DATA ANALYTICS FOR BUSINESS

Final Project Report

Submitted To:

Dr. Manisha Sharma

Group 7

I017 – Dushyant Sharma

I022 – Shriya Pandey

I024 – Guneet Singh

I047 – Shyam Patel

I053 – Sanjali Dawar

INDEX

ABSTRACT

Employee attrition is predictable under stable conditions, i.e wherein a set pattern can be derived from specific parameters impacting the employee and the organization consistently. Some of these parameters could be predicted like retirement age or unforeseeable for example company performance, management shakeup, external funding, etc.

However, who will leave, when and why, can be answered dependent on analytical models developed as a result of data analysis. Through predictive algorithms, organization's gain better understanding and can attempt preventive measures for employee attrition.

The analytical model works by grouping/ classifying employee profiles based on various attributes such as age, sex, marital status, education level, work experience, distance from home, etc. and create different levels of risk of attrition. Occasionally, other parameters like performance over the years, pay raise, work batch, educational institution are also contemplated.

Predictive Attrition Model aids in taking preventive measures and also in making better hiring decisions. Determining trends in the candidate's performance out of past data is important in order to predict the future trends, as well as to board new ones. Moreover, HR can use the employee data to predict attrition, the reasons behind it and can take appropriate measures to prevent it.

INTRODUCTION

In the present competitive economy and its developing innovative technological specialisation, acquisition, study and analysis of data are offering ascend to new knowledge. Data has become an essential asset for most companies across multiple sectors, including those connected to business processes. All types of organisations benefit from the adoption of new advancements and collection, management and analysis of data bring numerous benefits in terms of efficiency and competitive advantage. In fact, analysing large amounts of data can lead to improvements in decision-making processes, the achievement of pre-set corporate goals and better business competitiveness.

The significance of skills, knowledge and constant learning ability has demonstrated to be principal for businesses. The application of artificial intelligence in the field of HR permits organizations to transform data into knowledge by implementing predictive models: such models allow predictions on employees using past data collected by the company over years, thus reducing critical issues and optimising all HR activities.

In recent years, increasing attention has been centred around human resources, since the quality and abilities of employees establish a growth factor and a genuine upper hand to companies. Employee resignations are a reality for any organization. However, if the circumstances aren't handled properly, key staff members' departures can lead to a downturn in productivity. The organization may have to employ new people and train them on the tool that is being utilized, which is tedious. Most organizations are keen on knowing which of their employees are at the danger of leaving.

Organizations invest a lot of time and resources in employee recruiting and training, according to their strategic needs. Therefore, the employees (to a greater or lesser extent) represent a real investment for any organisation. When an employee leaves the company, the organisation is not only losing a valuable employee, but also the resources, specifically money and HR staff effort, that were contributed selecting and recruiting those employees and providing training to them for their related tasks. Consequently, the organisation must invest in recruiting, training and nurturing new staff to fill the vacant job positions. Training a new employee is a long and exorbitant process and it is of full interest of the organization to control and decrease the

employee attrition rate: attrition is basically defined as an employee resigning or retiring from a company. Additionally, highly motivated, satisfied, and loyal employees form the core of an organization and furthermore has an impact on the productivity of an organisation.

PROBLEM STATEMENT

To predict employee attrition with Machine Learning using KNIME and help the HR manager to retain the best talent by applying the correct strategies.

DATASET

The Dataset contains 35 columns with employee details. The link for the dataset is mentioned below:

https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv

Age	Attrition
Business Travel	Daily Rate
Department	Distance From Home
Education	Education Field
Employee Count	Employee Number
Environment Satisfaction	Gender
Hourly Rate	Job Involvement
Job Level	Job Role
Job Satisfaction	Marital Status
Monthly Income	Monthly Rate
Number of Company Worked	Over 18
Over Time	Present Salary Hike
Performance Rating	Relationship Satisfaction
Standard Hours	Stock Option Level
Total Working Years	Training Times Last Year
Work Life Balance	Year At Company

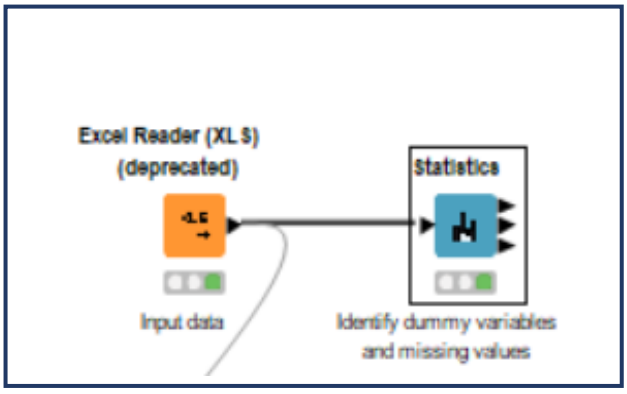
Years In Current Role	Year Since Last Promotion
Years With Current Manager	

The dataset contains target feature, distinguished by the variable Attrition: “No” represents an employee who did not leave the organization and “Yes” represents an employee who left the organization. This dataset will permit the machine learning system to learn from real data instead of through explicit programming.

DATA CLEANING

Data cleaning is one of the most important aspects of machine learning. It is usually complex and time consuming process. In fact, it has been calculated that on average this operation requires 60% of the time and energy spent on a data science project.

