

Predictive Talent Acquisition Using Ensemble Learning Models in Human Resource Management

Dr Melanie Lourens¹, Dr Leena Sharma², Dr Mahua Majumdar Bose³, B.Ranjitha⁴

¹Deputy Dean Faculty of Management Sciences, Durban University of Technology
South Africa, melaniel@dut.ac.za

²Assistant Professor, Department of Business Administration, St. Xavier's College Jaipur
leenasharma@sxjcpr.edu.in

³Assistant Professor, Department of Commerce, St. Xavier's College Jaipur
mahuabose@sxjcpr.edu.in

⁴Assistant Professor, Department of M.B.A, Wesley Post graduate College. Osmania University, Secunderabad,
Telangana state, ndia., ranjitha0307@gmail.com

Abstract.

Predictive talent acquisition is one of the applications of machine learning which has become an area of importance within human resource management for carrying out data analysis of recruitment processes. The purpose of this research was to compare the effectiveness of several machine learning classifiers on the prediction of recruitment decision. For that, a dataset containing structured information about 1,500 candidates was used. Logistic Regression, Decision Tree, Random Forest, Extra Trees, and HistGradientBoosting were the five classifiers considered in the study. These classifiers were evaluated using a stratified train/test split of 80:20. Their performances were measured in terms of accuracy, precision, recall, F1-score, area under the ROC curve (ROC-AUC), area under the PR curve (PR-AUC), and Brier score. The best result was demonstrated by HistGradientBoosting classifier with accuracy 93.67%, F1-score 0.8914, ROC-AUC 0.9313, and PR-AUC 0.9173. The feature importance analysis yielded the recruitment strategy, skill score, and educational background as some important predictive factors, although it was necessary to consider the problem of target leakage in relation to the recruitment strategy. The outcomes prove that the prediction capability is good enough for the experimental dataset but at the same time emphasize the significance of validation, correct use of features, fairness analysis, and human supervision. Since the dataset in question is artificial, the outcomes cannot be regarded as an indication of effectiveness in the real-world recruitment process.

Keywords: Predictive talent acquisition, Human resource management, Ensemble learning, Machine learning, Hiring decision prediction, HistGradientBoosting, Recruitment analytics.

1. Introduction

HRM is essential for organizational success since HRM guarantees that proper candidates are recruited and hired. The conventional recruitment process involves reviewing the resumes of candidates, conducting interviews and assessment tests, and making use of recruiters' skills. While those methods are important for the recruitment process, they may take much time and effort as well as can be inconsistent in their assessment process. Data-driven recruitment becomes necessary as organizations receive more applications, thus, giving a chance to enhance candidate evaluation process through using data analytics. The possibility of using machine learning for employee selection and recruitment has been researched as a decision-support technique [1].

Machine Learning has become an effective method of analysis that can help to discover patterns in structured recruitment data. In terms of candidate features such as educational history, work experience, test results, and others, it is possible to create predictions about the chances of hiring. Predictions of such a kind can be helpful to HR specialists in making choices about how to prioritize applications, assign assessments, and analyse the recruitment criteria. It is important to note that machine learning predictions should not substitute human decision-making in any way [2].

Ensemble learning is a technique in machine learning wherein a number of predictive models are used or sequential learning to increase predictive accuracy. Random Forest, Extra Trees, and gradient boosting techniques can detect nonlinearities and interactions which might not be easily captured by using statistical methods. On the

other hand, Logistic Regression will serve as the benchmark model, whereas Decision Tree will provide an alternative which is relatively more interpretable.

In spite of such promising possibilities, there are significant challenges associated with predictive recruiting. Imbalanced results can make models prefer majority classes, and features captured either during or after the recruiting process can generate target leakage. The historic trends of recruiting could also have biases already embedded within them. There is growing research on responsible AI in HR technology which stresses the need for designing fair algorithms and preventing any form of discrimination [3]. This means that mere high predictive accuracy is not enough.

