**Implementing Naïve Bayes’ Consequential Probability for Prediction and Prevention of Employee attrition**

**Abstract**

Employee attrition has been a nightmare for many organizations irrespective of their size or stature. Losing talent costs a company direly in terms of time and money which makes research in this field a necessity. This research sheds light on this topic and narrows down the reasons for attrition to allow prediction based on probabilistic methods using the Naïve Bayes probabilistic approach for consequential event occurrence in the HR (human resources) management domain.

***Keywords: attrition, Naïve Bayes, Prediction, Probability, attribute selection, relationship***

**Introduction**

Talent attrition has been one of the biggest problems faced by all organizations as when talent leaves, it leaves a dent in the company which takes both time and money to cope up with. This research is based on an analysis that is setup to understand the reasons for attrition by laying down the statistics of people who left the company and implementing that knowledge on the people who are still employed to assure that people who seemingly fall under similar circumstances are offered what they desire to stay in the company.

This report has been divided into 6 major sections which focus on providing a path towards narrowing down the reasons for attrition to gradually predicting how it can be stopped using the right model. Section 1 has been allocated to the process of data acquisition and preprocessing which leads into the assumptions that have consolidated the base for implementing Naïve Bayes in section 2. This is followed up by a study

on the state of the art techniques and the establishment of a null hypothesis to be checked for accuracy. This is followed up by business implications of this research focusing on how this will save organizations time and money. Section 5 shows the implementation techniques and methodologies which is followed up by the results and conclusions section. We conclude this research with a small proposal on how to further enhance the accuracy of this research in a future works section.

1. **Data Acquisition and Preprocessing**

The data used for the execution of this research has been sourced from Kaggle. The data comprises of 10 attributes, off which 5 have been chosen for this research based on their inter-dependencies leading to attrition. To implement Naïve Bayes technique efficiently, the data had to be manipulated to allow dichotomous outputs to all attributes. For instance, “satisfaction\_level” ranges from 0 to 1 (including float values), hence, to apply Naïve Bayes, the output was converted to satisfaction\_level > 0.5 or not. Table 1 shows a list of all the similar interpretations:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Range** | **Interpretation** |
| Satisfaction\_level | 0 - 1 | >0.5 or <=0.5 |
| Time\_spend\_company | 2-10 yrs | >5 or <= 5 |
| Average Monthly hours | 96-310 | >=190 or <190 |
| Number project | 2- 7 | >=4 or <4 |

Table 1.

To conclude the research, the interdependencies of attributes of the employees who left, i.e., left = 1 would be used to predict the signs of upcoming attrition possibilities.

This means that the test data that is being used is for people who have left the company and the accuracy of the null hypothesis will be checked based on that data.

The biggest advantages of reducing the columns being inspected is to narrow down the attrition to controllable and more effective reasons. Work accidents and when a survey was conducted last refer to events that have occurred in the past and the likelihood of another accident is not predictable, whereas when the survey was conducted is also susceptible to be of less value in the research.

1. **Assumptions:**

To accommodate the parametric requirements of this research, the assumptions that have been made are as follows:

* Satisfaction level above 0.5 is 1, else 0.
* Projects >= 4 are good, else less
* Working hours above 190 are good, else less
* All the employees that have left did so due to the reasons within the table, no outliers or external parameters considered.
* Attrition is caused due to 5 major factors, categorized into two segments of employees, this is presented in a more elaborated manner as the null hypothesis in section 3, sub-section C.

1. **Technical Investigation**

State of the art models to understand attrition implement Artificial Neural Networks (ANN) and Decision Trees (C&R tree), which resulted in astounding accuracies of 85.33% and 80.89% respectively **[4]**. One of the established implications of understanding the features leading to attrition is to be able to predict the upcoming employees who could be next in line to leave. This means that attrition prevention not only should be accurate, but also scalable, to allow predictions and be prepared for outliers, hence, robust. To get an output which maximized the accuracy of attrition prediction, this research proposes the use of a Naïve Bayes model implementing a classifier based technique, where the employees are not just marked based on the features of their attrition but also classified.

