Journal of Industrial and Systems Engineering
Vol. 8, No. 4, pp106-121
Autumn 2015

JISE

Case Study

A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)

Amir Mohammad Esmaieeli Sikaroudi[1]*, RouzbehGhousi[1], Ali EsmaieeliSikaroudi[2]

[1]Department of industrial engineering, Iran University of Science and Technology
Tehran, Iran
[2]Industrial & Manufacturing Engineering Department
Florida State University,USA

amir_esmaieeli@iust.ac.ir, ghousi@iust.ac.ir, aesmaieelisikaroudi@fsu.edu

Abstract

Training and adaption of employees are time and money consuming. Employees' turnover can be predicted by their organizational and personal historical data in order to reduce probable loss of organizations. Prediction methods are highly related to human resource management to obtain patterns by historical data. This article implements knowledge discovery steps on real data of a manufacturing plant. We consider many characteristics of employees such as age, technical skills and work experience. Different data mining methods are compared based on their accuracy, calculation time and user friendliness. Furthermore the importance of data features is measured by Pearson Chi-Square test. In order to reach the desired user friendliness, a graphical user interface is designed specifically for the case study to handle knowledge discovery life cycle.

Keywords: Employees' turnover; Data mining; Human resource management; Recruitment decision support system

1- Introduction

We consider turnover as the rate of employees' leave and replacement in a predefined period of time. Turnover has various forms. It can be voluntary or involuntary, functional or dysfunctional, avoidable or unavoidable followed by same financial consequences for all types. At macro level reasons of turnover are categorized as the enterprise ,the individual and the industry (Chalkiti & Sigala, 2010). Turnover may eventuate to positive results like functional turnover or to negative results like dysfunctional turnover. Functional turnover means that employees with poor performance quit their jobs and employees with good performance remain in their jobs. In turn, dysfunctional turnover means that employees with good performance quit their jobs and employees with poor performance stay in their jobs (Sexton, McMurtrey et al., 2005). Unavoidable separations include retirement, death, permanent disability, or a spouse changing jobs to a different community. Classification of turnover helps managers to courage or discourage specific turnover type (Holtom, Mitchell et al., 2008). The separation of experienced

*Corresponding author.
ISSN: 1735-8272,

employees from their job may damage the productivity and services of organization and it will take time to obtain the previous efficiency.

Performance of employees can influence both voluntary and involuntary turnover. Poor performance is certainly related to involuntary turnover and also good performance will avail alternatives that may eventuate to voluntary turnover. Job performance is a factor that relates to job satisfaction through expectancy theory. It implies that high performance leads to greater rewards and more job satisfaction(Zimmerman, 2008). Involuntary turnover can be considered in two categories which are layoffs and dismissals. Layoff is a consequence of retrenchment to cutoff costs and improvement of efficiency while dismissal is a consequence of poor performance (Iverson & Pullman, 2000). An important target of organizations is reduction of voluntary and avoidable turnover. Voluntary turnover is a decision making process related to "image theory". Generally individuals have insufficient sources to evaluate information and also employee's perception of their environment might be various(Beach, 1990).

Turnover causes many different types of costs for organizations. Costs of turnover are divided to direct costs such as advertising the position, replacement, recruitment and selection, temporary staff and management time, and indirect costs are morale related costs, pressure on remaining staff, costs of learning, product/service quality and organizational memory (Ongori, 2007). It has been proved that 15-30 percent of turnover costs are direct and about 70-85 percent of turnover costs are hidden costs such as lost productivity and opportunity (Boles, Dudley et al., 2012; Racz, 2000). Costs of an employee turnover continue until inducted performance of employee reaches to previous one. This interval is called "ramp-up" time byDeConinck and Johnson (2009) due to being unfamiliar with new system. Turnover in organization will affect customer satisfaction and service experience and causes overall dissatisfaction and inconsistency among employees. Inefficiency in organization is closely associated with high employee turnover (Hancock, Allen et al., 2013). Low employee morale, personnel conflicts, unhealthy work atmosphere, unwillingness for engaging, frequent employee absence, low performance of organization and perfunctory quality control are signs of job dissatisfaction and job stress that eventuate to resign if job alternatives are accessible. Wang et al. (2011) illustrate a through summary of employee turnover factors shown in figure 1.