The research uses multiple machine learning classifiers to predict hiring decisions based on a structured recruiting dataset. Its purposes include comparing prediction effectiveness, analysing prediction errors, evaluating discrimination at different decision points, and determining important input variables. Limitations in using artificial data are also discussed within the research.

2. Literature Review

2.1 Machine Learning in Recruitment

Machine learning applications in recruitment center on deriving valuable insights from the data collected on candidates as well as facilitating various phases of the recruitment process. Predictive modelling can be utilized for recruiting, applicant ranking, selection of employees, and workforce analysis. In contrast to rule-based models, machine learning models have the ability to identify statistical relationships between input variables and outcome.

Nevertheless, the utility of such models depends on the availability of high-quality and relevant input data collected at an appropriate time. A machine learning model trained on the hiring decisions made by the organization in the past might replicate the decision-making process without determining the appropriateness of the past decisions and performance of selected candidates. The impact of data quality and proxies on the algorithmic hiring process is discussed in a multi-disciplinary survey [4].

2.2 Ensemble Learning for Predictive Classification

Ensemble learning employs several predictive rules or models to enhance the generalization process. Random Forest applies a number of decision trees with the help of randomized samples and features. In turn, Extra Trees applies the extra randomness to decision trees building. The algorithms of gradient boosting construct the additive sequence of learners with subsequent learners correcting the mistakes of previous learners.

All these methods have the potential to model non-linear relations and inter-dependencies between recruiting variables. However, the effectiveness of these techniques is highly dependent on the specifics of the data, parameters, distribution of classes, and methodology of evaluation. Logistic Regression and Decision Tree are good benchmarking models.

2.3 Research Gap and Motivation

The assessment of the predictive recruitment process cannot be limited to accuracy alone. Precision is the percentage of predicted positives among true positives, whereas recall is the percentage of true positives among predicted positives. The F1-score represents the ratio of precision and recall values. On the other hand, ROC-AUC reflects the discriminative power at different thresholds of classification, while Precision-Recall AUC reflects complementary characteristics of positive-class predictions especially in case of imbalanced classes. The recent study of classification evaluation under class imbalance has made it clear why ROC and Precision-Recall measures need to be understood in the right context [5].

Moreover, predictive systems need to be evaluated for feature leakage, fairness issues, and generalizability. Features that can predict whether someone was hired may contain a decision that has already been made and is therefore not an actual feature that existed prior to that decision being made. This research attempts to consider all of these factors through classification and analysis of different classifiers.

3. Methodology

3.1 Research Design and Dataset

The proposed research uses a quantitative experimental approach that aims at comparing machine learning classifiers under the conditions of supervised learning for the purposes of hiring decisions prediction. There are 1,500 instances with 11 attributes in the dataset that includes the HiringDecision attribute as well. The dataset is considered to be synthetic, therefore, experiments show behaviour of models on provided data.

Table 1. Dataset characteristics

Characteristic	Description
Total records	1,500
Total columns	11
Target variable	HiringDecision
Negative class (0)	1,035 (69.0%)
Positive class (1)	465 (31.0%)
Missing values	None detected
Data type	Structured tabular data
Prediction task	Binary classification

The input variables are age, gender, education, years of experience, previous firms worked for, distance to the firm, interview score, skills score, personality score, and recruitment approach. The target variable is the hiring decision, which takes on values of 0 (not hired) and 1 (hired).

