Churn Prediction in MMORPGs using Player Motivation Theories and an Ensemble Approach

Zoheb Borbora, Jaideep Srivastava
Dept. of Computer Science and Engineering
University of Minnesota
borbo001@umn.edu,
srivasta@cs.umn.edu

Kuo-Wei Hsu
Dept. of Computer Science
National Chengchi University
Taipei, Taiwan
hsu@cs.nccu.edu.tw

Dmitri Williams
School for Communication and Journalism
University of Southern California
Los Angeles, USA
dcwillia@usc.edu

Abstract— **In this paper, we investigate the problem of churn prediction in Massively multiplayer online role-playing games (MMORPGs) from a social science perspective and develop models incorporating theories of player motivation. The ability to predict player churn can be a valuable resource to game developers designing customer retention strategies. The results from our theory-driven model significantly outperform a diffusion-based churn prediction model on the same dataset. We describe the synthesis between a theory-driven approach and a data-driven approach to a problem and examine the trade-offs involved between the two approaches in terms of prediction accuracy, interpretability and model complexity. We observe that even though the theory-driven model is not as accurate as the data-driven one, the theory-driven model itself can be more interpretable to the domain experts and hence, more preferable over a complex data-driven model. We perform lift analysis of the two models and find that if a marketing effort is restricted in the number of customers it can contact, the theory-driven model would offer much better return-on-investment by identifying more customers among that restricted set who have the highest probability of churn. Finally, we use a clustering technique to partition the dataset and then build an ensemble on the partitioned dataset for better performance. Experiment results show that the ensemble performs notably better than the single classifier in terms of its recall value, which is a highly desirable property in the churn prediction problem.**

I. Introduction

A massively multiplayer online role-playing game (MMORPG) is a popular genre of computer-based game which is characterized by a persistent virtual world maintained by the game developer. In an MMORPG, each player controls a game character and performs different activities in the virtual game environment. Such activities can include interactions with the environment such as fighting monsters as well as interactions with other players such as player vs. player matches, raids and trading items.

MMORPGs can generate substantial revenue in a variety of ways – while games such as World of Warcraft (WoW) have a subscription-based model in which players pay a regular fees, other games such as Lords of the Ring Online and Dungeons & Dragons Online are free-to-play. Common revenue models for free-to-play games include virtual item sales, subscription tiers and advertisements displayed within the games [2]. The estimated worth of the MMORPG market in the US, as of 2009, stands at $6 billion, with little letdown in growth expected going forward [3]. The huge revenue potential has attracted several game producers to this market segment and with increased competition, customer acquisition and retention is of major concern to game companies.

The focus of this paper is churn prediction in MMORPGs. In this work, we have used real-world data from Sony Online Entertainment's EverQuest II (EQII). It is a fantasy-based MMORPG where each player creates a character and embarks on a never-ending quest for advancement and exploration. Within the game, the character can adventure (complete quests, explore the world, kill monsters and gain treasures and experience) and socialize with other players. At the time the data used here were recorded, EQII was a subscription-based title, although it has since shifted to the free-to-play model.

We investigate the problem of churn prediction from the perspective of analyzing player motivation. In our approach, churners are defined based on not only their subscription information but also their activity signatures within the game. Players with greater motivation and involvement are less likely to leave a game. We, therefore, draw on different player motivational factors and build prediction models based on these factors. Three key contributions of this paper are as follows.

First, we build two classifiers – one of which has a small number of features drawn from the theories for player motivation and the other has a larger number of features extracted from data of player logs. We describe the synthesis between a traditional theory-driven approach and a data-driven approach to a problem and examine the trade-offs in terms of prediction accuracy, interpretability and model complexity.

Second, we generate and compare lift curves for the data-driven and theory-driven models and find that if we are only allowed to consider a small portion of test instances the theory driven-model has better lift. On the other hand, if we are allowed to consider a larger portion, but not all, of test instances, the data-driven model performs better.

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

Third, we explore how ensemble techniques can improve on the performance of churn prediction models. It has been shown that an ensemble, a committee of classifiers aggregated to provide overall predictions, usually outperforms single prediction models (i.e. individual classifiers). We utilize a clustering technique to partition the dataset first and then build a classifier on each segment. Experimental results show that the ensemble performs notably better than the single classifier in terms of its recall value, which is a desirable property in the churn prediction problem because it means that the classifier is able to identify a larger percentage of the churners correctly. In a typical churn prediction problem, failing to identify potential churners (false negatives) can be much more costly than wrongly identifying non-churner as a churner (false positives).

II. Background and related work

Churn is an important problem for any business with repeat customers as it directly affects revenue. As such it has been analyzed in a wide range of industries, particularly in the telecom sector [5-13], but also in other domains such as retail business [14], banking [15,22], Internet service providers [16], service industries [17], P2P networks [18], insurance [20], credit card [21] and MMORPGs [19]. Reference [4] gives an overview of the current research on churn analysis and prediction – the main focus being churn in digital social networks and how it differs from churn in the telecommunication networks.

