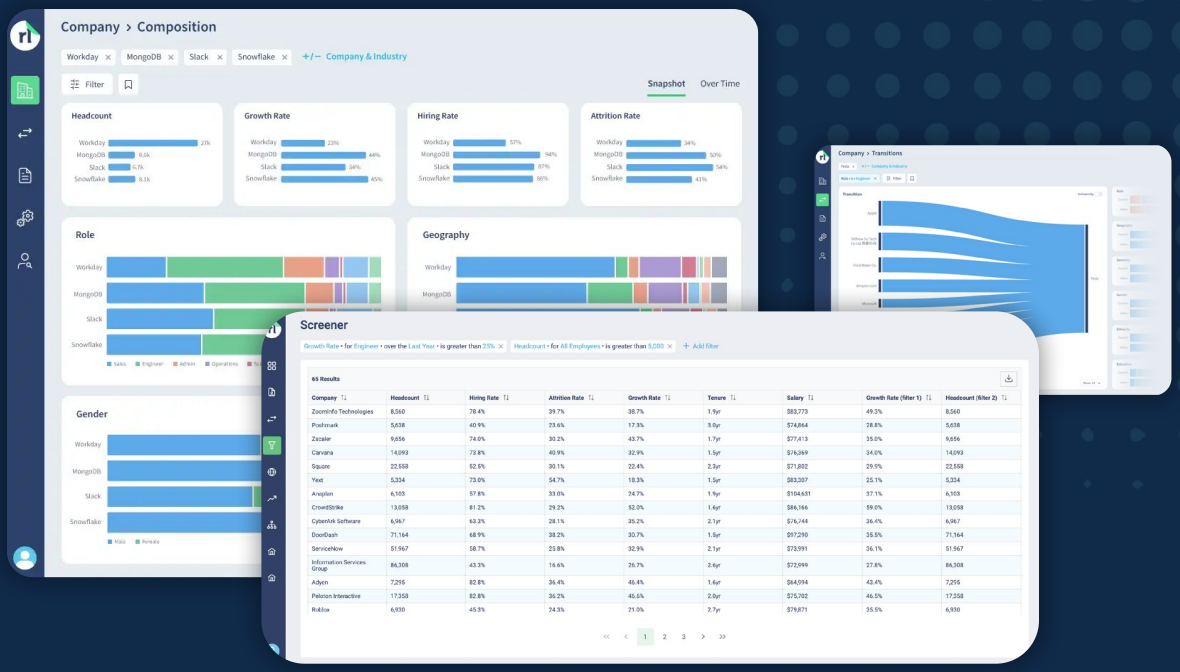
LinkUp
JOB MARKET DATA

revelio labs

Supplemental Data: Modeled Salaries Methodology



MARCH 2023

LinkUp
JOB MARKET DATA

revelio labs

About LinkUp and Revelio Labs

Revelio Labs and LinkUp have worked to expand their current data offering to include comprehensive salary ranges for every job listing contained within LinkUp's database of employment data.

These salary ranges are model-based using various sources of salary data. The models are trained using salaries in publicly-available visa application data, self-reported data, and job listings, and take into account job title, company, location, years of experience, and seniority, as well as the year observed. The training includes data down to the month level for more recent years.

COVERAGE

Salary ranges based on this model have been applied to all LinkUp job listings dating back to 2007.

Location

Salary ranges are based on U.S. salary data. If a role is hiring for multiple locations, each location has an applicable salary range distinct to that location.

Salaries for non-U.S. listings are calculated by multiplying applicable U.S. salaries by scaling factors that account for occupation and geographic variables.

SOURCES

Sources for the salary data used to train the model include:

Job listings

Job listings used to calculate salary ranges are collected from aggregator sites: Indeed, Monster, and LinkedIn. LinkedIn data includes both stated salaries and modeled salary ranges.

H-1B filings

Salary information for H-1B visa filings by company, title, and location.

Self-reported salary data

Employee self-reported salary data is collected from Layoffs.fyi and Glassdoor.

DATA DELIVERY

The data based on this model is currently available via these data delivery methods:

- Snowflake
- Amazon Web Services (AWS) S3

Additional delivery methods can be supported on an as-needed basis, including Azure and Google Cloud.

The Model

The model is trained independently on each dataset using two different modeling structures:

Standard XGBoost Model
The XGBoost model is trained to predict both the expectation value and the variance associated with the prediction, under the assumption of a normal salary distribution.

Bayesian Additive Regression Tree (BART) Model
The BART model allows for the construction of a true non-parametric probability distribution of salaries associated with a job posting, which gives us even more detail about the range of possible salaries.

