the VSF data. This indicated how good Weibull characterized the outliers in the VSF data. A good result would give us much more confidence to use the trained Weibull distribution to detect VSF anomalies.

To achieve the measurements above, we conducted two tests. For the first metric, we performed the K-S test. The K-S statistical test is a powerful tool to measure the goodness of the fit between a distribution and data samples. In short, we sorted and configured the dataset as a step function, and computed the percentile of each value.

$$F(x_i) = |\{x_j \mid x_j < x_i\}| \quad \textit{for each } x_j \textit{ in the data}$$

After that, for each point x_i, we measured the gap between the $F(x_i)$ and assumptive distribution’s CDF. We recorded the maximum gap and looked up the corresponding p-value, which indicated the significance level that we should accept the null hypothesis that the data set followed the given distribution (Stephens).

For the second metric, we designed our empirical validation procedure, which verified that if the VSF tail distribution was consistent over the time. For example, if the threshold value, which corresponded to the 95 percentile of the first k weeks of VSF data distribution, was around 60 percentile of the next k weeks of VSF data data distribution, it was a red sign that we could use an anomaly threshold from past in future prediction. Thus, we adopted the idea of cross-validation here (Kohavi). At the beginning, we randomly partitioned the VSF data into two sets of equivalent size. Then, we learned the best Weibull parameters from the first partition and selected the threshold according to the probability distribution gained from training. In our experiment, it was the 95 and 99 percentile, which covered most outliers in the first partitions. After that, we used the same threshold on the second partition and measured its percentile. We compared the two percentiles to get an empirical goodness of fit on the VSF data tail.

Result - Weibull analysis

For the Weibull distribution on the entire VSF data set, visually it seemed to be a good fit, but the K-S test result contradicted our intuition. If we ran the K-S test on the VSF data and the best Weibull parameters obtained, the p-value we got was 6.35e-5. It meant that we should reject the hypothesis that the data followed the given Weibull distribution at 10% significance level. The K-S test result seemed to indicate that Weibull distribution was not an appropriate choice for this data set, but the empirical tail test showed some positive signs. Following the method described earlier and repeating the test, we achieved, on average, 96.5 percentile on the second partition by using 95 percentile on the first partition, and 99.1 percentile by using 99 percentile on the first partition. From this result we learned that VSF data had a consistent tail distribution over the time, and our Weibull distribution was able to catch the percentile threshold on the skewed tail.

To figure out the reason for the extremely small p-value returned from K-S test, in **Figure 5** we plotted the CDF of learned Weibull distribution and the step function corresponding to the sorted VSF data points used in the K-S test. We found that

1. It was the outliers that contaminated the K-S test result. Those two curve remained close to each other within the first 70% of cdf, but became larger when VSF data curve hit those outliers.
2. The large gap happened because when seeking the best Weibull distribution parameters, the algorithm considered all the data points, including those extreme outliers, as normal points and tried to stretch the distribution to reach those extremes. The result was a compromise that

failed to simulate either the distribution of outliers or the normal part of VSF data.

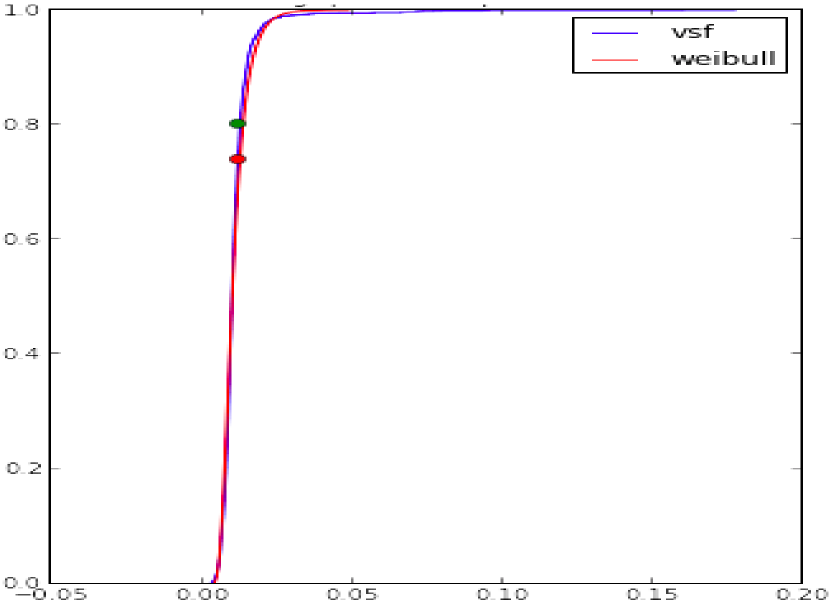

Figure 5. K-S test

Phase 2. Outlier pruning

The analysis of K-S test result above showed us that it were the outliers that affected the p-value returned from K-S test. Although the VSF data had a skewed long-tail distribution, it was a bad idea to use all the data in the tail to compute our Weibull maximum likelihood estimators. Following the assumption that the normal part of VSF data followed a Weibull distribution (right skewed, but not as extreme as the entire VSF data set), we started seeking a pruning method to exclude part of outliers to rectify best Weibull fit.