A wide-array of techniques has been used for churn analysis. For example, logistic regression models [11, 20-22], decision tree models [6, 8, 11, 14], neural networks [8, 11, 12, 14], and support vector machine (SVM) [13-15] are available. Besides these traditional approaches, alternative techniques have also been explored for churn analysis and prediction, such as survival analysis [9] and genetic algorithms [6, 7]. Social Network Analysis (SNA) has emerged as an important technique for studying complex, real-world networks and researchers have started using SNA methods as an alternative or extension to customer churn prediction [4]. The effect of different behavioral and structural features on the user's churn likelihood in an online social network has been analyzed in [37]. Reference [5] demonstrates a simple diffusion-based approach that exploits social ties to identify a significant fraction of churners in a social network and reference [19] proposes a churn prediction model that uses a modified diffusion model to propagate the social influence (which has both a positive and a negative component) of a player in a social network. With the ability to predict customer churn, it becomes possible to calculate the lifetime value (LTV) of a customer. Reference [23] proposes an LTV model considering the past contribution, potential value and churn probability of a customer.

A. Social science perspective

Player motivation is one of the most important factors that can help in analyzing and predicting churn behavior. Bartle's original motivation taxonomy [24] of early text-based game players has served as the industry standard, but has no empirical basis. In it, he proposed that there were several "types" of players, and that these were mutually exclusive. A more recent and empirically based taxonomy by Yee [25] uses validated scales to detect player motivations, but does not preclude multiple "types" from existing within a single player.

TABLE I. A Model of Player Motivations [25]

Achievement	**Social**	**Immersion**
Advancement Progress, Power, Accumulation, Status	***Socializing*** Casual Chat, Helping Others, Making Friends	***Discovery*** Exploration, Lore, Finding Hidden Things
Mechanics Numbers, Optimization, Templating, analysis	***Relationship*** Personal, Self-disclosure, Find and Give, Support	***Role-playing*** Story Line, Character History, Roles, Fantasy
Competition Challenging Others, Provocation, Domination	***Teamwork*** Collaboration, Groups, Group achievement	***Customization*** Appearances, Accessories, Style, Color schemes
		Escapism Relax, Escape from Real Life, Avoid Real Life problems

Nick Yee used a factor analytic approach to create an empirical model of player motivations [26]. The analysis revealed 10 motivation subcomponents that grouped into three overarching components (achievement, social, and immersion) with underlying relationships between motivations and demographic variables (age, gender, and usage patterns), as illustrated in Table I.

B. Data mining perspective

Over the years, ensemble techniques have been an active topic of data mining research. For classification tasks, an ensemble technique constructs a group of effective member classifiers and aggregates outcomes from them. Effective member classifiers are those fine and diverse. The former means that a member classifier is expected to provide reasonable performance, while the latter means that the correlation between outcomes from two member classifiers is expected to be small. It is commonly admitted that member classifiers in an ensemble need to be diverse in order to improve their performance by aggregation. The intuition is that if some member classifier commits an error on some specific data sample then other diverse member classifiers would have a higher chance to correct the error.

From one point of view, ensemble techniques have become popular for data mining practitioners because every classification algorithm has its own limitations, such as the way it generates decision boundaries and its capability to tolerate noise. From another viewpoint, the idea behind all ensemble techniques is that, if each member classifier has expertise in analyzing samples specific to some portions of a given data set then the final outcomes aggregated from all member classifiers would become more reliable and stable. Consequently, one potential advantage of using ensembles is the enhancement of stability and robustness of the resulting classification models.

III. THEORY-DRIVEN AND DATA-DRIVEN APPROACHES – A SYNTHESIS

In this section, we describe the idea of synthesis between a theory-driven and a data-driven approach to a problem [28]. A classical theory-driven approach would typically rely on statistical hypothesis testing. Let us consider an example where we intend to assess the truth of the hypothesis that achievement-orientation is a key factor of player motivation. In an MMORPG, a measure of achievement-orientation of a player could be his/her rate of progress (acquiring points or leveling up) within the game. To test the hypothesis, we could take samples of motivated (non-churners) and non-motivated (churners) players and perform a two-sample significance test with the null hypothesis as the mean rate of progress for the two populations are equal, and the alternative hypothesis – the mean rate of progress of motivated players is higher than that of non-motivated players. The null hypothesis is rejected if the P-value of the test statistic is less than or equal to the pre-specified significance level; otherwise, we fail to reject the null hypothesis.

A data-driven (e.g. data mining) approach, on the other hand, is different from the hypothesis-driven approach described above. Game logs collect an entire range of the online activities of players in great detail. In a data-driven approach, one can choose a meaningful subset of the observed variables or features and perform tasks such as classification, clustering or association rule mining (or frequent pattern mining). For example, we can pose churn prediction as a binary classification problem and construct models with appropriate features constructed from the game logs. Using appropriate metrics, one can evaluate the effectiveness of the model in predicting churners and/or non-churners.