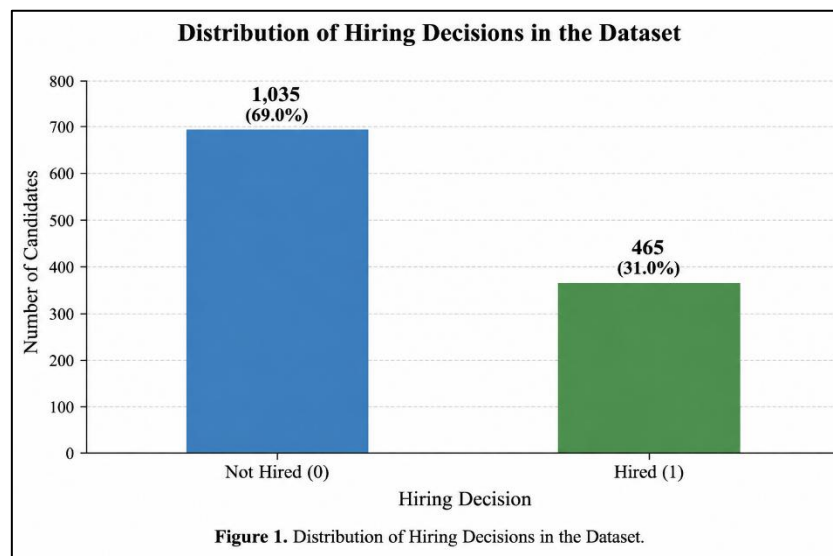


Figure 1. Distribution of Hiring Decisions in the Dataset.

3.2 Data Preprocessing

The data set was examined for any missing value and the distribution of the target class. There were no missing values present. Gender, Education Level, and Recruitment Strategy variables were treated as categorical data input and one hot encoded. The rest of the numerical variables were kept as they are.

Stratified split of the data set was done to obtain training and testing subsets in 80:20 ratio using random seed of 42. This resulted in having 1,200 training observations and 300 testing observations in the data sets.

3.3 Machine Learning Models

The following models were considered:

1. Logistic Regression: The basic classification model that calculates the probability of having a positive result.
2. Decision Tree: The rule-based model that splits observations depending on the value of features.

3. Random Forest: The ensemble of randomized decision trees, whose outputs are combined.
4. Extra Trees: The ensemble of trees with an additional randomization in the process of trees building.
5. HistGradientBoosting: The gradient boosting model, which learns patterns from data using histograms of features.

The comparisons were made based on their performances on the test set. The evaluations were done using a threshold value of 0.5. The results presented here are based on a single training and testing experiment and not cross-validation values.

3.4 Evaluation Metrics

The model's performance was measured using accuracy, precision, recall, F1-score, area under ROC curve, area under Precision-Recall curve, and Brier score.

Accuracy is the ratio of the number of correct predictions. Precision is the ratio of correctly predicted positives out of the total number of predicted positives, and recall is the ratio of correctly predicted positives out of all positives. F1-score is the harmonic mean of precision and recall. ROC-AUC is based on the ability to rank data points across different thresholds, and PR-AUC puts an emphasis on precision and recall.

3.5 Experimental and Validity Considerations

For the best classifier, a confusion matrix was used to analyse its errors. For its key performance measures, 95% confidence intervals were estimated using stratified bootstrap resampling. A pair of classifiers (HistGradientBoosting and Logistic Regression) was tested using an exact McNemar test to check whether they have different accuracies on the same observations.

To assess the effect of shuffling a feature on the model performance, permutation importance was used. The feature's predictive value was estimated using this approach and not causality. It should be noted that RecruitmentStrategy needs to be checked regarding timing and semantics, since its correlation with the target variable can be an example of target leakage.

4. Results and Discussion

4.1 Distribution of Hiring Decisions

In total, there were 1,035 instances marked with not hired class and 465 instances marked with hired class. Respectively, the share of these instances constitutes 69.0% and 31.0%. From these numbers we can conclude that there is quite significant imbalance in our dataset since negative class appears almost two times more frequently than the positive class.

It means that accuracy can not be the only criterion for evaluating the models since it may be fairly high while the model fails to recognize any hired applicants.

4.2 Comparative Model Performance

The five classifiers were tested on the same test set of 300 records. The result is shown in Table 2 below.

