

# Project Report for Employee Turnover Prediction

## **Submitted by:**

Aayushi Vora

Shruti Sankolli

Akhil Sreehari

Pravesh Humane

## **Contents**

1. Abstract
2. Introduction
3. Data Exploration
4. Data Preprocessing
5. Feature Ranking
6. Correlation
7. Variability in the Data
8. Cross Validation Technique
9. Predicting Probability of Left using Classification Technique
10. Summary of Models Result
11. Inferences
12. Retention Strategies
13. References

## 1. Abstract

Employee turnover is a key issue for any organization because of its adverse impact on workplace productivity and long-term growth strategies. Accurate predictions enable organizations to take actions for retention planning of employees. This study focuses on providing a data mining model that can accurately estimate future attrition. An example of employee turnover prediction model leveraging classical classification techniques is developed. Model outputs are then discussed to design and test employee retention policies. The data models and the retention strategies constitute the main value of this study.

## 2. Introduction

Employee Retention is one of the primary measures of the health of an organization. Key employee retention is vital to the long term health and success of any business. It is in the best interest for organizations to account for future attrition and develop sound employee retention strategies as employee turnover has negative impacts on issues ranging from workplace morale and productivity, to disruptions in project continuity and to long term growth strategies. This study proposes a data mining solution to target this problem.

In this study, we apply 8 classification techniques on a sample employee dataset and provide a comparative study of the performance of these techniques. We then analyze the results to generate inferences from the data and propose retention strategies. We supplement the classification results with visualizations that provide a better understanding of the variability in the data and help us make relevant inferences.

The intention of this study is to equip management with a sound data mining solution to constrict employee turnover and provide them insights to facilitate better decision-making.

## 3. Data Exploration

### Predictors and Target

This study uses a sample HR Analytics dataset from Kaggle. The data set has 15000 rows and provides 9 predictors that we analyze to predict the target 'Left'.

The predictors can be explained as:

- Satisfaction Level: A numeric predictor that takes values between 0 and 1 and describes the average satisfaction level of the employee.
- Last\_evaluation: A numeric predictor that describes the last evaluation value for the employee.
- Number\_Project: A numeric predictor that describes the number of projects completed while at work.
- Average\_Monthly\_hours: A numeric predictor that describes the average monthly hours at workplace.
- Work\_Accident: A factor predictor that takes value 0 if the employee did not have an accident and value 1 if the employee had an accident while at work.
- Promotion\_Last\_5years: A factor predictor that takes value 0 if the employee did not have an accident and value 1 if the employee had an accident while at work.
- sales: A categorical variable denoting the department the employee is a part of.
- salary: A factor predictor with values low, medium and high describing the salaries of the employees.
-

The target can be explained as :

- Left: A factor variable that takes the value 0 if the employee has not left the organization and value 1 if the employee has left the organization.

The figure below shows the sample 5 rows of this dataset.

	satisfaction_level	last_evaluation	number_project	average_monthly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low

#### 4. Data Preprocessing

It is important to process the data so as it to make it fit for use. The HR Analytics dataset is clean with no null values, however it needs some cleaning and renaming so as to make it fit for the data mining algorithms.

The data pre-processing steps we applied are:

- **Renaming the columns for better understanding.**

The columns in the dataset were renamed for better understanding. The 'sales' column was renamed to 'department'. This column indicates all the departments in the organization from which the data is collected.

- **Changing categorical to numeric:**

- Department:** The Department column had categorical values like 'IT', 'hr', 'management' and so on. These departments were mapped to numerical values like 1,2,3 and so on. This is how the departments are mapped:

Department	Mapped_Department_Numerical_Values
accounting	1
hr	2
IT	3
management	4
marketing	5
product_mng	6
RandD	7
sales	8
support	9
technical	10

- b. Salary:** The Salary column had categorical values like low, 'medium', 'high' and so on. These salary levels were mapped to numerical values like 1,2,3 and so on. This is how the salary is mapped:

Salary	Mapped_Salary_Numerical_Values
low	1
medium	2
high	3

- c. Checking for Null values:** Null values in the dataset can give incorrect results. Thus it is imperative to check for null values in the dataset. For the HR Analytics dataset, the dataset is clean and has no null values. No pre-processing was needed for this part.