If we think in terms of interpreting the outcomes of patterns discovered from a data-driven approach, there could be three possibilities [28], as follows:

- The discovered pattern can confirm an existing theory, in which case we can think of it as an interpretation of the theory from a different perspective.
- If the discovered pattern contradicts an existing theory, this could mean a potential breakthrough and could imply that the existing theory needs to be re-examined.
- If a novel pattern is discovered, this could mean discovery of new knowledge.

In this paper, we investigate both theory-driven and data-driven approaches to solving the churn prediction problem. Theories of human behavior usually do not account for a large number of simultaneous variables. They simply are not able to be complex enough because it is difficult for the human mind to handle more than about seven factors at once [35], and perhaps fewer [36]. That is, any theory has to be both usable (which machine learning models can do at any scale) and understandable (which has an upper bound of complexity).

In our first approach, we construct 14 features based on intuition but without the constraint of size. We call this approach as data-driven. In the theory-driven approach, we take only four features derived from the theories of player motivation – three of these features are achievement-oriented and one is socialization-oriented. We run classification algorithms on these two datasets and then observe and compare results from these two approaches. Finally, we use the theory-driven approach to examine the predictive power of each of the player motivation factors.

IV. ENSEMBLE APPROACH

Ensemble methods are a classification technique in which individually trained classifiers are combined when classifying novel instances [29]. They combine several data-driven models in order to obtain a better composite global model [30]. In this paper, we have implemented an ensemble and used it for the churn prediction problem. The flowchart in Figure 1 outlines the steps involved in training the ensemble.

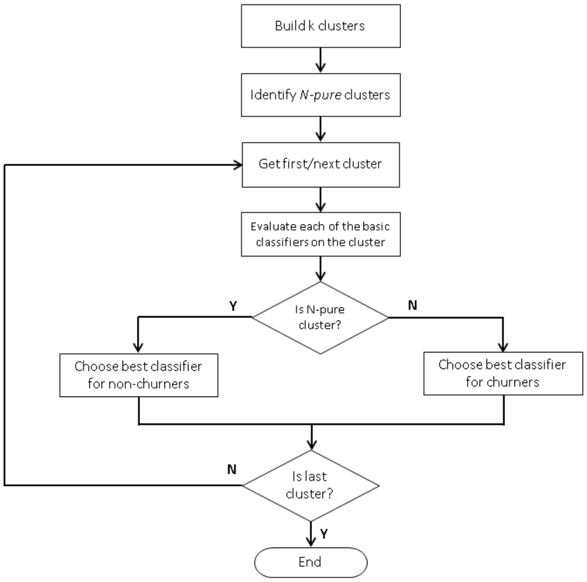
Build k clusters
Identify *N-pure* clusters
Get first/next cluster
Evaluate each of the basic classifiers on the cluster
Is N-pure cluster?
Y
N
Choose best classifier for non-churners
Choose best classifier for churners
Is last cluster?
N
Y
End

Figure 1. Flowchart for training an ensemble

In the first step, K-means algorithm is used to partition the dataset into *k* clusters. When choosing the best classifier for a cluster, we would like to keep the class distribution of the cluster in mind. If a cluster is dominated with instances of the negative class (non-churners), we would like to use a classifier which performs well on the negative class. We call such a cluster which is purer with respect to the negative class as an N-pure cluster. In the second step of building an ensemble, we identify the N-pure clusters. These are the clusters for which a) number of non-churners is greater than twice the number of churners and b) the cluster entropy is less than a threshold. We have chosen the threshold to be 0.4 (as we shall see later, Table V gives an indication of why 0.4 was used as the threshold – all the clusters with entropy value below 0.4 were dominated by instances of the negative class). Next, we evaluate several basic classifiers on each of the clusters (i.e. segments of the dataset generated by K-means). In case of N-pure clusters, we identify the classifier which performs best on non-churners as the designated classifier for that cluster. For other clusters, we choose the classifier which performs best on the positive class. At the end of the process, we have a classifier associated with each of the clusters.

Once the ensemble is built, we classify a new instance in the following way: First, we find the nearest cluster to which the instance belongs. We then use the classifier associated with that cluster to classify the instance.

V. Experiments and Analysis

A. *Experimental setup*

For the experiments, we have used player activity logs from Sony Everquest II for the time period February to June, 2006. The activity logs collect an entire range of the online activities of players such as individual and group quests, monster kills, player deaths, in-game trade activities, spells cast, failures by individuals or groups, player vs. player interactions. The dataset consists of 7,891 instances of churners (the positive class) and 8,578 instances of non-churners (the negative class).

Churners are identified as players who cancelled their subscription and did not subsequently return to the game. In addition, we also consider players for whom no activity has been observed in the two months prior to the date of analysis as churners. This is a meaningful addition to the hard subscription-based definition since extended periods of inactivity are indicative of player disinterest and churn likelihood. Consequently, for this dataset, churners are the players who have cancelled their subscriptions and stopped playing the game and/or those with no recorded activity in the months of May and June, 2006. A player session is defined as a contiguous period of player activity. Since the activity logs only record player actions, we had to define player-sessions using a simple heuristic. A session consists of sets of activities which are separated by no more than 30 minutes. Using this definition, we calculate both session lengths and inter-session lengths for a player. A single player can play the game with multiple characters or roles and we identify the primary character as the one which has the highest playtime (sum of all session lengths) among all of the player's characters.

