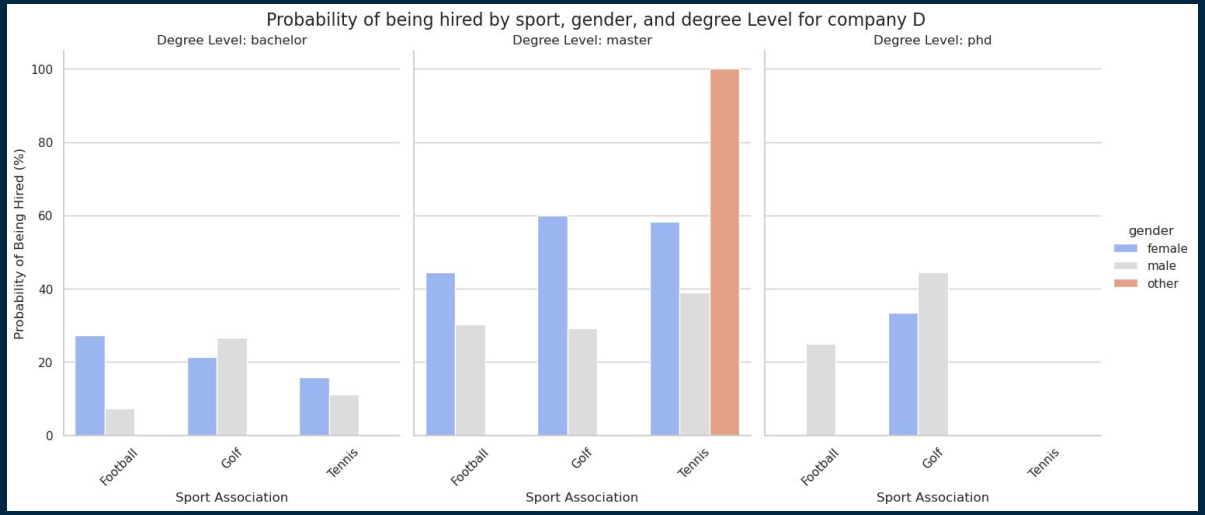
HR ADVISORY REPORT

Group 4[redacted] - [redacted] Aldric de Jacquelin [redacted]

OUR TARGET DATA



Our targeted company is company D and the targeted sports are Golf, Football and Tennis. 391

FEATURES SELECTED

We use information gain to choose five most important features and then use them for our models.

Feature Name	Information Gain Score
ind-iniversity_grade	0.312
ind-exact_study	0.148
ind-programming_exp	0.073
ind-languages	0.055
age	0.045

VALIDATION APPROACH

Test/train split with different sizes, including:

- 90% train 10% test
- 80% train 20% test
- 70% train 30% test
- 60% train 40% test
- 50% train 50% test

It was chosen to decrease the costs, since cross validation with multiple cuts requires more power.

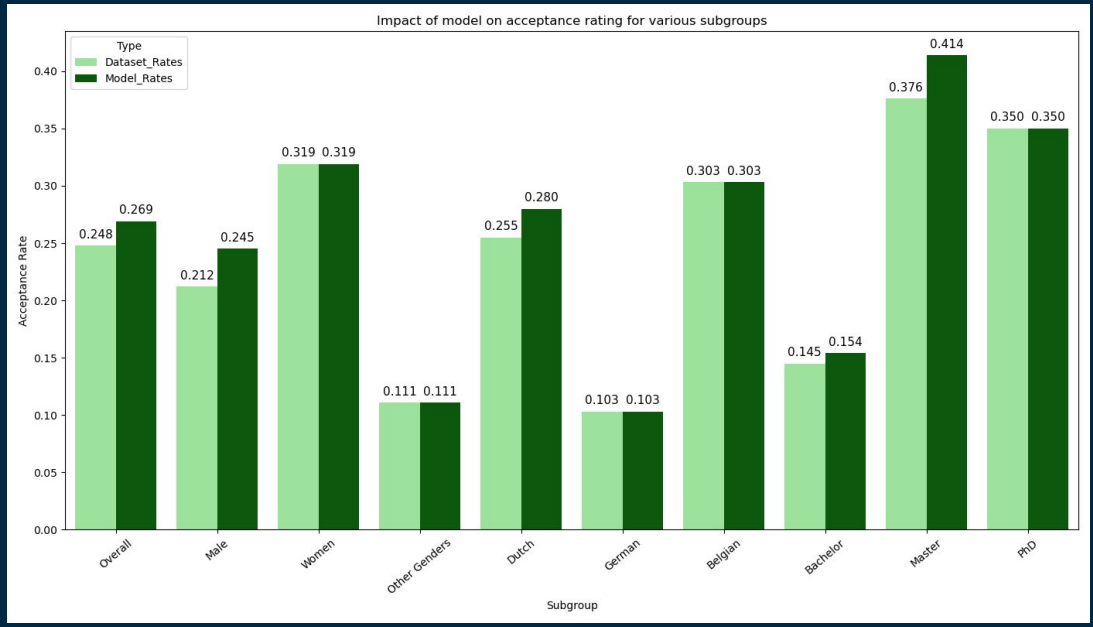
Hyperparameters: max five categories with max ten values for tweaking.

