Customer Lifetime Value Prediction in Non-Contractual Freemium Settings: Chasing High-Value Users Using Deep Neural Networks and SMOTE

Rafet Sifa
Fraunhofer IAIS
rafet.sifa@iais.fraunhofer.de

Julian Runge
Humboldt University Berlin
julian.runge@hu-berlin.de

Christian Bauckhage
Fraunhofer IAIS
christian.bauckhage@iais.fraunhofer.de

Daniel Klapper
Humboldt Univ. Berlin
daniel.klapper@hu-berlin.de

Abstract

In non-contractual freemium and sharing economy settings, a small share of users often drives the largest part of revenue for firms and co-finances the free provision of the product or service to a large number of users. Successfully retaining and upselling such high-value users can be crucial to firms' survival. Predictions of customers' Lifetime Value (LTV) are a much used tool to identify high-value users and inform marketing initiatives. This paper frames the related prediction problem and applies a number of common machine learning methods for the prediction of individual-level LTV. As only a small subset of users ever makes a purchase, data are highly imbalanced. The study therefore combines said methods with synthetic minority oversampling (SMOTE) in an attempt to achieve better prediction performance. Results indicate that data augmentation with SMOTE improves prediction performance for premium and high-value users, especially when used in combination with deep neural networks.

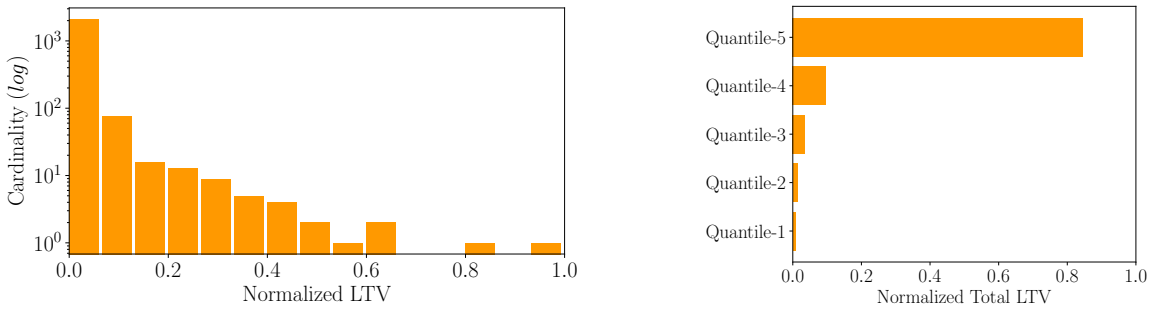
1. Introduction

Freemium is the dominant pricing model for digital products: A basic version of the product or service can be used for free and premium upgrades are available against payment of a fee. Contractual, subscription-based freemium models are widespread and e.g. used by Dropbox, Spotify and many providers of digital news. Here, customers upgrade a single time to a premium plan and then pay in monthly installments. Some firms, e.g. Tinder, extend their contractual upgrades with ad-hoc, non-contractual premium offers that allow for repeat purchasing at the customers' discretion. Purely non-contractual freemium models are common in products with large network effects such as Skype, many dating and lifestyle apps, and in digital (free-to-play) games [1–3].
Non-contractual upgrades are usually made in in-app purchases (IAPs) and often referred to as micro-transactions as they tend to be made in small amounts and repeated manifold by engaged players [1]. One key feature of non-contractual freemium and sharing economy settings is that a small subset of all users can finance the free provision of the product or service for the rest of users [4]. In free-to-play games, frequent repeat purchasers are often referred to as whales, relaying to their disproportional revenue contribution. To avoid any potentially demeaning connotations, we will refer to such users more objectively as *high-value users*.
High-value users indeed carry huge value for companies as their existence can make the difference between operating at a profit or a loss [5]. Companies hence often strive to provide the best possible experience to these users. The goal is to achieve a high share of premium users and to spur repeat purchases. High-value users are commonly defined based on their revenue contribution which is generally captured in the notion of customers' Lifetime Value (LTV) [5–7].
Firms wish to have accurate predictions of LTV as early as possible after users' product adoption to tailor their marketing efforts. The predicted LTV is used for a number of marketing initiatives:

- *User acquisition*, i.e. the task of acquiring new users on digital advertising networks [5]: LTV informs marketing managers how much they can afford to expense for acquisition of a new user. For user acquisition campaigns to be profitable, LTV needs to be higher than the amount expensed to have a new player install a game [8]. LTV predictions hence allow for a profitability assessment of different marketing campaigns. Budget size and allocation and targeting parameters can then in turn be adjusted for more optimal targeting.

- *Customer service*: High-value users, i.e. users with high predicted LTV, can be offered preferential treatment by customer service agents

URI: http://hdl.handle.net/10125/50002
ISBN: 978-0-9981331-1-9



(a) Histogram of LTV for premium users

(b) Sum of LTV per LTV-ranked quantile

Figure 1: A visual exploration of LTV in our non-contractual freemium setting: (a) illustrates the log-scaled distribution of normalized LTV per user; the majority of premium users pays a small amount. (b) shows the normalized sum of LTV (and hence revenue contribution) per LTV-ranked quantile; more than 80 % of the revenue comes from 20 % of premium users.

to ensure a great product experience or they can be enrolled in loyalty programs and other tailored marketing initiatives [2].

- *In-product advertising*: Companies strive not to harm the product experience of high-value users and may hence choose not to expose users with high predicted LTV to in-product advertising.
- *Pricing and promotions*: Marketers may choose not to expose players with a high predicted revenue contribution to price promotions with excessive discounts as this may harm their revenue contribution. At the same time, they may want to support users with lower predicted LTV by extending targeted offers to them to foster their engagement with the product.