The features used in the theory-driven approach are described below along with their information-gain (IG) value. Information gain measures how well a given attribute separates the training examples according to the target classification and is given by the expected reduction in entropy caused by partitioning the examples according to the attribute [31]. All the features are calculated for the player's primary character.

Achievement-oriented features

- **A1**-Rate of quest participation (IG: 0.1306)
- **A2**-Rate of monster kills (IG: 0.0491)
- **A3**-Rate of gaining experience points (IG: 0.0385)

Socialization-oriented feature

- **S1**-Rate of group interactions (IG: 0.0588)

The motivation behind choosing rate is that it would be a stronger indicator of the intensity of engagement in an activity than just the absolute value. For example, say A and B are two players both of whom gain a total of 1,000 experience points over the entire duration from February to June. If A acquires those experience points in a single month and B does it over a period of three months, one could say that A is more engaged in the game and that would be reflected by the A's higher rate of gaining experience points as compared to B.

The following is a list of features for the data-driven approach. Except **D9**, all the other features are calculated for the player's primary character. The features are listed in decreasing order of the information gain (IG) value.

- **D1**-Total session length (IG: 0.2077); this is the effective playtime over the entire duration.
- **D2**-Total experience points gained.(IG: 0.1462)
- **D3**-Number of quests participated in (IG: 0.1384)
- **D4**-Number of monster kills (IG: 0.1343)
- **D5**-Number of deaths (IG: 0.1259)
- **D6**-Number of times the primary character was part of a group activity, such as group quests (IG: 0.1162).
- **D7**-Total inter-session length (IG: 0.1147)
- **D8**-Number of other characters interacted with (IG: 0.1138)
- **D9**-Total number of characters controlled by the player account (IG: 0.1107).
- **D10**-Average character level (IG: 0.1002). As players complete quests and gain experience points, their character level goes up. This feature is a measure of the primary character's average level for the duration.
- **D11**-Number of levels advanced during that time period (IG: 0.0999)
- **D12**-Number of churners the character has interacted with (IG: 0.0404).
- **D13**-Out of all the interactions with other players, what percent of it was with churners (IG: 0.0173)
- **D14**-Out of all the total experience points acquired in group activities, what percent of it was acquired with churners (IG: 0.015)

In order to better understand the dataset and distribution of the two populations, we standardized the 14 data-driven attributes and performed PCA (principle component analysis). A PCA plot along the first two principal components (that corresponds to the two variables capturing most the variance in the data or the two most important uncorrelated new features transformed from the original feature space) is presented in Figure 2 in order to explore and visualize the international structure of the data. Churners are represented by the red 1s and non-churners by the blue 0s. As evident from the PCA plot, the two classes are not well-separated (even in a feature space where features are transformed so that they best explain the variance in the data) which makes this a non-trivial classification task. The plot also highlights the fact that, as the instances of the two classes are not well separated, density-based techniques such as one-class SVM would probably not perform well on this dataset. A one-class SVM finds a boundary that separates volume of high density from volumes of low density [27] and uses that boundary for identifying outliers.

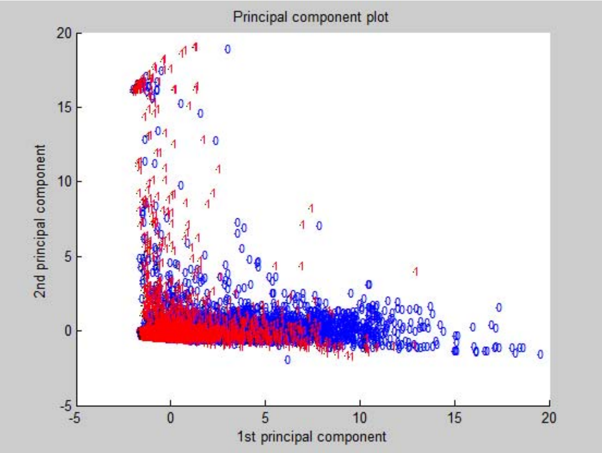

Figure 2. PCA plot of churners and non-churners

B. *Experiment results and analysis*

1) *Experiment 1-* The purpose of this experiment is to do a comparison between the data-driven and theory-driven models. Table II shows the 10-fold cross-validation results given by a C4.5 decision tree classifier on Weka [33] (J48 is the implementation of C4.5 on Weka) for different feature combinations. In the table, precision, recall, and F-meausre are given in percentage. We have used a decision tree for this experiment because we intend to investigate model complexity that could be represented by the size of a tree. Please note in Table II a tree size of X/Y indicates the total number of nodes in the tree is Y with X leaf nodes.

