

A Model for Gamers' Revenues in Casinos

Jehoshua Eliashberg

Sam K. Hui

Raghuram Iyengar*

June 18, 2009

*Jehoshua Eliashberg is the Sebastian S. Kresge Professor of Marketing and Professor of Operations and Information Management at the Wharton School of the University of Pennsylvania, Sam K. Hui an Assistant Professor in Marketing at Stern School of Business of New York University, and Raghuram Iyengar is an Assistant Professor in Marketing at the Wharton School of the University of Pennsylvania. The authors are listed in alphabetical order. Please address all correspondence to Jehoshua Eliashberg (eliashberg@wharton.upenn.edu). The authors benefitted from comments by Sunil Gupta and Christophe Van den Bulte.

A Model for Gamers' Revenues in Casinos

Abstract

We develop a model for predicting gamers' revenues using data provided by a major casino operator. Our model captures the underlying revenue generation process. Specifically, we model both gamers' visit and gambling behavior. By incorporating latent individual-level skill parameters, our model allows casino managers to better understand the role that "skill" or "luck" play in driving the revenue from each player. It also provides the casino operator with a better characterization of players (high skill, high rollers etc.) in addition to generating more accurate predictions of future revenue from each player. We estimate the model using Bayesian methods. We show that predictions of future player revenues using our model are more accurate than those obtained from several alternative models, which capture revenues but ignore the underlying revenue generation process.

1. Introduction

Consumers engage in various leisure activities. For instance, during 2005-2007, Americans made 3 visits per month on average to a mall with each trip lasting more than 1 hour and involved a spending of more than \$75 (Connolly and Rogoff 2008). Recent estimates indicate that Americans went to the cinema about 5 times a year on average (Nationmaster 2009) and spent more than \$9 Billion in a year (Motion Picture Association of America 2009). In the gaming industry, which is the focus of this paper, the average trip frequency to a casino is about 6 visits per year (Harrah's Report 2006) with a spending of \$90 per trip (Morse and Goss 2007). In 2006, the gaming industry generated more than \$90 Billion in revenues (American Gaming Association 2006). Of this total, casinos accounted for around \$59 Billion with the remaining generated from alternatives such as online gambling, state lotteries and parimutuel wagering. Recent estimates indicate that online gambling generated about \$5.9 Billion, state lotteries accounted for about \$24 Billion while parimutuel wagering generated about \$3.5 Billion (American Gaming Association 2006). Hence, the casino environment is the dominant revenue generator. Despite the size and significant revenues it generates, so far the casino industry has received little attention by marketing academics.

The casino industry can be characterized as driven by database marketing. Other such industries, for instance, are retail and movie exhibition chains. From the perspective of a retail chain, understanding consumer behavior involves when consumers will visit a store (Morrison and Schmittlein 1981; Helsen and Schmittlein 1993), what brand they buy and what quantity (Gupta 1988). Similar issues are also of great interest to casino operators. Similar to the behavior of movie goers which has been studied extensively (see Eliashberg, Alberse and Leenders 2006 for a review), gamers come to the physical casino facility for excitement and fun (Abt, Smith and Christiansen 1985; Kallick et al. 1979). However, unique aspects of the casino context which deserve research attention are: i) Unlike the retail chain, in the casino industry, there are significantly more cross-market outlet visits and ii) a casino operator observes individual gamer-level data unlike the aggregate level of box-office data that a movie theater owner observes (Swami et al. 1999) and iii) the active participation of gamers as opposed to their passivity in

both the retail and movie contexts. Such active behavior from gamers often requires certain skills that casino operators may find costly to ignore. In comparison with other forms of gambling and in particular to online gambling, there is much greater data available related to gamer behavior within a casino environment (Cotte and Latour 2009). Having access to more data allows researchers to study closely gaming behavior where gamer skill is essential as well as the opportunity to relate such latent skill to gamer demographics and the casino's expected revenue.

Other database driven industries which have also received attention in the past are catalog, credit card and telecommunication industries (see Blattberg, Kim and Neslin 2008; Kumar and Reinartz 2006 for several examples). In the catalog and credit card industry and, to a lesser extent in the telecom industry, the time between customer transactions is a critical characteristic that needs to be explicitly modeled. In the casino industry, the gamers' inter-visit times is also a critical issue. However, the casino context is unique in the sense that gamers' inter-visit times may depend on their previous outcomes (e.g., previous wins/losses). The other key challenge in the catalog /credit card industry is predicting the transaction volume from a customer given the transaction timing. This is also a major challenge in the casino industry. However, unlike the two industries, a unique feature of the casino context is the underlying revenue generation process. The net revenue from a gamer depends on multiple factors such as the choice between table games and slots, how much to wager in each, the house edge and the gamers' latent skill level. We test empirically the performance of our proposed modeling approach, which captures the underlying revenue generation process against others (e.g., Fader, Hardie and Lee 2005a) that do not explicitly incorporate it.

Similar to managers in other database driven industries, casino operators also seek to better understand their customer base and to predict their future expected revenues. Achieving this goal involves two major challenges. First, given the gambling context, the revenues that each player generates in a trip is highly uncertain. As will be shown later in this paper, observed past revenue from a player does not provide a reasonable estimate for her future revenue. For an accurate prediction of future gamer revenues, it is critical to model the underlying revenue generation process and, in particular, tease apart

the “chance” component from the “skill” component of gambling outcomes (Croson, Fishman and Pope 2008). Doing so will allow a casino manager to identify the low-skill players who are likely to provide a higher expected revenue for the casino.

Second, a consumer’s decision to visit and gamble in a casino may be dependent on the past outcome of her previous visit(s). Some previous research in psychology (Clotfelter and Cook 1993; Guryan and Kearney 2008) and economics (Stigler and Becker 1977; Becker and Murphy 1988) have shown that the outcomes of past gambles can affect future gambling behavior. Thus, to obtain an accurate prediction of future revenue, there is a need to consider the *dynamics* of how the outcome of previous trip(s) affects future visit and wagering behavior.¹ For instance, after a gamer loses money in a trip (thus generating revenue for the casino in that trip), she may become less enthusiastic about gambling and may reduce her wager amount on her next trip (thus, the expected revenue will go down on the next trip).

In this paper, we address all the aforementioned challenges and develop an integrated model that captures the underlying revenue generation process. Specifically, we jointly capture players’ visit and wagering behavior. At the heart of our model is a latent “skill” parameter for each player, which affects the quality of the in-play decisions she makes (e.g., whether to stand/hit in blackjack). This in turn increases or decreases the effective odds (and hence the expected return) of her wager and changes the casino’s expected revenue. As will be discussed later, our model allows us to estimate the latent skill for each player through her previous wagering and win/loss patterns. This provides the casino manager with the opportunity to tease apart whether a certain gambling outcome is due to skill, or due solely to chance. Furthermore, our model allows the outcome of previous visit to affect both a consumer’s rate of future visit (thus the waiting time before her next trip) and the amount that she wagers on the next trip. Finally, we also incorporate some demographics information into our model, allowing us to examine how, if at all, demographic information is related to casino visit and wager behavior.

¹ There is past work in marketing especially within a retail context which explores how past choices of consumers can affect their current decisions (e.g., Lattin 1987; Seetharaman 2004).

We apply our model to player-level data obtained from a major casino operator, owning properties in seven different geographical regions in the U.S. The dataset contains information concerning when players' visit a casino, their aggregate trip-level wager amount in slots and table games and net wins/losses. We estimate our model using Bayesian methods. We find that our model provides more accurate, individual-level prediction of future player revenue than predictions based on several alternative models. In addition, we find that some demographics information are indeed related to visit and gambling behavior.

Our research diverges from previous research in the casino context by focusing on capturing gambling behavior and predicting revenue on the individual level. Instead of modeling individual-level revenue, previous studies have modeled industry-level annual revenues using time-series methods (Cargill and Eadington 1978; Nichols 1998). In a separate stream of research, Baker and Marshall (2005) model players' wagered amount and their choice of gambling venues. They do not model how wagers are linked with wins, and how players' skill can affect their wins from table games. In addition, they use cross-sectional survey data (rather than player-level longitudinal play data) and hence cannot address heterogeneity in skill level across players.

The remainder of this paper is organized as follows. Section 2 outlines our modeling framework. Section 3 describes the data. Section 4 reports our empirical results. Section 5 discusses managerial implications in regards to individual-level revenue prediction. Section 6 concludes with future research directions.

2. Model

In this section, we develop a model that captures the overall revenue generation process. We begin with a description of gamers' visits followed by their wagering behavior. Our overall modeling framework is shown in Figure 1.

[Insert Figure 1 about here]

We begin by specifying a model of visit in Section 2.1, which captures when a gamer will visit a casino, and which region she visits. As we discussed earlier, these decisions are influenced to a large extent by the gamer’s gambling wins/losses on her previous trip. Once a gamer enters a casino region, we model the total amount of wagers she makes, and how she divides her wager between slot and table games in Section 2.2. Next, we model the gambling outcomes in slot games and table games separately, in Section 2.4 and 2.5, respectively. This part of our model is critical as it captures how wagers are translated into wins/losses. The underlying drivers of gambling outcomes are the effective “house advantage” (the expected loss per dollar wagered by the player) that is driven by the odds that the casino offers, the player’s choice of wager, and the player’s skill level (in the case of table games). The concept of house advantage is discussed in detail in Section 2.3.