Predicting the LTV of high-value users correctly is of particular importance. These users contribute the majority of overall revenue - e.g. the top 10% of paying users in [2] contribute 60% of revenue - and hence 'co-finance' product use and the development of future products for all other users [4][1]. Figures 1 and 2 visualize this phenomenon for the freemium setting under study in this paper, where we observe that more than 80 % of the revenue comes from 20 % of premium users.

Furthermore, identifying future high-value users and predicting their LTV correctly bears strongly on overall prediction performance as errors here will proliferate over-proportionally to predictions of average and summed LTV in customer cohorts.

[1]Bear in mind that only a small share of users pay at all [1, 9] and hence less than one percent of users generate 60% of revenue. [10] even find that one percent of users generates 85% of revenue.

1.1. Contribution

The present study is situated in the freemium gaming industry and makes use of data from a large free-to-play game (we use the terms freemium and free-to-play game interchangeably [11]) to generalize to non-contractual freemium and sharing economy settings. Despite the relevance of LTV predictions in this context, to our best knowledge, there is only one study on the topic [10]. This paper frames the prediction problem and applies a number of common machine learning methods for the prediction of individual-level LTV on unique data from a large free-to-play game. It then combines said methods with synthetic minority oversampling (SMOTE) [12] to achieve better prediction performance. We find that deep multilayer perceptrons combined with SMOTE achieve the best prediction performance for premium and high-value users. In doing so, we extend the study of Voigt and Hinz [10] by using a number of common machine learning methods rather than stochastic models, making use of non-purchasing related information and explicitly addressing the high imbalance and skewness present in freemium datasets.

2. Predicting LTV in Non-Contractual Freemium Settings

The marketing literature abounds of conceptual [6, 7], methodological [13, 14] and empirical [15, 16] studies of customers' LTV. Methodologically, most empirical contributions are routed in stochastic models of customer behavior [10, 14, 16] or regression approaches [15, 17]. The datasets at use mostly describe monetary transactions of customers.

The present study makes use of data from a large

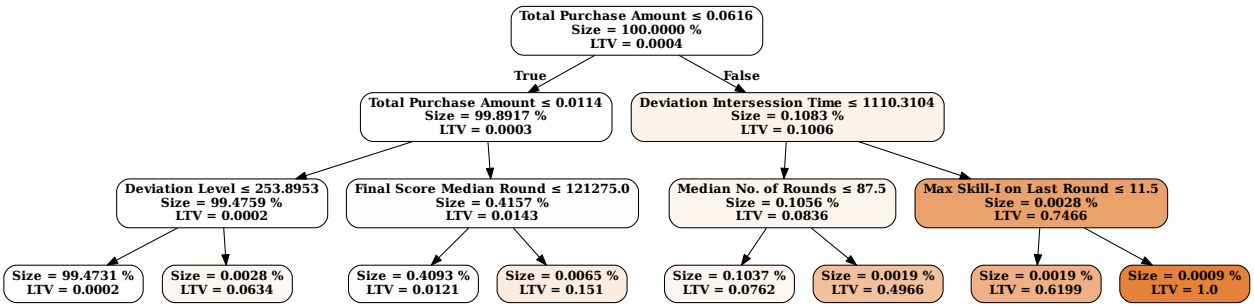


Figure 2: An example regression tree for individual-level LTV prediction in a freemium game. User activity related data contribute to predictions beyond data on a user's purchase history (please note that monetary predictors and LTV are normalized). The high skewness of the LTV distribution and the related low share of high-value users can be noted in the tree's leaves.

free-to-play game. There are numerous studies that investigate the prediction of player behavior in (freemium) games. E.g. [2, 18–21] investigate the prediction of player disengagement/churn, [22–24] focus on player retention and [1] predict players' purchase decisions in mobile free-to-play games. Player engagement and purchasing are at the core of players' value to a company, and so is their combined outcome: Monetary LTV. To date, to our best knowledge, there is only one study approaching LTV prediction in such settings [10].

Non-contractual freemium models can be mostly found in products with strong network effects (e.g. Skype or dating apps) and free-to-play games. They allow for repeat purchases, the number and monetary value of which is at the users' discretion (compare to [10, 16]). Only a small subset of users, usually in the single digit percent space [1, 9], ever make a purchase. Data are hence highly imbalanced [25] and there is abundant non-purchase related data that describe user behavior in depth and that can be valuable for prediction [1].

The stochastic models proposed by the marketing literature [14] are somewhat ill-equipped to deal with this kind of data as they were conceived for purchase related data. It would certainly be interesting to extend them to highly detailed usage data, but this is beyond the scope of the current study – especially as their performance in applied settings remains uncertain [26].

We wish to emphasize that our methodological thinking deviates substantially from [10] who build on stochastic models and disregard non-purchasing related information. While purchasing related information carry a lot of predictive value, it has been shown that non-transactional data do contribute to prediction performance additionally [1]. We hence expect their inclusion to lead to superior prediction performance. Figure 2 indicates that they indeed do as user activity related predictors show up near the top of the tree. It will be interesting to compare both approaches in future research.

In the spirit of [27], a marketing study that finds bagging and boosting techniques to perform well for customer behavior prediction, and [1] who find Random Forests with SMOTE to perform best for purchase prediction in free-to-play games, we adopt tree-based (ensemble) methods combined with approaches to deal with the strong imbalance in the data [12]. We also include regression techniques as these have been reported to work well [15, 17]. We opt for using simple linear regression as a benchmark. And we finally add deep multilayer perceptrons (Deep-MLP) to the mix because of their recent superiority in a number of applications [28, 29]. [2] further find them to perform well for user behavior prediction in freemium games.