TABLE II. J48 PERFORMANCE FOR DIFFERENT FEATURE SETS

Feature sets	***Tree size***	***Precision (%)***	***Recall (%)***	***F-measure (%)***
D1..D14	*243/485*	*69.3*	*84*	*76*
A1,A2,A3,S1	*30/59*	*67.1*	*76.2*	*71.3*
A1,A2,A3	*19/37*	*67.2*	*73.2*	*70.1*
S1	*4/7*	*64.3*	*55.4*	*59.5*

Below are the main observations from this experiment:

a) Comparison of data-driven and theory-driven: The first row in Table II shows the results for the 14 data-driven features. This feature set produces the best results (F-measure=76) but the model complexity is also the highest, as evident from a tree size of 243/485. The second row shows the results for the 4 theory-driven features – here, we see a 4.7% drop in F-measure to 71.3. However, the model complexity is substantially reduced (tree size 30/59). Thus we observe that even though the theory-driven model is not as accurate as the data-driven one, the theory-driven model itself can be more interpretable to the domain experts and hence, more preferable over a complex data-driven model.

In keeping with the arguments put forth in [34], we have discovered a more accurate and complex model, extracted a more comprehensible approximation to it (based on domain knowledge) and using the Occam's razor argument we are favoring the simpler (more comprehensible) model because simplicity is a goal in itself.

b) Impact of different motivational factors: The last two rows of table II indicate the individual discriminating power of the two motivational factors (achievement and socialization). An interesting observation here is that achievement-oriented features alone produce an F-measure of 70.1 whereas achievement and socialization features taken together gives an F-measure of 71.3. This is only a 1.2% increase – so, we can conclude that the theory-driven model is dominated by achievement-orientation.

c) Improvement over previous results : Kawale et al have proposed a churn prediction model based on social influence among players and their personal engagement in the game [19]. They get a precision and recall values of 50.1 and 29.8 respectively on the same dataset as used in this paper [19]. Using new theory-driven features and an updated definition of churners, our best prediction model gives a precision score of 69.3 and recall score of 84 (refer Table II) – an increase of 38.3% in precision along with an increase of 181.9% in recall and thus, significantly outperforming the modified diffusion model from reference [19].

2) *Experiment 2-* In this experiment, we generate and compare lift curves for the data-driven and theory-driven models. A lift curve is an important tool for direct marketing when a subset of customers are to be contacted [32]. The lift is a measure of a predictive model calculated as the ratio between the results obtained with and without the predictive model [22]. The steps involved in generating the lift curve is described in the next section.

We first divide the original dataset into training and test instances in the ratio of 2:1 (i.e. 66.7% for training and 33.3% for test), while keeping intact the ratio of churners to non-churners within each set. Table III depicts breakdown of the original dataset for the experiment. Next, we train a C4.5 decision tree classifier on the training set and use the trained model to classify instances on the labeled test set. During classification, along with the predicted label, we also assign a churn probability to each instance. The classified instances are then sorted in decreasing order of their churn probabilities to generate the lift curve for the prediction model.

TABLE III. BREAKDOWN OF DATASET

	Training Set	Test Set	Full dataset	% of total
Churners	5261	2630	7891	47.91
Non-churners	5718	2860	8578	52.09
Total	10979	5490	16469	100

Figure 3 shows lift curves for the data-driven model and the theory-driven model and the lift over a random-guess model. For example, when we consider the first 20% of test instances (e.g. contact the first 20% of players), random guess would identify 20% of churners, but the data-driven model and the theory-driven model would identify 30% and 40% of churners, respectively.

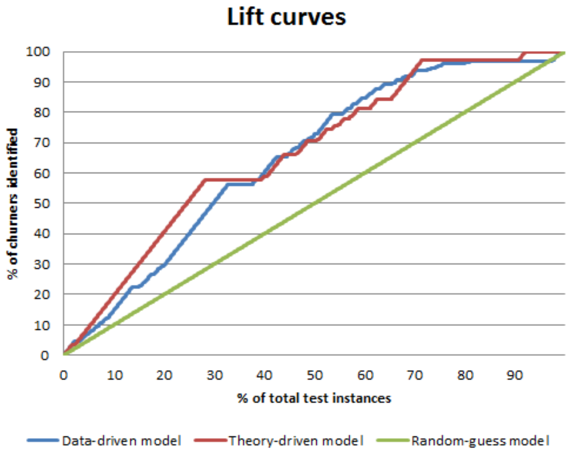

Figure 3. Lift curves for the two models

Figure 3 also presents a trade-off between when to use the theory-driven model and when to use the data-driven model. We observe that with 25% of the total test instances, we can reach 50% of the potential churners with the theory-driven model but we can only reach around 40% of the churners if we were to use the data-driven model. Thus, the theory-driven model performs better than the data-driven model for the top 25% of instances with the highest churn probabilities. On the other hand, if we consider more test instances (e.g. contact more players), we could reach slightly more potential churners using the data-driven model than the theory-driven model. When the portion of the considered test instances is between 40% and 70%, the data-driven model identifies more churners. Between 70% and 75%, the theory-driven model is slightly better; between 70% and 90%, both are the same; after 90%, the theory-driven model is better. As a result, if we are only allowed to consider a small portion of test instances (for example, if resources are limited and we are only allowed to contact a small portion of players), we could consider using the theory-driven model. On the other hand, if we are allowed to consider a larger portion, but not all, of test instances, we could consider using the data-driven model. Thus, if a marketing effort is restricted in the number of customers it can contact, the theory-driven model would offer much better return-on-investment by identifying more customers among that restricted set who have the highest probability of churn, whereas the data-driven model might be preferred if marketing resources are more abundant.

