Predictive Data Mining Group Project for Employee Attrition Dataset

Zilan Huang
MSc Computer Science
zh4u17@soton.ac.uk

Bang Du
MSc Computer Science
bd1m17@soton.ac.uk

Zihan Wang
MSc Computer Science
zw1u17@soton.ac.uk

Yang Zhou
MSc Computer Science
yz12a17@soton.ac.uk

Da An
MSc Computer Science
da1g17@soton.ac.uk

Mardhiyyah Rafrin
MSc Computer Science
mr1n17@soton.ac.uk

ABSTRACT

This paper is going to find the result or the solution of the reason why employees turnover base on an employee attrition dataset. The project found the ways to clean and process the data as well as solving the problems this dataset had. After these data being processed, the project will use these processed data to train and find the results on different predictive models. The result will be used to perform internal data analysis to give the solution and the suggestion to the employee attrition problem.

KEYWORDS

Coefficient Correlation, Information Gain, MDS, Imbalanced Dataset, PCA, Logistic Regression, Decision Tree, LDA, SVM, Random Forest.

1 INTRODUCTION

This project used the dataset collected from IBM Watson Analytic which provides the employee attrition data from human resources department. The dataset contains 35 features and around 1,400 data to be used which is a very small dataset and it will be very challenging to be analysed, implementing technical data mining skill in order to build the predictive models and to analyse the result from the model. The project is going to find the approach to deal with this kind of problems and being able to have the evaluation to the predict result in the end.

2 DATA ANALYSIS & PROCESSING

2.1 Coefficient Correlation

Correlation coefficient is the amount of linear correlation between research variables. It is calculated by the product difference method. [7] It is also based on the dispersion of the two variables and the respective averages, and the degree of correlation between the two variables is reflected by multiplying the two dispersions. Definitions:

$$r(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var[X]Var[Y]}}$$

Analyze the correlation coefficient of the training set data and obtain the correlation coefficient graph.

According to figure 1, there is no strong correlation between most of the features and Attrition, only features 'Age', 'JobInvolvement', 'JobLevel', 'MonthlyIncome', 'OverTime', 'StockOptionLevel', 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole' and 'YearsWithCurrManager' have a strong correlation with Attrition.

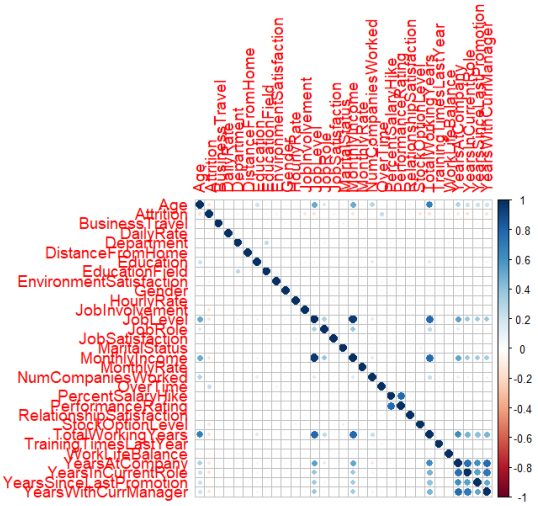

Figure 1: Correlation coefficient of Train Dataset.

Only 'OverTime' is strongly positively related to Attrition, and the rest is negatively correlated.
These features also have strong correlation directly, which makes it necessary to consider Multicollinearity when building a model.

2.2 Information Gain and Information Gain Ratio

Information entropy is a very common measure of purity of the sample collection. [6] It is defined as:

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$

The smaller the $Ent(D)$ value, the higher the collection purity.

