



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Workforce Analytics Architectures for Predicting Attrition and Capacity Gaps in Large Enterprises

Rohini Tirumalaraju

Independent Researcher India. Email: rohinitirumalaraju83@gmail.com

Abstract

Attrition prediction is usually presented as a modelling problem. In a large enterprise it is first an architecture problem. The signals that carry information about whether a person will leave are distributed across a human capital management system of record, a performance module, a talent and career interest module, a learning platform, system access logs and purchased labour market data. None of those systems was designed to be joined to the others, they refresh on incompatible cycles, they carry different identifiers for the same person, and they record events at a different time from when the events occurred. This paper is a practitioner architecture account of a workforce analytics platform built across six such sources for a workforce of about fifteen thousand people in more than forty locations. It describes the identity resolution layer that makes any downstream analysis possible, the provenance and refresh cadence of each feature, and what breaks when a source changes. It then argues that the analytically interesting target is not attrition but regretted attrition, because predicting departures is easy and predicting the ones that hurt is the actual problem: reported figures show total attrition close to flat near twelve percent while the regretted share of departures fell from twenty two percent to eight percent over four years. The paper treats a flight risk score as a governed artefact rather than a measurement, sets out what such a score is and is not licensed to do, and reports the departure model's evaluation exactly as described, including one case where the reported precision does not reconcile with the reported confusion matrix. Capacity gap forecasting is presented as the operational output that makes prediction actionable, and internal mobility as the intervention lever rather than a reporting dimension. Six signals the organisation found predictive and six it found weak are reported with their conditional rates. All figures are the author's own operational reporting from a single enterprise, and no controlled evaluation was performed.

Keywords: workforce analytics, people analytics, employee turnover, regretted attrition, flight risk scoring, capacity planning, internal mobility, data provenance, concept drift, algorithmic governance

I. Introduction

The literature on employee turnover is nearly a century old and unusually well developed. Meta-analytic work has established which antecedents predict voluntary departure and how strongly [10], theory has explained why embedded people stay even when dissatisfied [15], and the unfolding model has shown that many departures begin with a discrete shock rather than a slow decline in satisfaction [12]. A centenary review consolidates the field and is candid that prediction at the level of the individual remains weak relative to explanation at the level of the population [11].

What that literature does not supply is a design. An enterprise that wishes to act on turnover research needs to know, for each of its people, and on a weekly or monthly cycle, what the relevant antecedents currently are. That is not a modelling question. It is a question about where each signal lives, how often it refreshes, whether it can be reliably attached to the right person, and what happens to every downstream feature when a source system is upgraded. The people analytics literature has begun to say the same thing from the other direction. Reviews of the field find that its central constraint is not analytical technique but data infrastructure and the skills to build it [2], that credible human capital analytics depends on data quality work that is unglamorous and expensive [14], and that the evidence base for value creation is thinner than the enthusiasm [13, 19]. Practitioners who have built such systems argue that the discipline fails when it starts from available data rather than from a business question [18].

This paper is an account of building such a system in one large enterprise, and its argument is stated plainly at the outset. Attrition prediction is an architecture problem before it is a modelling problem. The model is the smallest part of the work and the least fragile part of the result. What determines whether the system produces anything usable is the layer beneath it: identity resolution across systems that do not share keys, reconciliation of contradictory status records, alignment of event dates against recording dates, and honest accounting of which features are complete enough to score an individual rather than a cohort.

Four further claims follow from the account and organise the rest of the paper.

First, the target matters more than the accuracy. Total attrition is a poor objective because most of it is either desirable or unavoidable. The distinction that carries decision value is between total attrition and regretted attrition, the departures the organisation would have prevented if it had known. Functional and dysfunctional turnover were separated in the literature four decades ago [8], but the distinction is routinely collapsed in practice into a single headline percentage.

Second, an individual score is not a measurement. A flight risk score computed from hand set weights is an operational prompt with a governed set of permitted uses. Treating it as a probability, or letting it reach a performance or compensation process, converts a support tool into a mechanism that can help produce the outcome it forecast.

Third, prediction is inert without a capacity output. A list of people who might leave is not actionable. A forecast of which roles in which locations will be short of capacity in thirty, sixty and ninety days is actionable, because it maps onto requisition timing, internal candidate identification and coverage planning.

Fourth, the intervention that works is mobility, not persuasion. Where the organisation could offer a genuine internal move, retention followed. Where it could offer only a conversation, it largely did not. This is consistent with evidence that external hiring is more expensive and initially lower performing than internal mobility for equivalent roles [4].

The paper closes on governance, because individual level predictions about identifiable employees are sensitive personal data and carry a duty of care that aggregate analytics does not. The account describes an access model, a purpose limitation rule, and a bias audit in which the model correctly reproduced a real pay inequity. That last case is the clearest illustration in the whole account of why excluding protected attributes from a feature set does not make a model fair [17].

2. Scope and Evidentiary Basis

This section states what this paper is, because the rest of it should be read against that statement rather than against the conventions of an empirical study.

This is a practitioner architecture account from a single large enterprise. It describes one system, built once, by the team the author led. It is not a controlled evaluation, a comparative study, or a replication. No experiment was designed, no treatment and control populations were constructed, no pre registration exists, and no independent party instrumented or audited the measurements reported here. Where the paper reports a number, that number is the author's own operational reporting drawn from internal dashboards, model performance summaries, recruiting invoices, employee records and exit interview classifications. Readers should treat every figure as unverified operator supplied input rather than as measured fact.

The distinction between what is measured and what is described matters throughout. The architecture, the data model, the reconciliation logic, the scoring rules, the access control model and the intervention workflow are described: they are design artefacts and the account of them is direct. The outcome figures are measured only in the ordinary business sense that somebody counted them in a system of record, not in the scientific sense that an instrument was validated and a protocol followed. That distinction is not pedantic. The regretted attrition classification, which carries most of the paper's outcome claim, rests on a human judgement recorded at exit about whether the organisation would have wanted to retain the person, made by people who knew the analytics programme was being assessed.

The population is about fifteen thousand active employees across eight business units, more than forty locations and twelve countries with a significant presence. Coverage is not uniform. Contingent workers are reported at roughly half coverage. Several countries have sparse data because of local confidentiality law and partial system adoption. Reliable history extends three to four years, following system implementation in the two years before the analytics work began.

That last figure is the first of several internal inconsistencies in the source material, and this paper resolves them explicitly rather than silently. The history depth is stated variously as twenty four months, as three to four years, and as four years of model training data. This paper uses three to four years, the value given in the dedicated scope answer, and notes that the model is elsewhere described as trained on 2021 to 2023 with 2024 held out, which is consistent with three years of training history. The complete set of contradictions found, together with the value chosen in each case and the reason, is recorded in a data handling note accompanying this paper. Several derived percentages were dropped altogether because they do not reconcile with the components they are computed from. Where a derived figure was dropped, this paper states the components and leaves the arithmetic to the reader rather than publishing a number that fails its own check. No reconciling figure has been invented anywhere in this paper.

Three specific consequences follow. Total value and return on investment figures appear in the source material in at least five mutually inconsistent forms, so no single return multiple is asserted here as a finding. Prediction quality is reported exactly as described, including a reported precision that does not follow from the reported

confusion matrix. And the account of model drift is explicitly framed in the source material as a hypothetical timeline rather than an observed episode, so this paper presents it as a design requirement derived from a threat to validity, not as a reported event.

3. Related Work

Turnover theory. Meta-analytic work established that job satisfaction, organisational commitment, perceived alternatives and search behaviour predict voluntary turnover, with effect sizes that are reliable in aggregate and modest at the individual level [10]. Job embeddedness extended the account by showing that fit, links and sacrifice predict staying independently of satisfaction, which is why dissatisfied people often remain [15]. The unfolding model showed that many departures follow a shock, an event that prompts a person to consider leaving, rather than a gradual decline that a survey would detect [12]. The centenary review consolidates a hundred years of this work and is explicit that the field predicts populations better than persons [11]. The functional taxonomy separated turnover the organisation regrets from turnover it does not, and argued that a single aggregate rate systematically overstates the problem [8]. That distinction is the direct antecedent of Section 6 of this paper. Practitioner facing syntheses have catalogued the specific misconceptions that survive in organisations despite the evidence, including the belief that pay is the dominant lever and that engagement survey scores identify leavers [1]. Section 10 reports an independent practitioner encounter with several of those same misconceptions.

People analytics. Reviews of HR analytics find limited evidence that it delivers the value claimed for it, and attribute the gap partly to weak research design in the field and partly to the difficulty of the underlying data work [13]. A widely cited critique argues that human resources functions are poorly placed to meet the data challenge because they lack the analytical and engineering capability, and that the binding constraint is infrastructure rather than algorithms [2]. Work on credible human capital analytics stresses data quality, governance and the credibility of the analyst as prerequisites for influence [14]. A scoping review maps the conceptual boundaries of people analytics and identifies employee privacy and consent as an unresolved boundary condition rather than an implementation detail [19]. Practitioner accounts argue for starting from a business question and for accepting that most of the effort is integration rather than analysis [18]. This paper is consistent with all of these and adds a specific architectural description of what that integration effort consists of.

