EMPLOYEE RETENTION BOOSTER MODELS

Data Mining Musketeers - Group 5
Aakash Shanavas
Prajakta Ingle
Purva Khandelwal
Rajesh Gowda Chiratahalli Prakash Gowda
Riya Bansode
Rujuta Thorat
Saranya Rajagopalan

CONTENT

BACKGROUND OF PROBLEM

"In order to build a rewarding employee experience, you need to understand what matters most to your people." – *Julie Bevacqua.*

There are only few people who begin and end their careers in one company. Most people move on after a time, or the company forces them to do so. People leaving organizations seems like a quite straightforward matter, but there is a lot to unpack with employee attrition.

Among all employee related problems, employee attrition is one of the key problems in today's scenario despite the changes in the external environment. In simplicity, attrition is the gradual reduction in number of employees through resignation, death and retirement.

This can be broadly categorized in following two ways:

- **Voluntary attrition:** When an employee chooses to leave the company, that is voluntary attrition. This can include any reason an employee leaves on their own accord, whether it's truly voluntary or not.
- **Involuntary attrition:** When the company decides to part ways with an employee, this is involuntary attrition. This can be through a position elimination, for example, due to reorganization or layoffs, for cause such as poor performance.

MOTIVATION FOR SOLVING THE PROBLEM

Attrition is the silent killer that can swiftly disable even the most successful and stable of organizations in a shockingly spare amount of time. While many companies put an emphasis on the expensive process of hiring and recruiting, there is not enough attention towards solving the issues that cause top talent to leave.

Most work we do in the field of human resource management is to help organizations understand what is most important to their employees, with the goal of making improvements to increase employee efficiency and engagement and strategize employee retention.

Each year companies hire multiple employees and invest time and money in training them. Companies spend huge capital in training programs to upskill their employees and make them ready for their specific business. But what happens when after all these investments, employee voluntarily resign, and the company cannot obtain return on their investment? When a well-trained and well-adapted employee leaves the organization for any of the reason, it creates vacuum in the organization. It creates a great difficulty for a Human resource personnel to fill this gap. Modern Human resource managers are taking serious steps to reduce the employee attrition rate and it has been a pivotal challenge for today's Managers. Here is where we wish to step in and lend them a helping hand in this seemingly uphill task.

SOLUTION METHODOLOGY

In this section, various methods or techniques used in the paper to predict employee attrition have been discussed along with their respective diagrams. Thus, with an impetus to accurately predict the attrition in employees and the reasons/factors responsible for the attrition we provide a detailed analysis of different approaches. The parameters which closely substantiate the build of attrition among the employees are explored and the numbers of parameters considered are crucial for concluding the result. Thus, a detailed evaluation of different data mining techniques is done to accurately predict the attrition and the parameters leading to employee attrition. The modelling process consists in selecting models that are based on various machine learning techniques used in the experimentation. In this case the following predictive models are used

TWO-CLASS BOOSTED DECISION TREE: Decision Trees are very popular amongst classifier algorithms due to their ease of interpretations and implementations. From the training data set, the algorithm builds a tree in which each node is an attribute, and the branches represent the corresponding attribute values. A problem faced by decision trees is instability in which small changes in the input training samples may cause dramatically large changes in output classification rules.

TWO-CLASS LOGISTIC REGRESSION: Logistic regression is a statistical method which is used to analyze a dataset which consists of one or more independent variables that determine an outcome. The outcome has only two values i.e., either yes or no or other binary outcomes like 1 or 0, true or false. It was developed to estimate the likeliness of a binary response based on more than one independent feature. Logistic regression permits anyone to say that the existence of a risk factor enhances the probability of a given case. Like every other regression analysis, logistic regression is also a predictive analysis.