Based on this, we can get the information gain for each feature. It shows how much the purity will improve if we use this feature to split the dataset. The feature with higher information gain usually affect the result more. However, the information gain is affected by the number of discrete values, so we have used the information gain ratio to analyze the data. It uses the information gain divided by its intrinsic value to eliminate the effect of the number of discrete values. The features with highest information gain ratio are shown as below on figure 2:

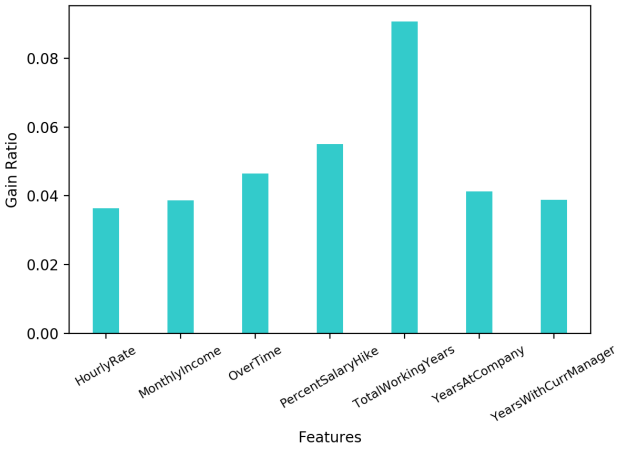

Figure 2: Features with highest information gain ratio.

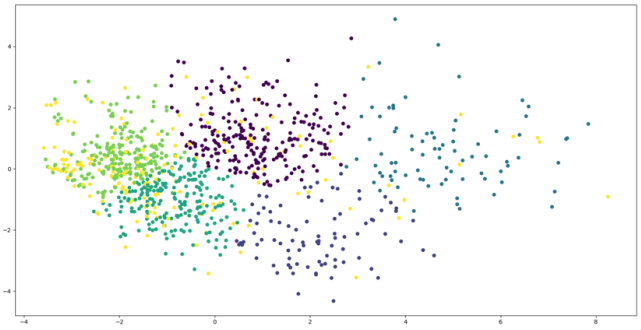

Figure 3: The Result of MDS.

2.3 Multi-Dimensional Scaling (MDS)

2.3.1 Description. MDS is a clustering algorithm which can map the data of high dimensional space into low dimensional space. The goal of MDS is to maintain the dissimilarity of the data in the process of dimensionality reduction, and also to understand that the distance relationship in the high dimensional space and the distance relationship in the low-dimensional space remain unchanged. In other words, the people can visually see the relationship between raw data through MDS.

2.3.2 result. After the normalization, the dataset was clustered by MDS. The yellow points in figure 3 are departing staff and the rest 5 color are currently working employees. Through the analysis of clustering results, it is easy to find that the green and blue area has more yellow points. This means that the separated employees mainly concentrate on these two clusters. By analyzing the samples in these two clusters, we can find that the proportion of people in sales department in higher than other clusters and the proportion of employees who work overtime is higher than other clusters. Besides, the Relationship Satisfaction of these two clusters is lower than other clusters. Therefore, the result illustrates that department, overtime and Relationship Satisfaction are the three mainly reason which effect the demission of employees.

3 DATA PROCESSING

3.1 Clean Dataset

Train set has 1470 rows and 35 colums, test set has 471 rows and 34 colums. Firstly, perform a null query on the data set. The strategy worthwhile to deal with NA is to complete the complement of high-value and low-missing value, and to delete the features with low feature and high loss. After querying the training set and test set, it is found that there is no null value in the data. The data does not have a null value, so there is no need to delete the feature or complete the null value. Then, calculate the slope of the feature data and use the skew function to query the skewness, logarithmic exchange of features with skewness exceeding 75%, which makes the data distribution tend to normal and let the models more efficient and accurate. After logarithmic transformation, data needs to be quantified. Finally, Use the get_dummies function to quantify the data, which quantizes non-digital features based on the number of variables by one-hot encoding. After one-hot encoding, features are expanded, so the number of features is magnified and data cleanup has also been completed.