To implement the correct model and achieve optimum accuracy, this research has been broken down into 3 different categories:

* Algorithmic Viewpoint
* Hierarchical Viewpoint
* Analytical Viewpoint

**3.1. An Algorithmic Viewpoint**

The technique to execute this research has been derived using the traditional ensemble method **[1]**, where a supervised model is implemented and various classifiers are generated to narrow down the suitable techniques for research.

The Naïve Bayes probabilistic approach is considered when the probability of the occurrence of an event is to be studied with respect to the probabilities of occurrence of other dependent events **[2]**. This makes the Naïve Bayes a strong predictive model as it considers many attributes of a data set together including their dependencies.

The reasons supporting the selection of Naïve Bayes model for this research are listed as follows:

* High performance against noisy data
* Can be scaled, i.e., a small portion can be researched and a prediction for a large model can be generated
* We can use basic classifiers to allow us to reject or accept based on theory and our knowledge **[3]**
* Allows us to combine the output of two models, a feature implemented immensely in the execution of this research.

**3.1.1 Naïve Bayes**

This research implies the usage of the classic Naïve Bayes probabilistic approach along with Naïve Bayes classifiers. The pros of using classifiers is to enable a simple accept or rejection theory [7], where we can now determine whether a field is going to affect attrition or not based on a dichotomous pattern that has been implemented on the data.

The Naïve Bayes combines the probability of occurrence of one events which lead to a final event and is chosen for this research for the following reasons:

* This model is setup to get the likelihood of one event when other events happen, hence can take up any number of fields that are linked to an outcome, such as attrition in this case.
* This model is ranked best [**Brett LANTZ**] in terms of dealing with bad data and null values as is completely ignores any weird outliers while calculating probability.

The equation for Naïve Bayes links the probability of the events to happen with respect to each other occurring:

Hence,

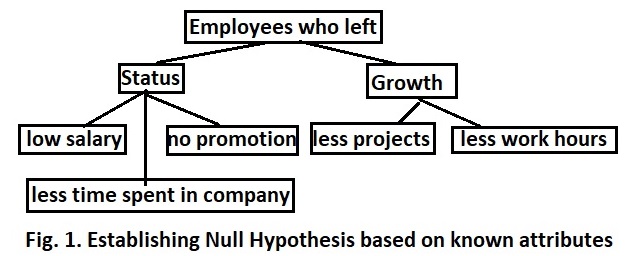
P (X| A ∩ B ∩ C) = P (A ∩ B ∩ C| X) P(X)

P (A ∩ B ∩ C)

This means that the probability of a secondary event happening with respect to a first event is directly proportional to the likelihood of the reverse event sequence and inversely proportional to the probability of the second event occurring independently.

**3.2 A Hierarchical Viewpoint: Classification**

This section of the research is focused on the cessation of the dataset complexity and implementing a simple hierarchy based classification technique. As elaborated in figure1, the group of attributes has been narrowed down to attrition based on 2 features; **status** and **growth**.



Status refers to the employees seeking more benefits from the company in terms of salary, promotion and this is linked to the time the employees have spent in the company. Growth refers to the employees who left the company due to less projects and being underworked.

The classification is done based on feature selection as described in table 1 and the reasons for selecting the specific values are based on traditional work place scenario analysis such as 36- hour week, making 190-hour month a feasible standard for this research.

To further simplify the implementation of Naïve Bayes, this kind of an approach has been implemented on all the fields which are not dichotomous in nature initially. This allows classification and Naïve Bayes to work hand in hand as desired and help incorporate a certain level of simplicity to an otherwise complicated process.

The process of using multiple techniques to achieve high quality output in a research is coined as **Stacking** **[5]**, where a classifier is combined with a machine learning technique (such as Naïve Bayes, as proposed in this research) to increase the accuracy of the research.