We use f1 scoring system for final comparison, since it takes into accounts false positives and false negatives.

DECISION TREE

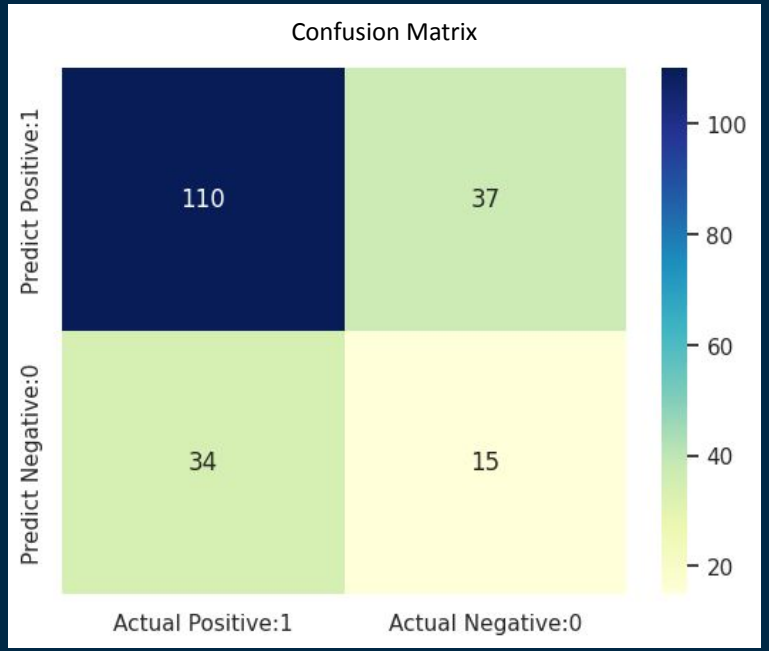
Model biases have been identified across gender, nationality, and education:

Acceptance rate rises slightly overall.
There's a potential gender bias favoring males.
Possible geographical bias leans towards Dutch candidates.
Master's degree holders are distinctly favored.
Adjustments are needed for a more equitable model.



LOGISTIC REGRESSION

- Accuracy (86%): Reflects the overall correctness of the model in classifying instances.
- Precision (92%) Represents the model's correctness when it predicts the positive class. In other words, 92% of the times the model predicted "True", it was correct.
- Recall (89%): Measures the model's capability to identify all actual positive instances. The model correctly identified 89% of all the actual "True" cases.

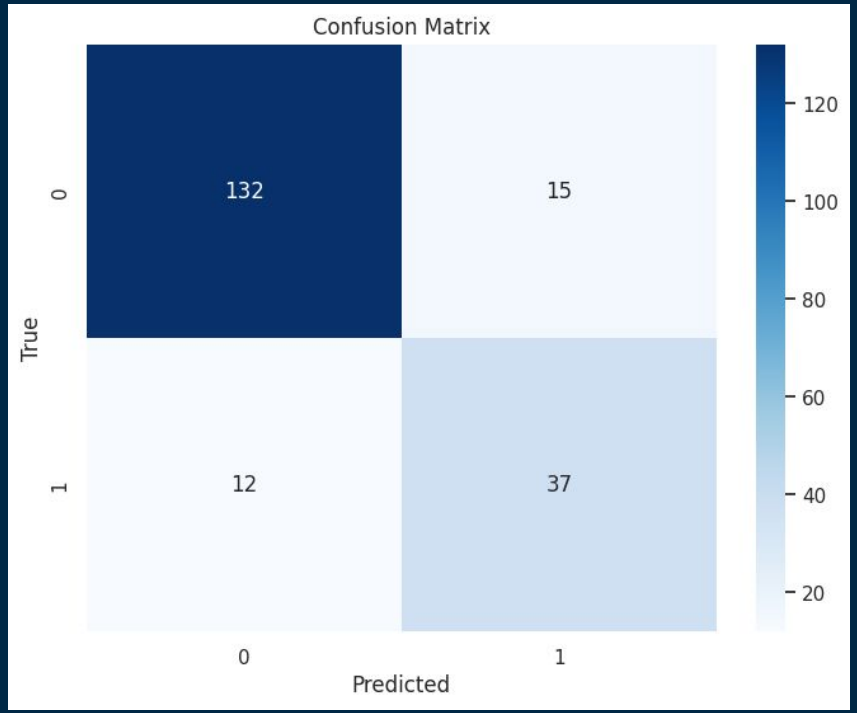


KNN

The k-NN model's performance changes as we use more data. As we add more examples, accuracy drops a bit. This is a known issue with k-NN models.

