2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT 2021)

Bhopal, India
18 – 19 June 2021

Pages 1-447

IEEE Catalog Number: CFP2118P-POD
ISBN: 978-1-6654-4624-2

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP2118P-POD
ISBN (Print-On-Demand):	978-1-6654-4624-2
ISBN (Online):	978-1-6654-2306-9
ISSN:	2329-7182

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

2021 10th IEEE International Conference on Communication Systems and Network Technologies

CSNT 2021

Table of Contents

Track-B: Computer Networking, Wireless Communication, Ad Hoc Networking and Vehicular Networking

Track C: Data Mining, Big Data, Text & web mining and Social Networking

Track D: Digital System Design, Fabrication and Robotics

Track E: Electrical Power System Design, Grid and Solar Devices

Track F: Image Processing, Signal Processing and Applications

Track -G: Internet of Things and Sensor Networks

Track -H: Machine Learning, Deep Learning and Soft Computing

Track -I: Cryptography, Security and BlockChain

Track -J: Intelligent Systems, Software Testing and IT Applications

Author Index

2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT) | 978-1-6654-2306-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/CSNT51715.2021.9509601

A Novel Approach to Improve Employee Retention Using Machine Learning

Abhay Charan Patro
Department of CSE
Birla Institute of Technology Mesra
Ranchi,India
abhaycharanpatro@gmail.com

Saad Aziz Zaidi
Department of CSE
Birla Institute of Technology Mesra
Ranchi,India
seanazarrin@gmail.com

Aaradhya Dixit
Department of ECE
Birla Institute of Technology Mesra
Ranchi,India
aaradhyadixit@ieee.org

Manish Dixit
Department of CSE
Madhav Institute of Technology and Science
Gwalior.India
manishdixit@ieee.org

***Abstract*—Employee attrition is one of the biggest challenges faced within business organizations. With the advancement in machine learning, various research papers brought into light the numerous reasons why an employee leaves, this paper puts forward a machine learning pipeline that not only predicts employee attrition but also suggests a minimum cost approach for the company so that the employee does not leave.**

Keywords-Attrition; HumanResource; Machine Learning; Slack features; Willingness factor; H2O AutoML; Random Forests;Extremely Randomized Trees (XRT); Generalized Linear Models; XGBoost GBM; H2O GBM; Deep learning

I. INTRODUCTION

As defined by FurstPerson[1], attrition is a reduction of staff be it, voluntary or involuntary. Attrition is usually viewed negatively and as a burden for employers. According to FurstPerson, large enterprises, more often than not, don't calculate the turnover rate, downplaying it's costs. In a study conducted by furstperson[1], the financial toll employee turnover can take on an organization, in general, across several major industries, can cost anywhere from $1,500 to $16,650 per agent. According to salary.com's Compdata surveys [10], the total attrition in 2018 was 19.3% out of which 14.2% was voluntary in nature.The attrition in 2018 saw a 3.6% increase compared to 2014's total attrition of 15.7%.

Machine learning seems to be an ever progressing field that shows no signs of slowing down, with the monumental leaps in the way we recognize, process and find patterns in data, lost dreams seem within reach. Machine learning has taken care of predicting employee attrition.

Using these predictive models made and machine learning as our foundation, we tried to answer the question of what specific changes would an employee need to prevent them from leaving the company and thus reducing the attrition rate.

II. PROBLEM

The Question of why an employee Quits [2] has been answered in quite an extensive manner, with various dependency graphs, correlation matrices and the like. The topic is quite important and thus multiple methods like decision trees[3],data-mining techniques [4][5] etc, have been employed for predictions, but all they could do was predict or try and show what factors had the most impact on the quitting employee. Here we try to build on our predecessors' incredibly insightful work to answer the pertinent question :-

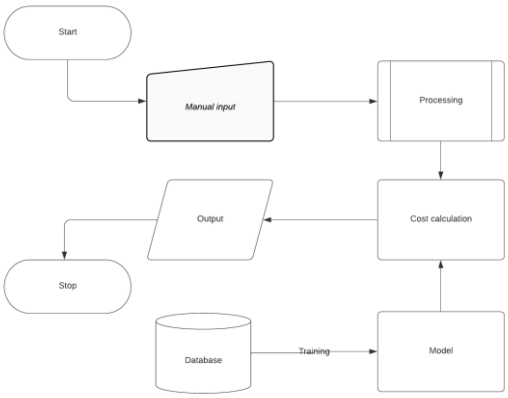
What changes could have stopped the quitting employee?