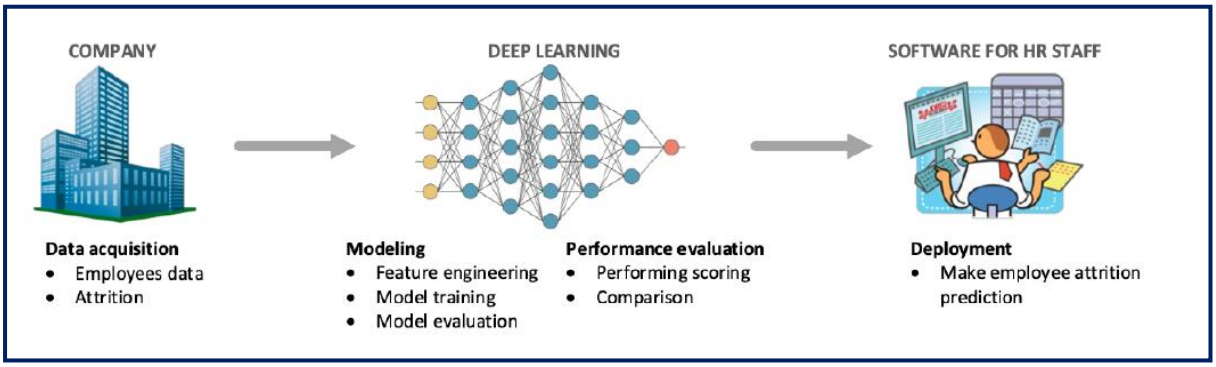
Firstly, the data relevant to the target was selected from the initial dataset; characteristics considered less significant or redundant were removed, such as the progressive number of the employee (1, 2, 3, . . .), flags marked over 18s (the variable “age”), hourly and weekly rates (monthly rates are also present). Then, “null” and “undefined” values or duplicate records were identified, since they could impact the correct training of the model and, result in, producing inaccurate predictions. Neither null or undefined values were found in any variable nor any duplicate observations emerged.



METHODOLOGY

To build a predictive model to predict employee attrition the below mentioned steps are followed:

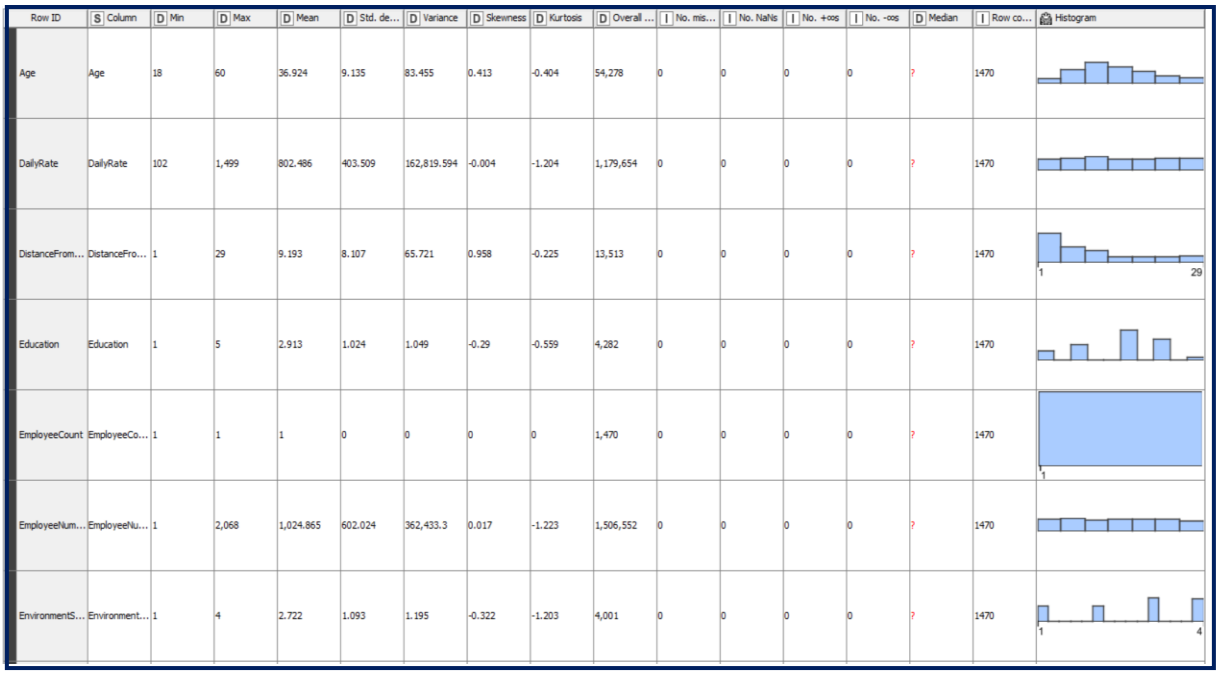
- Collection of employee dataset, which includes current and past employee observations
- Applying various data cleaning techniques to prepare the dataset
- Start with descriptive analysis of data to identify the key factors and trends that contribute to attrition
- Elaborate the dataset for training and testing phase and try various classification algorithms to process it
- Based on the results collected with test data, compare performance metrics of machine learning models and select which model best fits and gives the most accurate results for the given problem and release HR support software that implements the classification model.



DATA EXPLORATION

Here we generated the descriptive statistics of the dataset in order to observe the characteristics of all variables. We considered the below mentioned variables:

- Count, Unique,
- Mean, Standard deviation,
- Top, frequency,
- Minimum and maximum values (min/max)
- 25%/50%/75% percentile, etc.

