Employee Turnover and Customer Churn Prediction Using Machine Learning

Belal Adel [1], Habiba Amr[2], Noran Essam[3], Rana Dahshan [4]
Faculty of Computer Science
Misr International University, Cairo, Egypt
belal2011213[1],habiba2008121[2], noran2007486[3], rana2015051[4]{@miuegypt.edu.eg}

***Abstract*—This research proposes new algorithms used to detect employee turnovers using machine learning through different algorithms. It compares different algorithms efficiencies to each other's. Different types of classifiers in machine learning are used including KNN, SVM, Naive-Bayes, Logistic Regression, Random Forest, AdaBoost, Decision tree. The result tables are classified according to AUC, CA, FI, Recall and Precision.**

I. Introduction

Employee turnover is when an employee decides to resign and leave the company. Similarly, customer churning occurs when a customer decides to stop dealing with a business. This is also called customer attrition, which is when someone chooses to stop using the company's products or services, which basically means the customer stops being a customer. It can be measured using the customer churn rate. These happen to be serious problems for companies. In the case of employee turnover, the loss of talented employees usually sets the company back, and failing to retain customers decrease the company's competitiveness. Churn is an act of collective turnover in a company's staff or customers. Existing or current employees leave the company and new employees are hired. The churn rate is usually calculated as the percentage of employees leaving the company over some specific time. Predicting the turnover will save the company lots of money, and can also help it avoid the phenomenon, and analyze the characteristics of the turnover. Therefore, the company collects data about employees and customers to be able to predict the turnover. Using machine learning, we propose a way for these companies to predict the turnover, we also use datasets collected by these companies to explain how this way or model works in real life and how it would be useful for them to use it to develop better recruitment plans and promotion rules. We mainly focus on voluntary turnover which happens when the employee voluntary resigns or leaves the company.
The remainder of this paper is organized as follows: Section 2, discusses related work. In Section 3, we describe our methodology for this study, and our approach in detail. In Section 4, we discuss the results of our experiments. Finally in Section 6, we discuss the conclusion.

II. Related work

Ming Fang, Jiahao Su, Jiamin Liu, Yuxi Long, Renjie He, and Tao Wang studied the event of employee turnover on a Chinese state-owned enterprise and proposed a method to predict employee turnover rate using a conditional semi-Markov (SMK) model for calculating the conditional amount of employee turnover. They demonstrated how the model works in reality and how it can support a management system when making any human resources-related decisions. [1]

Manas Rahman and V Kumar at the Department of Computer Science at the Central University of Kerala in Periye, Kasaragod have observed the struggles of banking systems to retain their clients and keep them engaged. They used a churn modeling dataset of banking records where they selected relevant and appropriate features that describe the clients's relation to their bank and trained several classifiers to predict whether or not a client is likely to exit. [2]

This research studies the problem of employee turnover from a completely new perspective by modeling users' historical job records as a dynamic bipartite graph. Specifically, they propose a bipartite graph embedding method with temporal information called dynamic bipartite graph embedding (DBGE) to learn the vector representation of employees and companies. Their approach incorporated temporal information embedded in consecutive work records. Experiments show that Their approach achieves better performance in the link prediction and visualization task than other graph embedding methods that don't consider temporal information. [3]

Heng Zhang, Lexi Xu, Xinzhou Cheng, Kun Chao, Xueqing Zhao, from the Network Technology Research Institute, China United Network Communications Corporation, Beijing, P.R.China used a machine learning technique to sort out the characteristics of employee turnover, they used the GBDT algorithm and LR algorithm to fit the characteristic model which influences employee turnover phenomenon. They carry out the employee turnover prediction for real companies which provides a reference for other companies to help them reduce the employee turnover rate of their employees. [4]

Researchers from Chongqing University and Guilin University of Electronic Technology in China have tackled this field where they focused on historical events of turnover behaviors of employees and its longitudinal data instead of focusing on employee centered turnovers. The researchers

proposed a new prediction algorithm called Cox-RF that views it more from an event centered perspective. It combines statistical results of survival analysis with ensemble learning and so simplifies the problem to a much more traditional supervised binary classification problem that's well known to be solved. [5]

III. Proposed Methodology

We attempt to fit each of Logistic Regression, Ada boost, naive bayes, KNN, Decision Tree, Random Forest, and Support Vector Classifier on three datasets. The first dataset contains 10000 sample records of a banking system's clients in 10 features and whether or not they have left the bank. [6] The second dataset contains 15000 sample records from the human resources department of a corporation in 9 features and whether or not they resigned. The third dataset contains 7044 reocrds of a telco company's customers in 19 features and whether or not they churned (unsubscribed or left). [7]

We created a testing framework that standardized the datasets and testing conditions for the different models. The three datasets needed preprocessing where categorical values had to be ranked, quantized, or encoded into numeric values for the machine learning models to use. The preprocessed dataset was then split into training and testing records where training records were used to fit the models and testing features were inputted into the model to classify them. The resulting predictions are then compared to the true label values to construct a confusion matrix and then score the model's performance using different measures of accuracy.