**3.3 Analytical Viewpoint: Hypothesis Specification**

On conducting an initial analysis on the data, research showed that the people who had left have a few attributes in common which can be looked it to narrow down the reasons for attrition.

These features were further précised down to 2 segments of workers, those who left as they needed a status based motivation and those who needed a personal growth based motivation to stay.

A **null hypothesis** based on initial results from the analysis suggests:

*“****The attrition rates primarily depend upon the attributes such as registered satisfaction level (low), number of projects (low), average monthly hours (low), promotions overdue (high), salary scale (low) and the time spent in the company (less)****”*

To test this hypothesis, the research will be conducted and an accuracy scale (to compare the output with state of the art methods) will be established based on the output of the research.

1. **Business Implications**

This research will help consolidate the employee working environment by allowing the HR department to analyze the characteristics of their workers and allocate appropriate baits to them, this would not only help the organization retain talent but also help them break down the work and projects more efficiently. If automated, this system could well replace the traditional HR internal XRM systems and be used to track all employee behaviors leading to attrition. Despite all the predictions and accuracies, some parameters such as external influence and rival organizations are not a factor in the dataset and hence excluded from the research.

This research will not only focus on learning but also on stopping attrition, the proposed techniques for employee retention would depend on the result from the test data, that has been limited to the people who left the company and applying the model with a high accuracy on the impactful attributes to the details of the people who are still employed.

1. **IMPLEMENTATION**

For the successful implementation of this research, the following execution path has been designed:

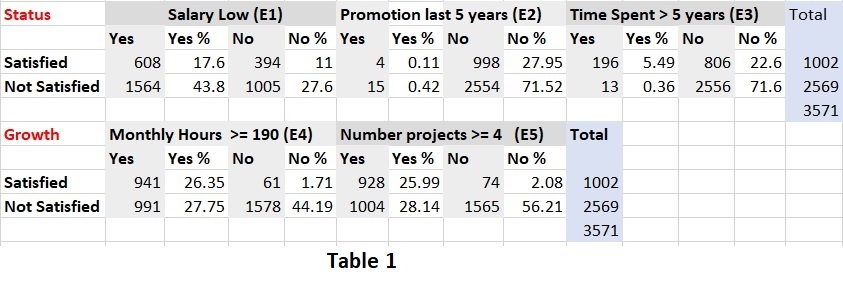
* Data cleaning for null values and eliminating independent attributes from the data.
* Classifying the relevant attributes into the 2 categories, Status based and Growth based.
* Converting the relevant numeric data into dichotomous data to allow the implementation of Naïve Bayes.
* Designing a frequency table and a likelihood table for the attributes being hypothesized to lead to maximum attrition.
* Combining the probability of attrition derived in both the categories based on the relevant relationship they share to check the validity of the null hypothesis.
* Predict the employees susceptible to attrition and prepare retention plans and offers wherever necessary.

**Tools used:**

* For cleaning the data, R programming language was implemented in R studio
* For basic classification and count of occurrences of events as specified in the null hypothesis, the data was served to SQL Server and simple SQL queries were run to get the count.
* To produce the results in a presentable manner and apply color coding to distinguish between attributes, Microsoft Excel (tables 2 and 3)
  1. **Attribute input selection criteria**

The attributes chosen for this research are based on 2 different mindsets of people, which can be linked towards obtaining a 360-degree view of the problem. Attrition is mostly linked to a few attributes at once but the employees in this research have been divided into segments, who can be classified as Status seeking and Growth seeking employees. This is done to be able to tackle the attrition problem more efficiently as the employee who falls in either category could be given more of what leads to a lack in motivation for them.