2.1 Model of visit

We begin by modeling *when* a player visits a casino region and *which region* she visits. We closely follow previous work on inter-purchase time models (e.g., Gupta 1991). Figure 2 illustrates the data for a single player along a timeline.

[Insert Figure 2 about here]

Let the data collection time period be denoted by time $t = 0$ to $t = T$; let player i make N_i trips during this time, and let the time span between the $(k-1)$ ’th visit and the k ’th visit be w_{ik} . Note that the inter-visit time before the first visit (w_{i1}) and after the last visit ($w_{i(N_i+1)}$) are not fully observed as our data is both left- and right-censored. We denote the observed inter-visit time before the first trip and after the last one as $w_{i1}^* (= t_{i1})$ and $w_{i(N_i+1)}^* (= T - t_{iN_i})$, respectively.

We use the competing risk framework (Klein and Moeschberger 2003) to model a player’s inter-visit time and the choice of gambling region. We assume that for each player i and trip k , there are latent inter-visit times associated with all J regions. The gamer chooses to visit the region that has the shortest inter-visit time. Formally, for gamer i , region l and trip k , we assume that the latent inter-visit time (s_{ilk})

is generated by an exponential distribution with λ_{ilk} as the rate of visit. Let gamer i visit region $z_{ik}=j$ on an uncensored trip k ($k \in \{2, \dots, N_i\}$). The likelihood for such an observation can be written as follows:

$$l(w_{ik}, j) = l_{ik} = f_{\text{exp}}(w_{ik} | \lambda_{ijk}) P(s_{ilk} > w_{ik} \quad \forall l \neq j). \quad [1]$$

Here f_{exp} denotes the exponential probability density function. Assuming that the visit to each region is affected only by the underlying rate to that region, we obtain:

$$l_{ik} = \left(\lambda_{ijk} e^{-\lambda_{ijk} w_{ik}} \right) \prod_{l \neq j} e^{-\lambda_{ilk} w_{ik}} = \lambda_{ijk} \exp\left(-w_{ik} \sum_{l=1}^J \lambda_{ilk} \right). \quad [2]$$

Next, we focus on the first and last censored observations. Due to the memoryless property of the exponential distribution, the likelihood of the first observation is also given by an expression similar to Equation [2]. The last observation ($k = N_i + 1$) is right-censored. Thus,

$$l_{i(N_i+1)} = \exp\left(-w_{i(N_i+1)}^* \sum_{l=1}^J \lambda_{il(N_i+1)} \right). \quad [3]$$

The likelihood of the entire data (i.e., for the N_i trips) available for player i can be written as:

$$L_i = \left(\prod_{k=1}^{N_i} \lambda_{ijk} \right) \exp\left(- \left(\sum_{k=1}^{N_i} w_{ik} \sum_{l=1}^J \lambda_{ilk} + w_{i(N_i+1)}^* \sum_{l=1}^J \lambda_{il(N_i+1)} \right) \right), \quad [4]$$

We link the rates for a trip to each region l to past win/loss using the following specification:

$$\lambda_{ilk} = \lambda_{il} \exp(\nu R_{ik-1}). \quad [5]$$

Here, λ_{il} is the base rate of visit to region l and R_{ik-1} is a player- and trip-level covariate that captures win/loss for player i from her most recent trip (i.e., trip $k-1$) to any region. R_{ik-1} takes value 1 if player i wins money in her previous trip, and -1 otherwise. An inclusion of this variable allows a player's win/loss from past trips to influence the rate of visit for the current trip. The parameter ν is the sensitivity to this past covariate – a positive coefficient indicates that past wins lead a player to have a shorter inter-visit time while past losses lead her to have a longer inter-visit time.

2.2 Model of wagering behavior

As defined in the model of visit, player i visits region z_{ik} during her k 'th trip. Let y_{ik} denote her wager amount in that trip. This total amount is allocated between slots and table games. We denote the amount wagered in slots as y_{ik}^{SLOT} and the amount wagered in table games as y_{ik}^{TABLE} . Further, the proportion of total wager spent on slots is denoted by ϕ_{ik} . Formally,

$$y_{ik} = y_{ik}^{SLOT} + y_{ik}^{TABLE}, \text{ where } y_{ik}^{SLOT} = y_{ik}\phi_{ik}. \quad [6]$$

We model the total wager amount and proportion as:

$$\log(y_{ik}) = \alpha_i + \beta_{z_{ik}} + \kappa R_{ik-1} + \varepsilon_{ik}^y, \text{ where } \varepsilon_{ik}^y \sim N(0, \sigma_y^2), \quad [7]$$

$$\text{logit}(\phi_{ik}) = \mu_i + \eta_{z_{ik}} + \varepsilon_{ik}^\phi, \text{ where } \varepsilon_{ik}^\phi \sim N(0, \sigma_\phi^2). \quad [8]$$

As defined earlier, R_{ik-1} is a player- and trip-level covariate that captures win/loss for player i from her most recent trip (i.e., trip $k-1$) to any region. The parameter κ is the sensitivity to this past covariate – a positive coefficient indicates that past wins lead a player to wager more in the current trip while past losses lead her to wager less. The parameters α_i and $\beta_{z_{ik}}$ capture the gamer- and region-level effect respectively on the total wager amount. Similarly, in Equation [8], which captures the proportion of wager amount spent on slots, the term μ_i is a player-specific effect and $\eta_{z_{ik}}$ is a region-specific effect.

2.3 House advantage

The term house advantage (or house edge) refers to the expected revenue by the casino per dollar wagered by a player (Epstein 1997; Packel 2006). To illustrate, suppose the player wagers \$ y , then, the house advantage, denoted as r , can be written as follows:

$$E(x| r, y) = r * y, \quad [9]$$

where $E(x| r, y)$ denotes the revenue that the casino expects to make from the \$ y wager amount.

More specifically, consider a slot game called “21 Bell” in which three separate wheels, each containing eight different symbols (7, Bar, Melon, Bell, Plum, Orange, Cherry and Lemon) are spun. If

the result of a \$1 wager spin is any of the “winning combinations” (first column of Table 1), the player receives a payoff (second column of Table 1).

[Insert Table 1 about here]

We can calculate the associated house edge using the following expression:

$$\text{House Advantage } (r) = 1 - \sum_i (\text{PAYOFF}_{\text{combination } i} \times \text{PROB}_{\text{combination } i}). \quad [10]$$

The house advantage associated with the 21 Bell slot machine is 5.55%. This indicates that in the long run, a player will tend to lose 5.55% of her total wagered amount.

The house edge may differ across slot machines. In some cases, different casinos may offer differing house advantages even for the same slot machine. For the same 21 Bell slot machine, some casinos may offer a different payoff table from that shown in Table 1. For instance, they may offer a jackpot for the combination 777, which pays \$500 (instead of \$200 as shown in Table 1). In this case, the house advantage for the 21 Bell slot machine will be 1.80%.

In many table games, a gamer is required to make decisions during the course of the play. In such games, the house advantage also depends on how good the player’s decisions are. For instance, consider the game of blackjack, which is the most popular table game among players (Harrah’s Report 2006). In blackjack, a player is dealt two cards in the beginning and then given the opportunity to select more. The goal of a player is to beat the dealer by getting as close to a total of 21 without going over (referred to as “busting”). After obtaining each card, the player is given the option of either “hitting” (i.e., getting an additional card) or “standing” (i.e., not getting any additional card). As this game involves a player’s decisions, the effective house advantage depends on whether the player hits/stands at the right situation. If she makes each decision correctly, the house advantage is around 0.5% (Griffin 1999). However, a deviation from the overall correct strategy results in a much larger house edge. In particular, if a “never bust” strategy is followed (i.e., stand whenever a total of 12 or more is observed), the house advantage rises to 3.9%. Thus, players with lower skill will incur a higher house advantage (as compared to players

with higher skill) and thereby give the casino higher expected revenues. This suggests that a characterization of such player skill is important for revenue predictions.

To summarize, the house advantage that a gamer faces depends on several factors. In slots, it depends on the choice of game and the odds offered by a casino. In table games, in addition to these two factors, the house edge also depends on player-specific skill. In the next subsection, we show how players' wagered amount link to their wins/losses.

2.4 Modeling net wins/losses in slot machines

To illustrate the nature of the net wins/losses data that the casino collects, consider the example shown in Table 2.

[Insert Table 2 about here]

Table 2 shows the within-trip data for a player (A) who visits Region 1. In this trip, he plays on two different slot machines – one with a house advantage of 0.2 and another with a house advantage of 0.3 (column labeled as “Slot Machines/ House Advantage”). He wagers \$100 in the former slot machine and \$200 in the latter machine. Thus, he bets a total of \$300 over the entire trip. The column labeled as “Expected Revenue” shows the revenue that the casino expects (theoretically) to make from this particular combination of the wagered amounts and the house edges of the two chosen slot machines. The casino then expects to make a total of \$80 from this player's betting in slots during the given trip. The final column (“Actual Revenue”) is the actual revenue that the casino makes. Note that the actual revenue may be either higher or lower than the expected revenue. In addition, the actual revenue may be positive (the casino receives money) or negative (the player makes money). The net actual revenue that the casino makes in this example from the player over the trip is \$20 (= \$40-\$20). The last row labeled as “Aggregated Information” is the information from this player summarized at the trip level — he wagers a total of \$300, the casino's expected revenue is \$80 and its total actual revenue is \$20. This is part of player-level data that many casino operators keep track of.