TWO-CLASS NEURAL NETWORK: A neural network is a set of interconnected layers. The inputs are the first layer and are connected to an output layer by an acyclic graph comprised of weighted edges and nodes. To compute the output of the network for a particular input, a value is calculated at each node in the hidden layers and in the output layer. The value is set by calculating the weighted sum of the values of the nodes from the previous layer. An activation function is then applied to that weighted sum.

After identifying the objectives and adequately preparing and analyzing the dataset to be used, we proceeded with the design of the prediction model to identify employees that would potentially leave the company. In the construction phase of a model that implements a supervised learning algorithm, it was necessary to have a training-set available that consisted of instances of an already classified population (target), in order to train the model to classify new observations, which will constitute the test-set. Then, the model must be trained on a consistent number of observations in order to refine its prediction ability. The precision of the machine learning algorithms increases with the amount of data available during training. Ideally, one would have two distinct datasets: one for training and a second to be used as a test. The original dataset was divided into two parts with an 80:20 ratio, one used for training, and one used for testing.

- Train set contained **80%** of the dataset. This information was dedicated to the training phase in order to allow the model to learn the relationships hidden in the data; the train-set contains 1176 observations.
- Test set contained the remaining **20%**. This information was dedicated to the test and validation phase in order to evaluate the general performance of the model and to calculate errors between predicted and actual results; the test-set contains 294 observations.

EVALUATION METRICS

For classification problems, **Area Under Curve (AUC)** can be trusted for performance measurement. It is considered as one of the most important evaluation metrics for assessing any classification model's performance. This report uses Area Under Curve (AUC) as the evaluation criteria for the above-mentioned models. Different models are implemented, tested and then evaluated based on the AUC score. When it comes to evaluating how well a model performs there are multiple metrics that can be used. To choose the evaluation metric to best evaluate your model, it is vital that you understand what each metric calculates.

ACCURACY: Accuracy is an evaluation metric that allows you to measure the total number of predictions a model gets right. Accuracy will answer the question, what percent of the model predictions were correct? Accuracy looks at True Positives and True Negatives. The formula for accuracy is below:

Accuracy = (True Negative + True Positive) / (True Positive + False Positive + True Negative + False Negative)

PRECISION: Precision evaluates how precise a model is in predicting positive labels. Precision answers the question, out of the number of times a model predicted positive, how often was it correct? Precision is the percentage of your results which are relevant. Precision is a good evaluation metric to use when the cost of a false positive is very high, and the cost of a false negative is low. The formula for precision is below:

Precision = True Positive / (True Positive + False Positive)

RECALL: Recall calculates the percentage of actual positives a model correctly identified (True Positive). When the cost of a false negative is high, you should use recall. The formula for recall is below:

Recall = True Positive / (True Positive + False Negative)

DESCRIPTION OF YOUR DATASET

We have used the "IBM HR Analytics Employee Attrition & Performance" data set available on Kaggle. This is a fictional data set created by IBM data scientists. We have associated a fictional Organization "WR Health" for this data.

- This Employee Engagement Data which is created for two years has been useful in unearthing various insights that follow through our project. The data set has a total of 35 columns, 34 being predictor variables, while "Attrition" being the response variable.
- We weighed in most possibilities for using the data at hand. This data can be used to create demographic of staff who resigned which will provide information about attrition in different department, locations, job position, age, and educational background.
- Likewise exit Interview data can be used to analyze the reasons behind voluntary resignation, thereby informing about the employee's workplace experience, job satisfaction level, work life balance rate, satisfaction with regards to the compensation program.