2.1. The prediction problem

[4, 9, 30, 31] all highlight the relevance of social interaction for engagement, monetization and virality in freemium games and products more generally. While we deem it highly relevant to investigate how social interaction shapes users' LTV in freemium settings and how one can reasonably quantify positive externalities such as viral activity to include them in LTV, the focus of the present study is on monetary LTV. In line with managerial literature [5, 8], we aim at predicting users' undiscounted cumulative spend until day 360 of product use[2] and we disregard cost as it is virtually zero at the margin [9, 10].

As firms strive to have predictions available early

[2]Some companies prefer different time horizons, e.g. 180 or 720 days of product use. This largely depends on monetization dynamics of the firm's products and financial/strategic planning horizons. In our opinion, it is unlikely that prediction performance of different methods substantially differs between the different mentioned time horizons; we hence adopt the time horizon used by the data sponsor of this study.

Table 1: Dataset description

Feature Type	Descriptor(s)
Telemetry	Number of sessions
	Number of rounds
	Number of days
	Number of purchases
	Total purchase amount
	Total playtime
	Total absence time
	Total inter-session time
	Total inter-round time
	Total weekday number
	Social network connections
	Total score
	Number of lives
	Amount of game currency
	Number of cleared game elements
	Difficulty level
	Moves count
	Level outcome
	Skill measure I
	Skill measure II
	Skill measure III
	Level type
Temporal	Time between daily first and last session
	Inter-day time distribution
	Inter-session time distribution
Composite	Correlation coefficients on time[1]
	First order trends[1]
	Maximum, mean, median and deviation on time[1]
	Activity ratios and entropy per unit[2]
Meta	Country segment
	Device type
	Operating system
	Acquisition type

[1] Calculated for daily and session-wise distributions of features in **Telemetry** and **Composite**
[2] Calculated for the game types and difficulty levels in **Telemetry**

in the customer journey to inform aforementioned marketing actions [8], we use datasets of seven days of user behavior observations in the prediction. This is further sensible as user behavior in free-to-play games displays a strong weekly periodicity. Formally, we can denote the prediction problem as

$$LTV = y_{day8-day360} = \mathrm{f}\,(\boldsymbol{x}_{dayZ}), \qquad (1)$$

where

- $y_{day8-day360}$ represents the total amount spent by a user on IAPs during the prediction period, i.e. between day eight and day 360 after product adoption.
- $\boldsymbol{x}_{dayZ} \in \mathbb{R}^m$ describes a user behavior observation set for given m features until day Z after the installation of the game, for our case $Z = 7$.
- $\mathrm{f} : \mathbb{R}^m \rightarrow \mathbb{R}$ is the hidden function we will be approximating.

Grouping the target values under a vector $\boldsymbol{y}$, and following [15], our aim is to evaluate different approximators of f that individually minimize the normalized rooted mean squared error (NRMSE) that we denote as

$$NRMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_i - \hat{y}_i)^2}\;\bar{y}^{-1}, \qquad (2)$$

where $\bar{y}$ represents the mean of $\boldsymbol{y}$.

2.2. Data Pre-processing and Feature Extraction

Sifa et al.'s [1] is one of the few studies focusing on the prediction of purchases in free-to-play games. One reason for the paucity of research on this topic may be the unavailability of detailed data on monetary transactions. In this work we used a dataset containing behavioral and monetary information of 120,000 players of a large free-to-play game. The game builds on a widespread casual game mechanic and is distributed on Apple's app marketplace. It is highly representative of large-scale social games [31] and has been played by close to 100 million players across platforms to date. Game monetization, as standard in free-to-play games, is based on IAPs where the player purchases in-game currencies that can then be used to acquire in-game goods.

Before building our prediction models, similar to [1, 19, 22], we define game agnostic and gameplay features for each player that can be grouped under four categories. In category *Telemetry* we consider primary activity-related metrics such as number of sessions, rounds and days played or the total play and inter-session time within the window of observation. Additionally, we consider game specific features such as the amount of in-game currency purchased, the number of lives and the current game and mastery level of players.

The category *Temporal* covers features that capture temporal patterns of behavior, such as time between first and last session and inter-day and inter-session time distributions. To capture variation in time and represent ratios of the features in the previous two categories, we calculate players' game-type distributions and correlation coefficient, maximum, mean, median and standard deviation on time. We call this category *Composite* as we create composite features from pure telemetry and temporal data.

The final category in our data set is *Meta* which contains features such as country of origin, type of device, operating system and if a player was acquired through digital marketing or not.

We refer the reader to Table 1 for an overview of the dataset. In the next section we will address a method to regularize regression models towards predicting non-zero entities better by generating synthetic premium users during the training phase.

3. Imbalance in Behavioral Datasets and Synthetic Minority Oversampling

Prediction objectives in behavioral datasets are often characterized by high class imbalance. In the case of digital games, this imbalance is mostly caused by rare classes rather than rare events [25] as the number of observed entities is usually very large. Often, prediction of rare entities carries the higher value proposition. This is certainly the case for premium and high-value users in freemium settings. Other examples include the prediction of fraudulent behavior [32], detecting purchase decisions [1] or predicting player churn [2, 19, 33].

