Internship Report

Alejandro Jiménez Rico

5 July 2018

The Company: Social Point

Social Point is a video game developer, specialised in free-to-play mobile games and social network games, founded in October 2008 by two engineering students of Barcelona and later acquired by Take-Two Interactive in 2017, it is considered one of the most successful start ups in Spain.

Throughout its lifetime, the company has shown great interest in Data Science and Analytics. Leading and sponsoring many data science meet-ups, tech events and analytics conferences.

My Position: Analytics Intern

Within the company, one of the most vibrant and exciting departments is what they call *Data Products*. This department manages all tasks related to develop tools, data-oriented insights and models. This is the place where the magic of Machine Learning is casted.

Along many other tasks, this department handled petitionts from most of the other departments, including the analytics teams from each game, marketing, accountability and financial planning.

Since most of the team is extremely busy handling all the duties and mentioned, the arrival of an intern carrying no responsibilities within the company was a great opportunity to undertake less day-to-day ordinary tasks, and pursue more long-term oriented research. Even though from time to time I had to perform more rutinary jobs in order to reduce the workload of my partners, most of my time I spent in the company was studying patterns, developing models and building fresh new interactive tools.

My Tasks: Summary of the most relevant ones

Automated Dashboard for Model Accuracy

One of the first tasks that I was asked to do is to measure and analyise the different accuracy metrics of the predictive models that the team was running, in a daily basis. In order to do that, I wrote an R script that reads the predictions of those models and compares them to the real data, computes different accuracy metrics such as *Area Under the Curve* in a ROC plot and the F-Score, among others.

Moreover, aiming to automate this task as much as possible, I wrote another R script that took all those metrics and built a pretty dashboard that highlighted all metrics, their time-evolution and made some analysis upon them.

From that moment forward, one of my roles within the team was to generate that dashboard everyday, and report the situation of the running models.

Time Series Forecasting: YouTube Campaigns Attribution Analysis

One of the disadvantages of working in a such a transversal department, is that you have to deal with petitions from many other departments of the company. More often than they should, some departments tend to ask

your team to release them from dull and boring tasks that were extremely time-consuming. Those tasks usually had to do with numbers, and so they expected the *Data Science* guys to deal with it.

As it is expected, the interns ended up having one of those tasks in their hands. The task consisted in analyse a YouTube campaign that the Marketing team had ran a few weeks ago for a given Game. They paid a YouTuber a humungous amount of money to persuade him to speak goood things about one of the games of the company. And so the Marketing team wanted us to tell them how many installs of the game were thanks to that YouTuber. The usual methodology was to compute the average installs from the week prior to the campaign and compare that number with the installs from the week of the campaign. This was a dull, monotonous task and full of errors.

Fortunately, the Team Lead gave me enough time and confidence to address the issue more *scienc-y* and using a more mathematical approach.

Therefore, making use of Autoregressiv Integrated Moving Average (ARIMA) models, I developed a Time Series predictive model that tried to predict the number of installs of theh week of the campaign, utilising solely data previous to the campaign. This way, if the model was accurate enough, the difference between the prediction of the model and the actual number of installs, could be fairly attributed to the YouTube Campaign.

Moreover, I built a R Shiny App that was able to do all of this automatically and generate pretty beautiful reports answering all marketeers questions. Since this task had to be done periodically - because Marketing likes to do a lot of campaigns - the App turned out to be quite useful, relieving the team from a lot of workload and with far less error than the previous methodology.

Evaluation of Predictive Models and their Profitability

One of the most importants models in the Analytics team of a free-to-play games developer company is the one that tries to predict whether a player is going to pay some real mone in the game. The vast majority of players do not spend a penny in the game. And trying to identify and predict those players that were suitable to pay money was one of the biggest priorities of the analytics side of the company.

One of the major problems when facing this kind of situation is that every model trained in order to minimise its error, is going to predict that *all* players are non-payers. No one. This happens because stating that radical prediction assures an utterly high accuracy, such as 0.9999. The same kind of problem is present in medical diagnosis and epidemiology. If your model simply states that all people on the planet are not sick, it is being extremely accurate, since most people are perfectly healthy. Nonetheless, the model turns out to be absolutely useless.

One of the proposed solutions was to train the algorithms to minimse False Negatives instead of overall Accuracy. This could happen to be dreadful also, because the model simply tends to overestimate the payers and ends labelling everyone as payer, hence minimising False Negatives and skyrocketting False Positives. Different approaches on that consisted in building somewhere-in-the-middle metrics as the F-Score, that weights both types of errors based on a β parameter. But this can be extremely tricky, since the election of that parameter could have enormous consequences in the usability of the model.

Since both False Negatives and False Positives have different costs associated to the company, one of my tasks consisted in analyse the economic consequences of prioritising one error over the other one, or whether it existed some sweetspot in the middle of both that we could find, based on the β parameter from the F-Score.

The lectures from the *Data Visualisation* subject of the Master turned out to be quite useful for this task, since I made use of some of the Bootstrap techniques I learned there in order to balance the data set and get a biased population more close to what was useful for the company. This approach eventually led to the most useful solution for the problem.

Conclusion

I have been in the company from February 2018 to July 2018, and during my estance I have gotten the opportunityt to really sharpen my coding and programming skills, to improve my analytical mindset and to expand my mathematical tools in order to solve real-world problems with messy and unustructured data. I learned how to better formulate the proper questions when facing a problem in order to deconstruct complex issues into more solvable smaller pieces.

Overall, I see myself more prepared and with fruitful experience, after my internship in Social Point. I feel very lucky to have been given the opportunityt to work with such an amazing team of people, in a creative and relaxed environment that has given me the freedom and material to develop my skills, pursue my curiosity and challenge myself proving that I am capable of make constructive, worthwhile and significant contributions to a company.

For those and more reasons, I strongly recommend to any student from the Master, interested in aiming a career on Data Science, to pursue an internship in this company.