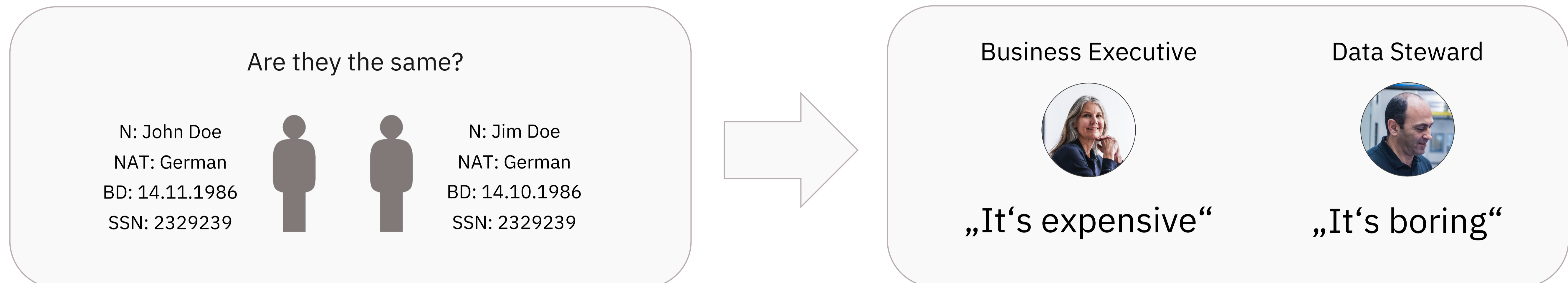


Machine Learning in Master Data Management Systems

Clerical Task Resolution



Training Data

Task resolution history with contains decision of data stewards and comparison data of the two suspected duplicates.

