P4 HR advisory report - model comparison

D. Paplauski, P.M. Ketterman,
R.M.A van Nee, V. Li - Group 14

2023

GOALS AND OBJECTIVES

We aim to gain a deep understanding of Business B's recruitment and hiring data from the previous year, in order to create a model with which to help university graduates find their perfect job and help Business B find talent that is well suited to their needs.

We will:

- Collect, describe and explore the data
- Prepare the data for use in our decision tree model
- Model the data using training and test splits, then validate that model
- Explore the impact of this model on selection rates

UNDERSTANDING THE DATA

First, we select and describe the data, we used the Utrecht Fairness Recruitment dataset [1], and filtered it using Pandas DataFrames to retrieve data for the sports we are interested in recruiting athletes from for Business B. We then describe the data using what we retrieved.

Checking for NaN values in the data:

Id	0
gender	0
age	0
nationality	0
sport	0
ind-university_grade	0
ind-debateclub	0
ind-programming_exp	0
ind-international_exp	0
ind-entrepeneur_exp	0
ind-languages	0
ind-exact_study	0
ind-degree	0
company	0
decision	0

Checking for Null values in the data:

Id	0
gender	0
age	0
nationality	0
sport	0
ind-university_grade	0
ind-debateclub	0
ind-programming_exp	0
ind-international_exp	0
ind-entrepeneur_exp	0
ind-languages	0
ind-exact_study	0
ind-degree	0
company	0
decision	0

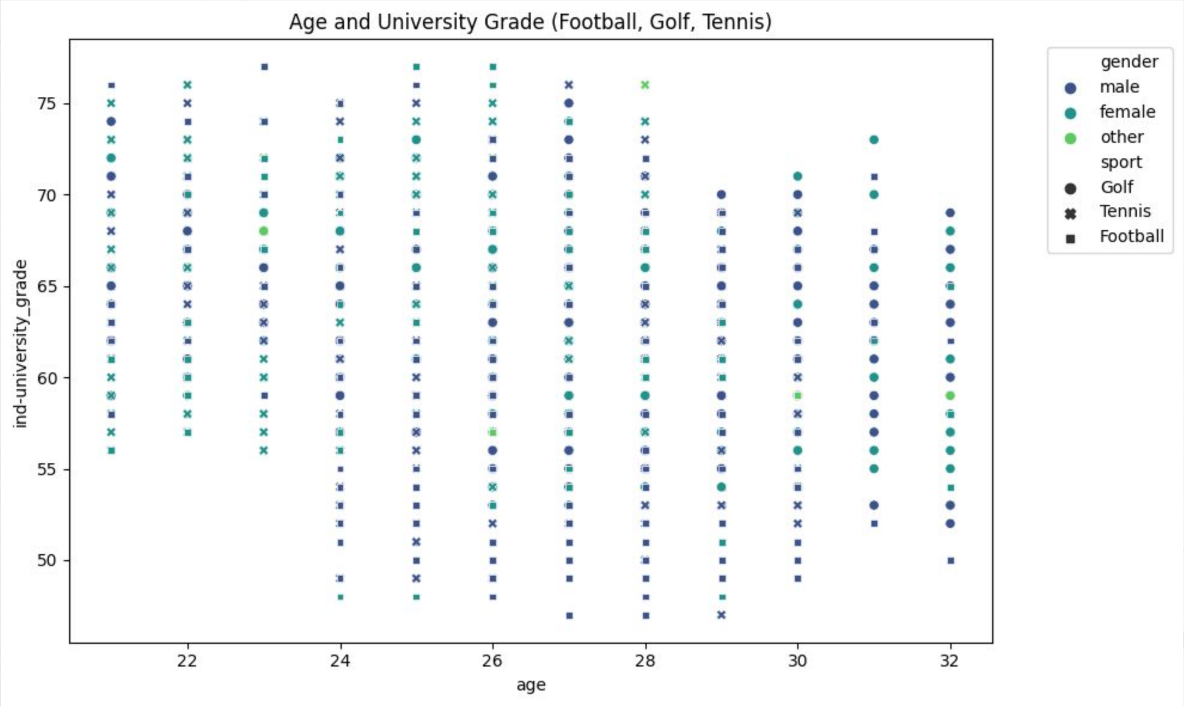
UNDERSTANDING THE DATA (2)

Next we checked the types of data in the table, and noticed we had a mix of values, some boolean, some numeric, some nominal and some ordinal. This meant that we would have to re-encode or scale some of our data so that it could be used properly in our model. For Business B, after filtering the data for our target sports: Football, Golf, and Tennis we ended up with 410 usable rows of data, with 15 columns of values, eight of those columns containing indicators. The table below demonstrates the first 10 rows described after re-encoding the boolean data, gender, and degree type. Please note that gender is not a binary encoding, we chose to use a numeric nominal representation.

Id	gender	age	nationality	sport	ind-univer	ind-debat	ind-progra	ind-intern	ind-entrep	ind-langua	ind-exact_	ind-degree	company	decision
x5572d	1	26	Dutch	Football	66	0	0	0	1	2	1	1	B	FALSE
x3158e	1	27	Belgian	Football	55	1	0	0	1	3	0	2	B	TRUE
x9413d	0	26	Dutch	Football	66	1	1	0	0	2	0	1	B	FALSE
x9400d	1	23	Dutch	Tennis	67	0	0	0	1	2	1	1	B	FALSE
x5254f	0	26	Dutch	Golf	70	0	1	0	0	1	1	2	B	FALSE
x7432e	1	24	Dutch	Football	62	1	0	1	1	3	0	2	B	TRUE
x3386e	1	32	Dutch	Golf	60	0	1	0	0	1	0	1	B	FALSE
x4103d	0	28	Belgian	Football	66	0	1	0	0	1	1	2	B	FALSE
x8601c	0	21	Dutch	Tennis	71	0	1	0	0	2	1	1	B	FALSE
x1131d	0	26	Belgian	Tennis	69	0	1	0	0	1	1	2	B	FALSE

