Comparative Study Of The Machine Learning Techniques For Predicting The Employee Attrition

Mrs. A. Aafiya Thahaseen, Dr. P. Nalini, Mrs. M.Rabiyathul Fathima, Mr.A.Mohammed Aslam, Mr.S.M.Abdul Rahman, Mr.S.B.Shajahan ,Mr. A.M.Niyas
Al-Ameen Engineering College,Erode.

Abstract

In HR Analytics, employee attrition was one of the most significant business issues. Companies spend a lot of money on employee training because they know how much money they will bring in in the future. When an employee leaves the organization, the company suffers a loss of economic cost. In the mass recruiting industry, attrition is very high. We have chosen a sample of IBM USA's employee information for this research. Using the data value idea, we discovered that employee attributes such as work role, overtime, and job level have a significant impact on turnover. To forecast the chance of attrition of any new worker, we constructed and tested numerous classification methods such as LDA, logistic regression, lasso categorization, ridge classification,random forests , and decision trees. We compared the models using multiple model assessment criteria and discovered that LDA provided the highest reliability, logistic provided the most precision, and ridge provided the maximum recall. The comparative analysis enables the company to choose a model that best suits its needs. It will assist a business in determining the likelihood of attrition for any new individual they may hire. It would also assist employers in deciding on compensation increases and perks modifications for existing workers, and tweaking them in such a way that the person is retained in some way.

Key words:work role, overtime, economic cost, logistic regression, lasso categorization, ridge classification,random forests.

1.Introduction

Staff turnover has risen to prominence in organisations due to its detrimental effects on a variety of issues, including workplace productivity and morale, project consistency, and long-term growth plans. One method firms deal with this challenge is by employing ml algorithms to estimate the likelihood of employee attrition, allowing leaders and HR the vision to implement proactive measures to retain staff or plan for transition.

However, the ml algorithms that have been employed to overcome this problem in the past have failed to account for the distortion for most HRIS information. Most businesses have not placed a high priority on investing in effective HRIS systems that record an employee's information during the course of his or her employment. One of the most significant problems is a lack of awareness of the strengths and costs. The return on investment in HRIS is currently difficult to calculate [1]. As a result, there is more noise in the information, which reduces the generalisation capacity of these methods. The topic of staff turnover is described in this study, as well as the important machine learning techniques that were utilised to solve it. This study makes a unique contribution by investigating the use of severe XGBoost as an enhancement over typical algorithms, particularly in terms of its capacity to generalise on noisy data, which is common in this sector. This is accomplished by analysing data from a worldwide retailer's HRIS and considering the attrition issue as a classifier, which is then modelled using supervised methods. The conclusion is achieved by comparing the XGBoost classifier's higher accuracy to that of other algorithms and explaining why it outperforms them.

Acquisition, investigation, and analysis of information are producing new knowledge in today's prosperous economy and its expanding technical specialism, referred to as the "modern economy." Technological advances are not only a data source, but also an enabler of data analysis, allowing for the processing of enormous data sets and the extraction of data from them. Information became a key asset for most businesses in a variety of industries, including those that deal with business operations. The use of new technology benefits all sorts of organisations [1], and information gathering, management, and analysis provide significant advantages in terms of quality and market benefit. In fact, analysing massive amounts of data can enhance decision-making processes and contribute to the attainment of goals.

There are various areas in which machine intelligence adoption has an impact on a company's decision-making processes [4,5]. Human resources (HR) has received more interest in recent times, as the quality and abilities of people are a growth element and a true competitive advantage for businesses [6]. Indeed, after gaining traction in marketing and sales, neural network is now being used to help companies make judgments about their personnel, with the goal of basing HR management decisions on factual facts rather than subjective factors [7–9]. Companies strive to maximise their profitability in general. Because they can rely on on-call, infrequent, and permanent work in organisations where workers

execute simple tasks, they can turn to on-call, infrequent, and temporary labour. However, in firms where employees execute increasingly specialised duties, the employee's specialisation and consistency of performance become critical. The value of skills, information, and the ability to acquire new things has shown to be crucial for organisations. Machine learning in HR allows firms to turn data into knowledge by using forecasting analytics: these models allow for worker projections based on data acquired by the business over the past years, decreasing essential concerns and optimising all HR tasks.

