Alberta Gambling Research Institute Conference 2019: Blurred Lines in Gambling Research

Browne, Matthew; Clark, Luke; Cunningham, John; Hilbrecht, Margo; Johnson, Mark R.; Quilty, Lena C.; Rodda, Simone; Sanders, James; Tavares, Hermano; Tremblay, Joël...

Alberta Gambling Research Institute

http://hdl.handle.net/1880/110151
conference proceedings

Downloaded from PRISM: https://prism.ucalgary.ca
Applying Data Science to Behavioural Analysis of Online Gambling

Dr Luke Clark
AGRI annual conference
30 March 2019
Disclosure Statement

The Centre for Gambling Research at UBC is supported by the British Columbia Lottery Corporation (a Crown Corporation) and the Province of BC government.

LC holds a Discovery Award from NSERC. The online gambling project also received a further project grant from the BC Ministry of Finance.

Honoraria (speaker travel / reviewing): Svenska Spel (Sweden), Victorian Responsible Gambling Foundation (Australia), National Center for Responsible Gaming (US), Gambling Research Exchange Ontario.
Thanks to…

Xiaolei Deng  
Dr Tilman Lesch  
Kent Macdonald

Data provided by the BCLC. With thanks to Kahlil Philander and the Data Analytics team

Project funding from BC Ministry of Finance, Gaming Policy & Enforcement Branch
Online Gambling

“Which forms of gambling do you think are harmful?” (Angus Reid poll)

- BC has a single, state-run online gambling platform (since 2004)
- Public concern over e.g. 24/7 access
- Unlike most land-based gambling, every click is recorded...

(Angus Reid poll in Vancouver Sun, 13 Oct 2016)
4 questions:

• What is the most popular form on a diverse, state-run platform like PlayNow?

• Pareto effects: how much gambling is due to the ‘vital few’?

• What proportion of the ‘vital few’ have gambling problems?

• How well can machine learning predict self-exclusion status using behavioural markers?
The PlayNow.com dataset

• De-identified data from PlayNow.com provided by BCLC, only accessible to BC residents

• Dataset #1: One month (June 2015) from 5 sections of platform:
  – 41,401 users placing 48 million bets
  – 5 sections: Lottery, eCasino, Sports, Bingo, Keno

• Dataset #2: One year of data (Oct 2014 – Sept 2015) for eCasino (online slots & table games):
  – 30,902 users placing 564 million bets

• (160gb PostgreSQL database maintained on ComputeCanada server)
So which games are most popular?

- 30.0% users access PlayNow to buy lottery products...
- 79.7% of users access PlayNow to access eCasino games
- 97% of individual bets are in eCasino (slots & table games)
- 0.9% of bets are in other games
The Pareto Rule

- The law of the vital few: 80% of activity from the top 20% most active consumers

In online gambling:
- Fiedler (2012): 6 months of poker data from 55 million transactions. Top 10% generate 91% of rake.

- Tom et al (2014): 12 months of bwin data, 80% of net loss attributable to top 7% most active. Approx half of vital few score positive on BBGS.
Pareto calculations in the eCasino

Can be calculated in multiple ways! We ask “what proportion of activity is due to top 20%?”

Can be calculated for number of bets or spending/loss
Pareto varies by time window

- Pareto estimate increases with data window
- Plateau after ~12 months at approx 90% for net loss, 92% for total bets
- Top 20% stay in / return to the system. The other 80% are more transient (leaving & joining)
Self-Exclusion Status

- 2,465 users have a record of self-exclusion. Note: 1) could be prior to or within the data window, 2) may have enrolled in casino or online

<table>
<thead>
<tr>
<th>VSE, n (%)</th>
<th>Top 20%</th>
<th>Remaining 80%</th>
<th>$X^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranked by total bets</td>
<td>832 (33.8%)</td>
<td>1633 (5.7%)</td>
<td>316.7</td>
</tr>
<tr>
<td>Ranked by net loss</td>
<td>1007 (40.9%)</td>
<td>1458 (5.1%)</td>
<td>755.5</td>
</tr>
</tbody>
</table>

→ the 20% most active users contains a substantially higher proportion with a self-exclusion record
Interim Conclusions

• Online gambling engagement varies massively by game types (lottery = most users, eCasino (slot machines) = most bets)

• The 20% most active gamblers generate >80% of the bets and spend

• Pareto estimates from < 1 year of data substantially underestimate the effect. The top 20% generate 92% of bets and 90% of spend.

• Elevated rates of self-exclusion indicate disordered gambling in the vital few
Predicting Problem Gambling from Behavioural Data

- Machine learning generates an algorithm for classifying subjects into 2 groups
  - Complex multivariate relationships
  - Includes interactions between predictors
  - Robust to non-normality & outliers
  - Cross-validation reduces ‘over-fitting’

- Multiple performance measures: classification accuracy, sensitivity & specificity, AUROC
High Risk Algorithms

• **Philander (2014 IGS):** *bwin* public dataset (European)
  530 account closers, separating gambling problems (n = 176) vs other reasons for closure
  Input variables: daily aggregates + age, gender
  50-55% AUROC across 9 models, barely above chance

• **Percy et al (2016 IGS):** BetBuddy / GTECH European data: 176 self-excluders vs 669 controls
  33 input variables, incl. demogs and trajectory (change from baseline period)
  79% AUROC for random forest model
  (see also Haeusler 2016 IGS, Excell et al 2014 report 3)
The dataset

- 1 year eCasino dataset: 30,902 users in BC, from Oct 2014 – Sept 2015; 564 million bets
- Primary model: 1323 with self-exclusion status vs 3000 controls, all placed >200 bets.
- 18 input variables, reflecting daily- and session-aggregates of frequency (e.g. Sessions per Day), intensity (e.g. Money Bet per Session), variability (e.g. Variance in Money Bet per Session)
- Feature selection algorithm used to reduce input variables to 9 that are sufficiently independent
- 10-fold cross-validation to establish performance
**Results**

- Primary model: gradient-boosted decision tree, using feature selection (9 from 18 input variables).

<table>
<thead>
<tr>
<th></th>
<th>Prediction: Self Excluder</th>
<th>Prediction: Control</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self Excluders</td>
<td>395</td>
<td>928</td>
<td>1323</td>
</tr>
<tr>
<td>Controls</td>
<td>196</td>
<td>2804</td>
<td>3000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.74</td>
<td>0.30</td>
<td>0.94</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Secondary models used:
- i) Balanced dataset (1323 vs 1323) (AUROC 75%)
- ii) All 18 variables (AUROC 76 - 79%)
- iii) Random forest (AUROC 75 – 76%)
- iv) Alternative data threshold >10 sessions (AUROC 75 - 79%)
Transparency

In our primary model, 25% of variance in predictive signal is from Variance in Money Bet per Session

Risk Scores

<table>
<thead>
<tr>
<th>Risk group</th>
<th>Scores</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0-27</td>
<td>2016</td>
</tr>
<tr>
<td>Medium</td>
<td>27-54</td>
<td>828</td>
</tr>
<tr>
<td>High</td>
<td>54-81</td>
<td>156</td>
</tr>
</tbody>
</table>

$\eta^2=.005$ (small)

$\eta^2=.024$ (small)

$\eta^2=.208$ (large)
Combining sources of risk data

- Behavioural ‘play’ data (frequency, intensity, variability…)
- Demographics (gender, age, time of day, geography)
- Financial behaviour (deposits, withdrawals, account switches)
- Operator communication (complaints, text analysis)

Future Directions

• Going deeper than daily aggregates to analyze behaviour *within* a gambling session → markers of loss chasing (see Smith et al 2009 poker, Xu & Harvey 2014 sports; Leino et al 2016 EGMs)

• Drilling down on self-exclusion (prospective, different lengths, casino vs online)

• Need for *independent datasets* to test generalization (see Gowin et al 2019 relapse in stimulant dependence)

• How should operators *respond* to users identified as high-risk? (account closure, direct contact, marketing blocks, RG tools...)

Identification ≠ Intervention !!
Email luke.clark@psych.ubc.ca
Web www.cgr.psych.ubc.ca
Twitter @LukeClark01 @CGR_UBC