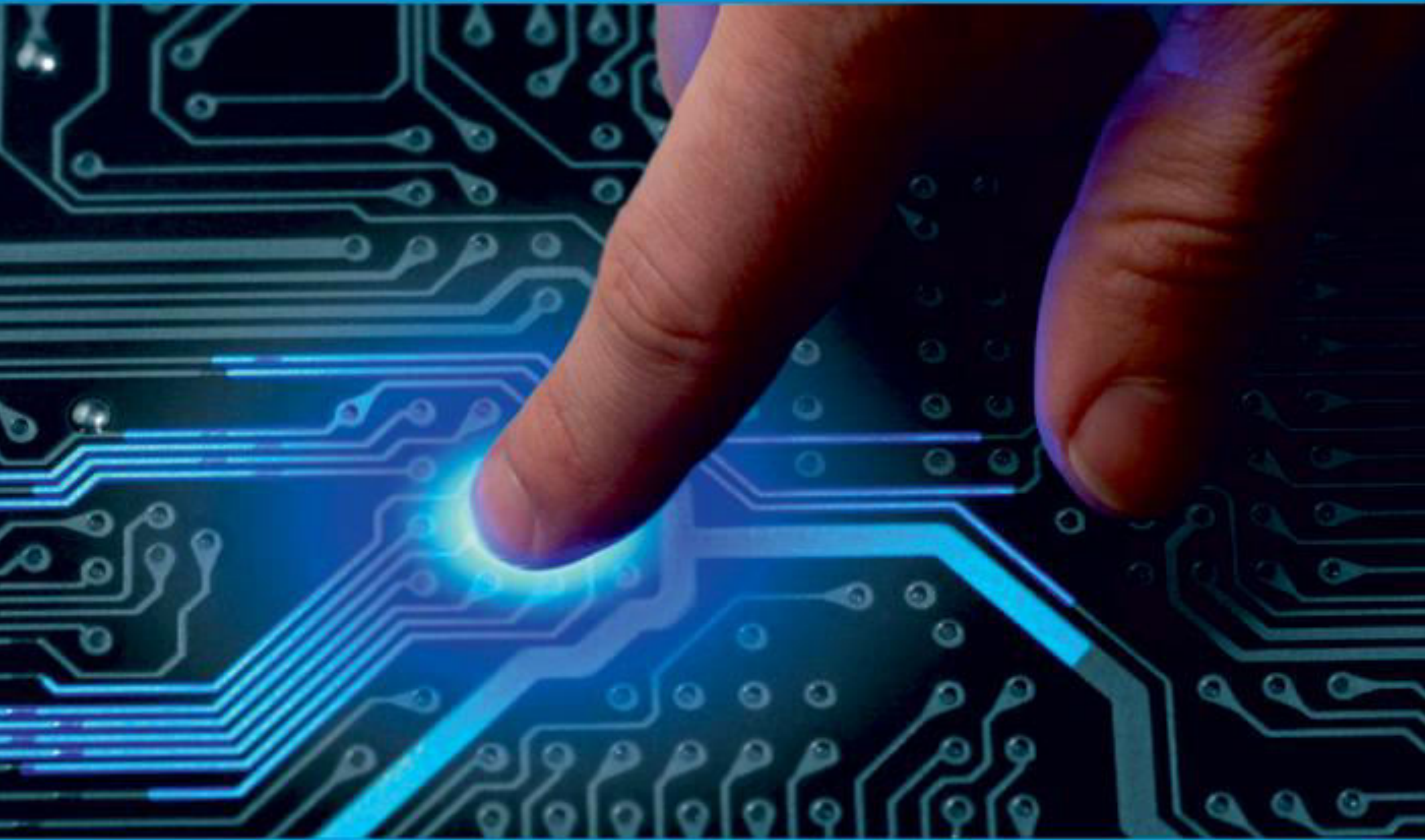
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 11, November 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

Employee Turnover Prediction Using Machine Learning

Prof.Neelam Gaikwad[1], Bhirud Gunjan[2], Barhe Harshal[3], Dendage Kishor[4], Shelar Pramod[5], Barole Chaitanya[6]

Assistant Professor, Dept. of Computer Engineering, GS Moze College of Engineering, Balewadi Pune, India.