Due to the more or less generalizable approaches to finding optimal parameters, conventional supervised machine learning models usually tend to lean towards the majority class especially when we are dealing with highly unbalanced datasets. SMOTE [12] is a flexible data augmentation method that can be adapted to behavioral datasets in a straightforward manner [1]. It creates synthetic entities of the minority class during the model training phase to regularize the prediction models to avoid overfitting and to learn structures representing minority entities. In many ways, SMOTE resembles distortion-based model regularization techniques [34, 35]. In this section, we will shortly study the adaptation of the algorithm for LTV prediction.

Since collectively and sequentially structured behavioral datasets usually contain numerical (such as number of sessions/logins to the app) as well as nominal attributes (such as country of origin), in the following, we consider an ordered set based data representation. Formally, we start by letting $\boldsymbol{y} \in \mathbb{R}^n$ contain the vector of LTV values for the n players in our training set, grouping the data indices of numeric (including the LTV value) and nominal attributes respectively as under $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{B}}$, we represent a data point as an $\hat{m} = |\boldsymbol{\mathcal{V}}| + |\boldsymbol{\mathcal{B}}|$ dimensional ordered set that we refer to as $\S$ and the data set as $\boldsymbol{\mathcal{X}} = \{\S_1, \S_2, \ldots, \S_n\}$. This is, followed by separating the minority data points that we will be using to generate synthetic entities as $\boldsymbol{\mathcal{X}}_p = \{\S_{\boldsymbol{\mathcal{J}}[1]}, \S_{\boldsymbol{\mathcal{J}}[2]}, \ldots, \S_{\boldsymbol{\mathcal{J}}[\hat{n}]}\}$ where $\boldsymbol{\mathcal{J}} = \{i \mid \forall i \in [1, 2, \ldots, n] \wedge y_i \neq 0\}$ with the cardinality $\hat{n} = |\boldsymbol{\mathcal{J}}|$.

Next, for a predefined number of times (which we denote as $\tilde{n}$), we select a random data point $\S_i$ from $\boldsymbol{\mathcal{X}}_p$, one of its k nearest neighbors (which we denote as $\S_{neighbor}$) and create a point as combination of both. There are two major points to be stressed regarding this generation process:

- *Defining a dissimilarity* (or similarity) coefficient [36] between two selected points: In this study, we make use of an additive dissimilarity coefficient combining Euclidean distance and set difference weighted dissimilarity for numerical and nominal features respectively. The set difference-based similarity can be obtained by considering a composite metric on $\boldsymbol{\mathcal{B}}$ [1, 12] as a so called *punishment multiplier* to increase the dissimilarity between points sharing a smaller number of nominal attributes. In essence, the dissimilarity between two given data points $\S_i$ and $\S_j$ can be obtained as

$$d_{ij} = \sqrt{\boldsymbol{u}_{ij}^T \boldsymbol{u}_{ij} + c^2 \sum_{l=1}^{|\boldsymbol{\mathcal{B}}|} \mathrm{q}(b_{il}, b_{jl})}, \qquad (3)$$

  where $\boldsymbol{u}_{ij} = \S_i[\boldsymbol{\mathcal{V}}] - \S_j[\boldsymbol{\mathcal{V}}]$ is the difference vector of the numerical attributes of $\S_i$ and $\S_j$, $|\cdot|$ represents the set cardinality, $\boldsymbol{b}_i \in \mathbb{N}^{|\boldsymbol{\mathcal{B}}|}$ indicates the enumerated nominal attributes in $\S_i[\boldsymbol{\mathcal{B}}]$, q is the discrete metric [37] and defined as

$$\mathrm{q}(a, b) = \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{otherwise.} \end{cases} \qquad (4)$$

  Additionally, the punishment multiplier c can be calculated as

$$c \leftarrow \underset{l \in \boldsymbol{\mathcal{V}}}{\mathrm{median}} \{\mathrm{g}(\boldsymbol{x}_{pl})\}, \qquad (5)$$

  where $\boldsymbol{x}_{pl} \in \mathbb{R}^{\hat{n}}$ represents the (numerical) values of the lth attribute for data points belonging to the minority class and $\mathrm{g} : \mathbb{R}^{\hat{n}} \rightarrow \mathbb{R}$ and can picked to be the standard deviation [1, 12]. It is worth mentioning that, since we are interested in finding k nearest neighbors for each active point (i.e. the *dissimilarity ranking* between all of the minority players), we can as well consider the squared value of d_{ij} in (3) when selecting a similar player [38]. Furthermore, both d_{ij} and d_{ij}^2 are valid symmetric dissimilarity coefficients [36] with zero lower-bounds (also considering *self dissimilarity*).

- *Mixing the points* to generate new ones: We can combine two users by creating a nonnegative

mixture vector $\boldsymbol{h}_i = \{h_{ij} | \forall j \in [1, 2, \ldots, |\boldsymbol{\mathcal{V}}|] \wedge h_{ij} \in [0, \ldots, 1]\}$ containing probabilities (for instance selected from uniform distribution) and considering the convex mixture

$$\S_{synth}[\boldsymbol{\mathcal{V}}] \leftarrow \boldsymbol{h}_i \odot \S_i[\boldsymbol{\mathcal{V}}] + (\mathbf{1} - \boldsymbol{h}_i) \odot \S_{neighbor}[\boldsymbol{\mathcal{V}}], \quad (6)$$

where $\S_{synth}$ is the synthetic data point that we are generating and $\odot$ is the Hadamard product. As a final step $\S_{synth}$ inherits the nominal attributes of the active data point as $\S_{synth}[\boldsymbol{\mathcal{B}}] \leftarrow \S_i[\boldsymbol{\mathcal{B}}]$.

We list the main steps of SMOTE adapted for LTV prediction in Alg. 1, where we for convenience rewrote (3) in terms of euclidean norm and set difference.

