Human Resource Analytics

Chandrasekar Rajasekar - crajase
Dinesh Prasanth M K - dmolugu
Vishal Murugan - vmuruga
Vivek Mani - vmani2

1 Background

1.1 Introduction

Every company wants to hire the best candidates and retain them over a long period. They provide attractive compensations in order to retain them. However, many such best employees switch over to other companies for various reasons; poor performance due to heavy workload, less promotions, low job satisfaction levels, salary. This will incur a huge loss for the company; the company has to hire a new resource and train him - takes several days to months to gain the knowledge/experience as the ex-employee.

This project aims to analyze the mindset of 15,000 employees (including people who have left the company) and predict the employees that the company is about to lose. Each employee is tested against the various features that include: Satisfaction level, Last rating, Number of projects completed, Number of years served, Work accidents, Promotions, Department, Salary

1.2 Literature Survey

The paper compares different supervised learning approach to solve a classification problem. The paper compares Decision Tree, Naive Bayes, Perceptrons, kNN, SVM and rule based with respect to different performance parameter like accuracy, speed of learning/classifying etc. However the dataset used to solve these problem are different for each classifier. For instance all attributes in the dataset used in decision tree problem have only categorical variable whereas all attributes in Perceptron learning are of type ratio - (which is not quite possible with real world data). Real-world data is a mix of Nominal, Ordinal, Interval, and Ratio.

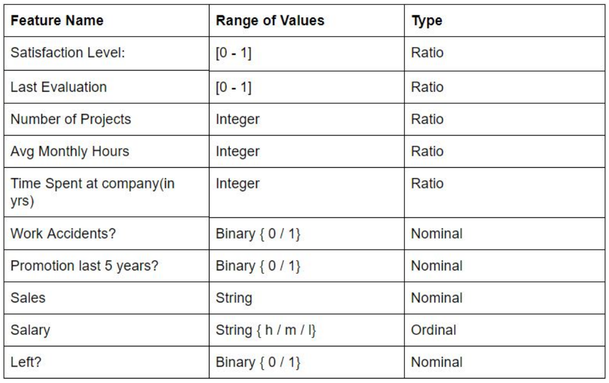
We propose to compare different supervised model by solving classification problem on same dataset - transforming data so that it fits each model. We compare the result of all the model. In this way, we compare the performance of each model in real world data not just dataset that is suitable for that model. Also, due to this change, our model comparison may not be the same as the one given in the paper, and this change may be biased toward the problem we are trying to solve.

2 Methods

The dataset for this problem has 10 attributes with about 15,000 employee observations. Luckily, since the data was found to be clean, we did not have to clean the dataset. We started to analyze the dataset using ID3, Naive Bayes and we try to get the best model using a cost matrix created for this kind of problem.

2.1 Data Transformation

Features and their types are in figure 1



Feature Name	Range of Values	Type
Satisfaction Level:	[0 - 1]	Ratio
Last Evaluation	[0 - 1]	Ratio
Number of Projects	Integer	Ratio
Avg Monthly Hours	Integer	Ratio
Time Spent at company(in yrs)	Integer	Ratio
Work Accidents?	Binary { 0 / 1}	Nominal
Promotion last 5 years?	Binary { 0 / 1}	Nominal
Sales	String	Nominal
Salary	String { h / m / l}	Ordinal
Left?	Binary { 0 / 1}	Nominal

Figure 1: Dataset Attribute Table

For neural networks, it is required that our attributes should be ratio since we multiple each attributes with their weights. In addition, the neural network also work with binary attributes (can be nominal provided the attribute should only contain 0/1).

Our dataset mostly contains Ratio or binary nominal which suits for Neural networks. However the ’Sales’ attribute and ’Salary’ attribute contains string which has to be changed to numeric values without affecting its Nominal or Ordinal Properties. For example, the attribute ’Sales’ was changed from a ’Nominal String’ to a ’One hot vector’. Similarly, the attribute ’Salary’ was changed from an ’Ordinal String’ and mapped to a map of Integers in sorted order {h=>3, m=>2, l=>1}.

2.2 Decision Tree

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item’s target value (represented in the leaves). The Measure of Misclassification in the tree such as Gini or Information Gain is used to select the best attribute for the tree. We built two decision trees one using GINI error and other using Information Gain to build the tree.

2.3 Naive Bayes

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes’ theorem with strong (naive) independence assumptions between the features.

2.4 Support Vector Machine

A support vector machine constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

2.5 Logistic Regression

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

3 Experiment And Results

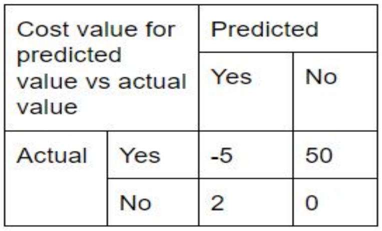
We will be analyzing many models and find out the best model to apply which would give us the best classification for the given dataset.

3.1 Dataset

We will be using a Human Resources Analytics dataset from Kaggle. This dataset has 10 attributes and 14999 rows or observations. We have used 80% training data and 20% test data for all models. Data set: Kaggle Dataset.

3.2 Cost Matrix

The cost matrix in figure 2 was used to evaluate the models. When we predict that an employee might leave the company, the company could try to compensate them with more money or benefits.



