Introduction:

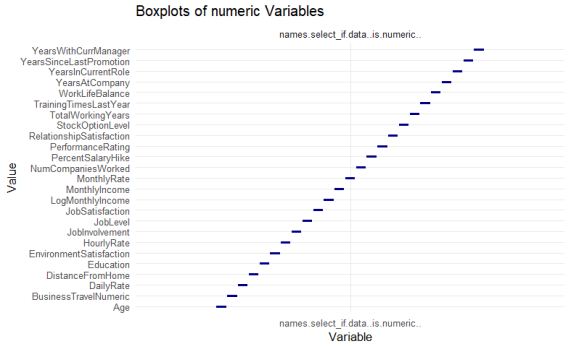
Our study will use the IBM attrition dataset with the target variable “Attrition” being the binary variable of whether an employee is retained or attrited (stay or leave). (the dataset is extensively explained in the rmd file).

Preprocessing and visualization:

In the preprocessing stage we have applied many techniques to handle irregularity with the dataset, firstly we removed four variables that are constant like the hours and the fact that they are not minors which are all the same additionally the identifier variable. Secondly, we checked for Missing values but there was none, hence no action was needed in that regard.

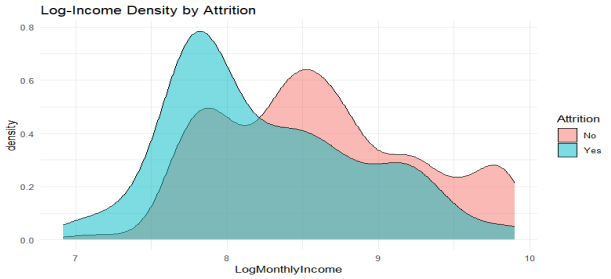
We also applied a few feature engineering techniques to some of the variables, for instance the Income variables are right skewed (a few employees earn significantly more than the majority). This skewness can negatively impact linear models we will apply. for this reason we will have to Log transform them. Other feature engineering we will apply is codifying of the ordinal variables like the travelling frequency, that is because we want to retain the ordinal nature of these variables.

Moreover we had to handle the outliers of the dataset and whether they raised a bias in the data so we have checked the skewness of the variables to determine the method we want to use to impute them.

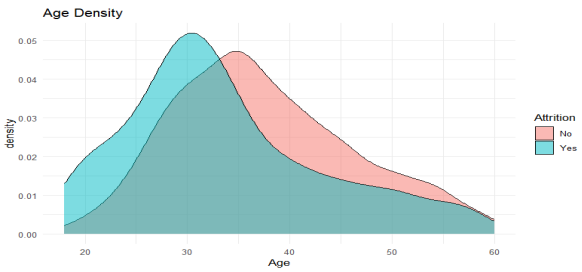


From the boxplots above we can see that many of our variables are skewed and since the 3-Sigma method assumes normality of the data we can not use it, we will proceed with using the 3 * IQR method to remove the outliers. Retaining 93% of the dataset

We now visualized some of the variables which seemed to have had some importance towards our study,

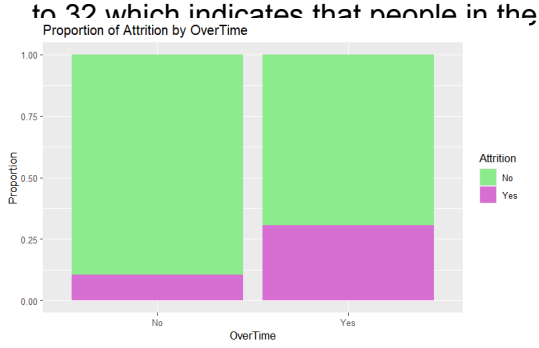


This graph shows that when employees have higher income there is a higher chance of staying, on the other hand we can also say that lower income results in



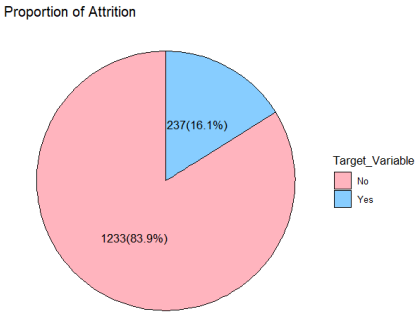
them leaving the company.there is a turning point where beyond that point in the income scale people are more likely to stay.