Group 14

DATA EXPLORATION 1

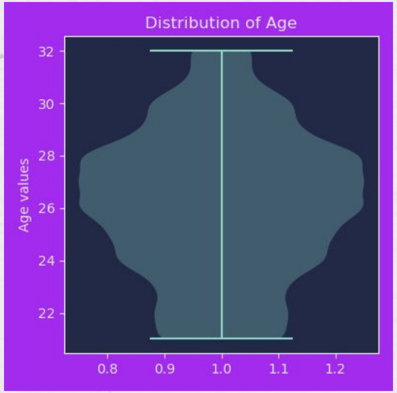


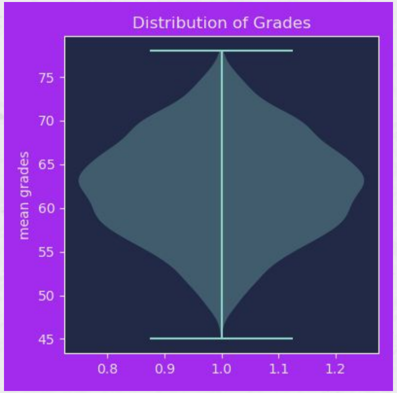
This scatterplot represents the performance scores of people from different age groups, their age is on the x-axis and their performance score is on the y-axis. The gender is represented by different color, and the sport is represented by the marker shape. We can see that from ages 21 to 23 the lowest grade is 56, while the highest is 77. Students who are from 24 to 28 are the most representative in the category, since the grades go from the lowest to the highest ones.

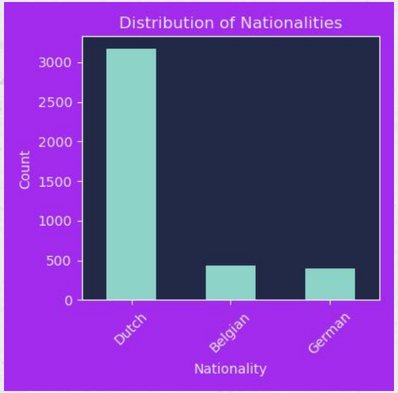
2023

Group 14

DATA EXPLORATION 1







Distribution of Age

On the violin plot presented above we can see that the age in our dataset varies from 21 to 32. The width of the violin plot provides insight into the density of the data, and the most common ages are between 26 to 28, while less common ages are 31 and 32.

Distribution of Grades

On the violin plot we can observe the distribution of grades, where the lowest is 45 and the highest 76. Most common value is 64 and two least common values are 76 and 45.

Distribution of Nationalities

On the bar chart above we can see that the majority of people have Dutch nationality (around 3000), while Belgian and German nationalities are around 500 people each.

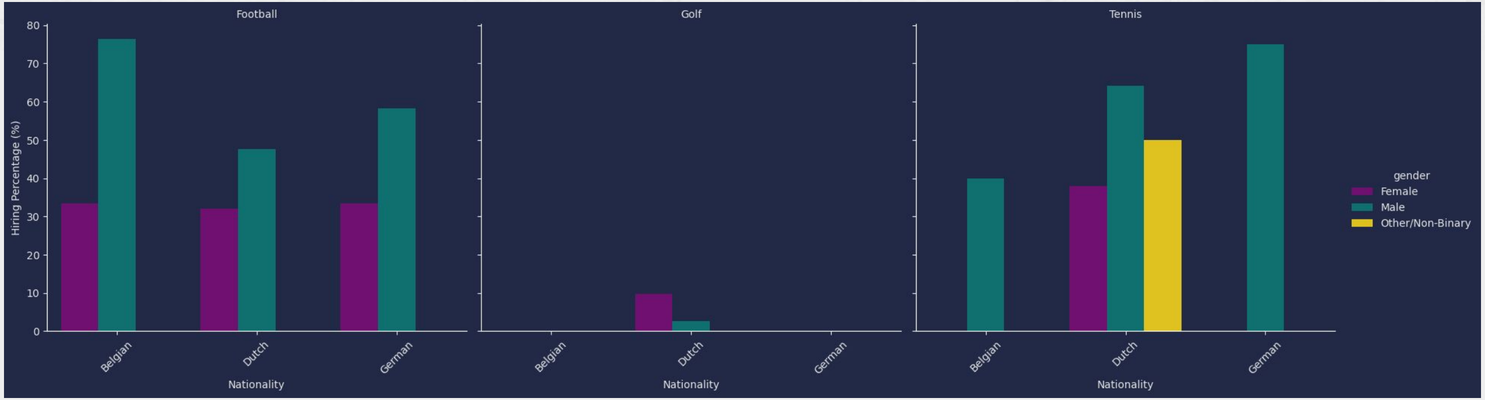
2023

DATA EXPLORATION 1

In the table below we can see the minimum and maximum values of our dataset. Where minimum value for age is 21, while maximum is 32; the minimum university grade is 45 and the maximum university grade is 78. All the other values from our dataset were not numeric ones, thus they are not presented in the table.

Metric	Min Value	Max Value
Age	21	32
University Grade	45	78

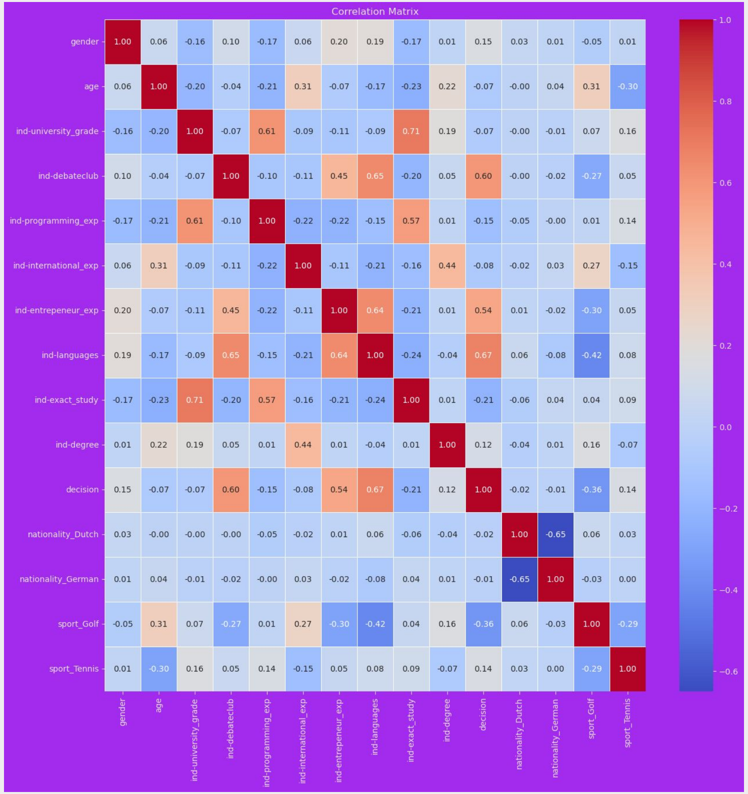
DATA EXPLORATION 2



In the graph above we can see hiring percentages within football, golf and tennis based on your nationality and gender. For instance, we can see that female players are less likely to be hired in football and tennis, while in golf female hiring percentage is higher than male hiring percentage. In terms of nationalities Belgian male (76%) is more likely to be hired than Dutch male (around 50%) in football, while in tennis Belgian male (40%) is less likely to be hired than German male (70%)

2023

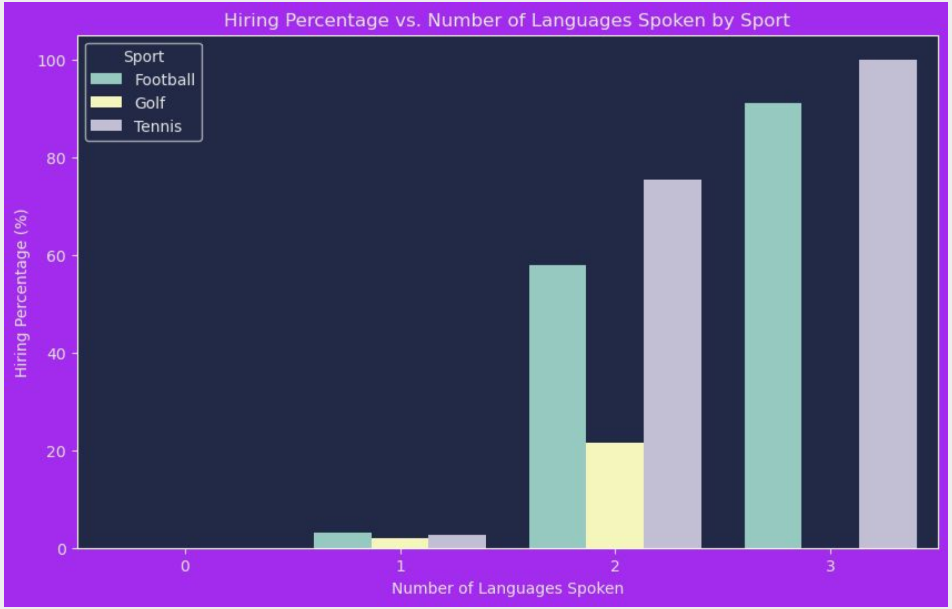
DATA EXPLORATION 2



This matrix is to check if there is some linear relation between any of our columns, to see if there's anything worth exploring further. We see that there are three columns, ind-debateclub, ind-entrepreneur_exp, and ind-languages that seem to have higher correlation than expected for hiring decision, so we will explore these further in the next plots.

2023

DATA EXPLORATION 2

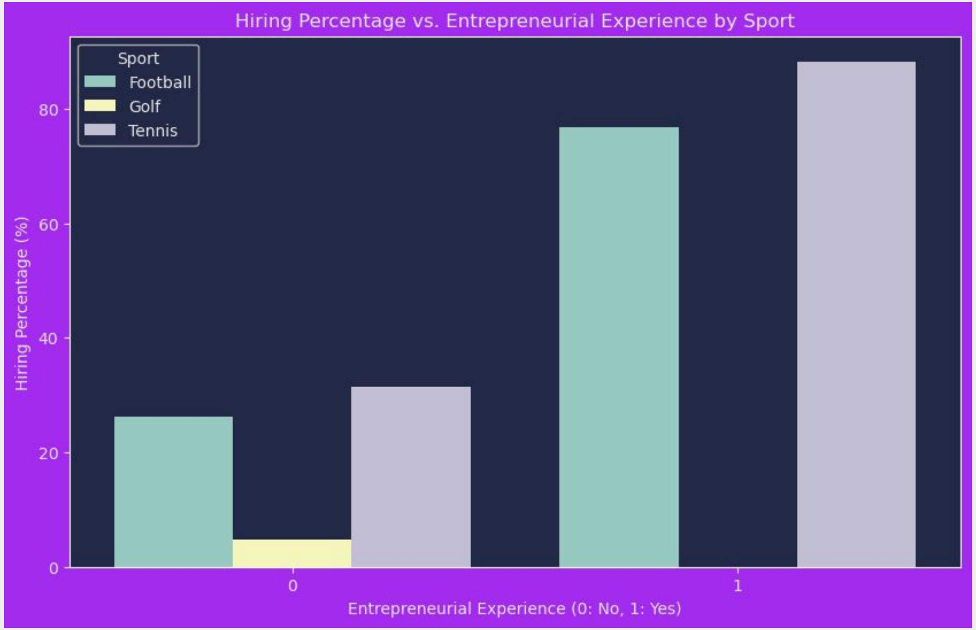


This graph highlights how linguistic proficiency influences hiring trends in sports, revealing both the importance of languages and potential biases across different sports. For instance, if a person speaks one language hiring percentage is only about 5% for all three sports, while with the knowledge of 3 languages the hiring percentage grows up to 90-100%, however only within football and tennis. Somehow, hiring percentage is 0 when knowing 3 languages in golf, but with 2 languages the percentage grows up to 25%.

2023

Group 14

DATA EXPLORATION 2

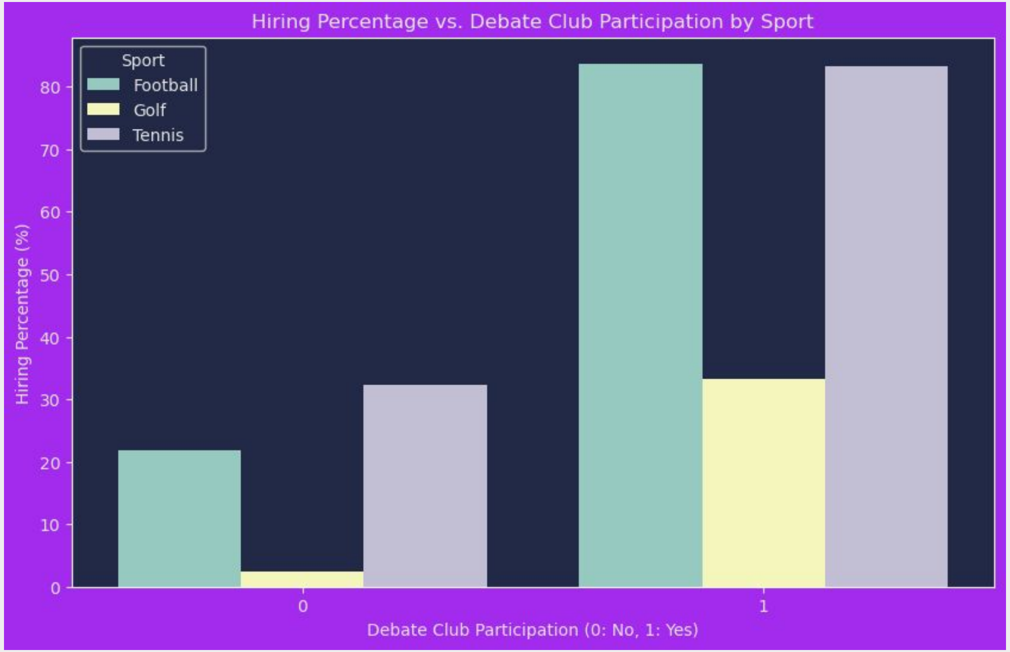


We plotted the hiring proportion versus entrepreneurial experience and found a clear connection. Athletes without prior entrepreneurship experience had a hiring rate of roughly 25–30% for sports like football and tennis. For those with expertise, this number rose to an astonishing 80–90%, a remarkable increase. It's interesting to note that golf departed from this pattern, indicating that recruiting practices in this sport may be affected or relevant by entrepreneurial experience differently. This information highlights the important role that entrepreneurship experience has in recruiting decisions within specific sports industries.

2023

Group 14

DATA EXPLORATION 2



This graph illustrates the impact of debate club participation on hiring percentages in various sports. It showcases whether being part of a debate club makes a difference in hiring decisions for different athletic disciplines. We can see that Athletes without prior debate club participation had a hiring rate of roughly 25–35% for sports like football and tennis. For those with experience, this number rose to an astonishing 80–90%, while for golf it also showed that hiring percentage rose when having experience, however the percentage is not as high.

2023

MODELS SELECTION METHOD

Objective:

Our main objective is to choose the most accurate model for predicting the decision from a variety of variables in our dataset. We search for a model that uses no more than 4 indicators.

Data Processing:

After creating a clean copy of the filtered dataset to preserve the original, we listed all columns and removed 'decision', 'grade_bin', 'Id', 'company', 'nationality', 'age', and 'sport' due to their roles as target variables, identifiers, irrelevance to the model or ethical concerns.

Feature Combination:

We explore feature combinations up to a maximum size of 4 since using more indicators could be very expensive for the company.

2023

DECISION TREE

Data Splitting:

We divided the data into training (70%) and testing (30%) sets for each combination of features. This makes sure that our model is assessed using hypothetical data, giving us a more accurate picture of how well it performs.

Hyperparameter Tuning:

We train the model on the test set and assess its performance using a Decision Tree classifier with 'gini' and 'entropy' criteria for each feature combination. We also make use of a grid of parameters, which we looped through to find the best combination. The parameters we also made use of were max tree depth, the minimum split of samples, and the minimum number of samples of leaf nodes in the tree. These hyperparameters help avoid overfitting and assist in limiting complexity. The tree size may also be affected in some cases.

Performance Evaluation:

Accuracy is the chosen performance metric. We evaluate the accuracy of each model against the highest level of accuracy currently attained. We update our best accuracy, the best set of indicators (features), the best criterion, and the maximum depth of the tree if a model outperforms the prior best. It is worth noting that we also account for F1 score and precision

DECISION TREE RESULTS

The results output of the initial model gave us a set of best parameters, indicators, and a set of scores based on the parameters and indicators that performed best.

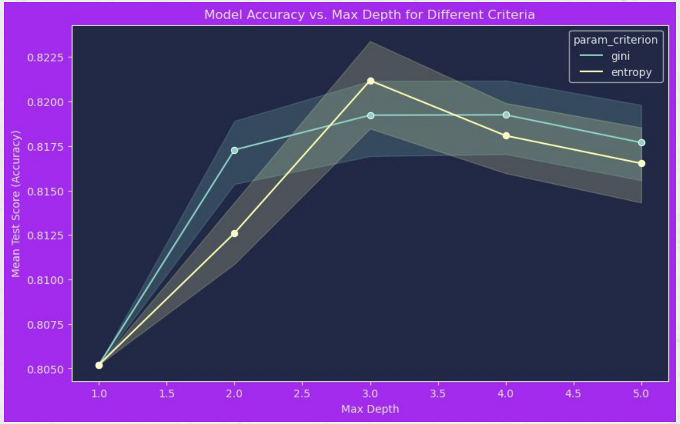
- Best parameters:
  - criterion: gini, max_depth: None, min_samples_leaf: 10, min_samples_split: 40}
- Best indicators:
  - ind-university_grade, ind-debateclub, ind-entrepeneur_exp, ind-languages]
- Accuracy score: 0.862
- Precision score: 0.784
- F1 score: 0.825

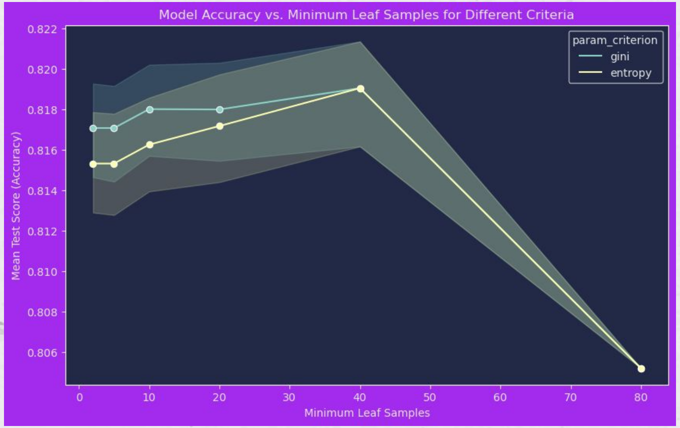
2023

Group 14

DECISION TREE RESULTS

Using these line plots, we can demonstrate the range of accuracy values we observe in our data depending on which criterion formula we use on our data. Comparing them against different hyperparameters helps demonstrate how we arrived at the best-fit results from the model. For example the Accuracy vs. Max Depth chart clearly shows that both criterion perform best with a max depth of three and 40 Minimum Leaf Samples.



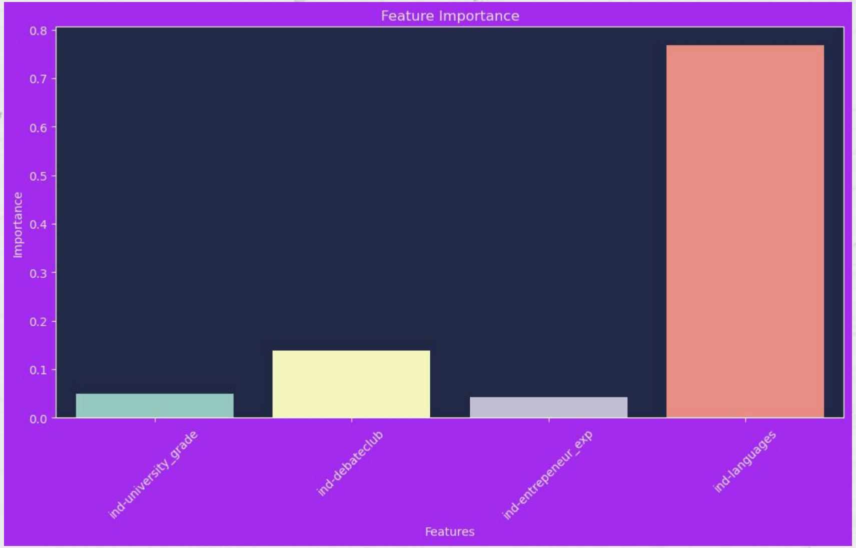


* A leaf sample minimum in this case indicates the minimum number of times a leaf must be sampled for a node split to occur.

2023

DECISION TREE RESULTS

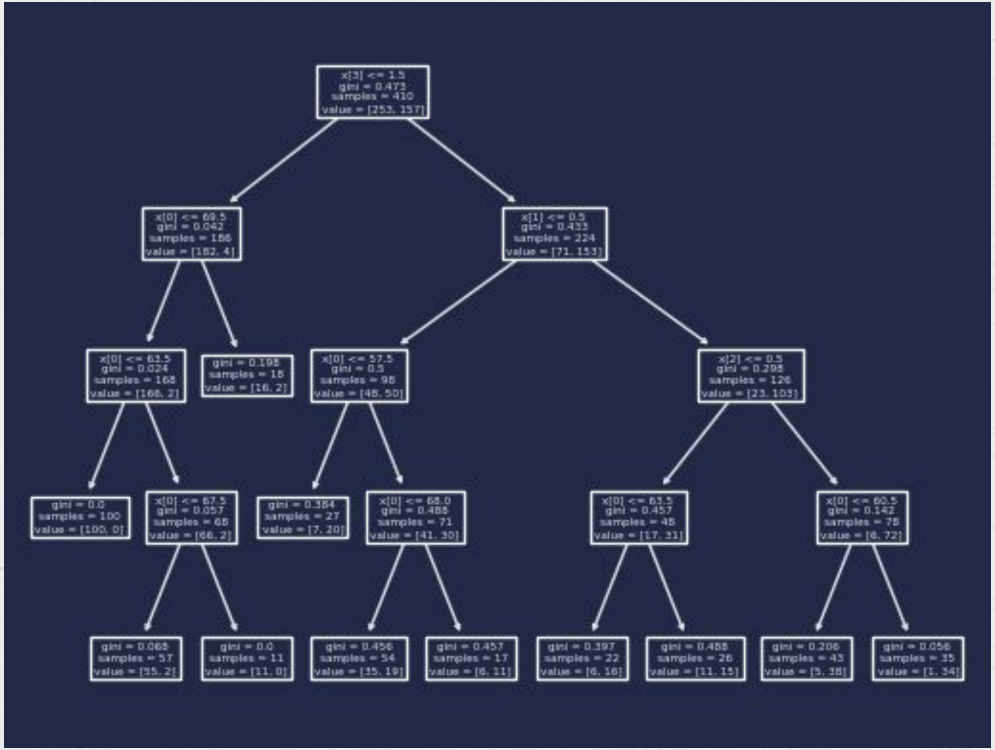
We also created a bar chart that shows feature importance. We can notice that some features clearly play a more pivotal role in the model's decisions, as evidenced by their taller bars. For instance, the most significant one is ind-languages and its importance is 0.8, while the least important is ind-entrepeneur_exp (less than 0.05).



2023

DECISION TREE RESULTS

A decision tree is a flowchart-like structure in which each internal node represents a decision or a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class label.

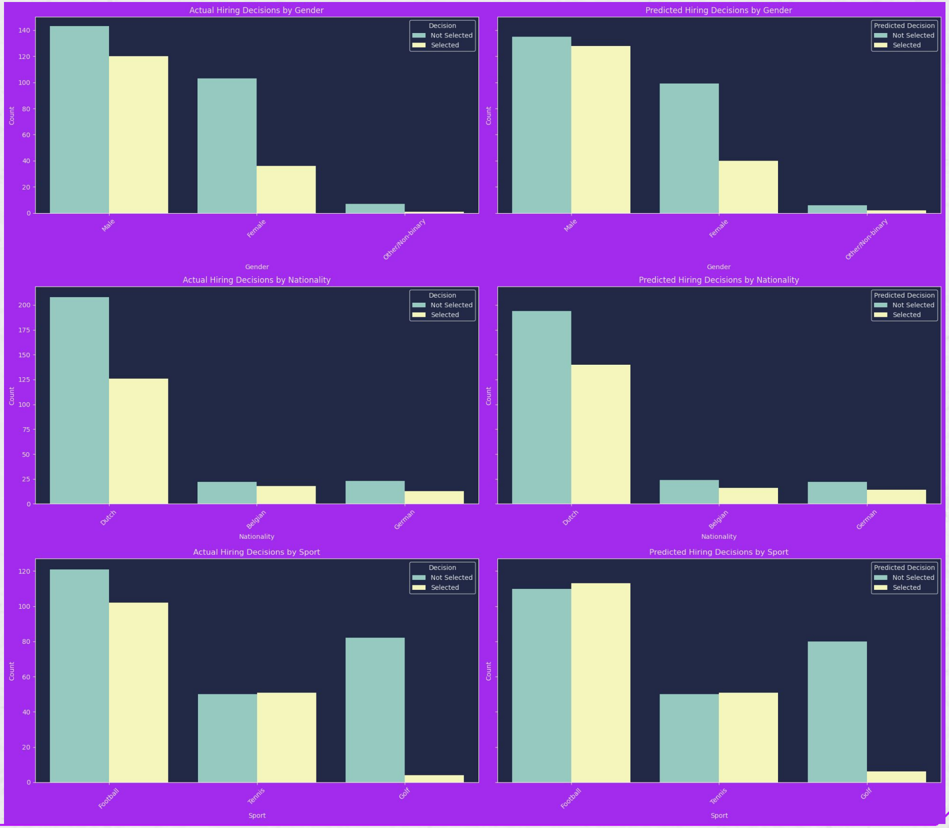


2023

Group 14

FINAL MODEL EXPLORATION

In analyzing the results from our final model, we observe that selections across various factors were broadly balanced. Noteworthy deviations occur in instances grouped by nationality and gender. Specifically, individuals of Dutch nationality exhibit a marginally higher propensity to be selected when applying our model. Similarly, male candidates demonstrate a slightly higher likelihood of being hired than their female or non-binary counterparts. These observed impacts are considered mixed within our analysis framework for two primary reasons. Firstly, the proportional shifts in selection across other categories align closely with those within the 'Not Selected' category when compared with actual hiring decisions, which bodes well for our model. Secondly, the disparities in the hiring predictions for Dutch individuals and males do seem to indicate significant biases in the source data, skewed to hiring these two categories.



2023

K-nearest neighbors algorithm results

The results output of the initial model gave us a set of best parameters, indicators, and a set of scores based on the parameters and indicators that performed best.

- Best parameters:
  - Number of neighbours - 7, best metric - euclidean, best weights - distance.
- Best indicators:
  - Debate club participation, Entrepreneur experience, number of languages and degree.

- Accuracy score: 0.854
- Precision score: 0.769
- F1 score: 0.816

K-nearest neighbors algorithm

Data Splitting:

We divided the data into training (70%) and testing (30%) sets for each combination of features. This makes sure that our model is assessed using hypothetical data, giving us a more accurate picture of how well it performs.

Hyperparameter Tuning:

We train the model on the test set and assess its performance using a K-nearest classifier. We make use of a grid of parameters, which we use to find the best combination. The parameters are: number of neighbours, weights and metrics.

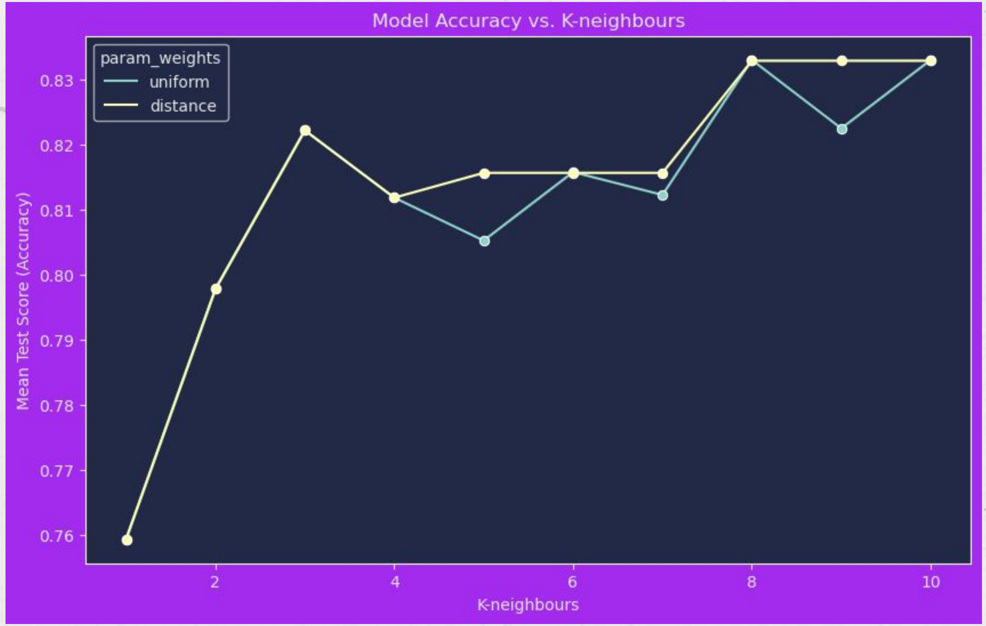
Performance Evaluation:

Accuracy is the chosen performance metric. We evaluate the accuracy of each model against the highest level of accuracy currently attained. It is worth noting that we also account for F1 score and precision

2023

MODEL ACCURACY

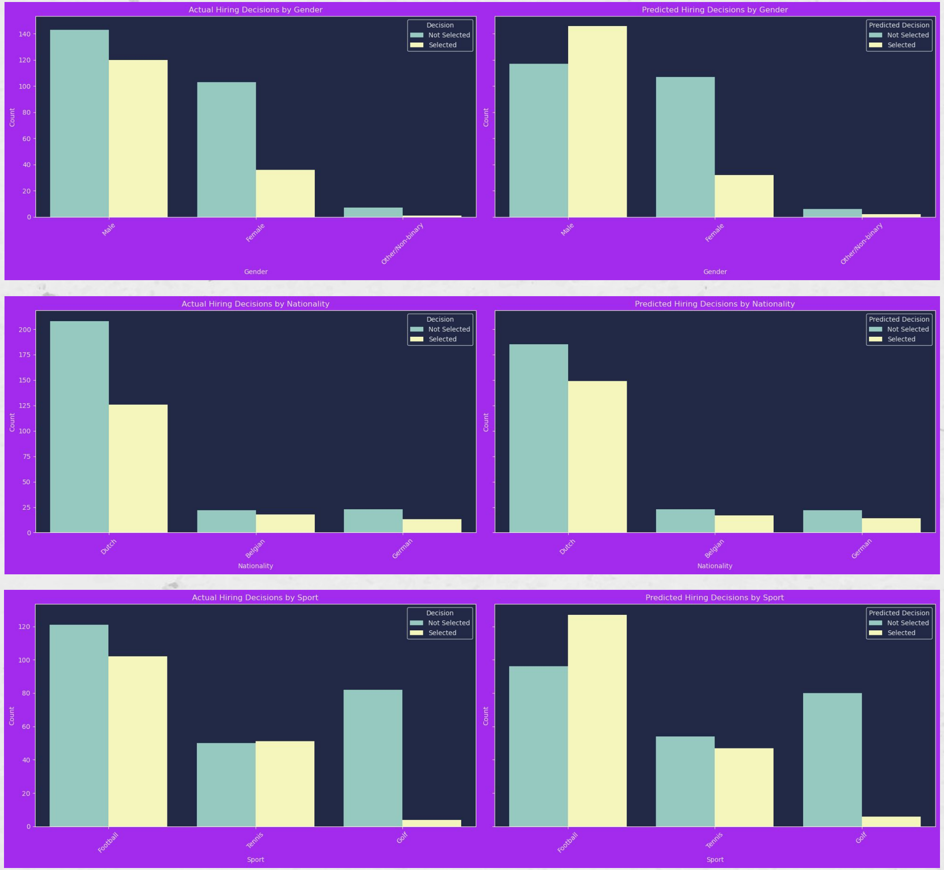
The graph is a line plot, showing the relationship between the number of neighbours, and the accuracy test score. The accuracy starts high when K_n is near 0, drops sharply around K_n = 2, and then climbs again as K_n approaches 10. Y-Axis, ranging from 0.74 to 0.82, represents the model's accuracy. X-Axis is from 0 to 10, represents the number of neighbours. The model seems to work best when 'C' is around 3. Since high K tend to overfit, 3 is the best parameter to use.



Group 14

FINAL MODEL EXPLORATION

In analyzing the results from our final model, we observe that selections across various factors were broadly balanced. Noteworthy deviations occur in instances grouped by nationality and gender. Specifically, individuals of Dutch nationality exhibit a marginally higher propensity to be selected when applying our model. Similarly, male candidates demonstrate a slightly higher likelihood of being hired than their female or non-binary counterparts. These observed impacts are considered mixed within our analysis framework for two primary reasons. Firstly, the proportional shifts in selection across other categories align closely with those within the 'Not Selected' category when compared with actual hiring decisions, which bodes well for our model. Secondly, the disparities in the hiring predictions for Dutch individuals and males do seem to indicate significant biases in the source data, skewed to hiring these two categories.



2023

Logistic Regression

The results output of the initial model gave us a set of best parameters, indicators, and a set of scores based on the parameters and indicators that performed best.

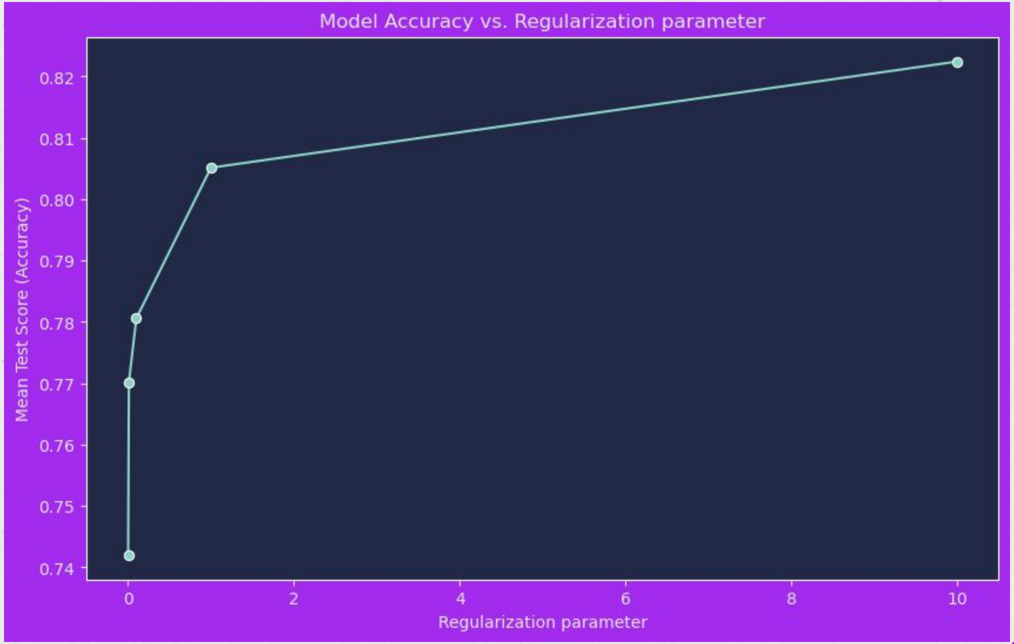
- Best parameters:
  - {'C':10}
- Best indicators:
  - Programming experience, languages, exact study, degree

- Accuracy score: 0.862
- Precision score: 0.809
- F1 score: 0.817

Group 14

MODEL ACCURACY

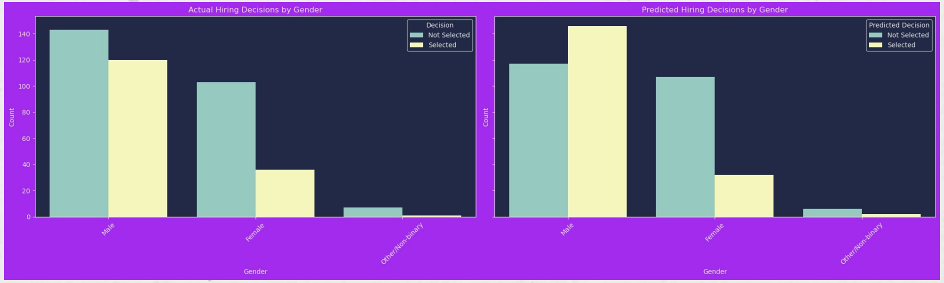
The graph visualizes the impact of different regularization parameters on a model's accuracy. The x-axis represents the regularization parameter values, while the y-axis displays the corresponding mean test scores or accuracy values. The relationship is depicted as a line plot with distinct circular markers indicating each data point. As the regularization parameter varies, the line plot shows how the model's accuracy changes, providing insights into the optimal regularization parameter for maximum accuracy.

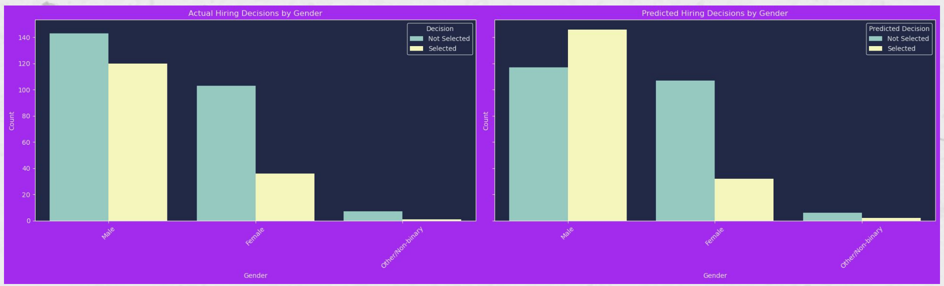


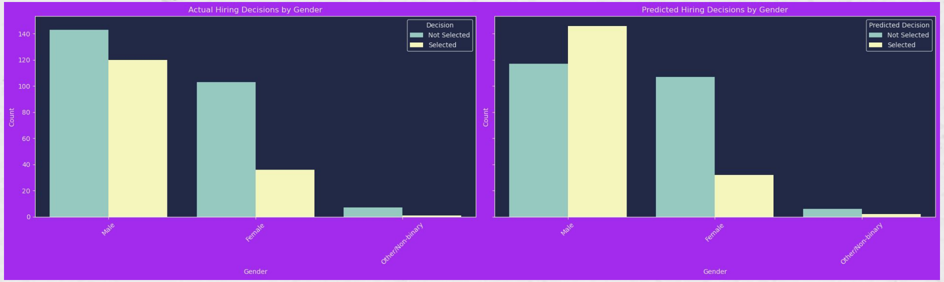
Group 14

FINAL MODEL EXPLORATION

In analyzing the results from our final model, we observe that selections across various factors were broadly balanced. Noteworthy deviations occur in instances grouped by nationality and gender. Specifically, individuals of Dutch nationality exhibit a marginally higher propensity to be selected when applying our model. Similarly, male candidates demonstrate a slightly higher likelihood of being hired than their female or non-binary counterparts. These observed impacts are considered mixed within our analysis framework for two primary reasons. Firstly, the proportional shifts in selection across other categories align closely with those within the 'Not Selected' category when compared with actual hiring decisions, which bodes well for our model. Secondly, the disparities in the hiring predictions for Dutch individuals and males do seem to indicate significant biases in the source data, skewed to hiring these two categories.







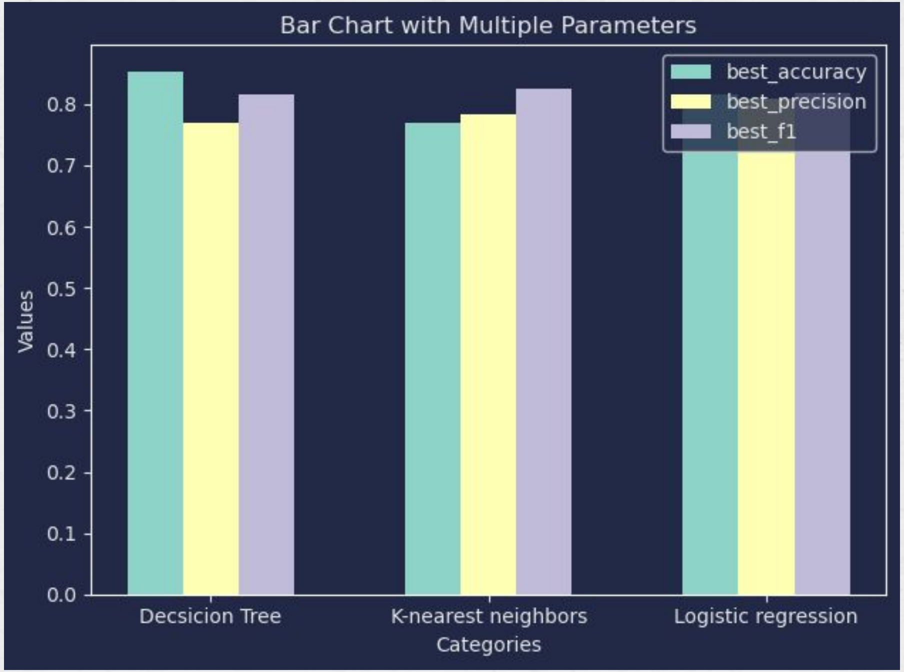
2023

Group 14

FINAL PLOT

This bar chart offers a comparative look at how three different algorithms work against three parameters.

We decided to judge performance of the models by F1 metric. Since it balances the trade-off between precision and accuracy,and it's useful when both, false positives and false negatives matter.



REFERENCES

[1] van Otterloo, S. (2022) Utrecht Fairness Recruitment Dataset, Kaggle. Available at: https://www.kaggle.com/datasets/ictinstitute/utrecht-fairness-recruitment-dataset (Accessed: 10 October 2023).

[2] Hotz, N. (2023) What is CRISP DM?, Data Science Process Alliance. Available at: https://www.datascience-pm.com/crisp-dm-2/ (Accessed: 10 October 2023).