In our dataset (described in greater detail in a later section), we have aggregate trip-level information for each player (e.g., \$300, \$80, \$20) but do not have any intra-trip details (i.e., \$100, 0.2, \$40; \$200, 0.3, -\$20). Thus, consistent with the information in the dataset, for a gamer i , during trip k and chosen region z_{ik} , we employ an *aggregate* house advantage associated with slots (r_{ik}^{SLOT}):

$$r_{ik}^{SLOT} = \frac{\text{Aggregate Trip-level Expected Revenue in Slots}}{\text{Aggregate Trip-level Wagered Amount in Slots}}. \quad [11]$$

Given the above metric, the aggregate house advantage for player A in Table 2 is 0.27 (= \$80/\$300). This aggregate house advantage can vary across players, trips and regions depending on the odds offered in different regions and differing allocations to slot machines.

We model the aggregate trip-level house advantage as a function of a person-level effect, which captures differences in the choice of slot machines across players, and a region-level effect that captures the differences in odds offered across different gambling regions. Formally,

$$r_{ik}^{SLOT} = \tau_i^{SLOT} + \psi_{z_{ik}}^{SLOT} + \varepsilon_{ik}^{r_{SLOT}}, \text{ where } \varepsilon_{ik}^{r_{SLOT}} \sim N(0, \sigma_{r_{SLOT}}^2). \quad [12]$$

Here, τ_i^{SLOT} denotes a person-level parameter, $\psi_{z_{ik}}^{SLOT}$ is a region-level parameter for the chosen region, and $\varepsilon_{ik}^{r_{SLOT}}$ captures any unexplained remaining effects.

Let x_{ik}^{SLOT} denote the actual net revenue to the casino from player i wagering in slots during the k 'th trip in region z_{ik} . Within our notation, $x_{ik}^{SLOT} > 0$ indicates that the casino makes money, while $x_{ik}^{SLOT} < 0$ indicates that the casino loses money. Following Equation [9], the total trip-level expected amount that the player wins/loses in slots machines is:

$$E(x_{ik}^{SLOT} | y_{ik}^{SLOT}, r_{ik}^{SLOT}) = y_{ik}^{SLOT} r_{ik}^{SLOT}, \quad [13]$$

where y_{ik}^{SLOT} is the amount wagered on slots and r_{ik}^{SLOT} is the trip-level aggregate house advantage. Next,

given the skewness of the payoff in each slot wager, we approximate the total payoff as the expected payoff (in Equation [13]) plus an error term which is skew-t distributed, as follows²:

$$x_{ik}^{SLOT} = y_{ik}^{SLOT} r_{ik}^{SLOT} + \sigma_{X_{SLOT}} \varepsilon_{ik}^{x_{SLOT}}, \text{ where } \varepsilon_{ik}^{x_{SLOT}} \sim \text{Skew} - t. \quad [14]$$

2.5 Modeling net wins/losses in table games

Table games involve active in-play decision making on the part of players. As described earlier, the house advantage that each player receives may depend on her “skill” level. To model the skill level of a player, we posit, consistent with industry practice, the existence of a prototypical or average (we use the term interchangeably) gamer who has an average level of skill. Given the assortment of table games and bets that a particular player selects to engage in during a trip, the aggregate house advantage for the prototypical player can be calculated by combining the exact same assortment of games and bets with the house edges that she would receive. We illustrate this notion with a hypothetical example.

Consider Table 3 that shows a particular assortment of wagered amounts in different table games (Games 1, 2 and 3) for a player i during a trip to Region 1.

[Insert Table 3 about here]

The column (“Prototypical Player House Advantage”) shows the house advantage that the *average* player receives for this exact assortment of wagering amounts and chosen games. For the total bet of \$300, the “Expected Revenue” is calculated as before, i.e., Total Aggregate Expected Revenue is \$60 (=100*0.2 + 100*0.1 + 100*0.3). As discussed earlier, the casino keeps track of aggregate trip level information for each player (e.g., \$300, \$60, \$28) but not any intra-trip details (i.e., \$100, 0.2, \$40; \$100, 0.1, \$8; \$100, 0.3, -\$20). Thus, we denote the “Aggregate prototypical house advantage” in table games based on the trip level information of player i (assortment of games and wagers) to region z_{ik} combined with the house advantage received by the prototypical player as \bar{r}_{ik}^{TABLE} . This is calculated as follows:

² We also modeled the total payoff using a normal distribution. Given the skewness in the payoffs, a skew t -distribution (Jones and Faddy 2003), with fatter tails than the normal distribution and more skewed than the t -distribution, gave a better fit.

$$\bar{r}_{ik}^{TABLE} = \frac{\text{Aggregate Trip-level Expected Revenue based on house edge received by an } \textit{average} \text{ player}}{\text{Aggregate Trip-level Wagered Amount in Tables for player } i} \quad [15]$$

Note that the aggregate prototypical house advantage for a player in table games is based on two drivers: i) the selection of games and wagers of the player, and ii) the house edges received by an average skilled player for that same assortment of games and wagers. For this example, \bar{r}_{ik}^{TABLE} is $\$60/\$300 = 0.2$. This aggregate prototypical house advantage may vary across players, trips, and regions depending on the types of table games offered in different regions and differing allocations to table games. We model

\bar{r}_{ik}^{TABLE} as follows:

$$\bar{r}_{ik}^{TABLE} = \tau_i^{TABLE} + \psi_{z_{ik}}^{TABLE} + \varepsilon_{ik}^{\bar{r}_{TABLE}}, \text{ where } \varepsilon_{ik}^{\bar{r}_{TABLE}} \sim N(0, \sigma_{\bar{r}_{TABLE}}^2). \quad [16]$$

Here, τ_i^{TABLE} is a person-level parameter, $\psi_{z_{ik}}^{TABLE}$ denotes a region-level parameter associated with the chosen region, and $\varepsilon_{ik}^{\bar{r}_{TABLE}}$ captures any unexplained effects.

In table games, players' skill can alter the house advantage that they actually receive. We model this aspect and specify the actual aggregate house advantage received by player i as follows:

$$r_{ik}^{TABLE} = \bar{r}_{ik}^{TABLE} - \gamma_i. \quad [17]$$

Here, r_{ik}^{TABLE} denotes the actual aggregated table house advantage received by player i in trip k , and γ_i represents his level of skill. We estimate this skill from our model. The parameter γ_i can range from $-\infty$ to $+\infty$. If γ_i is positive (negative) then the actual house edge for player i will be lower (higher) than the house edge for an average player.

We use the Central Limit Theorem (see Appendix) to approximate the distribution of the revenue that the casino wins/loses from gamer i in table games on a trip k (x_{ik}^{TABLE}) by a normal distribution:

$$x_{ik}^{TABLE} = y_{ik}^{TABLE} r_{ik}^{TABLE} + \varepsilon_{ik}^{x_{TABLE}}, \text{ where } \varepsilon_{ik}^{x_{TABLE}} \sim N(0, y_{ik}^{TABLE} \sigma_{x_{TABLE}}^2). \quad [18]$$

Equation [18] links the win/losses to actual aggregated table house advantage and, via Equation [17], to the player-level skill. This completes the model for wagering decisions and how wager amount influences players' wins/losses.

2.6 Heterogeneity across players

Till now we focused on one player (player i) and showed the model specification. Next, we incorporate heterogeneity across players. Let $\ln(\lambda_i)$ be $\{\ln(\lambda_{i1}), \ln(\lambda_{i2}), \dots, \ln(\lambda_{iJ})\}$, where λ_{il} is the base rate of player i 's visits to region l . We assume that all individual-level parameters in the model are drawn from a common multivariate normal distribution - $(\ln(\lambda_i), \alpha_i, \mu_i, \tau_i^{SLOT}, \tau_i^{TABLE}, \gamma_i)' \sim N(\mu^{IND}, \Sigma^{IND})$. The use of this multivariate prior distribution achieves two objectives. First, the gamer-specific base rate parameters across regions are allowed to be correlated. Second, the parameters in the model of visits (visit frequency) and wager/wins are linked. The estimation results will indicate to the strength of this linkage.

Similarly, the region-level parameters are drawn from a different multivariate normal distribution, $(\beta_l, \eta_l, \psi_l^{SLOT}, \psi_l^{TABLE})' \sim N(\mu^{REG}, \Sigma^{REG})$. For model identification, both vector μ^{REG} and the sum of each of the parameters across the J regions ($\sum_l \beta_l, \sum_l \eta_l, \sum_l \psi_l^{SLOT}$ and $\sum_l \psi_l^{TABLE}$) are set to 0. We adopt a hierarchical Bayesian framework for simulation-based inference using MCMC methods. The priors for the unknowns and the full conditionals are available from the authors upon request.

