mosaic
DATA SCIENCE

http://mosaicdatascience.com/

info@mosaicdatascience.com

Data Science White Paper

Predicting Employee Churn

This white paper examines a machine learning approach to predicting employee churn and optimizing for retention.

Machine Learning

Mosaic data scientists collaborate with customers, digging deep into the data to inform design and deployment of custom ML tools that make a difference.

Artificial Intelligence

Mosaic integrates powerful AI tools into clients' existing technology stack to solve complex business challenges

Business Analytics

Mosaic helps corporations of all shapes and sizes take advantage of their data, transforming their decision-making processes.

INTRODUCTION: RETENTION IS NOT (MERELY) A PREDICTION PROBLEM

Management scientists have studied for decades how to recruit, motivate, and retain employees. These three issues are closely related, so that one cannot effectively address any of them without addressing all of them to some degree. In particular, a highly motivated employee is far less likely to leave; while recruiting an employee who poorly matches the organization culture or job requirements will make it difficult, and perhaps undesirable, to retain the employee. More generally, an organization can only focus on retention to the degree that it consistently recruits well-qualifies candidates. In that case retention becomes an aspect of motivation. Once the employer knows how to motivate employees, the remaining question is determining the optimal retention rate.[1]

It's important to realize that the optimization problem[2] implicit in a retention-improvement project is not to retain all employees for as long as possible. Rather, the problem is **optimal employee retention**. That means retaining desirable employees (those having adequate skills, cultural fit, productivity, ethical commitments, etc.) for as long as they remain desirable, while allowing, encouraging, or requiring undesirable employees to depart in a way that minimizes cost (or to become desirable employees). Defining cost minimization is nontrivial, because it must account for several actual or potential costs: litigation, loss of goodwill, compensation and benefits, skills training, medical or drug rehabilitation, recruitment, accident risk, etc. All of which makes optimal employee retention a hard problem—much harder than, say, predicting which employees are likely to leave in a given time period, or predicting when a specific employee is likely to leave.

In the remainder of this white paper we outline a scientific approach to the use case of predicting employee churn. The approach lets the data science function apply its craft profitably to employee retention, while framing retention as part of a larger optimization problem involving recruitment and motivation. Throughout the paper we'll focus on the example of driver turnover at large trucking companies, which historically have 100% annual driver turnover rates.

The cost of turnover for carriers is high. If the cost of hiring a driver averages $5,000, a company with 200 drivers and a 100% turnover rate would spend $1 million a year on recruitment alone. High turnover also makes it difficult for fleets to operate efficiently, maximize utilization of tractors and trailers and meet customers' service expectations.[3,4]

Again, the objective for these companies is probably not to retain all drivers. In particular, there is a high rate of drug abuse among truck drivers, so that some drivers (habitual users) are presumably unqualified for their jobs.[5] At the same time, some driver motivations (time at home, better pay) are well known, making it possible to quantify at least part of the cost-benefit analysis.

PREDICTING EMPLOYEE CHURN REQUIRES MODELING EMPLOYEE DECISIONS

Recent popular accounts of data science suggest that explanatory (causal) modeling matters less in the presence of large datasets.[6] This can be true when two other things are also true. First, the big dataset must help you build a highly accurate predictive model.[7] Second, the predictive model must suffice to solve the underlying optimization problem. For example, a stock-market trading model may be accurate enough for the model's owner to make a profit buying stocks whose prices the model predicts will rise. The model's owner doesn't need to know what actually makes the prices rise. They just need an accurate prediction to occur with enough lead time to capitalize on the prediction.

In contrast, predicting employee churn at moderate accuracy the quarter in which a truck driver will quit their job only helps reduce turnover under additional conditions that are not generally true. In particular, the class of drivers likely to stick with an employer for much longer than normal must be of significant size. When that's true, one can consider using a predictive model before or after hiring. Each potential application makes additional assumptions.

Pre-hiring tenure prediction. One can use a predictive model to screen job candidates before hiring, aiming only to hire candidates likely to stay for substantially longer than average. Here the model must be able to predict turnover before the hiring decision, when one has no on-the-job behavioral data. Given the ubiquity of high turnover across the industry, screening with a modestly predictive model may have very limited effect. The way forward would be to build a highly accurate predictive model that only relies on data available before hiring.

Post-hiring causal modeling. Some of the predictive model's independent variables may be causal variables (not merely correlates) that the employer can manipulate at plausible cost after hiring. Here the open questions are which variables are causal; and how much of an effect manipulating the causal variables will have, and at what cost. The way forward is to transform the predictive model into a causal model that quantifies (in expectation) the connection between causal variables, costs, and tenure outcomes.[8]

Let's explore each possibility in turn.