3.2 Deal with Severe Imbalance Dataset

3.2.1 Techniques of Balancing Data. The main challenge of training the prediction models is that the dataset is surprisingly imbalanced (yes : no = 6 : 1). There are several attempts to deal with this problem. In this project, however, we try to compare the result of two approaches. The first approach is that implemented is to generate new samples in the class which are under represented. The minority class is oversampled by using Synthetic Minority Over-sampling Technique (SMOTE) [2], which extra training data are obtained by taking each minority data class sample and introducing synthetic example along line segments joining any/all of the k minority class nearest neighbors. Therefore, this technique can prevent overfitting and causes the decision boundaries for the minority class to be span into the majority class space.

The second approach to fight the severe imbalanced dataset problem is bagging technique. Bagging [1] is a bootstrap ensemble technique which produces individuals for its ensemble by training each classifier on a random redistribution of the training data. Each classifier's training set is generated by randomly drawing, with replacement, N examples - where N is the size of the original training set. As a result, many of the original examples may be repeated in the resulting training set while others may be eliminated. Each individual classifier in the ensemble is generated with a different random sampling of the training set. The classification algorithm used for this technique is called Bagging Classifier.

3.2.2 Implementation. For over-sampling technique, we use decision tree algorithm to predict the output. First learning process is done on real dataset. With the same structure of tree, the over-sampled data is trained. Both trained models are tested with new

Table 1: Results of Decision Tree learning on real data and over-sampled data

Real Data				
Class	Precision	Recall	F1	Sup
0.0	0.91	0.88	0.90	370
1.0	0.47	0.54	0.50	71
Total/Avg	0.83	0.84	0.83	441

Over-Sampled Data				
Class	Precision	Recall	F1	Sup
0.0	0.88	0.88	0.88	370
1.0	0.38	0.38	0.38	71
Total/Avg	0.80	0.80	0.80	441

Table 2: Counfusion Metrics of the decision tree results

Real Data		
	Detected Negative	Detected Positive
Actual Negative	327	43
Actual Positive	33	38

Over-Sampled Data		
	Detected Negative	Detected Positive
Actual Negative	326	44
Actual Positive	34	27

Table 3: Results of Bagging Classifier learned on real data and balanced data

Imbalanced Data				
Class	Precision	Recall	F1	Number of Data
0.0	0.86	0.99	0.92	370
1.0	0.72	0.18	0.29	71
Total/Avg	0.84	0.86	0.82	441

Balanced Data				
Class	Precision	Recall	F1	Number of Data
0.0	0.92	0.85	0.88	370
1.0	0.43	0.59	0.50	71
Total/Avg	0.84	0.81	0.82	441

Table 4: Counfusion Metrics of the result of bagging classifier

Real Data		
	Detected Negative	Detected Positive
Actual Negative	365	5
Actual Positive	58	13

Balanced Data		
	Detected Negative	Detected Positive
Actual Negative	315	55
Actual Positive	29	42

data which is highly imbalanced. The comparison of testing data between two classes is 370 : 70.

The same step is used to implement bagging technique. The bagging process is integrated directly with learning process of bagging classifier, therefore bagging classifier is used. We build two trained model, classifying real data with and without bagging technique.

The measurement of performances is based on precision, recall, and F1 Score. Precision (Positive predictive value) is the fraction of relevant samples among the retrieved samples.On the other hand, recall (sensitivity) is the fraction of relevant samples that have been retrieved over the total amount of relevant samples. These measurements inform the relevance of data.

3.2.3 Results. The tables 1 and 2 show the comparison of result of predicting new data which the models have learned with and without balancing techniques.

The bagging technique result in table 3 shows an increasing number of true positive in the small class, from 13 samples to 44 samples. The recall, therefore, increases from 0.18 to 0.59 compare to the result of oversampling method, this technique is more reasonable to skew the decision boundaries to the majority class.

Learning on over-sampled data, the prediction of new data show is unexpected. The precision of small class (class 1) slightly decrease and the recall is slightly increase compare to the prediction model learning Real data. In table 4, comparing the confusion metrics between real data and over-sampled data, the number of true positive decrease 11 samples of data. Therefore, the technique is not considerable for this data.