2.7 Model-based predictions

As our model contains past win/loss covariates, a simulation procedure is required to obtain model-based predictions of various metrics of interest, such as wager amounts, and the associated wins/losses. The prediction of total expected revenue, which accrues from a gamer, incorporates all the intermediate predictions of how many times a gamer visits a region, how much she wagers and her wins/losses. We explain how we generate the expected revenue.

For a time period T^* , the casinos' expected revenue that player i will generate for region l ($\text{Rev}_{il}(T^*)$) is given by:

$$\text{Rev}_{il}(T^*) = \sum_{k=1}^{N_i(T^*)} x_{ik} I_{\{l\}}^k, \quad [19]$$

where, $N_i(T^*)$ denotes the number of times player i visits any of the J regions during the period T^* , x_{il} is the total revenue generated from both table games and slots in region l and $I_{\{l\}}^k$ is an indicator function which takes a value of 1 if player visits region l and is 0 otherwise.

For each player and each posterior draw of her parameters from the MCMC chain, we generate a total of 200 paths where a path for the player represents her visits to the various regions during the period T^* and the associated wagers/wins during each visit. Using the simulated wins/losses in each trip, we also generate the indicator variable that captures the effect of past wins/losses on gamer behavior during the next trip.

We generate 200 such paths for each gamer and each sample of her individual-specific parameters. We then average over all 200 paths. Finally, we incorporate the uncertainty in the individual-specific parameters by averaging the probabilities over all sampled values of the parameters from the MCMC chain.

3. Data

Our dataset is provided by a major casino operator in the United States. It contains records of 1571 customers over a 28-month period from December 2004 to April 2007. During this time period, these customers made a total of 8641 trips, where a trip is a visit to a casino region. For each player-trip combination, we have information on when the trip takes place, the visited region, and for both slots and table games separately, the total amount that the gamer wagers, her trip-level aggregate house advantage and the net actual amount that she wins/loses. For a subset of 400 players, we also have demographics data (age and gender) along with the above discussed transactional information.

Table 4 shows the overall data aggregated at the region level. For reasons of confidentiality, we cannot name these regions but these are regions such as Las Vegas and Atlantic City. For each of the seven regions, the table shows the number of observed visits, the total wagered amount, and the total amount of gambling revenues.

[Insert Table 4 about here]

Table 5 shows the summary statistics of the data for the players. There is substantial variability in terms of both visit frequencies and wagering behavior. The number of visits that a player makes varies from 1 to 122 over the 28-month period, with an average across players of around 6 visits. In terms of wagering behavior, the mean aggregate amount wagered in slots (tables) across all the trips per player is around \$37,800 (\$1000). Further, not reported in the table, gamers tend to lose around 5.6% (17.6%) of their total wagered amount in slots (table games).

[Insert Table 5 about here]

A closer investigation reveals that about 40% of the players in the data (635 players) make a single visit during the data period. For the remaining players, the mean inter-visit time is 165 days (around 5 months). More than 25% of these players visit more than one region. This percentage of customers making cross-region trips is consistent with industry reports (see CIO magazine, <http://www.cio.com.au/index.php/id;609782439>).

4. Empirical Results

We begin with a comparison of the goodness-of-fit of four model specifications based on their log-marginal likelihood (LML). We compare our proposed model (Model 1) with three nested versions – Model 2 ignores past covariates (i.e., $\kappa = v = 0$) but includes the construct of latent skill, Model 3 ignores the skill level (i.e., $\gamma_i = 0$) but allows for past covariates and Model 4 excludes both past covariates and skill.

4.1 Models comparison

We use the MCMC draws from the simulation-based inference to calculate the LML for each model. Higher LML denotes a better model. The LML for Model 1 is -62139.2, for Model 2 is -62259.0, for Model 3 is -62614.0 and for Model 4 is -67095.0. Thus, Model 1 is best supported by the data based on the criterion described in Kass and Raftery (1995). A comparison of the log-marginal likelihoods shows that inclusion of both the construct of latent player skill and past covariates are important within our modeling framework.

We further compare all four models on the percentage differences between the fitted values from our model for the various metrics of interest aggregated over all players and regions and the observed values. The fitted values are obtained by applying the simulation procedure described in Section 2.7 in-sample. Table 6 shows these comparisons and indicates that our proposed model best fits the data. The results also show that past covariates and latent skill contribute to the model in distinct ways. An inclusion of past covariates helps in improving the fit of the wager amount, which in turn has an effect on revenues from both slots and table games. An inclusion of the latent skill level improves the fit of the revenues from table games. This latter result has face validity as player skill only affects his revenues from table games and does not directly influence his revenues from slots.³

[Insert Table 6 about here]

4.2 Parameter estimates

We begin by examining the player-level parameters (μ^{IND}) for our proposed model.⁴ Table 7 shows the estimated parameters together with their 95% posterior interval. As expected, the parameter related to amount of wagering (α) is positive. The coefficient capturing the proportion of wager spent on slots (μ) is positive as well and implies that, on average, players allocate more of their wagering amount

³ We also validated our proposed model by comparing the model-fitted wagers and revenues (for both slots and table games) with their observed values. We found that our model provided a good fit to the data – the observed values were within the 95% posterior intervals around the model-fitted values. Due to space constraints, we do not report these results. They are available from the authors upon request.

⁴ Due to space constraints, we do not report the details on the region-level parameters for our proposed model. These estimates as well as those for Models 2, 3 and 4 are available from the authors upon request.

to slots than to table games. We find that the average skill level (γ) is marginally negative. This suggests that on average, gamers in our database incur a higher house advantage than the prototypical player. In addition, after controlling for region-level effects, the average (across players) house advantage within slot machines (τ^{SLOT}) is around 0.10 (i.e., 10%) while the house advantage faced by the prototypical player in table games (τ^{TABLE}) is about 0.18 (i.e., 18%). These findings are also consistent with industry reports on the average house advantage of around 20.5% (Vogel 2007). As the logarithm of player-level rate parameters for the seven regions is normally distributed across players (see Section 2.6), we report its population mean for all the seven regions. Finally, not reported in the table, we find that parameters κ and ν , which capture the effect of past covariates, are positive and significant. The parameter κ is estimated to be 0.22 with a 95% posterior interval of (0.19, 0.26) while the parameter ν is estimated to be 0.13 with a 95% posterior interval of (0.09, 0.17). These results suggest that past wins influence gamers to both visit a casino faster (i.e., decrease their inter-visit time) and wager more.

[Insert Tables 7 about here]

Table 8 shows the estimated correlations among the player-level parameters. We find a positive but low correlation ($\rho = 0.25$) between wager amount (α_i) and skill level (γ_i).⁵ This indicates that an observable player-level characteristic such as wagering amount is not a perfect indicator of player skill. We also find a moderate level of correlation among several of the visit parameters ($\lambda_{i1}, \dots, \lambda_{i7}$) across different regions. For instance, the correlation between λ_{i3} and λ_{i7} is 0.45 and the correlation between λ_{i2} and λ_{i5} is -0.56. These correlations have face validity as Region 3 and Region 7 are geographically close to each other and hence a gamer might consider visiting both regions. Region 2 and Region 5 are far apart. This indicates that there is dependence in gamers' visit patterns across regions. Finally, in general, the individual-level parameters associated with the model of wagering behavior and winning are weakly correlated with the visit parameters. For instance, only the visit parameter to Region 1 (λ_{i1}) is moderately

⁵ Cohen (1988) suggests that a correlation of less than 0.3 is weak, between 0.3 to 0.5 is moderate and greater than 0.5 is strong.

correlated with the wager amount ($\rho = 0.38$) while the remaining show weak correlation. Similarly, the strongest correlation between the visit parameters and skill level is only 0.28 (this is for visit parameter to Region 6). This suggests that a player's frequency of trips is also not a good metric for capturing player skill.

[Insert Table 8 about here]

These findings indicate that i) there is dependence in gamers' patterns across regions and ii) observable player characteristics such as frequency of trips and wagering amount are not perfect indicators of player skills.

5. Managerial implications

5.1 Individual-level revenue prediction

An important task for casinos is to predict gamer-level future revenues. For making the predictions, we divide our data time period into two equal halves, and use only the first half of the data (14 months) to calibrate our proposed model. We then create three holdout datasets of 5, 10 and 14 months respectively. We compare our model-based out-of-sample predictions with those from three alternative models, which are based on current practice as well as on the extant modeling literature.

The first alternative model (termed as "Model A: Historical Revenues-based Model") uses observable player-level revenue within the calibration data and assumes that a proportional level of revenue will be present in the holdout dataset. This is similar to the prediction models often used within other database marketing contexts where measures of customers' present behavior are predictors of their future behavior (Berry and Linoff 2004; Parr Rud 2001). For this model, the proportionality constant is dependent on the number of months in the holdout vis-à-vis the 14 months in the calibration. For instance, for the holdout dataset of 5 months, the alternative model assumes that the level of revenue for each player will be $0.357 (= 5/14)$ times the revenue in the calibration dataset.⁶ Similarly, for the holdout

⁶ We also used a different methodology to make revenue predictions for the alternative model. For the holdout dataset with 5 (10) months, we assumed that the level of revenue for each player was the same as present in the last

dataset of 14 months (the same length as the calibration dataset), the alternative model assumes that the level of revenue for each player will be the same as that in the calibration dataset.

The second alternative model (termed as “Model B: Gamma-Exponential Visits with Stochastic Net Revenues Model”) overcomes the limitations of the previous deterministic model by stochastically modeling both gamer visit behavior and revenues conditional on visit. We use modeling assumptions similar to past work (e.g., Fader, Hardie and Lee 2005a, 2005b). For the visit model, the inter-visit time for each gamer is assumed to be exponentially distributed and the gamer-specific visit rate is gamma distributed across gamers. Note that, unlike our proposed model, the gamer-specific visit rate in this model is time invariant. For the revenue model, we assume that the revenue in each trip from a gamer varies randomly around an unobserved mean value, which is distributed across gamers. As the revenue can be positive (casino makes money) or negative (gamer makes money), we assume that the actual trip revenue for each gamer is normally distributed around his/her mean value and specify a normal heterogeneity across gamers in the mean value of revenue. This model, unlike our proposed model, ignores the underlying revenue generation process i.e., it does not decompose the revenue into wagers, how the wagers are combined with the relevant house advantage and the skill level of players to give rise to the net revenue.

The third model (termed as “Model C: Gamma-Exponential Visits with Stochastic Theoretical Revenues Model”) is a variant of the second model. Here, we model gamer- and trip-level *theoretical* revenues, which are based on wager amounts and the house edges of the games (See sections 2.4 and 2.5), instead of their actual revenues. Thus, we assume that theoretical revenue from each gamer from a trip varies randomly around an unobserved mean value, which is different across gamers. As before, the theoretical trip revenue for each gamer is normally distributed around his/her mean value and this mean value of revenue is assumed to be normally distributed across gamers. The visit model is the same as that in Model B. As we model the theoretical revenues, this model does decompose the revenues as arising

5 (10) months of the calibration dataset. The predictions from the alternative model based on this assumption were worse than the reported predictions. These results are available from the authors upon request.

from wagers and the associated house advantage but does not incorporate the heterogeneity in skill across gamers as our proposed model does.

We employ three metrics to evaluate the models' predictions: i) predicted ranking of players in terms of their revenues, ii) the root mean squared error (RMSE), and iii) cumulative lift curve and the corresponding Gini coefficient (Gini 1921). For the rank based comparison, we evaluate the performance of the predicted ranks vis-à-vis the actual ranks using Spearman's rank correlation coefficient (Lehmann 2006), which ranges from +1 (perfect agreement to actual ranking) to -1 (perfect disagreement to actual ranking). For the RMSE calculation, in the proposed model, we predict the total revenue from each player across all the seven regions and then compare it with its observed value. (For the alternative models, please see the above description of how we predict player-level revenues.) A cumulative lift curve (also called a Lorenz curve) curve is a graphical representation that allows a comparison of the percentage of one variable against the percentage of another (Blattberg, Kim and Neslin 2008; Gastwirth 1972). Of particular interest to casino managers is the percentage of total revenue that the casino collects from the top 10 %, 20 %, 50 % (and so on) of all players. We determine player rank either by (i) our full model, or (ii) alternative models.

Table 9 shows the comparison among all four models based on Spearman's correlation and RMSE across the three holdout datasets.

[Insert Table 9 about here]

The ranking based on our model is better than the ranking based on the alternative models. For instance, when the holdout dataset is 14 months, the ranking based on our model has a Spearman correlation of 0.59 while those based on Model A, Model B and Model C have a lower Spearman correlation of 0.43, 0.44 and 0.45, respectively.⁷

⁷ Similar to the deterministic revenue model, we also use other observable statistics such as frequency of visits, wager amounts and other combinations to develop player ranking. The ranking system based on our proposed model is better (i.e., has a higher Spearman correlation) than the ranking systems based on any of these observable statistics. In particular, for the holdout dataset of 14 months, the ranking system based on visit frequency had a correlation of 0.36 and the ranking based on frequency and wagered amount had a correlation of 0.46. The results of other

We find another indication for the superiority of our full model when we compare the RMSE across the four models. For each holdout dataset, we find that the RMSE from our model is smaller than those from the alternative models. For instance, when the holdout dataset is of 14 months, the RMSE of our model is 5556.23 while those from Model A, Model B and Model C are 6261.73, 6575.92 and 6531.45, respectively. These differences indicate that there is significant improvement from using our model.

We make a third comparison among the four models based on their Lorenz curves. Figures 3a-3c show Lorenz curves that compare our model-based predictions with those from the alternative models for each of the three holdout datasets. An “ideal” Lorenz curve would be one which starts out flat (the gamers who are predicted to provide low revenue actually generate low level of revenue) and with an increase in predicted revenue, the model starts to capture a greater percentage of the observed revenue. Thus, the farther the Lorenz curve for a model is from the 45 degree line, the better the model is in identifying high revenue players. The area between the 45-degree line and the Lorenz curve can be interpreted as the ability of the model to differentiate between high- and low-revenue-generating players. In addition, the Gini coefficient associated with a Lorenz curve is the ratio of two areas – the area between the 45-degree line and the Lorenz curve and the overall area under the 45-degree line, which is $\frac{1}{2}$ (Blattberg, Kim and Neslin 2008). A model’s Gini coefficient close to 0 indicates that its predictions are no better than random while a coefficient close to 1.0 indicates that it provides perfect predictions.

[Insert Figures 3a-3c about here]

As can be seen in the figures, for each of the three holdout datasets, the Lorenz curve for our proposed model (the solid line) lies below the Lorenz curves for all three alternative models. Table 9 shows the Gini coefficients corresponding to the Lorenz curves of all four models and each of the three holdout datasets. A comparison shows that our proposed model has a higher Gini coefficient compared to all three alternative models. For instance, in the dataset with 5 months holdout, the Gini coefficient of our

combinations are available upon request. All combinations indicate that predictions based on observable player metrics are inferior relative to our model predictions.

proposed model is 0.884 while those for Models A, B and C are smaller at 0.804, 0.826 and 0.838, respectively.

All three comparisons indicate that our model provides better individual-level revenue predictions compared to the predictions based on alternative models of the type used in past research and current business practice. This clearly shows that incorporating the underlying revenue generation process and the heterogeneity in skill across gamers are critical for accurate predictions of future value of gamers.

5.2 Demographic characterization

Another important managerial goal is to understand how player-level demographics relate to various behavioral aspects such as amount of wagering, skill level and others.

To study how demographics information is related to our model parameters, we re-estimate our proposed model for a subset of 400 players (as described in Section 3.1) and include their demographics (age, gender) in the heterogeneity specification across players.⁸ We assume that the player-specific parameters are drawn from a multivariate normal distribution i.e.,

$(\ln(\lambda_i), \alpha_i, \mu_i, \tau_i^{SLOT}, \tau_i^{TABLE}, \gamma_i)' \sim N(Z_i \mu^{IND}, \Sigma^{IND})$ where Z_i contains player demographics. This is a typical way of specifying heterogeneity across players with demographics (Rossi and Allenby 2003).

We estimate our model and find several interesting results, which are shown in Table 10. We find that older players wager more and are inclined to allocate more of their wagers on slot machines vis-à-vis table games. We also find that women have a tendency to wager more than men and, compared to men, they allocate more of their wagers to slots versus table games. We also find some directional, though not significant, evidence that men have higher level of skill than women. Finally, there is no evidence that age or gender affects the rate of visit to a region (with the only exception being Region 3 where we find that older players are more likely to visit).⁹ Having additional data with respect to demographics such as home

⁸ While we were supplied with all the trip-related information of the players, due to privacy reasons, the casino operator did not feel comfortable sharing individual-level demographics information for more than 400 players. In addition, it shared only age and gender information for these 400 players.

⁹ To examine further the usefulness of having the player-level age/gender demographics, we also estimated a model without demographics on the data from the 400 players. We used the MCMC draws from the simulation-based inference to calculate the LML for the two models. The LML for the model with demographics was -39610.60 while

address, profession, level of education, relationship status might yield finer differences and it certainly deserves further research attention.

[Insert Table 10 about here]

6. Conclusions and Future Research

We developed a new integrative model for gamers' behavior in a casino setting and incorporated various context-specific features. Our integrated framework contains several components: a model for wager amount and net wins/losses in slots and table games, a model for the proportion that he wagers on slots and table games respectively, a description of the odds offered by different regions, a specification of player's latent skill level and the model for visit and choice of region.

The results indicate that our modeling framework, which decomposes the underlying gamer revenue generation process, is quite useful. At the aggregate level, it provides a very good fit to the data in terms of region-level wager and revenue amounts in slot machines and table games as well as in terms of the total number of gamer visits. At the individual level, the integrated model provides better evaluation of future player revenue vis-à-vis those from more typical models.

With suitable modifications, our modeling approach can be applied to other marketing contexts. For instance, in the context of salesforce management, a salesperson's compensation is based on his/her sales performance (John and Weitz 1989). It is unclear how much of the sales are due to product attributes (e.g., a product being good sells by itself) and how much is due to his/her latent skill level in convincing buyers. An estimation of such skill thus becomes vital for capturing true differences in performance across salespeople. Another context is the area of hotel revenue management. Hotels are interested in predicting the amount of revenue they would receive from each customer (Bitran and Mondschein 1995).

for the model without demographics was -39622. In comparing the player-level revenue predictions from both models (with/without demographics) with the observed player-level revenue, we found that the Spearman rank correlation between the predicted player ranks based on the model with (without) demographics and actual player ranking was 0.87 (0.85), the RMSE from the model with and without demographics, across all seven regions, were very similar (around 12357), and the areas under the Lorenz curve for the model with (without) demographics were 0.359 (0.357).

However, customers differ in their usage of services such as mini bar, business center, room internet service and others, which affect the overall revenue received from them. Clearly, for accurate prediction of revenues it is important to capture such underlying differences in service usage among customers. A similar problem occurs for other services such as car rentals where customers differ in their usage of services such as in car navigation system, satellite radio, accident insurance and others.

Within a boarder perspective, our research is related to other studies in entertainment industry. While most past research in marketing that focuses on the entertainment industry addresses the motion picture and home entertainment sectors (Eliashberg et al. 2006), some attempts worth noting in other sectors include Broadway shows (Reddy et al. 1998) and theater performance (Putler and Lele 2003). Our contribution in this paper is also to a non-motion picture sector -- the gaming sector.

The gaming industry offers modeling challenges as well as a large volume of rich data that allow modelers to generate relevant managerial insights. The current analysis is a first step in that direction. Our model can be further extended by considering the following research opportunities:

- (i) *Share of wallet*: In order to target the customer base, casinos should use their proprietary database along with a model such as the one proposed in this paper. Augmenting it, however, with primary survey data from customers about their share of gambling wallet that is, what proportion of the overall wagers is a casino getting from a player's gambling budget in a given trip (e.g., Du et al. 2007) may influence the relative attractiveness of gamers. Another avenue for estimating the share of wallet is collecting *intra-trip* gamer behavior data and modeling it. For instance, if we have data on a gamer's appearance in the casino facility of interest on different days within a given visit, a model may be developed for inferring his patronage of other casino facilities (Chen and Steckel 2009). Consequently, the gamer's share of wallet can be estimated. This is an important direction for future research.
- (ii) *Promotional activities*: Typically, casinos attract gamers by sending out a diverse array of promotions (e.g., complimentary hotel stays/meals, cash backs) and giving

them compensations while they visit a casino. Here, we did not have data on promotions and compensations. It would be interesting to investigate the effect of such promotions on the type of customers who are induced to visit (high skilled / low skilled) and their wagering behavior.

- (iii) *Consumer learning*: We treat “skill” as a static, individual-level trait. Another possibility is to model skill as a dynamic construct, which may change over time based on a player’s experience with a certain game. For example, a player may become more skilled with blackjack as she becomes more experienced with the game. With our current dataset, however, given that we observe a small number of trips per gamer, estimating a model with dynamically changing skill level will likely leading to over fitting.
- (iv) *Consumer demographics*: In this study, we examined the role of age and gender. However, incorporating other demographics such as relationship status, home address, profession and level of education represents another fruitful area of research.
- (v) *Casino floor design*: Casino industry managers are also interested in issues such as how best to configure their games on the casino floor (Bayus and Gupta 1985), the effect of environmental cues on gaming behavior (Griffiths and Parke 2003) and the effect of prices, number of table games at a facility on the per capita wagering on slot machines (Thalheimer and Ali 2003). By incorporating such casino-level covariates into our model, we can provide some insights into casino floor planning.

References

- Abt, Vicki, James F. Smith and Eugene M. Christiansen (1985), *The Business of Risk: Commercial Gambling in Mainstream America*, Lawrence: University Press of Kansas
- American Gaming Association (2006), *Industry Information*, available at http://www.americangaming.org/Industry/factsheets/statistics_detail.cfv?id=7, last accessed on February 3, 2009.
- Baker, Robert G.V. and David C. Marshall (2005), "Modeling Gambling Time and Economic Assignments to Weekly Trip Behavior to Gambling Venues," *Journal of Geographical Systems*, 7, 381-402.
- Bayus, Barry L. and Shiv K. Gupta (1985), "Analyzing Floor Configurations for Casino Slot Machines," *Omega*, 13, 6, 561-567.
- Becker, Gary S. and Kevin M. Murphy (1988), "A Theory of Rational Addiction," *Journal of Political Economy*, 96 (4), 675-700.
- Berry, Michael J.A. and Gordon S. Linoff (2004), *Data Mining Techniques*, 2d ed. Indianapolis, IN: John Wiley and Sons.
- Bitran, Gabriel and Susana V. Mondschein (1995), "An Application of Yield Management to the Hotel Industry considering Multiple Day Stays," *Operations Research*, 43, 3, 427-443.
- Blattberg, Robert C., Byung-Do Kim and Scott A. Neslin (2008), *Database Marketing: Analyzing and Managing Customers*, Springer, New York, NY.
- Cargill, Thomas F. and William R. Eadington (1978), "Nevada's Gaming Revenues: Time Characteristics and Forecasting," *Management Science*, 24, 12, 1221-1230.
- Chen, Yuxin and Joel H. Steckel (2009), "Modeling Credit Card 'Share of Wallet': Solving the Incomplete Information Problem," Working Paper, New York University.
- Clotfelter, Charles T. and Philip J. Cook (1993), "The Gambler's Fallacy in Lottery Play," *Management Science*, 39 (12), 1521-1525.
- Cohen, Jacob (1988), *Statistical Power Analysis for the Behavioral Sciences*, New Jersey: Lawrence Erlbaum Ass.
- Connolly, John and Brandon Rogoff (2008), "Keeping Track of U.S. Mall Visits," *Research Review*, 15, 2, 5-9. Available at <http://www.icsc.org/srch/rsrch/researchquarterly/current/rr2008152/US%20Mall%20Visits.pdf>, last accessed February 9, 2009.
- Cotte, June and Kathryn A. Latour (2009), "Blackjack in the Kitchen: Understanding Online versus Casino Gambling," *Journal of Consumer Research*, 35, 742-758.
- Croson, Rachel, Peter Fishman and Devin G. Pope (2008), "Poker Superstars: Skill or Luck," *Chance*, 21, 4, 25-28.
- Du, Rex Y., Wagner A. Kamakura and Carl F. Mela (2007), "Size and Share of Customer Wallet," *Journal of Marketing*, 71 (April), 94-113.
- Eliashberg, Jehoshua, Anita Elberse and Mark A. A. M. Leenders (2006), "The Motion Picture Industry: Critical Issues in Practice, Current Research and New Research Directions," *Marketing Science*, 25, 6, 638-661.
- Epstein, Richard A. (1997), *The Theory of Gambling and Statistical Logic*, Revised Edition, Elsevier Science.

- Fader, Peter S., Bruce G. S. Hardie and Ka Lok Lee (2005a), "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," *Journal of Marketing Research*, 42, 415-430.
- Fader, Peter S., Bruce G. S. Hardie and Ka Lok Lee (2005b), "'Counting Your Customers' the Easy Way: An Alternative to the Pareto/NBD Model," *Marketing Science*, 24, 2, 275-284.
- Gastwirth, Joseph L. (1972), "The Estimation of the Lorenz Curve and Gini Index," *The Review of Economics and Statistics*, 54, 306-316.
- Gini, Corrado (1912), "Measurement of Inequality of Incomes," *The Economic Journal*, 31, 124-126.
- Griffin, Peter (1999), *The Theory of Blackjack*, 6th Edition, Huntington Press.
- Griffiths, Mark D. and Jonathan Parke (2003), "The Environmental Psychology of Gambling," In G. Reith (Ed.), *Gambling: Who Wins? Who Loses?* 277-292, Prometheus, New York.
- Gupta, Sunil (1988), "Impact of Sales Promotions on When, What and How much to Buy," *Journal of Marketing Research*, 25, 4, 342-355.
- Gupta, Sunil (1991), "Stochastic Models of Interpurchase Time with Time-Dependent Covariates," *Journal of Marketing Research*, 28, 1, 1-15.
- Guryan, Jonathan and Melissa S. Kearney (2008), "Gambling at Lucky Stores: Empirical Evidence from State Lottery Sales," *American Economic Review*, Forthcoming.
- Harrah's Report (2006), *Profile of the American Gambler*, available at <http://www.harrah.com/harrahs-corporate/about-us-profile-of-gambler.html>, last accessed on February 3, 2009..
- Helsen, Kristiaan and David C. Schmittlein (1993), "Analyzing Duration Times in Marketing: Evidence for the Effectiveness of Hazard Rate Models," *Marketing Science*, 12 (4), 395-414.
- John, George and Barton A. Weitz (1989), "Salesforce Compensation: An Empirical Investigation of Factors Related to the Use of Salary versus Incentive Compensation," *Journal of Marketing Research*, 26, 1, 1-14.
- Jones, M.C. and M.J. Faddy (2003), "A Skew Extension of the t-distribution, With Applications," *Journal of the Royal Statistical Society, Series B*, 65, 1, 159-174.
- Lattin, James M. (1987), "A Model of Balanced Choice Behavior," *Marketing Science*, 6, 1, 48-65.
- Kallick, M., D. Suits, T. Dielman and J. Hybels (1979), *A Survey of American Gambling Attitudes and Behavior*, Research Report Series, Ann Arbor, MI: University of Michigan, Institute for Social Research.
- Kass, Robert E. and Adrian Raftery (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773-795.
- Klein, John P. and Melvin L. Moeschberger (2003), *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd Edition, Springer, New York, NY.
- Kumar, V. and Werner J. Reinartz (2006), *Customer Relationship Management: A Databased Approach*, John Wiley and Sons, Inc.
- Lehmann, Eric L. (2006), *Nonparametrics: Statistical Methods Based on Ranks*, Springer, New York, NY.
- Morrison, Donald G. and David C. Schmittlein (1981), "Predicting Future Random Events Based on Past Performance," *Management Science*, 27 (September), 1006-1023.
- Morse, Edward A. and Ernest Goss (2007), *Governing Fortune: Casino Gambling in America*, University of Michigan Press.

- Motion Pictures Association of America (2009), *Research and Statistics*, available at <http://www.mpa.org/researchStatistics.asp>, last accessed on February 6, 2009.
- NationMaster (2009), *Cinema Attendance by Country*, available at http://www.nationmaster.com/graph/med_cin_att_percap-media-cinema-attendance-per-capita, last accessed on February 6, 2009.
- Nichols, Mark W. (1998), "The Impact of Deregulation on Casino Win in Atlantic City," *Review of Industrial Organization*, 13, 713-726.
- Packel, Edward (2006), *The Mathematics of Games and Gambling*, 2nd Edition, The Mathematical Association of American.
- Parr Rudd, Olivia (2001), *Data Mining Cookbook*, New York: John Wiley and Sons.
- Putler, Daniel S. and Shilpa Lele (2003), "An Easily Implemented Framework for Forecasting Ticket Sales to Performing Arts Events," *Marketing Letters*, 14, 4, 307-320.
- Reddy, Srinivas K., Vanitha Swaminathan and Carol M. Motley (1998), "Exploring the Determinants of Broadway Show Success," *Journal of Marketing Research*, 35, 370-383.
- Rossi, Peter E. and Greg M. Allenby (2003), "Bayesian Statistics and Marketing," *Marketing Science*, 22, 3, 304-328.
- Seetharaman, P.B. (2004), "Modeling Multiple Sources of State Dependence in Random Utility Models: A Distributed Lag Approach," *Marketing Science*, 23, 2, 234-242.
- Stigler, George J. and Gary S. Becker (1977), "De Gustibus Non Est Disputandum," *American Economic Review*, 67 (2), 76-90.
- Swami, Sanjeev, Jehoshua Eliashberg and Charles B. Weinberg (1999), "SilverScreener: A Modeling Approach to Movie Screens Management," *Marketing Science*, 18, 3, 352-372.
- Thalheimer, Richard and Mukhtar M. Ali (2003), "The Demand for Casino Gambling," *Applied Economics*, 35, 907-918.
- Vogel, Harold L. (2007), *Entertainment Industry Economics: A Guide for Financial Analysis*, 7th Edition, Cambridge University Press, New York, NY, 353-419.

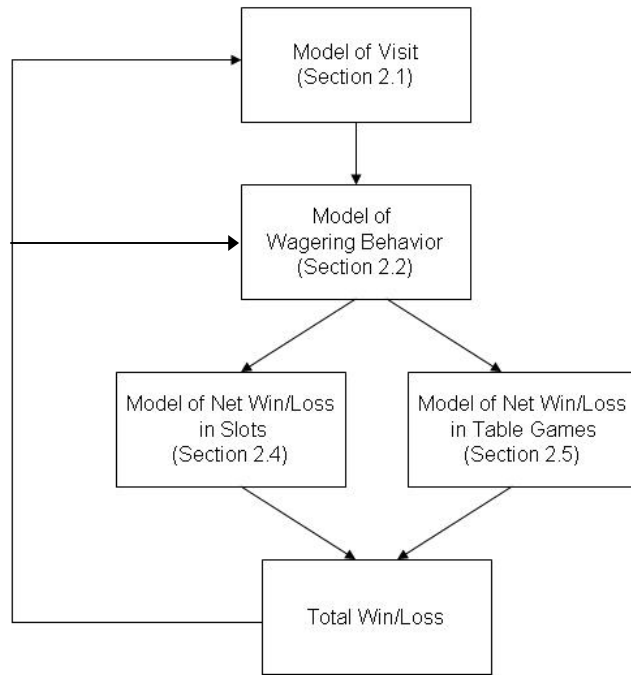


Figure 1: Overview of our modeling framework.

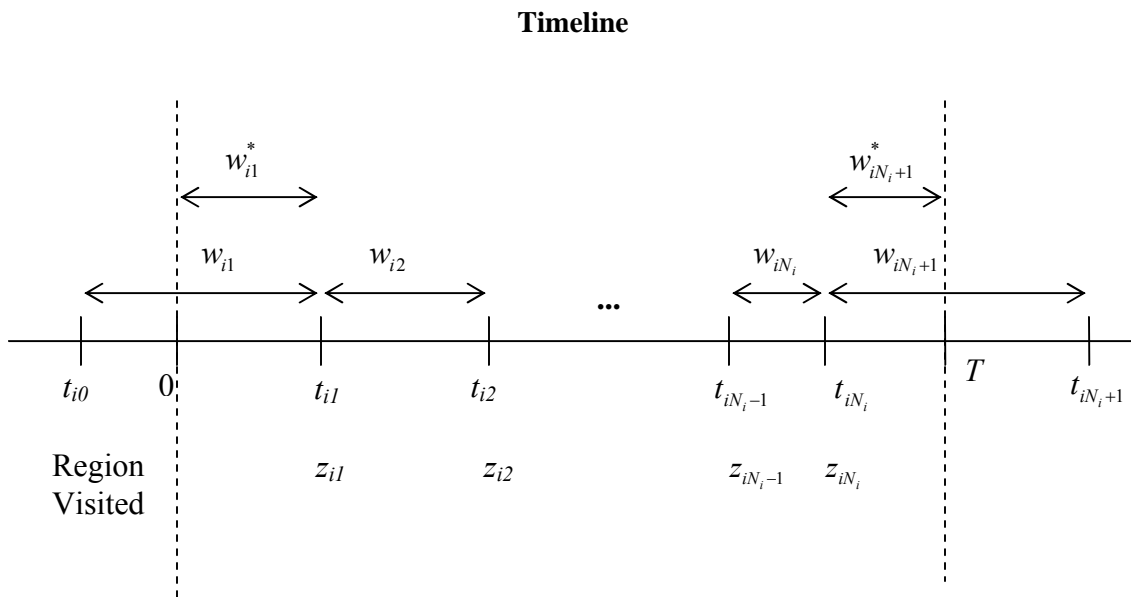


Figure 2: Timeline to illustrate the notations.

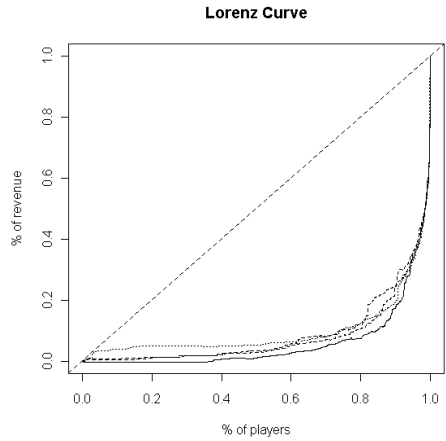


Figure 3a: (5 months holdout) Lorenz curves for our model (solid), alternative model A (dotted), alternative model B (dashed line) and alternative model C (dashed-dotted line)

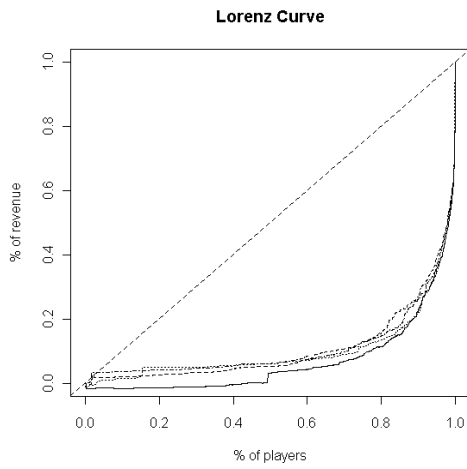


Figure 3b: (10 months holdout) Lorenz curves for our model (solid), alternative model A (dotted), alternative model B (dashed line) and alternative model C (dashed-dotted line)

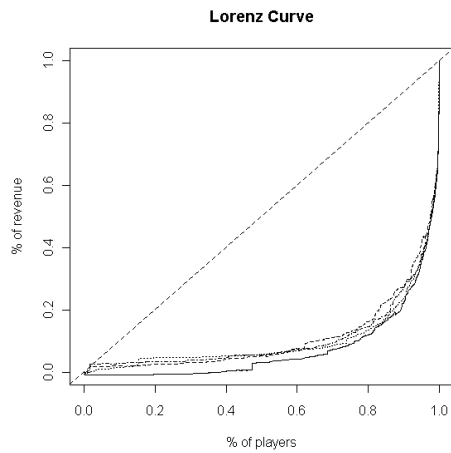


Figure 3c: (14 months holdout) Lorenz curves for our model (solid), alternative model A (dotted), alternative model B (dashed line) and alternative model C (dashed-dotted line)

Combination	Payoff (z_i) (in \$)	Probability (p_i)	Payoff x Probability ($p_i \times z_i$) (in \$)
3 sevens	200	0.000125	0.025
3 bars	100	0.00075	0.075
3 melons	100	0.001	0.1
Melon, melon, bar	100	0.0005	0.05
3 bells	18	0.005	0.09
Bell, bell, bar	18	0.000625	0.01125
3 plums	14	0.007875	0.11025
Plum, plum, bar	14	0.002625	0.03675
3 oranges	10	0.0125	0.125
Orange, orange, bar	10	0.003125	0.03125
Cherry, cherry, any	5	0.03	0.15
Cherry, any, any	2	0.07	0.14
Expected payoff			0.9445

Table 1: The payoff structure for 21 Bell slot machine.

Slot Machine	Wagered Amount (\$)	Slots Machines / House Advantage	Expected Revenue (\$)	Actual Revenue (\$)
1	100	0.2	20	40
2	200	0.3	60	-20
Aggregated Information	300		80	20

Table 2: Intra-trip data within slots for player A visiting region 1.

Table Game	Wagered Amount (\$)	Prototypical Player House Advantage	Expected Revenue (\$)	Actual Revenue (\$)
1	100	0.2	20	40
2	100	0.1	10	8
3	100	0.3	30	-20
Aggregated Information	300		60	28

Table 3: Intra-trip data within table games for player i visiting region 1.

Region	Number of Visits	Total Wager Amount (\$, 000)	Total Revenue (\$, 000)	Wager Amount Per Visit (\$)	Average Revenue Per Visit (\$)
R1	745	17230.2	557.5	23127.8	748.3
R2	1841	11746	944.6	6380.2	513.1
R3	315	4437.6	193.3	14087.7	613.5
R4	438	1601.9	136.6	3657.2	311.9
R5	2831	13359.7	964.6	4719.1	340.7
R6	1888	8783.7	553.1	4652.4	293
R7	583	3919.4	216	6722.8	370.5

Table 4: Summary statistics of the dataset by each region.

Gamers' Data	Mean	S.D.	Minimum	Maximum
Total number of visits	5.5	10.9	1	122
Total Wager Amount - Slot machines (\$,000)	37.8	567.2	0	21650.9
Total Amount Won - Slot machines (\$,000)	-2.1	12.7	-342.4	14
Total Wager Amount - Table games (\$,000)	1	8.8	0	186.6
Total Amount Won - Table games (\$,000)	-0.2	1.6	-33.4	13.4

Table 5: Summary statistics of the dataset across players.

Metrics	Model 1	Model 2	Model 3	Model 4
Wager Slots	2.50	5.30	-1.71	3.49
Wager Tables	2.90	-7.45	-3.91	-7.14
Revenue Slots	10.71	25.85	11.26	24.50
Revenue Tables	-3.41	-6.75	-9.10	-12.04

Table 6: Model fit. For each model, the numbers denote the % difference between the fitted value and the observed values.

Parameter	Mean	95% Posterior Interval
Wager Amount (α)	6.10	(5.99, 6.19)
Proportion on Slots (μ)	4.40	(4.21, 4.59)
House Edge in Slots (τ^{SLOT})	0.10	(0.09, 0.10)
House Edge in Tables (τ^{TABLE})	0.18	(0.16, 0.20)
Skill Level (γ)	-0.13	(-0.24, -0.02)
Rate - region 1	-8.54	(-8.68, -8.38)
Rate - region 2	-9.09	(-9.24, -8.93)
Rate - region 3	-11.03	(-11.54, -10.62)
Rate - region 4	-10.94	(-11.40, -10.55)
Rate - region 5	-8.86	(-9.09, -8.66)
Rate - region 6	-8.69	(-8.82, -8.55)
Rate - region 7	-12.13	(-15.53, -11.39)

Table 7: Posterior mean for player-level parameters.

Parameter	1	2	3	4	5	6	7	8	9	10	11	12
1) Wager Amount (α)	1.00											
2) Proportion on Slots (μ)	0.41	1.00										
3) House Edge in Slots (τ^{SLOT})	-0.04	0.14	1.00									
4) House Edge in Tables (τ^{TABLE})	0.06	0.07	0.02	1.00								
5) Skill Level (γ)	0.25	0.06	0.01	0.10	1.00							
6) Visit Rate - Region 1 (λ_1)	0.38	-0.02	-0.07	-0.06	-0.04	1.00						
7) Visit Rate - Region 2 (λ_2)	-0.04	0.02	0.02	0.04	-0.06	-0.05	1.00					
8) Visit Rate - Region 3 (λ_3)	0.17	0.03	-0.06	-0.08	-0.20	0.30	-0.23	1.00				
9) Visit Rate - Region 4 (λ_4)	0.27	0.27	-0.02	-0.01	0.09	-0.18	-0.07	0.10	1.00			
10) Visit Rate - Region 5 (λ_5)	0.19	0.13	-0.03	0.04	-0.01	0.03	-0.56	-0.03	-0.03	1.00		
11) Visit Rate - Region 6 (λ_6)	0.21	0.16	0.07	0.01	0.28	-0.17	-0.46	-0.23	0.03	-0.13	1.00	
12) Visit Rate - Region 7 (λ_7)	0.16	0.13	0.03	0.03	0.02	0.16	-0.13	0.45	-0.41	-0.06	-0.02	1.00

Table 8: Correlation matrix for player-level parameters.

Model	Holdout - 5 months			Holdout - 10 months			Holdout - 14 months		
	Spearman Correlation	RMSE	Gini	Spearman Correlation	RMSE	Gini	Spearman Correlation	RMSE	Gini
Proposed Model	0.45	2622.80	.884	0.54	5171.61	.864	0.59	5556.21	.854
Historical Revenues-based Model (Model A)	0.34	2843.43	.804	0.41	5687.59	.796	0.43	6261.73	.794
Gamma-Exponential Visits with Stochastic Net Revenues Model (Model B)	0.40	2977.08	.826	0.44	5923.19	.786	0.44	6575.92	.782
Gamma-Exponential Visits with Stochastic Theoretical Revenues Model (Model C)	0.40	2953.47	.838	0.44	5869.15	.780	0.45	6531.45	.788

Table 9: Comparison of the model-based predictions from the proposed model and alternative models.

Parameter	Intercept	Age	Gender
1) Wager Amount (α)	8.03 (7.61, 8.45)	0.01 (0.005, 0.02)	-0.29 (-0.48, -0.11)
2) Proportion on Slots (μ)	1.25 (-1.43, 4.01)	0.11 (0.06, 0.15)	-5.03 (-6.18, -3.88)
3) House Edge in Slots (τ^{SLOT})	0.10 (0.01, 0.20)	0.00 (-0.01, 0.01)	0.00 (-0.01, 0.01)
4) House Edge in Tables (τ^{TABLE})	0.24 (0.03, 0.47)	0.00 (-0.01, 0.01)	0.02 (-0.07, 0.13)
5) Skill Level (γ)	0.14 (-0.51, 0.63)	0.00 (-0.01, 0.01)	0.12 (-0.21, 0.46)
6) Visit Rate - Region 1 (λ_1)	-7.70 (-8.72, -6.69)	0.00 (-0.02, 0.01)	0.43 (-0.08, 0.87)
7) Visit Rate - Region 2 (λ_2)	-10.09 (-12.31, -7.63)	0.01 (-0.03, 0.04)	-0.25 (-1.24, 0.96)
8) Visit Rate - Region 3 (λ_3)	-10.49 (-12.25, -8.96)	0.03 (0.01, 0.05)	0.02 (-0.65, 0.69)
9) Visit Rate - Region 4 (λ_4)	-9.81 (-12.69, -6.77)	-0.01 (-0.05, 0.04)	-1.03 (-1.99, 0.12)
10) Visit Rate - Region 5 (λ_5)	-11.44 (-14.15, -9.09)	0.03 (-0.01, 0.07)	-0.01 (-1.16, 1.09)
11) Visit Rate - Region 6 (λ_6)	-10.61 (-12.91, -8.28)	0.01 (-0.02, 0.05)	-0.25 (-1.20, 0.67)
12) Visit Rate - Region 7 (λ_7)	-11.93 (-14.27, -9.49)	-0.01 (-0.05, 0.03)	0.31 (-1.09, 1.45)

Gender is coded as 1 for male and 0 for female.

The numbers in parenthesis are the 95 % posterior interval around the mean.

Table 10: Posterior mean for player-level parameters including demographics.

Appendix

I. Application of central limit theorem

Let X be the total net amount of loss by the gambler, Y be the total amount of wager. Let the expected net value of a unit (dollar) wager be $Z_i = 1-r$ (i.e., r denotes the house edge), and let $\text{var}(Z_i) = \sigma^2$.

By definition, we have $X = Y - (Z_1 + Z_2 + Z_3 + \dots + Z_Y)$. Hence, the mean and variance of X can be derived as follows:

$$E(X) = Y - \sum_{i=1}^Y E(Z_i) = Y - Y(1-r) = Yr$$

$$\text{Var}(X) = \sum_{i=1}^Y \text{Var}(Z_i) = Y\sigma^2$$

Since X is the sum of a series i.i.d. random variables, its density converges (in distribution) to a normal distribution. Thus, we have: $X | Y, r, \sigma^2 \sim N(Yr, Y\sigma^2)$.