DARDEN
BUSINESS PUBLISHING
UNIVERSITY of VIRGINIA

UVA-QA-0828
Oct. 14, 2014
DRAFT

KAGGLE'S BIKE-SHARING DEMAND COMPETITION

By 2013, Kaggle—the world's leading crowdsourcing site for predictive analytics—had grown to almost 100,000 participants who had submitted more than 270,000 entries in hundreds of contests. It had hosted forecasting competitions for organizations such as the U.S. Census Bureau, NASA, Amazon, Merck, Facebook, Yelp, Microsoft, and MasterCard. Considering the growing popularity of data science and machine learning, the Kaggle team decided to host a Bike Sharing Demand competition "for the machine learning community to use for fun and practice." The goal was to predict the total number of bikes rented during each hour in a holdout sample.[1]

Kaggle Background

Founded in 2010 as a crowdsourcing site for predictive modeling, Kaggle announced in November 2011 that it had raised $11 million in Series A financing. The round was led by venture capitalists in Silicon Valley; participants included Stanford Management Company, which invests and manages Stanford University's endowment. PayPal founder Max Levchin and Google chief economist Hal Varian participated in the round as well. Kaggle was currently hosting the largest medical prize ever—the $3 million Heritage Health Prize.[2]

The Kaggle platform allowed organizations to crowdsource predictive models from the world's greatest data scientists. Kaggle posted competitions to a community of thousands of PhD-level data scientists located around the world. Competitors accessed training data from Kaggle's website and built predictive models using these data. The majority of participants used the statistical software R. Once a competitor submitted his or her model, it was evaluated for accuracy instantaneously using hidden testing data. The participant's score was immediately ranked on a leaderboard. After the competition ended, the leader was awarded the prize and was often given the opportunity to share with the community the details of the winning model.

[1] https://www.kaggle.com/c/bike-sharing-demand/data (accessed Oct. 7, 2013).

[2] VerticalNews.com, "Algorithms; Kaggle Raises $11 Million in Series A Financing Led by Index Ventures and Khosla Ventures," November 16, 2011.

INTERNAL WORKING DRAFT. NOT FOR DISTRIBUTION OUTSIDE DARDEN. This case was prepared by Associate Professor Kenneth C. Lichtendahl Jr.

Bike-Sharing Systems

The three-year-old Capital Bikeshare system was the first of its kind in the United States. By mid-2014, the system included 337 stations around Northern Virginia, the District of Columbia (DC), and Montgomery County, Maryland. More than 2,700 bikes were in service,[3] and members could check bike availability online (or through the app Spotcycle) at any time for any station.

Annual memberships cost $75, and monthly memberships were priced at $25. The three-day and daily membership fees were $15 and $7, respectively. The first 30 minutes was free for all members. Each of the next 30 minutes was charged at an increasing rate. For annual and monthly members, the second 30 minutes cost $1.50, the third 30 minutes cost $3, and each 30 minutes thereafter cost $6. For a three-day or daily member, it was $2 for the second 30 minutes, $4 for the third 30 minutes, and $8 thereafter.[4]

Unlike New York City's popular bike-sharing system, Citi Bike, the Capital Bikeshare system was close to becoming financially sustainable. Citi Bike had more than 100,000 annual members, while Capital Bikeshare had about 24,000 annual and monthly members. The big difference was that Capital Bikeshare had a higher percentage of daily and three-day users—mostly tourists—taking trips longer than 30 minutes. Bike-sharing systems made most of their money from high overage fees racked up by these casual users. In 2013, 81% of Capital Bikeshare's trips longer than 30 minutes were made by casual users, compared with only 69% of Citi Bike's.[5]

Still, Capital Bikeshare cost DC $392,000 and Virginia's Arlington County $440,000 in fiscal year 2013. These local governments funded bike sharing in the capital region and viewed it as another form of public transportation along with the region's train and bus systems—Metrorail and Metrobus, respectively. From the local government perspective, Capital BikeShare was a financial success. Arlington County recovered 74% of the Capital Bikeshare expenses it funded, which compared very favorably to the 27% of Metrobus's expenses it recovered.

One of the big challenges many bike-sharing systems faced was load imbalance. Stations were either starved of bikes or too congested. A station with no bikes meant missed revenue and dissatisfied users. A completely full station was a problem because users could not drop off bikes there. The imbalances fluctuated during the day. Starvation and congestion increased between 8 a.m. and 10 a.m. For bike-sharing systems, rebalancing bikes was a significant expense.[6] To

[3] http://cabidashboard.ddot.dc.gov/cabidashboard/#Home (accessed Oct. 10, 2014).
[4] https://www.capitalbikeshare.com/pricing (accessed Oct. 10, 2014).
[5] http://www.washingtonpost.com/blogs/wonkblog/wp/2014/04/01/why-dcs-bikeshare-is-flourishing-while-new-yorks-is-financially-struggling (accessed Oct. 10, 2014).
[6] http://research.microsoft.com/apps/video/default.aspx?id=226601 (accessed Oct. 10, 2014).

combat load imbalance, Capital Bikeshare used six vans that traversed the capital region for 20 hours per day, redistributing bikes more evenly across the system.[7]

Forecasting Bike Demand and Usage

In addition to load imbalance, another challenge was forecasting hourly demand on any given day. The number of bikes needed in the system as a whole was, in large part, determined by peak demand. Because demand could not be observed per se (e.g., a potential user who showed up at an empty station may have taken the Metrorail or Metrobus), forecasting hourly usage was the next best thing.

Understanding demand better was a pressing issue for Capital Bikeshare in early 2014. In January 2014, the system's sole station and bike equipment provider, the Montreal-based Public Bike System Co., filed for bankruptcy. By April 2014, an August 2013 order for eight new stations and 71 new bikes had not arrived, and a February 2014 order for 40 stations and 400 bikes was on hold.[8]

Kaggle's Bike Sharing Demand competition generated a great deal of interest among data scientists after it was made public on May 28, 2014. A PhD candidate at the University of Porto in Portugal, Hadi Fanaee Tork, provided the dataset to Kaggle for the competition. The dataset was based on data Tork had obtained from Capital Bikeshare. It contained hourly information on time of day, season, day of the week, holidays, rain, temperature, humidity, wind speed, bike rentals by casual users, bike rentals by registered users, and total bike rentals.

The task was to predict hourly total bike rentals from the 20th day through the last day of each month. The training set comprised data on the first 19 days of each month. By October 10, 2014, with seven months to go in the competition, Kaggle had 944 submissions on its Public Leaderboard. The leader, named "Alliance," had a root-mean-square logarithmic error of 0.24976.

[7] http://www.washingtonpost.com/wp-srv/special/local/how-capital-bikeshare-has-grown (accessed Oct. 14, 2014).

[8] http://www.washingtonpost.com/local/trafficandcommuting/capital-bikeshare-expansion-hindered-by-bankruptcy-of-montreal-based-bike-vendor/2014/04/12/d42c8a2a-bf23-11e3-b195-dd0c1174052c_story.html (accessed Oct. 10, 2014).