Methods. Time to event methods are the natural formalism for turnover, since the quantity of interest is when a person leaves rather than whether they ever do, and the proportional hazards model remains the standard tool [7]. The account described here uses a fixed horizon binary formulation instead, predicting departure within ninety days, and Section 8 discusses what that choice costs. Departure within a short horizon is a rare event, roughly three percent per quarter at the rates reported here, and rare event prediction is sensitive to class imbalance in both training and evaluation [5]; because the imbalance is extreme, overall accuracy is nearly uninformative and precision, recall and calibration must be reported separately. Where a score is presented to a manager as a percentage the relevant property is calibration rather than discrimination, and calibration must be checked rather than assumed [16], since gradient boosted tree ensembles of the kind used here are strong discriminators and are not calibrated by default [6]. Finally, the assumption that the relationship between features and outcome is stable over time is unsafe in a workforce setting, where reorganisations and reductions in force change the meaning of every feature at once. This is concept drift, and the literature on detecting and adapting to it applies directly [9].

Governance. The legal and ethical literature on employment algorithms is the part of this field that has moved fastest. Disparate impact can arise from a model that contains no protected attribute at all, because retained features proxy for excluded ones and because the training labels themselves encode past discrimination. Empirical work on algorithmic hiring found a wide gap between fairness claims and fairness practice, and argued for disclosure of what is actually measured rather than reliance on the removal of protected attributes [17]. Reviews of workplace privacy set out why employee monitoring is a contested terrain in which the employer's legitimate interest and the employee's reasonable expectation are genuinely in tension, and why consent obtained in an employment relationship is weak consent [3]. Data protection scholarship has clarified how limited the individual's right to an explanation of automated processing actually is under European law, which matters for any multinational scoring employees across jurisdictions [20]. Section 11 reads the described governance model against this literature and finds it strong on access control and weak on employee facing transparency, which remains a design intention rather than a deployed capability.

4. Why Prediction Is an Architecture Problem

The pre analytics state described in the source material is not a state of ignorance. It is a state in which the relevant information existed and was not joinable. Career aspirations were recorded in development plan documents nobody read. Learning enrolments that revealed a person's intended direction sat in a learning platform that did not talk to the performance module. Requisitions were filled externally without any check against a population of internal candidates whose interest was documented somewhere else. Attrition was managed through exit interviews, which are by construction posterior, and through annual engagement surveys, which introduce a twelve month lag between measurement and insight.

The illustrative failure the account opens with makes the point precisely. A high performing analyst in one function had documented interest in a second function, had completed relevant certification, and resigned to take a role in that second function at a competitor. Every signal needed to predict and prevent that departure existed in a system the organisation owned. None of those systems was joined to another. The cost components reported for that single case are between seventy five thousand and one hundred fifty thousand dollars of external recruiting and onboarding cost and a further one hundred twenty thousand dollars or more of lost productivity across a four month ramp for the replacement. The source material also states a total of two hundred thousand to three hundred fifty thousand dollars for that case, which does not follow from those components. The components are reported here and the stated total is not used.

Figure 1 shows the resulting architecture. It has four layers, and the argument of this section is that the second layer is where the work is.

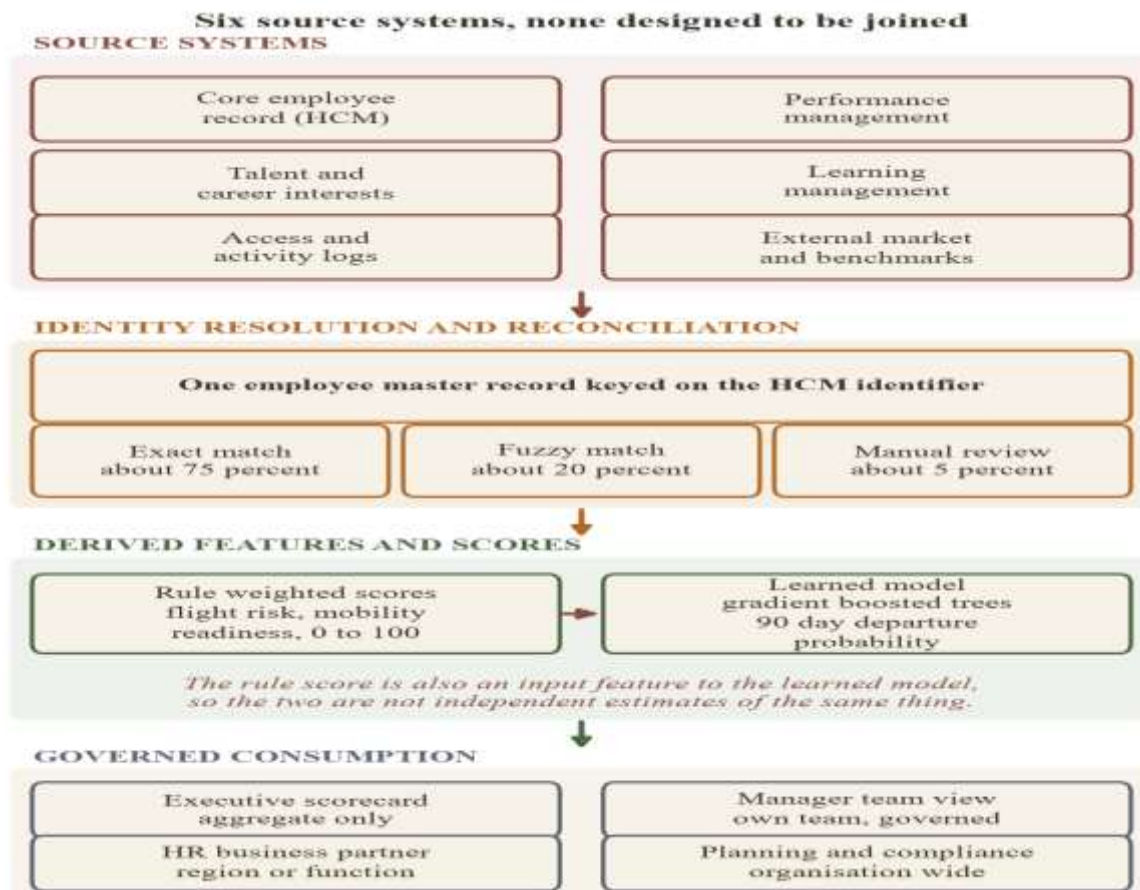


Figure 1. The six source systems feeding the platform, the identity resolution layer between them and any feature, the derived score layer, and the governed consumption layer. Note that the rule weighted score is itself an input feature to the learned model, so the two are not independent estimates of the same quantity.

Identity is the first problem and it is not trivial. The same person exists as an employee identifier in the human capital system, a worker identifier in the performance module, a candidate identifier in recruiting, a user string in the learning platform, and a member identifier in benefits, with names rendered differently in each. Matching proceeded in three tiers: exact matching on name and date of birth or on an explicit integration identifier, covering about seventy five percent of the population at a reported ninety nine point five percent accuracy; fuzzy matching on surname, initial and hire date, covering about twenty percent at a reported ninety two percent accuracy with the remainder requiring manual override; and manual reconciliation by a specialist for

about five percent. The source material describes the second tier as twenty percent of the remaining unmatched records rather than of the population, which cannot be reconciled with the three tiers summing to the whole. This paper uses the reading in which the three tiers are shares of the total population, because that reading reconciles.

The consequence of getting identity wrong is not a degraded model but a nonsensical one. If a person's performance record attaches to one identity and their learning record to another, the platform reports one individual with strong performance and no development activity and a second with development activity and no performance history, when in fact there is one person doing both. The reported end state is ninety eight percent of records correctly matched, with the remaining two percent flagged and excluded from analytics rather than used with uncertainty. Excluding them is the right call, because a record that might belong to two people cannot be scored at all: the error is not noise around a true value but attribution to the wrong person.

Status is the second problem. A person marked terminated in the system of record may continue to appear as active in recruiting and may show activity in the learning platform after their departure date. Without a designated source of truth for termination, departed people inflate the active population and every rate computed against it is wrong in the same direction. The rule adopted is that once the human capital system marks a person terminated, every other system masks their data, and a weekly reconciliation flags any system presenting a person the system of record does not know.

Reference data is the third problem. A person's official position, their job code and the pay scale they are administered under can disagree, because position updates lag compensation changes. Analytics keyed on the wrong one of the three produces a defensible looking answer to the wrong question.

Time is the fourth problem and the most interesting. Events happen before they are recorded. A person decides to leave, tells their manager two to four weeks later, and is processed as terminated some weeks after that, and throughout the notice period they appear in the system as an active employee. A model trained on the state of the system at a point inside that window learns that this person was staying, and never sees the behavioural signals they exhibited in the weeks before they resigned, which is exactly the period the model needs. The fix described is to key everything on event date rather than recording date, to include departed people in training with a lookback window reaching back before their resignation, and to carry known systematic lags explicitly for the sources that batch rather than stream. The source material reports a comparison between a model without lag adjustment and a model with it, at sixty five and seventy nine percent predictive accuracy, a difference of fourteen percentage points favouring the adjusted model. That comparison is presented through generic labelled organisations rather than as an instrumented experiment at the enterprise studied, and it is reported here on that basis.