PRE-EMPLOYMENT SCREENING

A purely predictive model is likely to be most useful in screening candidates during hiring. Of course, traditional forms of screening (such as drug screening and background checks) are nearly universal.[9] But recently a number of data brokers have entered the market with a wide variety of data that records and summarizes individuals' offline and online behaviors. Some data brokers enrich or interpret the raw data with propensity models, risk scores, etc. Enriching applicant data with third-party data may be an important step towards a highly accurate pre-employment tenure-prediction model. Cutting-edge big data models use historical online behavior, transaction and credit history, social networks, lifestyle variables, etc. to predict individual outcomes such as health risks and drug use. We have every reason to believe some of these variables (combined with variables traditionally available before hiring) may also predict a candidate's tenure if hired.

The superabundance of variables available in third-party datasets[10] means the data science function must explore many potential models using many different combinations of variables, to hope to find the most potent model possible. For example, when Google built a model to predict the spread of influenza in the U.S., it tested 450 million models combining up to 100 variables, ultimately choosing a 45-variable model whose correlation coefficient was between 0.85 and 0.97 during development and validation.[11]

Searching large spaces of possible models means one of two things. One can adopt a Google-like brute-force search strategy that exhaustively searches all possible[12] combinations of variables for a given model type (in the influenza study, Google used a simple linear probit model). This requires some form of embarrassing parallelism (such as the MapReduce algorithm implemented in Hadoop) spread over a great deal of hardware (possibly temporary, virtualized hardware on a cloud infrastructure provider such as Amazon EC2). Or, one can use a clever algorithm (possibly a stochastic-search method, such as genetic programming or simulated annealing) to search intelligently for a good model without testing most of the possibilities. Such search algorithms are more flexible and powerful than embarrassing parallelism, because they can explore different model forms (non-linear as well as linear) as well as different combinations of variables. (Mosaic has considerable expertise with stochastic search.) Non-parametric models can be searched efficiently as well, while matching or improving on the goodness of fit of linear models.

One complexity in using a candidate-screening algorithm is that some of the algorithm's input variables may correlate with attributes that one cannot use during screening, as a matter of law.[13] Even if the proscribed variables are not model inputs, other model inputs may correlate with the proscribed variables, resulting in a model that effectively screens for the proscribed variables. This creates an extra challenge for data scientists: not just verifying that individual input variables do not correlate with the proscribed variables, but also verifying that the model as a whole does not penalize membership in a protected group.[14]

In sum, an employer may hope to improve turnover rates by

1. enriching traditional resume variables with third-party data,

2. using big data and/or intelligent-search technologies and algorithms to construct an optimal model predicting employee tenure, and

3. hiring only candidates the model predicts will have good tenure.

MODELING EMPLOYEE RETENTION AFTER HIRING

A model predicting employee churn is only useful to the degree that it is actionable. To be actionable once hiring has occurred, the model's independent variables must influence retention, not merely correlate with it. And the variables must be within the employer's control at reasonable cost.

There is now a large body of applied social-science research that has identified causal retention variables. For example a University of Alberta study of employee attitudes found that support groups encouraging healthcare employees to focus on finding meaning in their work reduced turnover by 75%.[15] This result is not new. Frederick Herzberg's classic, oft-cited 1987 Business Review article, "One More Time: How do you Motivate Employees?" summarizes the state of social-science knowledge of employee motivation in 1987. It distinguishes "hygiene factors" from motivators. Hygiene factors are necessary to avoid job dissatisfaction (which leads to high turnover), while motivators lead to job satisfaction (and retention). Achievement, recognition, intrinsic work satisfaction, and other aspects of meaningful work are strong motivators. Interestingly, the most common reasons truckers cite for leaving an employer closely match the variables in Hertzberg's lists.[16]

The motivational variables relevant to employee retention operate to different degrees in different employment contexts. For example, inadequate pay is the reason American truck drivers cite most frequently for leaving an employer, but the same variable is not a reason Finnish anesthesiologists cite for changing careers.[17] Moreover, the variables may have significant interactions. For example, recognition is a strong motivator, while relationship with supervisor is mostly a hygiene factor. But lack of recognition by a supervisor may also contribute to a poor supervisory relationship.