Algorithm 1 Synthetic Minority Oversampling with Nominal and Numerical Attributes for Lifetime Value Analysis

// Separate attributes based on their types
Let $\boldsymbol{\mathcal{V}}$ contain indices of numerical attributes in $\boldsymbol{\mathcal{X}}$
Let $\boldsymbol{\mathcal{B}}$ contain indices of nominal attributes in $\boldsymbol{\mathcal{X}}$
// Filter the premium players from the data set
Let $\boldsymbol{\mathcal{J}} = \{i \mid \forall i \in [1, 2, \ldots, n] \wedge y_i \neq 0\}$ with $|\boldsymbol{\mathcal{J}}| = \hat{n}$
Let $\boldsymbol{\mathcal{X}}_p = \{\S_{\boldsymbol{\mathcal{J}}[1]}, \S_{\boldsymbol{\mathcal{J}}[2]}, \ldots, \S_{\boldsymbol{\mathcal{J}}[\hat{n}]}\}$
// Define a multiplier for nominal deviations
$c \leftarrow \underset{l \in \boldsymbol{\mathcal{V}}}{\text{median}}\{\text{g}(\boldsymbol{x}_{pl})\}$
Set $\boldsymbol{\mathcal{U}}$ to contain synthetic players.
// Create $\tilde{n}$ synthetic players
for $i \in [1, 2, \ldots, \tilde{n}]$ **do**
  // Randomly select a premium player
  $\S_i \leftarrow Select(\boldsymbol{\mathcal{X}}_p)$
  // $\forall j \in [1, 2, \ldots, \hat{n}]$ calculate the dissimilarities to $\S_i$
  $d_{ij} \leftarrow \sqrt{\left\|\S_i[\boldsymbol{\mathcal{V}}] - \S_j[\boldsymbol{\mathcal{V}}]\right\|^2 + c^2 \left|\S_i[\boldsymbol{\mathcal{B}}] - \S_j[\boldsymbol{\mathcal{B}}]\right|}$
  // Randomly select one of the k nearest neighbors
  $\S_{neighbor} \leftarrow Select(\boldsymbol{\mathcal{X}}_p, k, d_{i1}, d_{i2}, \ldots, d_{i\hat{n}})$
  //Initialize a random mixture vector $\boldsymbol{h}_i$
  $\boldsymbol{h}_i = \{h_{ij} | \forall j \in [1, 2, \ldots, |\boldsymbol{\mathcal{V}}|] \wedge h_{ij} \in [0, \ldots, 1]\}$
  // Create a synthetic player by mixing both vectors
  $\S_{synth}[\boldsymbol{\mathcal{V}}] \leftarrow \boldsymbol{h}_i \odot \boldsymbol{x}_i$
  $\S_{synth}[\boldsymbol{\mathcal{V}}] \leftarrow \S_{synth}[\boldsymbol{\mathcal{B}}] + (\mathbf{1} - \boldsymbol{h}_i) \odot \S_{neighbor}[\boldsymbol{\mathcal{V}}]$
  // Inherit the nominal attributes from $\S_i$
  $\S_{synth}[\boldsymbol{\mathcal{B}}] \leftarrow \S_i[\boldsymbol{\mathcal{B}}]$
  // Add the synthetically generated player to $\boldsymbol{\mathcal{U}}$
  $\boldsymbol{\mathcal{U}} \leftarrow \boldsymbol{\mathcal{U}} \cup \S_{synth}$
end for
Train regression models with $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{U}}$

4. Results and Discussion

This section presents the results obtained from benchmarking four different regression methods on our dataset with and without SMOTE based data augmentation. We will start by introducing the general setting we used for experimentation and present the prediction results for different user segments. Next, as also illustrated for the classification problems in [1, 22, 23], we will show a ranked list of important features from our decision tree based models to understand how informative the different features are for the prediction of LTV. Finally, we will turn our attention to the use of our findings to practitioners, in particular how they can be used by managers to customize marketing iniatitives.

4.1. Setting

We compare the prediction performances of mean squared error minimizing Random Forests (RF), Linear Regression (LR), Decision Trees (DT) and Deep Multilayer Perceptrons (Deep-MLP) with and without SMOTE augmentation. We apply four different perspectives: Correctly predicting LTV of (a) all users, (b) premium users and (c) high-value users that are within the fifth (highest value) quantiles (see Fig. 1b); and (d) assessing the hit rate of sorted LTV values produced by the models. The first three perspectives aim to analyze the stability of the models from general to specific player types and we will use NRMSE (see Eq. (2)) to quantify the quality of the fit. The last extends this assessment to user recommendations targeted on a ranking. To concisely quantify the performance, we will refer to the recall value of each LTV-ranked list.

Furthermore, we used fully connected Multilayer Perceptrons of four hidden layers with neurons ranging in $[50, \ldots, 140]$ in each layer. To avoid overfitting and obtain better generalization [39] we apply dropout regularization [40], which randomly blinds different portions of the weights of neural networks during training. Unless it is explicitly specified all of the results presented are based on tenfold cross validation. In order to ensure usability of the proposed system in real world analytics applications, both the SMOTE-based data augmentation as well as the model building phase have been performed individually for each cross validation split [22]. Models with highest NRMSE for high value users were selected.

4.2. LTV Prediction Results

Table 2 lists the average results of a tenfold cross-validation for the four chosen algorithms, once

Table 2: Normalized Rooted Mean Squared Error Values from 10 fold cross validation for the entire (All), premium (Prem.) and high value (H.V.) players.