WORKFLOW

Salary ranges based on this model have been applied to all LinkUp job listings dating back to 2007.

Location
Salary ranges are based on U.S. salary data. If a role is hiring for multiple locations, each location has an applicable salary range distinct to that location.

Salaries for non-U.S. listings are calculated by multiplying applicable U.S. salaries by scaling factors that account for occupation and geographic variables.

INPUT FEATURES

Job title

The model uses a fast-text model to embed the raw job title, as well as a cleaned version of the job title into an embedding space that aligns with the Revelio job taxonomy.

Company

Includes company-level embeddings based on transitions:

> Example: During a particular month, the model observes that a lot of Facebook engineers are moving over to Meta. Those companies would be tied closely together during analysis because they share a significant portion of the workforce, so their salary offerings are likely to be competitive with one another. The model would also incorporate the directionality of those migration flows.

Location

The model uses city-level data, going one level deeper than the level of Metropolitan Statistical Areas (MSAs). Population and demographic statistics, such as average income and median house prices, are sourced from government databases and used as covariants

Seniority

The seniority metric is created using an ensemble model.

First, information about an individual's current job, including their title, company, and industry, are used to generate an initial seniority score based off labeled seniority data.

Second, details about individuals' job histories, such as the duration of their previous employment, aggregate transitions between positions, and likely career paths,are taken into account to create a second seniority value.

Finally, the average age of a position relative to the occupation type or industry is used to create a third seniority score.

The scores from these models are averaged together to arrive at a continuous seniority metric for an individual. To convert this continuous seniority metric into an ordinal value, we gather samples of seniority predictions corresponding to recognizable keywords such as "junior", "senior", "director", etc. and map the metric to the most likely bin. This allows us to attach meaning to the raw metric values, and to bin seniorities into discrete buckets.

LinkUp JOB MARKET DATA

reveliolabs

Seven Ordinary Seniority Levels

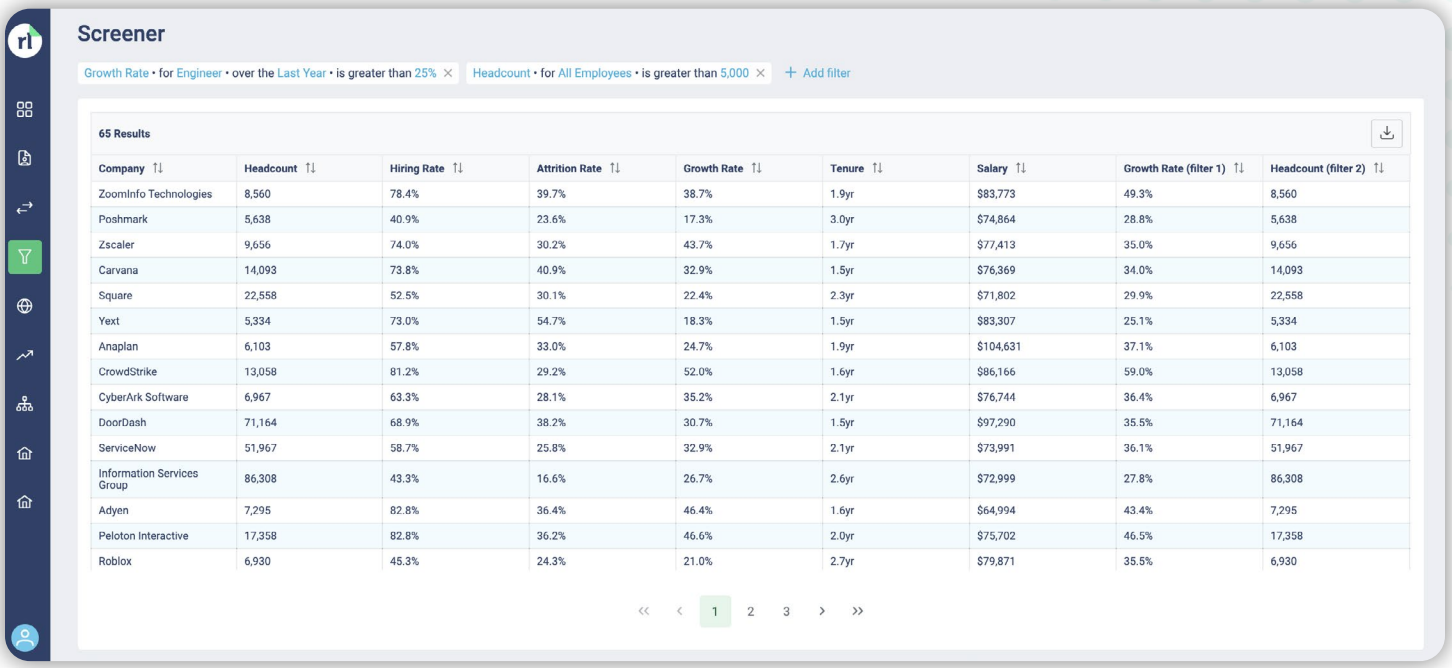
1. Entry level / Intern (Ex. Accounting Intern, Software Engineer Trainee, Paralegal)
2. Junior Level (Ex. Account Receivable Bookkeeper, Junior Software QA Engineer, Legal Adviser)
3. Associate/Analyst Level (Ex. Senior Tax Accountant; Lead Electrical Engineer; Attorney)
4. Manager Level (Ex. Account Manager; Superintendent Engineer; Lead Lawyer)
5. Vice President Level (Ex. Chief of Accountants; VP Network Engineering; Head of Legal)
6. Director Level (Ex. Managing Director, Treasury; Director of Engineering, Backend Systems; Attorney, Partner)
7. C-suite Level (Ex. CFO; COO; CEO)

