Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 3, July 2021: 5120-5128

Research Article

Data Science Process Pipeline to Solve Employee Attenuation: Attrition prediction

Ramesh Karnati[1], C Deekshitha Reddy[3], Sofia[4], Arathi J[5]

ABSTRACT

Industries are the backbone of the country visibility at global market. Employs and their capabilities are the back bone of industry growth. Therefore, most of the industries have started Human Resource Department to recruit and train employs for contributing company with huge investment. In other side, it become difficult to retain their employs because of death, retirement and resignation due to good offers from other companies. Therefore, it become more difficult to the HRD to transform their experience, knowledge and investment with new persons, it is widely recognized as Employ Attrition or Attenuation. In this scenario, HRD team is more responsible in identifying the reasons whether an employee can continue or leave company. This paper focus on data science process pipeline to identify the factors for employ attrition and predict whether an employ leaves the company or not. The proposed model is integrated with the various prediction models and returns the predicted value which is the best. In addition to that it is also investigated with Top K-best features selected from chai square test analysis and train the model to get accurate prediction value. The experimental analysis shows with accuracy values of the prediction methods for the given dataset

Keywords: Attrition, Attenuation, Prediction methods, accuracy.

I. INTRODUCTION

Employees are the biggest benefit to any organization. Therefore, most of the organizations have invested more money on their recruitment and skills development. Currently, retention of the employee is the huge challenge to the organization. Attenuation/Attrition is a serious problem for all the firms over various industries. Anything that involves human effort will be affected by the reduction of the workforce. Whenever employees leave an organization, they carry with them immense knowledge, experience. It is difficult for the organization, particularly for the human resource managers to fill the void. Sometimes, the human resource managers may not find a candidate like the previous one. To hire someone new, the company has to put in a lot of time and money. These days, the organizations have understood the significance and relevance of employees and are taking necessary steps to understand the reasons for the employee attrition. The aim of this research is to predict the employee attenuation (so that the company can take necessary actions to retain the employee) and know the reasons behind the employee to leave the organization. The research has emphasized the

[1] Associate Professor, Department of Information Technology Vardhaman College of Engineering.Email: ramesh.krnt@vardhaman.org
[2,3,4,5] Student, Department of Information Technology,Vardhaman College of Engineering

factors like age, daily rate, business travel, education, department, distance from home, education field, employee count, environment satisfaction, employee number, gender, hourly rate, job involvement, job level, job role, job satisfaction, marital status, monthly income, monthly rate, number of companies worked, over time, percentage salary hike, performance rating, relationship satisfaction, standard hours, stock option level, total working hours, number of trainings last year, work life balance, years at company, years in current role, years since last promotion, years with current manager, which affect employee attrition. Understanding the attenuation is crucial for every organization. Different industries might have different reasons for the employee to leave. So thorough research has to be done by the organization to get better understanding. Mere identification of factors would do no better for both the employees as well as employer. This study focuses on predicting the attrition and collect the factors that affect attrition and has a further scope where the prediction can be done large datasets and the factors affecting can be actually tested in different sectors and other related factors can be embellished to observe the less attrition rate.

Emery and Trist's research [5], they have defined employee turnover/attrition as when an individual joined a company, the communication between the company and the individual has to increase. If the communication could not increase to an appropriate extent, the individual's former experience with previous company would be so-called Guiding Crisis and the individual would give up on the current company eventually. Bluedorn (1982) [3] has given explanation to employee turnover as the individual put an end to playing a role in the company and left the relevant areas of the company. According to Hutchison and Purcell (2013) [5], the repercussions of employee attenuation depends upon the magnitude of the company (if the company size is small, then the company has to face severe consequences), and the employees.

Employee turnover can be categorized into two categories. The first one is voluntary that the decision taken by employee to leave the company and join in the competitive company. This kind of leave is considered as unethical. On the contrary, involuntary turnover means ousting of employees.

The importance of managing the talented people has always been in the picture. Irrespective of industry, no organization cannot bear to lose their employees at a rapid rate. And this is the reason why employee turnover or employee attrition has caught the observation of researchers.