Model	NRMSE		
	All	Prem.	H.V.
RF	22.61	3.04	1.57
LR	22.10	3.07	1.60
DT	29.47	3.37	1.69
Deep-MLP	21.80	3.03	1.57
RF+SMOTE	22.52	3.03	1.55
LR+SMOTE	22.90	3.02	1.57
DT+SMOTE	34.65	3.66	1.67
Deep-MLP+SMOTE	21.97	2.90	1.48

with SMOTE data augmentation and once without. NRMSE can be interpreted as the RMSE in percent of the predicted criterion [15]. RF, LR and Deep-MLP display comparable prediction performance with Deep-MLP having a slight edge over the other methods. The DT regressor performs worst, both with and without SMOTE. In particular it does not benefit and likely overfits when data is augmented with SMOTE. Importantly, we observe that SMOTE improves the prediction performance of multilayer perceptrons for premium and high-value users, by four and six percent respectively. If we look at all users, SMOTE slightly improves the performance of RF which is in line with prior research [1].

The behavioral and background data at hand are characteristic of the datasets usually available to industry in such settings.

Relevant data that are not covered by the background heterogeneity and product use observations in our dataset (see table 1 for an overview) may address personality and other psychological characteristics of users, their income and further socioeconomic covariates.

Additionally, a large part of the error stems from non-zero predictions for users with actual zero LTV. This is apparent in the NRMSE dropping significantly when we look at premium and high-value users only (the two rightmost columns in Table 2). It appears worthwhile to investigate in future studies how incorporating data from different sources might improve the predictions of high-value users as well as the actual zero instances.

4.3. Variable Importance Analysis

One advantage of decision tree based approaches for classification and regression is that they allow to read off the important features (which are informative for the prediction objective) directly from the built models [1, 22]. We extend the variable importance insights from figure 2 with an additional approach: Analyzing the prediction results for each of the features and comparing the reduced *loss* over all of the estimators in the random forest models. In figure 3 we compare the average normalized feature importance values for the top 20 features for random forest models trained on datasets augmented with SMOTE. [1] find that the number of purchases and the amount spent were almost equally informative for the prediction of future purchase decisions (with the former being more informative). Our results indicate that the amount spent during the last purchase prior to prediction time point is the most informative feature to predict LTV. This is followed by the frequency of the purchases made within the observation window. These findings are in line with a recent study on LTV in freemium settings [10].

Additionally, for the game under study, we observe that features encoding user activities related to in-game currency are highly ranked predictors of the LTV of a user. These involve the round-based mean and deviation, the temporal correlation and status on the last and median round. We interpret them to reflect the relation between users' spending on IAPs and the evolution of their stock of in-game currency (remember that in-game currency is sold in IAPs).

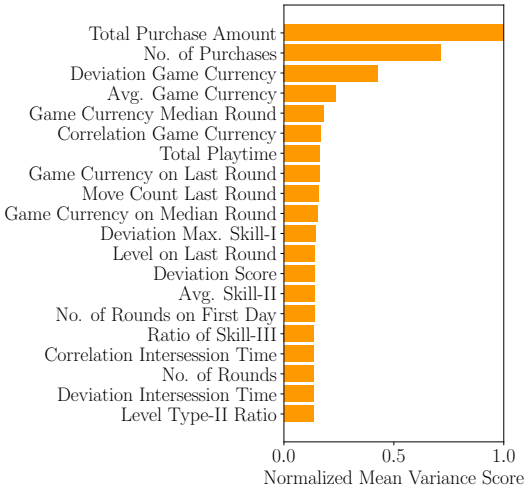


Figure 3: Normalized mean variance scores for feature importance from random forest based regression models that are trained with augmented datasets. In line with prior research [1, 10], results indicate that the purchase amount as well as the number of previous purchases are the most informative features for determining future LTV of a player. Features encoding player in-game activities also rank highly.

In line with the findings of [1], our results also underline the importance of product use related features such as different measures of skill. This speaks to our methodological difference from [10] who build on stochastic models and disregard non-purchasing related information. While purchasing related information carry the bulk of predictive value, product use related information add to prediction performance. A direct comparison of results is not possible due to choice of different error measures. We deem it highly interesting to benchmark stochastic models and the presented machine learning methods in future research. An extension of stochastic models to take product use related information into account could be valuable to further improve their prediction performance in freemium settings.

4.4. Informing Marketing Initiatives: Hit Rates

Having noted the overall prediction results and variable importance through decision tree based models, we now turn our attention to the use of the presented methods for marketing initiatives. Often, a ranking of users in terms of revenue contribution forms the main decision basis. Coming back to the marketing initiatives mentioned in the introduction of the paper, examples of managerial actions informed by a user ranking in terms of LTV are:

- *User acquisition*: Shift marketing budget to campaigns that contain large shares of high ranking users.
- *Customer service*: Handle requests from high ranking users with priority.
- *In-product advertising*: Do not expose top x% of the ranked list to advertising.
- *Pricing and promotions*: Offer large bundles to high ranking users, offer small bundles to low ranking users.

To assess the validity of the ranking yielded by the different methods, we sort users into a ranking based on the predictions of each method. We use this ranking to evaluate how well the respective method is able to capture future premium users by analyzing the hit rate (or recall) of premium players along the sorted LTV predictions. Similar to the prediction results we showed in table 2, here the values are calculated from the predictions of the unseen test entities and averaged over tenfold cross-validation.