## 5. Feature Ranking

Feature Ranking is important part of Machine Learning because in this process, we find the most meaningful inputs. In order to do Feature Selection, we have many methods available like [as per <https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/feature-selection-data-mining>] :

1. Naïve Bayes
2. Decision Trees
3. Neural Network
4. Logistic Regression
5. Clustering
6. Linear Regression

For this dataset, we have used Decision Tree Classifier. Based on this we get following 3 top features:

1. Satisfaction Level
2. Time Spent in Company
3. Evaluation

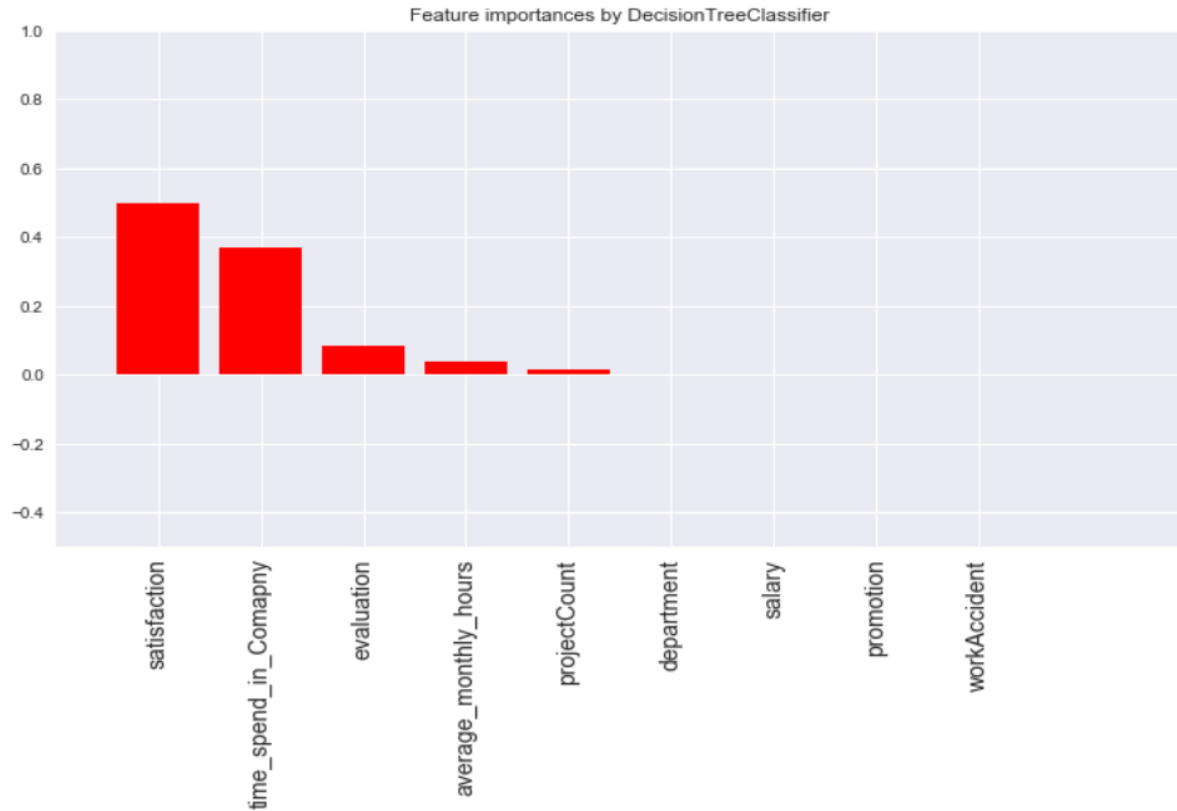


Fig. 2 Feature Ranking

The top 3 features helped us in calculating the Logistic Regression. As we know, by reducing the variance it will be more interpretable to understand what goes in our model when we use minimal number of features.

## 6. Correlation

Correlation is a statistical technique that can show if and how the variables are related. Correlation values can range from -1 and 1. -1 means the negative strong correlation and 1 means positive strong correlation. 0 means no correlation. This correlation matrix shows the relation between the predictors and the target.

From the figure below, we see that the highest correlation is between the Satisfaction level and left(Turnover).

