Advanced Data Analysis in R Final Project

Team 14: Ariel Siman Tov, Tal Klein, Ido Villa

GitHub file link: https://github.com/Idovilla/Final-Project-Team-14-.git

Introduction:

General data science question: "What are the key personal attributes and qualifications that significantly contribute to the candidate being hired to an office work in Netherlands?"

Our focus in is this analyzing work is to analyze the "office work" market in the Western world, with a specific emphasis on the Netherlands as a representative country.

The Netherlands, as a country, is known for its well-developed infrastructure, strong economy, and diverse job market. Utrecht, the fourth-largest city in the Netherlands, is a major hub for businesses, including many national and international organizations. Utrecht is familiar with various industries and sectors, ranging from finance and technology to healthcare and research. Therefore, we find Utrecht as a city that is suitable to represent the "office work" market in the Netherlands.

The prevailing opinion in the job market today is that a candidate has a higher chance of being hired if he has a wider "toolbox" and knowledge/experience in the field he is interested in working in. Some key attributes and qualifications that are generally valued in the Dutch job market in particular include:

- Educational background - The Netherlands places a strong emphasis on education and having a relevant degree.
- Language skills: Proficiency in both Dutch and English is highly desirable in many job roles.
- Experience and skills: relevant work experience and specialized skills.

This analysis is important because it provides valuable insights for job seekers, recruiters, and employers by identifying the factors that play a crucial role in candidate selection. By understanding the specific attributes and qualifications that are highly valued, job seekers in Netherlands can focus on developing those skills and presenting themselves as strong candidates. Recruiters and employers can use this information to improve their hiring processes, refine job descriptions, and assess candidates more effectively.

This problem is difficult because the world is developing and changing. The skills and attributes that were considered relevant in the past may no longer hold the same importance in today's job market. This creates a need for updated and current analyses to understand the characteristics that are valued by employers when making hiring decisions.

Our research aims to provide valuable insights into the key features and abilities that hold the utmost relevance in today's job market in Nederland. By conducting this study, we seek to gain a comprehensive understanding of the specific characteristics that employers prioritize when making hiring decisions.

Data:

Our dataset[1] consisting of recruitment decisions from 4 major companies in Utrecht, the fourth-largest city in the Netherlands. Each candidate in the dataset has various criteria and general descriptions, along with information about whether they were accepted or rejected by the company they applied to.

The data includes 4000 samples that represent each candidate. Each company has 1000 candidates, as well as 15 columns. The columns in the dataset provide details about the candidates, such as their unique identifier (id), gender, age, nationality, main sport, university grade, participation in debating/social clubs, programming experience, international experience, entrepreneurial experience, fluency in additional languages, study background (science-oriented or not), highest completed degree, the company they applied to, and the decision outcome (whether they were hired or not).

In the analysis of the data and building the models, we used all the columns except the "Id" column, because this column is used as a unique key to identify an entity in the data. In addition, we will remove the company column because in Part 3 of the analysis we will verify the model, according to this column.

Methods and Results:

In order to answer our research question, we chose to work according to several steps:

1. We selected the best classification model to achieve the most accurate results, by comparing two well-known classification models.
2. We found the features that are most significant in the model. For the relevant features found, we expanded a detailed analysis.
3. We verified our findings by comparing between the four sub-samples (indicating different companies).

[1] https://www.kaggle.com/datasets/ictinstitute/utrecht-fairness-recruitment-dataset

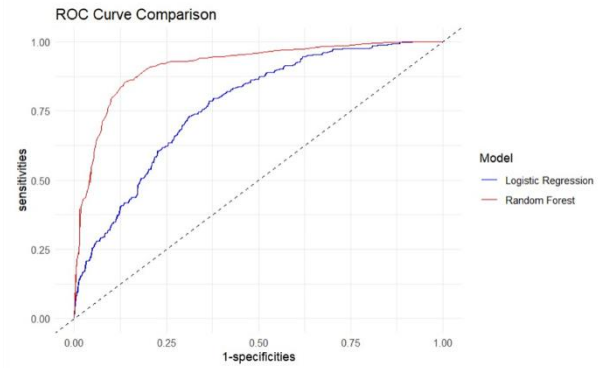
Part 1:

We chose to compare two well-known classification models, "Logistic Regression" (which we learned as part of the course) and "Random Forest" (which we learned about ourselves) on all the data in general to find which model will reach more accurate results. Through this comparison, we got a first impression of the data and its characteristics.

Observations:

It can be seen by comparing the ROC_AUC, Accuracy, Recall, F1 Score and Precision between the two models, that the Random Forest is better and leads to more accurate results. **Therefore, we chose to continue the analysis using this model.**

After choosing the Random Forest model, **we performed a Cross-Validation[2] analysis** to make sure that the Random Forest model doesn't cause overfitting.



Logistic Regression

Area under the curve: 0.7694

Precision: 0.4171779
Recall: 0.5991189
F1 Score: 0.4918626
Accuracy: 0.719

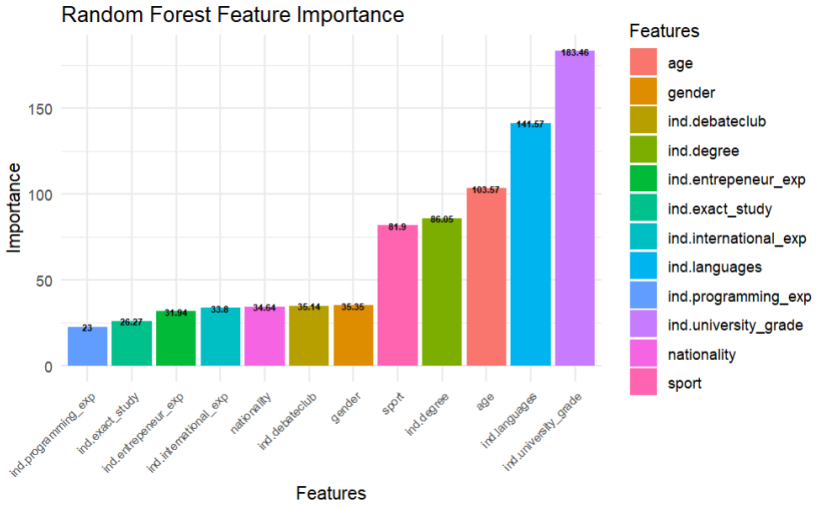
Random Forest

Area under the curve: 0.9135

"Precision: 0.802867383512545"
"Recall: 0.687116564417178"
"F1 Score: 0.740495867768595"
"Accuracy: 0.843"

Part 2:

We calculated which are the features that have the greatest influence on the decision to accept a candidate for a job by the Importance (Random Forest build in method) that based on the "MeanDecreaseGini"[3] calculation.

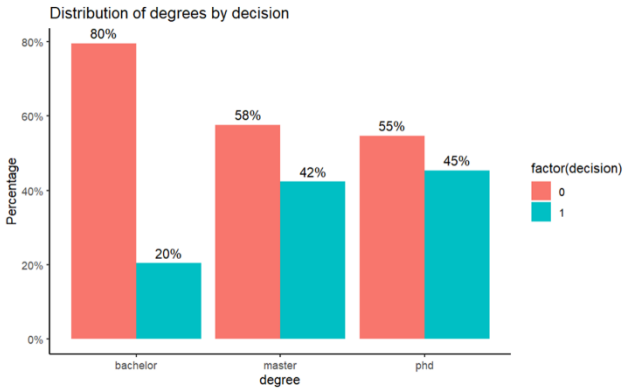


Observations:

According to the "MeanDecreaseGini" calculation value, the top 5 parameters which have the greatest influence on the decision variable are: grade, number of languages, age, degree, and sport.

It can be seen from the importance analysis that from the sixth category onwards, there is a significant decrease in the "MeanDecreaseGini" calculation value, which remains relatively the same in the other categories.

An extended analysis for 3 of the 5 main categories:



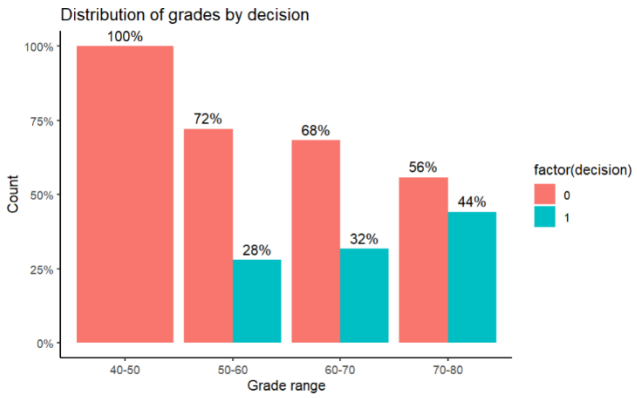
Observations:

*The probability of people who are hired increases as they have a more advanced degree.

*A master's degree confers a high percentage of job acceptance. We assume that a master's degree is more common among the population as well as in our data.

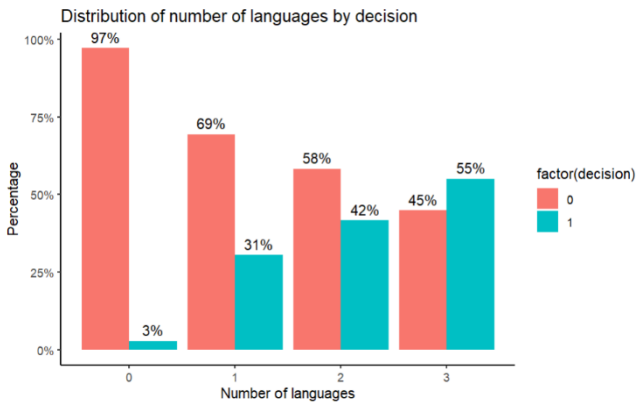
[2] Full analysis of the model results attached in the appendices.

[3] Explanation attached in the appendices.



Observations:

*The higher the grades, the greater the chance of being hired.



Observations:

*The percentage of people who are hired increases as they have more knowledge of different languages.

Part 3:

To validate the results of the classification model, we made a comparison between the four different companies. We assumed that if the same features will be significant in each sub population, the model will be more robust. This way, we can make a comparison in another level, against the analysis performed on the whole data.

Observations:

*The "grade" and "languages" features are significant in **all four companies.** The "degree" feature is significant in three companies, **in accordance with the results obtained for the importance analysis performed on all the data.**

*We noticed that the age feature does not appear in any of the four companies as one of the five important columns, in contrast to the fact that it appears as an important category in the analysis of the entire sample. This can happen because when you re-run Model X on subsets of the data, the model may find other, more informative features to predict the decision variable for the subsets.

Result summary:

According to the data analysis we performed on the data, it can be understood that the key personal attributes and qualifications that significantly contribute to the candidate being hired are **grade, degree, and know several languages.** These conclusions support and strengthen the basic hypothesis prevalent today. We reached these conclusions after a multi-stage validation of the model and results.

Limitations and Future work:

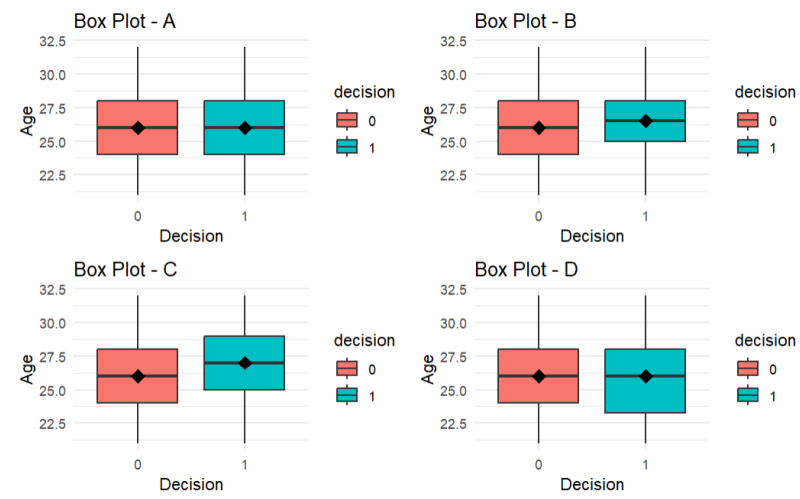
We believe that the primary difficulty in our data analysis lies in the limited number of companies and the small number of candidates within each company, as well as the limitation to information relevant only for the Dutch employment economy. Having a larger sample size would enable us to obtain more precise and reliable results. Furthermore, another challenge we faced was the lack of information about the specific industries in which these companies operate, which could potentially impact our findings. However, by analyzing the key features associated with each company and their candidate selection process, we can make informed assumptions about their fields of activity. Additionally, if we had more time, we would be interested in exploring additional databases that contain information on other companies and include supplementary features. By merging these datasets, we could achieve more comprehensive and accurate conclusions.

Appendices:

1. **The "MeanDecreaseGini" value** for a feature is computed as the average reduction in Gini impurity across all decision trees in the random forest when that feature is used for splitting. Features that lead to larger reductions in Gini impurity are considered more important for the classification or regression task.
   By examining the "MeanDecreaseGini" values, we can identify which features have the most significant impact on the random forest model's performance. Features with higher values are generally more influential in the model's predictions, while features with lower values contribute less to the overall prediction power.

2. **Gini Impurity** is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree. More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.[4]

3. **Extended analysis for the age category against all four companies:**



It can be concluded from this plot that there is no difference between the median and the STD between the 4 different companies in the segmentation of the age characteristic.

4. **Cross-Validation full analysis results:**
   We performed a cross-validation analysis with the Random Forest model in which we divided the data into 10 folds.

Calculation of the MSE for each fold:
```
[1] 0.3968627 0.3708099 0.3741657 0.3712743 0.4387482 0.4123106 0.4000000 0.4056951 0.3963675
[10] 0.4218351
```

Mean of the MSE calculations: `[1] 0.3988069`

Variance vector of the MSE calculations: `[1] 0.0005017354`

Calculation of the ROC_AUC for each fold:
```
[1] 0.7830031 0.7969196 0.7929826 0.8095238 0.7232650 0.7756771 0.7992991 0.7802891 0.7749152
[10] 0.7683150
```

Mean of the MSE calculations: `[1] 0.780419`

[4] https://www.learndatasci.com/glossary/gini-impurity/