3) *Experiment 3-* In this experiment, we use the ensemble described earlier on the dataset and observe the results. As mentioned earlier, the ensemble takes as input the number of clusters *k* into which the dataset needs to be partitioned. Since we do not know the number of naturally ocuring clusters in the dataset, we can estimate this by running some preliminary experiments as described next.

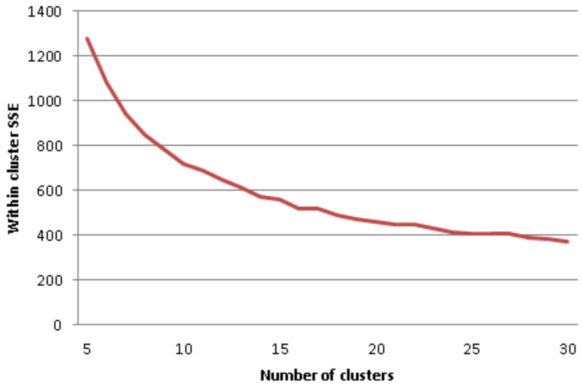

Figure 4. Within-cluster SSE vs. number of clusters

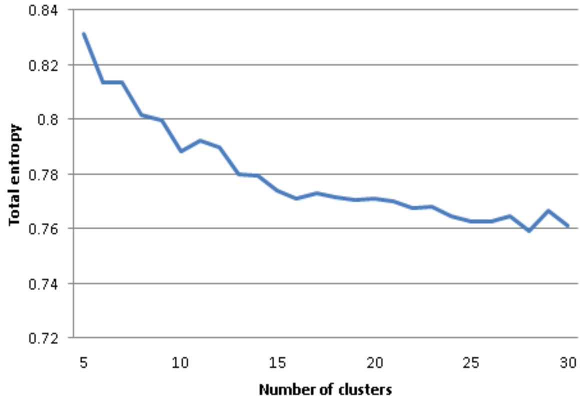

Figure 5. Total entropy vs. number of clusters

Using K-means we partition the dataset into increasing number of clusters in successive runs. Figures 4 and 5 show how the within cluster Sum of Squared Errors (SSE) and the total entropy change as the number of clusters is increased. As one might expect, both the within cluster SSE and total entropy generally decrease with an increase in the number of clusters. To get an estimate of the number of clusters to choose as input for the ensemble, one might look for the knee of the curves. For our experiments we run the ensemble with for different number of clusters as input and observe how the overall performance of the ensemble is affected (refer Table IV).

Entropy characterizes the impurity of an arbitrary collection of samples [31] and total entropy of a set of clusters is the weighted average of individual cluster entropies.

$$Cluster\ entropy, E_i = -\log_2 p - \log_2(1-p)$$

where, p = proportion of positive samples (churners)

$$Total\ entropy = \sum_{i=1}^{k} w_i E_i$$

where, k is the total number of clusters
$w_i = N_i/N$ is the fraction of samples in the ith cluster
E_i is entropy of the ith cluster

TABLE IV. ENSEMBLE CLASSIFIER RESULTS

No. of clusters	Precision	Recall	F-measure
5	66.23	91.94	76.99
10	66.49	91.08	76.87
15	67.58	89.80	77.12
20	67.6	90.13	77.25

Table IV shows the 10-fold cross-validation results of the ensemble for 5, 10, 15 and 20 clusters. For the ensemble, we have used J48 (C4.5 decision tree), JRip (Repeated Incremental Pruning to Produce Error Reduction, or RIPPER), SMO (Sequential Minimal Optimizaton) with RBF Kernel, Naïve Bayes and k-Nearest neighbor (called IBk in Weka) as base classifiers, using the 14 data-driven features. The best overall result, 77.25 shows a 1.25% improvement over the best result, 76 of a single J48 classifier (refer Table II) on the entire dataset. Though this may not be a substantial gain in overall performance, we observe a considerable improvement in terms of the recall value of the classifier. The best recall value for the ensemble classifier is 91.94 – this is a 7.94% increase over the best recall value, 84, for a single classifier (refer Table II). A high recall value is a desirable property in a churn prediction model because it means the classifier is able to identify a larger percentage of the churners correctly. In a typical churn prediction problem, failing to identify potential churners (false negatives) can be much more costly than wrongly identifying non-churner as a churner (false positives). In reality, a 7.94% increase in recall could be translated into millions of dollars in cost savings.