UG Student, Dept. of Computer Engineering, GSMCOE, Savitribai Phule Pune University, India

ABSTRACT: Employee turnover is a big challenge to the organizations and companies. The employee turnover is the major issue in all organizations and companies. This model is used to predict the intensions of employee turnover during requirement process. The various algorithms are used in this model they are Logistic Regression Method, Random Forest Method, KNN, Extreme Gradient Boosting Method, Decision Tree Algorithm, Support Vector Machine, Naïve Bayes, and Linear Regression. The used data set includes the most essential features, which are considered during the requirement process of employee and may lead to turnover. This feature is salary, age distance from home, marital status, and gender. The KNN based model exhibit better performance in terms of accuracy probability percentage of turnover intentions of the workers. Therefore, the model can be used to aid human resource managers to make precautionary decisions; whether the candidate employee is likely to stay or leave the job, depending on the given relevant information about the candidate employee. To understand the employee needs from his salary. Try to predict his daily essentials his EMI's and other his daily needs, any difficulties is he facing during work. The company must look to the employee needs. The prediction means to know earlier about anything before it gets executed. From its behaviour needs intensions and many other things. In the human resources management, employee turnover is very important for the company operations since the leave of key employees can bring great loss to companies' .Employee turnover indicates the staffs decides to leave the company. Along with the fast development of economic and industries.

KEYWORDS*: Prediction Models, Employee Turnover, Machine Learning Algorithms*

I.INTRODUTION

The human resources department spends a lot of money and efforts on dealing with employee turnover, since the leave of excellent employees will cause huge losses to the company. Therefore, it is important to study and predict the turnover behaviour of employees. In recent years, there has been a massive increase in the competition among companies in sustaining in the business .The profits of the company can be improved by company efficiency. Staff retention is more important than acquisition of new staff. Employee turnover reflects the staffs decide to leave the company. There are a series of data, which records useful information of employee.

Turnover intention, which is an employee's reported will-in gness to leave the organization within a defined period of time, is considered the best predictor of actual employee turnover. n this paper, we model employee turnover intention using a set of traditional and state-of-the-art machine learning(ML) models and a unique cross-national survey collected by Effectory2 , which contains individual-level information. The survey includes sets of questions (called items) organized by themes that link an employee's working environment to her willingness to leave her work. Our objective is to train accurate predictive models, and to extract from the best ones the most important features with a focus on such items and themes. We train three interpretable (k-nearest neighbour, decision trees, and logistic regression) and four black-box (random forests, Boost, Light GBM, and Tab Net) classifiers. Wean laze the main features behind our two best performing models (logistic regression and Light GBM) across multiple folds on the training data for model robustness. We do so by ranking the features using a new procedure those agree-gates their model importance across folds.

Models have been built based on KNN and RF algorithms to predict the probability percentage of turnover intention of candidates before recruitment.The data was carefully studied and selected to include only the essential features. In this paper, we model employee turnover intention using a set of traditional and state-of-the-art machine learning.

The rest of this paper is organized as follows: Section1.Introduction; Section2. Related work; Section 3. Data mining; Section 4.Research Methodology; Section 5.Analysis; 6. Conclusion; 7. References. In the related work section we study about the work which we have done in this project. In the data mining section we study about the feature selection and feature engineering.

II. LITERATURE REVIEW

In this paper, modified approaches using various data mining techniques are collected to analyze the employee attrition rate at various levels. The study related to data mining for extracting the employee's attrition rate used in various models and the comprehensive literature review of various researcher's works are stated below; Qasem A, A.Radaideh and Eman A Nagi, has applied data mining techniques to build a classification model to predict the performance of employees. They adopted CRISP-DM data mining methodology [4] in their work. The Decision tree was the main data mining tool used to build the classification model, where several classification rules were generated. They validated the generated model; several experiments were conducted using real data collected from several companies. The model is intended to be used for predicting new applicants' performance.