It is observed that the analysis on these collection of factors for attrition could help to identify the circumstances that lead to an employ leaving the organization and to take decision on knowledge transformation from such employs. In this paper, the factors that influences attrition are studied and appropriate factors are chosen for prediction model. We have integrated prediction models and chose the best model as per the accuracy over the test data and train that model with the remaining data.

2. LITERATURE REVIEW

Abelson, M 1984 et al. [2] have thrown light on the positive side to the employee attrition. The study [2] focus on how attrition of poor performers is beneficial to the organization. Magner et al 1996 [13] have identified that employees should be aware of the issues happening within the organization. If there is no limpidity, the employees would opt to leave the company. Labov, 1997 [10] stated that the firms which are having good communication with their employees

have suffered less attrition rate than organizations which are having poor connection with their employees. **Sahu and Gupta, 1999 [20]** have emphasized that extent of service, expectation reality match, turnover perception and outside career opportunities are reasons for deciding to exit or to stay with the organization. **Abbasi and Hollman, 2000 [1]** talked about reasons that increases the rate of employee attenuation. Unsafe employment environment, hiring practices, lack of acknowledgement, and lack of competitive remuneration systems are the reasons which results in employee attenuation. **Zuber, 2001 [22]** has found out that the employees are not interested to work at the company where the employment conditions are unpredictable. The employees want to work at risk-free and predictable workplace. **Khatri et al, 2001 [9]** has emphasized that the causes of employee attenuation are age, income, gender, their position of work, nature of work. **Boxall, P., Macky, K., & Rasmussen, E. ,2003) [4]** have stressed that work life balance is the most crucial factor in retaining the employees and they would forsake the organization to obtain a better work life balance. The study also emphasized that the employee attenuation involves several factors but not one. **Ramlall, 2003 [18, 19]** accentuated that deficient compensation, remuneration below present market rate and deficiency in the equity is the customary reason because of which employees quit an organization. **Ma, Cheng and Wang, 2003 [12]** has cited that senior employees or employees who have been working for years might get exhausted and have low level of exhilaration which might lead to resignation or abandon the company.
Robbins,2003 [19] identified that if the employees are dissatisfied with the remuneration, they seek to leave the company and join somewhere where they are getting paid better. **Owence, Pinagase, Merey, 2004 [17]** has highlighted that reasons for employee attenuation are job dissatisfaction, quixotic expectations, and other better remunerated jobs. **Firth et al., 2004 [6]** has underlined that employees leave their work place if they experience job stress, job dissatisfaction and lack of commitment. **Mullins, 2005 [14]** has highlighted that employees have to get the acknowledgement and appreciation they deserve. The employees also have to be acclaimed for the work they do. **Oldham and Hackman, 2005 [16] highlighted** that employees quit their jobs only when they face consistent problems with regard to work related matters. For example, dominant s presented the necessity of understanding employee expectations. This kind of information is useful for the HR Managers to know the reasons and significance of their employees.

Liu and Wang, 2006 [11] has noted that the age, gender, tenure are obliquely affecting the employee to leave the company. Their study also discovered that women employee attenuation rate is higher than the men because women are confined to family duties such as pregnancy or taking care of the household. **Mathis and Jackson, 2007 [15]** has stated that the existing employees have to work more to compensate of those who have left the organization. **Radzi (2009)** has emphasized that the employees will continue in the company if they are treated impartially regarding results and process. **Zhang, 2016 [21]** has underlined that the significant factors for the employee attenuation relies on work burden, tractable work hours, type of culture. **Ibrahim M et al. [8]** have done research on customer churn prediction in telecommunications industry which is quite similar to the employee attrition. This research study has highlighted how to get awareness that a customer is leaving the company because necessary actions are to be taken by the company towards that customer.
The above literature focus on identifying the factors that could affect the decision of employee. It has not been investigated deeply on predicting whether an employ can leave the company in the future or not. The rest of the modules describes the methodology that can predict employ attrition.

3. METHODOLOGY

This paper discus data science process to solve employee attrition. The steps involved in the proposed are visualized in fig 1. It is started with preprocessing the dataset, extracting features that are suitable for prediction, some portion of the dataset is used for training predicting model, find the accuracy over the test case samples and decide the best model as per the accuracy of the model.