Years of Experience

The model accounts for the years of experience the role requires, as salary is commonly commensurate with experience levels.

Year Observed

When the model applies salary data to a job listing, the date the listing was made available determines how the model calculates corresponding salary information.



Company	Headcount	Hiring Rate	Attrition Rate	Growth Rate	Tenure	Salary	Growth Rate (filter 1)	Headcount (filter 2)
ZoomInfo Technologies	8,560	78.4%	39.7%	38.7%	1.9yr	$83,773	49.3%	8,560
Poshmark	5,638	40.9%	23.6%	17.3%	3.0yr	$74,864	28.8%	5,638
Zscaler	9,656	74.0%	30.2%	43.7%	1.7yr	$77,413	35.0%	9,656
Carvana	14,093	73.8%	40.9%	32.9%	1.5yr	$76,369	34.0%	14,093
Square	22,558	52.5%	30.1%	22.4%	2.3yr	$71,802	29.9%	22,558
Yext	5,334	73.0%	54.7%	18.3%	1.5yr	$83,307	25.1%	5,334
Anaplan	6,103	57.8%	33.0%	24.7%	1.9yr	$104,631	37.1%	6,103
CrowdStrike	13,058	81.2%	29.2%	52.0%	1.6yr	$86,166	59.0%	13,058
CyberArk Software	6,967	63.3%	28.1%	35.2%	2.1yr	$76,744	36.4%	6,967
DoorDash	71,164	68.9%	38.2%	30.7%	1.5yr	$97,290	35.5%	71,164
ServiceNow	51,967	58.7%	25.8%	32.9%	2.1yr	$73,991	36.1%	51,967
Information Services Group	86,308	43.3%	16.6%	26.7%	2.6yr	$72,999	27.8%	86,308
Adyen	7,295	82.8%	36.4%	46.4%	1.6yr	$64,994	43.4%	7,295
Peloton Interactive	17,358	82.8%	36.2%	46.6%	2.0yr	$75,702	46.5%	17,358
Roblox	6,930	45.3%	24.3%	21.0%	2.7yr	$79,871	35.5%	6,930

LinkUp
JOB MARKET DATA

revelio labs

FAQ

What happens if Revelio incorporates more sources or otherwise refines the model in the future (how does this impact point-in-time data)?

In the case that new data sources are added, or the models are refined, salaries are recalculated for all historical job openings. Depending on the degree to which models change, versioning is supported, meaning that Revelio can maintain older versions of the model until such a time that the client is ready to update.

Is it easy to add new dataset inputs in the future?

Yes. Because the model is an 'ensemble model,' or every dataset is treated as independent, the model scales very well.

What are some potential reasons for differences between modeled and posted ranges in job descriptions?

Because stated salary ranges can be very broad and employers tend to set the ceiling high to avoid weeding out potential applicants, the modeled ranges are generally tighter and, in our view, can be more informative to what the company is actually paying for a given position.

Why not just extract ranges from postings?

Companies only recently began including salary ranges on any meaningful scale due to new legislation. Even still, this remains largely limited to certain locations and is still not a widely adopted best practice yet.

Using Revelio's modeled data gives us history that we could not get from simply parsing, which will benefit clients who want to look at trends over time. Additionally, as noted above, stated ranges can still tend to be too broad to be truly informative or overstate what a company is actually paying most employees in a given position.

Does the data include other compensation metrics or just base salary?

The data only includes base salary data for now but there are plans to include total compensation in the future.