TABLE V. CLUSTER STRUCTURE (K=20)

Cluster#	Size	Churner: Non-churner	Entropy	Classifier
19	525	4:521	0.065	J48
15	441	4:437	0.075	J48
0	482	5:477	0.08	J48
8	179	3:176	0.123	J48
11	351	7:344	0.14	J48
12	635	18:617	0.19	J48
18	382	15:367	0.24	JRip
17	127	14:113	0.5	IBk(3)
5	525	107:418	0.73	NaiveBayes
4	470	104:366	0.76	IBk(3)
6	137	100:37	0.84	J48
2	2945	2084:861	0.87	JRip
16	363	111:252	0.89	IBk(3)
14	888	596:292	0.91	J48
1	3105	2053:1052	0.92	J48
13	1570	970:600	0.96	J48
9	991	587:404	0.98	SMO
3	359	159:200	0.99	JRip

Table V gives cluster size and distribution of classes within each cluster. The clusters are ranked in increasing order of purity, as given by cluster entropy. The highlighted clusters are *N-pure* and for these clusters we use the classifier which does best on the negative class. The last column in Table V shows the base classifier being used for that cluster.

We observe from Figure 2 (the PCA plot) that the two classes are not well-separated. As we can see in Table V that the two classes are well-separated in some clusters (with low entropy values). In addition, Table V also shows that different groups of churners or non-churners require different classification algorithms. As one example, clusters 4 and 5 are similar in size and entropy value, while 3-nearest neighbor algorithm performs the best for cluster 4 and naïve Bayes algorithm performs the best for cluster 5. As another example, clusters 9 and 3 are similar in entropy value, while sequential minimal optimization algorithm outperforms others for cluster 9 and RIPPER algorithm outperforms others for cluster 3. This implies that there may not exist a single classification algorithm able to effectively classify churners or non-churners in all groups. This also implies that different groups churners or non-churners present different behavior patterns. These two together imply that a single model built globally on the whole dataset would not be as good as a set of models built locally on clusters. This argument is supported by our experiments where we observed an improved recall score of the ensemble over a single model.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have looked at the problem of churn prediction in MMORPGs and found that player motivation is a key factor in understanding churn behavior. Our churn prediction model based on theories of player motivation significantly outperforms a diffusion-based model on the same dataset. Within the context of the problem, we have done a comparison of data-driven and theory-driven approaches and tried to give some insight into how the two approaches can work together. We found that the theory-driven model can be more interpretable to the domain experts and hence, more preferable over a complex data-driven model (which is only slightly more accurate). We have also used an ensemble approach to further improve the results of the basic model.

As future work, we will examine the interplay of the different motivational factors and investigate the feasibility of coming up with behavioral signatures of different population segments based on this analysis. We believe that such analysis will provide valuable insight and serve as a helpful tool for marketing and sales analysts. Furthermore, we will refine the ensemble approach, using other algorithms and different evaluation measures for training and choosing the base classifiers.

ACKNOWLEDGMENT

The research reported herein was supported by the AFRL via Contract No. FA8650-10-C-7010, and the ARL Network Science CTA via BBN TECH/W911NF-09-2-0053. The data used for this research was provided by the Sony Online Entertainment. We gratefully acknowledge all our sponsors. The findings presented do not in any way represent, either directly or through implication, the policies of these organizations.

REFERENCES

[1] Blizzard Press Release (Oct 7,2010) *WORLD OF WARCRAFT SUBSCRIBER BASE REACHES 12 MILLION WORLDWIDE* Retrieved January29, 2010 from http://us.blizzard.com/en-us/company/press/pressreleases.html?101007

[2] http://freetoplay.biz/2007/08/02/top-10-revenue-models-for-free-to-play-games/, Retrieved March 30, 2011

[3] A. Bagga. *The Emergence of Games As A Service. Industry Report*, ThinkEquity LLC, Sn Francisco, May 4, 2009.

[4] M. Karnstedt, T. Hennessy, J. Chan, P. Basuchowdhuri, C. Hayes, T. Strufe, *Handbook of Social Network Technologies and Applications*, Springer US, 2010, pp. 185-220

[5] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. *Social ties and their relevance to churn in mobile telecom networks*, *EDBT '08*, pages 668–677, 2008

[6] J. Ferreira, M. B. R. Vellasco, M. A. C. Pacheco, and C. R. H. Barbosa. *Data mining techniques on the evaluation of wireless churn*, *ESANN*, pages 483–488, 2004

[7] B. Huang, B. Buckley, and T. M. Kechadi. *Multi-objective feature selection by using NSGA-II for customer churn prediction in telecommunications*. *Expert Syst. Appl.*, 37(5):3638–3646,2010

[8] S.-Y. Hung, D. C. Yen, and H.-Y. Wang. *Applying data mining to telecom churn management*. *Expert Syst. Appl*, 31(3):515–524, 2006

[9] J. Lu. *Predicting customer churn in the telecommunications industry – an application of survival analysis modeling using sas,* In SAS Proceedings, SUGI 27, pages 114–127, 2002