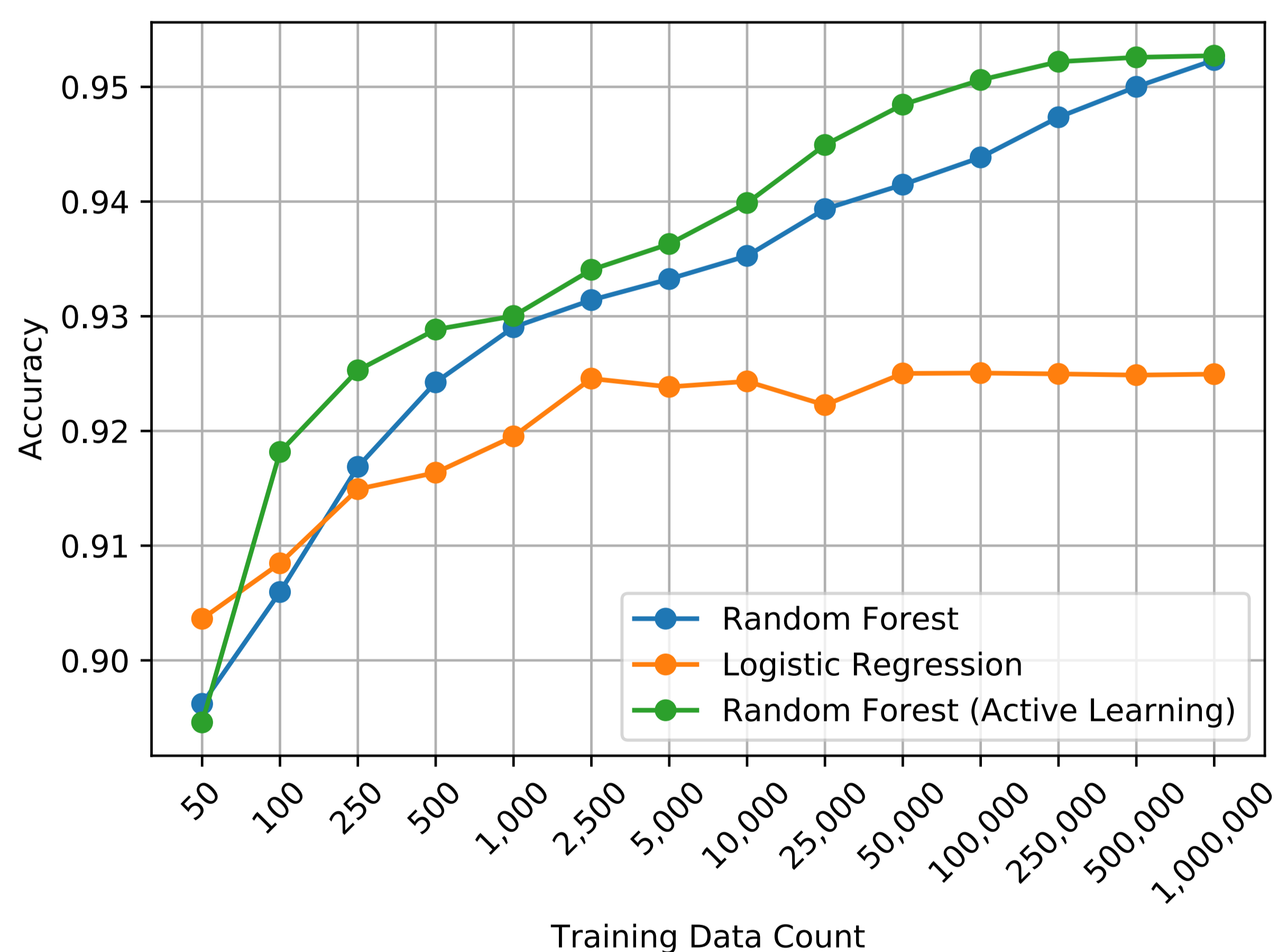
```
MEMRECNO, MEMRECNO2, CAUDTIME, RULETYPE, XNM, AXP, SSN, DOB, SEX, FPF2, OVERALL_CMPSCORE
29955364, 45928598, 2015-01-02 08:07:44, S, +0.66, +0.13, +0.00, +4.47, +0.26, -3.00, 2.5
33087603, 45928598, 2015-01-02 08:07:44, S, +0.66, +0.13, +0.00, +4.47, +0.26, -3.00, 2.5
46274721, 46331036, 2015-01-02 08:10:07, S, +8.27, +4.71, +5.01, +4.55, +0.26, +0.00, 22.8
30214332, 46331062, 2015-01-02 08:10:07, S, +8.27, +4.71, +0.00, +4.55, +0.26, -2.00, 15.7
46220762, 46315567, 2015-01-02 09:35:55, D, +8.07, +4.71, +0.00, +4.45, +0.35, -6.00, 11.5
25754083, 46264503, 2015-01-02 15:32:23, D, +2.28, +1.33, +0.00, +4.53, +0.35, -3.00, 5.4
25754083, 46262360, 2015-01-02 15:32:23, S, +8.27, +1.33, +0.00, +4.53, +0.35, -2.00, 12.4
```

Data Pre-Processing

Data is skewed, we evaluated different sampling methods to balance the data. Random Oversampling showed the best results.

Zero values describe a situation where the attribute is missing. To compensate we introduced artificial features indicating if the comparison value is 0.

