Predicting Employee Attrition

Kyle Rodriguez

Problem Identification Overview

High Cost of Turnover

Replacing an employee costs 1.5-2x their salary on average

Lost Productivity

Two thirds of turnover costs are intangible like lost knowledge and internal strain

Identify At-Risk Employees

Predicting turnover could minimize costs and increase retention

Proactively identifying employees likely to resign would allow companies to reduce turnover costs and retain talent.

Data and Method

This synthetic dataset contains 18 months of daily performance and attrition data for a factory with 508 workers.

Due to employee turnover, there were a total of 687 persons in the dataset. The dataset had around 412,000 rows and 42 distinct features.

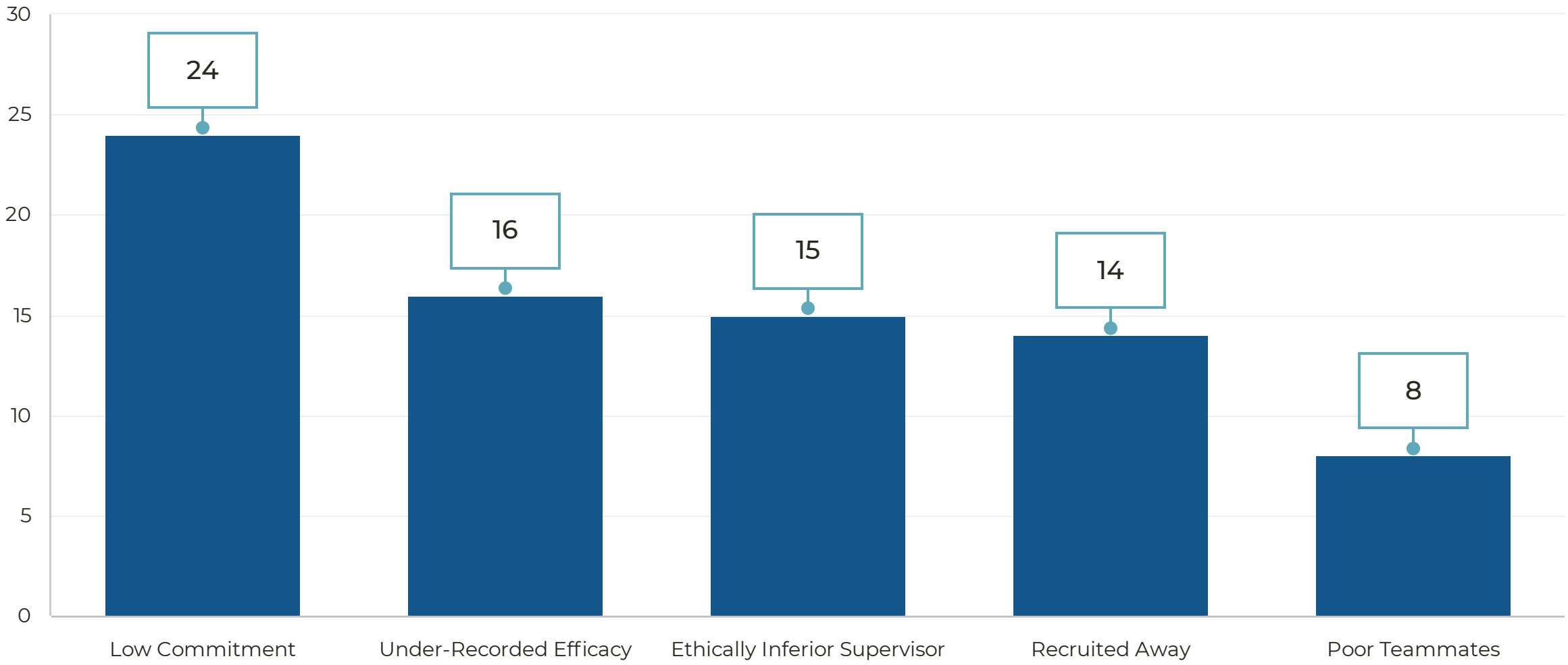
sub_ID	sub_age	sub_health_h	sub_commitment_h	sub_perceptiveness_h	sub_dexterity_h	sub_sociality_h	sub_goodness_h
3000001	40	0.895	1.0	0.659	0.592	0.799	0.501
3000001	40	0.895	1.0	0.659	0.592	0.799	0.501
3000001	40	0.895	1.0	0.659	0.592	0.799	0.501
3000001	40	0.895	1.0	0.659	0.592	0.799	0.501
3000001	40	0.895	1.0	0.659	0.592	0.799	0.501

EDA: Employee Hidden Traits

This dataset contained hidden trait values on a range from 0 - 1 for both employee and supervisors.