5. Data Sources and Feature Provenance

Each of the six sources contributes different features, at a different cadence, with different completeness, and with a different failure mode. Treating them as interchangeable inputs to a feature matrix is the error that produces confidently wrong scores. Table 1 records the provenance of each source as reported.

Table 1. Provenance, refresh cadence and reported completeness of the six sources. Completeness values are as reported for the current record. The weaknesses in the third column are those the source material identifies, not an independent assessment.

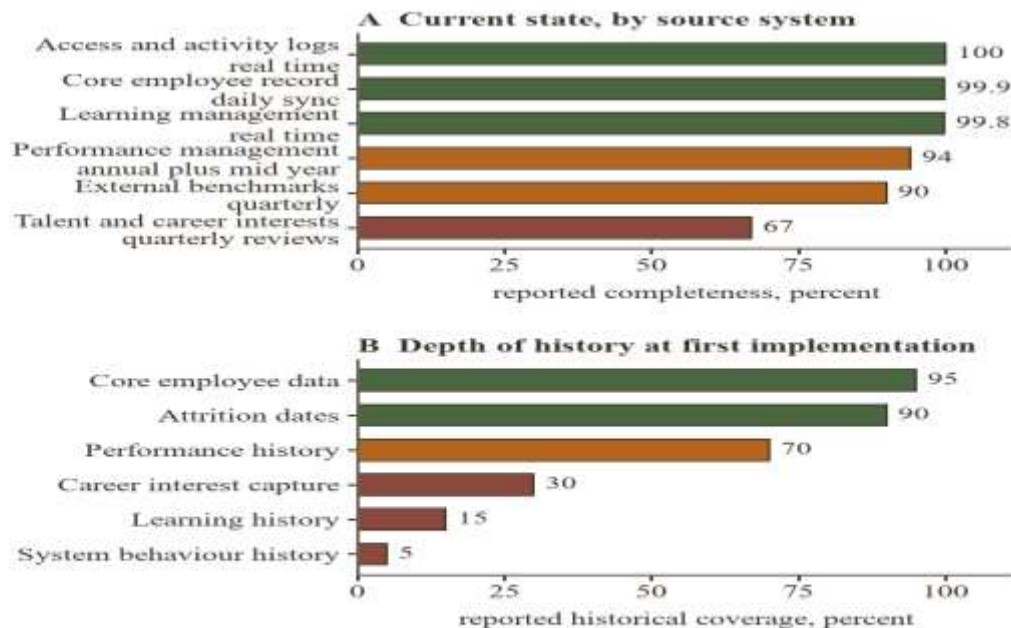
Source	Cadence, completeness	Reported weakness
Core employee record , human capital management	Daily sync. 99.9 percent	Five systems, five identifiers for one person. If the mapping fails the record splits and analysis breaks
Performance management , same platform	Annual, plus mid year and continuous feedback. 94 percent	Some roles are excluded from review by contract, so no rating is not the same as no performance
Talent and career interests , same platform	Quarterly talent reviews and mobility events. 67 percent	Manual entry, loosely enforced, so a blank field cannot be read as an absence of interest
Learning management	Real time on completion. 99.8 percent	Reliable in the present, but history predates the platform and is largely absent

System access and activity logs	Real time. 100 percent	Logging began recently, so the per role baseline the strongest signal needs has little history behind it
External market and benchmark data	Quarterly. 90 percent	Publication delay silently ages the denominator of every compensation ratio

Two features of this table determine what the platform can honestly do.

The first is the completeness gradient. The two features carrying the most reported predictive weight, the compensation ratio and the change in system access activity, sit at opposite ends of the reliability range. The compensation ratio depends on external market data refreshed quarterly at ninety percent completeness, so a person can be scored as underpaid against a benchmark that is a quarter stale. Access activity is logged at complete fidelity in real time but has the shallowest history of any source, which means the per role baseline that a forty percent drop is measured against is itself estimated from a short window. Neither is a reason not to use the feature. Both are reasons why the score derived from them should carry a confidence band, which the account does provide.

The second is the difference between current completeness and historical depth. These are distinct measures and the source material reports them separately, with values that appear to contradict one another until the distinction is made explicit. Figure 2 shows both without merging them.



Panel A is reported per source completeness of the current record. Panel B is described as typical for a 15,000 person organisation implementing analytics for the first time, not as measured at the enterprise studied. The two sets are not merged, and the same source can appear high in A and low in B.

Figure 2. Reported completeness by source. Panel A is the current state of the record per source.

Panel B is the depth of usable history at first implementation, which the source material describes as typical for an organisation of this size implementing analytics for the first time rather than as measured at the enterprise studied. A source can be near complete in Panel A and nearly absent in Panel B, and system behaviour data is exactly that case.

The historical depth figures are the harder constraint. A learning platform that records ninety nine point eight percent of completions today supplied about fifteen percent of the completions that occurred before it was installed. Access logging is complete today and covers about five percent of the relevant history. Career interest capture, the feature most directly relevant to mobility, stands at about thirty percent historically. This is the ordinary condition of a first analytics implementation and it has a specific analytical consequence: the features with the shortest history are precisely the behavioural ones that the account reports as most predictive, so the model's strongest signals are the ones with the least evidence behind them.

The mitigations described are conventional and are described honestly as imperfect. Missing values for departed people were imputed from cohort means over comparable employees matched on role, tenure, performance and location, with the imputation flagged and confidence intervals widened. Three model variants were built under conservative, moderate and liberal imputation, and agreement across the three was used as a confidence signal rather than as validation. For newly hired people with no history at all, trajectories of a small number of comparable earlier hires were used as a proxy, which the source material itself labels honest guessing with wide uncertainty. Resampling was used to make fuller use of a short panel, with the explicit acknowledgement that resampling adds no information.

The threshold rules stated alongside those mitigations are unusually disciplined for a practitioner account. Two to three years of history is described as adequate for cohort level analytics and risky for individual predictions; under two years is described as adequate for aggregate analytics only, with individual scoring withheld. That is the correct shape of rule, and it draws the line where the duty of care actually changes, which is the point at which a number becomes attached to a person.

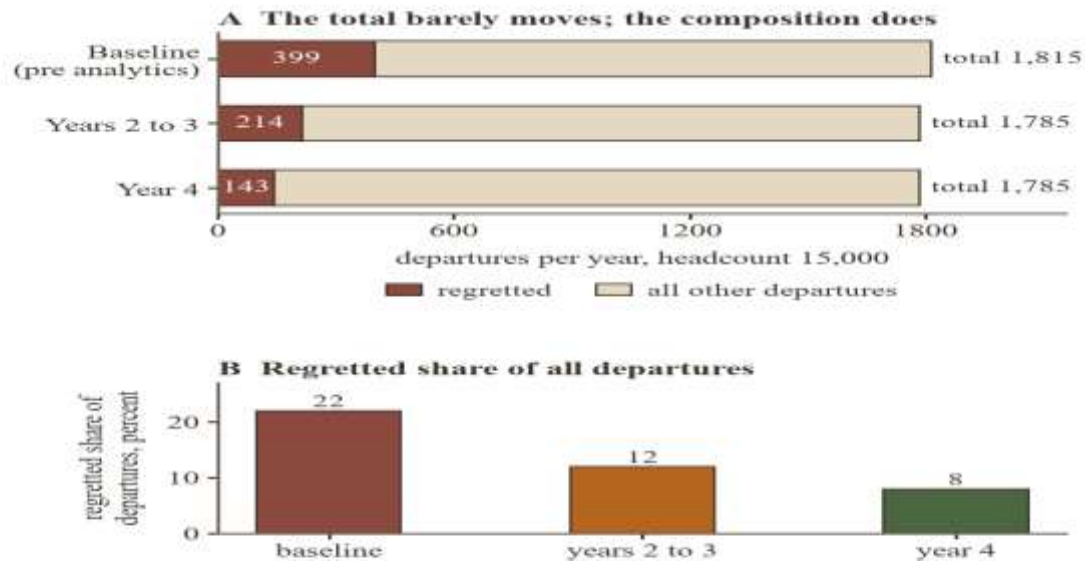
Feature stability under source change. The provenance question that matters operationally is not where a feature comes from but what happens to it when its source changes. A performance rating scale revision breaks a three year trajectory feature, because the trajectory compares periods measured on different scales. A learning catalogue renumbering breaks the mapping from courses to competencies. A migration of the tooling that access logging observes resets every per role activity baseline simultaneously, which presents to the model as a synchronised collapse in engagement across the population. That last case shows why drift must be monitored at the source level and not only at the level of aggregate model accuracy [9]: a detector watching only the output would see the model degrade without knowing that the cause was a tooling change rather than a change in behaviour.