IV. Experimental results

This experiment was conducted using a testing environment we created in python.

A. Datasets and resources

The trained models were used to predict how likely a customer would churn or how likely an employee would quit according to the given dataset.

B. Used classifiers and their main advantage

1) k-Nearest Neighbours:
The k-Nearest Neighbors classifier learning method that keeps all of the training data for classification.
As shown in equation 1 the knn use this equation to be calculation the shortest distance between the 2 points where the (K) identify how many distances to be taken in consideration and the k must be an odd value where even can cause a an issue cause it may turn out to be the both equal[8].

$$\sqrt{\sum_{i=1}^{k}\left(x_i - y_i\right)^2} \tag{1}$$

2) Support Vector Machine:
SVC is a classifier that is defined informally as separating hyperplanes. A hyperplane is a one-dimensional subspace that is smaller than its surrounding space. This indicates that a two-dimensional hyperplane is a one-dimension separator (line). A two-dimensional separator is a three-dimensional hyperplane (plane)
Points above the hyperplane Satisfy

$$W^T x + b > 0 \tag{2}$$

Points bellow the hyperplane Satisfy

$$W^T x + b < 0 \tag{3}$$

we can calculate the support vector machine using the following equation equation 4

$$\max_{\alpha} W(\alpha) = \sum_{k=1}^{m} \alpha_k - \frac{1}{2} \sum_{k,\ell}^{m} \alpha_k \alpha_\ell y_k y_\ell \left(K\left(\mathbf{x}_k, \mathbf{x}_\ell\right) + \frac{1}{C} \delta_{k,\ell} \right) \tag{4}$$

3) Naïve Bayes:
The nave Bayes model is a Bayesian probability model that has been drastically simplified. Consider the likelihood of an end result given numerous associated evidence variables in this model. The likelihood of the end outcome, as well as the chance of the evidence variables happening if the final result happens, is stored in the model. The chance of one evidence variable occurring in the presence of the end result is considered to be independent of the probability of other evidence variables occurring in the presence of the end result.
The Naïve Bayes get calculated by the following equation in equation 5 [9]

$$\mathrm{P}(A \mid B) = P(B \mid A)^* P(A)/P(B) \tag{5}$$

4) logistic Regression:
Logistic regression is a analytic technique for predicting a binary answer, based on prior observations of a data set. A logistic regression model can predict a dependent data variable by studying the relationship between one or more variables.
logistic Regression can be calculated using the following formula includes in equation 6

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta(\text{ Age }) \tag{6}$$

5) Random forest:
   a) In Random Forest, n random records are chosen at random from a data
   b) For each sample, a separates decision tree is assembeled
   c) Each decision tree will provide a result.

d) Final output is considered based Average for Clasification and regresion respectively.

Random Forest can be calculated using the following formula includes in equation 7 [10]

$$H(N) = -\sum_{i=1}^{i=d} P(\omega_i) \log_2 (P(\omega_i)) \quad (7)$$

6) Adaboost: AdaBoost (Adaptive Boosting) fits a series of weak learners by using variable weighted training data. It predict the originale dataset and equaly weighting all of the observation. If the first learner's prediction is incorrect, the observation that was incorrectly predicted is given greater weight.
adaboost can be calculated using the following formula includes in equation 8:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \quad (8)$$

7) Decision tree: A decision tree is a sort of supervised machine learning that predicts output based on the answers to a series of questions.
Decision tree can be calculated using the following formula includes in equation 9:[10]

C. Results

1) Accuracy:
The accuracy is calculated by adding the true postive to the the true negative and all of that divded by the sum of them all which are the true postive and truse negtive and false postive and negtive which are showed on equation 9.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{[11]} \quad (9)$$

2) Precision:
The Precision is calculated by getting the true postive and divideing it byt the sum of the postive part the true and the negive one which are showed on equation 10

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (10)$$

3) Recall:
The Recall is calculated by getting the true postive and divideing it byt the sum of the true postive and false negtive which are showed on equation 11

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (11)$$

4) F1 Score:
The f1 is calculated by getting the precision and multiplying it with the recall then divide it by the addition of them and multiply the whole outcome by 2 which are showed on equation 12

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

TABLE I
TESTING RESULTS FOR DATA SET 1

	AUC	CA	F1	Recall	Precision
Logistic Regression	0.527	0.803	0.129	0.075	0.468
KNN	0.504	0.74	0.148	0.116	0.206
SVM	0.57	0.45	0.353	0.768	0.229
Decision Tree	0.686	0.793	0.49	0.509	0.473
Random Forest	0.705	0.859	0.556	0.451	0.725
Gaussian Naive Bayes	0.521	0.795	0.12	0.072	0.368
Ada Boost	0.706	0.856	0.556	0.461	0.699

TABLE II
TESTING RESULTS FOR DATA SET 2

	AUC	CA	F1	Recall	Precision
Logistic Regression	0.607	0.77	0.379	0.298	0.52
KNN	0.948	0.946	0.893	0.952	0.84
SVM	0.701	0.696	0.524	0.709	0.415
Decision Tree	0.98	0.983	0.964	0.975	0.953
Random Forest	0.989	0.995	0.988	0.98	0.997
Gaussian Naive Bayes	0.737	0.72	0.564	0.768	0.445
Ada Boost	0.94	0.959	0.912	0.904	0.919

D. Result Table

In this section, we tabulated the testing scores of the 7 classifiers on the 3 different datasets.
On the upcoming Tables: (I,II, III)

We will display our model testing results including the Accuracy, Precision, Recall, F1 Score.

TABLE III
TESTING RESULTS FOR DATA SET 3

	AUC	CA	F1	Recall	Precision
Logistic Regression	0.755	0.819	0.646	0.617	0.677
KNN	0.648	0.735	0.493	0.429	0.58
SVM	0.636	0.667	0.478	0.571	0.411
Decision Tree	0.659	0.711	0.521	0.534	0.509
Random Forest	0.675	0.782	0.522	0.449	0.622
Gaussian Naive Bayes	0.752	0.692	0.604	0.883	0.459
Ada Boost	0.74	0.815	0.626	0.582	0.678

V. CONCLUSION

We have concluded that the best classifier for the first data set is Ada boost with 0.856 accuracy and F(0.556). As for the second data set, the best classifier is random forest with 0.995 accuracy and F(0.988). And finally for the third data set, the best classifier is logistic regression with 0.819 accuracy and F(0.646)

REFERENCES

[1] M. Fang, J. Su, J. Liu, Y. Long, R. He, and T. Wang, "A model to predict employee turnover rate: Observing a case study of chinese enterprises," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 4, no. 4, pp. 38–48, 2018.

[2] M. Rahman and V. Kumar, "Machine learning based customer churn prediction in banking," in *2020 4th International Conference on Electronics, Communication*

and Aerospace Technology (ICECA), 2020, pp. 1196–1201.

[3] X. Cai, J. Shang, Z. Jin, F. Liu, B. Qiang, W. Xie, and L. Zhao, "Dbge: Employee turnover prediction based on dynamic bipartite graph embedding," *IEEE Access*, vol. 8, pp. 10 390–10 402, 2020.

[4] H. Zhang, L. Xu, X. Cheng, K. Chao, and X. Zhao, "Analysis and prediction of employee turnover characteristics based on machine learning," in *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, 2018, pp. 371–376.

[5] Q. Zhu, J. Shang, X. Cai, L. Jiang, F. Liu, and B. Qiang, "Coxrf: Employee turnover prediction based on survival analysis," in *2019 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2019, pp. 1123–1130.

[6] A. Aggrawal, "churn-modeling.csv," Feb 2018. [Online]. Available: https://www.kaggle.com/datasets/aakash50897/churn-modellingcsv

[7] M. LLC. (2018) MS Windows NT kernel description. [Online]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

[8] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Using knn model for automatic text categorization," *Soft Computing*, vol. 10, no. 5, pp. 423–430, 2006.

[9] M. Panda and M. R. Patra, "Network intrusion detection using naive bayes," *International journal of computer science and network security*, vol. 7, no. 12, pp. 258–263, 2007.

[10] L. Khaidem, S. Saha, and S. R. Dey, "Predicting the direction of stock market prices using random forest," *arXiv preprint arXiv:1605.00003*, 2016.

[11] C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, and O. Jo, "Covid-19 patient health prediction using boosted random forest algorithm," *Frontiers in public health*, vol. 8, p. 357, 2020.