Companies devote a significant amount of time and money to staff recruitment and training in order to meet their strategic objectives [12]. As a result, personnel (to a lesser degree) constitute a significant investment for businesses. When a person quits the firm, it not only loses a qualified resource, but it also loses the resources, namely money and Human staff time, that were spent in recruiting, choosing, and training those individuals for their particular tasks. As a result, the company must continue to invest in hiring, training, and introducing new employees to fill open roles. Training a new worker is a time-consuming and costly procedure, hence it is in the company's interest to keep the dropout rate low: An employee's turnover is defined as the loss of employment. Furthermore, satisfied, extremely motivated, and loyal personnel are the heart of a business and have an effect on its productivity. Some writers in the literature recommend keeping just happy and motivated employees because they are more creative, productive, and perform better, which leads to and sustains greater business performance [13–15]. Employee satisfaction statistics are strongly predictive of both breakups and defections, even after controlling for income, hours, and typical personal and job characteristics, since job unhappiness has been demonstrated in the economic research to be a good predictor of intention.

2. Methods

Classification has two separate implications in deep learning. We can well be given a series of data with the goal of determining if the data contains classes or groups. Alternatively, we could be positive that there are a specific set of classes, and the goal is to develop a rule or procedures that will allow us to classify additional observations into one of the current classes. The former is referred to as Unlabeled Data, whereas the latter is referred to as Supervised Learning [19]. Because the data is divided into two classes – active and ended – this work treats classification as reinforcement methods.

2.1 Logistic Regression

One of the most fundamental linear programming for categorization is the logistic linear extrapolation entropy classifier. Linear regression is a type of regression that is particularly useful for predicting binary or classification factors. To prevent over-fitting, it's frequently combined with regularisation in the form of fines depending on the L1-norm or L2-norm. For this study, I used an L2-regularized logistic regression. This method calculates probability values by establishing a model for the situation and estimating the parameters in that model. In (1), the model's shape is as follows: [] p(churn|w) = (1) The maximum - likelihood method is used to estimate the variables w. [20]

2.2 Naïve Bayesian

Nave Bayes is a prominent categorization algorithm that has gotten a lot of attention because of its ease of use and effectiveness [21]. With the premise that all factors are uncorrelated of each other, Nave Bayes accomplishes categorization based on assumptions. The classifier simply needs a limited quantity of training approach to predict the characteristics required for classification. It can also deal with both actual and arbitrary data [22]. The following is the logic behind applying the Bayes' rule to deep learning: We utilise the learning information to analyze estimates of P (X|Y) and P (Y|X) to prepare an overall objectives fn: X Y, that is the same as P (Y|X) (Y). New X samples might then be created using these projected posterior distribution and Bayes' rule.

2.3 Random Forest

A common tree-based ensemble supervised learning is the Random Forest method. Bagging has been the type of 'assembling' utilised here. Sequential trees in bagging are not dependent on previous trees; each is built independently to use a bootstrap samples sample of the set of data. Finally, a simple consensus vote is used to make a prediction. Rf vary from regular trees in that each branch is split according to the optimal split between all variables in the latter. Each cluster in a randomized forest is divided using the strongest correlation from a subset of variables picked at random at that location [23]. This is resistant to over-fitting because of the added layer of unpredictability [24].

2.4 KNN

Nearest Neighbor Classifier works on the principle of classifying datasets based on the category of their closest neighbours. Because it is typically useful to include more than

one neighbour, the technique is known as k-NN Classification [25]. The two phases of Knn classifier are determining neighbouring datasets and then settling on a categories depending on the neighbours' classes. Distance measures such as Distance measure (used in this study), Manhattan distance, and others could be used to find the neighbours. The class can be determined by a majority vote of neighbours or by a weighting system that is inversely proportionate to distance. Before creating the KNN-based classifier, the information was resized to the [0, 1] band.

2.5 LDA

Discriminant analysis entails developing one or even more classifier functions to maximise the variation between the groups in comparison to the variance within the categories [14]. People derive a phasic or z-score, which would be a convolution of two or more variables that will distinguish best between two (or more) separate groups or categories, is what Classification Algorithm is all about. The classifier functions' z-scores are then used to evaluate the likelihood that a certain member or event corresponds to a class. It's vital to remember that the characteristics utilised in LDA must be continous or quantitative in nature.