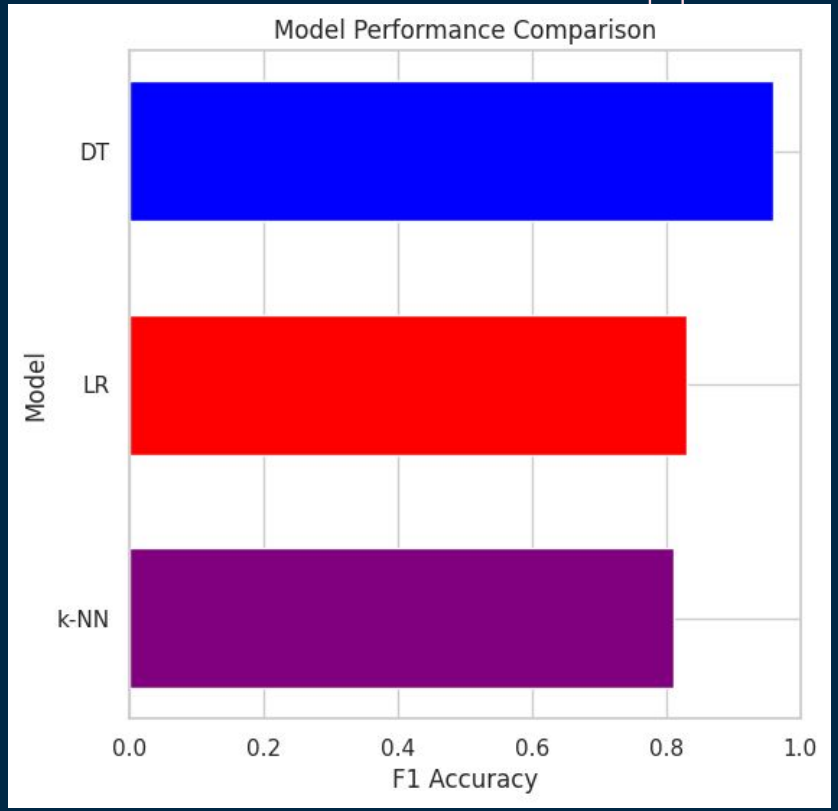
However, our model is still good at telling the difference between good and bad candidates, with a score of 0.78 out of 1.

But, as we try to find more good candidates, there's a higher chance we might pick some who aren't a good fit.



MODEL COMPARISON

Finally, the comparison of models illustrates that Decision Tree has the best results with over 90% f1-score accuracy, meanwhile both Logistic Regression and k–NN are above 80%.

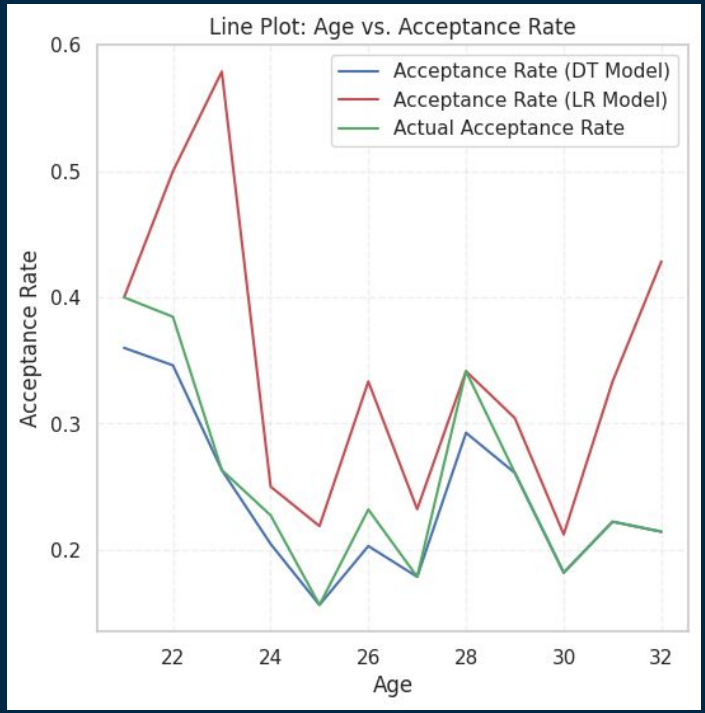


MODEL COMPARISON

The line plot illustrates how Decision Tree Model and Logistic Regression Model follow the acceptance rate trend within the Company D.

Decision Tree model is quite close to the original acceptance rate, with only slight leniency.

Logistic Regression Model has larger bias towards younger people, from 21 to 24, and older, from 30 to 32. Overall it amplifies the acceptance rates of the company.



ADVICE TO COMPANY

- Use the Decision Tree Model
- In the future do not use “Age” feature as it can introduce bias and unfairness
- Only used along with recruiter, not on its own!
- If the company’s goal is to omit any possible prejudice and bias in the algorithm, we then would advice to opt to k-NN model since it ignores the features completely
- The models do not get rid of discrimination within the company, they only imitate trends that are already there.

Do you have any questions?

THANKS