Cost value for predicted value vs actual value		Predicted	
		Yes	No
Actual	Yes	-5	50
	No	2	0

Figure 2: Cost Matrix

- True Positive: If we correctly identify that the employee is leaving, the company's efforts to keep the said employee might be effective in saving a lot of future costs (for training a new employee).
- False Positive: The employee might get good benefits but it would be of loss to the company as the employee would still be working without that compensation.
- False Negative: If the employee is leaving the company and the model is unable to predict it, then this incurs a huge loss to the company. So this type of error has a huge cost so as to only produce models with good recall.
- True Negative: This tells us that we do not need to worry about this employee leaving the company and which this is important to us, the impact this type of event has over errors and the frequency of this event occurring leads it to having a low positive cost.

3.3 Decision Tree

We have performed decision tree using ID3 and using GINI index.
Using the following configurations with Gini:

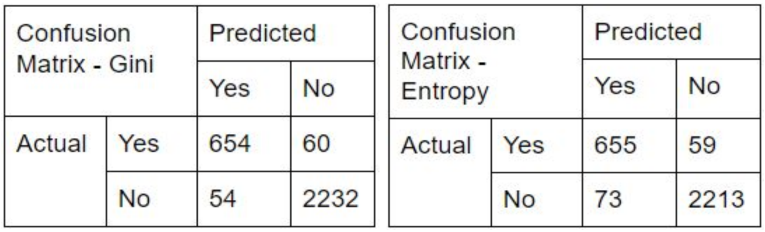
- class_weight='balanced'
- criterion='gini'
- min_weight_fraction_leaf=0.01
- splitter='best'

The accuracy for the above model = 96.2%
The total Cost of the model [Conf matrix * cost matrix] = -162 The confusion matrix is in figure 3

Using the following configurations with ID3:

- class_weight='balanced'
- criterion='entropy'
- min_weight_fraction_leaf=0.01
- splitter='best'

The accuracy for the above model = 95.6%
The total Cost of the model [Conf matrix * cost matrix] = -179 The confusion matrix is in figure 3



Confusion Matrix - Gini		Predicted	
		Yes	No
Actual	Yes	654	60
	No	54	2232

Confusion Matrix - Entropy		Predicted	
		Yes	No
Actual	Yes	655	59
	No	73	2213

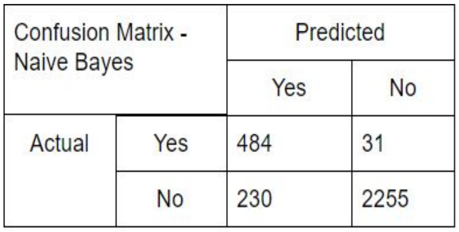
Figure 3: Confusion Matrices

3.4 Naive Bayes

Naive Bayes is built using klaR package in R. The naive bayes classifier is built using the following parameters:

- Cross validation using 10-fold cross validation

For this model, we get accuracy of 91.3%.
The total cost of this model = confusion matrix * cost matrix = -410 The confusion matrix is in figure 4



Confusion Matrix - Naive Bayes		Predicted	
		Yes	No
Actual	Yes	484	31
	No	230	2255

Figure 4: Confusion Matrix for Naive Bayes

By using this model, it is observed that False Negative is less (ie) probability of our model predicting that an employee will stay but actually leaves the company is l ow. This factor is the most important as this incurs the huge loss to the company.

3.5 Support Vector Machine

under construction

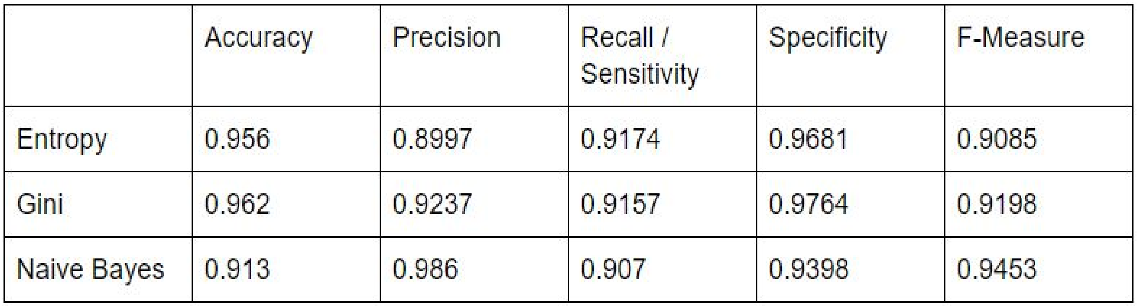
3.6 Logistic Regression

under construction

The comparison of all the experimented models are in figure 5

4 Conclusion

From figure 5 we can infer that using ID3 using information gain (or entropy) is better than GINI and Naive Bayes by a slight margin. This is because we want to maximize recall in our case, as we would not want to wrongly predict employees leaving the company. Based on the references we have read, we think that SVM and linear regression might give slightly better results for this data set.



	Accuracy	Precision	Recall / Sensitivity	Specificity	F-Measure
Entropy	0.956	0.8997	0.9174	0.9681	0.9085
Gini	0.962	0.9237	0.9157	0.9764	0.9198
Naive Bayes	0.913	0.986	0.907	0.9398	0.9453

Figure 5: Comparison of models

References

[1] Kotsiantis,S. B.(2007) *Supervised Machine Learning: A Review of Classification Techniques.*

[2] Tin Kam Ho. & M. Basu.(2002) *Complexity measures of supervised classification problems, IEEE.*