3.1 Data Science Pipeline Process

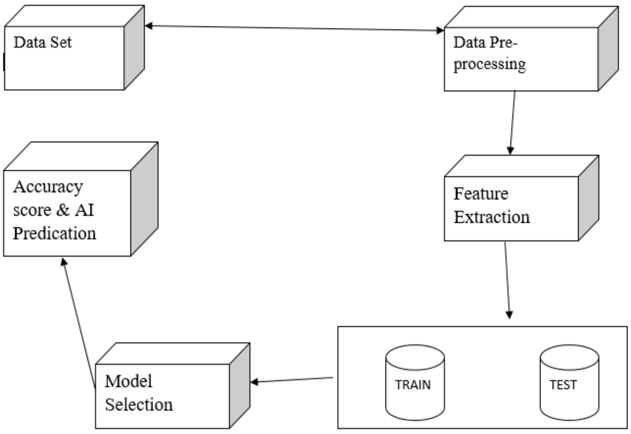
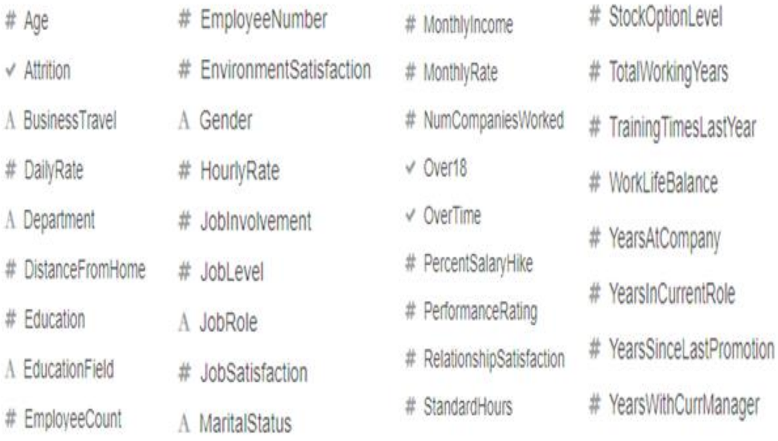


Fig 1. Data Science Pipeline Process

3.1.1 Dataset: The relevant features are collected and represented interms of attributes that are comma separated values represented in the fig 2.



Age
✓ Attrition
A BusinessTravel
DailyRate
A Department
DistanceFromHome
Education
A EducationField
EmployeeCount
EmployeeNumber
EnvironmentSatisfaction
A Gender
HourlyRate
JobInvolvement
JobLevel
A JobRole
JobSatisfaction
A MaritalStatus
MonthlyIncome
MonthlyRate
NumCompaniesWorked
✓ Over18
✓ OverTime
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StandardHours
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager

Fig 2. Features of dataset

3.1.2 Preprocess the data:

The dataset that has been read is messy and not understandable to the system. The dataset might contain null values as well. To make the data understandable, preprocessing is done. The categorical columns have to be converted into numerical columns and null values have to be removed. For preprocessing, we have checked for null values (there are no null values in the dataset) and categorical columns have been converted to numerical columns.

3.1.3 Feature extraction: The irrelevant features which do not contribute to the prediction have been dropped from the dataset. The dropped features or columns are Daily Rate, Education Field, Employee Count, Employee Number, Hourly Rate, Monthly Rate, Over 18, Relationship Satisfaction, and Standard Hours. All the columns except above are extracted for prediction.

After EDA analysis on the dataset, it is identified that there exists certain relationship between age, monthly income, total working years, gender, and position of employee and attrition factors of employees.

3.1.4 Training and Testing: After feature extraction, the dataset is shuffled to get unbiased predictions. The dataset is partitioned into training and testing with 80% and 20%. This procedure is used to fit the parameters and testing is used to evaluate the performance of the model.

3.1.5 Model Selection: The models are built using the algorithms random forest, support vector classifier, k-nearest neighbor, logistic regression, naive bayes, decision tree and stochastic gradient descent. The algorithm which yielded highest accuracy rate has been considered for prediction.

Naïve Bayes:

It is simple technique based on the assumption that all the features are independent, calculates probabilistic values that are posterior probability for classification. In this model selection, Gaussian distribution is used to handle the data distribution.