EXPLANATORY VARIABLES: Below is the data dictionary of the dataset which enlists all predictor variables:

Variable	Type	Description
Age	Numerical-Discrete	Age of an employee
BusinessTravel	Categorical	Travel pattern of Employee
DailyRate	Numerical-Discrete	Rate in dollars per hour for an employee.
Department	Categorical	Department under which Employee Works
DistanceFromHome	Numerical-Discrete	Distance between employee's home and office
Education	Categorical	The education background of employee 1- "Below College", 2- "College", 3- "Bachelor", 4- "Master", 5- "Doctor"
EducationField	Categorical	Field of Specialization
EmployeeCount	Unitary	It gives count of Employee.
EmployeeNumber	Numerical-Discrete	Employment id of Employee
EnvironmentSatisfaction	Categorical	How comfortable employees are working in the office environment. 1- "Low", 2- "Medium", 3- "high", 4- "Very high"
Gender	Binary	Employee's Gender
HourlyRate	Numerical-Discrete	Hourly earnings in Dollars
JobInvolvement	Categorical	Employee's involvement in the assigned work 1- "Low", 2- "Medium", 3- "high", 4- "Very high"
JobLevel	Categorical	Level of Employee's job in Organization
JobRole	Categorical	Role that employee caters as part of his job
JobSatisfaction	Categorical	Employees satisfaction to the tasks assigned. 1- "Low", 2- "Medium", 3- "high", 4- "Very high"
MaritalStatus	Categorical	Marital Status
MonthlyIncome	Numerical-Discrete	Monthly income of employee in dollars
MonthlyRate	Numerical-Discrete	Income provided as part of salary
NumCompaniesWorked	Numerical-Discrete	Total companies the employee worked for
Over18	Unitary	If employee is above 18 or not
OverTime	Numerical-Discrete	Hours spent over time on biweekly basis
PercentSalaryHike	Numerical-Discrete	Salary hike in percentage for employee
PerformanceRating	Categorical	Employees rated based on their performance. 1- "Low", 2- "Good", 3- "Excellent", 4- "Outstanding"
RelationshipSatisfaction	Categorical	Rated based on relationship satisfaction with other employees. 1- "Low", 2- "Medium", 3- "high", 4- "Very high"
StandardHours	Unitary	Standard biweekly hours of company
StockOptionLevel	Categorical	Stock which employee can purchase
TotalWorkingYears	Numerical-Discrete	Total work years including all companies
TrainingTimesLastYear	Numerical-Discrete	Number of months spend training last year
WorkLifeBalance	Categorical	How employee manages time at work with life 1- "Bad", 2- "Good", 3- "Better", 4- "Best"

YearsAtCompany	Numerical-Discrete	Total years worked in Company
YearsInCurrentRole	Numerical-Discrete	Number of years win current role
YearsSinceLastPromotion	Numerical-Discrete	Number of years since last promotion
YearsWithCurrManager	Numerical-Discrete	Number of years worked under current Manager

RESPONSE VARIABLE:

Attrition: Defines the predicted variable as a Boolean output "Yes/No"

DATA EXPLORATION:

- Variable Identification: Step 1 in data exploration where we identify the input, output variables, data type and whether its categorical, binary, discrete or continuous. As you can see in data description this step is covered. We did Univariate Analysis to analyze each variable independently. This will be done to find out characteristic of each variable like how disperse or tight the data is, missing values, or outliers. Tools like bar chart, boxplot, histogram can be used for the analysis.
- We analyzed variables bivariate way to find out if there exists any linear or non-linear relation between two variables and the strength of correlation between these two variables. Highly correlated variables that are misleading the output of our model are removed. Correlation

(Figure1) heatmaps were used to find the same. Vertical scale on the right, shows a transition from **blue** (least correlated) to **brown** (most correlated).