6. Total Attrition Versus Regretted Attrition

Predicting departures is not difficult. Roughly twelve percent of a fifteen thousand person population leaves each year; the rate is stable, and a model that predicts the base rate will be approximately right about the aggregate every year. The difficulty is that most of that twelve percent is either desirable or beyond influence. Some departures remove poor performers, some resolve genuine misfit, some follow relocation or family circumstance, and some are involuntary. Acting on total attrition means spending retention effort indiscriminately across all of it.

The distinction the account draws, and the one the turnover literature drew four decades ago [8], is between total attrition and regretted attrition, which the source material defines as departures the organisation would have wanted to prevent and believes it could have prevented. This is the analytically interesting target for three reasons. It is a much smaller population, so intervention is affordable. It is the population where intervention plausibly works, since the person did not fundamentally want to leave. And it changes independently of the total, which means it can improve without the headline number moving at all.

Figure 3 shows that last point directly.



Counts and shares reconcile against the stated totals to within one person. A separate dashboard figure of 3.2 percent regretted is excluded because it cannot be reconciled with any of these departure counts.

Figure 3. Reported departure counts and the regretted share of them. Total departures move by about thirty people out of roughly eighteen hundred across the whole period. The regretted share falls from twenty two percent to eight percent. Counts and shares reconcile against the stated totals to within one person

At baseline, reported total attrition is twelve point one percent, or 1,815 people, of whom twenty two percent, about 399 people, were classified as regretted. In the second and third years total attrition is eleven point nine percent, or 1,785 people, of whom twelve percent, 214 people, were classified as regretted. In the fourth year the total is unchanged and the regretted share is eight percent, or 143 people. Every one of those figures reconciles against its own components. The reduction from 399 to 143 is a fall of about sixty four percent in the regretted count, and the source material's stated sixty four percent is correct for those components.

Two figures in the source material for this same quantity were not used. A dashboard reports regretted attrition at three point two percent described as a share of total turnover, which cannot be reconciled with any of the departure counts above and appears elsewhere in the same material as a share of headcount instead, where 480 of 15,000 is indeed three point two percent. A dashboard also reports an eighty eight percent improvement from twenty two percent to three point two percent, which is arithmetically about eighty five and a half percent even taking both endpoints at face value. Both were dropped. Separately, the baseline regretted count appears once as 330 people, which implies a total turnover of about 1,500 rather than 1,815. This paper uses 399, from the answer that supplies the raw figures, and reports the discrepancy.

The measurement problem in the target. The regretted classification is a judgement, not an observation, and it is the weakest link in the outcome chain. It is assigned during or after the exit process on the basis of whether the person was wanted and whether the organisation believes it could have retained them, and both halves of that judgement are affected by knowing the answer. Once an analytics programme exists and is being assessed on regretted attrition, the same departure can be reclassified as unavoidable rather than preventable without anybody acting in bad faith. The source material supplies partial and genuinely useful corroboration: exit interview reasons shifted from thirty eight percent citing an absence of growth to eighteen percent, and from twenty two percent citing compensation to eight percent, while the share citing an external opportunity fell only from fifty six to forty eight percent. That pattern is consistent with the controllable reasons being addressed and the less controllable ones remaining, and it is stronger evidence than the classification itself because it is a report by the departing person rather than a judgement about them. It remains self reported at exit.

Why the composition shift is the right claim. The honest reading of Figure 3 is not that the programme reduced attrition. It did not, in any material sense; the total moved by well under one percentage point and the account does not claim otherwise. The claim is that the composition of departures shifted toward the kind an organisation cannot prevent. If the regretted classification is even roughly reliable, that is the correct thing to

have achieved, and it would be invisible in a headline attrition rate. An organisation that manages to the aggregate number would have concluded that four years of work produced nothing.

7. Flight-Risk Scoring and Its Limits

The platform produces two individual level rule weighted scores and one model output. The flight risk score starts from a base determined by performance rating, on the stated reasoning that stronger performers are more externally marketable and weaker performers have fewer options, then adds or subtracts for a compensation ratio below zero point eight five or above one point one five, for tenure in the first year or in the three to five year band or beyond ten years, for engagement below or above stated thresholds, for the presence or absence of development activity, for recent internal applications or recent advancement, and for the scarcity of the person's skills in the external labour market. The total is clipped into a nought to one hundred range and mapped to five bands from very low to critical. The mobility readiness score is built the same way from cross functional interest signals, skill readiness, performance, time in current role and breadth of network.

Rule scores of this kind are often dismissed in favour of learned models, and that dismissal is wrong here for a specific reason: the weights are the organisation's hypotheses, written down. Each term corresponds to a claim a practitioner can argue with, which is what evidence based practice in this field asks for [18]. The account states this as its first lesson for organisations starting out, that one should begin from a business hypothesis and test it rather than instructing an algorithm to find patterns.

The scores nonetheless have four properties that constrain what they license, and the account is candid about some and silent on others.

A score is not a probability. The flight risk score is a clipped sum of integer weights on a nought to one hundred scale. The departure model separately outputs a probability of departure within ninety days. These are different objects and the source material interchanges them, presenting an eighty five point score as an eighty five percent flight risk in a manager facing view. That is a category error with a practical cost, because a manager who reads eighty five percent will act as though the departure is near certain, and no such claim is supported. If a number is presented to a manager as a percentage then it must be calibrated, meaning that people scored at seventy six percent should depart at approximately seventy six percent [16]. The account does describe monthly calibration checking and does present a band of plus or minus eight points rather than a point value, which is the right instinct. The calibration evidence supplied, however, is at the aggregate rate level rather than the individual level, and Section 12 separates those.

The bands are ambiguous at their boundaries and the labels are applied inconsistently. The five bands are stated as nought to twenty, twenty to forty, forty to sixty, sixty to eighty and eighty to one hundred, so a score of exactly twenty belongs to two bands. A worked manager facing example then labels a score of eighty five as high when eighty five falls in the critical band, which in the stated governance triggers a different and more senior action. For a rule system whose entire advantage is auditability, neither defect is cosmetic. Clipping compounds this: summing the maximum available positive adjustments exceeds one hundred, so the highest risk individuals are compressed into an indistinguishable ceiling, which is exactly where the score is supposed to discriminate.

The rule score is an input to the learned model. The account lists the calculated flight risk score among the input variables of the departure model, alongside several of the features the rule score is itself computed from. The two are therefore not independent, and the model's reported feature importances cannot be read as a decomposition of independent contributions. This is visible in Figure 1 and it is an ordinary design decision, but it means the comforting observation that the rule score and the model agree carries no information.

Table 2 reports the model evaluation exactly as described, alongside the metrics that follow from the reported confusion matrix.

Quantity	As reported	Recomputed from the matrix
Evaluation basis	Train 2021 to 2023, test on 2024	A genuine temporal holdout
Population evaluated	14,010	Consistent with a ninety day window at the stated annual rate
Actual departures	470	470, being 320 plus 150. Reconciles
Predicted departures	475	475, being 320 plus 155. Reconciles
Recall	68 percent	68.1 percent, being 320 of 470. Reconciles

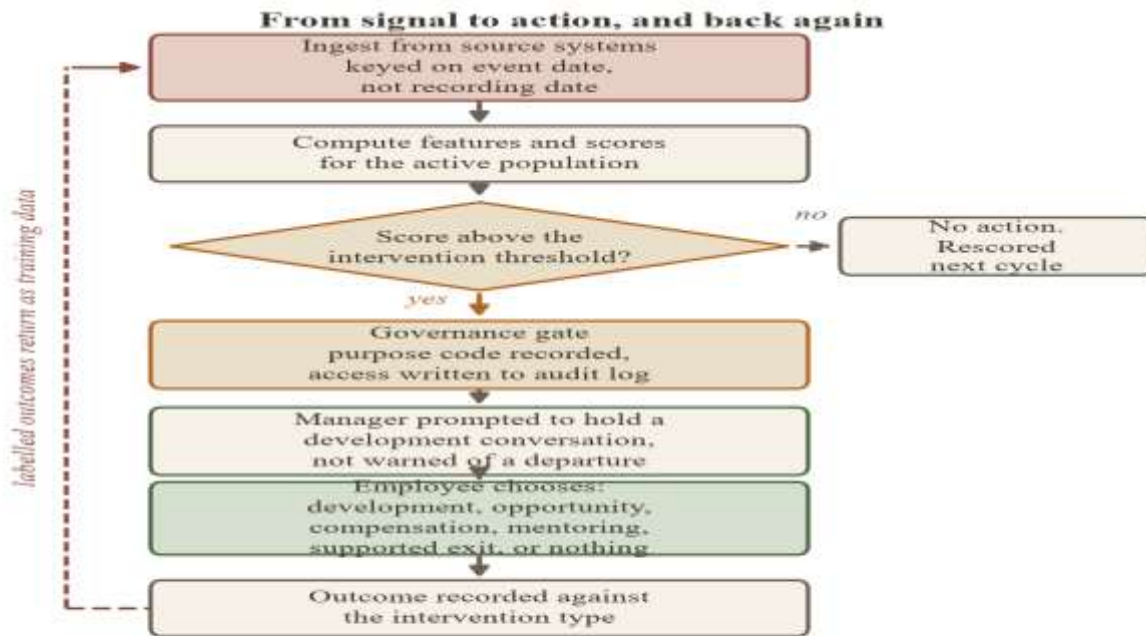
Precision	76 percent	67.4 percent, being 320 of 475. Does not reconcile
Accuracy	76 percent on test	97.8 percent on the matrix, an imbalance artefact, uninformative

Table 2. Reported evaluation of the ninety day departure model against the metrics implied by the reported confusion matrix. Every quantity reconciles except precision. Gradient boosted trees are not calibrated by default [6], and overall accuracy should not be used, since a model predicting that nobody leaves would score above ninety six percent on this population. A backtest re predicting the prior year is reported at seventy three percent with no components supplied.

The precision discrepancy is worth stating plainly rather than smoothing over. The reported confusion matrix is internally consistent: rows, columns and margins all sum correctly, and recall computed from it matches the reported sixty eight percent to a tenth of a point. Precision computed from the same matrix is 320 of 475, or about sixty seven percent, against a reported seventy six percent. One of the two must be wrong and nothing in the source determines which, so this paper reports both and uses neither as a headline. The practical difference is real but not dramatic: at sixty seven percent, about a third of flagged people would not have left rather than about a quarter.

The class imbalance also explains why overall accuracy should never be quoted here. With 470 departures in 14,010 people over ninety days, a model predicting that nobody leaves attains better than ninety six percent accuracy while being useless. Rare event prediction requires precision, recall and calibration reported together, and requires care in training as well, since naive resampling to correct the imbalance distorts the very probabilities a manager facing percentage depends upon [5].

Figure 4 shows the loop the score sits inside, which is where its remaining limits live.



Schematic of the described control flow. No timing or conversion measurements are encoded in this figure.

Figure 4. The described control flow from ingestion through scoring to intervention and back into training data. Two properties are structural rather than incidental: the score reaches a manager only through a purpose gate, so what the manager receives is a prompt rather than a prediction; and the loop trains on outcomes it has already intervened in.

That second property is the deep problem with any deployed individual risk score and it is not solvable by better modelling. Once a score triggers an intervention, the outcome that returns to the training set reflects both the person's original trajectory and the intervention. If interventions work, the model is trained toward underestimating risk for exactly the people it flagged. If an intervention is handled badly, and Section 11

describes precisely how it can be, the score contributes to producing the departure it forecast. Neither effect can be estimated from operational data alone, because the counterfactual was never observed. Establishing it would require withholding intervention from a randomly selected group of flagged employees, which is a study design with obvious ethical cost and which was not performed here.

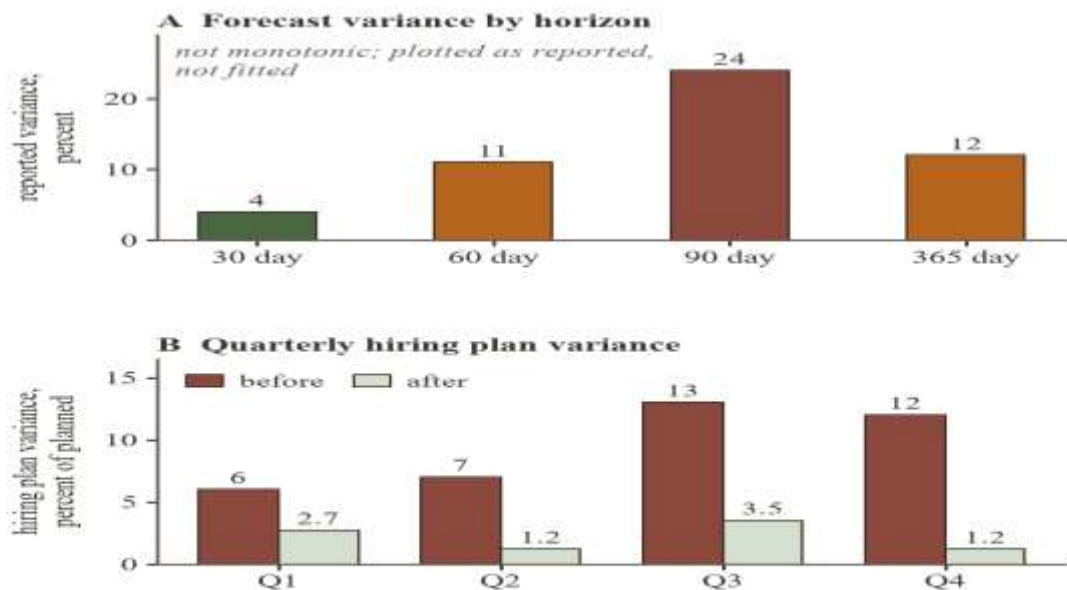
8. From Prediction to Capacity-Gap Forecasting

A ranked list of people who may leave is not an operational artefact, because nobody has a decision whose input is that list. The decisions that exist are whether to open a requisition, when to open it, whether to look internally first, and how to cover a gap that will open before a replacement can be productive. Capacity gap forecasting is the transformation that turns the prediction into those decisions.

The capacity model described takes as inputs the current open requisitions, average time to fill by role and location, predicted promotions derived from mobility readiness scores, predicted departures from the departure model, planned absences including parental leave and sabbatical, and a skill demand forecast from planned projects. Its output is a gap by role, location and horizon at thirty, sixty and ninety days. A worked departmental output combines one predicted departure, one planned internal promotion and one six month leave against three new projects that increase headcount need, and resolves to a stated gap of five positions. Reaching five from those components requires treating the internal promotion as capacity neutral within the department, which is reasonable and is not stated. The attached recommendation is what makes the output operational: begin recruiting now against a forty five day typical time to fill, identify internal candidates already assessed as ready, and plan coverage through cross training or contingent staffing.

This composition is what distinguishes a capacity forecast from an attrition forecast. Most of the capacity gap in any given quarter is not attrition at all; it is promotion, leave and new demand. An attrition model on its own predicts one term in a sum of four, and the other three are more predictable than it is. That is a strong argument for building the capacity layer even where the departure model is weak, and it also bounds how much improving departure prediction can improve planning.

The forecasting posture described is a rolling horizon rather than an annual plan. The annual plan previously locked in November was, by the third quarter, a forecast made in a different world. The replacement is a monthly refreshed rolling forecast with explicitly graded confidence: high at thirty days, good at sixty, moderate at ninety and low beyond. Figure 5 shows the reported variance by horizon and the reported quarterly hiring plan variance before and after.



The before quarters average about 9.5 percent, not the 13 percent stated alongside them, so the stated average is reported here but not used. Annual totals were close in both states; the gain is in intra year dispersion.

Figure 5. Panel A shows reported forecast variance at each horizon. It is plotted as bars rather than as a curve because the series is not monotonic and fitting a curve would invent a shape not reported.

Panel B shows reported quarterly hiring plan variance before and after. The before quarters average about nine and a half percent, not the thirteen percent stated alongside them, so the stated average is reported and not used.

Panel A carries an instructive irregularity. Variance rises from four percent at thirty days to eleven at sixty and twenty four at ninety, then falls to twelve percent at three hundred sixty five days. A forecast worse at ninety days than at a year is not paradoxical once the quantities are distinguished: the annual figure is a total, and totals benefit from errors in opposite directions cancelling, while the ninety day figure is a claim about a specific quarter whose errors have no longer period to absorb them. Panel B shows the same effect from the other side. The quarterly misses before the programme were six, seven, thirteen and twelve percent against a plan of four hundred fifty hires per quarter, while the annual total came within about five hires of eighteen hundred. The plan was accurate in aggregate and wrong every quarter, which produced the reported symptoms: hiring bunched into one quarter and starved in another, with recruiting premiums paid for urgency and onboarding delivered in waves.

The gain reported afterwards is therefore correctly stated as a reduction in intra year dispersion rather than an improvement in annual accuracy. Quarterly variance falls to two point seven, one point two, three point five and one point two percent, averaging two point one percent, which is what the components give. The stated before average of thirteen percent does not follow from six, seven, thirteen and twelve, whose mean is nine point five, so the comparison is best made against the quarterly figures directly.

What the fixed horizon formulation costs. Predicting departure within ninety days is a binary classification at a fixed horizon, and it discards timing information that capacity planning needs. Two people at the same ninety day probability, one likely to leave in three weeks and one in eleven, present entirely different coverage problems, because the difference is on the order of a typical time to fill. A time to event formulation estimates a hazard over time and answers the timing question directly [7], and it also handles the censoring that a fixed horizon binary handles badly: a person who is still present at the end of the window is not a negative case, they are a case whose outcome is not yet known. The account does not use survival methods, and the worked capacity outputs quietly need them, since they attach dates and confidence to individual predicted departures. The gap between the model's output and what the capacity layer consumes is bridged by judgement.

9. Internal Mobility as Intervention

The obvious use of an attrition prediction is to try to persuade someone to stay. The account's more interesting finding is that persuasion is the weakest available intervention and that internal mobility is the strongest.