III. CONCEPT

This paper has been inspired by Google's deep dream[9], wherein, using gradient ascent you try to change not the weights but the input itself to increase the activation of nodes, essentially flipping the neural net. Our concept for this pipeline, with its processes to change the input features for the desired outcome, was originally intended for a neural net, but owing to insufficient data, it was adapted for other models. This research seeks to help employers retain their employees, significantly reducing costs for the business organization in the long run.

We propose a pipeline that trains on the data of the past, predicting the future, while providing an appropriate minimum cost of, if you so wish, averting said future. In the context of our paper, this avertable future is Employee attrition. Here, we know the desired outcome, what we don't, is the change in our input that'll bring about the desired outcome to fruition

978-1-6654-2306-9/21/$31.00 ©2021 IEEE
DOI: 10.1109/CSNT.2021.119



IV. METHODOLOGY

Figure 1. Flowchart of the proposed pipeline

A. Pre-Processing

The data used for this purpose was “IBM HR Analytics Employee Attrition & Performance” [6]. It consisted of 34 features and 1470 synthetic examples. The feature type and the number of unique values in them is given in Fig. 2.

Employee count, Over18 and StandardHours were removed from the feature list because they all had the same value in all the training examples. EmployeeNumber was also removed from the feature list because it was just a unique identifier for each employee. Moreover 'DailyRate','HourlyRate','MonthlyRate' were also removed because of their relation with MonthlyIncome.This brought the feature list down to 27 features.

Then the features 'BusinessTravel’, ’Gender’, ’OverTime’, ’Marital Status’ were label encoded and the features ‘Department’, ’EducationField’, ’JobRole’ were one-hot encoded.

The data was then randomly shuffled to a 80:20 split for training and testing respectively.

B. Training

Trying to find a model that could most effectively help us answer our question was tedious, we tried and tested multiple classification algorithms to train our model to accurately classify employee attrition, recreating results from [7] was the first step, we then moved on to change our pre-processing techniques and could obtain a much better AUC, precision, recall and F1_score in all our testing algorithms, but it was not enough, bear in mind the results by others had been astonishing as well given that the dataset was so imbalanced, an accurate representation would be, for every ‘YES’ there were almost 15 ‘NO’s. With odds like these an F1_score of ~0.5 was good.

But we decided to train various models using H2O’s Auto-ML. H2Os Auto-ML is an algorithm for automatic machine learning on tabular data, part of the H2O machine learning platform[8].

Feature	Feature Type	Unique values
Age	Numerical	43
BusinessTravel	Numerical	3
DailyRate	Numerical	886
Department	Categorical	3
DistanceFromHome	Numerical	29
Education	Numerical	5
EducationField	Categorical	6
EmployeeCount	Numerical	1
EmployeeNumber	Numerical	1470
EnvironmentSatisfaction	Numerical	4
Gender	Categorical	2
HourlyRate	Numerical	71
JobInvolvement	Numerical	4
JobLevel	Numerical	5
JobRole	Categorical	9
JobSatisfaction	Numerical	4
MaritalStatus	Categorical	3
MonthlyIncome	Numerical	1349
MonthlyRate	Numerical	1427
NumCompaniesWorked	Numerical	10
Over18	Categorical	1
OverTime	Categorical	2
PercentSalaryHike	Numerical	15
PerformanceRating	Numerical	2
RelationshipSatisfaction	Numerical	4
StandardHours	Numerical	1
StockOptionLevel	Numerical	4
TotalWorkingYears	Numerical	40
TrainingTimesLastYear	Numerical	7
WorkLifeBalance	Numerical	4
YearsAtCompany	Numerical	37
YearsInCurrentRole	Numerical	19
YearsSinceLastPromotion	Numerical	16
YearsWithCurrManager	Numerical	18