[10] B. M. Masand, P. Datta, D. R. Mani, and B. Li. *CHAMP: A prototype for automated cellular churn prediction*. *Data Min. Knowl. Discov*, 3(2):219–225, 1999

[11] M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, and H. Kaushansky. *Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry*. IEEE Transactions on Neural Networks, 11(3):690–696, 2000

[12] C.-F. Tsai and Y.-H. Lu. *Customer churn prediction by hybrid neural networks*. *Expert Syst. Appl.*, 36(10):12547–12553, 2009

[13] Y. Zhao, B. Li, X. Li, W. Liu, S. Ren, *Customer Churn Prediction Using Improved One-Class Support Vector Machine*, Advanced Data Mining and Applications, Springer Berlin / Heidelberg, vol. 3584/2005, pp. 731, 2005

[14] Y. Xie, X. Li, E.W. T. Ngai, and W. Ying. *Customer churn prediction using improved balanced random forests*. *Expert Syst. Appl.*, 36(3):5445–5449, 2009

[15] K. Coussement and D. V. den Poel. *Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques*. *Expert Syst. Appl*, 34(1):313–327, 2008

[16] B. Q. Huang,M. T. Kechadi, and B. Buckley. *Customer churn prediction for broadband internet services*. In DaWaK, volume 5691 of Lecture Notes in Computer Science, pages 229–243, Springer, Berlin, 2009

[17] G. Nie, G. Wang, P. Zhang, Y. Tian, and Y. Shi. *Finding the hidden pattern of credit card holder's churn: A case of china.* In *ICCS '09*, pages 561–569, Springer, Heidelberg, 2009

[18] O. Herrera and T. Znati. *Modeling churn in P2P networks*. In *Annual Simulation Symposium*, pages 33–40. IEEE Computer Society, 2007

[19] J. Kawale, A. Pal, J. Srivastava, *Churn Prediction in MMORPGs: A Social Influence Based Approach*, Proceedings of the 2009 IEEE Social Computing (SocialCom-09). Symposium on Social Intelligence and Networking (SIN-09). Vancouver, Canada, August 29-31, 2009.

[20] K. Morik and H. Kpcke. *Analysing customer churn in insurance data – a case study*. In PKDD '04, pages 325–336, 2004

[21] G. Nie, G. Wang, P. Zhang, Y. Tian, and Y. Shi. *Finding the hidden pattern of credit card holder's churn: A case of china.* In *ICCS '09*, pages 561–569, Springer, Heidelberg, 2009

[22] T. Mutanen. *Customer churn analysis - a case study*. Technical report, Helsinki University of Technology, System Analysis Laboratory, 2006

[23] H. Hwang, T. Jung, and E. Suh. *An ltv model and customer segmentation based on customer value: a case study on the wireless telecommunication industry*. Expert Syst. Appl., 26(2):181–188, 2004

[24] R. Bartle, *Hearts, clubs, diamonds, spades: Players who suit MUDs*. Journal of MUD Research, 1 (1), 1996. Retrieved December 9, 2010, from http://www.mud.co.uk/richard/hcds.htm

[25] N. Yee, *A model of player motivations,* March 15, 2005. Retrieved December 9, 2010, from http://www.nickyee.com/daedalus/archives/001298.php?page=4

[26] N. Yee, *Motivations of play in online games*, CyperberPsychology and Behavior, 2006.

[27] D. Tax, R. Duin, *Support vector domain description*, Pattern Recognition Letters 20: 1191-1199, 1999

[28] J. Srivastava, IARPA VWE presentation, University of Minnesota, September, 2010, unpublished

[29] D. Opitz, R. Maclin, *Popular ensemble methods: an empirical study*, Journal of Artificial Intelligence Research 11 (1999) 169-198

[30] J. Ghosh. *Multiclassifier systems: Back to the future*. Keynote Talk, 3rd Int'l Work-shop on Multiple Classifier Systems, Cagliari, June, 2002a

[31] T. M. Mitchell, *Machine Learning,* The Mc-Graw-Hill Companies, Inc., 1997

[32] C. X. Ling, C. Li, *Data Mining for Direct Marketing: Problems and Solutions*. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, pp. 73-79, (1998).

[33] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); *The WEKA Data Mining Software*: An Update; SIGKDD Explorations, Volume 11, Issue1.

[34] P. Domingos: *The Role of Occam's Razor in Knowledge Discovery*. Data Min. Knowl. Discov. 3(4): 409-425 (1999)

[35] G. A. Miller. *The magical number seven, plus or minus two: some limits on our capacity for processing information*. Psychological Review 63 (2): 81–97, 1956.

[36] J. Farrington. *Seven plus or minus two*. Performance Improvement Quarterly 23 (4): 113–6, 2011

[37] M. Karnstedt, M. Rowe, J. Chan, H. Alani, C. Hayes. The Effect of User Features on Churn in Social Networks. Proceedings of the ACM WebSci, 2011