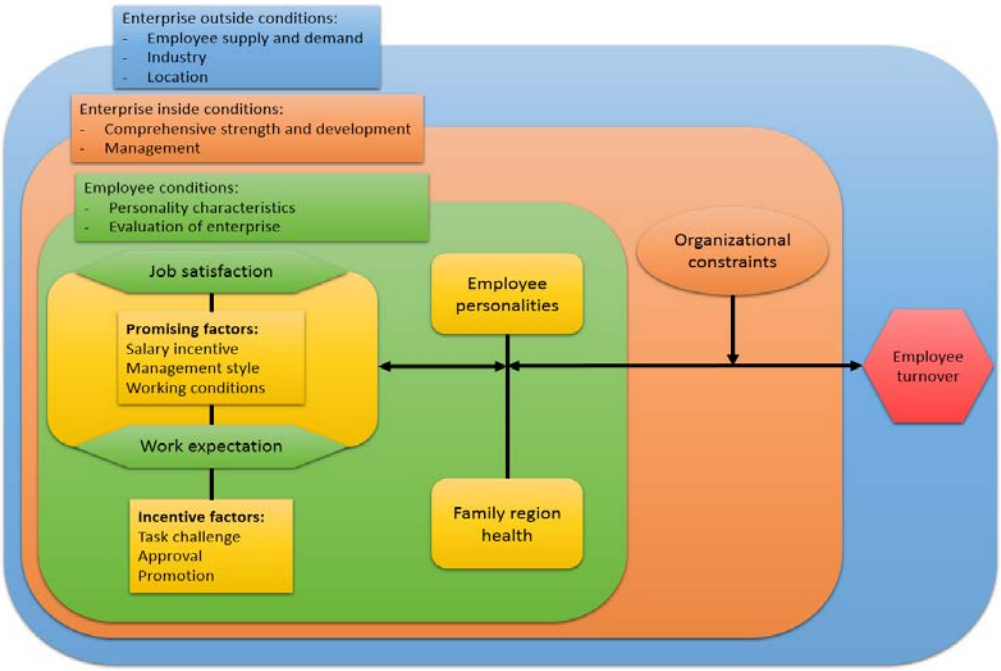

Figure1. Employee turnover factors

Turnover can be considered as a subgroup of human resource management (HRM). HRM function is to motivate employees and enhance workforce effectiveness. Integrating information technologies and HRM will provide smarter work. Martinsons (1997) considers this integration as knowledge based system in HRM that its applications are summarized in figure 2.

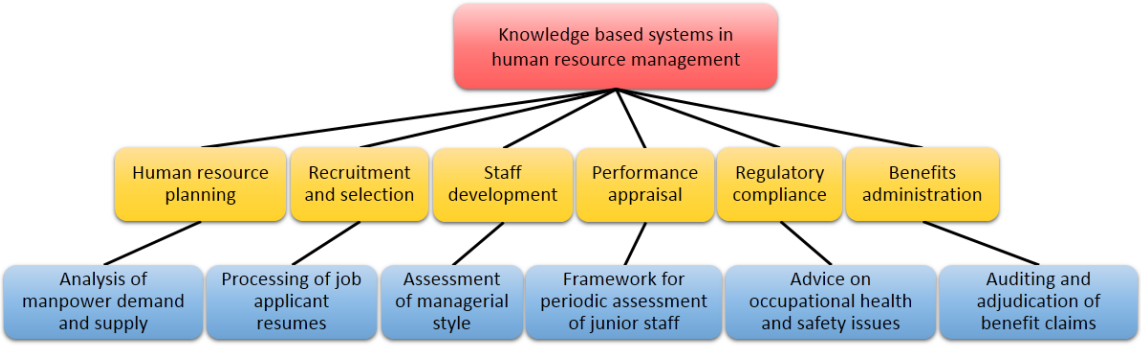

Figure2. Applications of knowledge based system in human resource management

Employee turnover survey is one of the issues of recruitment and selection part of HRM and it is done through analysis of existing and previous activities of employees and resumes of applicants. Table 1 summarizes the previous researches on turnover based on data mining (DM) methods. These Researches may concentrate on organizational or managerial characteristics such as organization size, organization site (like construction), working environment, relationships, anxiety, job security and etc.

Table 1. Relevant papers of human resource management by data mining methods.

Title	Models implemented	Authors
Data mining to improve human resources in construction company	Decision tree	Chang Youzheng (2008)
Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry	Decision tree (Chi-squared automatic interaction detection)	Chien and Chen (2008)
Employee turnover: a neural network solution	Neural network	Sexton et al. (2005)
Data mining for selection of insurance sales agents	Discriminant analysis, decision tree (C4.5) and neural network (feed-forward)	Cho and Ngai (2003)
Job performance prediction in a call center using a naive Bayes classifier	Naive Bayesian classifier	Valle, Varas et al. (2012)
Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals	Neural network and self-organizing map	Fan, Fan et al. (2012)

This study aims on prediction of voluntary turnover of employees based on their characteristics by DM concepts. In personnel recruitment, the aim of industry focuses on learning rate, productivity, and stability of workers but unstable employees are more detrimental and wasteful for firms. A poor hiring procedure can eventuate to loss of money, opportunity, credibility and even the current employees. Careful recruitment is essential for adverse workplace conditions to mitigate adversity and reduce employee turnover (Cottini, Kato et al., 2011). Prediction of unstable employees can reduce a great part of future loss of firms. In addition to hiring procedure, a frequent monitoring is needed to reveal changes in turnover trend in future. This monitoring and revisions in long term should have minimal complexity and robust accuracy on classifications. Hence a comparison of models with respect to tuning parameters and time consumption is required.

This paper is a data mining approach for prediction of quit among staff with an empirical research on a manufacturing company staff. Data mining concept, its subfields and models rarely are utilized in subjects like satisfaction and turnover.

2- Methods

2-1- Data mining

Data mining is a stage of the knowledge discovery process containing analyses and methods for finding implicit patterns, trends and relationships in data. The target is to acquire required knowledge for future decision makings. Data mining analyses are anomaly detection, association analysis, clustering and prediction modeling (Linoff & Berry, 2011). Association analysis is a research on finding implicit correlation among items, simultaneous events or frequent patterns such as consumption of complementary goods. It has wide range of usage such as in market management, telecommunication networks and inventory control. Prediction and classification models contain various model types such as decision trees, artificial neural networks and regressions. Clustering in contrast is an unsupervised method. Unsupervised methods don't have a class or target feature and all data features seen as distances instead of categories. Clustering procedure can be connectivity based, centroid based and density based.

2-2- Multilayer perceptron

Multilayer perceptron (MLP) is a simple artificial neural network that is feed forward. This kind of ANN has three layers of input, hidden and output that each layer is fully connected to next layer. Every layer contains nodes (neurons) that number of nodes in input layer depends on number of non-class/target features and in output layer depends on number of class/target features. Number of hidden layer nodes is user input but it is usually between 10 to 20 nodes. Except input layer, each node has an activation function. Each connection between two nodes has a weight that input is multiplied and passed through it and there is a summation before each activation function. Hidden layer facilitate nonlinearity and flexibility fit on data. MLP utilizes back propagation algorithm as learning part. Back propagation is gradient descend of each weight with respect to error of output and target values and it updates weights of ANN to reach less error.

2-3- Probabilistic neural network

Probabilistic neural network (PNN) is a kind of ANN that has similar structure to MLP but it has an extra class layer between hidden layer and output layer. Hidden layer contains one node for each case in the training data set. Hidden nodes are not fully connected to class nodes. In fact, class layer is also called summation layer because it doesn't contain activation function and simply sums input values of each category. Output layer in PNN has one node that decides category of input data. Activation functions in hidden nodes are usually radial basis function (RBF) (Specht, 1990).

2-4- Support vector machine

Support vector machine (SVM) first is introduced by Cortes and Vapnik (1995) for classification and regression purposes. SVM is a hyperplane that divides data into separate classes or fits a function on data and it has a kernel those effects on the shape of the classification function that is converted to hyperplane in feature space. SVM minimizes training data error and generalization error simultaneously. In order to minimize generalization error hyperplane margins are minimized. This means reduction of train error and reduction in hyperplane complexity. SVM learning process is quadratic optimization hence it will reach to a unique model each time implemented on the same data. SVM has various types of kernel that Gaussian RBF is one of the most popular kernels.

2-5- Classification and regression tree

Classification and regression tree (CART) recursively partitions on a nominal target category to reach a tree structure. Input of CART can be nominal or numerical that the term regression is associated to numerical input. As decision tree grows, a feature must be identified to split on it. So all features are compared to each others to select the best feature. This comparison can be done by Gini index that measures pureness of feature separation. CART stopping rule occurs when target feature in last separations are insignificant(Breiman, Friedman et al., 1984).

2-6- K-nearest neighbor

K-nearest neighbor (KNN) is a classifier that specifies categories based on training data. Input data in KNN are numerical. K nearest training points of each input point is discovered and the dominant category is assigned to input data. If the number of categories is equal, then neighbor points are weighted by their distance to find dominant category in them. In this model, induction is delayed to run time hence it is considered as lazy classifier.

2-7- Naive Bayes

Naive Bayes (NB) is a classifier that is based on Bayes' theorem. This classifier has a strong and unrealistic assumption of independency of all features in data. NB estimates parameters like mean and variance of each feature in training data. In nominal features, the number of occurrences of class value is divided by the number of total occurrences, and in numerical features Gaussian distribution is used. Output of NB considers the most probable class.

2-8- Random forest

Random forest (RF) is an ensemble learning method that constructs multiple decision trees. Each decision tree can be implemented by CART procedure. RF samples randomly training data with replacement on constructing each decision tree that is called bagging. Each decision tree returns a class and then bagging combines them to reach a unique decision(Breiman, 2001).

2-9- Apriori

Apriori is an association rule algorithm and can be considered as a subset of rule induction methods. Apriori deals with item sets and frequent rules. Item sets can be separated to consequences and antecedents to form if-then rules. One of the main aims in association rule is the reduction of candidate item sets that apriori utilizes purifying process to avoid extra calculations (Rekesh & Remekrishnen, 1994). Purifying in apriori is done by support and confidence thresholds that are defined by user. Support is the portion of an item set in all data. For instance 20 percent support means that in all data 20 percent contains that item set. Confidence measures whether antecedents of a rule have the rules ‘consequences or other consequences. Other consequences reduce confidence of the rule.

2-10- CN2 algorithm

CN2 rule induction method is a combination of AQ and ID3 algorithms. This method utilizes beam search algorithm that identifies shortest path in a graph. Beam search algorithm has pruning tree characteristic of ID3 algorithm hence it can limit rules count. CN2 in contrast to decision trees, constructs ordered if-then rules hence it has AQ algorithm systems advantage. The weighted relative accuracy measure is used as search heuristic that showed significant improvement in handling minority classes and reduction in number of generated rules (Clark & Niblett, 1989; Lavrač, Kavšek et al., 2004). Each rule has quality and coverage that quality is measured by accuracy on train dataset and coverage is similar to support in apriori algorithm.

3- Empirical study

Sample data in this study is a list of Arak company staff characteristics that comprises age, working experience in the current company, education, marriage state, socio-demographic characteristics, technical skills, physical fitness, etc. Arak data passed through cleaning, purifying and transformation processes as the initial steps of knowledge discovery. This research is based on the Cross-Industry Standard Process for DM (CRISP-DM) which is the basic of DM framework.

3-1- CRISP-DM Methodology

CRISP-DM that stands for Cross Industry Standard Process for Data Mining is based on the problem-solving strategy that can well fit the DM process, containing six phases. This process is based on six iterative main sectors which respectively are: business understanding, data understanding, data preparation, modeling, evaluation and deployment (Chapman, 2000). These steps are illustrated in figure 3.

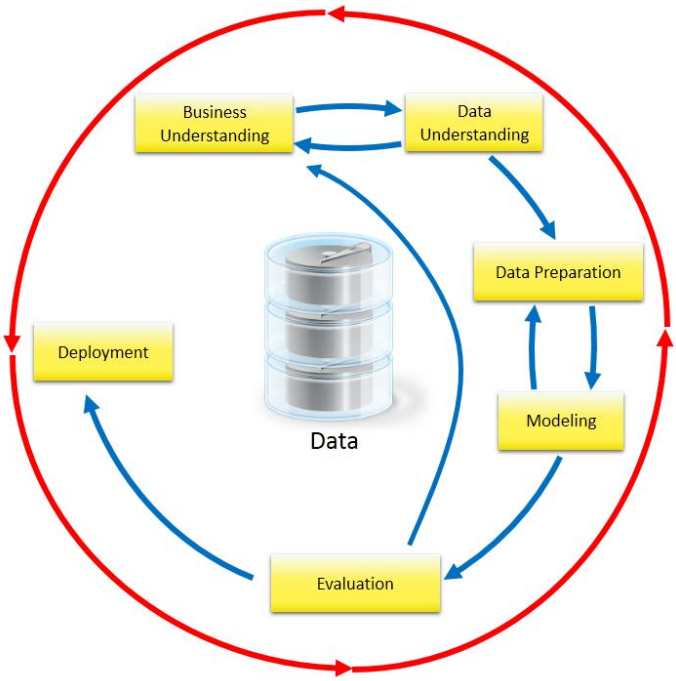

Figure 3. Crisp data mining methodology

3-2- Business Understanding

This research is performed on the employees that are working in a company in Arak province of Iran. This company is a supplier of automobile manufacturing's and products automotive parts. It consists of four sections which are assembly, press, welding and painting.

Training skillful staff is so costly in this company. So having a proper strategy toward keeping the best employees is essential. Lack of such strategy is one of the major problems of this company. This research aims to find the causes of this problem and making a Decision Support System (DSS) which is capable of detecting the employees that are highly probable to leave their jobs by means of DM algorithms. This DSS will help the managers to make the best decisions toward these kinds of employees.

3-3- Data Understanding

The data used in this research is obtained by the human resource management division of the company. This data is gathered in two years 2012 and 2013. Arak company human resource management data features are illustrated in table 2.

Table 2. Data features and their types.

Field number	Field type	Field description
1	Binary	Status (leaved or remained)
2	Continuous	Working experience in the current company (in days)
3	Continuous	Age (in years)
4	Set	Education
5	Binary	Marital status
6	Binary	Veteran status
7	Continuous	Duration of relevant work experience (in years)
8	Set	Compatibility of the work experience with the current job
9	Continuous	Number of job changing
10	Set	Social interaction ability
11	Set	Perseverance and interest to work
12	Continuous	Knowledge of working conditions and laws (between 1 to 100)
13	Continuous	Average working experience in other companies (in years)
14	Set	Technical skills
15	Set	Compatibility of body with job

It should be mentioned here that "Veteran Status" is so important in Iran since serving the military is compulsory for all of the males there and if one is exempted from military service in Iran that shows he had a problem. Therefore usually companies are reluctant to accept the ones who are exempted from military service as their employees.

3-4- Data Preparation

The whole database consists of two distinct databases. The first one was the data regarding to the employees who leaved their jobs and the second one was regarding to the employees who remained in their jobs. These two databases were merged together by considering their shared fields and a whole complete database was made.

The data used in this research is incomplete and needs to be purified. The records which consists null values are omitted from the database. The database of the employees who remained in their jobs consisted almost 80 percent of the whole database. Sampling training data may contain few data with minority class. This problem not only causes the outputs of models to have high variances, but also causes errors in the models due to existence only one class in training data. In order to obviate this problem, each of the two classes is equally sampled from data and sampling procedure is combined with the evaluation section.

3-5- Modeling

In the modeling process the first step is defining the target. This research aims to detect the employees who may probably leave their jobs. The first field in the table 2 is the target, i.e. status of employee. The inputs were selected by using the Pearson Chi-Square method. Considering that the target is categorical and the inputs are both categorical and continuous, this method is capable to test the independence of the target and the input fields as the predictors. The importance of each field is calculated as (1-P), in which p is the p value of the Pearson Chi-Square test between the candidate predictor and the target. Importance value of each input is shown in table 3.

The predictor variables were selected based on their importance value. The predictors with importance value more than 0.9, were selected and utilized in the next steps. Unimportant predictors are excluded.

Our MLP model has one hidden layer with 10 neurons and 1000 iteration limit. PNN leaning utilizes dynamic decay adjustment that utilizes two activation thresholds of activation function to avoid conflicts with rules of different classes which are θ^+ and θ^- with values of 0.4 and 0.2 respectively (Berthold & Diamond, 1995). SVM utilizes Gaussian RBF kernel with RBF sigma value of ten. CART used gini index on separation of nodes and minimum number of records in each leaf was 50 records of training data. Also in order to prune tree, the cost of tree's complexity is measured to balance misclassification risk and the complexity of the tree. KNN uses 3 nearest neighbor as input in euclidian space. Random forest randomly sampled 70 percent of training data with same CART configurations for each decision tree and constructs 50 decision trees.

3-6- Evaluation

In order to evaluate the prediction of models, k-fold cross validation is implemented to reduce bias of sampling data and ensuring model error randomness. K-fold cross validation divides data into k subsets randomly that one subset is used as training data and k-1 subsets are used as test data. This process is repeated k times to cover all data. In this paper 10-fold cross validation is implemented. Mean of 10 accuracies and construction time of each model is used to evaluate model performance. Rule induction models are evaluated by quality of rules and number of rules for each class of target feature.

3-7- Implementation

To implement a knowledge-based system a shared data format is required. This format for instance should have features in column format and just first row should be header of that feature and null values should be in respect to file format. For example in weka arrf format, question mark and in excel spreadsheet, blank cell. If data had simple text format a standard sign for null values should be defined. If data input process is managed well, conversion of data files and queries will be easily implemented. Furthermore Wang et al. (2011) discussed on data integrity, consistency, access security, basics of DSS and database specialized for employee turnover context. Sequentially modeling comes after data integrating, data cleaning and null handling. New employees' resume are entered to models in order to facilitate decision process of managers.

4- Results

4-1- Modeling results

Table 3 shows importance of each feature in database that guides to preclusion of irrelevant features for next steps. Features importance with respect to the target feature, status, is calculated by statistical measures, implied that features such as number of previous job changes, knowledge about the working conditions, perseverance and interest to work, compatibility of body with job were important. In contrast, features like social interaction ability, age and marital status were insignificant.

Table 3. Data features importance table

Field description	Importance
Number of job changing	1.00
Knowledge about the working conditions and laws	1.00
Perseverance and interest to work	0.998
Compatibility of body with job	0.991
Technical skills	0.984
Compatibility of the work experience with the current job	0.982
Average working experience in other companies	0.96
Duration of relevant work experience	0.912
Working experience in the current company	0.834
Level of education	0.706
Veteran status	0.541
Social interaction ability	0.364
Age	0.258
Marital status	0.254

Models prediction by 10 fold cross validation is shown in table 4. Random forest has the best prediction ability and also KNN and SVM are the worst in comparison. While training MLP with less than 1000 iterations and SVM with sigma value less than 8 they has significant decrease in accuracy. Recently researches on optimal kernel parameters such as Min and Lee (2005) that used grid search and k-fold cross validation and Liu, Zuo et al. (2012) that used class separation measure to reduce calculation time. Also SVM is unable to classify with polynomial and hyper tangent kernels. PNN with all possible values of θ^+ and θ^- has consistent accuracy, except very extreme values that are close to one or zero. KNN with higher values of k had similar results. CART with implementation of pruning tree has similar accuracies.

Table 4. Accuracies of implemented models

Prediction accuracy	MLP	PNN	SVM	CART	KNN	NB	RF
Fold1	0.82	0.76	0.74	0.74	0.76	0.92	0.88
Fold2	0.7	0.74	0.72	0.76	0.76	0.9	0.86
Fold3	0.78	0.76	0.86	0.78	0.8	0.8	0.9
Fold4	0.8	0.76	0.78	0.84	0.74	0.82	0.96
Fold5	0.88	0.78	0.66	0.82	0.7	0.88	0.86
Fold6	0.86	0.8	0.72	0.86	0.76	0.88	0.96
Fold7	0.82	0.78	0.72	0.72	0.74	0.9	0.94
Fold8	0.8	0.72	0.82	0.8	0.76	0.9	0.92
Fold9	0.86	0.78	0.72	0.84	0.72	0.92	0.9
Fold10	0.88	0.76	0.8	0.88	0.78	0.96	0.88
Average	0.82	0.764	0.754	0.804	0.752	0.888	0.906

Since the prediction accuracies of the above models are very similar, performance of models is compared by their processing time with full and partial data amount that shown in figure 2. Processing time of models are combined with 10-fold cross validation. This means that ten times sampling, training and testing are timing for each model. Also Random forest and CART didn't utilize any post prune technique in recording elapsed time of processing. SVM, PNN and MLP had significant increase in processing time as data size grows. Also SVM and PNN with some parameters took immense amount of time that could be considered unstable.

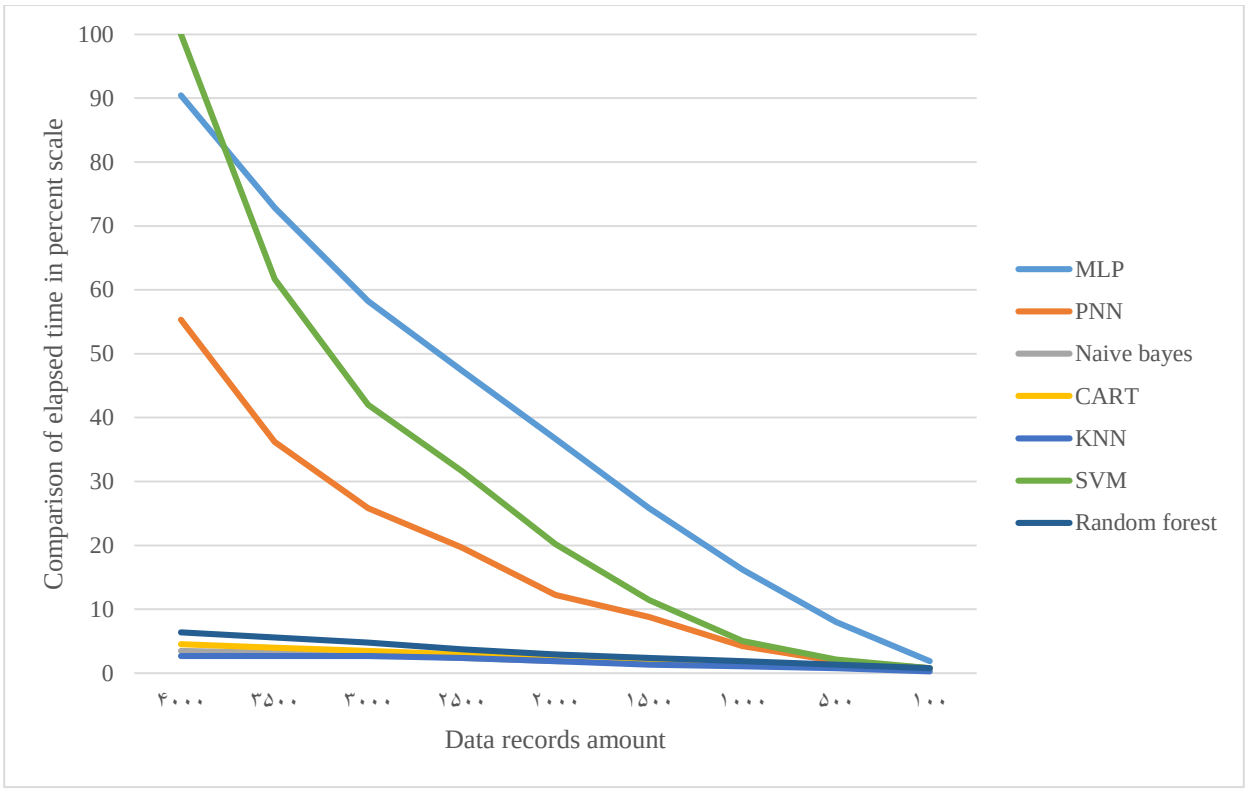

Figure 4. Elapsed time of full and partial data amount for implemented models

Figure 5 illustrates the number of rules extracted in support and confidence space for each of stayed and left classes with apriori algorithm.

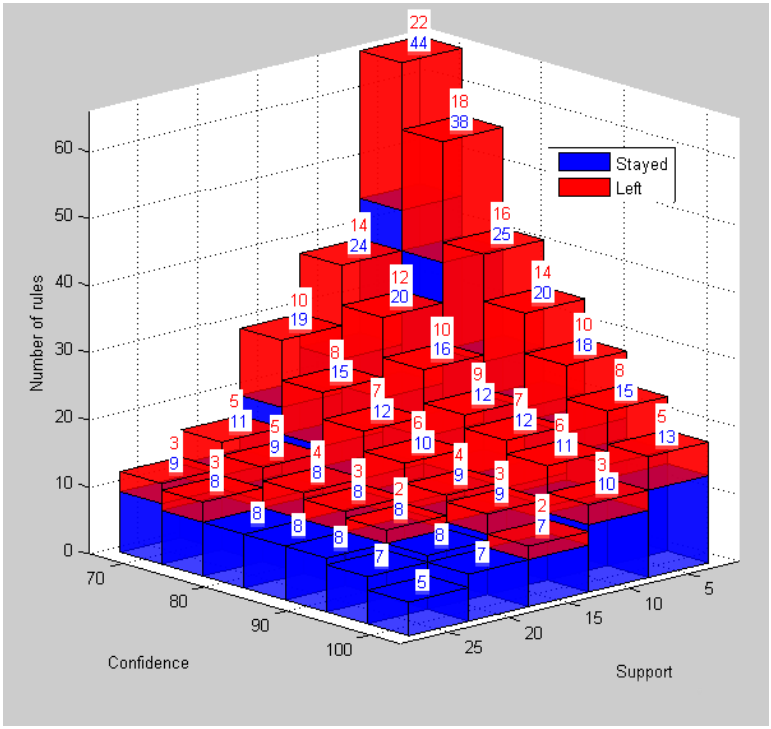

Figure 5. Number of rules extracted by apriori for a range of support and confidence limits

CN2 algorithm is implemented by various beam widths. Beam widths larger than three had exactly the same results. So the beam width equal to five is selected. Weighted relative accuracy measure is utilized to improve CN2 algorithm performance. Extracted rules are illustrated in table 5.

Table 5. Rules extracted by CN2 algorithm with weighted relative accuracy measure.

Rule number	Consequent	Antecedent	Support	Quality
1	Status = remained	Number of job changing = 0 and Perseverance and interest to work = good and Technical skills = average	5.85	95
2	Status = remained	Number of job changing = 0 and Knowledge about the working conditions and laws > 92 and Compatibility of body with job = average	6.58	92
3	Status = remained	Number of job changing = 0 and Technical skills = average	4.35	83
4	Status = left	Body compatibility = poor and Technique skills = no skill	3.6	82
5	Status = remained	Knowledge about the working conditions and laws > 92 and Level of education = diploma and Perseverance and interest to work = good	27.17	76
6	Status = left	Duration of relevant work experience <= 2 and Average of previous jobs experience < 1 and Working experience in current company <= 315.00	5.12	71
7	Status = left	Average working experience in other companies > 11.00	6.2	68
8	Status = remained	Knowledge of work conditions and laws > 70 and Perseverance and interest to work = good and Education degree = diploma and Working experience in the current company > 3000.00	21.15	65

4-2- Managerial results

The last step of CRISP-DM is deployment of constructed models. In order to utilize the constructed models, a graphical user interface (GUI) had been developed specifically for the automotive parts manufacturer. This GUI is illustrated in figure 6 that facilitates recruitment process and also triggers revival of CRISP-DM cycle. In figure 6 the GUI in part one, gets data from database in part five and the user inputs a new employee candidate. In part two recruitment candidates' data are concatenated to current employees' data and data preparation is implemented on it. In part three, feature selection is deployed. Extracted features are fed into the rule extraction and the prediction part. Part three illustrates the predictive models' meta nodes and part four is the core model training process with 10-fold cross validation. Part four feeds database in part five, and invokes the new rules and probability of new candidates' stability according to collected outputs of all models. At last, the database in part five sends data to GUI in order to summarize and visualize the results to facilitate decision making process of managers. Hence in this case study, this GUI is the front end part of the research.

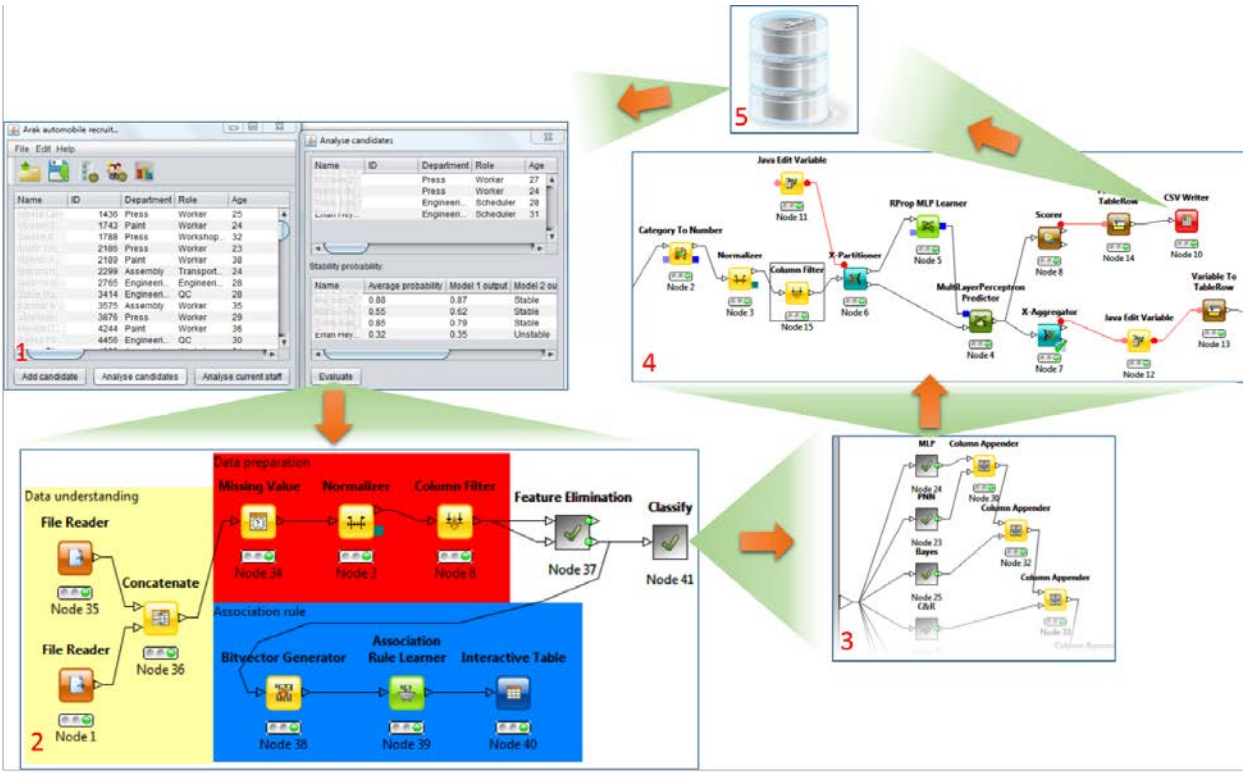

Figure 6. The specific process designed for the case study.

5- Conclusion

This study uses an industrial plant human resource management data in order to extract turnover pattern among employees. Integration between human resource management and DM will improve recruitment and decision making process. Based on the prediction model and association rule model, it is shown that turnover is predictable and by using DM, the management can rely on it as a DSS for human resource recruitment. Revised recruitment process can filter unstable staff in long term and make organization stable. As the CRISP-DM cycle restarts, importance of features can help revision of human resource database to discard unimportant features and to add new features.

User friendliness and time consumption are main factors for a robust DSS in knowledge based systems. All knowledge based systems require minimal tuning parameters to minimize complexity of implementation and need for experts. Results show that SVM, PNN and KNN are sensitive to parameters. In contrast, Naive Bayes is the most user friendly model that has a good performance in classification.

In order to construct rules, simple apriori algorithm not only is sensitive to parameters but also constructs many confusing rules for major classes in order to find minority class rules in the data. Hence simple apriori algorithm has the worst performance in user friendliness. Instead, CN2 algorithm with weighted relative accuracy measure always construct reasonable number of rules with consideration of minority classes and also CN2 with weighted relative accuracy measure is significantly faster than simple apriori algorithm.

Additionally, the processing time of models must be considered. PNN has significant increase in processing time as the amount of data grows. MLP is slow in training but it has the advantage of continual training. So in the future it can perform with less number of iterations. By consideration of accuracy, time and user friendliness, decision trees generally had the best performance.

In order to provide user friendliness, only low maintenance requirements of models are not sufficient. To achieve real user friendliness, a GUI is required to handle human resource management, reconstruction of models, visualize results, facilitate decision making and restart of this process.

References

Beach, L. R. (1990). Image Theory: Decision Making in Personal and Organizational Contexts. *European Journal of Operational Research, 47*(2), xv, 254 p. doi: 10.1016/0377-2217(90)90287-l

Berthold, M. R., & Diamond, J. (1995). Boosting the performance of rbf networks with dynamic decay adjustment. *Advances in neural information processing systems*, 521-528.

Boles, J. S., Dudley, G. W., Onyemah, V., Rouziès, D., & Weeks, W. A. (2012). Sales force turnover and retention: A research agenda. *Journal of Personal Selling & Sales Management, 32*(1), 131-140.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*: CRC press.

Chalkiti, K., & Sigala, M. (2010). Staff turnover in the Greek tourism industry: A comparison between insular and peninsular regions. *International Journal of Contemporary Hospitality Management, 22*(3), 335-359.

Chang Youzheng, G. M. (2008). Data Mining to Improve Human Resource in Construction Company. *International Seminar on Business and Information Management, 1*(19), 275 - 278.

Chapman, P., et al. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chien, C.-F., & Chen, L.-F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with Applications, 34*(1), 280-290. doi: 10.1016/j.eswa.2006.09.003

Cho, V., & Ngai, E. W. (2003). Data mining for selection of insurance sales agents. *Expert systems, 20*(3), 123-132.

Clark, P., & Niblett, T. (1989). The CN2 induction algorithm. *Machine learning, 3*(4), 261-283.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Cottini, E., Kato, T., & Westergaard-Nielsen, N. (2011). Adverse workplace conditions, high-involvement work practices and labor turnover: Evidence from Danish linked employer–employee data. *Labour Economics, 18*(6), 872-880. doi: 10.1016/j.labeco.2011.07.003.

DeConinck, J. B., & Johnson, J. T. (2009). The effects of perceived supervisor support, perceived organizational support, and organizational justice on turnover among salespeople. *Journal of Personal Selling & Sales Management, 29*(4), 333-350.

Fan, C.-Y., Fan, P.-S., Chan, T.-Y., & Chang, S.-H. (2012). Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Systems with Applications, 39*(10), 8844-8851. doi: 10.1016/j.eswa.2012.02.005

Hancock, J. I., Allen, D. G., Bosco, F. A., McDaniel, K. R., & Pierce, C. A. (2013). Meta-analytic review of employee turnover as a predictor of firm performance. *Journal of Management, 39*(3), 573-603.

Holtom, B. C., Mitchell, T. R., Lee, T. W., & Eberly, M. B. (2008). 5 Turnover and Retention Research: A Glance at the Past, a Closer Review of the Present, and a Venture into the Future. *The Academy of Management Annals, 2*(1), 231-274.

Iverson, R. D., & Pullman, J. A. (2000). Determinants of voluntary turnover and layoffs in an environment of repeated downsizing following a merger: an event history analysis. *Journal of Management, 26*(5), 977-1003. doi: 10.1016/s0149-2063(00)00065-9.

Lavrač, N., Kavšek, B., Flach, P., & Todorovski, L. (2004). Subgroup discovery with CN2-SD. *The Journal of Machine Learning Research, 5*, 153-188.

Linoff, G. S., & Berry, M. J. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*: Wiley.

Liu, Z., Zuo, M. J., & Xu, H. (2012). *Parameter selection for Gaussian radial basis function in support vector machine classification.* Paper presented at the Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE).

Martinsons, M. G. (1997). Human resource management applications of knowledge-based systems. *International Journal of Information Management, 17*(1), 35-53.

Min, J. H., & Lee, Y.-C. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), 603-614.

Ongori, H. (2007). A review of the literature on employee turnover. *African Journal of Business Management*, 049-054.

Racz, S. (2000). Finding the Right Talent Through Sourcing and Recruiting. *STRATEGIC FINANCE -MONTVALE-*, 38-44.

Rekesh, A., & Remekrishnen, S. (1994). Fast Algorithms for Mining Association Rules. *PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES*, 487.

Sexton, R. S., McMurtrey, S., Michalopoulos, J. O., & Smith, A. M. (2005). Employee turnover: a neural network solution. *Computers & Operations Research, 32*(10), 2635-2651. doi: 10.1016/j.cor.2004.06.022.

Specht, D. F. (1990). Probabilistic neural networks. *Neural networks, 3*(1), 109-118.

Valle, M. A., Varas, S., & Ruz, G. A. (2012). Job performance prediction in a call center using a naive Bayes classifier. *Expert Systems with Applications, 39*(11), 9939-9945. doi: 10.1016/j.eswa.2011.11.126.

Wang, X., Wang, H., Zhang, L., & Cao, X. (2011). Constructing a decision support system for management of employee turnover risk. *Information Technology and Management, 12*(2), 187-196.

Zimmerman, R. D. (2008). Understanding the impact of personality traits on individuals'turnover decisions: A meta-analytic path model. *Personnel Psychology, 61*(2), 309-348.