Proceedings, The Eleventh AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-15)

Predicting Purchase Decisions in Mobile Free-to-Play Games

Rafet Sifa
Fraunhofer IAIS
Sankt Augustin, Germany
rafet.sifa@iais.fraunhofer.de

Fabian Hadiji
TU Dortmund, goedle.io
Dortmund, Germany
fabian@goedle.io

Julian Runge
Wooga GmbH
Berlin, Germany
julian.runge@wooga.com

Anders Drachen
Aalborg University
Copenhagen, Denmark
drachen@hum.aau.dk

Kristian Kersting
TU Dortmund
Dortmund, Germany
kristian.kersting@cs.tu-dortmund.de

Christian Bauckhage
Fraunhofer IAIS
Sankt Augustin, Germany
christian.bauckhage@iais.fraunhofer.de

Abstract

Mobile digital games are dominantly released under the freemium business model, but only a small fraction of the players makes any purchases. The ability to predict who will make a purchase enables optimization of marketing efforts, and tailoring customer relationship management to the specific user's profile. Here this challenge is addressed via two models for predicting purchasing players, using a 100,000 player dataset: 1) A classification model focused on predicting whether a purchase will occur or not. 2) a regression model focused on predicting the number of purchases a user will make. Both models are presented within a decision and regression tree framework for building rules that are actionable by companies. To the best of our knowledge, this is the first study investigating purchase decisions in freemium mobile products from a user behavior perspective and adopting behavior-driven learning approaches to this problem.

Introduction

Free-to-Play (F2P) games is the formulation of the freemium business model for games: F2P games can be downloaded, installed and played for free, but players have the option to spend money on In-App Purchases (IAPs) like virtual currency or temporary boosts; or for example to eliminate the presence of advertisements, a smaller source of revenue for F2P games (El-Nasr, M.S. and Drachen, A. and Canossa, A. 2013; Fields and Cotton 2011; Luton 2013; Seufert 2014). In F2Ps, there is typically an imbalance in between players that spend money, which we will here refer to as **premium players**, and those who do not, here referred to as **non-spending players** (referring to financial value of players only), and which comprise the vast majority of F2P players (Swrve 2014; Lim 2012; Nozhnin 2012). The problems associated with imbalances in spending among customers is not unique to games (Chawla et al. 2002), nor is the need for predictions of user behavior (e.g. (Gupta, Lehmann, and Stuart 2004; Malthouse and Blattberg 2005)). However, F2P mobile games comprise a unique situation different from off-line retail situations or even subscription-based games, and therefore we cannot make assumptions about spending in F2P games based on knowledge from other contexts. Notably, F2Ps generate revenue and are played differently than MMOs. Also, because F2Ps follow the freemium model, the life cycle is different from non-fremium situations. Also, F2P games is the first time players face the same game before and after a purchase. Downloading a F2P is not the same commitment as buying in advance, and this is demonstrated by the extreme imbalances mentioned.

A key challenge in F2P mobile games is to convert players from non-spending to premium players, and by extension ensuring that the **Life-time Value** (LTV; also referred to as Customer Lifetime Value, CLV) of players is higher than the **User Acquisition Cost** (UAC). UAC is high and rising for mobile games. For example, Supercell[1] notes an average 4.36 USD cost per install on mobile platforms, an increase of 288% over 2 years. The problem is echoed by e.g. (Luton 2013; Seufert 2014; Swrve 2014; Takahashi 2014).

Given the imbalance, combined with rising UACs, the ability to detect, define and predict behavioral attributes of players as early as possible becomes an important factor for success in mobile market (Hadiji et al. 2014; Luton 2013; Runge et al. 2014). Predicting attributes of players is valuable both in terms of cutting costs and increasing revenue, and relates directly to the problem of LTV prediction. Predicting premium players is the first step in building LTV models. Predicting number of purchases the second, and value the third. The first two problems are here targeted experimentally: 1) Classification task: algorithms include Decision Trees (DT) (Quinlan 1996), Random Forests (RF) (Breiman 2001) and Support Vector Machines (Cortes and Vapnik 1995) are utilized, with Random Forests providing the best results. 2) Regression task: Poisson trees are learned for this task based on three different observation periods: 1, 3 and 7 days respectively. There is no current standard but

[1] http://www.superdataresearch.com/market-data/digital-game-user-acquisition-cost/

7 days appears to be broadly used in the F2P game industry (El-Nasr, M.S. and Drachen, A. and Canossa, A. 2013; Seufert 2014; Fields and Cotton 2011; Luton 2013; Nozhnin 2012). Both models are trained on generic behavioral and economic features.