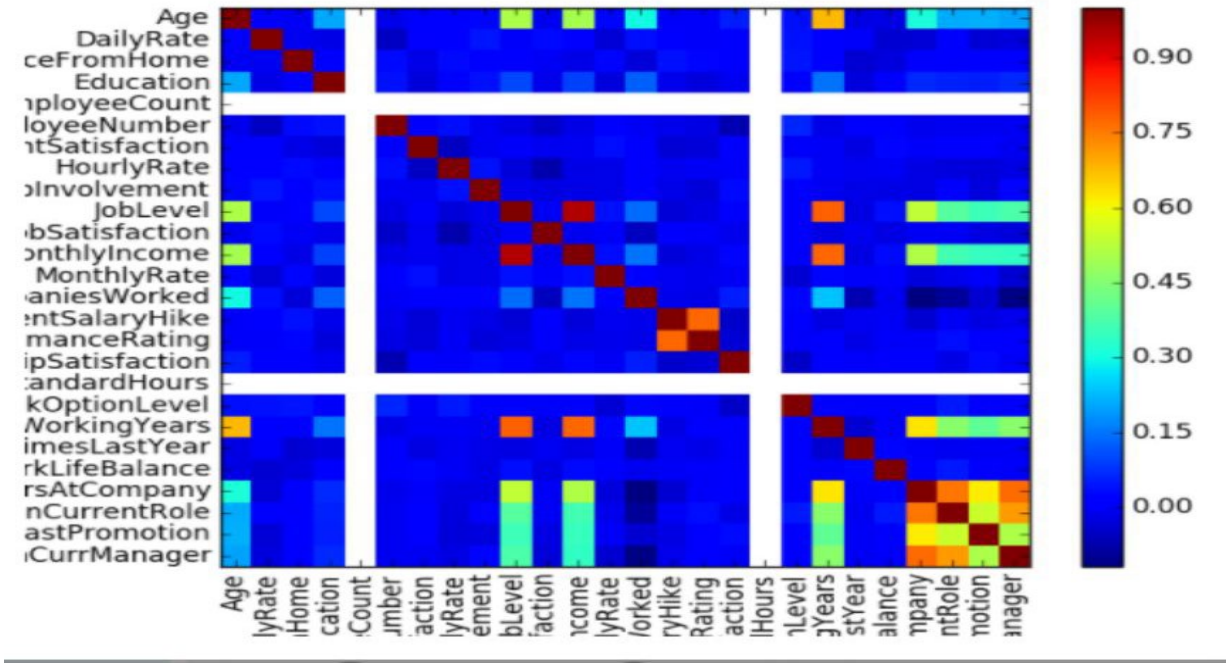

Figure 1

DATA MODELLING

PREDICTIVE MODEL

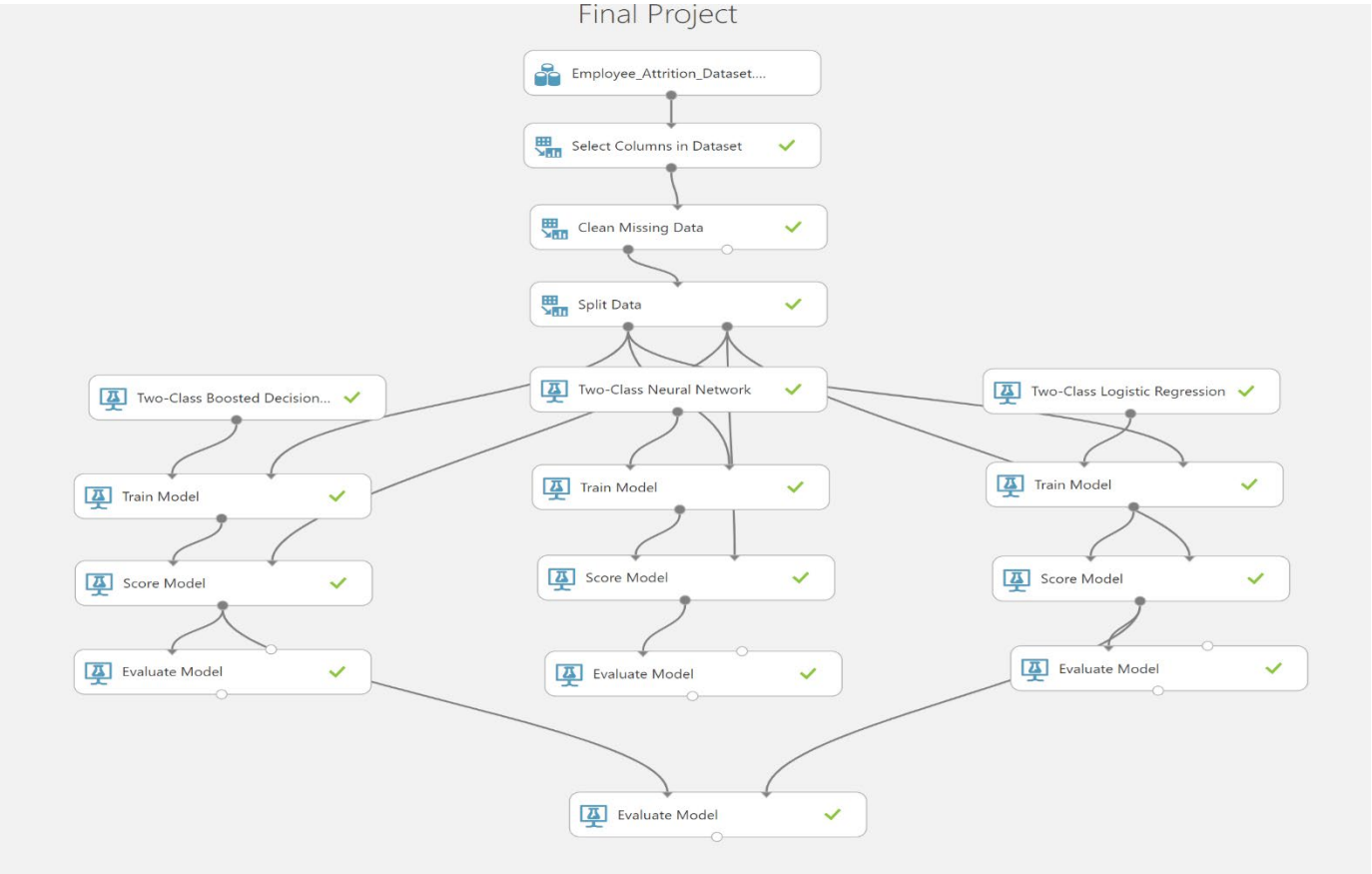

Figure 2

DATA CLEANING

Data preparation is one of the most vital facet of machine learning; it is usually complex and often requires rather a lot of time. We know missing values can create biasness in data or mislead the results of the model, as our data set had some missing values in some columns, so to clean the data we have used Clean Missing Data. The "null" and "undefined" values were identified and removed, since they could inadvertently influence the correct training of the model and, consequently, produce inaccurate predictions.

SPLIT DATA

Using Split Data, we have split our data into 80% as training data and 20% as testing data.

ALGORITHMS IMPLEMENTED

We have used below three algorithms to predict the results using 2 different experiments.

I. Selecting all the columns from the Data set.
II. Selecting Columns based the least correlation values from the dataset: - Daily rate, Distance from Home, Education, Employee Number, Environment satisfaction, Hourly Rate, Job Involvement, Job satisfaction, Monthly rate, Relationship Satisfaction, Stock option, level, Total working Years, Training Times Last Year, Work life balance.

TWO-CLASS BOOSTED DECISION TREE: By executing the experiment for two set of parameters a)Default values b)Specified Parameter changed values described in the table we get the below results

Algorithm	Parameters	TP	TN	FP	FN	TP+TN	Total = TP+TN+FP+FN	Accuracy	Misclassification Rate	Precision	TPR= Recall= Sensitivity	TNR= Specificity	F1 Score	AUC
All Columns														
Two-Class Boosted Decision Tree	Default - No of trees constructed -100	17	241	8	28	258	294	0.878	0.122	0.68	0.378	0.680	0.468	0.762
	Changed - No of trees constructed - 120	17	243	6	28	260	294	0.884	0.116	0.739	0.378	0.739	0.500	0.773
Selected columns														
Two-Class Boosted Decision Tree	Default - No of trees constructed -100	9	236	13	36	245	294	0.833	0.167	0.409	0.200	0.948	0.269	0.673
	Changed - No of trees constructed - 120	17	243	6	28	260	294	0.884	0.116	0.739	0.378	0.976	0.500	0.773