Finally, while survey data can measure the reported importance of a causal retention variable, it is far more difficult to measure the actual importance of each variable an employer might attempt to manipulate. Marketers wanting similarly to measure rigorously the efficacy of specific marketing techniques use control groups and hypothesis testing to measure campaign uplift.[18] Data scientists wanting to measure the efficacy of manipulating a set of retention variables may need to employ similar techniques, either conducting retention experiments within the employer or studying retention as a function of a set of variables across employers. There have been academic and governmental studies of driver turnover across companies.[19] Experiments within a company are possible, but as a practical matter are unlikely to explore anything like the entire space of possible combinations of variable values—even if management authorizes experimentation. Companies wanting to be rigorous will probably be limited to testing one variable's retention uplift at a time.

COMBINING PRE-EMPLOYMENT SCREENING AND TENURE MODELING INTO A SINGLE OPTIMIZATION MODEL

A big-data based pre-employment retention screening model should be developed and deployed before modeling and manipulating employee tenure, for two reasons. First, screening may significantly change the employee-tenure model. Second, pre-employment screening is likely to be much easier, in part because the model can be developed and validated without changing employer practices.

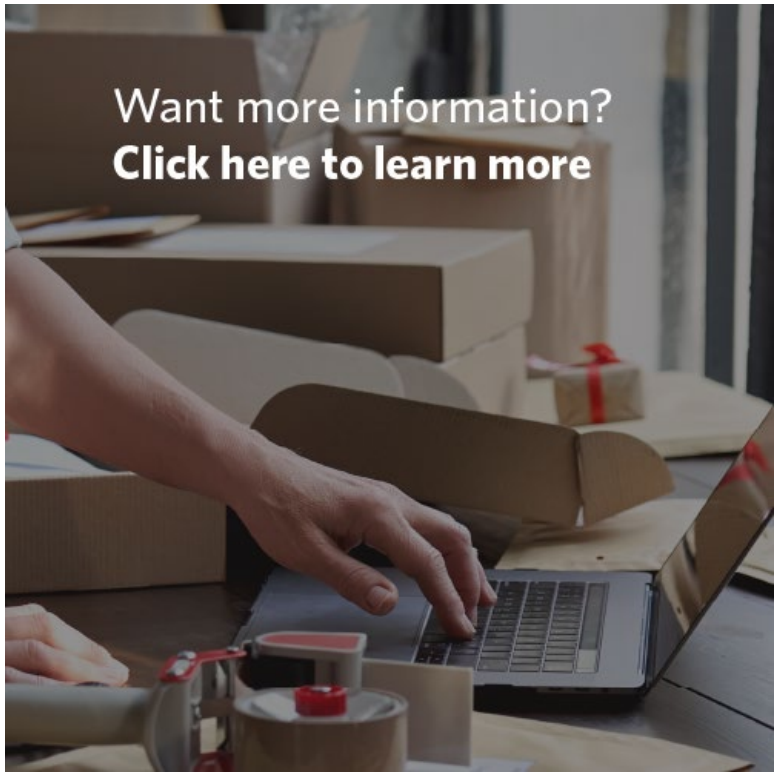
The main consideration in transforming a screening model into an optimization model is balancing the costs and benefits of including different data sources in the model. Some data may cost more than it contributes to cost reduction.

Treating retention modeling as an optimization problem involves two challenges: accounting for variable interactions, and evaluating the costs and benefits associated with each variable. Because variable interactions can be non-trivial, and because there are many possible interactions, the most a data scientist may hope to measure is the direction of the relationship between tenure and each input variable, and an approximate incremental effect size, ignoring possible interactions. If the data scientist anticipates specific interactions, the modeling process should test for them.

The costs involved in a tenure optimization model are likely to be relatively straightforward. For example, truck drivers frequently complain that dispatchers treat them poorly. Manipulating dispatcher behavior may amount to designing and implementing basic training and incentive programs for dispatchers that help them understand how the employer wants dispatchers to treat drivers, and tying dispatcher rewards to driver feedback. Even easier, drivers with more tenure may be paid higher wages. In contrast, the benefits of improving retention may reach beyond increased retention. For example, an employer may determine that adding a certain amount of quarterly vacation time to a driver's schedule doubles average tenure. The same practice might also decrease accident rates by 20%, and improve a driver's vehicle maintenance behaviors. The careful data scientist will consider all plausible effects of each variable manipulation, attempt to measure or estimate them, and include them within the cost-benefit optimization model.

STAKEHOLDER INVOLVEMENT

Finally, as we remark elsewhere[20], much of practical data science revolves around engaging stakeholders and decision makers in the optimization process. A data science team predicting employee churn to optimize employee retention must convince the human-resource function and general management that state-of-the-art screening and causal retention modeling and optimization support a strong business case. An exploratory study should quantify the business case in the process of engaging decision makers and stakeholders, so that (assuming the business case is in fact favorable) the organization supports the implementation project.



Endnotes

1. We know of two Fortune 500 employers whose annual turnover rate is 2%. Some executives at one of these organizations feel the rate is too low, because it does not bring enough fresh perspective into the organization, perhaps encouraging groupthink. This is one reason some turnover can be desirable, and predicting employee churn isn't a worthy use case.

2. And in business, every data science problem is an optimization problem, as we argue in our blog post "Sample Size Matters," . https://mosaicdatascience.com/2014/02/05/data-science-blog-sample-size-matters/

3. https://www.joc.com/trucking-logistics/labor/ata-reports-97-percent-truck-driver-turnover-rate_20131212.html (visited March 17, 2014)

4. $5,000 per driver is an oft-cited recruitment cost figure, predicting employee churn See e.g. http://xrscorp.com/blog/fleet-management/truck-driver-retention/

5. http://oem.bmj.com/content/early/2013/09/13/oemed-2013-101452.abstract (visited March 17, 2014).

6. Viktor Mayer-Schonberger and Kenneth Cukier, "Correlation," Big Data (Mariner, 2014), pp. 50-72, for example.

7. Indeed, the examples in Big Data illustrate that very high accuracy is a primary benefit of using very large datasets.

8. Whether the current turnover rate is suboptimal seems to be an open question. "As no trucking company has successfully demonstrated that the costs associated with attacking turnover can be offset by profits gained from increased retention, the assumption could be made that the level of turnover and retention is appropriate for the prevailing business climate in the motor carrier industry." http://fleetowner.com/fleet-management/driver-turnover-does-trucking-ignore-solutions (visited March 17, 2014)

9. http://www.rsiinsurancebrokers.com/12_12-hiring-and-retention-of-commercial-drivers/ lists traditional screening techniques (visited March 17, 2014)

10. Acxiom markets over 1,600 variables describing individuals and their households, for example.

11. Jeremy Ginsberg et/al, "Detecting Influenza Epidemics Using Search Engine Query Data," Nature Vol. 457 (Feb. 2009), available at http://static.googleusercontent.com/media/research.google.com/en/us/archive/papers/detecting-influenza-epidemics.pdf (visited March 17, 2014).

12. It does not appear that Google tested all available combinations, much less all theoretically possible combinations: 45! ≈ 1056 is much larger than 450 million ≈ 108. Google's data set was 50 million queries. Google tested combinations of up to 100 queries, and settled on a particular set of 45. The paper is somewhat vague about the exact method of choosing and evaluating candidate sets of queries, beyond saying that individual queries were scored by mean correlation across nine regions, and sets of top-scoring queries were tested.

13. The list of prohibited variables is at http://www.eeoc.gov/employers/upload/eeoc_self_print_poster.pdf (visited My 17, 2020).

14. The problem is not at all new to big data. There is considerable law-and-economics research on the problem of employment screening and discrimination. See for example George J. Borjas and Matthew S. Goldberg, "Biased Screening and Discrimination in the Labor Market," The American Economic Review (December 1978), pp. 918-922, available at http://www.hks.harvard.edu/fs/gborjas/publications/journal/AER78.pdf (visited March 19, 214). Mosaic encourages your organization to consult with an employment-law attorney before using any form of candidate screening.

15. http://www.sciencedaily.com/releases/2008/11/081126122317.htm (visited March 17, 2014)

16. http://www.truckinginfo.com/channel/fleet-management/article/story/2008/01/top-10-reasons-drivers-leave.aspx (visited March 17, 2014).

17. http://www.ama-assn.org/resources/doc/physician-health/icph2010-lindfors-leino-elovainio-nurmi-1.pdf (visited March 17, 2014)

18. Nicholas J. Radcliffe, "Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Models," Direct Marketing Journal (Direct Marketing Association Analytics Council, 2007) pp. 14–21.

19. Here are a few: http://www.atri-online.org/research/results/musical_chairs.pdf, http://www.memphis.edu/ifti/pdfs/cifts_examining_driver_turnover.pdf, and http://www.fmcsa.dot.gov/facts-research/research-technology/tech/driver-retention-safety.pdf (visited March 17, 2014).

20. https://mosaicdatascience.com/about-4/