Logistic Regression:

It is popular technique especially used when the give data contains categorical data. The basic principal idea is to fit the data into logistic function to prevent over fitting.

K-nearest neighbor:

It is distance based method that classifies the new feature value into the nearest k neighbor values with the help of Euclidian distance, Manhattan distance and Minkowski measures.

Decision Tree Algorithm:

It is an entropy based supervised method that translate rules into paths of tree based on the entropy value. It is used to predict the target feature by learning decision rules. It is widely used in predicting class label because of the way the rules are captured with measures Information Gain and Gini Index. Earlier versions are suffered from over fitting issues while building model but it is also achieved with Pre-Pruning and Post-Pruning.

Stochastic gradient descent

It is used along with linear regression to minimize the error by starting from a random number and travels towards the lowest function.

Support Vector Machine

It is popular technique is used for classification and predicting target class label. It became popular because it maximizes margin of the classifier. However, it is also suffering from local optima.

Random Forest

It consists of huge number of decision trees probably uncorrelated under one tree, each tree leads to a certain class predictor. Therefore, it gives more accuracy than decision tree models and other models.

The above popular techniques are embedded in the model selection, the best one is chosen based on the accuracy.

3.1.6 Accuracy score and prediction:

For testing, some portion of the data is considered for training, and the remaining data is considered for testing. To test the performance, well known measure accuracy is used to find the number of correct predictions made out of the all trail predictions. Fig 3 is presented with the accuracy of each model. It is clearly showing that Random forest accuracy is better than other approaches.

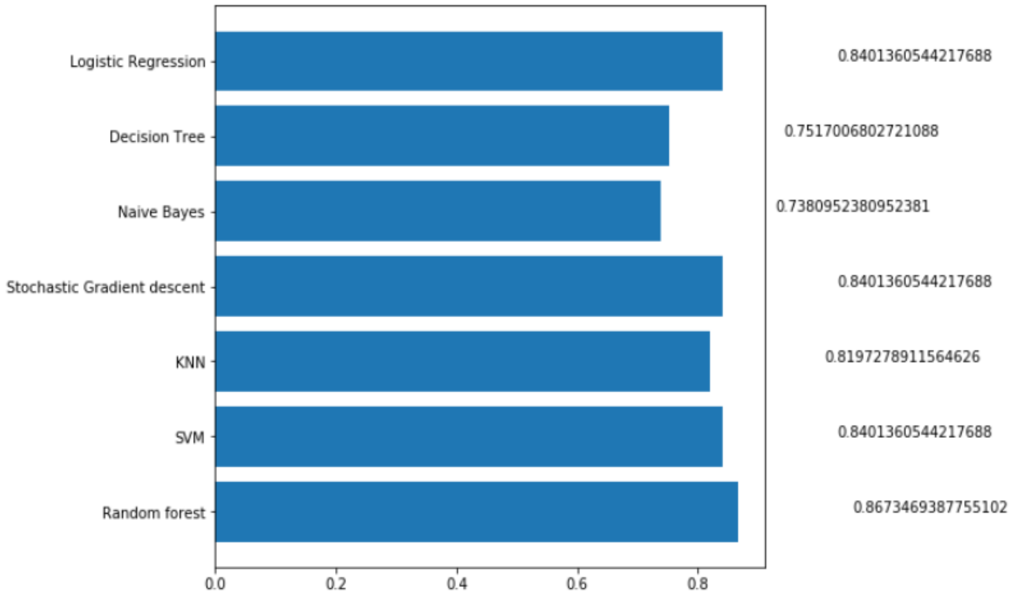


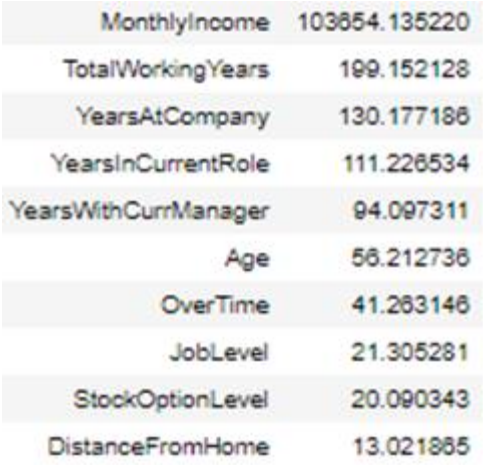
Fig 3. Accuracy Values of Prediction Methods