Tables 3 and 4 show the hit rates for the task of finding a premium player through the estimated LTV ranking and the area under the hit rate curve, both for the non-augmented and augmented datasets. Each row contains information about the recall and area of the curve *so far* calculated for a particular population ratio. To illustrate: The fourth row in Table 4 indicates that for each trained model, we consider one fifth of the data set of the LTV prediction-sorted players and calculate the percentage of future premium users that were "hit correctly". These values amount to 75.41, 63.88, 68.94 and 68.70%, for Deep-MLP, DT, RF and LR respectively on the data with SMOTE augmentation. Data augmentation with SMOTE helps increase the hit rates for Deep-MLP, DT and RF. For LR, results are mixed, but predominantly positive. Importantly, data augmentation with SMOTE combined with multilayer perceptrons allows to identify more than 80% of premium users with a 25% population ratio. This makes use of LTV ranked user lists in practice highly attractive. These 25% of users can be used for look-alike targeting in user acquisition[3], treated with priority in customer service, excluded from possibly annoying advertising, and be targeted with attractive upselling promotions.
The area under the hit curve similarly allows us to compare the evolution of the hit rate as we increase the considered population ratio. It largely confirms hit rate-based results in that multilayer perceptrons with SMOTE operate at a clear advantage over the other considered methods in identification of future premium and high-value users.

5. Conclusion

In non-contractual freemium settings, users can freely choose the number and size of their purchases of premium upgrades, examples are freemium games and dating, co-working and networking applications. In such settings, a small share of users usually drives revenue disproportionately and co-finances free provision of the product (and development of new products) for all users [4]. The existence of such high-value users can be essential to firms' survival. Firms hence wish to identify these users as early as possible to tailor marketing initiatives to retain existing and acquire new users of this kind.
Speaking to this managerial problem, we identify high-value users as those with high future revenue contribution and hence high LTV. In line with managerial literature [5,8], we define LTV as users' IAP spending until day 360 of product use. We frame the

[3]For more information we refer the interested reader to `https://www.facebook.com/business/a/lookalike-audiences`.

Table 3: Cross validated hit rates (in %) and hit curve areas for models trained on nonaugmented data

Ratio	Deep-MLP		DT		RF		LR	
5	43.15	188.53	35.50	148.07	35.04	138.48	42.52	189.67
10	54.08	483.95	46.85	397.86	50.38	392.76	53.53	479.30
15	61.48	832.11	53.88	703.52	60.67	728.75	60.93	823.26
20	67.58	1217.63	61.00	1047.38	67.82	1113.93	66.69	1205.41
25	72.29	1637.73	66.10	1430.50	73.70	1537.18	71.64	1620.25
30	75.87	2082.66	71.52	1843.42	78.03	1992.41	75.65	2061.47

Table 4: Cross validated hit rates (in %) and hit curve areas for models trained with augmented data

Ratio	Deep-MLP		DT		RF		LR	
5	46.24	193.00	36.38	171.11	40.16	175.48	40.13	161.87
10	59.92	515.66	45.85	418.32	52.95	458.34	53.68	446.15
15	68.86	903.89	54.42	719.67	62.00	803.35	62.70	797.04
20	75.41	1337.27	63.88	1074.10	68.94	1196.88	68.70	1192.48
25	81.08	1806.39	72.27	1481.85	74.44	1626.55	74.58	1622.10
30	85.20	2304.20	75.83	1928.55	78.78	2087.35	78.75	2082.52

related prediction as a regression of LTV on background information (particularly user's device and country) and seven days of in-product behavioral observations and apply four broadly used and powerful algorithms to solve it. To our best knowledge, we are the first to do so.

In line with existing literature [1, 10, 15], we find users' past purchases and their value to be highly predictive of LTV. We further find that features derived from product use improve predictive accuracy additionally.

Addressing the high imbalance present in freemium datasets where only a few users ever make a premium purchase, we apply data augmentation with SMOTE – creating synthetic premium and high-value users to facilitate their prediction. We find deep neural networks in combination with SMOTE to provide superior overall performance and outline how predictions can be used in managerial applications.

Finally, we want to briefly discuss limitations that open up opportunities for further research. The incorporation of social [4, 31] aspects of user and player behavior in LTV prediction is a beneficial avenue for future research. Further, it will be necessary to verify the present results on further products and games to assess their consistency and to enable the development of cross-product predictive systems [19, 41]. Last but not least, we deem it highly relevant to explore additional data sources for LTV prediction. The obtained NRMSE results indicate that the data at hand – that are characteristic of industry datasets in freemium environments – partially capture the data generating process; especially errors for users with zero actual LTV are highly prevalent. Additional socioeconomic and environmental covariates as well as information pertaining to psychological and personality attributes [42–44] may address relevant dimensions of consumer decision-making and could increase predictive performance tremendously.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments and the data sponsor for granting access to a unique dataset.

References

[1] R. Sifa, F. Hadiji, J. Runge, A. Drachen, K. Kersting, and C. Bauckhage, "Predicting Purchase Decisions in Mobile Free-to-Play Games," in *Proc. of AAAI AIIDE*, 2015.

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

[3] R. Sifa, A. Drachen, and C. Bauckhage, "Large-scale Cross-game Player Behavior Analysis on Steam," in *Proc. of AAAI AIIDE*, 2015.

[4] R. Bapna, J. Ramaprasad, and A. Umyarov, "Monetizing Freemium Communities: Does Paying for Premium increase Social Engagement?," in *MIS Quarterly*, 2017.

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

[6] S. Gupta, D. R. Lehmann, and J. Ames Stuart, "Valuing Customers," in *Journal of Marketing Research*, 2004.

[7] W. J. Reinartz and V. Kumar, "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration," in *Journal of Marketing*, 2003.