2.6 SVM

An SVM was a supervised learning method that can tackle both linear and nonlinear binary classifier problems and is based on computational learning concept [26]. For different classes, a svm classifier creates a high energy or collection of hyper-planes in capacity. The idea is that the hyper-plane with the maximum distance to the closest datapoints of any class achieves a good separation; the higher the margin, the smaller the classifier's generalisation error. As a result, it's also known as a decision boundary classifier. Before creating this model, the information was reduced to the [0, 1] range.

3. Proposed System

The information is first pre-processed from Kaggle so which we can extract critical characteristics like Month Income, Latest Promotion Date, Extra Pay, and other factors that are common in staff turnover. Factors considered, also known as predicted variables, are those that aid in the identification of elements that are dependent largely on employee-related factors. The id Number or employee count, for instance, has no bearing on the attrition rate. Data Exploration is a first step in the analytical process in which you summarise data attributes to anticipate who and when a worker will leave the company.

Using the random forest method, the system creates a forecasting models.

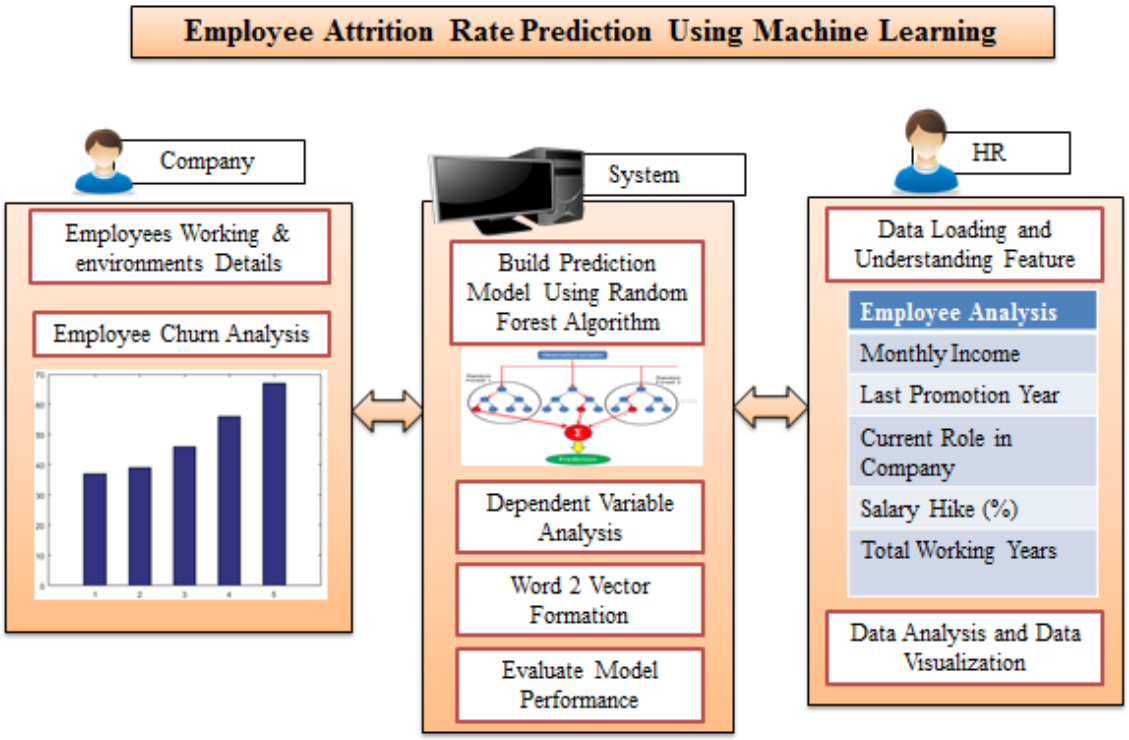

Figure 1. Architecture of the System

4. Dataset

This dataset was acquired from IBM Hr Management in order to assist in the solution of this challenge. This dataset includes all 35 characteristics, with Attrition serving as the reliant attribute. We've come to the notion that using this data, we could be able to solve this issue. These are the characteristics found in our data.