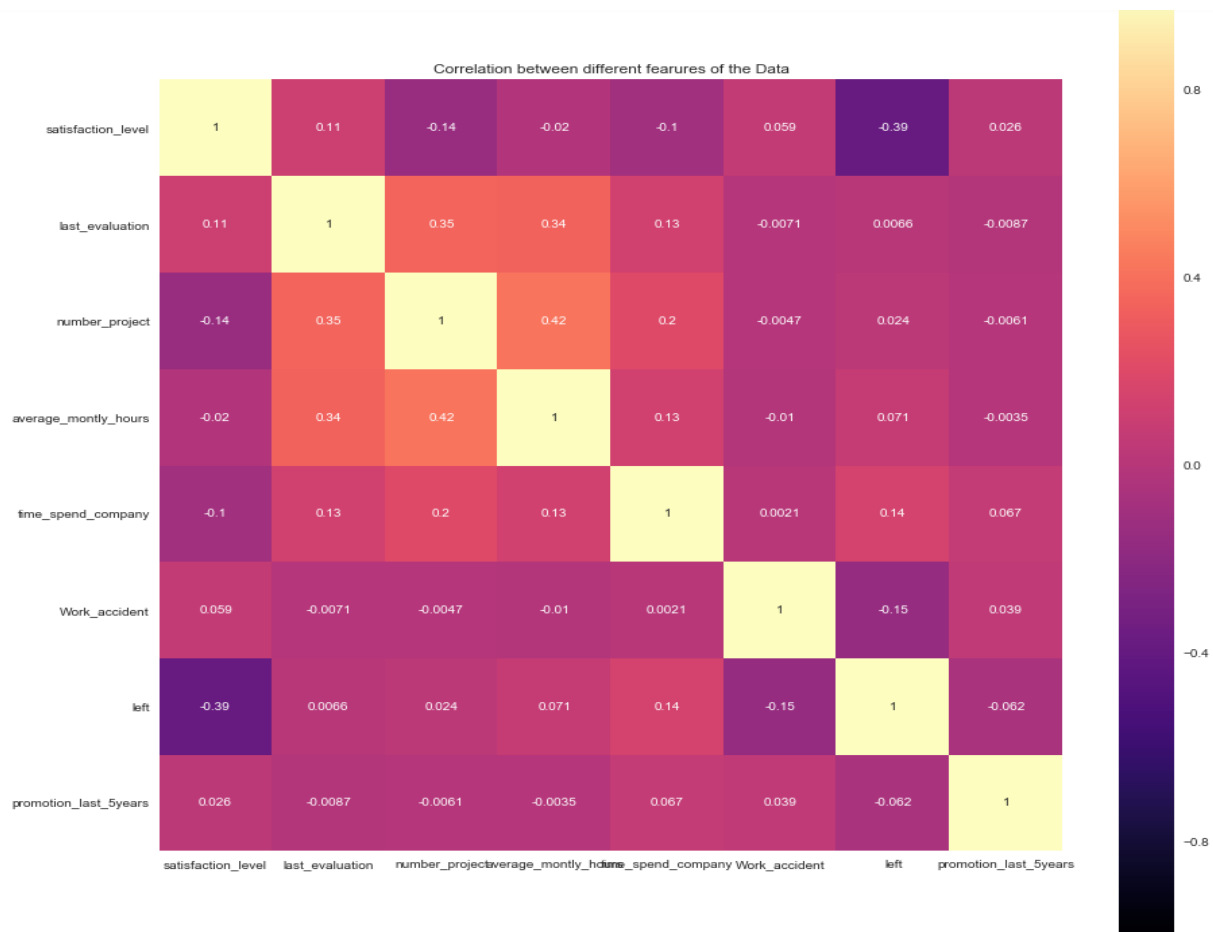


Fig. 3 Correlation matrix

The above below shows the result of correlation that is calculated using Weka. The predictor variable having highest correlation with target Left is Satisfaction Level.