One easy but brute-force way to finding the best pruning point was to do an exhaustive search from the most outlying point. At each point, we pruned all the data on its right, fit the Weibull distribution, and performed the K-S test. This was a correct way to go but turned out to be really inefficient. In order to find a faster way to reach best trimming point, we employed the “coarse-to-fine” technique to address the problem. Basically, we first used a coarse metric such as standard deviation to approximate the best trimming point, which essentially shortened our searching space. Then we would do an iterative search within the smaller range to get the best trimming point.

For the coarse metric, we had two choices -- standard deviation and MADe. **Figure 6** and **Figure 7** shows a plot of different thresholds of these two metrics.

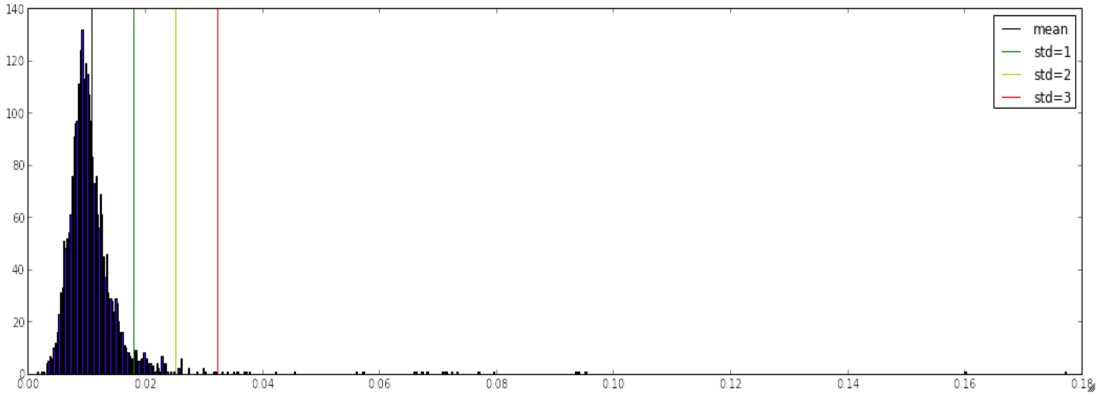

Figure 6. VSF data with different std thresholds

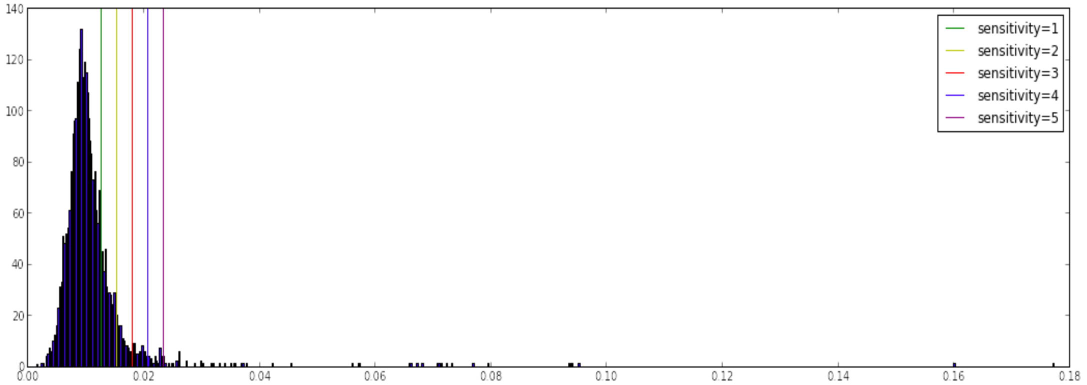

Figure 7. VSF data with different MADe thresholds

In both figure, x value is the VSF rate, and the y value is the frequency. **Figure 6** displays a plot of VSF data with the mean and standard deviations. The outliers appear all the way to the end on the right side. From left to right, the four vertical lines are mean, 1 std, 2 std, and 3 std. Similarly, **Figure 7** shows MADe thresholds with different sensitivities. Standard deviation is affected by outliers and stretched, however, because

MADe is a deviation measured based on the medium value, which is less vulnerable to extremes, it’s smaller than the std in our case.

For our experiments, on multiple subset weeks of VSF data set, we repeated the process of pruning -- fitting Weibull -- K-S test. We averaged the p-values to get a rough evaluation of the coarse trimming point.

Result - Outlier pruning

We used the pruning methods to take away outliers before fitting a Weibull distribution to VSF data. Below is a graph of p-values from Weibull K-S test after we pruned the outliers.

Pruning with std

std pruned	1.0	1.5	2.0	2.5	3.0
p-value	1e-12	0.0309	0.0808	0.0061	0.0099

Pruning with MADe

MADe pruned	1.0	1.5	2.0	2.5	3.0	4.0	5.0	6.0
p-value	1e-8	0.1219	0.3926	0.2030	0.0808	0.0432	0.0162	0.0082

Table 1. Coarse level pruning results

From **Table 1** we could observe that on average, MADe gave a better approximation to the best trimming point. The robustness of MADe to outliers enabled it to find a finer level of partition on the VSF distribution tail. Overall, we found that the p-value went up and down between 1.5 and 2.5 MADe pruning points. The best coarse level pruning result we got was by using 2.0 MADe as trimming point, which returned a p-value of 0.3926. The high p-value indicated that we should accept the null hypothesis that the data followed the estimated Weibull distribution at 1% significance level. This brought us two consequences:

1. The estimated Weibull parameters successfully characterized the pruned VSF data set
2. During our experiments, the best trimming point always fell between 1.5 and 2.5 MADe values. Following this observation in the future, we should be able to first use those two thresholds to sandwich the best pruning point, then did an iterative search within the small range. We would get an optimal Weibull fit after pruning from the best trimming point.

This coarse to fine trimming point searching technique turned out to work well and we used it as the default procedure to find best pruning value in this report. Compared to the brute force approach that iteratively tried each pruning point from the most extreme value on right hand side, our solution greatly reduced the search space at first place and speeded up the entire process. **Figure 8** shows an example plot the K-S test results we gained when performing the iterative pruning point search between 1.5 and 2.5 MADe threshold. In this graph, the leftmost point refers to the 1.5 MADe threshold in the same experiment as **Table 1**, and the rightmost point is the 2.5 MADe threshold. The x values are pruning thresholds and y values show the corresponding K-S test p-values. Doing this iterative search within the range helped us found the best trimming point, which gave a Weibull fit of p-value 0.564.

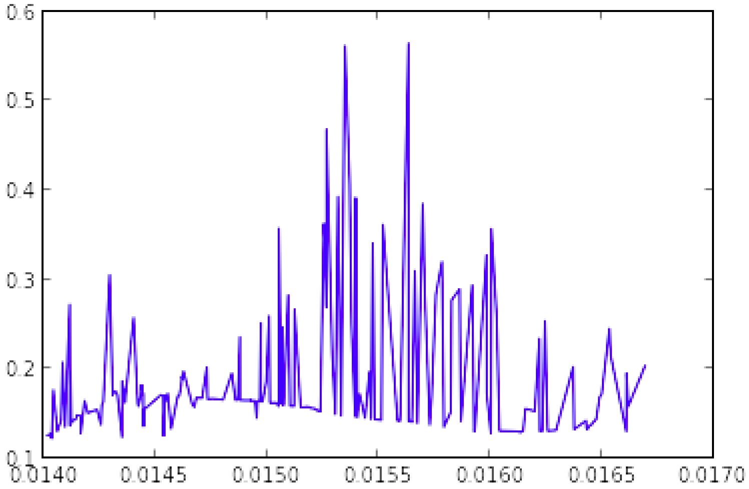

Figure 10. Iterative pruning results between 1.5 and 2.5 MADe

Phase 2. Online learning and Formal anomaly detection validation

Combining the coarse level squeeze by MADe and iterative search allowed us to find the best trimming point to fit a Weibull distribution to model the behavior of normal points in the VSF data. In this case, plugging in a normal VSF point into the probability density function would return a reasonable probability mass. while in contrast, the distribution assigned a relatively tiny probability to extreme outliers. A tiny probability indicated that the VSF point was very unlikely to occur, thus had a high possibility to be an anomaly. This was the rule that we followed when detecting anomalies with fitted Weibull distribution, and the probability threshold was decided to be the value which gave best performance on the training set.

The last step was to transfer the algorithm to an online learning method which gradually adapted new points and adjusted the parameters (Saad). Basically, we used a sliding window which consisted of previous 1-month of VSF points to train the predictor, and moved the window per 10 minutes to adopt new points. To achieve this, the simplest solution was to reestimate the Weibull parameters every time we slided the training window. However, a more efficient way to do it was to cache the last maximum likelihood estimator and adjust it based on the new gradient computed after moving the window. This saved the time from reprocessing the overlap part inside the window.

In order to have a measure of the predictor's performance, we conducted a formal accuracy test on the VSF data. Because we didn't have any anomaly info associated with the VSF data, we manually labeled the outliers. Basically, we categorized VSF data points into four different levels: 0 as no error, 1 as probably not an error, 2 as an error, and 3 as an extreme error. My work was to label the VSF data from points 15114 to 17129. There are 2016 points in total and each one was the VSF rate of the 10-min window. 2016 points corresponded to two weeks of data that we tested on.

Below are the three levels of VSF data points we manually labeled for testing. The x value is daytime and y value is the VSF rate. It's unfortunate that we can't automate this process because currently Conviva didn't have any labeled anomalous data set so they had to go back through logs and emails to confirm if a point was an anomaly. The labeling decisions came from human judgement, but we had reached a criteria at the beginning and tried to make the labeling as consistent as possible. Following are the anomaly plots for our validation week.

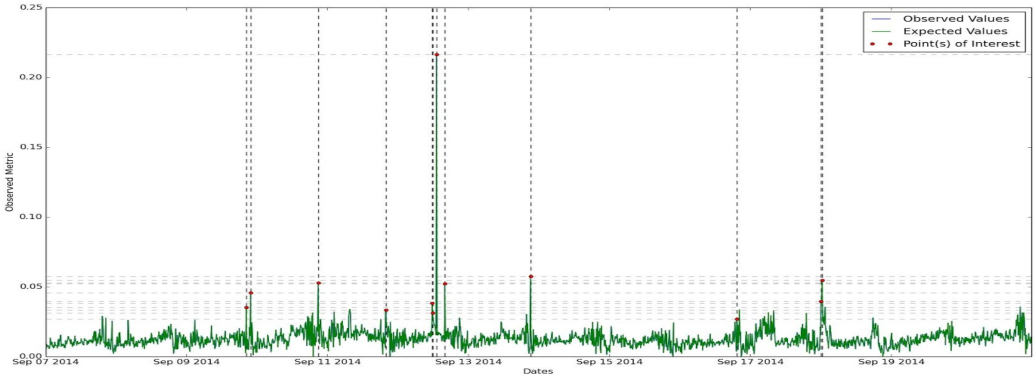

Figure 9. Level 3 anomaly

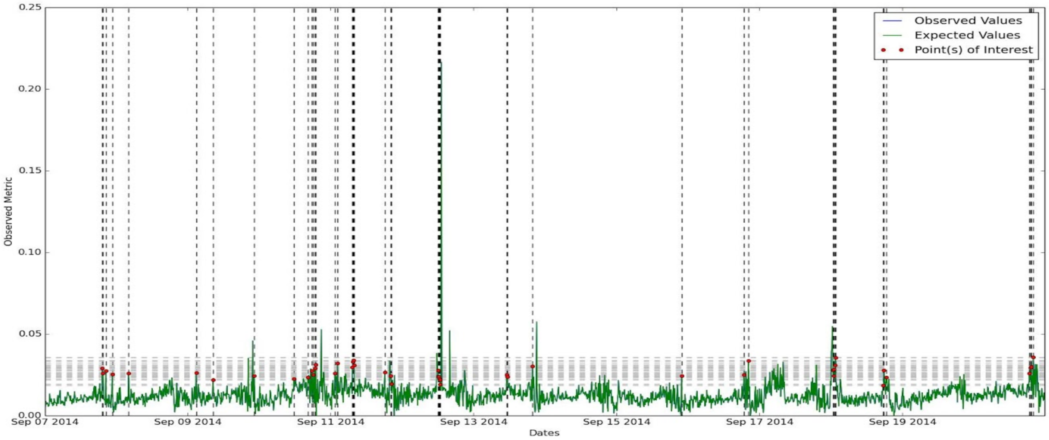

Figure 10. Level 2 anomaly

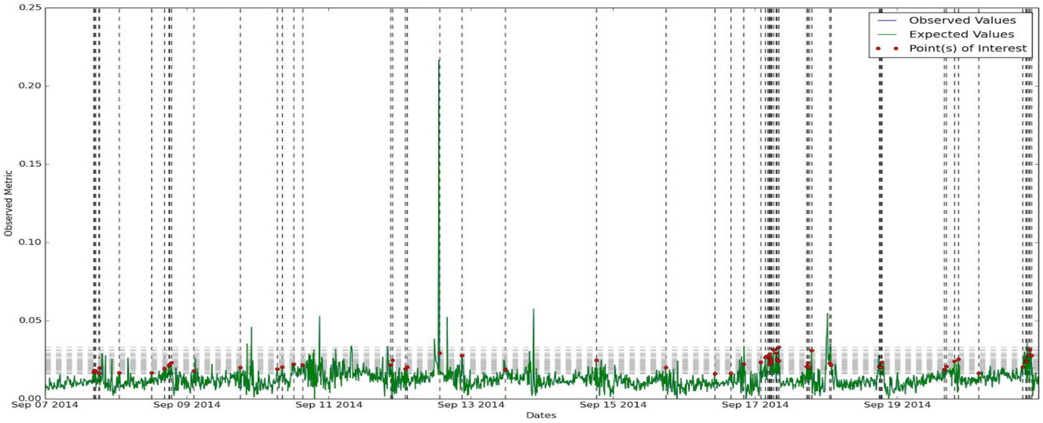

Figure 11. Level 1 anomaly

The first week among the two, as a validation set, was used to tune the best probability as predicting threshold, and we tested on the second week with tuned best predicting threshold.

During tuning and testing, we used precision, recall, and F1 score to evaluate the predictor's performance. The predicting precision is defined as the fraction of retrieved instances that are relevant, while recall is defined as the fraction of relevant instances that are retrieved (Powers). Taking the harmonic mean of them we got the F1 score which was a balanced measurement of test accuracy. Mathematically speaking,

$$TP = true\ positive;\ TF = true\ negative;\ FP = false\ positive;\ FN = false\ negative$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1 = 2 * Precision * Recall / (Precision + Recall)$$

Speaking in the context of Smart Anomaly Detection, a high precision means most fired alerts signal a true anomaly, and a high recall means most true anomalies are caught by the system. Since it's unsatisfactory to either wake an engineer up at night multiple times because of false alert, or mistakenly neglect a true anomaly which could jeopardize the service, we chose F1 as an unbiased measurement to incorporate both precision and recall. That's saying, we first trained the model and made predictions on the test data, after which we counted the number of true positive, true negative, false positive, and false negative. Then we tried different probability threshold and picked the one which gave highest F1 score. After we selected the sensitivity threshold which gave best F1 on the validation set, we ran the model on the second week to get a point F1 estimation as the final score.

Result - Formal anomaly detection validation

During the tuning, we used online learning and trained the Weibull distribution based on the previous one month of data. It was equal to 24*4*6*4 = 4032 points of 10-minute aggregated VSF data. Based on the previous experiment results, we started with a coarse squeeze of 1.5 and 2.5 MADe to approximate the best trimming point, did an iterative search to the pruning point, fit the Weibull distribution, set the anomaly probability threshold, and evaluate the performance. Candidate anomaly thresholds we used were 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8, 1e-9, 1e-10, 1e-11. Tuning with different predicting thresholds gave us different performance, and we collected them together into the F1 plot below.