3.3 PCA

Principal component analysis (PCA) is a technique for analyzing and simplifying data sets. Principal component analysis is often used to reduce the dimensionality of the data set while maintaining the largest contribution to the variance in the data set.[5] This is done by retaining the low-order principal components and ignoring the higher-order principal components. This low-level component can often retain the most important aspects of the data.
Through data cleansing, one-hot encoding, features are enlarged, and the number increases. At this time, PCA is required to reduce the dimensions of the data set.

4 PREDICTION MODEL

4.1 Logistic Regression

4.1.1 Description. Logistic Regression is a regression model based on linear regression. It has used a function to convert the result of linear regression to a binary dependent variable.[4] This makes the dependent variable categorical. The ideal function is unit-step function. However, it is not continuous. Logistic function is usually used as surrogate function of unit-step function, which

is shown as below:

$$y = \frac{1}{1 + e^{-z}}$$

Based on linear regression, this formula can be converted to:

$$\ln \frac{y}{1-y} = w^T x + b$$

Since y is a binary dependent variable. It can be rewritten as:

$$\ln \frac{p(y=1|x)}{p(y=0|x)} = w^T x + b$$

$$p(y=1|x) = \frac{e^{w^T x+b}}{1 + e^{w^T x+b}}$$

Finally, use maximum likelihood method to estimate the parameters w and b with these data. Since the

4.1.2 Implementation. For a better understanding of logistic regression, the code of algorithm is written by ourselves. Since the formula got by maximum likelihood method is a high-order differentiable continuous convex function, both of gradient descent method and Newton method are used in this part. Then we can get the optimal value for the parameters with these data. Finally, due to the imbalance of data sets, $y = 0.5$ cannot simply be used as a prediction threshold. Use m^+ as number of positive class and m^- as negative. If $\frac{y}{1-y} > \frac{m^+}{m^-}$, the prediction is positive, and if $\frac{y}{1-y} < \frac{m^+}{m^-}$, the prediction is negative.

4.1.3 Result. The accuracy of 10-fold cross validation is around 87%, which is reasonable. Logistic regression can be easily used and have a good performance to control the interference of noise. Meanwhile, it does not need to presume the data distribution in advance, thus avoiding the problem that the assumed distribution is inaccurate.

4.2 Decision Tree

4.2.1 Description. The decision tree is a tree-like model of decisions and their possible consequences. It can be used in data mining as a predictive modeling approach.[8] Its leaves represent class labels and branches represent conjunctions of features that lead to those class label. Once input a data, it will finally find its corresponding leaf node through the selection of multi-layer nodes. There are several ways to choose the optimal feature to split the tree. ID3 has used information gain, which make the node have highest purity. C4.5 has used gain ratio which use information gain divided by intrinsic value. Meanwhile, CART has used Gini index.

4.2.2 Implementation. For a better understanding of decision tree, the code of this part also did not use function from library. The first step is use recursion. In this way, we use gain ratio to split the tree. Because this can avoid that the feature with more discrete values has a higher information gain. Finally, we use pruning to reduce the degree of fitting.

4.2.3 Result. The accuracy of 10-fold cross validation is around 84%. Before the pruning, it has the problem of overfitting. The accuracy has been improved by pruning. Decision trees are easy to explain. However, there is a high probability of overfitting. Generally, a single tree gives low prediction accuracy for a dataset as compared to other algorithms.

4.3 LDA

4.3.1 Description. Fisher linear classifier is one of the basic methods of statistical pattern recognition. The classifier has the advantages of simple design, low computational complexity and mall memory footprint, which is one of the most commonly used methods in practical application. The basic idea of Fisher's test is to find the best projection to separate two classes when the eigenvector x is mapped from d dimensional space to this direction. This method actually involves the compression of feature dimension. The linear discriminant function of Fisher in one-dimensional space is:

$$J_F(w) = \frac{(m_1 - m_2)^2}{S_1 + S_2}$$

$$m_i = \frac{1}{N} \sum x, i = 1, 2$$

$$S_i = \sum_{x \in \xi} (x - m_i)(x - m_i)^T, i = 1, 2$$

m1 and m2 are means of two sample data, S1, S2 are the within-class scatters of the sample. The projection direction is:

$$w = S_w^{-1}(m_1 - m_2)$$

$$S_w = S_1 + S_2$$

In Fisher's decision function, the molecules is the square of distance between the two centers of two classes. The greater value means the better inter-class separability of the classes. The denominator is the dispersion degree of the two classes, the smaller value means data is more concentrated. When the JF(w) is maximized, the classifier can achieve best classifying quality.

4.3.2 Implementation and Result. for this project, the samples in each class are dimension vectors. After the normalization, 80% samples in dataset were trained and the rest 20% were test by trained model. By 10-fold cross-validation, the accuracy is about 86.7%. This shows that LDA discriminant obtained great results in the condition of less sample size. However, as a linear classifier, LDA cannot distinguish the samples well in junction of two classes. This is the limitation of LDA algorithm.

4.4 SVM

4.4.1 Description. Support Vector Machine is a general learning algorithm based on statistical learning theory. Compared with the traditional prediction model, Support Vector Machine has great generalization ability and nonlinear mapping ability which can solve the nonlinear feature and high dimensional feature problem in this case. [3] The high result of prediction model provides an efficient data mining measurement.

4.4.2 Implementation. Before the implementation, the linear classifier can help to produce the confusion matrix for the classification. To improve the result, a model established on certain function is implemented. The key parameters in the implementation is kernel, cost and gamma. Kernel is the genre of the Support Vector Machine. Radial basis function (RBF) kernel function is used in the data mining by setting the kernel parameter 'radial', which

Table 5: Searching for the Best Parameters

gamma	cost	error
$1 \times e^{-6}$	10	0.1625
$1 \times e^{-5}$	10	0.1625
$1 \times e^{-4}$	10	0.1625
0.001	10	0.1625
0.01	10	0.126136
0.1	10	0.147727

Table 6: 'n_estimators' test result

n_estimators	Score of Prediction Model
31	0.8571428571428571
800	0.87755102040816324
850	0.87755102040816324

there is no need for a linear hyperplane in the results. The division between categories is defined by a curved area, which has higher accuracy with the same training data. 'cost' is a function for violation of constraints and the gamma is a common parameter. The function tune.svm() is used for the best parameters in Support Vector Machine. According to table 5, the model has the minimum error rate when cost = 100, gamma = 0.001.

4.5 Random Forest

4.5.1 Description. Random Forest is an ensemble model that has multiple trees (n_estimators). The final prediction would be a weighting average (regression) or mode (classification) of the predictions from all estimators. For the preparation of the Random Forest Mode, the RandomForestClassifier used for the importance extraction. The assessment is to see how much each feature contributes to each tree in a random forest and then take an average. At last, comparing the features by the Gini index or OOB.

4.5.2 Implementation. In the first prediction model, the error rate is 14.03% which has further optimization. The main measurement of optimization is to adjust the parameter in the Random Forest function. The parameter 'n_estimators' represents the maximum number of iterations of the weak learner or the maximum number of weak learners. Generally, these parameters influence the fitting. And with the certain parameter, increasing the parameter cannot help improve the model. In this project, the statistic shown in the table 6 helps to figure out the suitable number for 'n_estimators'.

The 'max_features' decided to consider the number of features, which decides the decision tree generating time. In this project, the amount of data is not big enough. While 'sqrt' means only take up to $\sqrt{N}$ (N is the number of the feature), it is a suitable choice in this model. Also, maximum depth finally defined as 9 by ten-fold cross-validation to restrict the sub-tree depth. The overfitting and useless iteration will influence the effective prediction. For now, the prediction result is up to 88.48%. The least sample number restricts the minimum number of the leaf node. If the number of leaf nodes is smaller than this number, it will be pruned with the sibling nodes.

Table 7: Comparision to the Prediction Models

Model Name	Accuracy
Logistic Regression	87%
Decision Tree	84%
LDA	86.7%
SVM	90.24%
Random Forest	91.32%

4.5.3 Result. After the development of the Random Forest Model, the f1_score of the final model is up to 91.32%. And the most important influential factors of the employee are OverTime, MonthlyIncome, JobRole, Age, StockOptionLevel.

4.6 Comparision & Analysis

According to table 7, it is clearly seen that there are five different predictive models listing and forest has the best accuracy. First of all, the sample of the dataset is small. Therefore, by using one-hot encode, the number of features will be larger than number of sample data which has been extended to around 2000 features. Secondly, the dataset has nonlinear features. Random forest consist of multiple decision trees and decision trees has the ability to process nonlinear features. However, compare to random forest, decision tree does not have generalization ability and results in overfitting problem. In addition, other predictive models like linear regression, logistic regression and LDA. They are performing good in linear data, but it does not perform well compared with random forest and the dataset has nonlinear features. Therefore, random forest stands out from these predictive models and get the higher accuracy of 91.32% than the others. Furthermore, similar to random forest, SVM also has the ability to deal with nonlinear features, so it also has a very good result. To extend it, the performance of dealing high dimensional samples for SVM is good. Thus, SVM has got a high result of 90.24% in second place.

5 ANALYSIS TO THE RESULT

Through the model prediction, the result of nearly 90 correct rates is obtained. The training set and the test set are combined to perform internal data analysis. Due to the correlation analysis, it can be seen that features 'Age', 'JobInvolvement', 'JobLevel', 'MonthlyIncome', 'OverTime', 'StockOptionLevel', 'TotalWorkingYears', 'YearsAtCompany', 'YearsInCurrentRole' and 'YearsWithCurrManager' have strong correlations. With Attrition. Internal analysis focuses on analyzing these characteristics.

Figure 4 shows that 0 is a male and 1 is a female. The number of male evacuees is generally higher than that of females. Obviously, the number of people leaving after the age of 30 is the most, and the number of evacuations after 30 years of age is gradually decreasing. The number of employees who are younger than 20 and older than 55 is minimal.

In figure 5, the green part is the resignation and the red part is the working with job. Obviously, the lower the monthly income,

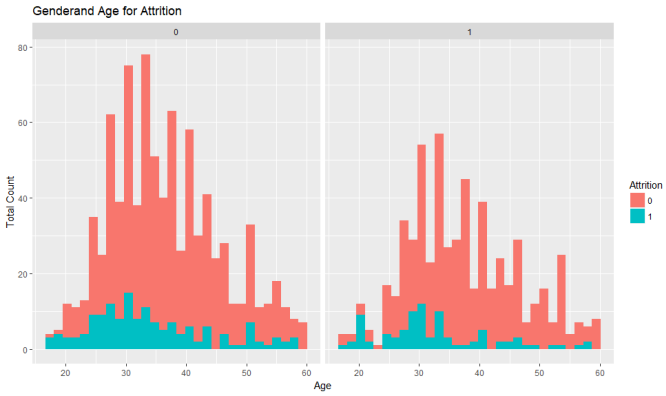

Figure 4: Gender and Age for Attrition.