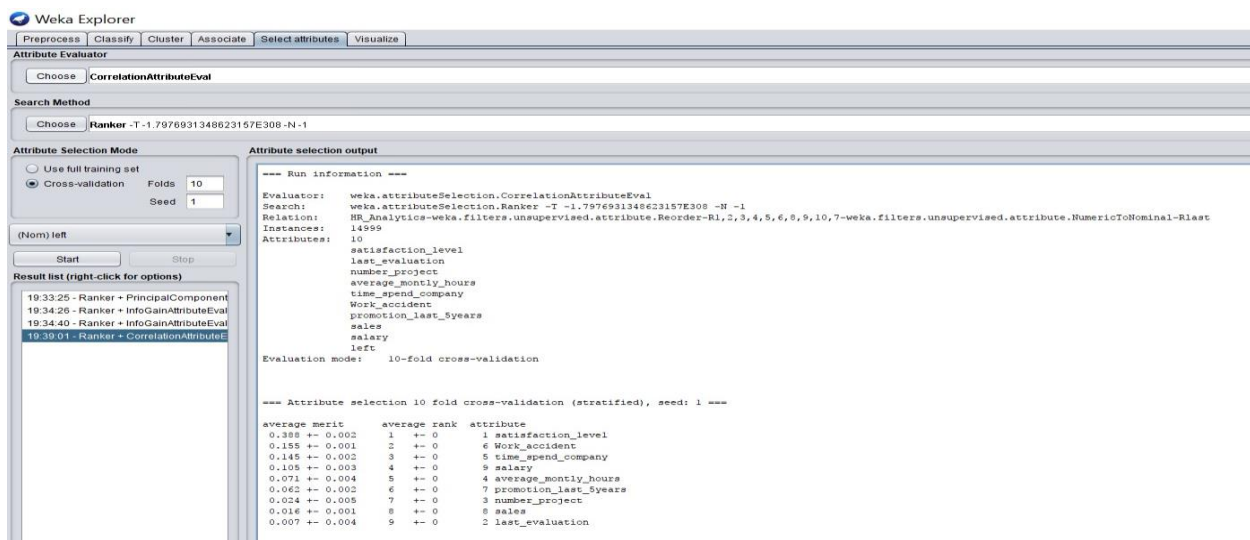


Fig. 4 Correlation in Weka

## 7. Variability in the data:

This section describes the various visualizations that we performed on the data.

In Fig. 5 we can clearly see that Employees with lower satisfaction level have left the company. Also, Employees who were part of either very few projects or those with higher number of projects have left the company. So, both the extremes left the organization in this case.

In 'Employee satisfaction', we can see that few Employees with higher satisfaction level have also left. Now this could be those who left for other significant reasons (maybe Avg. monthly hours or No. of Projects or so).

For 'Time Spent in the Company' we see that Employees who were in the organization for either less than 2 years or more than 6 years tended to stay and only those belonging to 3 to 6 years range have left. This is again understood when we analyze the count plot.

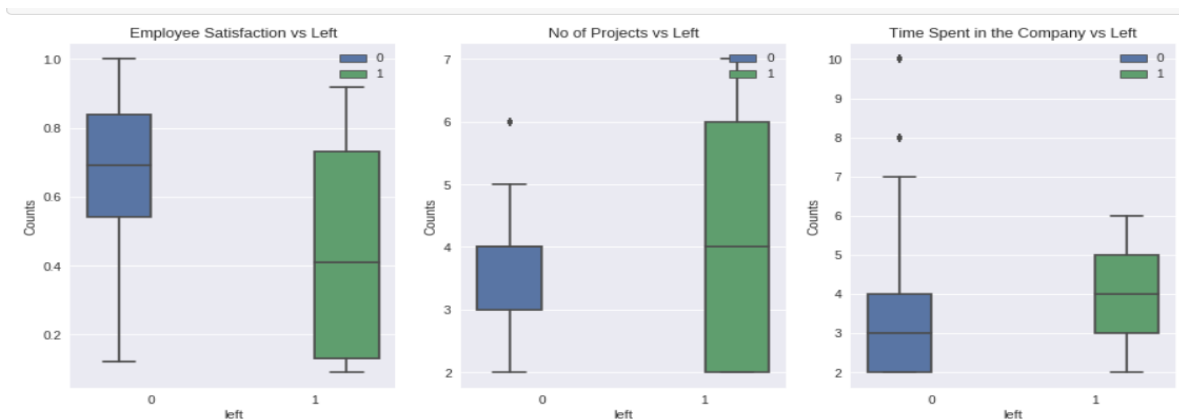


Fig. 5 Box plots of the higher ranked attributes vs the class attribute (Left)

In Fig. 6 we can see that the Employees who left were either under loaded with work or overloaded with work (Average Monthly hours). So, again we see that both the extremes in this attribute tend to leave. In case of 'Last Evaluation', Employees with either lower evaluation or people with higher evaluation have left the organization. This may be because Employees who got high ratings are also very talented and hence switching jobs. On the other hand Employees with low ratings ultimately left the company.