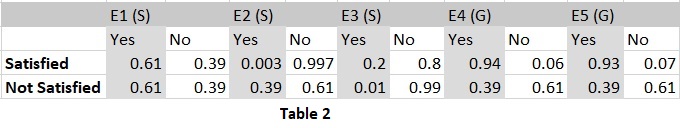
**5.2 Employee segmentation**

This section is dedicated to obtaining the best output in terms of understanding the mindsets of the employees who left the organization. This is a critical step towards offering the right employees what they seek in terms of their goals. To initiate the process, this research has the aforementioned Status affiliated group and the personal growth affiliated group. This allow us to not only understand where an employee stands, but also offer them more than one reasons to stay at once, which could add a personal touch and the employees feel more pampered. This is one of the most effective ways to optimize and personalize the interaction using a simple CRM system which could store the details of employees with their IDs or names along with the output of their statistics.

**5.3 Preparing the Data**

The data being included in the research is initially not dichotomous in any of the dependent variables, hence, relationships are established and a binary output based system is established. This is followed up by a basic level 1 segmentation where the employee information is used to understand the reasons for attrition. To allow this, out of the set of 14999 observations, 3571 are studied, which comprises of all the people who left the firm. The idea is to use this information and test the null hypothesis and if a high accuracy is achieved, implement the knowledge on the remaining employees in the organization to forecast who could be next in line for attrition and hence, offer them the appropriate bait to stay.

**5.4 Naïve Bayes Implementation**

As mentioned in section 3 sub-section 3.1, we approach this research using the Naïve Bayes model for chained probability evaluation. To understand how this will be implemented, this section of the research has been broken down into the following elementary process flow:

1. Designing frequency and likelihood tables
2. Determine the relationship between events based on the tables
3. Measure the accuracy of Naïve Bayes by establishing a relationship between the two segments.
4. Measure the accuracy of the output
5. Implement that to the remaining sample of currently employed workers to predict attrition and prepare measures for retention where necessary.

**5.4.1 Frequency and Likelihood Table**

The frequency of occurrence of the attributes leading to attrition (based on the null hypothesis) is registered in Table 1 and an aggregation on a percentage scale is presented. This is followed up by a **likelihood** table, that has been represented in table 3. Here, the outputs of table 1 are simply converted to a likelihood or probability of occurrence. Table 2 is a derivative of table 1, the process of generation and calculations behind the table have been elaborated in section 5.4.2.

**5.4.2 Setting up Naïve Bayes**

The evaluation of this research is based upon understanding the process that has been followed up in Naïve Bayes. As stated in section 2.4.1, likelihood of an event is based on the frequency of occurrence of the event in the test data. For instance, in table 2, the number of occurrences where salary is low and the person is not satisfied is 1564. The total occurrences in that domain is 2569, hence, the likelihood of this event occurring is 1564/2569 = 0.61. When predicting using the accuracy of this model, this likelihood is the probability of occurrence of a similar scenario, i.e., where the salary is low and satisfaction is low.

**5.4.3 Relationships Between Events**

As seen in tables 2 and 3, a broad scope has been left open for the conditions occurring in E (1:5). This is to accommodate the null hypothesis in terms of 4 sub-event classifications. To understand this, consider the following scenario:

E1 could occur when salary was low and the employee was satisfied or E1 could be the event as desired by the null hypothesis for this research where salary was low and employee was not satisfied (1564 occurrences, containing up to 43.8% of the total) and a probability of 0.61.

To demonstrate the null hypothesis in terms of adaptability to the Naïve Bayes approach, we consider only the conditions where employees who left registered a **“Not Satisfied”.**

We have two probabilities to calculate:

1. P (E1 ∩ -E2 ∩ -E3): Status
2. P (-E4 ∩ -E5): Growth
3. P (Status ∩ Growth)

The above probabilities relationships can be understood as stated below:

In probability 1, E1 refers to a low salary, -E2 refers to a no promotion in 5 years and -E3 referring to time spent being less than 5 years. The same logic applies to the second equation. The reason for choosing the intersection (∩) and not a union (U) is that the events are dependent and this research is focused on a precise result, union in this case could lead to a more than 100 percent accuracy [6].