3.2 Feature Selection:

It plays a key role in selecting features which contribute the most to the prediction. The selection can be automatic or manual. Feature selection helps to decrease the training time and more accurate results. To select the features, select ***k*** best has been used along with chi2 function. The $\chi 2$ test is used in statistics to test the independence of two events. More specifically in feature selection we use it to test whether the occurrence of a specific term and the occurrence of a specific class are independent. More formally, given a document D, we estimate the following quantity for each term and rank them by their score.

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

The feature scores are calculated using the chi2 function and the top k (10) features are listed in the fig 4.



MonthlyIncome	103654.135220
TotalWorkingYears	199.152128
YearsAtCompany	130.177186
YearsInCurrentRole	111.226534
YearsWithCurrManager	94.097311
Age	56.212736
OverTime	41.263146
JobLevel	21.305281
StockOptionLevel	20.090343
DistanceFromHome	13.021865

Fig 4. Top K Features of Chi square Analysis

The computed top-k features are used for training the model to predict the employee is going to stay in the organization or leave. The above values have been given to the corresponding features of the model without considering the feature selection and has obtained "Attenuation" as the output. The accuracy score of the feature selection model is greater than non feature-selection model shown in fig 5. The accuracy score obtained for feature selection model is: 0.8707482993197279 which is better than considering all the features.

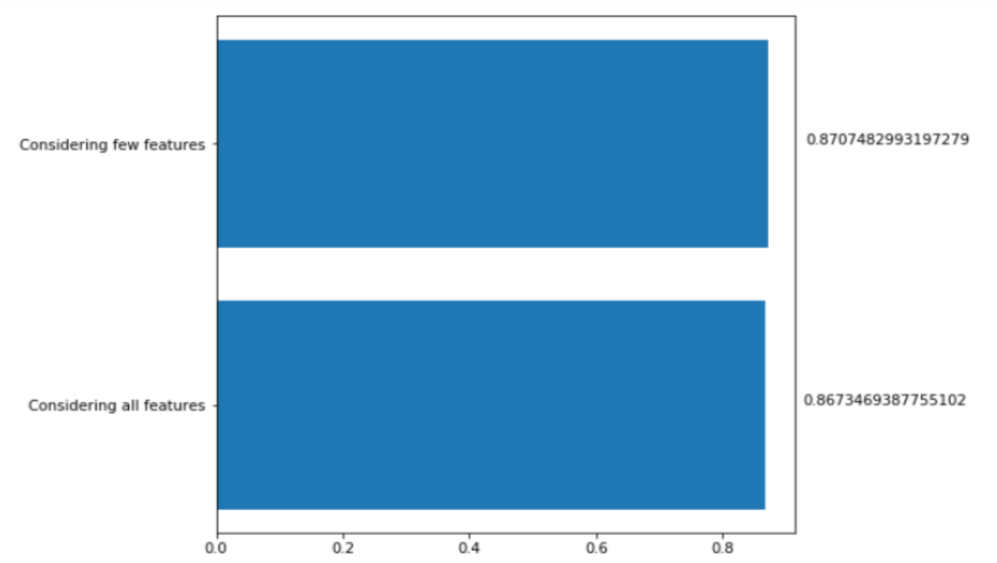


Fig 5. Accuracy w.r.t Top K and all the features

4. CONCLUSION

This paper identified the several key features that affect employee attenuation like monthly income, working hours, years at company. It has brought the importance of prediction model in attrition. In continuation, Data science process is proposed to address the issue. The experimental results of the proposed method shows the accuracy level of the various prediction methods in predicting whether employ leaves or not. And also it is investigated with accuracy levels of prediction methods with top-k features. In future work, the model needs to be tested on a larger dataset and other feature selection methods can also be considered as the present feature selection model.

REFERENCES

[1]. Abbasi, S. and Hollman, K. (2000), "Turnover: the real bottom line", Public Personnel Management, Vol. 29 No. 3, pp . 333-342

[2]. Abelson, M., B. Baysinger (1984), "Optimal and dysfunctional turnover: Toward an organizational level model," Academy of Management Review, Vol. 9 No.2, pp. 331–341.

