Fall 2021 Update: Using data-driven strategies to provide personalized job matches
Brian Heseung Kim and Benjamin L. Castleman

Several months ago, we announced a new project we have embarked on to augment traditional career advising for near-graduates in community college programs with data-driven, personalized job matches for students. In close partnership with the Virginia Community College System (VCCS), the Tennessee Board of Regents (TBR), and the Ascendium Education Group, we are developing a job recommendation algorithm and an accompanying intrusive career advising intervention to support community college graduates in identifying and applying for degree-relevant jobs which would provide stable employment and commensurate compensation. For more background and context on the project, please read our initial project release here.

As we approach the later stages of algorithm development and the beginning stages of the accompanying intervention design, we want to share our progress, challenges, and learnings in this process so far. In this blog post, we offer a brief overview of the design process for our algorithm, selected learnings, and next steps. In future blog posts, we will also offer a deeper technical dive into some of the specific analytic challenges we’ve encountered, with the intention of open-sourcing our approach and gathering feedback from the broader data science community.

Goals and Progress

To recap, our primary goal in this project is ultimately to improve the labor market outcomes of community college students, specifically by providing them with actionable, data-driven personalized information on available job postings as they embark on the job market immediately following graduation. More concretely, we want to help students filter down to available jobs that are actually relevant to them, and then prioritize applying to jobs that are mostly likely to provide them with well-paying, stable employment.

To this end, we have aggregated and harmonized data from several key sources:

1. **Student-level academic and demographic data** from our community college partners
2. **Historic employment and earnings data of all community college students**, also from our community college partners
3. **Job postings data** purchased from Burning Glass
4. **Occupation-specific average earnings data** at the state- and county-level from the Bureau of Labor Statistics (BLS)

Our general gameplan for the algorithm given this data context is then to take the universe of contemporaneous job postings from the Burning Glass data, **identify for each student which jobs are**
relevant to them both academically and geographically using the student-level data, and finally sort those jobs based on the best proxies for job quality available to us from the various data sources described above. This first process of identifying relevance has been fairly straightforward; Burning Glass data includes crosswalks that help us determine what programs of study are most relevant to which occupations, and so in the remainder of the post we focus primarily on the second step of determining job quality.

Ideally, every job in the Burning Glass data would have a concrete wage range associated with it, and we could then readily direct students to the highest paying jobs they are qualified for based on their program of study and degree level. Unfortunately, this is much less commonly the case: of the 1.5 million jobs we observe from 2010 through 2019 that are relevant to VCCS grads, only 11% have a listed salary. Thus, we need to find additional sources of information to rank relevant jobs for a given student.

For a given posted job, we then want to infer the likely quality of that job (most clearly proxied by the wage range) as best we can using the empirical data we have on the job’s occupation and employer. For occupations, we can leverage BLS data, and for employers, we can leverage historic employment and earnings data from past community college graduates. Note that the historic employment data we have reports what firm or organization graduates work at but not their specific occupation; this is true of Unemployment Insurance (UI) data systems in most states.

As it stands now, we can identify for graduates from more populous geographic regions and from larger programs of study employers who have historically paid graduates from their program or college well. For other graduates, this process is a little more complicated due to a variety of design challenges we’ve encountered. We describe a few of these issues at a high level below, and be on the lookout for future posts digging in on the technical specifics.

**Design Challenge #1: Specificity vs. Precision and Coverage**

One persistent challenge we’ve run into is the inherent trade-off in specificity versus precision and coverage. For example, using the BLS data, some occupations are common enough that the average earnings of these positions can be estimated as specifically as at the metropolitan statistical area (regions of grouped counties). This specificity would allow us to say something to the effect of, “People working as network administrators in the Charlottesville, VA metropolitan area earned average salaries of $87,140, and we can then recommend these currently-open network administrator jobs near Charlottesville, VA that are a good match for your program of study.”

However, there also exist many occupations for which average wages can only be estimated at the state- or even national-level. So the question surfaces: to what extent do we think the salary of a job posted in a given city/town in Virginia can be reasonably predicted by the Virginia average? We might
think the more granular measurements are better predictors, but we also reduce the precision of estimates (fewer and fewer individuals contributing to each estimate) and coverage (fewer and fewer occupations with sufficient sample) as we drill down.