We are adopting methods that are fast to implement, scalable and rely on accessible ML algorithms, in order to keep results relevant for the game industry. Compared to e.g. HMMs or NNs, decision trees allow more readable rules which can guide design and are favorable over more recent algorithms which are often not programmed for productive usage. Work such as (Wagner, Benlian, and Hess 2014; Lehdonvirta 2009) has explored freemium services but focus on products (e.g. music tracks) rather than user behavior. To the best of our knowledge, we are the first to take a customer-centric behavioral look at purchasing in freemium products, not the least F2P games. The insights presented are a first step in enabling companies to strengthen revenue streams from players through smart CRM and to increase engagement and entertainment in F2P mobile games through tailored game experiences.

Related work

While the publicly available prediction work in commercial games remains limited (Bauckhage et al. 2012; Drachen et al. 2013; Sifa et al. 2013), work in associated domains have a more established history, but due to space restrictions we here focus on key related work.

Churn prediction is a topic of interest in a variety of disciplines, e.g. insurance (Morik and Köpcke 2004) and retail banking (Mutanen, Ahola, and Nousiainen 2006). One of the earliest successful churn models was presented by (Mozer et al. 2000), in the area of wireless communication. It is a recent topic in F2P games: Hadiji et al. (Hadiji et al. 2014) formally defined the churn problem in games, and identified a range of behavioral features useful for prediction, as well as game-agnostic classifiers. (Runge et al. 2014; Rothenbuehler et al. 2015) investigated churn prediction in F2P games using a binary classification approach, via Hidden Markov Models to address temporal dynamics, and benchmarked several methods to predict churn.

Outside mobile games, Weber et al. (Weber et al. 2011) focused on retention modeling in Madden NFL 11. In MMORPGs, Kawale et al. (Kawale and Srivastava 2009) investigated churn in the MMORPG EverQuest II using social network analysis as the basis, proposing a churn model based on social influence among players. Nozhnin (Nozhnin 2012) focused on the first few minutes of gameplay in the MMORPG Aion, investigating triggers for churn; highlighting the challenge of feature selection. (Sifa, Ojeda, and Bauckhage 2015) investigated a generalized version of churn based on migration of users between products using a dedicated matrix factorization model. On the topic of IAP purchasing behavior in digital games Lehdonvirta (Lehdonvirta 2009) investigated which attributes of virtual items drive purchasing. (Sifa and Drachen 2014) modeled player engagement using lifetime analysis.

Outside games, the Pareto/NBD approach (Schmittlein, Morrison, and Colombo 1987) is commonly used to estimate the number of future purchases of a customer and then combined with the average expected profitability. Our regression model can be used similar to the Pareto/NBD for estimating LTV by combining it with estimates of the future purchase values. Gupta et al. (Gupta, Lehmann, and Stuart 2004) established that increasing LTV is key for financial performance of firms. Malthouse and Blattberg (Malthouse and Blattberg 2005) delved into the feasibility of future value estimation for customers from different industries. Borle et al. (Borle, Singh, and Jain 2008) used data from a membership-based direct marketing company, concluding that longer inter-purchase times correlate with larger purchase amounts and higher churn risk. Drawing on this we incorporate inter-purchase times as a feature.

Feature Type	Descriptor(s)
Telemetry	Country
	Device
	Move Count
	Active Opponents
	Logins & Game rounds
	Skill-1,2,3
	Reached Goals
	World Number
	Number of Interactions
	Number of Purchases
	Amount Spent
	Playtime
Specific	Last Inter-session Time
	Last Inter-login Time
	Inter-login time distribution
	Inter-session time distribution
Composite	Correlation on time*
	Mean and Deviation on Time*
	Country Segments

* Calculated for session-wise distributions of features in *Telemetry* and *Specific*

Table 1: Dataset description

Dataset

The work presented here builds on detailed (low-level) tracking data of over 100,000 players from a mobile free-to-play (F2P) puzzle game developed and published by Wooga. Title is kept confidential due to the purchase information in the data. We extracted a random sample from a week of new installs of this game and then follow these players through the game for 30 days. We observe their in-game behavior along three dimensions, logins, game rounds and purchases. A login is registered whenever a player starts the game client on their mobile device. This is required for the player to start a game round. There can be any number of game rounds for a login, including none. A purchase is registered when a player spends real money for an in-game purchase. In this game the only item on sale is the in-game currency that can then be spent in the game to have upgrades, boosts and longer game sessions. For our empirical analysis we merge the information from different tables to identify patterns and predict premium players. Tbl. 1 describes the processed dataset used for training and testing the prediction models. Furthermore, together with pure behavioral teleme-

try data, we incorporated specific and composite features to identify particular player behaviors for classification and regression. The former is introduced to capture the latest occurrence and the distribution of time between logins and sessions, whereas, the latter include mean, deviation and correlation in time to capture the evolution of other features and a feature for socio-demographic category of player's country.