Figure 2. Type of Feature and its unique values

model_id	auc	logloss	aucpr	mean_per_class_error
XGBoost_grid__1_AutoML_20210319_095526_model_1	0.837357	0.319802	0.609676	0.238802
StackedEnsemble_BestOfFamily_AutoML_20210319_095526	0.83687	0.301567	0.652458	0.230209
StackedEnsemble_AllModels_AutoML_20210319_095526	0.83056	0.308226	0.636741	0.229186
XGBoost_grid__1_AutoML_20210319_095526_model_9	0.82997	0.324921	0.606539	0.256299
XGBoost_grid__1_AutoML_20210319_095526_model_5	0.827343	0.325788	0.609031	0.247814
XGBoost_grid__1_AutoML_20210319_095526_model_3	0.822915	0.328112	0.598294	0.228271
GLM_1_AutoML_20210319_095526	0.822277	0.330172	0.628944	0.269102
XGBoost_2_AutoML_20210319_095526	0.821109	0.328804	0.589495	0.272386
GBM_5_AutoML_20210319_095526	0.821055	0.337032	0.5835	0.284165
GBM_grid__1_AutoML_20210319_095526_model_1	0.820527	0.340547	0.550585	0.270857
XGBoost_grid__1_AutoML_20210319_095526_model_10	0.819685	0.335535	0.5676	0.261984
GBM_grid__1_AutoML_20210319_095526_model_2	0.818686	0.333837	0.583345	0.288095
GBM_grid__1_AutoML_20210319_095526_model_7	0.817499	0.339303	0.578365	0.272623
XGBoost_3_AutoML_20210319_095526	0.817243	0.343133	0.576583	0.239663
XGBoost_1_AutoML_20210319_095526	0.816952	0.33673	0.584532	0.245381
XGBoost_grid__1_AutoML_20210319_095526_model_11	0.815224	0.333925	0.585591	0.238145
XGBoost_grid__1_AutoML_20210319_095526_model_7	0.814791	0.343491	0.560935	0.281118
GBM_grid__1_AutoML_20210319_095526_model_5	0.813283	0.344926	0.573433	0.284531
GBM_grid__1_AutoML_20210319_095526_model_6	0.812812	0.363997	0.526772	0.274012
GBM_2_AutoML_20210319_095526	0.812484	0.344781	0.533989	0.270964
XGBoost_grid__1_AutoML_20210319_095526_model_4	0.809135	0.346493	0.562634	0.24158
DeepLearning_1_AutoML_20210319_095526	0.808508	0.353982	0.516117	0.267777
GBM_4_AutoML_20210319_095526	0.80513	0.353018	0.520182	0.261327

Figure 3. A few top models of the leaderboard of the various models generated using H2O's AutoML

Model ID	Importance
GLM_1_AutoML_20210319_095526	0.6782996382634335
XGBoost_grid__1_AutoML_20210319_095526_model_1	0.3964367864527333
XGBoost_2_AutoML_20210319_095526	0.13192081915394016
XGBoost_grid__1_AutoML_20210319_095526_model_5	0.10976832022050043
XGBoost_1_AutoML_20210319_095526	0.06323860700302551
XGBoost_grid__1_AutoML_20210319_095526_model_3	0.062410970916512085
XGBoost_3_AutoML_20210319_095526	0.04553848325053091
DeepLearning_1_AutoML_20210319_095526	0.02526754006536049
GBM_grid__1_AutoML_20210319_095526_model_5	0.023648664303890343
XGBoost_grid__1_AutoML_20210319_095526_model_2	0.022977488834341198
DeepLearning_grid__3_AutoML_20210319_095526_model_2	0.01894327135590139
DeepLearning_grid__2_AutoML_20210319_095526_model_2	0.014568108799226352
XGBoost_grid__1_AutoML_20210319_095526_model_6	0.008854624036856702

Figure 4. The importance of various models in the stacked ensemble model made using all the 40 models