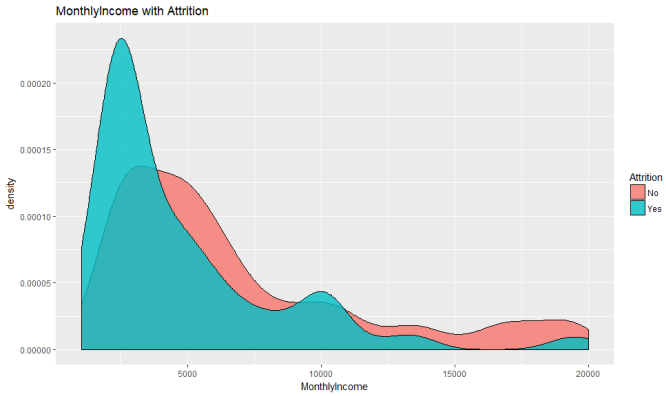

Figure 5: MonthlyIncome with Attrition.

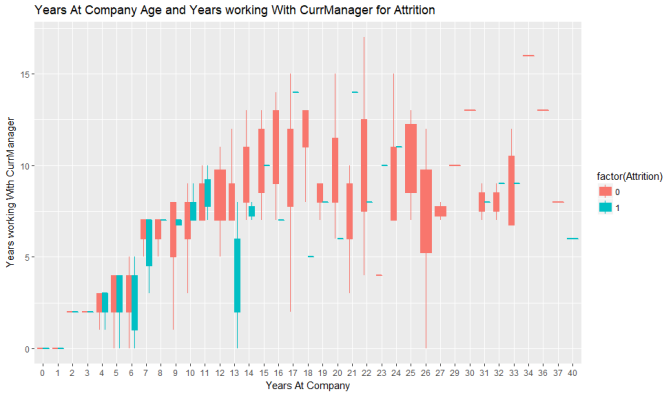

Figure 6: Years At Company Age and Years working With CurrManager for Attrition.

the greater the number of departed people, which means that the people with low income will have unstable job intentions, and the high-income earners will have more stable job intentions and fewer departures.

The green K line chart in figure 6 is resigned and red is in service. Obviously, the longer the service time is, the more willing it is to work, and the 4/13-year employee's resignation will be the strongest, and the willingness to work is most unstable.

6 SUGGESTION

On the one hand, according to the analysis, the monthly income and work time are the mainly reason which cause the employees leave office. However, increase the salary and reduce the work time will cause the cost increase. One solution is that the employer can optimize settlement regulation. The employer should quantify the workload of employees and pay the salary by the work time and work efficiency. If the employee know that he can get more pay for more work done, they will work with more active attitude more active and work more passionate.

On the other hand, the cooperate management needs to pay more attention to the employees who had been working in the company for from the year of 4th to 13th. The employees who had been working for 4 years started have the expectation from the company such as a better salary. In addition, the employees who had been working for 13 years may be have some disappoints to the company for a long time, so they are more likely to consider about finding a new job to meet their requirements. The company will lose their valuable employees who are skilful and can bring enormous benefits and values to the company.

7 CONCLUSION

To sum up, the project was firstly started from analysis of the dataset with some techniques like finding the coefficient correlation, information gain and MDS. Furthermore, based on these techniques, the project started to clean the dataset and finding the way to deal with the imbalance data which was implemented by using over-sampling and bagging technique. PCA, in addition to this session, was also used to reduce the dimension of the data. After cleaning and processing the data, the project was then going to train five prediction models and found the accuracies of different models by using 10-fold cross validation technique and the best result was from Random Forest which gave 91.32% of the accuracy. Based on the results, the project was then going to combine both training set and test set to perform the internal data analysis with their visualisation graph results as well as giving a suggestion to the solution.

REFERENCES

[1] L. Breiman. 1996. Bagging predictors. *Machine Learning* 24, 2 (1996), 123–140.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research* (2002).

[3] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[4] David A Freedman. 2009. *Statistical models: theory and practice.* cambridge university press.

[5] Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika* 28, 3/4 (1936), 321–377.

[6] Edwin T Jaynes. 1957. Information theory and statistical mechanics. *Physical review* 106, 4 (1957), 620.

[7] I Lawrence and Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* (1989), 255–268.

[8] J Ross Quinlan. 1986. *Simplifying decision trees.* Technical Report. MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB.