Game Analytics: How to Data Science like a boss?

So, I have spent the last two weeks working on those 100 parameters and 2 million rows or roughly 200 million data points. I went on testing all the possible combinations of parameters I can arrange. My eyes were glued to the very part of the screen where the F-score of my algorithm was displayed. I popped the champagne when the result finally hit the hot spot. And here I am, with the black box that is my algorithm, explaining with excitement to my manager that: “Look at those players, do you see them? They are very likely to leave to game and they are likely to generate you $X, so make sure they don’t and you’ll increase your future revenue by 5%. Don’t ask why, the numbers tell it all”

Well, I’ve just been successful on describing how Data Science is not in video games. OK I’m being provocative here, but the bottom line is, Data Science shouldn’t be conducted with “more data is better data” or “let machine do the whole work” mindset. Data Science, as predictive analytics, is not the science of forecasting. It is not about finding the perfect algorithm to predict human behaviors which by definition are volatile. It is not about mastering linear algebra/graph theory/ statistics/ probabilities. It is not about being a guru in SQL/R/Python/Spark. And it is definitely not about using all the available data and believing that somehow more numbers and letters will give you more prophetic power.

Therefore, what is Data Science in video games? Let me start with an example: Suppose I have GameStop year to year sales number for a large number of years. This enables me to forecast the sales within a certain interval, calculate my future cash flows, manage my retail production and ultimately allows strategic decision elsewhere, it may also yield a few bargaining arguments somewhere along the line. The case is simple, there is only one metric we perfectly understand, few data points and no major levees does it give you in terms of business strategy.

Now another case: I have just released a new game which has a terrible retention rate after a few days of release, therefore I need a fast feature that I can push in order to stop the bleeding. I decide to go with adding a new achievement, but which one? When? Privileging what category of players? In order to answer those questions, I will select the relevant metrics I have on the player, using the standard *market intelligence* I have as an expert of the industry and the understanding of *players’ psychology* as a gamer myself. Then is time for the Market Analytics 2.0 to be up on stage. A few magical keystrokes tell me that exact moment along the learning curve and game progression I should insert the achievement and hop, no more retention problem.

In other words, Data Science is the natural process that comes along the companies’ desire to become more and more proactive. In our example, the company doesn’t want to see all the players’ gone and then learn the lessons for their next title. They want timely response to improve the game retention rate and thus make every single one of their titles successful. Beyond the seemingly quantifiable revenue increase, what Data Science really gives us are insights allowing us to make better decisions, data-driven decisions: it is a way of having data back up your gut feelings as a ready expert of the domain and enable you to have more.

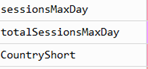
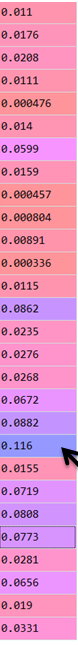
This is what I believe to be the right way to approach data in video games: Start the analysis with a precise question (“predicting if a player is going to leave the game in two weeks’ time?”) and a large range of data; Use the knowledge of the market and the players to do a preliminary trimming of the data; Build up a learning pipeline and refine the learning algorithm via metrics selection, data range correlation and model optimization; Use data to corroborate expected key metrics professionals of the industry usually believe to be reflective of the issue; Find out unexpected key metrics *and* understand why would they be relevant to the issue (what is the psychological reason? Is there a rational explanation? Or it is just one of those things that blow your mind); Use the newly found metrics to improve the end goal, add design features to stimulate those metrics and make forecast on generated revenue under hypothesis. The whole process should loop over itself from metrics/data prepping to algorithm optimization. If it helps, think of the whole process as a positive retroaction loop, the idea is to consider that a better understanding of the data and having a powerful algorithm are two sides of the same coin.

Now, I think a case study shouldn’t hurt. Let’s give the name X to my game here. So X is a medium-size game, has roughly 80 metrics and 200k players I can use, those metrics can be transaction-related or in-game behavior-related. X has a persistent level system and, for relevance reason, I want to look at players with level 3 or up and I want to predict if they are likely to lose interest completely in X two weeks from now. X has existed on old gen and just got ported to new gen, therefore I’ll use old gen data to train and test my model and try to predict what player’s interest is going to hit the point of no return two weeks from now on the new gen. This is step 1.

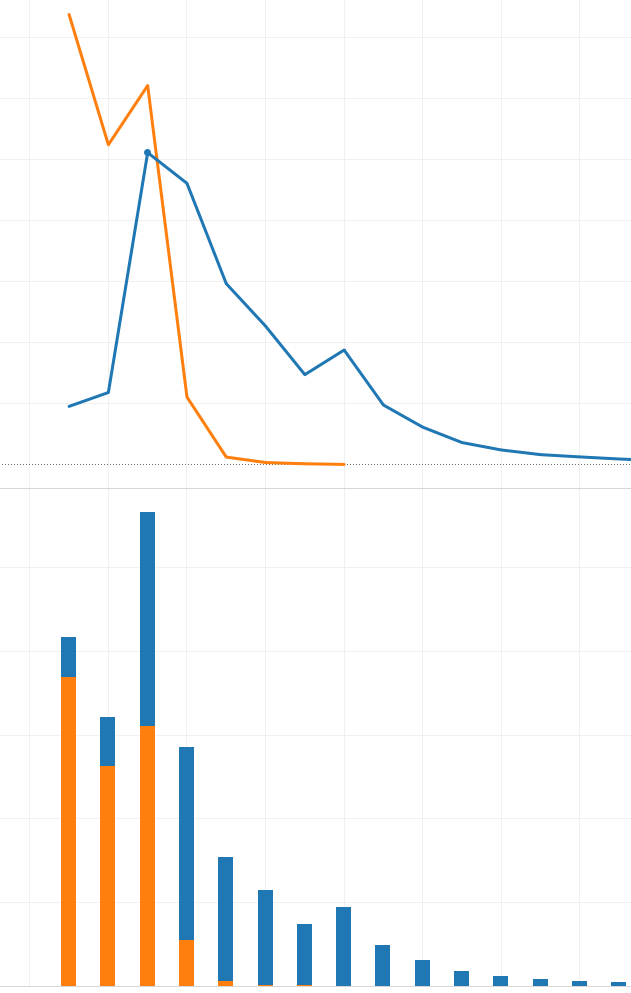
Now we’re led to step 2: the data trimming. This is particularly important since 80 features are only going to *overfit* our model, meaning it will yield so specific a model for my training set that it will give unreasonable (and false) predictions on my target set. Is the id of the player/session/event important? I don’t think so. Is the IP address relevant? Very unlikely. The country code in that? Possibly. The Persistent level of the player? Very likely. And the average session length? Definitely! This is often where lots of data-related company (and in general converted maths majors) make their first misstep: by not utilizing existing experts and believe that their algorithms will yield better insights regardless, it will at best waste some time but can lead in worse case to misleading conclusions.

The next step offers two strategies: Use all metrics, strip them one at the time and find the right combination, this is the top-down approach; The bottom-up approach, that I personally favor, consists of testing the predictive power of each parameter, or at least the most obvious ones, and adding processed metrics one by one until they all give a decent weight to the model (and you should find an explanation to the why as well). The data selection doesn’t stop at the metrics selection, it comprises the range of the selected data as well: Why would anyone compare the first month of the new game data with the first and second month data on another game? The comparison just doesn’t make any sense. In my case, I started with Total Time Played, Number of Sessions, Average Length of the Sessions, Number of Play Days, Successive Gap Between Last 3 Play Days etc… and went up to 26 features. The whole process allowed me to get to a decent score with the algorithm with an acceptable *precision/recall*.

Now come steps 4 and 5: the results analytics part. The weight of each parameter gives me insight on what metric influences most the likelihood of a player to leave the game. With no surprise, Total Time Played during the last few weeks is the most important metric, but the Number of Sessions, the Average Sessions Length as well. On a slightly less obvious way, the Exp and In-Game Currency Earning Speed are 3rd and 4th in terms of indicating player’s interest. Once it is exhibited by the algorithm, it has a straightforward explanation: A faster reward speed means a faster paced player (and defines the player) and if the player slows down, it means the player is less hooked by the game, and this begs for action. Last but not least, an unexpected metric: the reward out of every game played during the last three weeks is a major indicator (5th) of how players react to the game. This is the finding “new metrics” is about, but then I needed to understand it, and after some psychological research, the phenomena can be named as “*Casino Symptom*”, and outlines what professionals qualify as a need to “*Anchoring*” by utilizing the “*Optimism Bias*” in player’s behavioral pattern: Random events, hack ’n’ slash (Diablo/Path of Exile)’s exceptional rewards, week-end double Exp in MMORPG. All utilize this psychological trait to improve the genre’s gamification prowess. This is how I utilize the results to understand the data. Here is how the weights look like (some confidential parts are hidden).



*Features weight in learning* algorithm



*Level distribution, orange = predicted future leavers.*

Now, let me show a few ways to exploit predictive properties of our methodology (and understand players behaviors better in the process): the above graphs show the distribution of number of players per level, orange is the predicted future leavers (For confidentiality reasons, I cannot leave legends/scale on the graphics). The second pic on the first graph is level 5, and corresponds to an achievement point, then it takes 5 more times to go from level 5 to 10 which corresponds to the next achievement, and even though we are not saying those are exactly the players who are going to leave, we can safely state that level 5 aggregate most of the players likely to leave and maybe a new achievement between those two points won’t hurt, or any rewarding experience as that matters. A rough estimation allows us to say that one more level at level 5 increases the likelihood the player will stay more than two weeks from 45.1% to 80.6%. Of course, in the real world, increasing artificially the level of the player will cause all the percentages to go downward and doesn’t solve the gap problem, but again it is not about making sure every individual player’s leaving intent has been addressed, it is an overall strategy that is going to lift your retention curve slightly up by pinpoint the design loopholes and adding gamification techniques, which is the final objective.

To put all in a nutshell, I believe Data Science in gaming industry allows you to: Understand the huge amount of data you already have; Confirm standard assumptions and “obvious” metrics with data; Find new metrics or seemingly not relevant metrics to a targeted problem; Understand the underlying psychological, sociological or physiological reasons behind the ethereal relationships between metrics and players’ reaction to the game; Thus give more tools to evaluate a player in relation to the problem and allow *deterministic* measures to address the problem (but between us, it’s all about the conversion/retention right?). Your business mindset will be your best tools, and the more complex decisions you want to make, the more complex technical tools you’ll need to back up your analytics. In one sentence: *Big Data allows us to improve our understanding of consumers’ behaviors, to make data-driven decisions and have an edge over our competitors on how the market is going to evolve*. And in order to do so, Machine Learning algorithms should be considered the missing link from the last decade rather than the end product.