[3]. A. C. Bluedron, "The theories of turnover: Causes effects and eaning", Research in the Sociology of Organization, vol.35,1982, pp. 135-153.

[4]. Boxall, P., Macky, K., & Rasmussen, E. (2003). Labor turnover and retention in New Zealand: The causes and consequences of leaving and staying with employers. Asia Pacific Journal of Human Resources, 41(2), 196- 214. http://dx.doi.org/10.1177/10384111030412006

[5]. F.E. Emery, & E.C. Trist, "The Causal Texture of organizations". Human Relations, vol. 18,1965, pp. 21-31

[6]. Firth, Lucy, Mellor, David, Moore, Kathleen A and Loquet, Claude (2004). How can managers reduce employee intention to quit?, *Journal of managerial psychology*, *19*(2), 170-187.

[7]. Hora Gurdeep S (2005),“ Retaining Talentinthe Knowledge Economy ”, Icfai HRM Review http://www.ircc.iitb.ac.in/~webadm/update/archives/ 1_ lssue2004/hr_management.html

[8]. Ibrahim M. M. Mitkees , Sherif M Badr , Ahmed Ibrahim Bahgat ElSeddawy "Customer churn prediction model using data mining techniques"https://ieeexplore.ieee.org/document/8289798

[9]. Khatri, N., C.T. Fern and P. Budhwar, (2001). Explaining Employee Turnover in An Asian Context. *Human Resource Management Journal*, *11*(1): 54-74.

[10]. Labov, B. (1997). Inspiring employees the easy way, *Incentive*, *171*(10): 114-18.

[11]. Liu, Y.A. and Wang, F. (2006). A Study on the Influence Factors of Employee Turnover Intention. *Enterprise Economy*, **6**, 42-44.

[12]. Ma, S.J., Chen, J.Q. and Wang, L. (2003). A Study on the Causes of Employee Turnover. *China Human Resources Development*, **9**, 18-20.

[13]. Magner, N., Welker, R and Johnson, G. (1996). The interactive effects of participation and outcome favorability in performance appraisal on turnover intentions and

evaluations of supervisors. *Journal of occupational organizational psychology*, 69: 135-143.

[14]. Mullins, J.L. (2005). *Management and Organizational Behavior*. 4th Edition. London: Pitman Publishing

[15]. Mathis, R.B. and Jackson, J.H. (2007). *Human Resource Management.* 10th Edition. Singapore: Thomson Asia Pty Ltd.

[16]. Oldham, G.R. and Hackman, J.R. (2005), "How job characteristics theory happened", in Smith, K.G. and Hitt, M.A. (Eds), Great Minds in Management: The Process of Theory Development, , Oxford University Press, New York, NY, pp . 151-170.

[17]. Owence, C., Pinagase, T.G. & Mercy, M.L. (2014). Causes and Effects of Staff Turnover in the Academic Development Centre: A Case of a Historically Black University in South Africa. *Mediterranean Journal of Social Sciences* MCSER Publishing, Rome-Italy, *5* (11), June 2014

[18]. Ramlall, S. (2003), "Managing employee retention as a strategy for increasing organizational competitiveness", Applied HRM Research, Vol. 8 No. 2, pp . 63-72.

[19]. Robbins, S.P. (2003). *Organizational Behavior*: *Concepts, Controversies and*

[20]. *Applications*. (8th Ed). London: Prentice Hall.

[21]. Sahu, A. and Gupta, M . (1999), "An Empirical Analysis of Employee Turnover in a Software Organization", Indian Journal of Industrial Relations, Vol 35, No. 1, pp 55-73.

[22]. Zhang, Y.J. (2016). A Review of Employee Turnover: Influence Factors and Countermeasure. *Journal of Human Resource and Sustainability Studies*, 85-91. http://dx.doi.org/10.4236/jhrss.2016.42010.

[23]. Zuber, A. (2001). A career in food service cons: high turnover, *Nations Restaurant News*,(21):147-148

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{\left(N_{e_t e_c} - E_{e_t e_c}\right)^2}{E_{e_t e_c}}$$