Amir Mohammad EsmaieeliSikaroudi, RouzbehGhousi and Ali EsmaieeliSikaroudi et al, implemented knowledge discovery steps on real data of a manufacturing plant. They chew over many characteristics of employees such as age, technical skills and work experience. They used to find out importance of data features is measured by Pearson Chi-Square test. John M. Kirimi and Christopher Moturi et al, proposed a prediction model for employee performance forecasting that enables the human resource professionals to refocus on human capability criteria and thereby enhance the performance appraisal process of its human capital. RohitPunnoose and PankajAjit et al, explored the application of Extreme Gradient Boosting (XGBoost) technique which is more robust because of its regularization formulation. Data from the HRIS of a global retailer is used to compare XGBoost against six historically used supervised classifiers and demonstrate its significantly higher accuracy for predicting employee turnover.

Research Authors	Problem studied	Techniques Studied	Recommend
Jantan, Hamdan and Othman[9]	Data Mining techniques for performance prediction of employees	C4.5 decision tree, R.andom Forest, Multilayer Perceptron(MLP) and Radial Basic Function Network	C4.5 decision tree
Nagadevara, Srinivasan and Valk[10]	Relationship of withdrawal behaviors like lateness and absenteeism, job content, tenure and demographics on employee turnover	Artificial neural networks, logistic regression, classification and regression trees (CART), classification trees (C5.0), and discriminant analysis)	Classification and regression trees (CART)
Hong, Wei and Chen[11]	Feasibility of applying the *Logit*and *Probit*models to employee voluntary Turnover predictions	Logistic regression model (logit), probability regression model (probit)	Logistic regression model (logic)

Marjorie Laura Kane Sellers[12]	To explore various personal, as well as work variables impacting employee voluntary turnover		Binomial logit regression
Alao and Adeyemo[13]	Analyzing employee attrition using multiple decision tree algorithms	C4.5, C5, REPTree, CART	C5 decision tree
Saradhi and Palshikar[14]	To compare data mining techniques for predicting employee churn	Naïve Bayes, Support Vector Machines, Logistic Regression, Decision Trees and Random Forests	Support Vector Machines

III.RELATED WORK

We present the relevant literature around modelling and predicting turnover intention. Given our interdisciplinary approach, we group the related work by themes. The study of both actual and intended employee turnover has had a long tradition within the fields of human resource management and psychology. Traditional approaches for testing the determinants of employee turnover have focused largely on statistical significance tests via regression and ANOVA analysis, which are tools commonly used in applied econometrics. There has been a recent push for more advanced modelling approaches with the raise of human resource (HR) predictive analytics, where ML and data mining techniques are used to support HR teams. This paper falls within this line of work. Most ML approaches use classification models to study the predictors of turnover

The common approach among papers in this line of work is to test many ML models and to find the best one for predicting employee turnover. First, rather than reporting feature importance on a final model, we do so across many folds for the same model, which gives a more robust view on each feature’s importance within a specific model. Second, we go beyond the limited correlation-based analysis by incorporating causality into our feature importance analysis. Among the classification models used in the literature and from the recent state-of-the-art in ML, we will experiment with the following models: logistic regression, k-nearest neighbour, decision trees, random forests, XGBoost, and the more recent Light GBM, which is a gradient boosting method

Turnover data. Predictive models are built from survey data (questionnaires) and/or from data about workers’ history and performances (roles covered, working times, productivity). Given its sensitive information, detailed data on actual and intended turnover is difficult to obtain. Causal analysis. We note that this is not the first paper to approach employee turnover from a causality perspective, but, to the best of our knowledge, it is the first to do so using SCM. Other papers such as [25] and [48] use causal graphs as conceptual tools to illustrate their views on the features behind employee turnover. However, these papers do note Equip their causal models with any interventional properties. Some works, e.g., [4, 21, 61], go further by testing the consistency of their conceptual models with data using path analysis techniques. Still, none of these three papers use SCM, meaning that they cannot reason about causal interventions.

IV. DATA MINING

A.Overview of data mining.