Active Learning

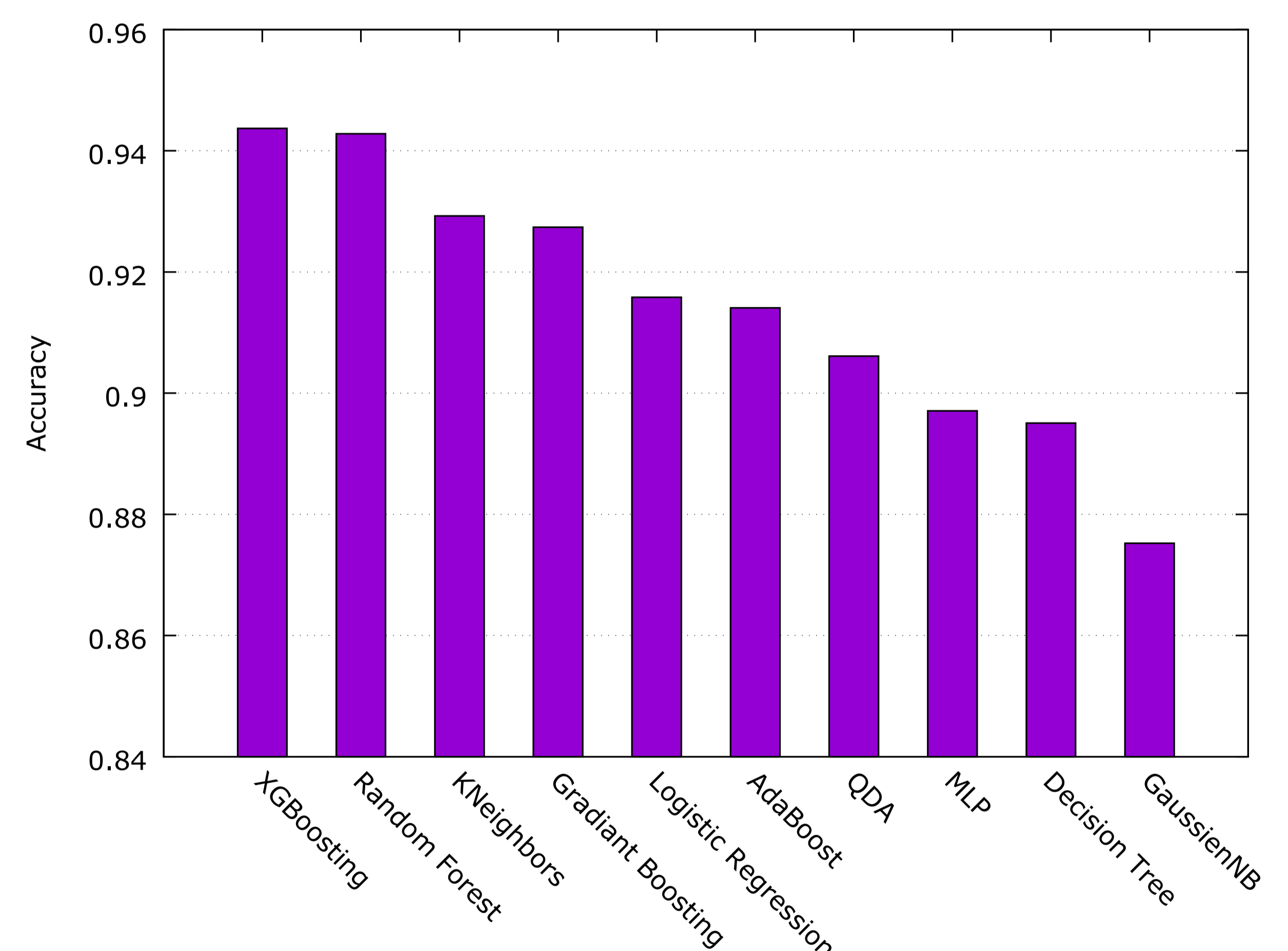


Two Step process:

1. Clustering: Using k-means to identify first 10 tasks to process by data stewards.
2. Active Learning: Actively suggesting the next 10 tasks with most information gain to process by data stewards.

Active learning lifts accuracy to about 92.5% with only 250 resolved tasks. At that point plain Random Forest has a accuracy of 91.7%.

Results



When preparing the input data with oversampling (Random Oversampling) and artificial features the prediction quality is

- Accuracy = 0.94
- Precision = 0.98 (same), 0.80 (different)
- Recall = 0.94 (same), 0.91 (different)

We showed that the ML approach works better than a highly tuned Matching Engine.