Classification Task

Similar to previous studies on churn analysis in mobile and social games as in (Hadiji et al. 2014; Runge et al. 2014), we begin with approaching the problem of understanding whether the player will become a premium user or not as a binary classification task. Given accumulated and series based history of player activities from $t \in \mathbb{N}$ days, our main goal is to predict whether the player is going to purchase an in-game item in the future. This approach has the advantage of providing insights regarding the game design, promotions and advertising for a better player experience and resource allocation. This can be formally represented as learning a function $f : H \rightarrow D$ that maps from the space of activities, denoted by H, to a binary space, denoted by D, for purchasing or not purchasing. In order to realize this mapping, we use classification algorithms from machine learning that include Decision Trees (Quinlan 1996), Random Forests (Breiman 2001), and Support Vector Machines (Cortes and Vapnik 1995). In the next section, we describe how trees can be used for regression while keeping the same domain as above (H), whereas, we change the target to numeric values that correspond to the number of purchases.

Considering the specific task of predicting future player purchases, we observe a skewed distribution towards non-purchasing behavior. That is, the percentage of future premium players lies below 2% and this produces a challenge when we are inclined to detect entities from this minority class. In order to tackle the problem of imbalanced datasets, we synthetically generated premium players following *Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC)* which is proposed by (Chawla et al. 2002). We chose SMOTE-NC since it is well established, automatically handles mixed data types and can be implemented easily and scalably since it is based on finding k-nearest neighbors. Given a desired rate $r \in \mathbb{R}$ and number of nearest neighbors $k \in \mathbb{N}$, SMOTE-NC generates synthetic data entities to populate the dataset with minority entities so that the ratio between number of minority and majority instances becomes r. This is performed iteratively by selecting a random minority entity to be used as part of the active entities. After that for each active entity a neighbor that also belongs to the minority class and among the top closest k neighbors of the active entity is randomly selected. Followed by that, a perturbed synthetic entity is created between the active entity and its neighbor. The main aim here is to help the classifiers to approximate the decision boundaries more accurately. It is important to note that, the perturbations only occur in numerical attributes whereas the synthetic entity inherits the active entity's nominal attributes. However, the distance measure defined for finding the top k entities punishes different valued nominal attributes by adding the squared median distance to the euclidean distance between the nominal attributes.

Learning Algorithm	Acc_	Recall	Precision	G-Mean	F-Score
RF & SMOTE-NC	**0.999**	**0.110**	**0.654**	**0.331**	**0.188**
DT & SMOTE-NC	0.981	**0.174**	0.127	**0.413**	**0.147**
RF	**1.000**	0.077	**0.750**	0.278	**0.140**
SVM Poly-Kernel	0.999	0.065	0.455	0.254	0.113
SVM RBF-Kernel	0.796	**0.574**	0.042	**0.676**	0.079
DT	**1.000**	0.006	**1.000**	0.080	0.013

(a) One-day Window of Observation

Learning Algorithm	Acc_	Recall	Precision	G-Mean	F-Score
RF & SMOTE-NC	**0.998**	**0.265**	**0.696**	**0.515**	**0.384**
DT	**0.998**	0.245	**0.679**	0.494	**0.360**
RF	**0.999**	0.211	**0.738**	0.459	**0.328**
DT & SMOTE-NC	0.979	**0.299**	0.179	**0.541**	0.224
SVM Poly-Kernel	0.995	0.184	0.375	0.428	0.247
SVM RBF-Kernel	0.821	**0.578**	0.046	**0.689**	0.085

(b) Three-day Window of Observation

Learning Algorithm	Acc_	Recall	Precision	G-Mean	F-Score
RF & SMOTE-NC	**0.997**	**0.439**	**0.643**	**0.662**	**0.522**
DT	**0.997**	**0.317**	**0.600**	**0.562**	**0.415**
RF	**0.998**	0.285	**0.700**	0.533	**0.405**
DT & SMOTE-NC	0.983	**0.472**	0.252	**0.681**	0.329
SVM Poly-Kernel	0.994	0.252	0.360	0.501	0.297
SVM RBF-Kernel	0.997	0.195	0.414	0.441	0.265

(c) Seven-day Window of Observation

Table 2: Binary classification of premium purchase behavior considering one-, three- and seven-day windows of observation (top-3 results for each category are highlighted and precision for floating point is kept at three). It is important to note that the negative accuracy does not change throughout the settings whereas the detection of premium users becomes better with more observations available. In all of the experiments, the combination of random forests and SMOTE-NC yielded substantially better results in all of the categories.