The traits influence the kind of events an employee/supervisor might perform. Of the employees that resigned, the average range their values fell within was between 0.7 - 0.9.

EDA: Reasons for Employee Resignations



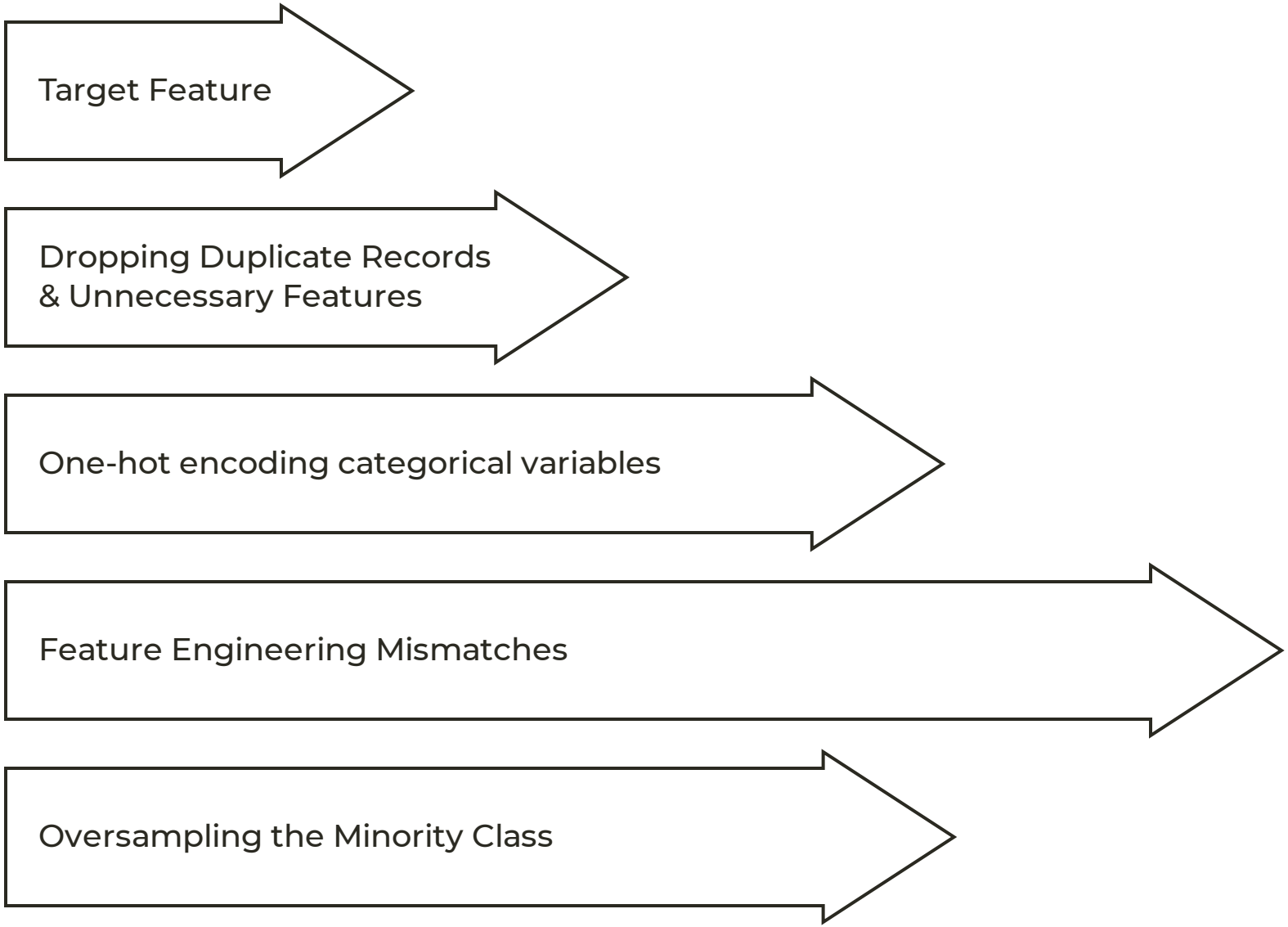
EDA: Mismatched Event Records

In total, there were 3,938 instances where a given event was recorded as something different by the supervisor than what actually took place - especially positive events.