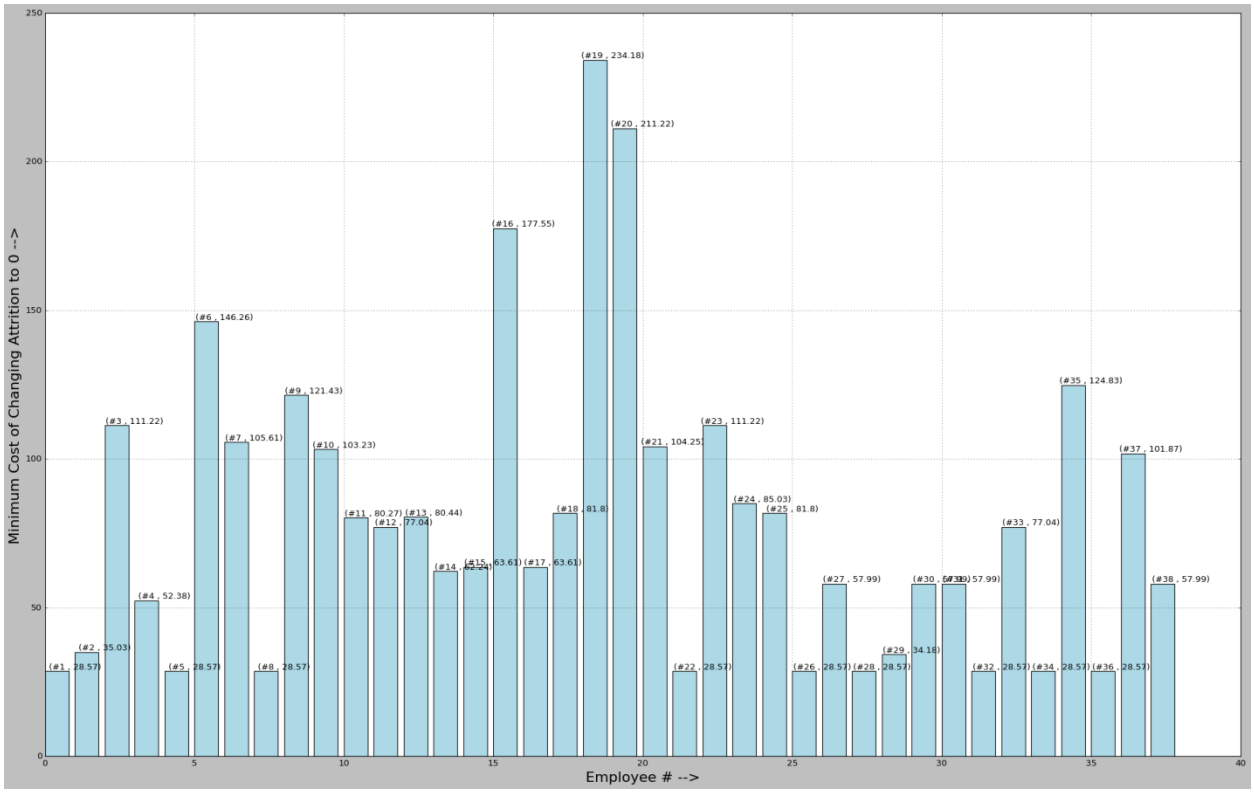


Figure 5. Minimum cost required to retain various employees

40 different models were trained using various algorithms such as Random Forest and Extremely Randomized Trees (XRT) models,Generalized Linear

Model	F1 Score
XGBoost_grid__1_AutoML_20210319_095526_model_1	0.4943820224719101
StackedEnsemble_BestOfFamily_AutoML_20210319_095526	0.6021505376344085
StackedEnsemble_AllModels_AutoML_20210319_095526	0.6086956521739131

Figure 6. F1 Score of the top 3 models of the leaderboard

Models,XGBoost,GBMs, and Deep Learning models. Apart from the 40 models, two more stacked ensemble models were created, one was created using the best model of the family and the other one using all 40 models.

From the leaderboard (Fig. 3) we calculated the F1 scores for the top 3 models (Fig. 6) and chose to use a stacked ensemble of all models as the main reference model as it gave the best F1 score. Then the attrition predictions were made on the test set.Now the main goal was to find the minimum cost of changes that needed to be done on a set of features so that an employee, predicted to leave, wouldn't.

C. *Cost calculation*

The feature-set would be decided by the business organization, and the changes they are willing to make. We refer to these features as slack features.

Slack Feature	Willingness Factor
EnvironmentSatisfaction	0.21
JobLevel	0.14
StockOptionLevel	0.28
PercentSalaryHike	0.21
WorkLifeBalance	0.14

Figure 7. Slack Features and their corresponding willingness factors used for this paper

Each slack feature has a willingness factor ω associated with it. The willingness factors are calculated by taking input from the business organisation, willingness percentages of the slack features, these describe the "intent" of the

organisation, in making changes to those slack features. Then the willingness percentages are normalized so that they all add up to 1, this gives us the willingness factors. For the purpose of this paper we chose the features and willingness factors as shown in Fig. 7. Then we needed to solve the problem of choosing the best set of changes from the many possible and to do that we came up with the following minimization goal :

Cost required for employee e:

$$min_{for\ all\ changes}\ (\ \sum_{x\ \epsilon\ slack\ features} (1 - \omega_x) * \Delta x_e\ )$$

We used this minimization goal individually for each employee to find the best set of feature changes that would suit the employee.

V. FUTURE PROSPECTS

The cost we calculated was minimized over the cost of changes in the various possible solutions. This cost has no real relation to a monetary value that the company might incur to prevent attrition, albeit this is what was originally envisioned for the pipeline, but couldn't be applied due to the reluctance of companies, with regards to data and varied ways in which accounting is done for a particular company. For a company that has records of its employees and their CTCs, combining these datasets should be no problem, we would then be able to calculate the exact monetary cost required for preventing employee attrition, given past accounting data, this could then be taken further, using real-time comparison of incurring cost, with other viable options to decide on the change.

This could be addressed in successive works.

VI. CONCLUSION

In this paper a solution to increasing attrition rates through machine learning is discussed by answering the question, what would make the employee stay ? Through this paper, we have tried to introduce a Machine Learning Pipeline, which business organizations can utilise by choosing their own slack features and willingness factors, complemented with a large database, to drastically reduce employee attrition.

REFERENCES

[1] S. Colberg, "What is Attrition & Attrition Rate, What Causes Attrition and Cost of Employee Turnover", Furstperson.com, 2021. [Online]. Available: https://www.furstperson.com/blog/causes-employee-attrition-cost-employee-attrition

[2] Frye A, Boomhower C, Smith M, Vitovsky L, Fabricant S. Employee Attrition: What Makes an Employee Quit?. SMU Data Science Review. 2018;1(1):9.

[3] Alao, D.A.B.A. and Adeyemo, A.B., 2013. Analyzing employee attrition using decision tree algorithms. Computing, Information Systems, Development Informatics and Allied Research Journal, 4(1), pp.17-28.

[4] Shankar RS, Rajanikanth J, Sivaramaraju VV, Murthy KV. Prediction of employee attrition using datamining. In2018 ieee international conference on system, computation, automation and networking (icscan) 2018 Jul 6 (pp. 1-8). IEEE.

[5] Yadav S, Jain A, Singh D. Early prediction of employee attrition using data mining techniques. In2018 IEEE 8th International Advance Computing Conference (IACC) 2018 Dec 14 (pp. 349-354). IEEE.

[6] "IBM HR Analytics Employee Attrition & Performance", Kaggle.com, 2017. [Online]. Available: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[7] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, "Predicting Employee Attrition Using Machine Learning Techniques," Computers, vol. 9, no. 4, p. 86, Nov. 2020

[8] v.3.32.0.5 Erin LeDell and Sebastien Poirier. H2O AutoML: Scalable Automatic Machine Learning. 7th ICML Workshop on Automated Machine Learning (AutoML), July 2020. URL https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf

[9] "Inceptionism: Going Deeper into Neural Networks", Google AI Blog, 2021. [Online]. Available: https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

[10] "Compdata 2018 Turnover Report", Www2.salary.com, 2021. [Online]. Available: https://www2.salary.com/turnover