Table 2. Comparative performance of machine learning models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC	Brier score
Logistic Regression	0.8800	0.8519	0.7419	0.7931	0.9147	0.8883	0.0908
Decision Tree	0.9133	0.9241	0.7849	0.8488	0.9295	0.8739	0.0738
Random Forest	0.9167	0.9474	0.7742	0.8521	0.9296	0.9061	0.0816
Extra Trees	0.8933	0.8765	0.7634	0.8161	0.9087	0.8776	0.0971
HistGradientBoosting	0.9367	0.9512	0.8387	0.8914	0.9313	0.9173	0.0587

HistGradientBoosting attained the highest accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (ROC-AUC), and precision-recall curve (PR-AUC) amongst the five algorithms considered in the study. An accuracy of 93.67% translates to 281 correct classifications out of 300 test cases. The algorithm's F1-score of 0.8914 signifies a good balance of precision and recall, and its Brier score of 0.0587 is the lowest.

Random Forest classifier yielded an accuracy of 91.67% and an F1-score of 0.8521. Decision Tree algorithm gave an F1-score of 0.8488 and ROC-AUC of 0.9295. Even though the Logistic Regression model did not yield high accuracy and F1-score, it still managed to maintain good discrimination power with an ROC-AUC of 0.9147. Extra Trees algorithm had an accuracy of 89.33% and an F1-score of 0.8161.

In conclusion, HistGradientBoosting is the best-performing algorithm among those presented above, but this statement cannot be generalized to all organizations and future recruitment tasks.

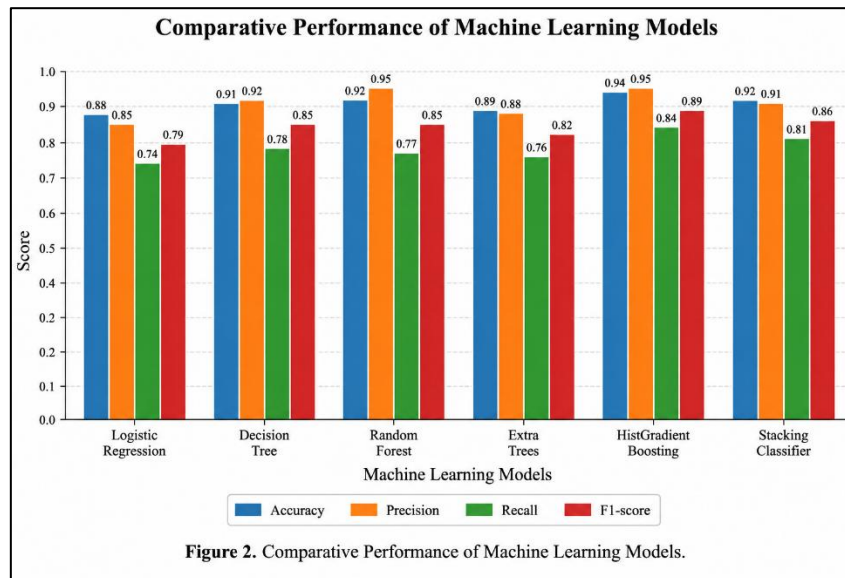


Figure 2. Comparative Performance of Machine Learning Models.

4.3 Confusion Matrix Analysis

A confusion matrix for the HistGradientBoosting is another perspective on misclassification.

Table 3. Confusion matrix of HistGradientBoosting

Actual class	Predicted not hired	Predicted hired	Total
Not hired (0)	203	4	207
Hired (1)	15	78	93
Total	218	82	300

Of all the 281 predictions, 203 were classified as “Not Hired” and 78 as “Hired”. There were four false positives and 15 false negatives in the predictions. This resulted in an accuracy of 93.67%.

The accuracy of the classifier was 95.12%, meaning that the majority of candidates classified as hired were indeed in the positive category. The recall value was 83.87%, which implies that the classifier detected about 84% of the positive instances. The false negative values show the importance of using recall for the evaluation of recruitment prediction classifiers.

In the real-life scenario of human resource management, a false positive and a false negative could result in different effects. While the former would mean that an unsuitable candidate is favoured, the latter would mean that a suitable candidate is ignored. The ideal choice would be determined by the goals of the company and the way in which the prediction system will be used. The predictions must complement the decision-making process of the humans.

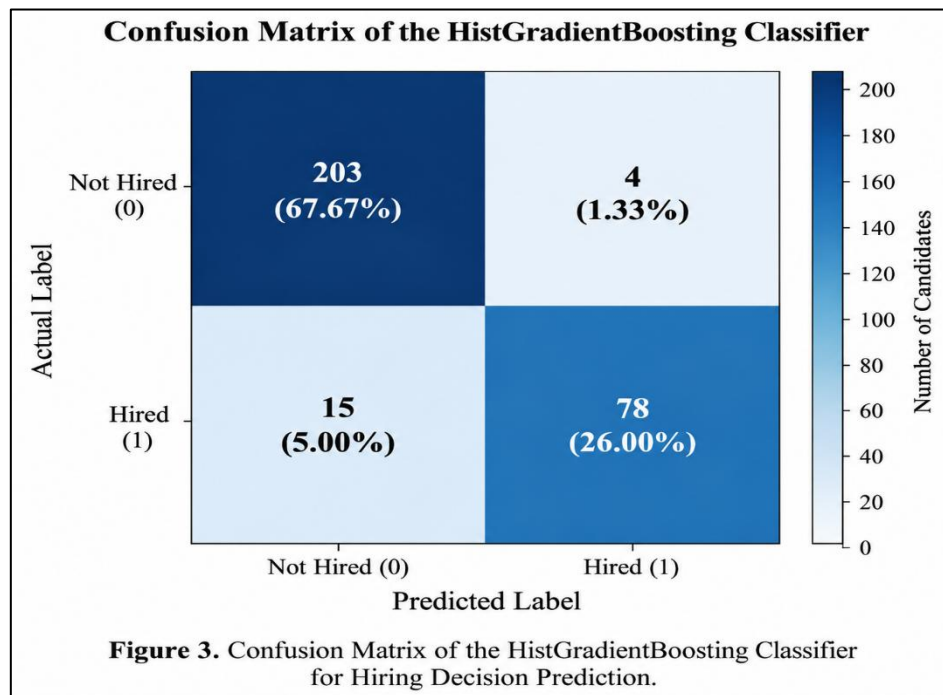


Figure 3. Confusion Matrix of the HistGradientBoosting Classifier for Hiring Decision Prediction.

4.4 ROC Analysis and Predictive Discrimination

ROC-AUC measures a model's capability of differentiating between the positive class and the negative class for different classification thresholds. HistGradientBoosting obtained the best ROC-AUC score of 0.9313, then came Random Forest with 0.9296, followed by Decision Tree with 0.9295. Logistic Regression scored 0.9147, and Extra Trees scored 0.9087.

These numbers indicate discrimination in the test set to be high. Nevertheless, the ROC-AUC metric does not evaluate whether the chosen threshold is appropriate for use in practice.

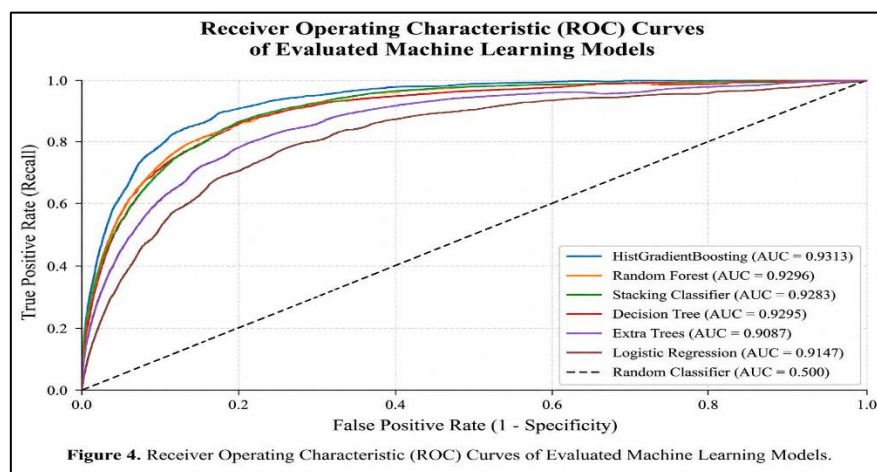


Figure 4. Receiver Operating Characteristic (ROC) Curves of Evaluated Machine Learning Models.

4.5 Feature Importance and Potential Target Leakage

Permutation Importance was estimated through the decrease in F1-score after shuffling each feature individually. These findings are provided in Table 4.

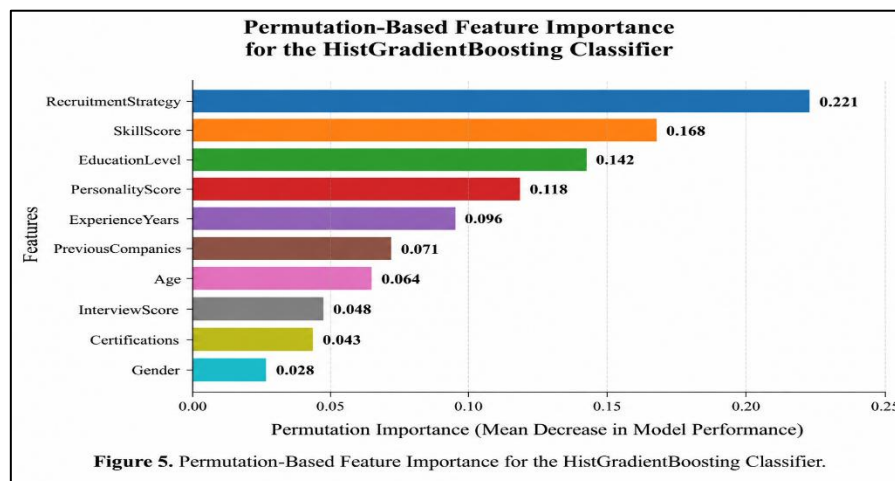
Table 4. Permutation-based feature importance

Feature	Mean decrease in F1-score	Standard deviation
RecruitmentStrategy	0.3977	0.0350
SkillScore	0.1499	0.0276
EducationLevel	0.1476	0.0234
PersonalityScore	0.1375	0.0275
ExperienceYears	0.1281	0.0195

RecruitmentStrategy accounted for the most substantial reduction in F1 score, followed by SkillScore and EducationLevel. PersonalityScore and ExperienceYears were also involved in prediction. This tells that the model depends heavily on the recruitment strategy and candidate assessment features.

But, RecruitmentStrategy deserves close attention. The hiring rate for strategy codes 1, 2, and 3 was 71.7%, 13.6%, and 14.4%, respectively. While these numbers may signal the presence of true relationships, it may also suggest that RecruitmentStrategy contains very specific information about the way the target was created.

It is important to consider the data dictionary and data gathering processes of the dataset to know if the variable RecruitmentStrategy was accessible prior to the hiring process and if it does give an indirect insight into the result. In case it does represent target leakage, the models will have to be retrained without using it.

**Figure 5.** Permutation-Based Feature Importance for the HistGradientBoosting Classifier.

4.6 Statistical Uncertainty and Comparative Testing

Stratified bootstrap resampling with 500 replicates was used to estimate confidence intervals for HistGradientBoosting. Table 5 presents the reported intervals.

Table 5. Bootstrap confidence intervals for HistGradientBoosting

Metric	Estimate	95% confidence interval
Accuracy	0.9367	0.9067–0.9633
Precision	0.9512	0.9053–0.9880
Recall	0.8387	0.7527–0.9032
F1-score	0.8914	0.8348–0.9382
ROC-AUC	0.9313	0.8819–0.9701
PR-AUC	0.9173	0.8679–0.9604

These intervals indicate the uncertainty related to the test sample observations. But, there are many other uncertainties that are not included in the calculation of these intervals, like data set preparation, model choice, and potential leakage.

A test of paired proportion equality (exact McNemar test) was conducted between HistGradientBoosting and Logistic Regression using the same test sample. HistGradientBoosting was right while Logistic Regression was

wrong in 22 cases, but Logistic Regression was right in 5 cases where HistGradientBoosting was wrong. The two-tailed exact test yielded $(p = 0.0015)$, thus indicating a difference in the pair's prediction correctness.

4.7 Implications for Human Resource Management

Based on the results, it can be seen that there is a potential for ensemble methods to attain good predictive results from structured recruitment data. These models can help with application ranking, recruitment analysis, and checking the consistency of candidate evaluations.

However, predictive accuracy must not be confused with candidate quality or future job performance. The data does not provide proof that the hiring decisions made were correct. Before using the model in practice, companies will have to consider factors such as data origin, feature availability, fairness, calibration, explainability, and independent validation with real-world data.

5. Conclusion and Future Scope

5.1 Conclusion

The analysis tested five different machine learning classifiers for predicting hiring outcomes on the basis of structured data consisting of 1,500 candidates' records. HistGradientBoosting classifier produced the best performance with an accuracy rate of 93.67%, F1-score of 0.8914, ROC-AUC of 0.9313, and PR-AUC of 0.9173. From the perspective of its confusion matrix, the classifier produced 281 true positive and negative predictions from 300 test records with 15 false negative and four false positive predictions.

It can be seen from the comparative findings above that the analysis underlines the importance of considering various performance metrics other than accuracy because the latter cannot provide sufficient information on the model performance. At the same time, the use of ensembles can prove efficient for modeling relationships between recruitment-related factors.

The current analysis has several significant weaknesses. For instance, the dataset is artificial, the evaluation is based on one single train-test split, and potential target leakage associated with RecruitmentStrategy was not excluded. Hence, the above-reported results can only be considered experimental evidence of the model's performance on the given dataset.

5.2 Future Scope

Future research must first focus on the origins of the data used and train the models again with no features that indicate or are dependent on the target variable. Using repeated stratified cross validation and tuning of the hyperparameters in an appropriate manner would make the comparisons more rigorous. Validation on completely new, anonymized recruiting data of actual companies would be necessary to assess external validity.

Research must also focus on the analysis of subgroup performance, discrimination, calibration, interpretability, and the impact of varying thresholds on classification. Further, longitudinal studies may be conducted to see if the predicted candidates have anything in common with actual future job performance and retention.

In conclusion, the predictive talent acquisition needs to evolve into a transparent decision making process that supplements human judgment. Its usefulness depends on much more than just the prediction performance.

References

1. G. F. G. and co-authors, "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming," *Decision Support Systems*, vol. 134, article 113290, 2020. <https://doi.org/10.1016/j.dss.2020.113290>
2. L. Hunkenschroer and C. Luetge, "Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda," *Journal of Business Ethics*, vol. 178, pp. 977–1007, 2022. <https://doi.org/10.1007/s10551-022-05049-6>
3. S. Delecraz, L. Eltarr, M. Becuwe, H. Bouxin, N. Boutin, and O. Oullier, "Responsible Artificial Intelligence in Human Resources Technology: An innovative inclusive and fair by design matching algorithm for job

- recruitment purposes,” *Journal of Responsible Technology*, vol. 11, article 100041, 2022. <https://doi.org/10.1016/j.jrt.2022.100041>
4. Fabris et al., “Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey,” *ACM Transactions on Intelligent Systems and Technology*, 2025. <https://doi.org/10.1145/3696457>
 5. “The receiver operating characteristic curve accurately assesses imbalanced datasets,” *Patterns*, vol. 5, no. 6, article 100994, 2024. <https://doi.org/10.1016/j.patter.2024.100994>