With this overview of a method to handle learning tasks in imbalanced datasets, we move to the presentation of our experimental results. We evaluate our findings in terms of being able to predict the premium users using the F-Score (also know as the F1-Measure) which is the harmonic mean over the precision and recall values for predicting the premium users. Additionally, we use the geometric mean of the classification accuracy values for each of the classes, which is known as G-mean in the literature, (Kubat, Holte, and Matwin 1997), to give equal weights to correctly classifying both classes. Since it is needed for the G-mean calculation and to show the straightforwardness of the task, we present the prediction accuracy values for the non-purchasing users as well. We trained decision trees with reduced error pruning and random forests by selecting the square root of number of possible features and 151 trees for both, normal and over-sampled, datasets for consistency. Both of the methods used Gini Index as their node impurity measure. Furthermore, we trained support vector machines (SVMs) with polynomial (Poly) and radial basis function (RBF) kernels by utilizing grid search over their parameters. We evaluate our methods

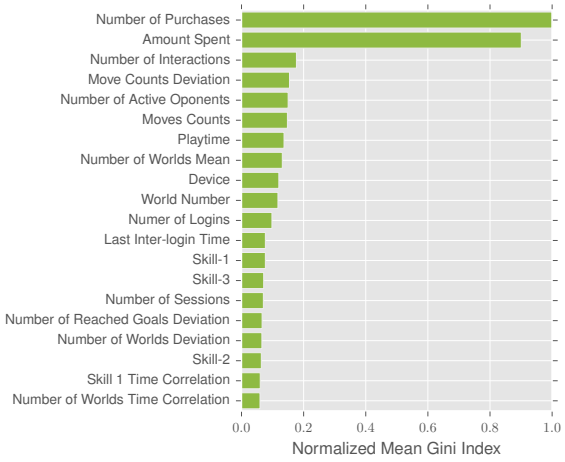

Figure 1: Normalized variable importance results in terms of mean Gini Impurity Index for predicting players' future status of purchasing using random forest algorithm. The results indicate that previous exposure to purchasing is a strong indicator of purchase activity in the future. Additionally in game interactions and activity related features and total playtime spent have significant indication of determining future purchase behavior. The results also show the value of having incorporated composite and user specific features to the classification task which capture user based events and describe the dynamics of the particular features.

using ten-fold cross validation over a 10,000 players sample and use the measures that are explained in the previous section for evaluation. The overall prediction performance results of the models for three types of observation window settings (one-, three- and seven-day) are shown in Tbl. 2.

Analyzing the results, we observe that tree based models, including models trained with and without synthetic oversampling, mostly outperformed SVMs in terms of G-Mean and F-Measure. Additionally, the more days we have available as our prior, the better the classification results which may be interpreted as a result of high class overlap between premium and non-purchasing users in the early days. Comparing the tree based learning techniques, random forests that are trained with synthetically oversampled datasets perform best with respect to predicting if the player is going to purchase in the future.

In terms of G-Mean values, we observe the best results for the models with SMOTE-NC where the recall (or the positive accuracy) values increase comparably to other methods trained with the pure data. It is important to note that synthetic oversampling has not always improved the categories for the models. In two of our experiments for single tree learning, SMOTE-NC worsens the generalization in terms of specificity. On the other hand, it has particularly improved the ensemble tree learning for predicting the future purchase behavior by increasing the recall values. This shows that for particular dimensions, SMOTE-NC does improve decision boundaries, hence, it provides a better approximation of the real decision boundary by extending the learned concept. Additionally, having obtained better generalization for minority class prediction through SMOTE-NC and random forests is an indication that for some dimensions, SMOTE-NC might create class overlaps which can be avoided by the random dimensionality sampling and the greedy heterogeneity based variable selection provided by the random forest algorithm.

A major advantage of using random forests in classification tasks is the *variable importance measure* that yields a comparable approach to weight the overall average value of the features in the classification task. Namely, for each feature the importance values indicate how much heterogeneity would be lost when a certain feature *was not used* in the classification. This is calculated by permuting the particular variable and obtaining the heterogeneity from the data instances that have not been used when building the model. Fig. 1 shows 20 variables with the highest normalized variable importance measures for the random forest algorithm (with the best F-Score) in terms of Gini Impurity Index. We observe from these results a high weight for previous exposure to purchasing. Namely, *if a player has previously bought an in-game item, they will be likely to purchase in the future as well*.

Moreover, we justify this finding in the following section when we estimate the number of future purchases using regression trees. Analyzing other highly ranked features, we observe social and game activity related features such as *Number of Interactions (with other players)*, *Move Counts* and *World Number*. Together with *Device* and *Playtime*, these are crucial factors for purchase activity detection. Furthermore, it is important to note the relative importance of the specific and composite time relevant features. Similar to its effect to players' churn behavior (Hadiji et al. 2014), *time dependency* is indeed a decisive factor for future purchasing.

Regression Task

Looking at the problem from the classification point of view make sense, in order to obtain an initial categorization of the players. However, ultimately we are not only interested in the distinction of premium players from non-spenders, but also adding a qualitative dimension to this question. To be more precise, we wish to predict the number of purchases for a more valuable LTV prediction. If incentives are given to the users, the value of the incentive should go along with the expected future return of the player. Therefore, we extend our model to the regression point of view, i.e., learning a model that predicts the average number of future purchases. We use the same features which have been used in the previous section but change our target variable from the binary to integral range.

The most straightforward model for the expected number of future purchases of a player is the mean of this quantity in the training dataset. To evaluate this simple baseline, we created a dataset consisting of 100,00 players and ran a ten-fold cross validation. Looking at the average over the folds, the expected number of future purchases after an observational period of one day is 0.057 per player. If we increase the number of observed days, the value decreases to 0.050 for three days of observations and to 0.037 for seven days. The results

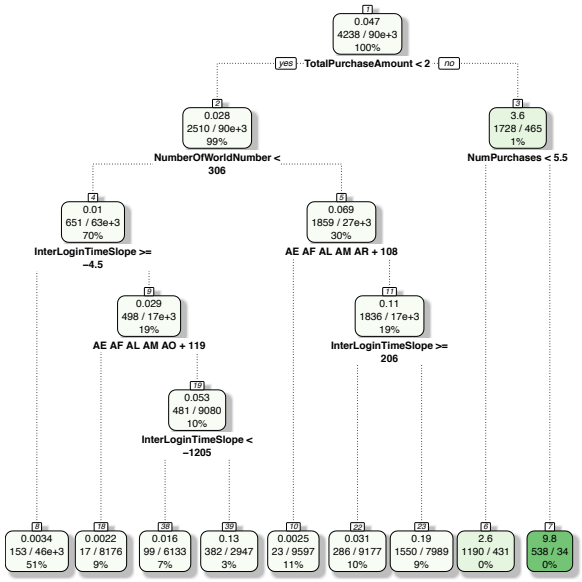

Figure 2: An example PRT learned for the task of predicting the future number of purchases of a player based on the a three day observation period. The color shades of the nodes indicate the value of the partition, i.e., darker shades correspond to higher expected values of the target variable.

using this approach for the prediction of the future number of purchases are shown in the column "baseline" in Tbl. 3.

We can do significantly better and apply machine learning techniques to learn a model with a lower error and more insights into groups of players. It is important to highlight at this point that we are interested in predicting a count value, i.e., a non-negative integral value, and not an arbitrary continuous value. Therefore, the Gaussian error assumption, which is assumed in ordinary least squares regression, is not appropriate. Predicting the future number of purchases of a player requires us to make a prediction over the natural numbers, i.e., $y \in \mathbb{N}$. Therefore, we suggest to learn a *Poisson Regression Tree (PRT)* and assume a Poisson distribution of the purchases.

We use PRTs in the following instead of other regression techniques because trees have shown in the previous section to work well in our problem setting. Nevertheless, we will below also compare our obtained results to Poisson regression. Similar to decision trees, regression trees are easy to interpret and learning is remarkably fast with existing implementations in various programing languages. We will now briefly explain how learning is done for PRTs and then present the results obtained on our dataset.

Given our dataset with a player's sessions and purchases, we construct a training dataset akin to the one used for the classification task. We feed this dataset to a Poisson regression tree learner which partitions the data based on the a likelihood ratio test (Therneau, Atkinson, and Ripley 2011). The result is a binary tree such as the one depicted in Fig. 2.

Given a new player and their observational feature vector of the first t days, we now use the PRT to obtain a prediction

Observation period	Baseline	PRT	F-Score
One day	0.91	0.89^*	0.17
Three days	0.86	0.80^*	0.35
Seven days	0.62	0.54^*	0.41

Table 3: RMSE for the predicted number of future purchases for different observation periods and F-Score values for classification based on the predicted mean. All results are averaged over ten folds. A "*" denotes that PRTs are significantly better than the baseline.

for that player by evaluating the tree. The prediction assumes a Poisson distribution of the data corresponding to the subset in the leaf and the mode of this distribution now determines the most likely number of future purchases by that player.

Using PRTs on the same datasets as the baseline above, we obtain an RMSE as depicted in the column "PRT" in Tbl. 3. For a single day of observations, the reduction in error size is small. However, this is a very challenging task as most players look identical. Having three days of observations at hand, the error is reduced by 7.89% compared to the naive baseline. The error is reduced to an even larger extend for seven days of observations. But even more interesting, we can now easily group the players according to their expected number of purchases by analyzing the trees. The trees learned for the different folds look quite similar. For three days of observations, each tree has about nine leaves on average if we use the default parameters of the tree learner and do not limit the depth of the trees artificially. The first splitting criterion is always "Amount Spent". This observation aligns nicely with the results depicted in Fig. 1 where this feature had the second highest variable importance. Even if this insight is not surprising on the first sight, it verifies common sense and the findings from the previous section: *players that have bought virtual goods in the first three days are more likely to consume above average in the future as well.*

Another interesting observation is the occurrence of the "Country" feature in the trees. Generally, the country feature is problematic because it has almost 200 different classes. Therefore, we created an additional feature "Country Segment" which was supposed to reduce the number of classes and thereby simplifying the training of the trees. However, instead of choosing this pre-clustered feature, the tree learner decided to group the countries on its own and hence created a new partitioning of the countries. For example, we have a segment "Oceania" consisting of Australia and New Zealand. However, only Australia follows the right path leading to higher means in the regression tree. We can find similar examples such as the Benelux countries, where only Belgium contributes premium players in our training database for the fold depicted in Fig. 2.

Looking at the distribution of all players in the leaves, we can observe that more than half of the players can essentially be neglected because it is very unlikely that they will ever purchase any virtual goods. We can probably also pay less consideration to the players at the upper end of the scale with a high number of expected purchases. These players might spend money in the game without further actions anyhow.

However, this is a tiny fraction of all players. An interesting group of players consists of the players with an expected number of purchases between 0.13 and 0.19 in Fig. 2 (leaves 39 and 23). These clusters contain a significant number of players who potentially turn into premium players. We have seen how PRTs provide us with improved estimates of the expected number of future purchases. Furthermore, we have also shown that the learned trees are easy to interpret and give interesting insights about player's purchase behavior. For a qualitative comparison of the obtained results, however, we also implemented a simple Poisson regression (McCullagh and Nelder 1989) with the same features. Poisson regression is akin to linear regression but defines the mean based on a log-linear model. The experiments showed that the "Country" feature with its almost 200 classes presents a problem to the weight learning. The feature did not only result in overfitting in certain cases but additionally resulted in issues with out-of-dictionary values for the country feature. Therefore, we learned a model without this feature and solely relied on our predefined country segmentation. Nevertheless, we still encountered problems during the prediction. The prediction overshoots for several players and hence, the RMSE is much higher. Averaged over all ten folds, we obtain an RMSE of 4.23 if the observation period is three days. The high error value is mainly due to one fold which has an extreme error. Removing this fold results in an error of 0.95 which is still meaningless as it is higher than the model just using the simple baseline in form of the mean value. This overshooting happens because the number of premium users is low and hence certain features might get high weights during the training to account for premium players. However, this might only be artifacts in the training dataset due to single players with a very high number of purchases. Nevertheless, a combination of multiple such features can quickly result in a large sum and hence the prediction overshoots. One can summarize that a Poisson regression approach requires further enhancements and adapted learning to be sensible to this extremely imbalanced problem. Still, the Poisson regression approach has a serious shortcoming compared to the regression trees. If weights are learned without regularization, all features are likely to obtain non-zero weights. Interpreting this large set of feature weights is much more difficult than understanding the paths in a PRT.

Since a PRT only represents a set of mean values for a Poisson distribution, we can use these mean values for a binary classification as well. We can reduce the prediction of a mean λ to a binary decision by setting $y = \texttt{True}$ if $\lfloor \lambda \rfloor \geq 1$ and `False` otherwise. Using the output of the regression model as classifier, we obtained F1 scores as shown in the last column of Tbl. 3. Although the learner optimizes a completely different criterion, the results are still comparable to the results obtained by the binary classifiers. However, one should also note that we used a larger training dataset compared to the experiments in the previous section. Looking at the results presented in this section, we can see that the findings from the previous section have been validated. For example, the amount of money spent in the first three days is also most important in the construction of the regression trees. Furthermore, the splitting on the country feature indicates that we should possibly revise our country segmentation according to the tree learner. However, this finding has to be verified on larger datasets across folds again. Across all folds, the tree learner finds similar models that can be used to group the players according to their expected number of purchases. This approach significantly decreases the prediction error compared to a naive baseline and also clearly outperforms a Poisson regression approach.

Conclusion and Discussion

Here two models for predicting premium players in F2P games are presented. Given the churn rates and imbalances in F2P games, any prediction model better than the baseline is valuable to the industry. The models and approaches presented provide a way to strengthen revenue streams from players through smart CRM and to increase engagement via tailored experiences. Thus the contribution is the first thorough investigation of purchase decisions in the unique situation of F2P from a player-behavior perspective, and the solution includes interaction- and economic features. Combined, this adds to our understanding of what drives purchasing behavior in F2P games and thus of player-game interaction within the context of F2Ps. Combining classification and regression allows a fine level of distinction of players and their expected behavior. Here, players are partitioned into groups based on their expected number of future purchases. By learning a descriptive regression model, error was reduced by roughly 13% compared to a baseline (7 day observations). The resulting models essentially provide information about what type of behavior (e.g World Number) encourage the players towards driving purchasing, as well as some circumstances (e.g. high Move Count) during which to try incentivizing players to make a purchase. On top of this, geographic and economic indicators are also implemented, indicating for example the importance of country as a predictor of purchasing in F2P games.

In more details, the models presented can also provide design information. Note that the tree in Fig. 2 shows interaction as well as economic features, e.g. World Number. These results indicate that designers of the game (and possibly others), can use them to optimize aspects of design. To take a few examples, the presence of Move Count in the model indicates that when people are intensely interacting with the game, a purchase is more likely, indicating when to provide offers. Playtime is also a driver of purchases - which means optimizing for total playtime rather than e.g. session length (Fig. 1). No. of Worlds [=levels] relate to progress - the more the player progresses in the game, the greater the potential for a purchase. The regression task holds similar results, e.g. country is a predictor which emphasizes localization and optimization to local markets. The work is also relevant for customer behavior modeling in marketing research. E.g. two components (#purchases, amount spent) of the Recency-Frequency-Monetary Value (RFM) framework (Reinartz and Kumar 2003; Fader and Lee 2005) - widely used in marketing - turn out to be the key drivers of purchasing in F2P games. It would be valuable to see research address this in more detail and add recency predictors. Future work will focus on improving the accuracy of the models as

well as target the potential of combining behavioral profiling, incentivization and prediction of spend, which in theory could be a powerful tool for F2P game companies.

References

Bauckhage, C.; Kersting, K.; Sifa, R.; Thurau, C.; Drachen, A.; and Canossa, A. 2012. How Players Lose Interest in Playing a Game: An Empirical Study Based on Distributions of Total Playing Times. In *Proc. IEEE CIG*.

Borle, S.; Singh, S. S.; and Jain, D. C. 2008. Customer lifetime value measurement. *Management science* 54(1):100–112.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Chawla, N.; Bowyer, K.; Hall, L. O.; and Kegelmeyer, W. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16(1):321–357.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

Drachen, A.; Thurau, C.; Togelius, J.; Yannakakis, G.; and Bauckhage, C. 2013. Game Data Mining. In El-Nasr, M.; Drachen, A.; and Canossa, A., eds., *Game Analytics: Maximizing the Value of Player Data*. Springer.

El-Nasr, M.S. and Drachen, A. and Canossa, A. 2013. *Game Analytics: Maximizing the Value of Player Data*. Springer.

Fader, P. S., H. B. G., and Lee, K. L. 2005. Frm and clv: Using iso-value curves for customer base analysis. *Journal of Marketing Research* 42(4):415–430.

Fields, T., and Cotton, B. 2011. *Social Game Design: Monetization Methods and Mechanics*. Morgan Kaufmann.

Gupta, S.; Lehmann, D. R.; and Stuart, J. A. 2004. Valuing customers. *Journal of marketing research* 41(1):7–18.

Hadiji, F.; Sifa, R.; Drachen, A.; Thurau, C.; Kersting, K.; and Bauckhage, C. 2014. Predicting Player Churn in the Wild. In *Proc. IEEE CIG*.

Kawale, J. A. P., and Srivastava, J. 2009. Churn Prediction in MMORPGs: A Social Influence Based Approach. In *Proc. CSE*.

Kubat, M.; Holte, R.; and Matwin, S. 1997. Learning When Negative Examples Abound. In *Proc. ECML*.

Lehdonvirta, V. 2009. Virtual item sales as a revenue model: identifying attributes that drive purchase decisions. *Electronic Commerce Research* 9(1-2):97–113.

Lim, N. 2012. Freemium Games Are Not Normal. http://www.gamasutra.com/blogs/NickLim/20120626/173051/Freemium_games_are_not_normal.php.

Luton, W. 2013. *Free-to-Play: Making Money From Games You Give Away*. New Riders.

Malthouse, E. C., and Blattberg, R. C. 2005. Can we predict customer lifetime value? *Journal of interactive marketing* 19(1):2–16.

McCullagh, P., and Nelder, J. 1989. *Generalized Linear Models*. Chapman and Hall.

Morik, K., and Köpcke, H. 2004. Analysing Customer Churn in Insurance Data–A Case Study. In *PKDD*, 325–336.

Mozer, M.; Wolniewicz, R.; Grimes, D.; Johnson, E.; and Kaushansky, H. 2000. Predicting Subscriber Dissatisfaction and Improving Retention in the Wireless Telecommunications Industry. *IEEE Trans. on Neural Networks* 11(3):690–696.

Mutanen, T.; Ahola, J.; and Nousiainen, S. 2006. Customer Churn Prediction-A Case Study in Retail Banking. In *Proc. of ECML/PKDD Workshop on Practical Data Mining*, 13–19.

Nozhnin, D. 2012. Predicting Churn: Data-Mining Your Game. *Gamasutra*.

Quinlan, J. R. 1996. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research* 77–90.

Reinartz, W. J., and Kumar, V. 2003. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing* 67(1):77–99.

Rothenbuehler, P.; Runge, J.; Garcin, F.; and Faltings, B. 2015. Hidden markov models for churn prediction. In *Proceedings of SAI IntelliSys*.

Runge, J.; Gao, P.; Garcin, F.; and Faltings, B. 2014. Churn Prediction for High-value Players in Casual Social Games. In *Proc. IEEE CIG*.

Schmittlein, D. C.; Morrison, D. G.; and Colombo, R. 1987. Counting your customers: Who-are they and what will they do next? *Management science* 33(1):1–24.

Seufert, E. 2014. *Freemium Economics: Leveraging Analytics and User Segmentation to Drive Revenue*. Morgan Kaufmann.

Sifa, R.; Bauckhage, C., and Drachen, A. 2014. The Playtime Principle: Large-scale Cross-games Interest Modeling. In *Proc. IEEE CIG*, 365–373.

Sifa, R.; Drachen, A.; Bauckhage, C.; Thurau, C.; and Canossa, A. 2013. Behavior Evolution in Tomb Raider Underworld. In *Proc. IEEE CIG*.

Sifa, R.; Ojeda, C.; and Bauckhage, C. 2015. User Churn Migration Analysis with DEDICOM. In *Proc. ACM RecSys*.

Swrve. 2014. Swrve Monetization Report 2014. http://landingpage.swrve.com/rs/swrve/images/swrve-monetization-report-0114.pdf.

Takahashi, D. 2014. The rising cost of acquiring mobile-app users is hitting devs like a hurricane. http://venturebeat.com/.

Therneau, T. M.; Atkinson, B.; and Ripley, B. 2011. *rpart: Recursive Partitioning*.

Wagner, T. M.; Benlian, A.; and Hess, T. 2014. Converting freemium customers from free to premiumthe role of the perceived premium fit in the case of music as a service. *Electronic Markets* 24(4):259–268.

Weber, B.; John, M.; Mateas, M.; and Jhala, A. 2011. Modeling Player Retention in Madden NFL 11. In *IAAI*.