5. Feature Selection

Feature extraction is regarded as the most important theory in the area of deep learning, and it has a considerable impact on the model's actual results. These characteristics can be utilised to easily teach your simulation and have a significant impact on its performance. The validity of the system can be harmed by trivial and irrelevant features. First and most important phase in model development must be feature extraction and data cleaning. Feature choice is the act of manually or automatically selecting those features that influence the most to your predictor variables or outcome variable that you are interested in

using various strategies. We found that the characteristics Worker Number, Employee Quantity, and Over18 have had no direct effect on our outcome parameter Arttrition after personally analysing the information. As a result, before using any feature selection technique, these characteristics were completely ignored. Importance of Feature Product importance assigns a value to each of your data's features; the greater the score, the more essential or relevant the information is to your single output. We will use Extra Tree Separator to retrieve the top characteristics for the dataset because Feature Significance is an intrinsic class that arrives with Tree Based Classifiers.

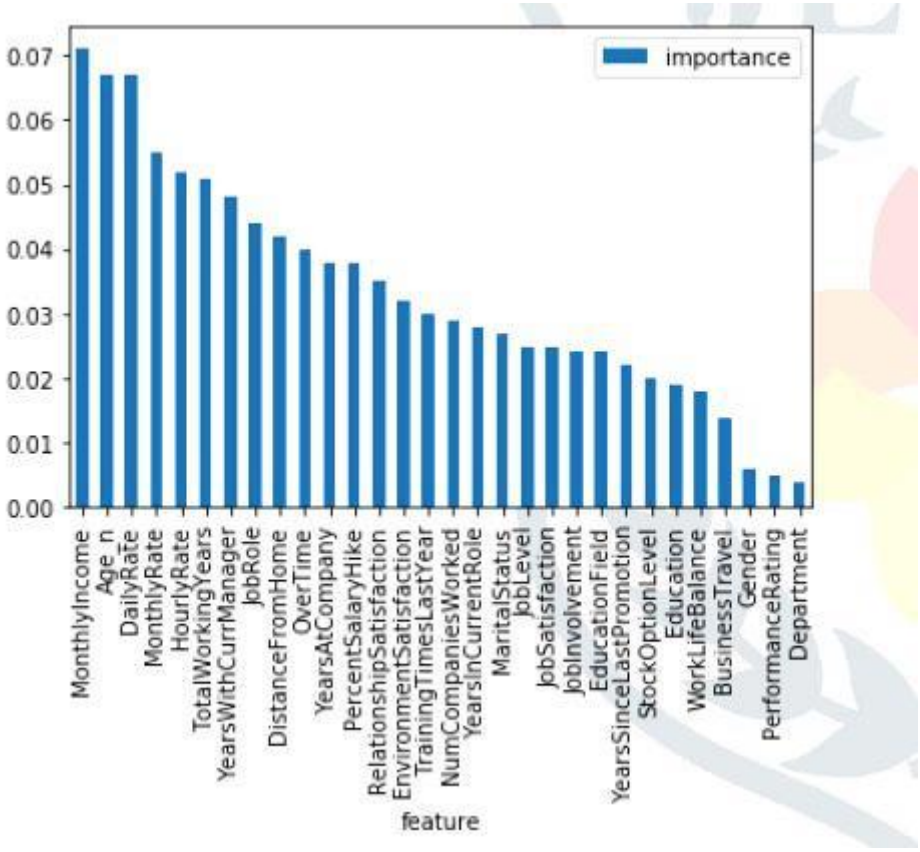

Figure 2. Importance of Feature

The figure above depicts the feature significance of each factor in our dataset. Using this feature significance method, we were able to determine that variables such as Monthly salary, Aged, Day rate, and Hourly rate are among the most important attributes. We also discovered that characteristics such as business trip gender, sector, and performance review have the slightest effect on our outcome variable Atrrition. As a result, we can ignore these characteristics in advance. The following components are used for model creation after

using feature selection technique.

6. Data Exploration And Analysis

We'll look at the connection between the characteristics and the outcome variable in this section. Because there are so many characteristics to choose from, we can't show each attribute's relationship. So, to demonstrate the relationship, we'll use the Age property as an instance.