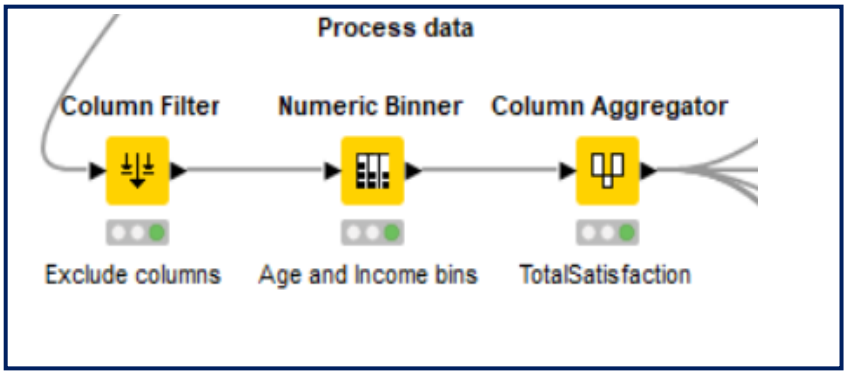


Row ID	Column	Min	Max	Mean	Std. de...	Variance	Skewness	Kurtosis	Overall ...	No. mis...	No. NaNs	No. +oos	No. -oos	Median	Row co...	Histogram
Age	Age	18	60	36.924	9.135	83.455	0.413	-0.404	54,278	0	0	0	0	?	1470	
DailyRate	DailyRate	102	1,499	802.486	403.509	162,819.594	-0.004	-1.204	1,179,654	0	0	0	0	?	1470	
DistanceFrom...	DistanceFro...	1	29	9.193	8.107	65.721	0.958	-0.225	13,513	0	0	0	0	?	1470	1 29
Education	Education	1	5	2.913	1.024	1.049	-0.29	-0.559	4,282	0	0	0	0	?	1470	
EmployeeCount	EmployeeCo...	1	1	1	0	0	0	0	1,470	0	0	0	0	?	1470	1 1
EmployeeNum...	EmployeeNu...	1	2,068	1,024.865	602.024	362,433.3	0.017	-1.223	1,506,552	0	0	0	0	?	1470	
EnvironmentS...	Environment...	1	4	2.722	1.093	1.195	-0.322	-1.203	4,001	0	0	0	0	?	1470	1 4

With Statisitcs view, it is evident that the variables *EmployeeCount, Over18*, and *StandardHours* have only a single value in the entire dataset; thus, we chose to remove them as they are not useful to predict the significance.

We then added a Column Filter node to exclude these variables along with variable EmployeeNumber variable as it is just an ID. Next, we generated some features in order to give more predictive power to our model, they are listed below:

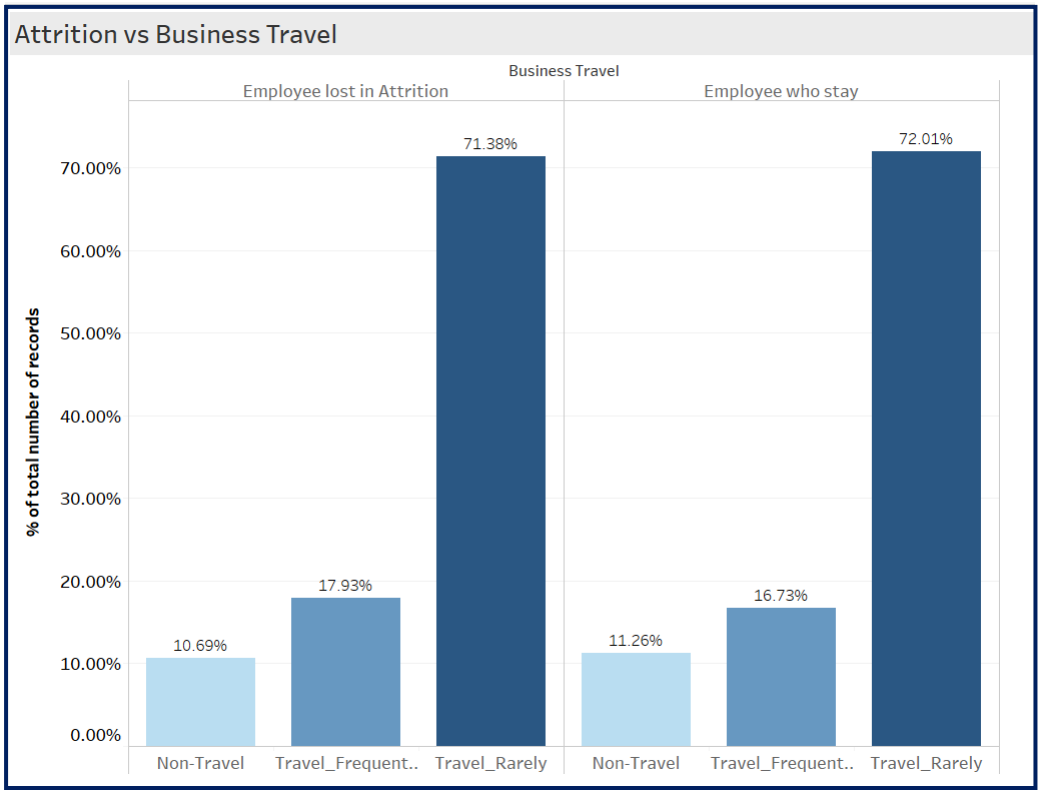
- We categorized Monthly Income, i.e from income between 0 to 6503 we labelled it as “low” and “high” if it was above 6503.
- We categorized Age, i.e for age between 0 to 24 we labelled it as “Young”, for 24 to 54 we labelled it as “Middle-Age” and above 54 we labelled it as “Senior”.
- We also aggregated the fields *RelationshipSatisfaction, EnvironmentSatisfaction, JobSatisfaction, JobInvolvement,* and *WorkLifeBalance* into a single feature (*TotalSatisfaction*) so as to provide an overall satisfaction value.



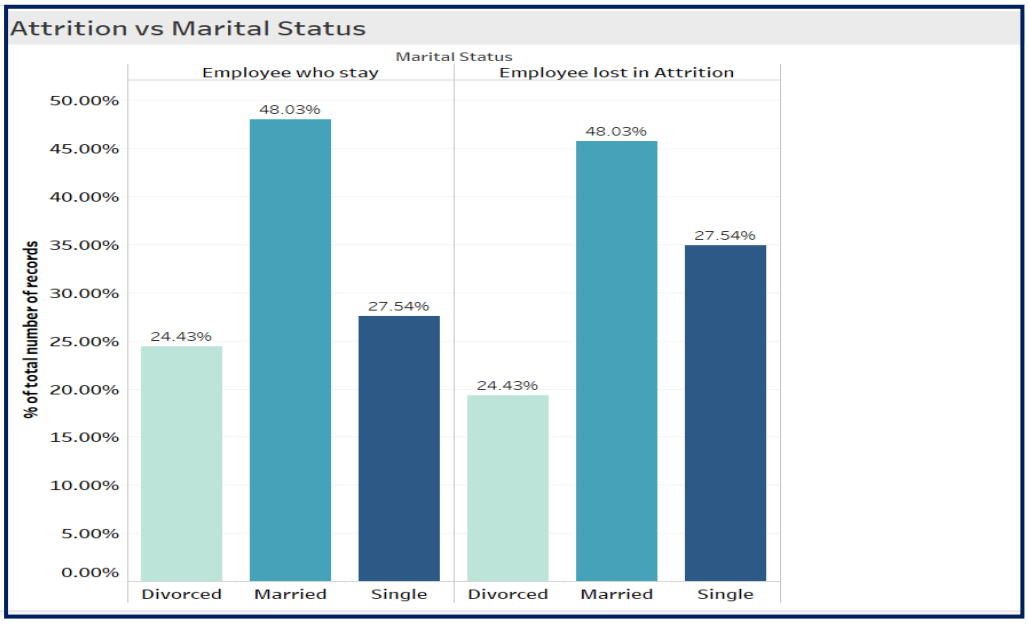
DATA ANALYSIS

At this point we analyzed the correlation between independent variables and the target variable, attrition.

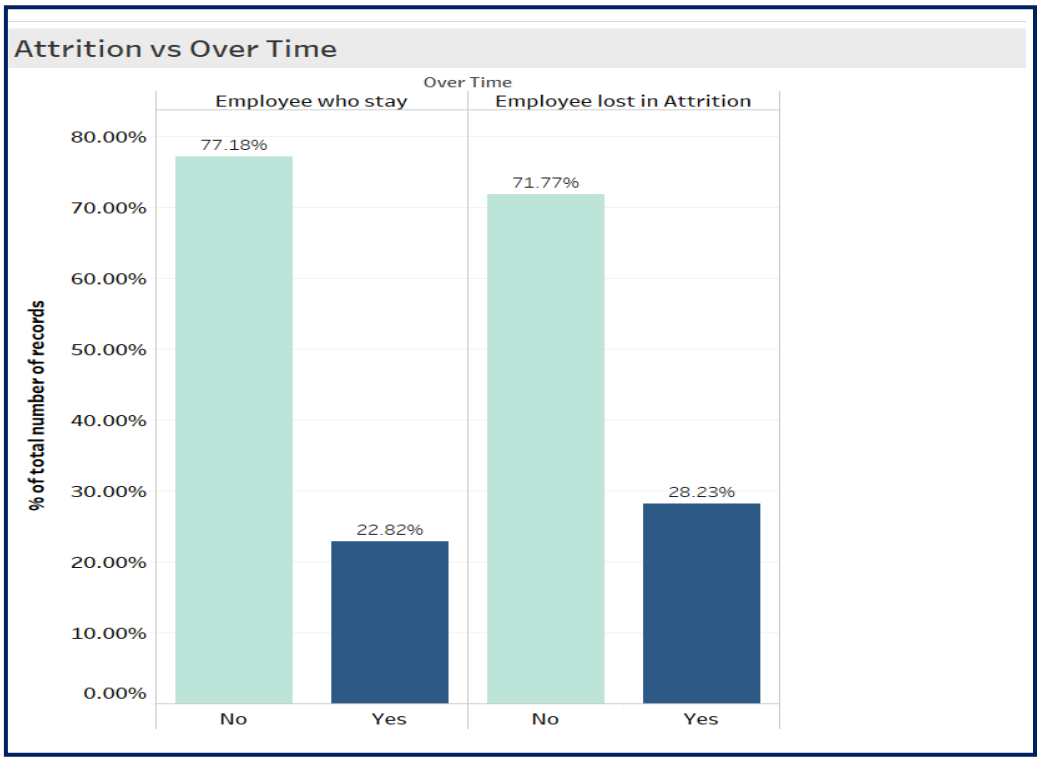
As a sample, we have listed a few below:



We can see that employees who travel a lot tend to leave the company more often; this will be an important variable for our model.



In a similar way, single people working overtime hours tend to leave at a higher rate compared to those who work regular hours and are married or divorced.

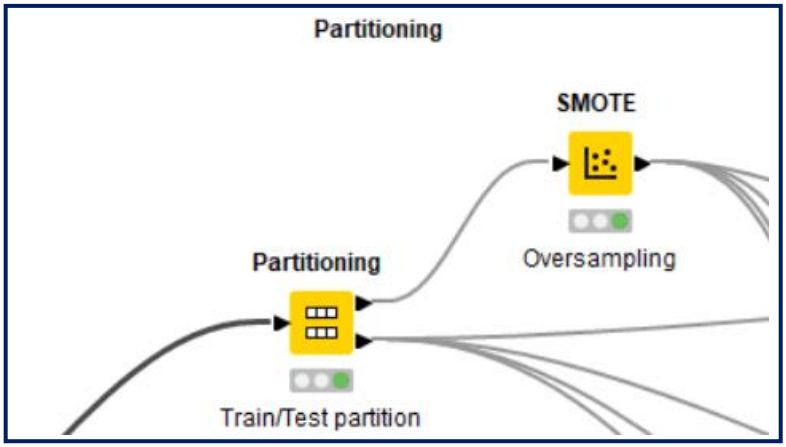


MODEL TRAINING

The dataset is imbalanced, when training models on such datasets, class imbalance impacts a learning algorithm during training by making the decision rules biased towards the majority class and improving the predictions on the basis of majority class in the dataset. To deal with this issue we have three methods listed below:

- Downsample the majority class/ Upsample the minority class.
- Assigning a larger penalty to wrong forecast from the minority class.
- Generating synthetic training examples.

Here we used the first approach of unsampling the minority class or downsample the majority class. For this, we split the dataset into training and test sets using an 80/20 split; 80% of data will be used to train the model and the rest 20% to test the accuracy of the model. Then we upsample the minority class, in this case the positive class so, we included the *Partitioning* & *SMOTE* nodes in KNIME.



After the partitioning and balancing, the data is ready to be the input of the ML models.

ML MODELS

Here we plan to train four different models: Naïve Bayes, Random Forest, Logistic Regression and Gradient Boosting.

Naïve Bayes: In the Naive Bayes model, there are NB classifiers that form a set of classification-based algorithms, which are based on Bayes' Theorem of Probability. It is used to describe a set of algorithms sharing a common feature, i.e. every pair of classified features is exclusively independent of one another.

The fundamental assumption in Naive Bayes is that each feature makes an equal and mutually exclusive contribution to the final outcome. This assumption is generally not applicable in real-world situations. The 'independence assumption' described is usually never true but the assumption generally works well in application.

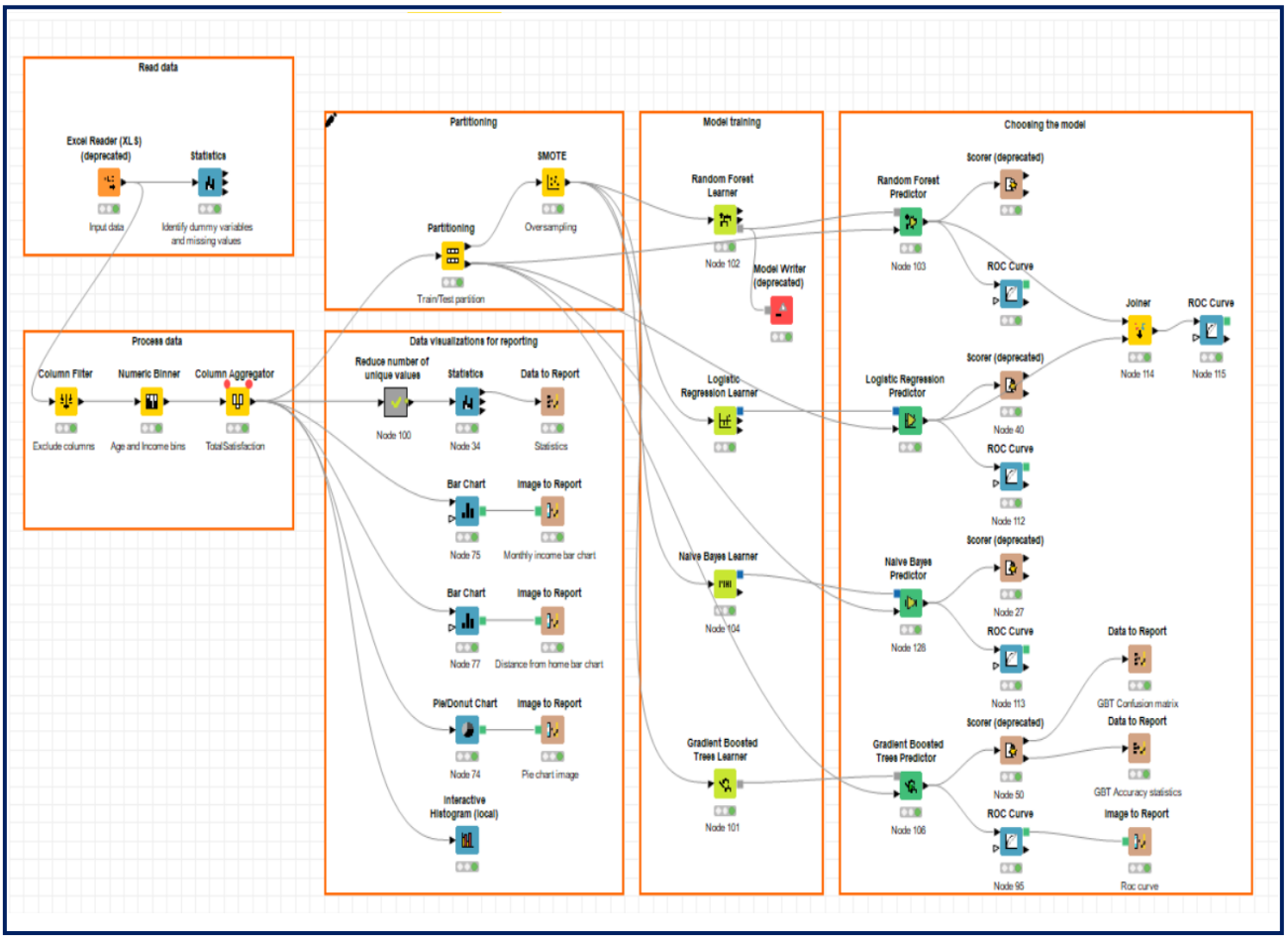
Random Forest: The widely used Random forest model is a supervised machine learning algorithm. The "forest" that is built, is an ensemble of several decision trees. It does not require hyper-parameter tuning; it is known for its high accuracy results. Popularly used for its simplicity and diversity. The Random Forest algorithm randomly selects observations & features and then constructs an ensemble of decision trees after which it averages the various results generated.

Logistic Regression: Logistic regression is a commonly used ML algorithm, which is used mainly to model the probabilities in classification problems (generally used when there are only two possible outcomes). It builds on the linear regression model which is then extended to accommodate classification-based problems.

Logistic regression is a widely prescribed solution to classification problems. As opposed to directly fitting a decision boundary/hyperplane, the model applies what is known as a 'logistic function' to fix the output of the equation between - 0 &1.

Gradient Boosting: Gradient Boosting ML model deals with converting weak learners into more powerful learners. In the process of boosting, each new tree is a superimposition on the modified version of our earlier dataset.

After the evaluation of the starting tree, the weightage of observations that are difficult to classify are increased incrementally and for those where classification is easy are lowered. The 2nd tree is later cultivated on this weighted data. The model is now a combination of (Tree 1 + Tree 2).

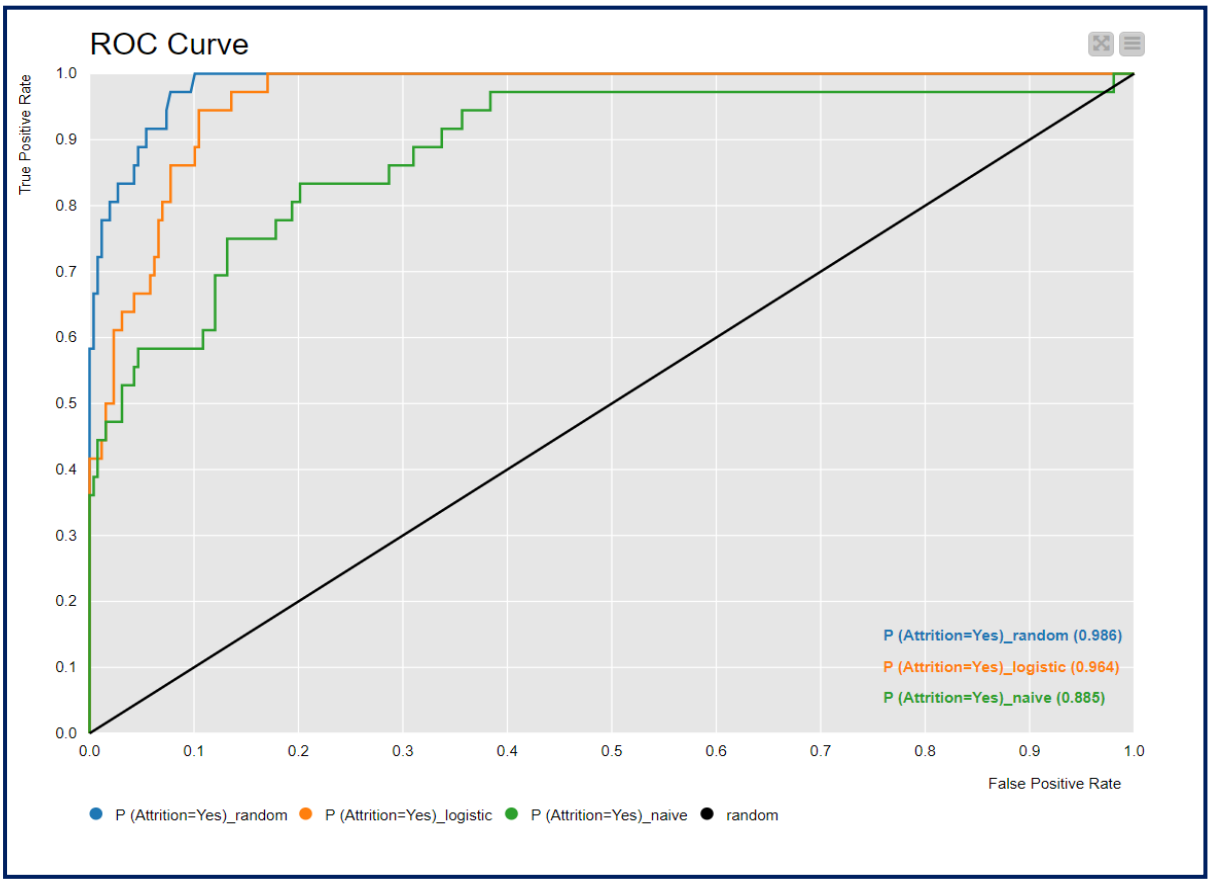


MODEL VALIDATION

Finally, after testing the models with the test set, we conclude that best model is the Random Forest (RF). We saved the trained model using the *Model Writer* node. The various parameters and its values in different ML model are listed below in the following table:

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	0.891	0.742	0.489	0.590	0.860
Naïve Bayes	0.827	0.462	0.511	0.485	0.820
Gradient Boost	0.881	0.667	0.511	0.578	0.879
Log. Regression	0.789	0.407	0.702	0.516	0.831

RF has the highest accuracy, which means it predicts 89.1% of the predictions accurately. In addition, and more importantly, it has the highest F1 score, which provides a balance among precision and recall, and is the measure to utilize if the sample is not balanced. The ROC curve is likewise a good measure to choose the best model. AUC stands for area under the curve, and the larger the AUC is the better is the model. Using the *ROC Curve* node, we visualized each ROC curve.

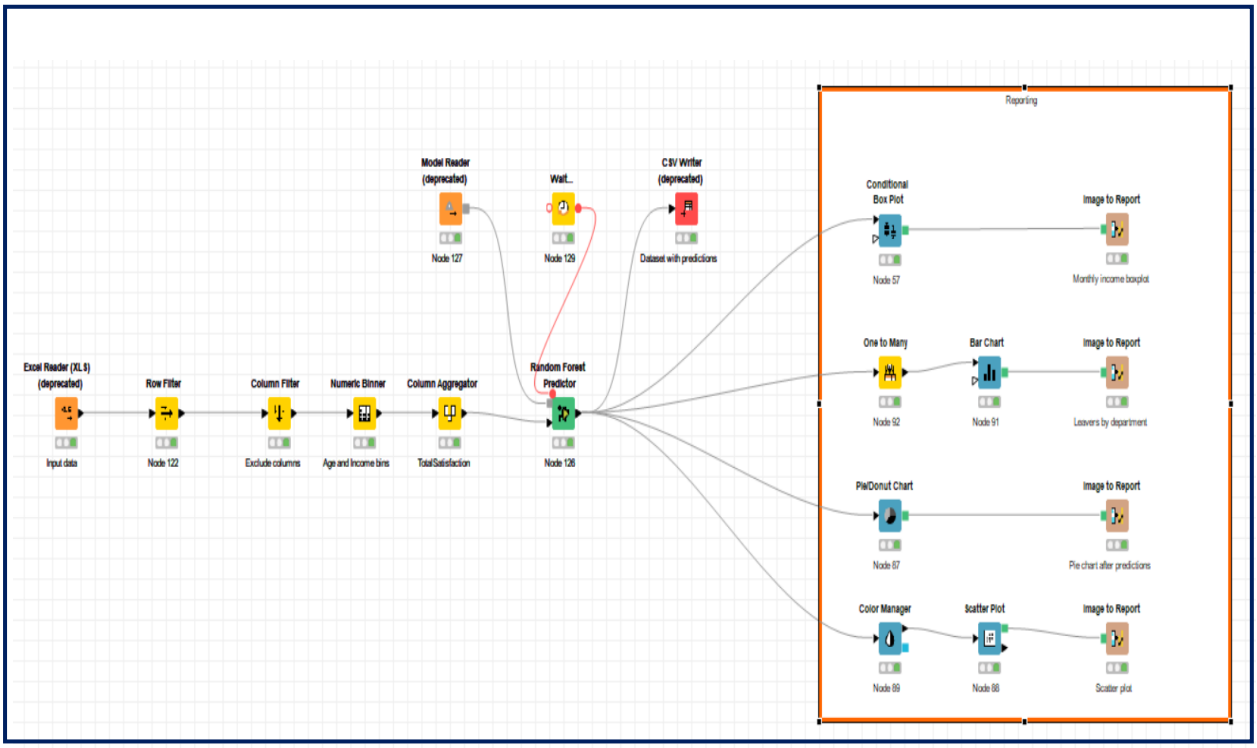


These values come from the confusion matrix, telling which of the predictions were correct (matrix diagonal) and which were not. We can also check the confusion matrix out of the RF model.

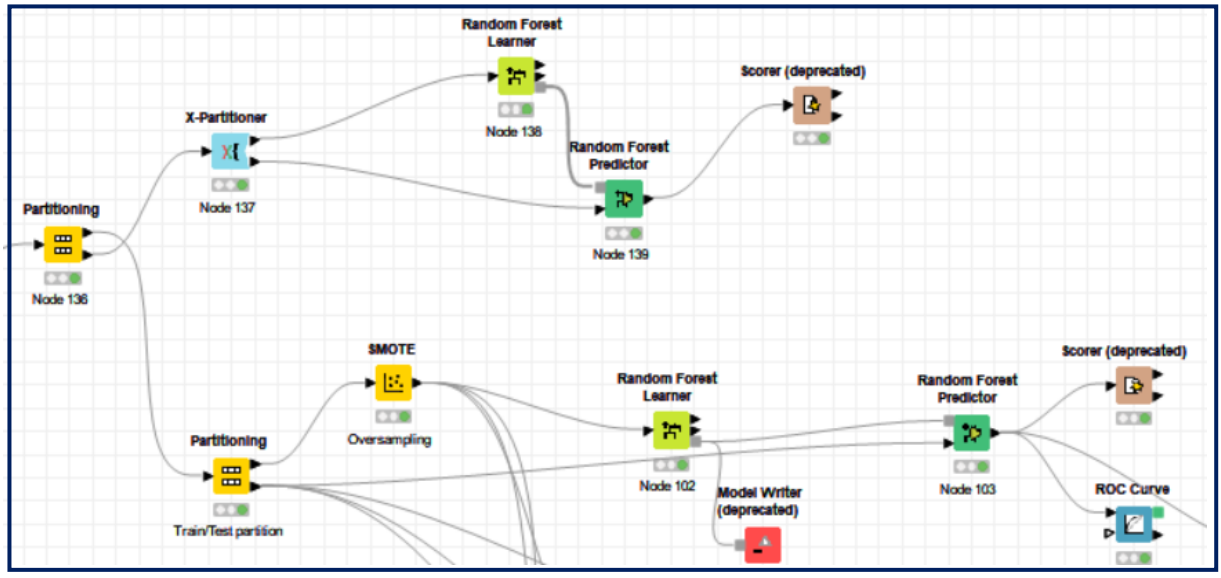
		Predicted Class: Yes	Predicted Class: No
Actual Class	Yes	23	24
	No	8	239

The Random Forest deals with the Bagging principle; it is an outfit of Decision Trees. The bagging method is utilized to increase the overall outcome by joining weak models. In the case of a classification problem, it takes the mode of the classes anticipated in the bagging process.

Once we have chosen the best possible model, the prediction of likelihood of attrition amongst the employees can be done accurately. We then applied the saved model to the current employees, and generate a new workflow that outputs the predictions.



To check the validity of the ML model we will use Partitioning and split the data into two equal parts and use it for validation of ML model “Random Predictor”. Then we used X-Partitioner to validate the model upto 10 validations and we check for the accuracy of the model it should be equivalent to the accuracy of the dataset used for training and testing.



From the accuracy statistics of both the scorer, we can see that the accuracy of both the model is approximately same.

Accuracy statistics - 0:9 - Scorer (deprecated)

File Edit Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
Yes	23	8	239	24	0.489	0.742	0.489	0.968	0.59	?	?
No	239	24	23	8	0.968	0.909	0.968	0.489	0.937	?	?
Overall	?	?	?	?	?	?	?	?	?	0.891	0.53

**for Validation of ML model*

Accuracy statistics - 3:134 - Scorer

File Edit Hilite Navigation View

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

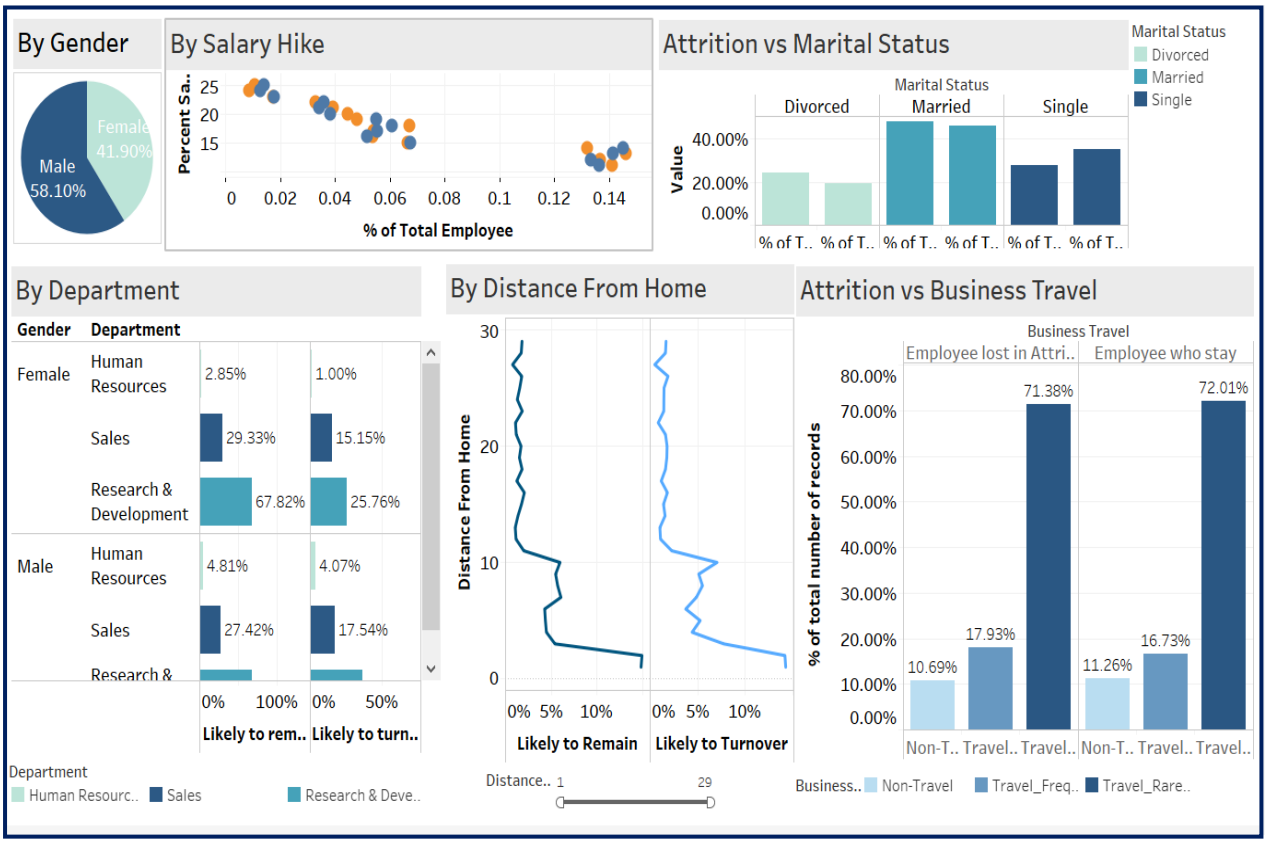
Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
Yes	2	0	62	10	0.167	1	0.167	1	0.286	?	?
No	62	10	2	0	1	0.861	1	0.167	0.925	?	?
Overall	?	?	?	?	?	?	?	?	?	0.865	0.251

**For Training and Testing*

DATA VISUALIZATION

Now we have the dataset with our current employees and their probability or likelihood of leaving the organization. For the Human Resource manager of the organization, what they need now is a dashboard in which we could perceive what we can expect regarding the future attrition in the organization and thus we will be able to adopt the correct strategy to retain the most talented employees.

So, we connected Tableau to this dataset and made a dashboard which contains analysis on the percentage of predicted attrition, business travel, analysis by gender, department, salary hike or by distance from home. We can further drill down to see the employees aggregated in each of these analyses. From the dashboard we can conclude that male employees who travel frequently, work in the HR department, have a low salary hike, and live far away from their place of work have a high probability of leaving the company.


By Gender
Female
41.90%
Male
58.10%
By Salary Hike
Percent Sa..
25
20
15
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
% of Total Employee
Attrition vs Marital Status
Marital Status
Divorced
Married
Single
Value
40.00%
20.00%
0.00%
% of T..
By Department
Gender
Department
Female
Human Resources
2.85%
1.00%
Sales
29.33%
15.15%
Research & Development
67.82%
25.76%
Male
Human Resources
4.81%
4.07%
Sales
27.42%
17.54%
Research &
0%
100%
50%
Likely to rem..
Likely to turn..
Department
Human Resourc..
Sales
Research & Deve..
By Distance From Home
Distance From Home
30
20
10
0
0%
5%
10%
Likely to Remain
Likely to Turnover
Distance..
1
29
Attrition vs Business Travel
Business Travel
Employee lost in Attri..
Employee who stay
% of total number of records
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
10.69%
17.93%
71.38%
11.26%
16.73%
72.01%
Non-T..
Travel..
Travel..
Business..
Non-Travel
Travel_Freq..
Travel_Rare..

CONCLUSION

We have detailed the different steps when executing advanced analytics use case in HR, employee attrition. We utilized the open-source tool KNIME to prepare the data, train different models, compare and choose the best out of them. With the model predictions, we made a dashboard in Tableau that would help any Human Resource manager to retain the best talent by applying the right strategies.

This work attempted to provide answers to some of the common questions of responsible human resources management:

- What are the key indicators that could signify that an employee could leave the organization?
- What is the probability that an employee can leave the organization?

For this, we applied some ML techniques in order to identify the elements that may contribute to an employee leaving the company and, above all, to predict the likelihood of individual employees leaving the company.

- Firstly, we assess the data statistically and then we classified it.
- Then the dataset was processed, and divided into the training and testing phase, guaranteeing the same distribution of the target variable.
- We selected four classification algorithms and, we carried out the training and validation phases, for each of them,.
- Now to further evaluate the algorithm's performance, the predicted results were collected and fed into the respective confusion matrices.

From these it was feasible to compute the essential metrics important for an overall evaluation (precision, recall, accuracy, f1 score, ROC curve, AUC, etc.) and to identify the most appropriate classifier to predict whether an employee was likely to leave the company or not. The algorithm that produced the best outcomes for the available dataset was the Random Forest Method: it revealed the best recall rate (0.489). The outcome obtained by the proposed automatic predictor demonstrate that the main attrition variables are monthly income, age, overtime, distance from home. The results obtained from the data analysis represent a starting point in the development of progressively efficient employee attrition classifiers. The

utilization of more numerous datasets or simply to update it periodically, the application of feature engineering to identify new significant characteristics from the dataset and the availability of additional information on employees would improve the overall knowledge of the reasons why employees leave their companies and, consequently, increase the time accessible to personnel departments to evaluate and plan the tasks required to alleviate this risk (e.g., retention activities, task redistribution, employee substitution, etc).

REFERENCES

- https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset?select=WA_Fn-UseC_-HR-Employee-Attrition.csv
- https://blog.floydhub.com/naive-bayes-for-machine-learning/
- https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab
- https://christophm.github.io/interpretable-ml-book/logistic.html
- https://analyticsindiamag.com/predictive-attrition-model/#:~:text=Predictive%20Attrition%20Model%20helps%20in,as%20to%20board%20new%20employees.