[8] J. Runge, "The Golden Curve: Determining Player Value in Freemium Apps." `https://thenextweb.com`, 2014.

[9] J. Runge, S. Wagner, J. Claussen, and D. Klapper, "Freemium Pricing: Evidence from a Large-scale Field Experiment," tech. rep., Humboldt University Berlin, School of Business and Economics, Institute of Marketing Working Paper, 2016.

[10] S. Voigt and O. Hinz, "Making Digital Freemium Business Models a Success: Predicting Customers Lifetime Value via Initial Purchase Information," *Business and Information Systems Engineering*, vol. 58, no. 2, 2016.

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

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002.

[13] D. C. Schmittlein, D. G. Morrison, and R. Colombo, "Counting Your Customers: Who Are They and What Will They Do Next?," in *Management Science*, 1987.

[14] P. Fader, B. G. S. Hardie, and K. L. Lee, "'Counting Your Customers' the Easy Way: An Alternative to the Pareto/NBD Model," in *Marketing Science*, 2005.

[15] B. Donkers, P. C. Verhoef, and M. G. De Jong, "Modeling CLV: A Test of Competing Models In the Insurance Industry," in *Quantitative Marketing and Economics*, 2007.

[16] P. S. Fader, B. G. S. Hardie, and K. Jerath, "Estimating CLV Using Aggregated Data: The Tuscan Lifestyles Case Revisited," in *Journal of Interactive Marketing*, 2008.

[17] Y. Ekinci, F. Ülengin, N. Uray, and B. Ülengin, "Analysis of Customer Lifetime Value and Marketing Expenditure Decisions through a Markovian-based Model," in *European Journal of Operational Research*, 2014.

[18] H. Xie, S. Devlin, D. Kudenko, and P. Cowling, "Predicting Player Disengagement and First Purchase with Event-frequency Based Data Representation," in *Proc. of IEEE CIG*, 2015.

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

[20] P. Rothenbuehler, J. Runge, F. Garcin, and B. Faltings, "Hidden Markov Models for Churn Prediction," in *SAI IntelliSys*, 2015.

[21] Á. Periáñez, A. Saas, A. Guitart, and C. Magne, "Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles," in *Proc. of IEEE DSAA*, 2016.

[22] R. Sifa, S. Srikanth, A. Drachen, C. Ojeda, and C. Bauckhage, "Predicting Retention in Sandbox Games with Tensor Factorization-based Representation Learning," in *Proc. of IEEE CIG*, 2016.

[23] A. Drachen, E. Lunquist, Y. Kung, P. Rao, D. Klabjan, R. Sifa, and J. Runge, "Rapid Prediction of Player Retention in Free-to-Play Mobile Games," in *Proc. of AAAI AIIDE*, 2016.

[24] M. Viljanen, A. Airola, T. Pahikkala, and J. Heikkonen, "Modelling User Retention in Mobile Games," in *Proc. of IEEE CIG*, 2016.

[25] G. M. Weiss, "Mining with Rarity: A Unifying Framework," in *ACM SIGKDD Explorations Newsletter 6.1*, 2004.

[26] M. Wuebben and F. v. Wangenheim, "Instant Customer Base Analysis: Managerial Heuristics Often 'Get It Right'," in *Journal of Marketing*, 2008.

[27] A. Lemmens and C. Croux, "Bagging and Boosting Classification Trees to Predict Churn," in *Journal of Marketing Research*, 2006.

[28] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, vol. 529, 2016.

[29] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level Classification of Skin Cancer with Deep Neural Networks," *Nature*, vol. 542, 2017.

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

[31] A. Alsen, J. Runge, A. Drachen, and D. Klapper, "Play With Me? Understanding and Measuring the Social Aspect of Casual Gaming," in *Proc. of AAAI AIIDE*, 2016.

[32] M. A. Ahmad, B. Keegan, J. Srivastava, D. Williams, and N. Contractor, "Mining for Gold Farmers: Automatic Detection of Deviant Players in MMOGs," in *Proc. of IEEE CSE*, 2009.

[33] S. K. Lee, S. J. Hong, S. I. Yang, and H. Lee, "Predicting Churn in Mobile Free-to-play Games," in *Proc. of IEEE ICTC*, 2016.

[34] C. M. Bishop, "Training with Noise is Equivalent to Tikhonov Regularization," *Neural Computation*, vol. 7, no. 1, 1995.

[35] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and Composing Robust Features with Denoising Autoencoders," in *Proc. of ACM ICML*, 2008.

[36] N. Jardine and R. Sibson, *Mathematical Taxonomy*. Wiley, 1971.

[37] J. R. Munkres, *Topology*. Prentice Hall, 2000.

[38] R. Sibson, "Order Invariant Methods for Data Analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 3, 1972.

[39] Y. Abu-Mostafa, "The Vapnik-Chervonenkis Dimension: Information Versus Complexity in Learning," *Neural Computation*, vol. 1, no. 3, 1989.

[40] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, 2014.

[41] N. Shaker and M. Abou-Zleikha, "Transfer Learning for Cross-game Prediction of Player Experience," in *Proc. of IEEE CIG*, 2016.

[42] M. Kosinski, D. Stillwell, and T. Graepel, "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior," *PNAS*, vol. 110, 2013.

[43] S. C. Matz, J. J. Gladstone, and D. Stillwell, "Money Buys Happiness When Spending Fits Our Personality," *Psychological Science*, vol. 27, 2016.

[44] P.-H. Chen, Y.-P. Tu, and K.-T. Chen, "On the Tiny Yet Real Happiness Phenomenon in the Mobile Games Market," in *Proc. of IEEE DSAA*, 2016.