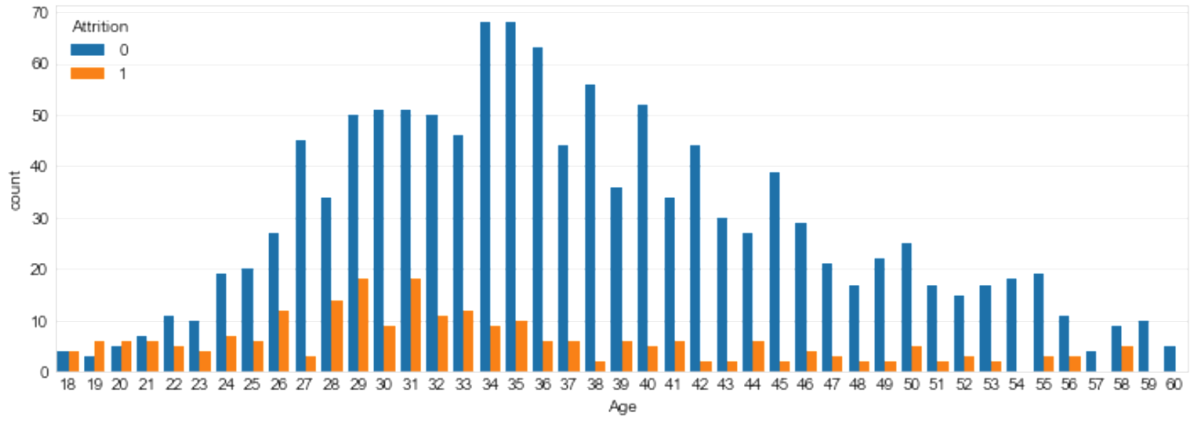

Figure 3. Data Exploration And Analysis

As can be seen in the graph above, the incidence of turnover is greatest across all ages between the ages of 29 and 31. Attrition is most likely to occur in those over the age of 50.

7. Out-Of-Balance Dataset

90% of the entries in the data are classified with the class YES, while the remaining 10% are classified with the class NO. This kind of dataset is referred to as an unbalanced dataset, but it can have a negative impact on the model's performance since it biases the model towards the class label of output variables. As a result, for this type of issue statement, dealing with a small datasets becomes a vital requirement.

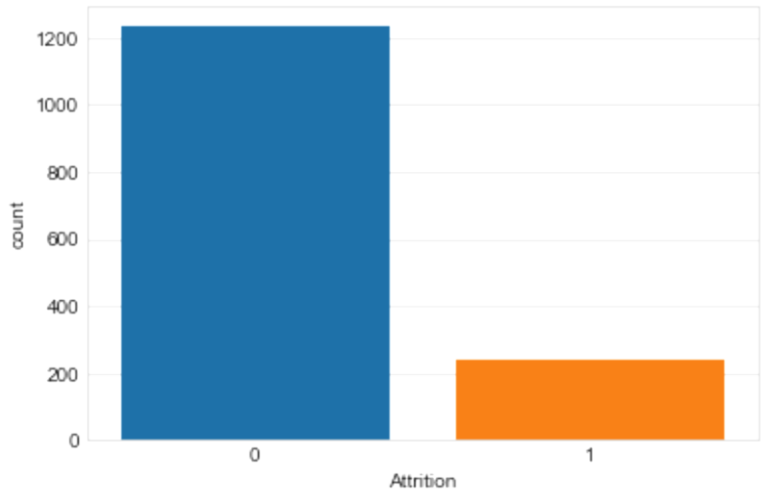

Figure 4. Out-Of-Balance Dataset

Data that is unbalancedRandom Over Sampling , Random Under Sampling, and Data Based Over Sample selection are some of the strategies for dealing with dataset instability.

To handle the imbalancness of our data, we are employing the over sampling strategy. Prior to the oversampling, 1233 entries were classified with class NO and only 236 data with class YES. We compared the amount of entries in both classes to 1233 instances after conducting over sampling, as indicated in the graph below.

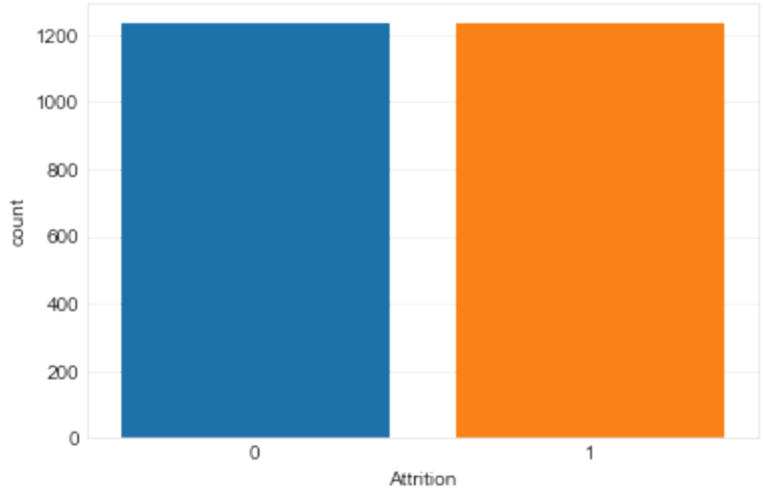

Figure 5. After oversampling, data distribution

8. Conclusions:

In this study, IBM data from the U. S. was used to anticipate staff turnover using 45 data mining methodologies and machine intelligence. Staff turnover is the most serious

corporate issues, hence the emphasis is on using data mining to use different algorithm and mixtures of many goal criteria to anticipate employee attrition efficiently and accurately. LDA, lasso regression, logistic regression, random forest, decision tree classification,support vector machine, ridge regression, and nave probabilistic classification techniques are used to analyse IBM worker attrition data. We discovered that the Lda Analysis Model, with an efficacy of 86.39 percent, outperforms mining methodologies in terms of accuracy.

REFERENCES:

[1] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in Proc. Int. Conf. Inventive Comput. Informatics (ICICI), pp. –, 2017.

[2] D. K. Srivastava and P. Nair, "Employee attrition analysis using predictive techniques," in Proc. Int. Conf. Inf. Commun. Technol. Intell. Syst., Springer, Cham, pp. –, 2017.

[3] S. S. Alduayj and K. Rajpoot, "Predicting employee attrition using machine learning," in Proc. Int. Conf. Innov. Inf. Technol. (IIT), pp. –, 2018.

[4] R. Jain and A. Nayyar, "Predicting employee attrition using XGBoost machine learning approach," in Proc. Int. Conf. Syst. Model. Adv. Res. Trends (SMART), pp. –, 2018.

[5] F. Fallucchi, et al., "Predicting employee attrition using machine learning techniques," Computers, vol. 9, no. 4, p. 86, 2020.

[6] D. Jain, "Evaluation of employee attrition by effective feature selection using hybrid model of ensemble methods," M.S. thesis, National College of Ireland, Dublin, 2017.

[7] S. Yadav, A. Jain, and D. Singh, "Early prediction of employee attrition using data mining techniques," in Proc. IEEE 8th Int. Adv. Comput. Conf. (IACC), pp. –, 2018.

[8] T. P. Salunkhe, "Improving employee retention by predicting employee attrition using machine learning techniques," M.S. thesis, Dublin Business School, Dublin, 2018.

[9] A. Mhatre, et al., "Predicting employee attrition along with identifying high risk employees using big data and machine learning," in Proc. 2nd Int. Conf. Adv. Comput. Commun. Control Netw. (ICACCCN), pp. –, 2020.

[10] S. Dutta and S. K. Bandyopadhyay, “Employee attrition prediction using neural network cross validation method,” Int. J. Commerce Manag. Res., vol. 6, no. 3, pp. 80–85, 2020.

[11] S. N. Khera and Divya, “Predictive modelling of employee turnover in Indian IT industry using machine learning techniques,” Vision, vol. 23, no. 1, pp. 12–21, 2018.

[12] S. Najafi-Zangeneh, et al., “An improved machine learning-based employees attrition prediction framework with emphasis on feature selection,” Mathematics, vol. 9, no. 11, p. 1226, 2021.

[13] P. K. Jain, M. Jain, and R. Pamula, “Explaining and predicting employees’ attrition: a machine learning approach,” SN Appl. Sci., vol. 2, no. 4, pp. 1–11, 2020.

[14] N. Bhartiya, et al., “Employee attrition prediction using classification models,” in Proc. IEEE 5th Int. Conf. Converg. Technol. (I2CT), pp. –, 2019.

[15] F. Barboza, H. Kimura, and E. Altman, “Machine learning models and bankruptcy prediction,” Expert Syst. Appl., vol. 83, pp. 405–417, 2017.