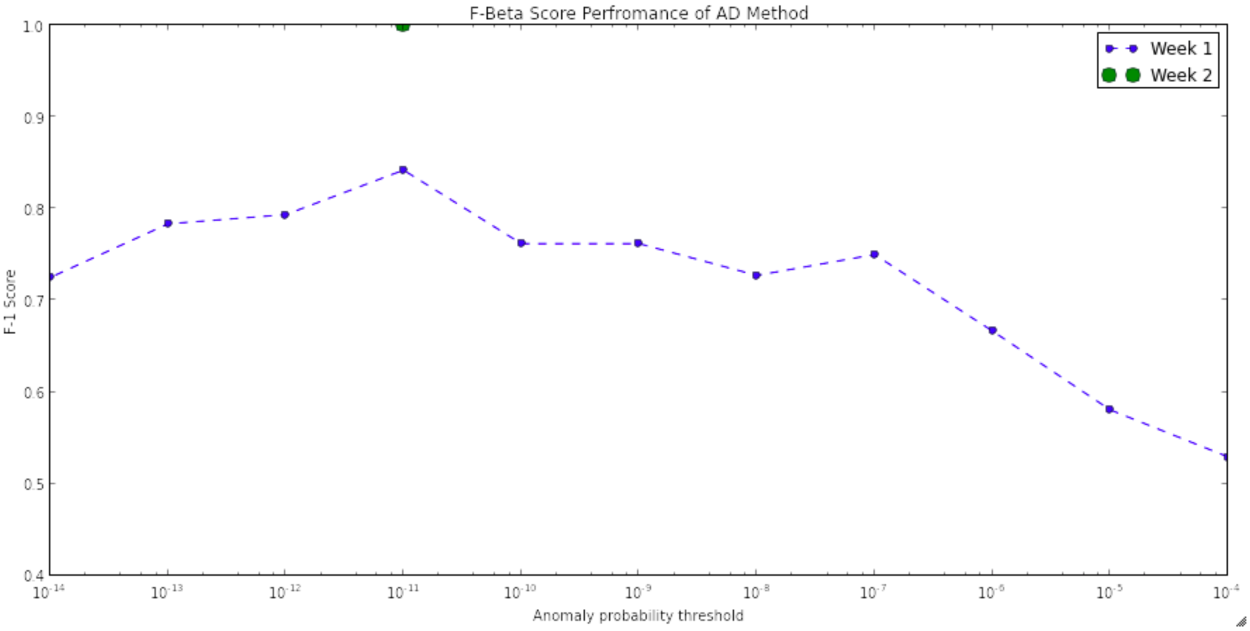

Figure 12. F1 plot for tuning anomaly thresholds on week1 and testing on week2

When graphing, we varied the anomaly probability threshold on the x-axis, and plotted corresponding F1 score on the y-axis. The dashed blue line shows the relationship between F1 scores and the anomaly probability thresholds on the first week of test set. From the result, we picked the threshold of 1e-11, and predicted on the second week of test set. We achieved an F1 score of 1.0, which is the green point on the graph. This indicated that our predictor performed reasonably well and had a good balance between precision and recall. Below is the detailed results.

	Predicted Positive	Predicted Negative
Actual Positive	3	0
Actual Negative	0	1005

Table 2. Confusion Matrix for Week 2

Accuracy	F1	Precision	Recall
1.0	1.0	1.0	1.0

Table 3. Additional Metrics for Week 2

Besides the test set F1 plot, we also plotted the receiver operating characteristic (ROC) curve for the predictor. By running the tests repeatedly with different sensitivity parameters, we graphed the ROC to demonstrate the performance of the predictor as we tweaked the discrimination threshold (Hanley). We also reported the area of undercurve, a 0-1 measurement of the predictor's accuracy compared to random guessing.

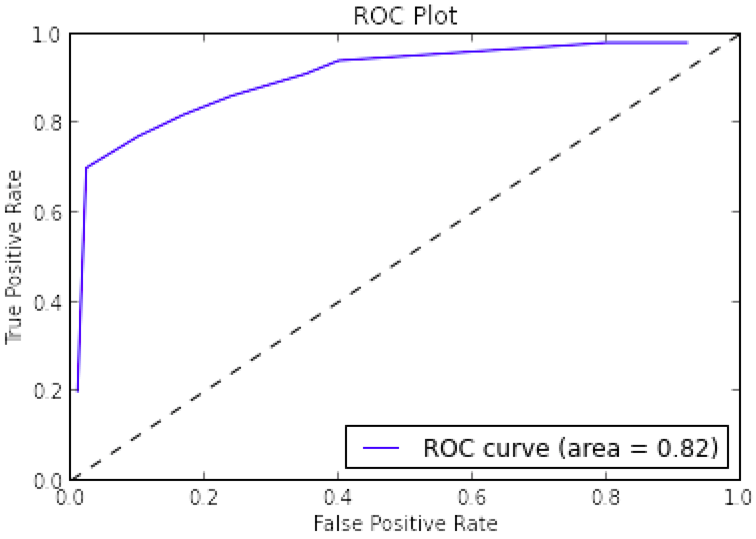

Figure 13. ROC plot for different anomaly thresholds on Week1

The ROC graph shows the true positive/false positive rates when picking different predicting probability thresholds to detect anomalies on the week1 data. This tells that how likely the threshold can be tuned to vary the true positive rate on the y-axis and the false positive rate on the x-axis. The straight line on the diagonal is the performance of a naive classifier which would guess all false at the bottom left corner and all true at the top right corner. An ideal classifier would gain a 100% true positive rate and 0% false positive rate, which stays at the top left corner. The better a binary classifier performs, the further it would stay away from the diagonal line in the graph, and thus closer to the top left corner. The area under the curve is a quantification of a predictor's discriminative capability, and an area of 1.0 represents a perfect discriminative classifier.

Our results above confirmed that MADe pruning, Weibull distribution, and online learning were a reasonable solution for detecting anomaly in VSF data set. Much Superior to using a simple, fixed empirical threshold, the combination of these methods took advantage of the moving average to dynamically catch the changing variance of the VSF data points and gave a much more accurate prediction over the anomalous points.

VII. Concluding reflections

Project Status

Online video analytics is an exciting field to explore. The project is still on its way and there is lots of potential future work that could be extended. My phase one work had provided the subscriber analysis with an insight into measuring customer engagement and the visualization tool benefited them by offering a data filtering technique. In phase two, my Weibull analysis on the VSF data served a parametric way to model the long tail of VSF. Using MADe pruning, Weibull analysis, and online learning, it provided a robust approach to accommodate varying variance of VSF data overtime and dynamically changed the anomaly thresholds to obtain high prediction accuracy.

Comparing to the original plan, we had made some changes according to the situations we were in. For example, at the beginning we also planned to commit some work to projects related to video content such as user customized video recommendations, but unfortunately due to the limited content related features of our data set and some technical difficulties encountered, we aborted the project and merge the efforts into anomaly detection.

Speaking for the Smart Anomaly Detection, I think we successfully finished most of the original goals. We showed that autoregressive function could be effectively used to detect anomaly on video attempts data, and MADe/Weibull analysis were proved to be good predictors on the VSF data.

Project management insight

Although we made a complete plan at the beginning, there were still many unpredictable problems we encountered over the process. The best learning here is to always keep a backup plan. An experienced project manager would be able to help the team pivot smoothly by encouraging team communication and analyzing different possible workarounds. It’s important to organize regular cross team discussion as a part of project management. Sometimes, there was a lack of communication between sub-teams and different sub-teams got stuck on the same problem. In this case, gathering them together and sharing the ideas could help brainstorm solutions and bring inspirations to both sides. This still holds true even when sub-teams run into different problems. Sometimes, those closely involved can not see clearly. An opinion from a spectator could point out a different perspective to look into.

Future work

A lot of work could be continued and optimized if we had more time. Researching a robust pruning method could be a good start for someone who is going to take over the project. Basically, the goal is to experiment different pruning methods and integrate the best one into the anomaly detection process. This would benefit the performance of our auto regressive function, MADe method, as well as Weibull analysis which utilized MADe to remove outliers. The anomaly detection techniques we implemented were online learning algorithms which accepted stream of data and updated the model parameters in real time. This caused problem when processing through anomalous data -- the parameters would be contaminated by the anomaly and further classified a normal point as anomaly before they were updated again to correct state. A possible solution here is to prune anomalous points and prevents the model from training on them, so the following question is how we should pick the threshold and make the pruning decision. This is significant to the stability and robustness of the model, and would boost the performance to a new level.

It’s a challenging and rewarding project, from which we learned how data scientist work in a real world project. Unlike the assignment we did in the course, which usually had clear guidance and was guaranteed to obtain a positive outcome if following the required constraints, doing data science in the real world is tricker and demanding. It requires extra work of research and many ideas can’t be validated on the paper. It’s a precious opportunity for me to practice both communication and cooperation skills. It taught me the way to be a professional data scientist.

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.

Reference

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

"Anderson-Darling Normality Test." ISixSigma. Six Sigma, n.d. Web. 03 May 2015. <http://www.isixsigma.com/dictionary/anderson-darling-normality-test/>.

Anderson, Theodore W. "Maximum likelihood estimates for a multivariate normal distribution when some observations are missing." *Journal of the american Statistical Association* 52.278 (1957): 200-203.

Anderson, T. W.; Darling, D. A. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. Ann. Math. Statist. 23 (1952), no. 2, 193--212.

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.

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.

Campanelli, Melissa. "Engagement Is next Phase in Marketing Communications

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

Chaffey, Dave. "Customer Engagement Interview with Richard Sedley of CScape - Smart Insights Digital Marketing Advice." Smart Insights. Smart Insights, 29 Apr. 2007. Web. 16 Mar. 2015. <http://www.smartinsights.com/customer-engagement/customer-engagement-strategy/customer-engagement-interview-with-richard-sedley-of-cscape/>.

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

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.

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.

Experian Summit." Direct Marketing News. Direct Marketing News, 18 Jan. 2007. Web. 16 Mar. 2015.

<http://www.dmnews.com/engagement-is-next-phase-in-marketing-communications-experian-summit/article/94175/>.

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.

Frederick, David. "S Curves Everywhere." David Fredericks IAIR BLOG. N.p., 21 July 2011. Web. 16 Mar. 2015. <https://instituteair.wordpress.com/2011/07/21/s-curves-everywhere/>.

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

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.

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.

Hugo. "Gradient Descent." Gradient or Steepest Descent. OnMyPhd, n.d. Web. 03 May 2015. <http://www.onmyphd.com/?p=gradient.descent>.

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.

John Funge, Mark Watson, Wei Wei, and David Chen. *Measure user quality of experience for a streaming media service*, June 16 2011. US Patent App. 13/329,038.

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.

Kejariwal, Arun. "Introducing Practical and Robust Anomaly Detection in a Time Series | Twitter Blogs." Introducing Practical and Robust Anomaly Detection in a Time Series | Twitter Blogs. Twitter Engineering Blog, 01 June 2015. Web. 16 Mar. 2015. <https://blog.twitter.com/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series>.

Kishore, Aditya. "Why Quality of Experience Is the Most Critical Metric for Internet Video Profitability." The Guardian. The Guardian, 7 Aug. 2013. Web. 16 Mar. 2015. <http%3A%2F%2Fwww.theguardian.com%2Fmedia-network%2Fmedia-network-blog%2F2013%2Faug%2F07%2Finternet-video-profitability-metric%3FCMP%3Dtwt_gu>.

Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (San Mateo, CA: Morgan Kaufmann) **2** (12): 1137–1143

Lauritzen,, Steffen. "Redirecting." Redirecting. N.p., 7 Nov. 2004. Web. 16 Mar. 2015. <http://www.google.com/url?q=http%3A%2F%2Fwww.stats.ox.ac.uk%2F~steffen%2Fte

aching%2Fbs2siMT04%2Fsi6bw.pdf&sa=D&sntz=1&usg=AFQjCNHlyWBRF89QhP9Gk72t15EUINY5yg>.

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.

"Life Data Analysis (Weibull Analysis)." Reliability. Weibull.com, n.d. Web. 16 Mar. 2015. <http://www.weibull.com/basics/lifedata.htm>.

Martz, Eston. "Why the Weibull Distribution Is Always Welcome | Minitab." Why the Weibull Distribution Is Always Welcome | Minitab. The Minitab Blog, 8 Mar. 2013. Web. 16 Mar. 2015. <http://blog.minitab.com/blog/understanding-statistics/why-the-weibull-distribution-is-always-welcome>.

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

M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: cluster computing with working sets. In Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, pages 10–10, 2010.

"Normal Distribution." Gale Encyclopedia of Psychology. 2001. *Encyclopedia.com.* 3 May. 2015<http://www.encyclopedia.com>.

“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.

Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

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.

Saad, David. On-line Learning in Neural Networks. Cambridge: Cambridge UP, 1998. Print.

Seo, Songwon (2006) A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. Master's Thesis, University of Pittsburgh.

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.

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

Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons". *Journal of the American Statistical Association* (American Statistical Association) **69** (347): 730–737.

Summerfield, Patti. "Redirecting." Redirecting. Srategy, 1 June 2006. Web. 16 Mar. 2015. <http://www.google.com/url?q=http%3A%2F%2Fstrategyonline.ca%2F2006%2F06%2F01%2Fmedia-20060601%2F&sa=D&sntz=1&usg=AFQjCNH8ckRxGVc9S_oi9Y0VkiJhtteiWQ>.

Summers, Nick. "HBO Extends Conviva Deal By Six Years To Improve HBO GO." TNW Network All Stories RSS. N.p., 23 Jan. 2013. Web. 16 Mar. 2015. <http://thenextweb.com/media/2013/01/23/hbo-extends-its-deal-with-conviva

Trautman, Erika. “5 Online Video Trends To Look For In 2015.” *Forbes*. Forbes Magazine, 08 Dec. 2014. Web. 16 Mar. 2015.

Tukey, John W. The Future of Data Analysis. Ann. Math. Statist. 33 (1962), no. 1, 1--67.

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

Varun Chandola, Arindam Banerjee, and Vipin Kumar. “*Anomaly detection: A survey*”. ACM Computing Surveys, 41(3):15:1{15:58, July 2009.

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

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.

Wadee S. Al halabi, Miroslav Kubat, and Moiez Tapia. 2007. Time spent on a web page is sufficient to infer a user's interest. In *IASTED European Conference on Proceedings of the IASTED European Conference: internet and multimedia systems and applications* (IMSA'07). ACTA Press, Anaheim, CA, USA, 41-46.

“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.

Wu, Y. "Gaussian Mixture Model." *Connexions* (2005).

"Youtube Statistics." YouTube. YouTube, n.d. Web. 06 Mar. 2015. <https://www.youtube.com/yt/press/statistics.html>

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.