The age with attrition density graph shows, The "yes" group peaks at ages 28 to 32 which indicates that people in the chance of leaving or being let go of. while the "No" group are concentrated at age groups 32 to 36 and has higher density towards the 60s and 50s.

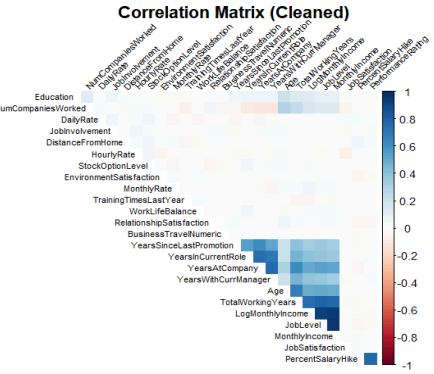


The overtime Barplot shows that employees who had done overtime were 3 times more prone to being attrited

The Job role with attrition barplot shows that more stress inducing roles usually result in higher attrition, in addition shows that the role of sales executive is extremely unstable



This biplot here shows that our data has a very skewed ration where 84% of the dataset is in the "No" category meaning that most our models would lean into predicting people who stay better than those who leave, this would suggest that it is better to focus on models prioritizing sensitivity and recall.



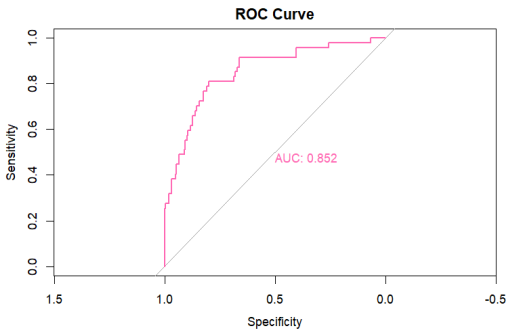
Finally, Our correlation shows a near-perfect positive correlation exists between `MonthlyIncome`, `JobLevel`, and `TotalWorkingYears` confirming that higher pay is related with more experience. In addition The high correlation between YearsAtCompany, YearsInCurrentRole, and YearsWithCurrManager, indicating these variables provide redundant and unnecessarily repeated information. There is also high variance between many variables which means that some regression models might struggle while ensemble based models like trees and forest would perform better.

Predictive classification:

Note: The models were run before and after the feature selection process and we will compare them below.

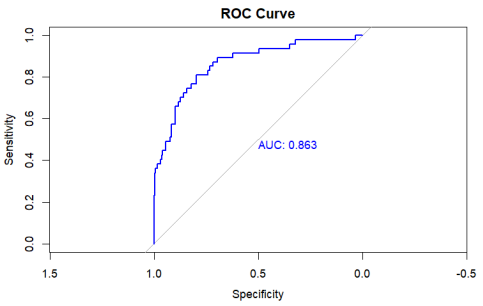
Logistic regression:

A simple model that provides interpretable results, using odds ratios for feature significance. It performs poorly with imbalanced datasets (such as our case), in addition the model's performance **after feature selection** was even poorer than before which tells us that all the features that were removed



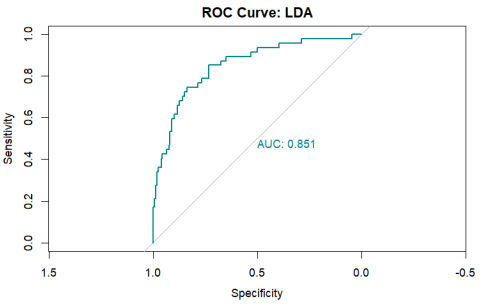
held great statistical significance. This model achieved **86% accuracy** and **95% sensitivity**, although it performed poorly on the specificity side with only **38%** meaning it is bad at predicting people who left. The **AUC** is **0.852** which is very strong although the cutoff probability is 0.5 which is too high for our imbalanced dataset.

Penalized Logistic regression:



This is a stronger variation of logistic regression, such as the definition this model performed slightly better than the baseline with **Higher Accuracy** Penalized Logistic Regression achieved **88.4%** accuracy, **Better Kappa** The Kappa score improved from 0.39 (Fair) to **0.46 (Moderate)** and The penalized model significantly reduced False Positives. Though they tied in **specificity** with **38%** and a slightly higher AUC**.** We also got very similar results with the **feature selection variables**.

Linear Discriminal Analysis (LDA):

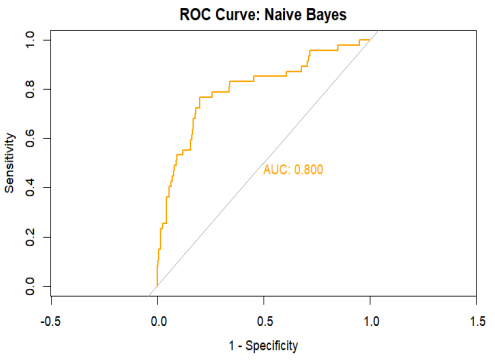


LDA is a powerful method for classification with a focus on maximizing class separability, this model performed better than the last 2 on all fronts it achieved higher **specificity 40.4%** and similar Accuracy with 86% which is better than the base model but lower than the penalized and an AUC of 0.85 tied with the two models. Ultimately it can be said that this is the better performing model until now, although it performed a little bit poorer with only the **"optimal selected variables**".

Quadratic Discriminal Regression:

QDA allows for a more flexible classification approach by modeling each class with its own covariance matrix. This model requires low collinearity between features which we do not have so the model failed under both the original dataset and the one with only the optimal features.

Naive Bayes:



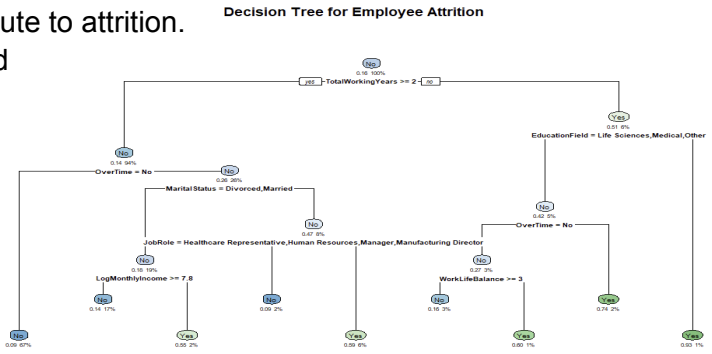
Naïve Bayes operates under the assumption of feature independence, which can work well in many situations despite this simplifying assumption. The **Specificity** here is much higher as it sits on **72.34%** which is considerably higher than LDA. meaning we were able to solve the issue of misidentifying the minority class "Yes". although there comes a tradeoff with the **accuracy** of the model as it has dropped down to **79.18%** and **AUC 0.8** which is lower than the other models but still very good considering the increase in specificity.

On the other hand the dataset with only the optimal selected features has performed poorer than the one with all the features which confirms that all our features are important to the classification process.

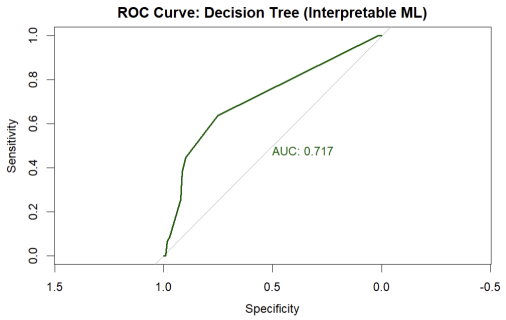
Interpretable Classification (machine learning):

Decision Trees

Decision trees classify employees through simple decision rules, making it possible to directly interpret how different characteristics contribute to attrition.



The estimated tree identifies TotalWorkingYears and OverTime as the most influential predictors, highlighting the importance of career stage and workload. Variables such as LogMonthlyIncome, MaritalStatus, and JobRole further refine the classification, indicating the role of economic factors.

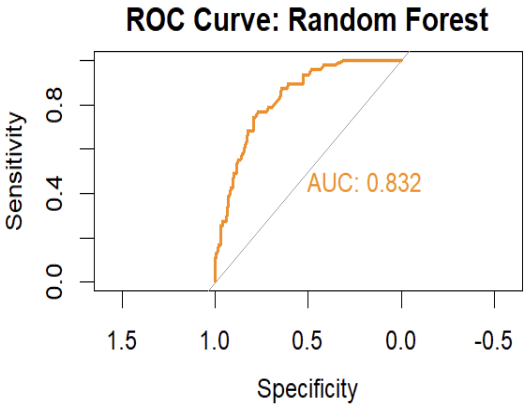


From a predictive perspective, the decision tree achieves a moderate performance (AUC = 0.717), which is lower than that of probabilistic models such as logistic regression and LDA. This reflects the bias–variance trade-off, where constraining model complexity improves interpretability at the cost of predictive accuracy.

From a risk learning perspective, misclassification costs are asymmetric, as failing to identify employees likely to leave can be more costly than false alarms. Although decision trees are not the most accurate models, their transparent structure provides actionable insights that support informed decision-making and targeted retention strategies.

Random Forest

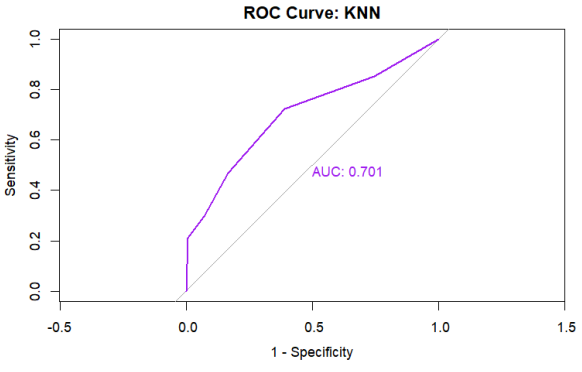
In addition to decision trees, we evaluated more complex machine learning models, namely **Random Forest**, which improves predictive performance by reducing variance through ensemble averaging. The model achieved higher predictive accuracy than the single decision tree, but lacked transparency and did not provide explicit decision rules. As a result, despite its improved performance, these models are less suitable when interpretability and explainability are required.



The ROC curves further highlight the differences between machine learning models. Random Forest achieves a high classification performance with an **AUC** of 0.832 and **Accuracy** : 0.856, substantially outperforming the single decision tree and approaching the performance of the best probabilistic classifiers. These results reinforce the bias–variance trade-off discussed in the course: Random Forest reduces variance at the cost of interpretability.

KNN

KNN is a model capable of capturing complex non-linear decision boundaries by relying on local neighborhood



information. KNN experienced a poor performance (optimal **k=13 and accuracy 0.854**) when running it on the full dataset (AUC=0.0.701) which is quite good but a bit less interpartable than random forest but had a poorer in performance when applying the selected variables (AUC=0.688), that is due to the fact that KNN relies on neighborhood and removing variables alters those relationships.

New Case Study:

We hypothesized 2 new employees, one with a large and the other with a low risk of attrition, and we will run the 3 best models from the above sections (Being, **LDA**, Penalized logistic regression and random forest) to predict their risk of attrition. And as expected the models were able to predict the high risk employee with probability (83~99%) and the low risk employee with probability of attrition 5%. The consistent high-confidence predictions across both linear and non-linear algorithms confirm that OverTime, JobRole, and Income are robust, reliable indicators.

```
new_employees <- data.frame(
  Age = c(23, 45),
  Gender = c("Male", "Female"),
  MaritalStatus = c("Single", "Married"),
  Education = c(1, 4), # 1=Below College, 4=Master
  EducationField = c("Marketing", "Life Sciences"),
  Department = c("Sales", "Research & Development"),
  JobRole = c("Sales Representative", "Manager"),
  JobLevel = c(1, 5),
  BusinessTravel = c("Travel_Frequently", "Non-Travel"),
  OverTime = c("Yes", "No"),
  MonthlyIncome = c(2500, 17500),
  DailyRate = c(400, 1200),
  HourlyRate = c(35, 90),
  MonthlyRate = c(5000, 22000),
  StockOptionLevel = c(0, 2),
  PercentSalaryHike = c(11, 22),
```

Conclusions:

For probabilistic models **Penalized logistic regression** proves to be the best model when it comes to predicting attrition, close to it are **LDA** and **Logistic regression**. With a poor performance in comparison from naive bayes although **naive bayes** might prove to be useful due its rigorous layout in a **risk sensitive** prediction in addition to that the **feature selection** resulted in a drop of performance for all predictive models.

Random Forest was the best performer from the machine learning models, achieving the highest overall accuracy and delivering good predictions for Attrition. It managed to handle the complexities of employee data without the high variance seen in the **Decision Tree** model, which overfitted the training data. The **KNN** model, optimized at **k=13**, achieved competitive accuracy (~84.3%) and successfully captured local clusters of at-risk employees. However, KNN's sensitivity makes it less useful for decisions compared to tree-based methods. Ultimately, Random Forest's stability and high predictive power make it the optimal choice for identifying potential attrition, while Logistic Regression remains crucial for explaining the underlying causes. **In similarity** to the Probabelistic models the **feature selection** was not useful to any of the models