The key of data mining is to deeply investigate/understand the data, including its type, characteristics, typical value, etc. Through deeply investigation on the data set or database, we can obtain the valuable information. Data mining technology can extract the knowledge, rules or high-level relationships from the database, including classification rules, clustering rules, association rules, and prediction rules and so on [21]-[23].In addition, data mining technique can assist researcher to analyse from multiple/different perspectives, so as to Obtain valuable information. The process of data mining is divided into data understanding, data cleaning, data mining and result processing

B.Key factors of data mining

There are various widely-used data mining algorithms, for example, C4.5, Means, SVM, Apriori, EM, PageRank, AdaBoost, KNN, Naive Bayes, CART

The four key factors to improve the accuracy of analysis Results are as follows:
1) Understanding of analytical data
2) The analysis and processing of error index in data
3) Feature engineering
4) Model fusion

C.The importance of data processing

Through the feature extraction, we can obtain unprocessed features. These unprocessed features have the following characteristics:
1) Dimensional disunity: features of different specifications cannot be compared together. Dimension less can deal with this problem.
2) Information redundancy: for some quantitative characteristics, the valid information contained is interval division, such as academic achievement. If only "pass" or “fail" is concerned, then quantitative test scores need to be converted to either "1" or "0", which indicates "pass" and" fail", respectively. We can use Binarization to deal with this problem.
3) Qualitative features cannot be directly used: some machine learning algorithms and models can only accept the input of quantitative features, so it is necessary to convert qualitative features into quantitative features. Dummy coding is usually used to convert qualitative features into quantitative features. Suppose there are Qualitative values, then this feature is extended to N kinds of features. When the original eigenvalue is the first qualitative value, the first extended eigenvalue is assigned as “1” and the other extended eigenvalues are assigned as“0”. The dumb coding method does not need to increase the work of tuning parameters, compared with the directly specified model. For the linear model, the nonlinear effect can be achieved by using the dumb coding feature.
4) Missing values: missing values need to be supplemented.
5) Low information utilization: different machine learning algorithms and models make different use of information in the data. In the linear model, the nonlinear effect can be achieved by using the dummy coding of qualitative features. Similarly, the polynomial of quantitative variables can achieve nonlinear effect.

V.RESEARCH METHODOLOGY

The turnover prediction framework is presented inFig.1. The use methodology comprises several phases; namely, data collection, data cleaning, data selection data pre-processing, benchmarking the algorithms, and evaluating the predicted outcomes.

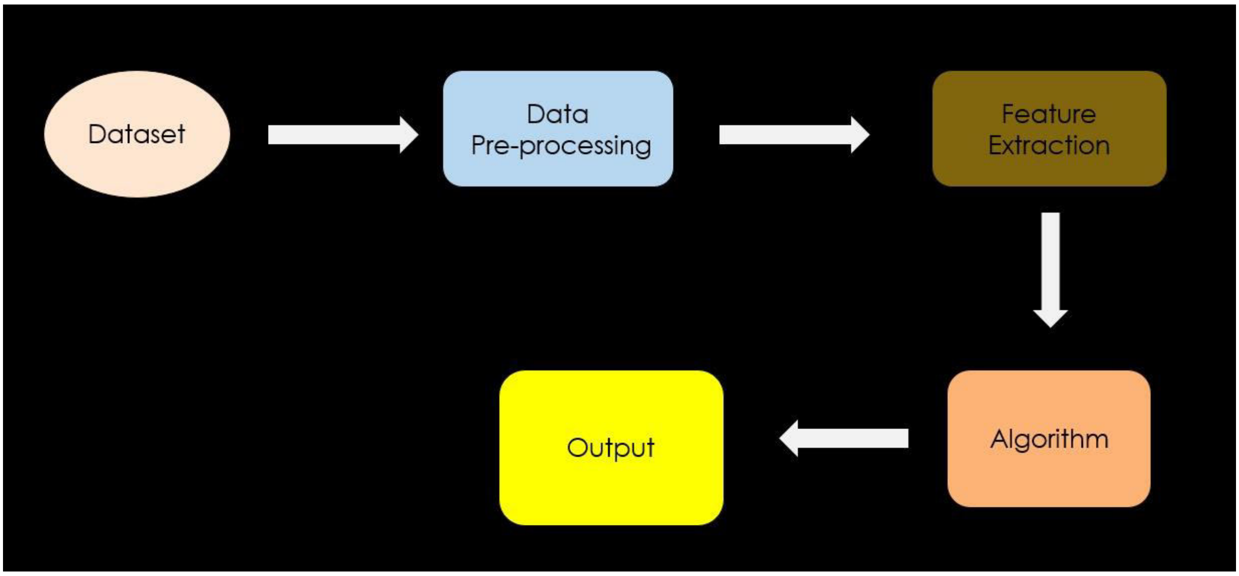


Fig. Turnover Prediction Framework

A.Dataset

The dataset was obtained from the Kaggle website (IBM, 2020). The IBM dataset comprises 1470 records with 34 features (6 categorical and 27 numeric), such as monthly salary, experience, distance from home, skills, nature of work, position etc.