Similarly, we encounter this issue when estimating an employer’s average compensation for past community college students. Because our historical data does not include occupational information (e.g., what job a given student had at a given employer we observe them working with), we can only really look at the employers as a whole. But even that said, we can at least try to estimate an employer’s average compensation per program. So if we see a Management-related job posted at Employer A, we might want to reference how well Employer A has paid Management students in the past, rather than just how well Employer A has paid all students in the past. Or maybe we’d rather reference how well Employer A has paid past Management students from the same college as a current graduate. As before, though, these more specific and relevant employer compensation estimates commensurately reduce the precision and coverage of any ensuing estimates.

While we can calculate program—and in some cases, program-by-college—specific estimates of past average employer compensation for some firms, the question remains: how do we pick between all of these estimates at varying levels of specificity and robustness, both in terms of occupations and employers?

Design Challenge #2: Reconciling Conflicting and Incomplete Data

Imagine you see two jobs available for a given graduate – both lacking a listed salary (or, equivalently, both listing broad salary ranges that almost entirely overlap). For one job, we see that it is an occupation that typically pays well in Virginia. For the other job, we see that the employer has historically paid other graduates of the same program well. Which one do we prioritize for the graduate?

This is the fundamental issue we observe across the majority of our data due to data points with varying precision and specificity. What we’ve opted to do, and are currently in the process of implementing, is creating a principled “ensemble” weighting scheme that allows us to form an aggregated index of likely job quality using whatever datapoints we do have available, while weighting more heavily whatever more specific data points are available (such as the posted salary range that appears with a job listing). We’ve crafted a process that we think allows us to complete this task with a defensible final weighting scheme. More on this to come!

Design Challenge #3: Data, Data, Everywhere, But...

Finally, we’ve run into some expected, but still difficult, data limitations. For example, there exist only so many relevant jobs for community college graduates in a given labor market: it is the case that
within a given quarter, we might not expect to see even one relevant posted job for graduates of certain programs (e.g., radiology technicians) in the more rural areas of Virginia – in which case, what value can we provide with the algorithm for these graduates? Similarly, we’ve found that linking the historical employment and earnings data from past community college students is more difficult than anticipated due to the lack of a standardized crosswalk of employer IDs that we can then link back to the Burning Glass data. Issues like the latter are more surmountable (using some creative fuzzy-matching processes, for example), while issues like the former are just realities of the job markets proximate to some community colleges.

**Going Forward**

The main implication of these data challenges is that we can offer an “optimal” algorithm – one that robustly and convincingly identifies, for each community college graduate, jobs that align with their program of study, are reasonably close to where they live, and that are most likely to compensate them well – mostly to graduates in more populous geographic regions and from larger programs of study.

In the next phase of our algorithm development, we will be formalizing the “ensemble” approach that provides each community college graduate with the most personalized and high-value job information we can based on their program of study and institution. This will extend our ability to provide meaningful, data-driven insights to a broader swathe of smaller programs and graduates attending colleges in more rural regions. This ensemble approach will require us to establish clear processes for ranking the jobs that we display to graduates and their career advisors; we will provide a second public post that describes and seeks feedback on our ensemble approach to the job matching algorithm.

We plan to complete work on the ensemble algorithm in Virginia by the end of 2021. We will then work with our partners at the TBR to replicate (and adjust as necessary) the algorithm in Tennessee. In parallel, we will work with behavioral scientists on our team and career advisors in Virginia and Tennessee to design an intrusive career advising intervention to deliver the job matches to students as effectively as possible, while also offering students support and guidance through the actual application processes as well.

We’re excited to share progress and insights into this work given the rapidly growing array of projects applying data science strategies to improving educational outcomes – an endeavor that, as we show here, is often easier said than done. If you’ve found this overview interesting and want to learn more about any of what we’ve described, or have insights you’d be willing to share on how you’ve addressed similar design issues in the past, please don’t hesitate to get in touch! We’re still very much looking for ways to improve, and we’re always learning more as we go.