**6. Results**

In terms of status based attrition:

* P (E1 ∩ -E2 ∩ -E3) ∝ 0.61\*0.61\*0.61 = 0.368

P (status) = 0.368/ (0.368+0.02) = 0.995

0.02 is for employees who were satisfied and left with a low salary.

* P (-E4 ∩ -E5) ∝ 0.61\* 0.61 = 0.372

P (growth) = 0.372/ (0.372+ 0.152) = 0.71

* To further extend the prediction, this research tries to narrow down the attrition to those employees who left and belong to both the above categories:

P (status ∩ growth) = p(status) \* p(growth)

P (status ∩ growth) = 0.71\* 0.995 = 0.706

**7. Prediction and Accuracy**

Since we now have a set notion of the attributes of people who left, we could run a simple analysis on the system database with the conditions in the null hypothesis and see how the statistics speak in comparison to the expected standards of this research. For a simple SQL query run based on the null hypothesis, the results revealed the following information:

Not Satisfied (Status) = 1546/ 3571

Not Satisfied (Growth) = 1526/ 3571

Summing up the above, based on the attributes narrowed down by the hypothesis, 3072 employees out of 3571 showed the attributes predicted to be the reasons of attrition based on the Naïve Bayes Model. The accuracy can be calculated as below:

Accuracy = (3072/3671) \* 100

= 86.03%

**8. Conclusion and Interpretation**

In the test data which comprised of employees who left the organization, there is a 0.706 probability of the null hypothesis being held true, i.e., they will have a low salary, nost been promoted in 5 years, less time spend in the company, not being given working hours and not being given many projects. With a prediction accuracy of over 86%, this model proves to be more accurate than ANN (85.33).

Despite the high accuracy, we cannot turn a blind eye towards the fact that a very restricted approach (in terms of flexibility) has been implied here and this could change in case a large data set is tested upon.

The Naïve Bayes model showed more accuracy for attrition prediction than any of the state of the art models. This could be the result of the data cleaning process being dichotomous to support the Naïve Bayes model, which ignored many close comings, such as a satisfaction of 0.49 is not satisfied but 0.51 is, hence, despite the high accuracy of this model, further research and analysis would be required.

**8.1 Limitations of this research:**

This research is limited by the lack of external reasons for attrition such as better opportunities in other firms.

The dichotomous behavior of data provides both pros and cons alike. Where it is good to declare a little satisfaction low as something that needs work, a 0.51 as satisfied is also a major exaggeration and needs work. This is why the use of ANN is better as compared to Naïve Bayes.

**9. Research Extension**

As mentioned in previous sections, this research uses a dichotomous method of implementing Naïve Bayes, however, a 3:n methodology of implementation can be established as well, where we can further break down the satisfaction levels and other similar fields to get a more precise result.

Once the analysis on people who left is complete, this research goes to predicting the people who fall under similar categories, that does amount to a majority of attrition prone employees but if the second and third highest attributes are ranked, this could change the game altogether as a higher accuracy could be achieved and more employee issues could be addressed and retained. A model could well be proposed to automate the process and approach employees directly with a “personalized touch” per say, based on the likelihood of a talented employee prone to leaving.

**REFERENCES:**

Dietterich, T.G., Ensemble Methods in Machine Learning.

Lantz, B. (2015) “Machine Learning with R,” United Kingdom, 2nd edition, pp. 1-425.

Murphy, K. P. (2006) “Naïve Bayes Classifiers,” pp. 1-8.

Frye, A., Boomhower, C., Smith, M., Vitovsky, L. and Fabricant, S. (2018) “Employee Attrition: What makes an employee quit?” *SMU Data Science Review*, Vol. 1, No. 1. pp. 1-28.

Dzeroski, S. and Zenko, B. (2004) “Is Combining Classifiers with Stacking better than selecting the best one?” *Machine Learning*, pp. 255-273.

Rish, I., Hellerstein, J. and Thathachar, J., “An analysis of data characteristics that affect naive Bayes performance”.