B.Data Pre-Processing

Machine Learning algorithms can typically process only numerical input. Hence, the qualitative variables (gender and marital status) were encoded into quantitative variables (One-Hot Encoding) to input an acceptable format for the machine.

VI. ANALYSIS

We found LGBM and LR to be the best performing models for predicting turnover intention, and in Section studied the driving features behind the two models. We also observe this in the data. In particular ,the Country attribute is correlated to each of the themes: the nonparametric Kruskal–Wallis H test [32] shows a p-value close to 0 for all themes, which means that we reject the null hypothesis that the scores of a theme in all countries originate from the same distribution

As a statistical analysis technique, survival analysis studies the time-dependent turnover event occurrence probability by considering both events and time. Thus, censored data can be fully used without a strong distribution hypothesis.

Basic Concepts: We introduce some basic concepts about survival analysis here.

Time: the time from the beginning of an observation to the event occurrence or the end of an observation. Event: death, failure, turnover, or other events of interest occur during an observation.

Censor: the non-occurrence of an event during an observation. The right censored data indicates that the observed object has left before the event occurrence and then the observation ends. The left censored data indicates that the event occurs before the object is observed.

VII. CONCLUSIONS

There is a positive correlation between distance from home and employee turnover, signifying that as the distance from home increases, the employee turnover increases. However, there is a negative correlation be-tween age and employee turnover, indicating that as age decreases, employees are more likely to leave the job. The percentage of employee turnover is inversely proportional to the salary. Single employees show a higher desire to leave, but divorced employees are more likely to stay in the job, where the highest turn-over percentage occurs in the single employees and the lowest turnover percentage occurs in the divorced employees.

Salary exhibits the highest feature affecting the turnover of employees and gender has the lowest effect.

REFERENCES

1. S. Pandey, P. Khaskel, “Application of AI in Human Resource Management and Gen Y"s Reaction”, International Journal of Recent Technology and Engineering, Vol. 8, 2019. pp. 10325-10331.
2. N. Govindaraju, “Demographic Factors Influence on Employee Retention”, International Journal of Engineering Studied and Technical Approach, Vol.7, 2018, pp.10-20.
3. R. Jesuthasan, “HR’s New Role: Rethinking and Enabling Digital Engagement”, Strategic HR Review, Vol. 16, No. 2, 2017, pp. 60-65.
4. Hassan, S.: The importance of role clarification in works Groups: Effects on perceived role clarity, work satisfaction, and turn overrates. Public Adm. Rev. 73(5), 716–725 (2013).
5. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, pp. 1137–1145. Morgan Kaufmann (1995)
6. Gabrani, G., Kwatra, A.: Machine learning based predictive model for risk assessment of employee attrition. In: ICCSA (4), Lecture Notes in Computer Science, vol. 10963, pp. 189–201. Springer (2018).
7. Angrist, J.D., Pischke, J.S.: Mostly Harmless Econometrics. Princeton University Press (2008).
8. Z. Han, et al. 2012. LTE FDD technology principle and network planning. Beijing: China Post and Telecommunications Press.
9. A. Y. Al-Dubai, L. Zhao, et al. "QoS-aware inter-domain multi cast for scalable wireless community networks," IEEE Transactions on Parallel and Distributed Systems, 26(11), pp.3136-3148, November 2015.
10. J. Liu, Y. Long, M. Fang, R. He, T. Wang, and G. Chen, “Analysing Employee Turnover Based on Job Skills.” ACM, 2018, pp. 16–21.
11. Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, “Employee Turnover Prediction with Machine Learning: A Reliable Approach. Springer, 2018, pp. 737–758.
12. A. C. C. de Jesus, M. E. G. Jnior, and W. C. Brando, “Exploiting linked in to predict employee resignation likelihood.” ACM, 2018, pp. 1764–1771

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING