Project: A - Succession Management WhitesT

Consultants: SALVATORI Alessandro – r0926525

WYNANTS Axl – r0861234

LAE

Leuven Analytics Experts & Partners

Part 1: Data exploration

Visual

Comments:

- There seem to be three groups of people who are leaving:
  1) People that have a high number of hours of work (>240) and extremely low job satisfaction level (around 0.1);
  2) People with low amount of hours of work (130-160) and average job satisfaction (around 0.4);
  3) People with high amount of hours of work (220-270) and high job satisfaction (0.7-0.9).
- It appears that most of the people that work at WhitesT have a good job satisfaction (>0.5) and work monthly hours ranging from 130 to 270.
- There are 10,000 data points, and after transformation: 5 continuous variables and 5 categorical variables.
- There is no missing data.

Part 2: Modelling

2.a. Models

Technique 1: Decision Tree (DT)

Motivation:

- Simple to understand, interpret and explain.
- Performs well, generating accurate predictions or outcomes even in the presence of outliers, noise, or fluctuations in the input data.

Technique 2: Gradient Boosted Trees (GBT)

Motivation:

- Facilitates the capture of complex relationships and interactions within the data.
- Tends to improve performance through the use of ensemble techniques.

Technique 3: k-Nearest Neighbors (k-NN)

Motivation:

- Simple lazy learner with low training cost.
- Non-parametric method, implying that it does not make any assumptions about the underlying distribution of the data.

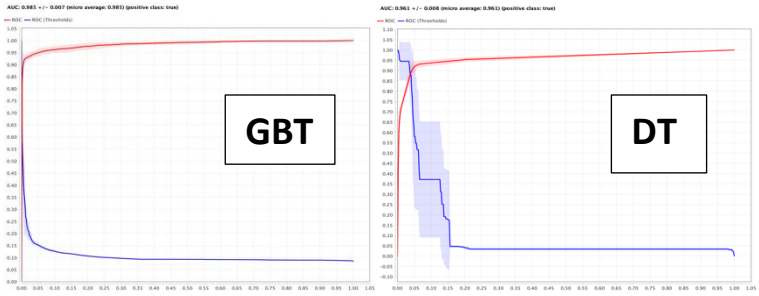
2.b. Evaluation

- In evaluation of the models, we opted to use the AUC because the dataset is quite unbalanced, so the usual accuracy measure can be uninformative in this case. The AUC was chosen since it provides a good trade-off between sensitivity and specificity, as well as allowing for cross-model comparison. AUC also provides a handy interpretation for our use case: it can be interpreted as the probability that a randomly chosen leaver gets a higher score from the model than a randomly chosen non-leaver. For all models, 10-fold cross-validation was used to detect possible overfitting. Moreover, for Decision Tree and Gradient Boosted Trees, the maximum depth was set to 7 to help with interpretability. The obtained AUC values for evaluation can be seen in the table in the Results section.

Part 2: Modelling

2.c. Results

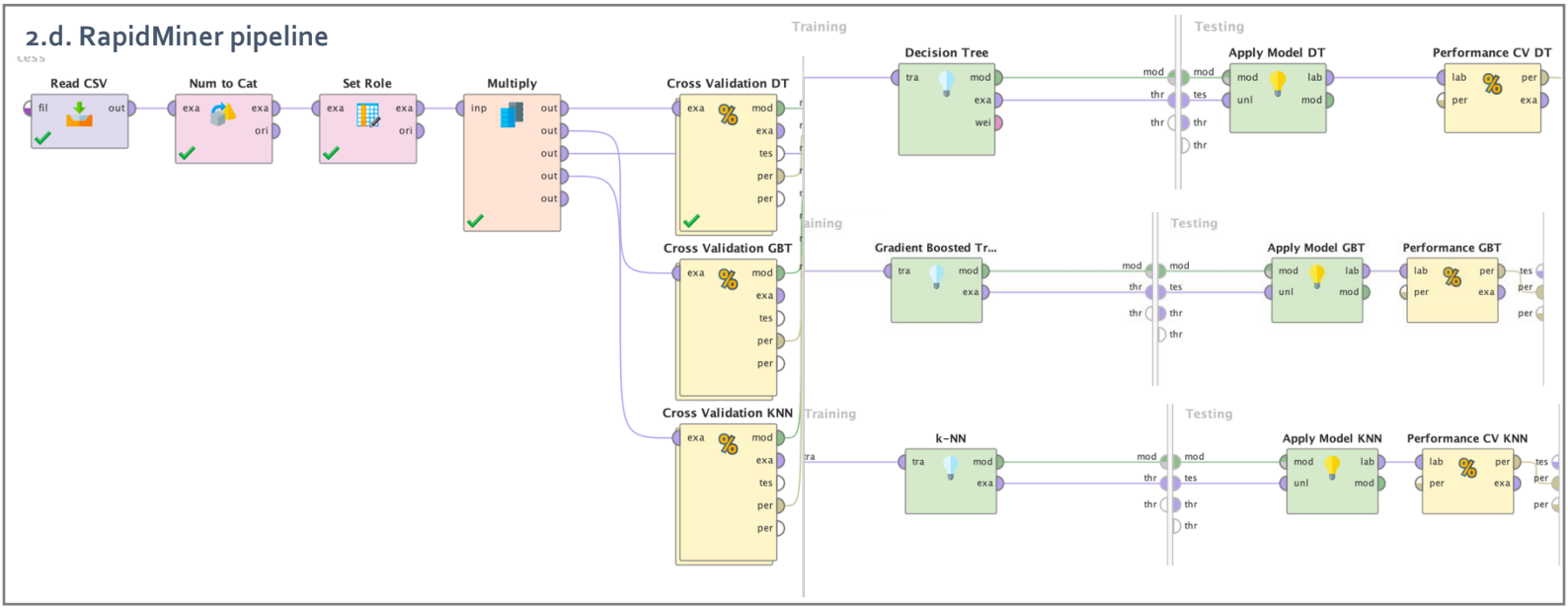
Model	DT	GBT	k-NN
AUC	0.961 +/- 0.008	0.985 +/- 0.007	0.974 +/- 0.004



Discussion:

- Best performance: Gradient Boosted Trees.
- Based on AUC (standardized metric), the best model is GBT, then k-NN, and the worst is DT, but the values are still all extremely good.
- Based on GBT, the most important variables are satisfaction, time at the company, and last evaluation, while based on the decision tree, the most important variables are satisfaction and number of projects.
- K-NN has worse accuracy but has good AUC value compared to DT, showing accuracy can be misleading.

2.d. RapidMiner pipeline



Part 3: Pre-processing

1. **Quality of Data**: Pre-processing is essential to enhance the quality of the raw data which is often noisy and contains missing values and outliers. Cleaning and correcting these issues contribute to a more reliable dataset, ensuring that analytics tasks are based on accurate and trustworthy information.
2. **Normalization and Standardization**: Pre-processing includes techniques like normalization and standardization to bring different variables to a common scale. This is crucial for models sensitive to the magnitude of variables, ensuring fair and meaningful comparisons between features and preventing certain features from dominating others in the analysis.
3. **Feature Engineering and Selection**: It allows for the creation of new features through techniques like feature engineering, making the data more informative for analysis. Additionally, it involves selecting relevant features and eliminating irrelevant ones, reducing dimensionality, and improving the efficiency and effectiveness of data analytics models.