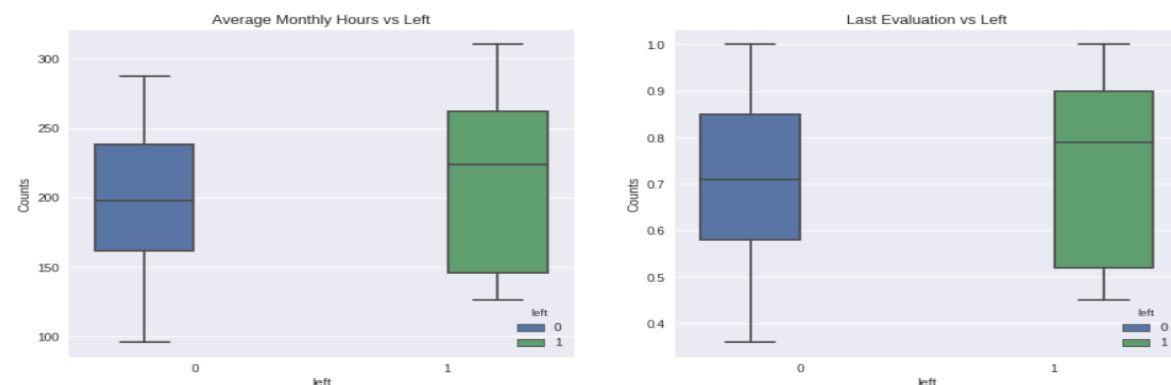


Fig. 6 Box plots of the higher ranked attributes vs the class attribute (Left)



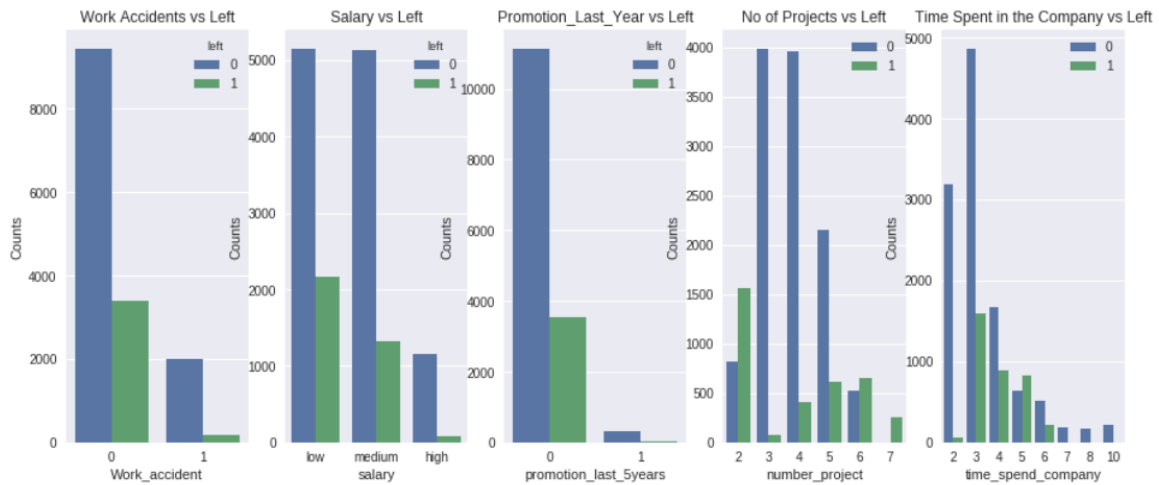


Fig. 7 Count plots of few more attributes vs the class attribute (Left)

In Fig. 7 we observe that- among the Employees who had work accidents, only few left the company. Whereas significant count of Employees who left fall in the zero work accidents group. So work accidents did not influence an employee's decision. In 'Salary' we observe that Employees with low and medium salaries tended to leave the company and the ones with higher salary rarely left. The count plots of 'No. of Projects' and 'Time Spent in company' convey similar statistics of what we saw in the box plots above. Overall, we observed that all the visualizations corroborated the findings of Feature rankings and Correlation matrix and thus enhanced the confidence on predicted values.

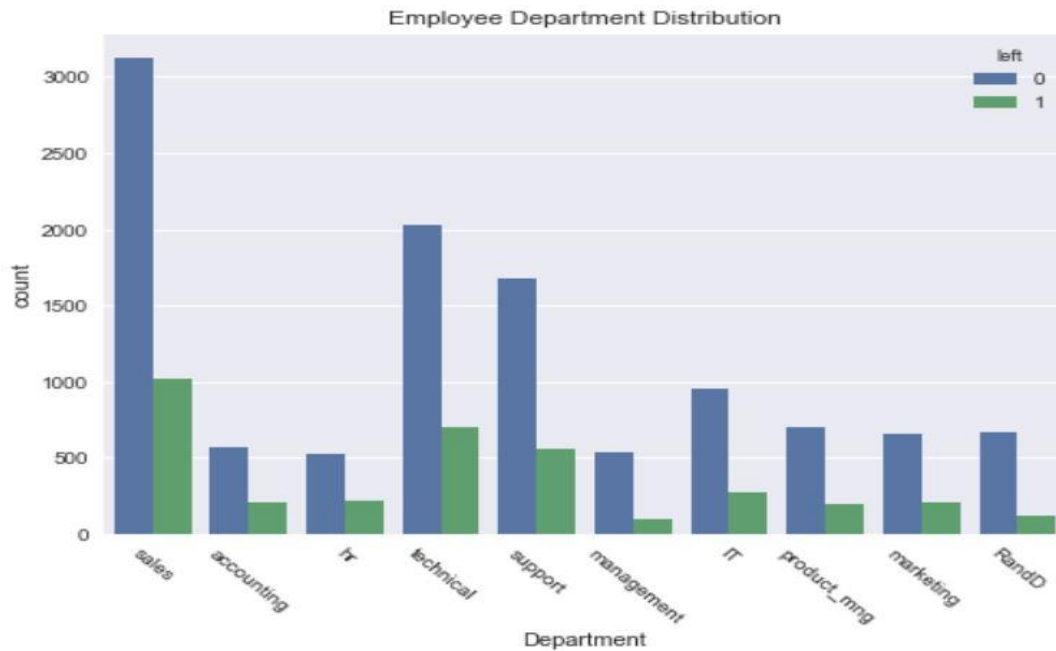


Fig. 8 Distribution of employees

From Fig.8, we can see the distribution of Employees who left across all the departments.

We see that the turnover is spread across almost evenly among the different departments with Sales having the maximum turnover.

## **8. Cross Validation technique**

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

For classification problems, one typically uses stratified k-fold cross-validation, in which the folds are selected so that each fold contains roughly the same proportions of class labels.

In repeated cross-validation, the cross-validation procedure is repeated n times, yielding n random partitions of the original sample. The n results are again averaged (or otherwise combined) to produce a single estimation.

## **9. Predicting Probability of Left using Classification Techniques**

In this study, we have used the following models to predict the probability of target “Left” using the 9 predictors.

### **- Evaluating Model Accuracy:**

The goal of any Machine learning model is to learn the patterns from the data. It is trained on the training data and then checked if it works well on the unseen test data. It is important to check if the model works well on the unseen data. We can do this by using the model to predict the answer on the held out data and then compare this predicted answer to the actual answer.

Many metrics are used to measure the accuracy of the model, that is measuring the predictive accuracy of the model.

Some metrics that can be used to test the classification models:

1. Sensitivity / Recall
2. Specificity
3. Precision
4. F1 Score

It is important to review these metrics to decide if the model is performing well.

### **- Understanding Precision and Recall**

Consider we are trying to check if a test can detect if someone has Cancer. Total people actually having cancer is 12 and some not having cancer. After the test, 8 were detected as having cancer. Of these 8, 5 people actually had cancer.

### - Precision

$$\text{Precision} = \text{True Positive} / \text{True Positive} + \text{False Positive}$$

$$= 5/8$$

### - Recall

$$\text{Recall} = \text{True Positive} / \text{True Positive} + \text{False Negative}$$

$$= 5/12$$

The models considered in this study are:

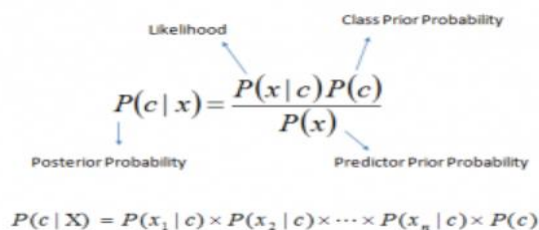
- a) Naïve Bayes
- b) Support Vector Machine
- c) Linear SVC
- d) Perceptron
- e) Logistic Regression
- f) Random Forests
- g) K-nearest neighbors
- h) Decision Trees

We are comparing the models on basis of their accuracy, precision and recall values.

### Naïve Bayes

Naïve Bayes is a simple classification technique that has attracted attention for its simplicity and performance. Naïve Bayes performs classification based on probabilities arrived, with a base assumption that all variables are conditionally independent of each other.

The underlying logic to using the Bayes' rule for Naive Bayes classifier is as follows:



$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$

To train a target function  $f_n: C \rightarrow X$ , which is the same as,  $P(C|X)$ , we use the training data to learn estimates of  $P(X|C)$  and  $P(C)$ . Using these estimated probability distributions and Bayes' rule new test  $X$  samples could then be classified.

For the HR turnover dataset, the Naive Bayes classifier calculates the posterior probability of the target Left using each of the 9 predictors with the assumption that all 9 predictors are independent of each other.

### Support Vector Machine

A SVM is a supervised learning algorithm that implements the principles of statistical learning theory and can solve linear as well as nonlinear binary classification problems.

A support vector machine constructs a hyper-plane or set of hyperplanes in higher dimensional space for achieving class separation. The intuition here is that a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class- the larger the margin the lower the generalization error of the classifier.

For the HR Turnover dataset, the support vector uses each of the 9 predictors and constructs a set of hyperplanes in 9-D space and chooses the hyperplane that provides the best class separation for the target 'Left'.

### Linear SVC

Linear SVM is a data mining algorithm for solving multiclass classification problems from large data sets that implements an original proprietary version of a cutting plane algorithm for designing a linear support vector machine. Linear SVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set.

Features of Linear SVC:

- Efficiency in dealing with large datasets (say, millions of training data pairs),
- Solving multiclass classification problems with any number of classes,
- Working with high dimensional data (thousands of features, attributes) in both sparse and dense format.
- No need for expensive computing resources.

### Perceptron

In machine learning, the perceptron is an algorithm for supervised learning of binary classifiers (functions that can decide whether an input, represented by a vector of numbers, belongs to some specific class or not). It is a type of linear classifier, i.e. a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector. The algorithm allows for online learning, that is, it processes elements in the training set one at a time.

In the modern sense, the perceptron is an algorithm for learning a binary classifier: a function that maps its input  $x$  (a real-valued vector) to an output value  $f(x)$  (a single binary value):

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

where  $w$  is a vector of real-valued weights,  $w \cdot x$  is the dot product  $\sum_{i=1}^m w_i x_i$ , where  $m$  is the number of inputs to the perceptron and  $b$  is the bias. The bias shifts the decision boundary away from the origin and does not depend on any input value.

### Logistic Regression

This is the framework similar to the multiple regression with predictors  $x_{1,i}, x_{2,i}, \dots, x_{k,i}$ :

$$\text{transformation}(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_k x_{k,i}$$

Logit Transformation can be written as:

$$\text{logit}(p_i) = \log_e \left( \frac{p_i}{1 - p_i} \right)$$

We can rewrite the first equation using the logit transformation of  $p_i$ :

$$\log_e \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

For this dataset, we take into consideration the top variables we found while calculating the feature importance. So, we consider these variables:

1. Satisfaction Level
2. Time Spent in Company
3. Evaluation

After selecting the variables, we calculated the coefficient values. The equation now looked like:

$$\text{Employee Turnover Score} = \text{Satisfaction}(-3.769022) + \text{Evaluation}(0.207596) + \text{YearsAtCompany}(0.170145) + 0.181896$$

When we take a row from this HR Turnover dataset, where an Employee with 0.7 Satisfaction and 0.8 Evaluation and worked 3 years has a 28% chance of turnover.

### Random Forests

Random forests are learning method for classification and regression by constructing a number of decision trees at training time. Then the output is the mode of the classes (of the classification) or the mean of the prediction (that of regression) of the individual trees.

It works as, there is a Training Set  $X=x_1, x_2, \dots, x_n$  and responses  $Y=y_1, y_2, \dots, y_n$  and for  $b_1, b_2, \dots, b_n$ :

1. Sample with replacement, take  $n$  training samples,  $X_b, Y_b$ .
2. Train a classification or regression tree  $F_b$  on  $X_b, Y_b$ .

After training, take the majority vote in case of classification trees.

When we take a row from this HR Turnover dataset, there are 10 trees generated.