Events	Recorded Differently by Supervisor	Actual Event	Difference
Efficacy	191,272	191,657	-385
Feat	3,387	3,937	-550
Idea	3,334	3,899	-565
Sacrifice	4,605	5,356	-751
Teamwork	4,090	4,814	-724

These mismatched events could be a source of frustration felt by employees, ultimately contributing to their resignations.

Data Preprocessing Steps of Note



Model Description

- ## Binary classification problem

  Used logistic regression for the first model due to the binary nature of predicting employee resignation.

- ## Model types

  Random Forest and Catboost models were chosen in addition to logistic regression.

- ## Feature engineering

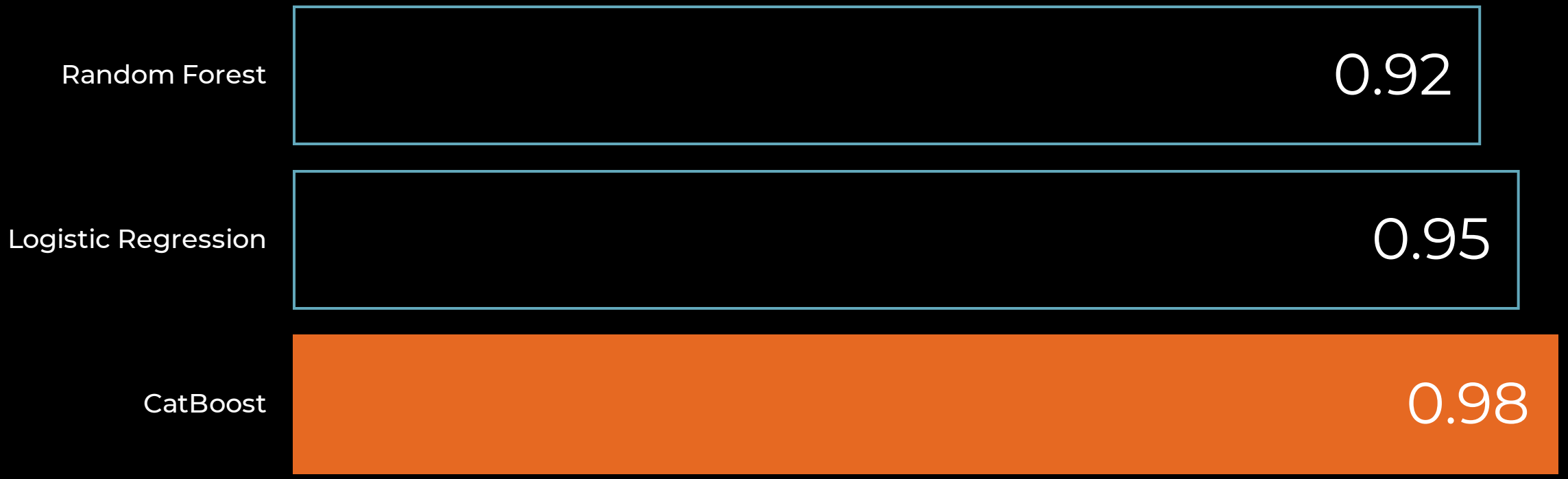
  Identified feature matrix X and target y from the dataset.

- ## Train test split

  Split data into training and test sets with 30% test size.

Model Performance

Accuracy scores of models on oversampled dataset



Model Performance

AUC scores of models on oversampled dataset

Model Findings

Important Features Ranked for the **Logistic Regression Model**

Feature	Importance
Being on Team 17	1
Being on Team 9	2
Supervisor with low perceptiveness score	3

Model Findings

Important Features Ranked for the **Random Forest Model**

Feature	Importance
Number of Under-recorded Efficacy Events	1
Number of Mismatched Events	2
Supervisor with low Goodness score	3

Model Findings

Important Features Ranked for the **CatBoost Model**

Feature	Importance
Number of Under-recorded Efficacy Events	1
Supervisor with Low Goodness Score	2
Supervisor with Low Commitment Score	3

Business Recommendations

1 Re-evaluate how metrics are recorded by Supervisors, aimed at minimizing mismatches.

2 Incorporate behavioral-based interview questions when interviewing managerial candidates to assess candidate's perceptiveness, goodness and commitment.

Taking an aggregate of the findings across the three models, the pain points that can be addressed include: Minimizing Event Mismatches by supervisors, Assessing Supervisor Perceptiveness, Goodness and Commitment.

Next Steps

Tuning

Tune model hyperparameters to optimize performance metrics

Hidden Traits

Incorporate proxy data for unrealistic hidden traits

Incorporate More Data

Add features like commute time and salary