Data Science and Privacy: Netflix Plays with Fire

Netflix should have known better. In 2006 it was young and innovative and sat atop a treasure chest of film rental and review history from its roughly 6 million users. These rental histories were most certainly private under US law.

A New Approach to an Old Industry

Netflix had been growing quickly since its inception in 1998 and it had a fresh approach to an established industry. It had a strategic competitive advantage in the form of reams of digital data at its fingertips and a hunger to bring to bear the best data science techniques available.

Why data science for a movie rental company? Historically, this was an industry with little customer interaction. Movie returns were dropped through a slot in the wall after 3 days. Netflix, in contrast, at least had the 'common courtesy' to ask customers how they liked each movie. This was the seed of a new data-driven opportunity.

Netflix, the company of the future, had actually taken a huge step back in time, but in a very good way. It was reclaiming small-store intimacy, building individual, interactive relationships with its customers. It knew what their interests were, what they searched for, which movies they browsed, previewed and rented, and, still better, it knew what they thought after viewing the films. Oh, and, it knew this for nearly 6 million people. Wow.

Why is this so important? The better Netflix understood what made its customers happy, the higher the chance it could recommend to each one (personally!) a film that

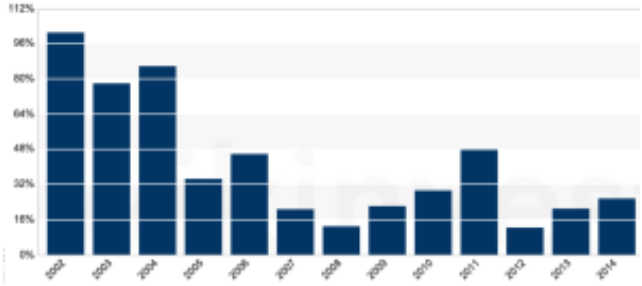
they would pay to rent (back in the day), or the more often they would return and also recommend the service to other people.

Enter Data Science

The only way to provide personalized recommendations for so many viewers was to automate the process. That's where data science came in (more specifically in this case, machine learning). The clever people at Netflix developed an algorithm into which they could input the film views and reviews of all of their customers. The algorithm would then make a good guess at what additional films any particular user was likely to enjoy.

Data Science was putting money in the bank for Netflix.

Unintentional Violation of Privacy Laws ?



But times were getting a bit tougher in 2006, and Netflix wanted to raise the bar. They wanted to get even better at recommending films that each customer was going to love.

Netflix had a eureka moment: Why not crowd-source the recommendation algorithm? With a cash prize of $1,000,000, Netflix invited the world to create an algorithm that would beat its own by at least 10%. Netflix would release a large amount of its user viewing and ratings data to the public to use in training their algorithms.

But, wait, wasn't this private information, protected in 1988 by an act of Congress? OK, then Netflix would remove the user names from the data. Thus, you could see that User 24601 had watched movies A, B, and C, along with his reviews of those movies, but not the name of User 24601. Netflix thought they were complying with privacy laws. Clever.

October 2, 2006. The Netflix Prize competition begins. Across the globe, eager minds bent on fame and fortune begin their work, and within just six days a team has beaten Netflix' own algorithm. Six days of outsourced research to surpass years of internal development! The 10% improvement goal wasn't yet reached, but results are looking promising.

November 2006. Only one month later, the first tremor hit. Two researchers at the University of Texas succeeded in linking some of Netflix' 'anonymized' data with external data, thus revealing the entire Netflex viewing history for certain individuals. Recall that Congress had declared this information to be private under the Video Privacy Protection Act of 1988.

At this point, people at Netflix HQ should have started losing sleep.

Sept 21, 2009. Netflix marinades in its potential privacy violation for three more years before the prize is finally claimed. Netflix announces plans for a follow-up competition. All that is soon to change.

December 17, 2009. Four Netflix users file a class action lawsuit against Netflix, alleging that Netflix had violated U.S. fair trade laws and the Video Privacy Protection Act by releasing the datasets. Four months later, Netflix has settled out of court (and cancelled their second competition).

Where did Netflix go wrong?

Ironically, Netflix failed to realize the power of data science. It warmed its hands by the flame, not realizing that it was actually standing in the fire.

Despite acting in good faith, Netflix failed to realize that data science techniques are able to de-anonymize data and convert non-PII into PII data. We are becoming increasingly aware of how this can be accomplished and of the development and proliferation of technologies that are increasing the exposure that companies unknowingly present and the risks that they unknowingly take.

In this age of increased privacy concerns, particularly with the pending EU Data Protection Regulation, the price for non-compliance has become too high to pay. Netflix settled out of court, but this will not always be possible.