Solving Customer Churn with Machine Learning

Case Study

PayPal

Summary

Paypal offers a global payments platform in 203 markets. The company boasts 173 million active customer accounts which resulted in 4 billion payments processed in 2014. Because Paypal relies on service fees as a percentage of payments made through its platform, the more active customers it has, the more revenues the company makes. As a result, customer churn is a critical business metric for Paypal, and the company has endeavored to minimize churn through a variety of marketing and product development programs. With H2O's powerful predictive modeling and machine learning, Paypal has been able to address churn when it happens, so the company can activate or reactivate consumers much faster and more efficiently than ever before.

The Challenge

For Paypal, consumer churn can have a big impact on its bottom line. Previously, the company looked at the problem in specific increments of time, noting that a customer who hadn't used its platform in that time period must have churned. Paypal would run a report that showed the churn date for all the customers that fell into this category as the date the report was run. The report also showed which features were the last ones to be used for all customers who churned during that time period. While this information was useful, it wasn't fully accurate, and, as such, the timeliness and effectiveness of Paypal's marketing efforts to win-back customers was less than ideal.

In reality, a customer's churn date needed to be closer to when they last interacted with the Paypal platform, not simply when a churn report is run. With machine learning, the data scientists at Paypal could predict if a customer will stay with the platform or if that customer will churn and when.

Additionally, because different customer segments may have different reactions to the platform features that caused them to churn, using machine learning would enable the scientists to get more specific feature importance results by customer rather than an aggregate.

The Solution

Paypal needed to redefine the metrics the company associated with churn, so the executive team could run KPIs and better understand the long-term health of the business. In addition, the operational teams, including marketing and product teams, needed more accurate, actionable information to help them run campaigns to retain or reactivate customers. To get the information each internal stakeholder needed quickly, Paypal's Senior Data Scientist, Julian Bharadwaj, and his team began developing a predictive model to show when a customer would churn, or not. During this process, the team jumped straight into using random forest and GBM with H2O, running through R.

"Integration was one of the key characteristics we were looking for," noted Bharadwaj. "Just in the data science world, Python is a big player and R is a big player. The fact that H2O integrates with both of them almost seamlessly was one of the deciding factors in our adoption of H2O," he continued.

As the models were developed, Bharadwaj looked at transaction and behavioral variables as well as

Use Case

Customer Churn

Industry

Financial Services

Challenges

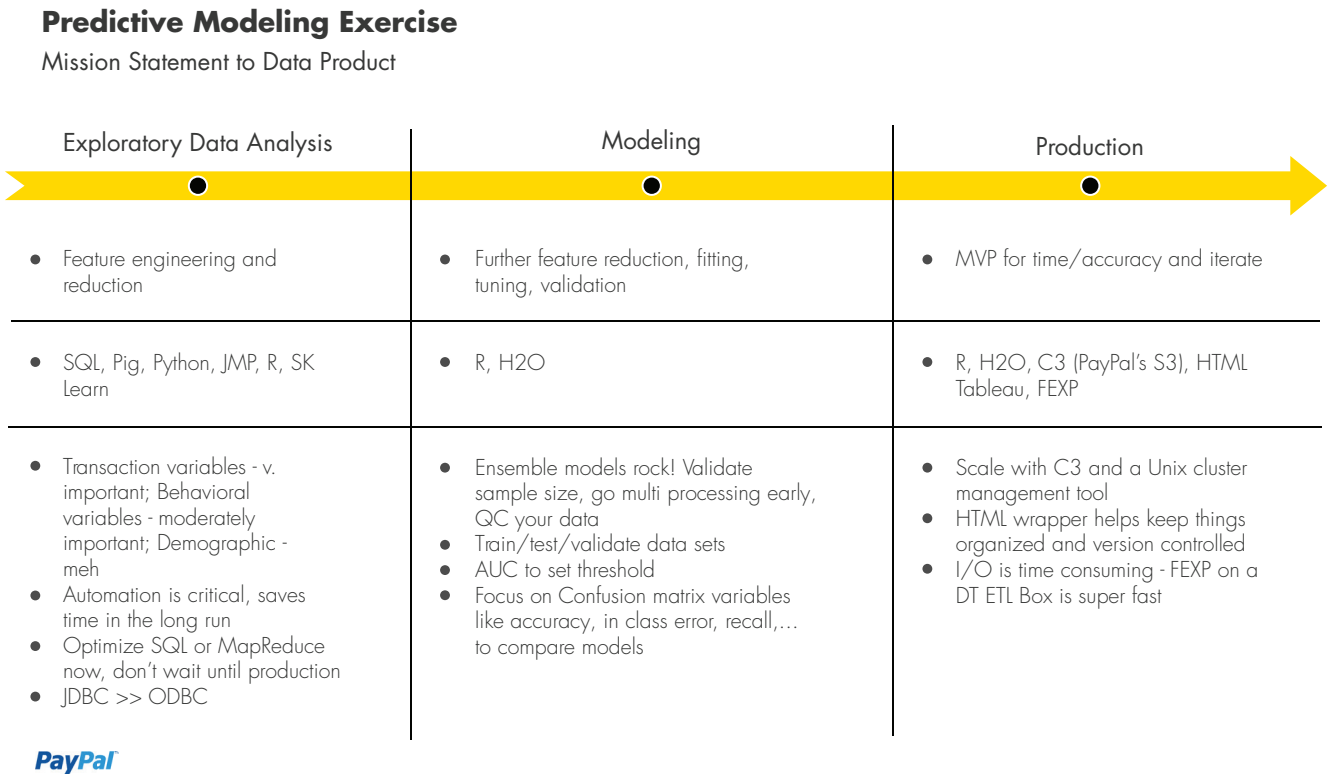
- Identifying if and when a customer will churn, and why, to improve the effectiveness of marketing campaigns
- Quickly delivering actionable information to operational teams to speed the development of new programs aimed at customer retention

Solution

- Developed predictive models leveraging EDA of existing churn reports and other datasets
- Integrated H2O with R and Python to run multiple models on entire customer base
- Created predictive modeling factory with H2O on Hadoop

Results

- Improved churn metrics and accuracy of information delivered to both executive and operational teams
- Increased speed at which models could be run, giving teams immediately actionable data
- Created more sophisticated and effective programs to reduce churn built around the output of the H2O machine learning algorithms



demographic data for customers who had churned. The first two turned out to be critical indicators of churn; the latter was not very useful, so the team dropped it. Using H2O made it fast and relatively easy to do that: the models could be modified across multiple parameters and run multiple times very quickly, so Bharadwaj could ensure the validity of the output.

"Just the way H2O multiprocesses and multithreads, the results that you get are really awesome," Bharadwaj remarked. "Using anything other than H2O meant that we couldn't run more than one model on a dataset and actually provide output in the time we needed," he said. "But with H2O, I can run a random forest with differently-tuned hyperparameters and several different GBMs on the entire population to get a model that wins."

Now in production, Paypal uses H2O on Hadoop to run a predictive modeling factory – large-scale, rapid modeling – that helps Paypal run more sophisticated and effective marketing programs that reduce churn

The Results

Immediately, Paypal began seeing great results. Initially, modeling on a laptop or a really big virtual machine took a long time when using R and ODBC on hundreds of thousands of rows of customer data. In fact, it wasn't unusual for modeling to take around 6 hours for a subset of customers, and scoring on the entire customer base would take close to 72 hours. When Bharadwaj switched to using H2O on Hadoop, those times diminished to 10 minutes and 5 minutes respectively to train and score on Paypal's entire customer base.

"When we started, it was a long, drawn-out process of testing," said Bharadwaj. "We used Python, then moved over to R as a platform, and then realized that for the volume of data we had, and for the complexity of the models we were fitting, those solutions took a long time. We had to figure out ways to do it quickly, and that's when we started exploring H2O." Bharadwaj continued, "What took me 6-7 hours, now took me less than 30 minutes on just development hardware."

"Just the way H2O multiprocesses and multithreads, the results that you get are really awesome. Using anything other than H2O meant that we couldn't run more than one model on a dataset and actually provide output in the time we needed."

– Julian Bharadwaj, Senior Data Scientist at PayPal

For Paypal, getting to production quickly was key to reducing churn. Often with restrictive lead-times, the marketing teams needed quality data as quickly as possible, so they could implement campaigns as effectively as possible. Those campaigns needed to be targeted to the right customers, at the right time,

with the right message, promotion, or offer. The data that Bharadwaj delivered to them certainly met those requirements: as designed, the data showed which customers would churn, when, and which features mattered most to them.

"It's been so successful that there is now a program built around the output of these machine learning algorithms," asserted Bharadwaj. "It's also given us a lot of ideas on where to use machine learning, so the inventory of projects we have going into next year has grown because people have seen the impact we made on consumer churn and how successful that program has been."

Bharadwaj concluded, "It means that what we do matters."

Modeling

Benchmarking on Random Forest and H2O's Distributed Random Forest

Software	Hardware	Performance	Data size
R, ODBC	1 processor, 32 GB RAM	Modeling - 6 hrs Scoring - 72 hrs	Train: hundreds of thousands of rows, score on entire consumer base
Revolution R, ODBC	8 processors, 32 GB RAM	Modeling - 2hr Scoring - 48+ hrs (did not complete)	Train on hundred of thousands of rows, score on entire consumer base
H2O, JDBC	3 machines 24 processors, 50 GB	Modeling - 30 min Scoring - 12 hrs (mainly I/O)	Train on hundred of thousands of rows, score on entire consumer base
H2O, JDBC	16 machines, 128 processors, 300 GB	Modeling - 20 min Scoring - 25 min (unzip)	Train on hundred of thousands of rows, score on entire consumer base
H2O, Hadoop	20 nodes	Modeling - 10 min Scoring - 5 min (about 4 min is I/O)	Train and score on entire consumer base !

Goal: Modeling - under 30 min
Scoring - under 1 hour
Enables multiple models daily - a true forecast!

About H2O.ai

At H2O.ai we see a world where all software will incorporate AI, and we're focused on bringing AI to business through software. H2O.ai is the maker behind H2O, the leading open source machine learning platform for smarter applications and data products. H2O operationalizes data science by developing and deploying algorithms and models for R, Python and the Sparkling Water API for Spark. Some of H2O's mission critical applications include predictive maintenance, operational intelligence, security, fraud, auditing, churn, credit scoring, user based insurance, predicting sepsis, ICU monitoring and more in over 5,000 organizations. H2O is brewing a grassroots culture of data transformation in its customer communities. Customers include Capital One, Progressive Insurance, Zurich North America, Transamerica, Comcast, Nielsen Catalina Solutions, Neustar, Macy's, Walgreens, Kaiser Permanente and Aetna.

Tel: +1.650.227.4572 sales@h2o.ai http://www.github.com/h2oai @h2oai

H2O.ai