The reported intervention outcomes make the point. Of about 1,200 people identified as high flight risk in a mature year, retention conversations were initiated with 1,150, and 680 of those people, about fifty nine percent, remained. By intervention type, offering a career opportunity retained seventy four percent, creating a development plan sixty eight percent, adjusting compensation sixty five percent, connecting a mentor fifty two percent, and holding a conversation with no specific intervention thirty eight percent. Two caveats are required. The cohort sizes behind each type are not supplied, so the weighted average cannot be checked against the overall fifty nine percent, and the intervention was chosen rather than assigned, so the ordering partly reflects which interventions were offered to whom. The source material also labels two different rows as most effective, an inconsistency worth noting since seventy four percent exceeds sixty eight percent.

Even with those caveats the spread from thirty eight to seventy four percent is wide, and its direction is consistent with turnover theory. Job embeddedness predicts staying through fit, links and sacrifice [15], and an internal move that increases scope while preserving accumulated links and organisation specific knowledge raises embeddedness on all three dimensions in a way a compensation adjustment does not. The unfolding model implies the same from the other direction: if a departure often begins with a shock prompting a person to reconsider their path, then supplying a credible alternative path inside the organisation addresses the shock rather than bidding against its consequence [12].

The architectural requirement for treating mobility as an intervention differs from the requirement for reporting it. Reporting needs a count of internal placements after the fact. Intervening needs three things available at the moment a requisition is created: documented career interest, an assessment of readiness, and the ability to surface both to the hiring manager inside the requisition workflow. The first is the least complete source in the entire platform, at sixty seven percent currently and about thirty percent historically, which sets a hard ceiling on how well mobility matching can work regardless of model quality.

The reported outcome is a shift in internal placements from 145 of 1,820 hires, about eight percent, to 224 of 1,798 hires, about twelve and a half percent, over four years. Every one of those percentages reconciles against its own numerator and denominator. The cost argument attached is that an internal placement costs ten to twenty five thousand dollars against seventy five to one hundred fifty thousand for an external hire, a difference the literature supports on both cost and early performance grounds [4]. One derived figure was dropped: the

source material states a cumulative improvement of 230 additional internal hires, which does not follow from 145 rising to 224, a difference of seventy nine placements per year at the end state. The components are given here and neither the 230 figure nor the annual saving computed from it is used. A smaller inconsistency is that the per placement saving appears as both eighty five thousand and one hundred thousand dollars; the lower figure is the one consistent with the stated cost ranges.

A structural caveat applies to the whole mobility argument, and the account raises it. An internal move fills one requisition and opens another, so the vacancy is relocated rather than removed. The net gain is real only if the cascade terminates in a role genuinely easier to fill externally, or if the retained person would otherwise have left. Where a regretted departure was prevented the gain is unambiguous; where it was not, an internal move is a transfer with transaction costs.

10. What Does Not Predict Attrition

The most useful practitioner content in the account is its list of signals that were expected to predict attrition and did not. Negative results of this kind are rare in this literature and they are precisely what an organisation starting out needs, because each of these weak signals is the basis of a widely sold intervention.

The source material assigns each signal an ordinal strength expressed as a percentage and labelled a correlation. Those labels are not used in this paper. They are internally inconsistent, with several signals placed in a band whose stated range does not contain their own value, they do not match the conditional rates reported for the same signals in the same passages, and a correlation of the magnitude claimed for a single behavioural indicator would be an extraordinary finding in this field. What the source material does report cleanly, and what is reported here, is the conditional rate for each strong signal against a stated comparison base. Table 3 gives those.

Signal, as operationalised	Reported outcome rate	Stated comparison base
System access activity falls by 40 percent or more from the per role baseline	62 percent depart within 90 days	12 percent annual attrition
Four or more disengagement indicators present together	84 percent depart within 26 weeks	12 percent annual attrition
Three or more disengagement indicators present together	71 percent depart within 26 weeks	12 percent annual attrition
Three or more cross functional access requests within 12 months	68 percent move internally within 12 months	8 percent internal mobility rate
Strong performer with a compensation ratio below 0.87	58 percent depart within 12 months	18 percent for strong performers
Two or more skills scarce in the external labour market	54 percent depart within 24 months	12 percent annual attrition

Table 3. Signals the account reports as predictive, with the conditional rate and comparison base as stated. Horizons differ by signal and are not comparable across rows. The ordinal strength scores the source material attaches to these signals are omitted, since they neither match these rates nor fall inside their own stated bands. The common property of the four strongest signals is that they are behavioural traces rather than reports. Nobody is asked anything. A person reduces their system activity, or requests access to a function they do not work in, and both are recorded as a side effect of doing the work. That property is what makes them predictive, and it is also exactly what makes them ethically loaded, which Section 11 takes up.

Against those, six signals are reported as weak.

Engagement survey scores. The expectation is that low scorers leave. The account reports almost no relationship, and the reasons it gives are sound. A survey administered annually cannot track a decision made over weeks. Survey items measure satisfaction and connection, which are not intentions. Many dissatisfied people have limited alternatives and stay. And responses are subject to social desirability, since nobody discloses an active job search on an instrument their employer administers. The turnover literature is consistent with this: satisfaction is a reliable population level antecedent with a modest individual level effect [10], and the unfolding model explains directly why a periodic satisfaction measure misses shock driven departures [12].

Stay interview responses. The same logic applies more strongly, because the question is asked directly by the employer. The account contrasts an exit interview citing a desire for new challenge with a stay interview six months earlier reporting enthusiasm about the team, and observes that both statements can be sincere. Stated and revealed preferences diverge, and an instrument that asks a person to disclose a private intention to the party with an interest in preventing it will systematically under report.

Manager intuition. Managers were asked which of their reports might leave within a year. The account reports that intuition performed poorly and gives a worked case in which a manager identified none of the actual departures. The mechanisms named are all documented biases: managers over predict departure for people they like less, employees actively conceal search behaviour from the person best placed to intervene, thresholds for concern vary across managers so predictions are not comparable, and sunk cost reasoning leads managers to under predict departure among people they have invested in. This is the sharpest illustration of the account's own conclusion, which is that the value of a data driven signal here is not that it is accurate in absolute terms but that it is better than the alternative it replaces, and unlike the alternative it is uniform across managers.

Goal completion. Incomplete goals were expected to indicate disengagement. The account reports a weak relationship and identifies the confounds precisely: goals go incomplete because of reassignment, leave or legitimate priority change, and the causality also runs backwards, since a person who has decided to leave stops completing goals. That reversal is not a modelling nuisance. It means the signal is partly a consequence of the outcome, and including it in a model that appears to predict is a form of leakage.

Tenure alone. Tenure has a strong marginal relationship with attrition, at reported rates of about twenty six percent in the first two years falling to about five percent beyond ten, but it is not usable alone because it does not separate. Most people in the highest attrition tenure band stay. The account is explicit that a strong performer who is underpaid may leave at ten years of tenure regardless.

Recent promotion. Promotion was assumed to secure retention. The account reports almost no relationship and offers a mechanism worth taking seriously: promotion raises a person's external market value and visibility to recruiters at the same time as it resolves one internal concern, so it can accelerate a departure it was intended to prevent.

Taken together these six weak signals match the misconceptions catalogued in the practitioner literature closely enough to be worth noting as an independent replication in a single organisation [1]. The pattern is consistent: every weak signal is either self reported to the employer, an artefact of an annual administrative cycle, or partly caused by the outcome it is supposed to predict. Every strong signal is a behavioural trace with a short refresh interval and no reporting incentive attached to it.

11. Governance, Privacy and Duty of Care

An aggregate attrition rate is a business metric. A prediction that a named person will probably leave within ninety days is sensitive personal data about that person, and it exists only because their employer observed their behaviour. The change in obligation between those two artefacts is the central governance fact of this field, and the account treats it as such. Figure 6 renders the described boundary.

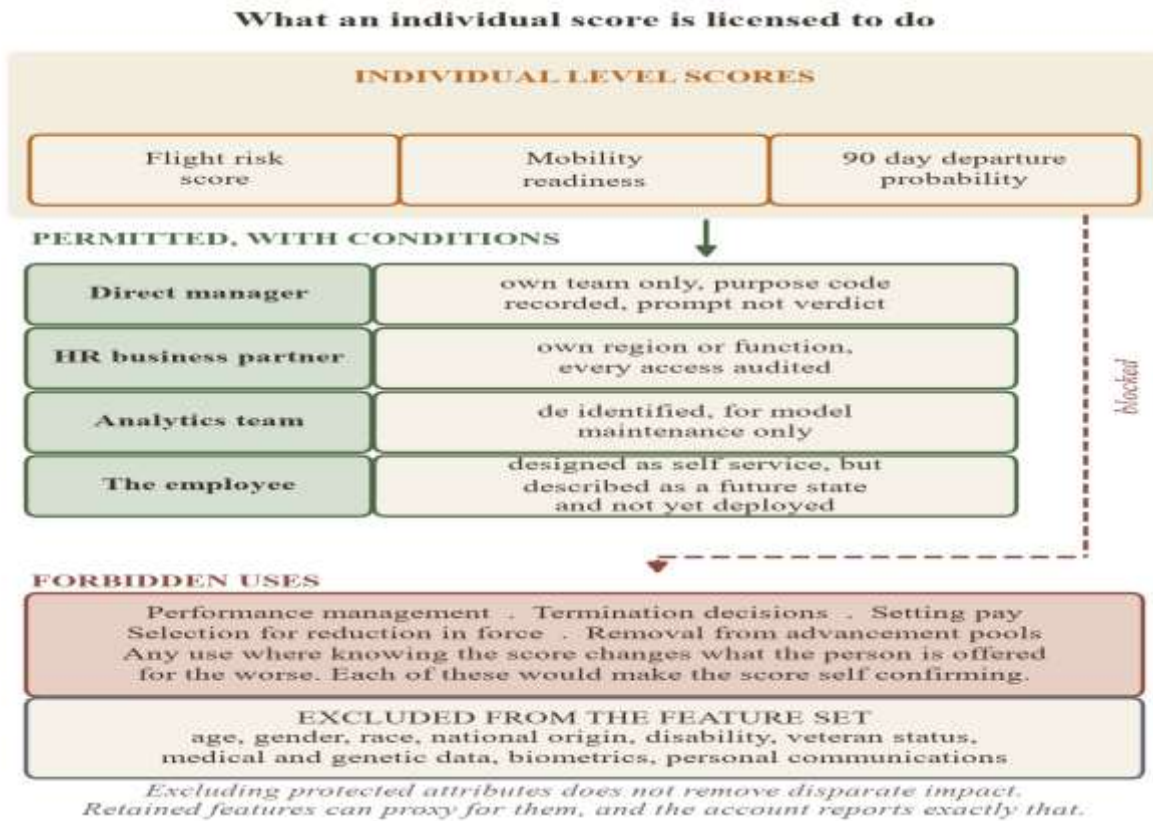


Figure 6. The described access and purpose boundary around individual level scores. Permitted consumers are listed with the conditions attached to each. Employee self service appears in the source material both as a live capability and as a future state, and is shown here as designed but not deployed, which is the more conservative reading.

Purpose limitation is the load bearing control. The rule stated is that an individual score may trigger a development conversation and may not enter performance management, termination decisions, compensation setting or selection for a reduction in force. The reasoning given for each exclusion is specific and correct. A departure forecast in a performance review conflates a prediction about the future with an assessment of past work. In a termination decision it becomes a pretext for acting first. In compensation it creates a perverse incentive to underpay people identified as likely to leave. In a reduction in force it targets people for a characteristic that is not performance. This is a stronger control than access restriction, because it constrains what may be done with the number by a person who is entitled to see it, and access control alone cannot do that.

Access control is layered and audited. Executives see aggregates without identities. Managers see their own team under an explicit governance rule and do not see compensation. Business partners see their own region or function, with every access written to an audit log recording who accessed what, when, under which purpose code, and what action followed. The analytics team works de identified. The audit runs quarterly and the reported result is instructive: of about 1,247 accesses in a quarter, eighty nine percent were followed by a documented coaching conversation and eleven percent had no follow up action and were flagged for review. Reporting the eleven percent is what makes that audit credible. An audit that always passes is not an audit.

The bias finding is the most important passage in the account. Protected attributes are excluded from the feature set. Age, gender, race, national origin, disability and veteran status are all withheld, on the twin grounds that models perform comparably without them and that including them creates legal exposure. A disparate impact test was then run, and the model was found to predict higher attrition for women by a reported eight point two percentage margin. The root cause analysis attributed this to a real pay difference rather than to a modelling artefact: women in the strong performer cohort had larger gaps between pay and market midpoint, and stratified figures show attrition among women falling from about twenty eight percent below a zero point eight five compensation ratio to about eleven percent above one point zero five, with men following the same gradient at consistently lower levels.

Three things follow and only the first is drawn out in the source material. First, the correct response is to fix the pay gap rather than to adjust the model, since the model is reporting something true. Second, and not stated,

the disparity persisted at every compensation band, so pay does not fully account for it, and the source material's claim that pay explains about seventy percent of the gap is not derivable from the stratified figures supplied and is not used here. The residual within band difference of two to ten percentage points is reported instead and left as an open question. Third, this case is a clean demonstration of the central finding of the fairness literature, which is that removing protected attributes does not remove disparate impact, because retained features proxy for them and because the historical labels themselves encode the disparity. A compensation ratio is a legitimate job related feature and it is also, in this organisation, partly a gender variable. Reported practice across comparable employment systems shows the same gap between fairness claims and fairness testing [17], and the account's willingness to publish an adverse finding is unusual and creditable. Transparency is the weakest part of the described model. The intended employee facing communication is clear and well drafted, stating what the analytics is for, which data is used, which is excluded, and what it will not be used for. But employee self service appears in one passage as a present capability and in another as a future state, and the conservative reading is that individuals could not, at the time described, see their own scores. That gap matters more than it appears. Everything else in the governance model protects the employee from misuse by others; nothing in it gives the employee any purchase on a number computed about them. Workplace privacy scholarship is direct that consent obtained inside an employment relationship is weak consent, because refusal is costly and the asymmetry is structural [3], and it is precisely in that circumstance that access to one's own data does the work consent cannot. Data protection scholarship further shows that the right to an explanation of automated processing is narrower than commonly assumed, so a multinational cannot rely on statutory rights to supply what it has not built [20]. Employee privacy and consent remain unresolved boundary conditions for the field rather than solved problems [19].

The self fulfilling prediction is the risk the architecture cannot exclude. The account describes both the intended and the failure path for the same situation. In the intended path a manager receives a prompt that a person is showing development signals, prepares by reviewing recent learning and access patterns, opens a conversation about interests rather than retention, and puts genuine options on the table including support for leaving. In the failure path the same manager receives a warning that the person is a flight risk and responds by withdrawing stretch assignments, removing them from promotion consideration, reducing their responsibilities and freezing their pay increase, on the reasoning that investment in someone leaving is wasted. The person, now genuinely disengaged by that treatment, leaves, and the prediction is recorded as correct. Nothing in the data architecture distinguishes those two paths. The difference lies entirely in how the prompt is framed, what the manager is trained to do, and whether the organisation's incentives reward retention or reward accurate forecasting. This is why Section 7 insisted that a score is a governed artefact rather than a measurement. The described mitigations, framing the alert as a development prompt, logging the action taken alongside the access, and auditing accesses without follow up, are the right ones. They are also the ones most likely to erode quietly under pressure, since none of them is enforced by the system.

12. Observed Outcomes

This section reports what the account reports, with derived figures that fail their own arithmetic excluded and the exclusions named.

Attrition by cohort. The cohort figures are the most internally consistent set in the whole account, and every stated improvement reconciles against its own components. By tenure, attrition fell from twenty six to twenty two percent in the first two years of service, from sixteen to thirteen percent at two to five years, from ten to nine percent at five to ten years, and was unchanged at five percent beyond ten years. By performance rating, attrition among people rated above expectations fell from eighteen to fourteen percent, with no change for those meeting or developing. By function, sales fell from twenty two to eighteen percent, operations from thirteen to eleven, finance from nine to eight, and technology from eleven to nine. Remote workers fell from sixteen to twelve percent against on site workers falling from eleven to ten. Elsewhere in the source material the first tenure band appears as twenty four percent rather than as twenty six before and twenty two after; the values from the dedicated cohort answer are used here.

The shape of that pattern is worth more than its magnitude. Improvement concentrated where the account says intervention was targeted, namely early tenure, strong performers, the highest attrition function and remote workers, and was absent where no intervention was directed. That is a weak form of validation, since a programme producing improvement everywhere including where it did nothing would more likely be reflecting a change in the labour market than a change in practice.

Model performance. As reported: seventy nine percent training accuracy, seventy six percent test accuracy, seventy six percent precision and sixty eight percent recall, from a temporal split training on 2021 to 2023 and testing on 2024, with a backtest re predicting the prior year at seventy three percent. Section 7 sets out why precision does not reconcile with the reported confusion matrix and why the accuracy figures should not be

used on a population this imbalanced. Reported feature importances are the compensation ratio at twenty one percent, access pattern change at eighteen, career interest outside the current function at sixteen, performance trajectory at fourteen, learning activity at twelve, engagement score at eleven and manager tenure with the employee at eight, which sum to one hundred exactly. Since the rule weighted score is itself an input alongside several of its own constituents, these should be read as importances within a correlated feature set and not as independent contributions.

Calibration, and what kind of calibration it is. A six month monthly series reports predicted against actual attrition of 1.15 against 1.12 percent, 1.18 against 1.21, 1.22 against 1.18, 1.25 against 1.32, 1.28 against 1.35 and 1.32 against 1.28, averaging 1.23 predicted against 1.24 actual. Every figure in that series reconciles with its own components. This is genuine evidence of one specific thing: the aggregate monthly rate forecast is well calibrated. It is not evidence that individual probabilities are calibrated, which is a different claim requiring the observed departure rate within each predicted probability band. The account describes checking exactly that monthly and does not report the result. The distinction matters because the aggregate can be well calibrated while individual scores are badly ordered, and it is the individual score that reaches a manager.

Retention of strong performers. Reported as 755 departures from 4,200 strong performers before, an eighteen percent rate, against 370 from 4,100 afterwards. Both counts reconcile with their stated rates and retained totals. The comparison is not clean, because the later figure is labelled as spanning two years while the earlier is annual, and because the cohort table reports post state attrition among strong performers as fourteen percent rather than the nine percent implied here. The direction is consistent across every version; the magnitude is not established and no single figure is asserted.

Financial outcomes. These are the least reliable figures in the account and are reported as such. Five mutually inconsistent statements of total value and return appear across the source material, and only one set is internally consistent: annual value of about sixty three point seven million dollars in the fourth year, composed of about thirty point seven million from prevented regretted departures at one hundred twenty thousand dollars each, nineteen point five million from internal placements at eighty five thousand dollars each, three point five million from smoother hiring, and ten million in estimated other benefits, against an annual cost of about five hundred fifty thousand dollars for platform and staff. Each component reconciles arithmetically with the counts and unit costs stated for it. That does not make them measured. The cost of one hundred twenty thousand dollars per departure is a stated average within a range of seventy five to one hundred fifty thousand, the counterfactual number of prevented departures is an inference from a classification made by people inside the programme, and the ten million dollars in other benefits is explicitly estimated. A competing statement of annual value at eighty point five million dollars is excluded because one of its three components appears to be a retention figure from a different year mislabelled as a planning benefit. No return multiple is asserted anywhere in this paper, because the source material supplies five and they cannot all be right.

13. Limitations

This is a single enterprise account. Every figure in it is the author's own operational reporting, drawn from internal systems and dashboards. No part of it was independently instrumented, audited or replicated, and no controlled evaluation was performed. Prediction quality is reported as described rather than independently validated, and one reported metric, precision, does not reconcile with the confusion matrix reported alongside it. Transferability to other organisations is argued from mechanism, not demonstrated.

Several specific limits follow from the material itself. Coverage is incomplete: contingent workers are at roughly half coverage and several countries are sparse, so the population scored is narrower than the population employed. Usable history is three to four years, which is thin for individual level prediction and, by the account's own stated threshold, close to the line at which individual scoring should be withheld. The features carrying the most reported predictive weight have the least historical depth behind them. The regretted attrition classification, which carries most of the outcome claim, is a human judgement made by people with an interest in the answer. The intervention outcome comparison is between chosen rather than assigned interventions and the cohort sizes are not supplied, so the ordering is confounded with selection. The account of model drift is presented in the source material as a hypothetical timeline, so no drift episode is claimed as observed here, and the lag adjustment comparison is described through generic organisations rather than as an instrumented experiment.

The governance limitations should be read as requirements rather than as achievements. Purpose limitation, audit logging and the bias audit are described as implemented. Employee facing self service, the mechanism that would give an individual any purchase on a score computed about them, appears in the source material both as live and as a future state, and this paper takes the conservative reading that it was not deployed. The described annual third party audit and quarterly employee trust survey are reported with target thresholds and no results, so they should be understood as an intended control regime rather than as evidence that the regime

worked. Most importantly, no part of the architecture can rule out the self fulfilling prediction described in Section 11. Establishing that a score does not help cause the departures it forecasts would require withholding intervention from a random subset of flagged employees, a design with an obvious ethical cost that was not undertaken and whose absence should be assumed rather than excused.

14. Conclusion

The engineering claim of this paper is that attrition prediction in a large enterprise is decided beneath the model. Six systems that were never designed to be joined must be reconciled to one identity, one status and one timeline before any feature means what it appears to mean, and the failures that matter most are attribution failures rather than accuracy failures. A record that might belong to two people cannot be scored at all.

The analytical claim is that the target is chosen badly by default. Total attrition is stable, largely unavoidable and mostly not worth preventing. The figures reported here show it moving by well under a percentage point across four years while the regretted share of departures fell from twenty two to eight percent. An organisation managing the headline rate would have seen nothing happen.

The operational claim is that prediction becomes useful only when it is composed into a capacity forecast, and that most of a capacity gap in any quarter is promotion, leave and new demand rather than attrition. This bounds how much better departure prediction can make planning, and it argues for building the capacity layer even where the departure model is weak.

The practitioner claim is that the strong signals are behavioural traces with short refresh intervals and no reporting incentive, while the weak ones are self reported to the employer, tied to an annual administrative cycle, or partly caused by the outcome they are meant to predict. That single distinction accounts for the whole ordering in Section 10, and it is the most portable finding here. The intervention that follows from it is mobility rather than persuasion, which is what the embeddedness literature predicts and what the reported spread from thirty eight to seventy four percent is consistent with.

The governance claim is the one to take away. The moment a system attaches a number to a named person, the obligation changes, and no amount of model quality discharges it. The controls that matter are purpose limitation, which constrains what an authorised viewer may do; access logging with recorded purpose and action; and adverse finding disclosure, of which the pay equity result reported here is a good example. Missing from the account is the control that gives the individual any standing: sight of their own score. An architecture can enforce the first three. It cannot enforce the framing that decides whether a prompt supports a person or forecloses them, and that framing is what determines whether the prediction helps or comes true.

References

- [1] D. G. Allen, P. C. Bryant, and J. M. Vardaman, "Retaining talent: replacing misconceptions with evidence-based strategies," *Academy of Management Perspectives*, vol. 24, no. 2, pp. 48 to 64, 2010. doi:10.5465/amp.2010.51827775
- [2] D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence, and M. Stuart, "HR and analytics: why HR is set to fail the big data challenge," *Human Resource Management Journal*, vol. 26, no. 1, pp. 1 to 11, 2016. doi:10.1111/1748-8583.12090
- [3] D. P. Bhawe, L. H. Teo, and R. S. Dalal, "Privacy at work: a review and a research agenda for a contested terrain," *Journal of Management*, vol. 46, no. 1, pp. 127 to 164, 2019. doi:10.1177/0149206319878254
- [4] M. Bidwell, "Paying more to get less: the effects of external hiring versus internal mobility," *Administrative Science Quarterly*, vol. 56, no. 3, pp. 369 to 407, 2011. doi:10.1177/0001839211433562
- [5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321 to 357, 2002. doi:10.1613/jair.953
- [6] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785 to 794, 2016. doi:10.1145/2939672.2939785
- [7] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 34, no. 2, pp. 187 to 202, 1972. doi:10.1111/j.2517-6161.1972.tb00899.x
- [8] D. R. Dalton, W. D. Todor, and D. M. Krackhardt, "Turnover overstated: the functional taxonomy," *Academy of Management Review*, vol. 7, no. 1, p. 117, 1982. doi:10.2307/257256
- [9] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1 to 37, 2014. doi:10.1145/2523813

10. [10] R. W. Griffeth, P. W. Hom, and S. Gaertner, "A meta-analysis of antecedents and correlates of employee turnover: update, moderator tests, and research implications for the next millennium," *Journal of Management*, vol. 26, no. 3, pp. 463 to 488, 2000. doi:10.1177/014920630002600305
11. [11] P. W. Hom, T. W. Lee, J. D. Shaw, and J. P. Hausknecht, "One hundred years of employee turnover theory and research," *Journal of Applied Psychology*, vol. 102, no. 3, pp. 530 to 545, 2017. doi:10.1037/apl0000103
12. [12] T. W. Lee and T. R. Mitchell, "An alternative approach: the unfolding model of voluntary employee turnover," *Academy of Management Review*, vol. 19, no. 1, p. 51, 1994. doi:10.2307/258835
13. [13] J. H. Marler and J. W. Boudreau, "An evidence-based review of HR Analytics," *The International Journal of Human Resource Management*, vol. 28, no. 1, pp. 3 to 26, 2017. doi:10.1080/09585192.2016.1244699
14. [14] D. B. Minbaeva, "Building credible human capital analytics for organizational competitive advantage," *Human Resource Management*, vol. 57, no. 3, pp. 701 to 713, 2017. doi:10.1002/hrm.21848
15. [15] T. R. Mitchell, B. C. Holtom, T. W. Lee, C. J. Sablinski, and M. Erez, "Why people stay: using job embeddedness to predict voluntary turnover," *Academy of Management Journal*, vol. 44, no. 6, pp. 1102 to 1121, 2001. doi:10.2307/3069391
16. [16] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625 to 632, 2005. doi:10.1145/1102351.1102430
17. [17] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: evaluating claims and practices," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 469 to 481, 2020. doi:10.1145/3351095.3372828
18. [18] T. Rasmussen and D. Ulrich, "Learning from practice: how HR analytics avoids being a management fad," *Organizational Dynamics*, vol. 44, no. 3, pp. 236 to 242, 2015. doi:10.1016/j.orgdyn.2015.05.008
19. [19] A. Tursunbayeva, S. Di Lauro, and C. Pagliari, "People analytics: a scoping review of conceptual boundaries and value propositions," *International Journal of Information Management*, vol. 43, pp. 224 to 247, 2018. doi:10.1016/j.ijinfomgt.2018.08.002
20. [20] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76 to 99, 2017. doi:10.1093/idpl/ix005