### KNN Neighbors

K nearest neighbors is a data mining algorithm which is non-parametric (doesn't make any assumptions) and lazy. It stores available cases and classifies the new incoming cases or data points based on a similarity measure (like distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its  $K$  nearest neighbors measured by a distance function. If  $K = 1$ , then the case is simply assigned to the class of its nearest neighbor.

Choosing the optimal value for  $k$  should be done after data exploration. A larger  $K$  value is more precise as it gives more accuracy. Cross-validation can also be used to determine a good  $k$ -value by using an independent dataset to determine it.

## Decision trees

A decision tree follows a tree structure in which each internal node represents a test on an attribute. Each branch in the tree represents the outcome of the test and each leaf node represents a class label. The nodes present in a decision tree are decision nodes, chance nodes and end nodes.

Pros:

- Easy to understand and interpret.
- Possible to add new scenarios. Can be integrated with other decision trees.
- They implicitly perform variable screening or feature selection.
- Non-linear relationships between parameters does not affect tree performance.

## 10. Summary of the Models Results

Model	Accuracy	Precision	Recall
Naive Bayes	0.7914	0.6730	0.7378
Support Vector Machine	0.95	0.909	0.915
Linear SVC	0.7302	0.5446	0.6673
Perceptron	0.7587	0.1384	0.00056
Logistic Regression	0.7687	0.3398	0.3598
Random Forests	0.9902	0.9646	0.9689
KNN k = 1	95.9	0.9245	0.955
KNN k = 1000	78.5	0.6413	0.1008
Decision Trees	98.16	0.981	0.9731

## 10. Inferences

We came across few inferences from both the visualizations and predictions we did on the dataset. They provide us with key insights into the data. HR managers can make use of these findings so they can retain more employees.

- From the employee Satisfaction v/s Left box plot we can find that mean satisfaction level for employees who left is much lower than the employees who have stayed. We can visualize that most of the employees who have left fall on the lower side of the satisfaction level spectrum although there are few exceptions (like there are employees who left with satisfaction level more than 0.5)
- Most of the employees who have stayed have an ideal number of projects between 3 and 4. We can notice from the No of projects v/s left the people who left also has way lesser or way more number of projects.
- From the average monthly hours v/s left box plot we notice that employees who left have worked much more than those who left every month. We can also find that they have a varied range even though mean is higher for people who have left.
- We observed that people who spent less than 3 years or more than 7 years are less likely to leave the organization. But those who worked between 3 years and 7 years seemed to show more tendency to leave the company. This is evident from the Time spent in the company v/s Left plot (Fig. 6).
- Last Evaluation v/s Left shows that the variability is lesser for people who have left. Means, the IQR is less for this group. On the other hand employees with lower and higher last evaluation scores tends to leave which is a little shocking.
- From the visualization we were not able to point out many inferences between work accidents at workplace and the employee leaving the company. This is actually good news for the corporate management since they can deduce that work accidents is not really a factor.
- As expected, we noticed that people with low and medium salaries are more likely to leave the job as compared to people with high salaries. As we see from the Salary v/s Left plot, people with higher salaries rarely left the company.

## 11. Retention Strategies:

The company can employ the following strategies:

- Provide flexibility to employees who work more than 240 hours a month like a compensatory time off (comp off) or give work relaxation. Avoid making employees work more than what a nominal amount since employees tend to leave the company if they are overworked
- Refrain from assigning employees on less than 2 or more than 4 projects at a time. Project managers and HR managers can make sure every employee has been assigned to ideal number of projects so that they can contribute to each project in a notable manner and not feel left out. If an employee feels that he hasn't been assigned enough projects or tasks, he/she might feel stagnant.
- Provide incentive to employees with lower satisfaction levels. Management should talk to employees who have lower satisfaction levels and see what the problem is and act on it so that they won't leave since they are pretty volatile.

- Groom employees who have spent between 3-7 years in the organization as they are the most volatile group. Make sure they have good work-life balance and benefits.

## 12. References

- <https://www.kaggle.com/ludobenistant/hr-analytics-1>
- <https://halshs.archives-ouvertes.fr/hal-01556746/document>
- <https://pdfs.semanticscholar.org/fa49/19810eae67e851ad13775b78c94217a7908.pdf>
- <http://www.saedsayad.com/>
- <https://www.openml.org/a/estimation-procedures/1>