TWO-CLASS NEURAL NETWORK: By executing the experiment for two set of parameters a)Default values b)Specified Parameter changed values described in the table we get the below results

Algorithm	Parameters	TP	TN	FP	FN	TP+TN	Total = TP+TN+FP+FN	Accuracy	Misclassification Rate	Precision	TPR= Recall= Sensitivity	TNR= Specificity	F1 Score	AUC
All Columns														
Two -Class Neural Network	Default - No of hidden nodes - 100 Learning rate - 0.1	17	215	34	28	232	294	0.789	0.211	0.333	0.378	0.333	0.354	0.754
	Changed - No of hidden nodes - 150 Learning rate - 0.5	17	228	21	28	245	294	0.833	0.167	0.447	0.378	0.447	0.410	0.754
Selected columns														
Two -Class Neural Network	Default - No of hidden nodes - 100 Learning rate - 0.1	22	219	23	30	241	294	0.820	0.180	0.423	0.489	0.905	0.454	0.763
	Changed - No of hidden nodes - 150 Learning rate - 0.5	17	225	24	28	242	294	0.823	0.177	0.415	0.378	0.904	0.395	0.700

TWO-CLASS LOGISTIC REGRESSION: By executing the experiment for two set of parameters a)Default values b)Specified Parameter changed values described in the table we get the below results

Algorithm	Parameters	TP	TN	FP	FN	TP+TN	Total = TP+TN+FP+FN	Accuracy	Misclassification Rate	Precision	TPR= Recall= Sensitivity	TNR= Specificity	F1 Score	AUC
All Columns														
Two-Class Logistic Regression	Default	18	247	2	27	265	294	0.901	0.099	0.900	0.400	0.900	0.554	0.824
Selected columns														
Two-Class Logistic Regression	Default	2	249	0	43	251	294	0.854	0.146	1.000	0.044	1.000	0.085	0.751

COMPARISION OF MODELS

On comparing the two best fit models i.e. Two-Class Logistic Regression and Two-Class Boosted Decision Tree, we get the below ROC curves, both the curves lean more towards the right and the AUC is 0.762. The accuracy and precision given by this model is 0.872 and 0.68 respectively. Also, the specificity is 0.680 with an average recall of 0.378 which is less than the recall rate of logistic regression model.

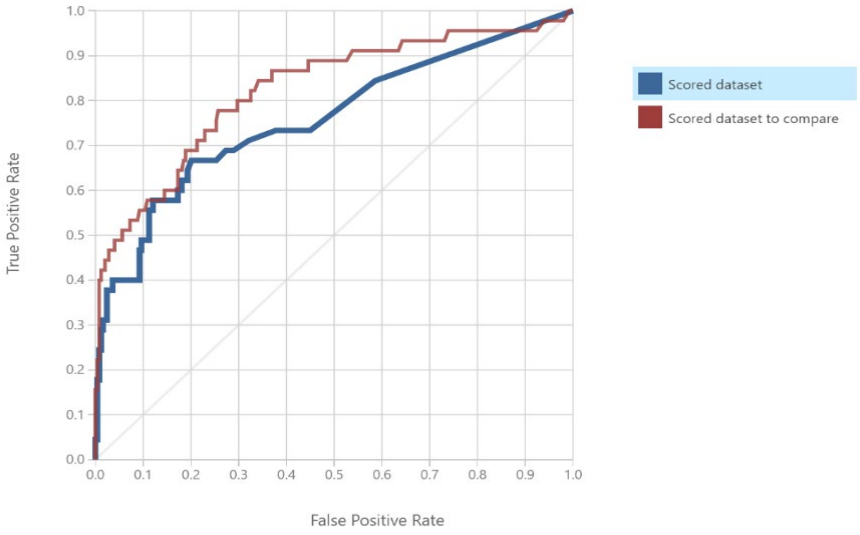

Figure 3

COMPARISION OF MODELS

In this experiment, we are interested in predicting the greatest number of people who could leave the organization. Two-Class Logistic Regression has helped the model to improve reasonably giving higher AUC, Accuracy and Precision which are better metric to justify because it indicates the classification performance of the model. Thus, the Two-Class Logistic Regression algorithm is identified as the best classification algorithm which can achieve the objective of the analysis.

SUMMARY SHEET SHOWING THE RESULTS OF ALL EXPERIMENTS

Algorithm	Parameters	Accuracy	Precision	Recall	F1 Score	AUC
All Columns						
Two-Class Boosted Decision Tree	Default - No of trees constructed -100	0.878	0.680	0.378	0.468	0.762
	Changed - No of trees constructed - 120	0.884	0.739	0.378	0.500	0.773
Selected columns						
Two-Class Boosted Decision Tree	Default - No of trees constructed -100	0.833	0.409	0.200	0.269	0.673
	Changed - No of trees constructed - 120	0.884	0.739	0.378	0.500	0.773
All Columns						
Two -Class Neural Network	Default - No of hidden nodes - 100 & Learning rate - 0.1	0.789	0.333	0.378	0.354	0.754
	Changed - No of hidden nodes - 150 & Learning rate - 0.5	0.833	0.447	0.378	0.410	0.754
Selected columns						
Two -Class Neural Network	Default - No of hidden nodes - 100 & Learning rate - 0.1	0.820	0.423	0.489	0.454	0.763
	Changed - No of hidden nodes - 150 & Learning rate - 0.5	0.823	0.415	0.378	0.395	0.700
All Columns						
Two-Class Logistic Regression	Default	0.901	0.900	0.400	0.554	0.824
Selected columns						
Two-Class Logistic Regression	Default	0.854	1.000	0.044	0.085	0.751

CONCLUSION AND RECOMMENDATIONS:

Employee turnover is inevitable, but if a company can find some ways to retain the proficient employees and let go of the unfit ones, it would help the company save huge financial losses and improve productivity. For any firm employee retention program is a highly important, but it is not an easy task. Many companies struggle to implement an efficient retention program.

Our model aims at using machine learning techniques to solve the employee turnover problem. Our model trying to understand patterns and trend that indicates the factors which are contributing the most towards employee leaving the firm. This can be helpful in predicting future chances of any employee leaving and based on his merits, actions can be taken towards his retention. We used association rules to discover frequently occurring item sets. Those item sets which are frequent and those confidence is above 60% were used to find reveal rules which will help us recommend some solutions to the firm.

I. **Rule 1: {educationalField=Marketing, OverTime=Yes, StockOption =0 } --> {Attrition=Yes}**
Rule2: {OverTime=Yes, PerformanceRating=4, StockOption =0 } --> {Attrition=Yes}

We can see those employees having marketing educational background, who are working overtime and not given any stock options, tend to leave. The reason behind this might be because they have marketing knowledge and can see the advantages of buying stock at 'grant price'. Also, they are performing overtime. Now what is Stock Option variable in our data? This is where an employee is given an option to buy certain number of shares of the firm at a preset price called as 'grant price'. If the performance ratings of these employees are high, then the firm will be willing to retain such employees. So, the firm can provide stock options to the employees of marketing department who are working overtime some compensation in form of Stock Options.

II. **Rule1 : { JobLevel=1, JobSatisfaction=1, YearsWithCurrManager=0 } --> {Attrition=Yes}**
Rule 2: {JobSatisfaction=1,StockOption=0,YearsWithCurrManager=0} --> {Attrition=Yes}
Rule3:{EnvironmentSatisfaction=1,TrainingTimesLastYear=2,YearsInCurrRole=0, StockOption=0} --> {Attrition=Yes}

From the above rules it is observed that employee whose job satisfaction or environment satisfaction is low from previous role or working under previous manager and are now assigned to new role or start to work under new manager are more likely to leave the company. This is an evidence for the obvious thought process of many employees who work for certain years in a project and is not satisfied. When he/she is offered new role or offered to work under new manager, they might look at it as a bad pattern of getting into yet another project in same company and remaining unsatisfied. And as a caution to avoid this pattern they might resign. Therefore, a

company can try to understand the needs of their employee. The company can conduct employee satisfaction surveys and calculate Employee satisfaction index to understand employee needs and their perspective of job/environment satisfaction. Company can have activities that the employees might enjoy, which in return boosts their productivity. Also, whenever an employee is assigned a new job role or manager, team building exercises can help. Company should always have a check on employee engagement to understand the level of their satisfaction.

III. **Rule1:{Department=Sales, JobLevel=1, YearsInCurrentRole=0} --> {Attrition=Yes}**
Rule2: {Department=Sales, MaritalStatus=Single, YearsWithCurrManager=0} --> {Attrition=Yes}
Rule 3: {JobRole=Sales Representative, YearsInCurrentRole=0} --> {Attrition=Yes}
Rule 4: {BusinessTravel=Travel_Frequently, Department=Sales, JobLevel=1} --> {Attrition=Yes}

From the association rules discovery, it seems that for our firm employee in sales department with less than 2 years of experience are more likely to leave the firm. Therefore, well performing salespeople can be given rewards and travel allowance or awards for recognizing their challenging work on achieving their target to provide job satisfaction. The Joblevel is 1 and hence the HR department can look if their need is higher package which HR can offer based on Merit.

IV. **Rule1: {Education=3, MaritalStatus=Single, OverTime=Yes} --> {Attrition=Yes}**
Rule 2:{Gender=Male, JobRole=Lab Technician, MaritalStatus=Single, YearLastPromotion=0} --> {Attrition=Yes}
Rule 3: {JobInvolvement=2, MaritalStatus=Single, OverTime=Yes} --> {Attrition=Yes}
Rule4: {MaritalStatus=Single, NumCompaniesWorked=1, OverTime=Yes} --> {Attrition=Yes}
Rule5: {Gender=Male, MaritalStatus=Single, OverTime=Yes} --> {Attrition=Yes}

Here we can see that if Unmarried and new to the Job, they try to learn more and stretch by doing overtime since it does not affect their family. But soon they would like to settle and start a family for self then they would want work life balance and hence we can retain such employees by bringing in shift culture if the work demands on call support for longer hours and the productivity will be improved by bringing this kind of initiatives which provide the employees work life balance without impacting their rating and increasing productivity to the company.

V. Association was found between employees with low performance rating, job satisfaction, less work life balance and monthly income. We can infer that the employee is not fit for the job and therefore cannot perform, so the above. Or because an employee is not satisfied and cannot balance professional and personal life, his performance is low. In both cases, the management can further evaluate the case and talk to his manager to find the exact cause. If the employee is not a fit, company do not have to put efforts to retain such employees and save capital on their income.

REFERENCES

https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

https://www.researchgate.net/publication/322896996_Employee_Attrition_and_Employee_Retention-Challenges_Suggestions

https://laptrinhx.com/news/employee-attrition-all-you-need-to-know-qb2LrRK

https://www.mdpi.com/2073-431X/9/4/86/pdf

https://www.business-science.io/business/2017/09/18/hr_employee_attrition.html

https://towardsdatascience.com/people-analytics-with-attrition-predictions-12adcce9573f

https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/