Predicting the Price of Used Cars using Machine Learning Techniques

Article · January 2014

CITATIONS
69

READS
20,126

1 author:

Sameerchand Pudaruth
University of Mauritius
79 PUBLICATIONS **584** CITATIONS

SEE PROFILE

International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764

http://www. irphouse.com

Predicting the Price of Used Cars using Machine Learning Techniques

Sameerchand Pudaruth[1]

[1]*Computer Science and Engineering Department, University of Mauritius, Reduit, MAURITIUS*

ABSTRACT

In this paper, we investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions. The predictions are then evaluated and compared in order to find those which provide the best performances. A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy. All the four methods provided comparable performance. In the future, we intend to use more sophisticated algorithms to make the predictions.

Keywords-car; price; machine learning; artificial intelligence

1. INTRODUCTION

Predicting the price of used cars in both an important and interesting problem. According to data obtained from the National Transport Authority [1], the number of cars registered between 2003 and 2013 has witnessed a spectacular increase of 234%. From 68, 524 cars registered in 2003, this number has now reached 160, 701. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. It is reported in [2] that the sales of new cars has registered a decrease of 8% in 2013.

In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financer. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to

seller/financers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financer at the beginning, the instalments will be higher for the clients who will certainly then opt for another seller/financer. If the residual value is over-estimated, the instalments will be lower for the clients but then the seller/financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of very high commercial importance as well. Manufacturers' from Germany made a loss of 1 billion Euros in their USA market because of mis-calculating the residual value of leased cars [3]. Most individuals in Mauritius who buy new cars are also very apprehensive about the resale value of their cars after certain number of years when they will possibly sell it in the used cars market.

Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance. Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven. Other factors such as the type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), safety index, its size, number of doors, paint colour, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, its physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well. Some special factors which buyers attach importance in Mauritius is the local of previous owners, whether the car had been involved in serious accidents and whether it is a lady-driven car. The look and feel of the car certainly contributes a lot to the price. As we can see, the price depends on a large number of factors. Unfortunately, information about all these factors are not always available and the buyer must make the decision to purchase at a certain price based on few factors only. In this work, we have considered only a small subset of the factors mentioned above. More details are provided in Section III.

This paper is organised as follows. In the next section, a review of related work is provided. Section III describes the methodology while in section IV, we describe, evaluate and compare different machine learning techniques to predict the price of used cars. Finally, we end the paper with a conclusion with some pointers towards future work.

2. Related Work

Surprisingly, work on estimated the price of used cars is very recent but also very sparse. In her MSc thesis [3], Listiani showed that the regression mode build using support vector machines (SVM) can estimate the residual price of leased cars with higher accuracy than simple multiple regression or multivariate regression. SVM is

better able to deal with very high dimensional data (number of features used to predict the price) and can avoid both over-fitting and underfitting. In particular, she used a genetic algorithm to find the optimal parameters for SVM in less time. The only drawback of this study is that the improvement of SVM regression over simple regression was not expressed in simple measures like mean deviation or variance.

In another university thesis [4], Richardson working on the hypothesis that car manufacturers are more willing to produce vehicles which do not depreciate rapidly. In particular, by using a multiple regression analysis, he showed that hybrid cars (cars which use two different power sources to propel the car, i.e. they have both an internal combustion engine and an electric motor) are more able to keep their value than traditional vehicles. This is likely due to more environmental concerns about the climate and because of its higher fuel efficiency. The importance of other factors like age, mileage, make and MPG (miles per gallon) were also considered in this study. He collected all his data from various websites.

Wu et al. [5] used neuro-fuzzy knowledge based system to predict the price of used cars. Only three factors namely: the make of the car, the year in which it was manufactured and the engine style were considered in this study. The proposed system produced similar results as compared to simple regression methods. Car dealers in USA sell hundreds of thousands of cars every year through leasing [6]. Most of these cars are returned at the end of the leasing period and must be resold. Selling these cars at the right price have major economic connotation for their success. In response to this, the ODAV (Optimal Distribution of Auction Vehicles) system was developed by Du et al. [6]. This system not only estimates a best price for reselling the cars but also provides advice on where to sell the car. Since the United States is a huge country, the location where the car is sold also has a non-trivial impact on the selling price of used cars. A k-nearest neighbour regression model was used for forecasting the price. Since this system was started in 2003, more than two million vehicles have been distributed via this system [6].

Gonggi [7] proposed a new model based on artificial neural networks to forecast the residual value of private used cars. The main features used in this study were: mileage, manufacturer and estimate useful life. The model was optimised to handle nonlinear relationships which cannot be done with simple linear regression methods. It was found that this model was reasonably accurate in predicting the residual value of used cars.

3. Methodology

Data was collected from <<petites announces>> found in daily newspapers such as L'Express [8] and Le Defi [9]. We made sure that all the data was collected in less than one month interval as time itself could have an appreciable impact on the price of cars. In Mauritius, seasonal patterns is not really a problem as this does not really affect the purchase or selling of cars. The following data was collected for each car: make, model, volume of cylinder (funnily this is usually considered same as horsepower in Mauritius), mileage in km, year of manufacture, paint colour, manual/automatic and price. Only cars which had their price listed were recorded.

Because many of the columns were sparse they were removed. Thus, paint colour and manual/automatic features were removed. The data was then further tweaked to remove records in which either the age (year) or the cylinder volume was not available. Model was also removed as it would have been extremely difficult to get enough records for all the variety of car models that exist. Although data for mileage was sparse, it was kept as it is considered to be a key factor in determining the price of used cars. A sample of the collected data is shown below in Table 1.

Table 1. *Sample Data Collection*

#	MAKE	CYLINDER VOLUME (CC)	YEAR	MILEAGE/KM	PRICE (RS)
1	TOYOTA	1300	2007	38000	410000
2	NISSAN	1500	2007	50000	325000
3	HONDA	1500	2005	59000	385000
4	TOYOTA	1000	2007	59000	360000
5	TOYOTA	1300	1989	62665	50000
6	TOYOTA	1500	2008	67000	615000
7	TOYOTA	1500	2008	69000	575000
8	TOYOTA	1490	2006	73000	450000
9	TOYOTA	1600	2006	82000	550000
10	TOYOTA	1000	2006	85000	325000
11	TOYOTA	1500	2000	113000	325000
12	TOYOTA	1500	2000	129000	218000
13	NISSAN	1500	2001	145000	195000

Initially, 400+ records were collected. However, after further pruning, for example, we kept only the three of the most popular makes in Mauritius, i.e. Toyota, Nissan and Honda. In particular, we removed all makes for which there were less than 10 records. Regarding the cylinder volume, for some cars, it was provided in a range. We then opted for the average value of the range. The final database contained only 97 records: Toyota (47), Nissan (38) and Honda (12). The values are then pre-processed in a form amenable to further processing using machine learning techniques. The minimum and maximum values for some numerical feature are shown in Table 2.

Table 2. *Minimum and Maximum Values*

#	CYLINDER VOLUME (CC)	YEAR	PRICE (RS)
Minimum	1000	1988	27, 000
Maximum	2160	2013	825, 000

4. Implementation and Evaluation

4.1. Multiple Linear Regression Analysis

The lack of mileage information for most of the cars did not allow us to use it to forecast the price. The Pearson correlation coefficient (r) was computed between different pairs of features [10]. The summarised results are shown below in Table 3.

Table 3. *Matrix of Pearson Correlation Coefficients*

	Cylinder volume	Year	Mileage	Price
Cylinder volume	1	-	-	0.33
Year	-	1	-0.33	0.81
Mileage	-	-0.33	1	-0.35
Price	0.33	0.81	-0.35	1

The value of r was found to be -0.33 between year and mileage, which means that there is a weak negative correlation between the year in which the car was manufactured and its mileage. The relation is not strong enough to use year information to predict mileage information. The relation between mileage and price is also very similar. On the other hand, there is a weak positive correlation between cylinder volume and price. This means that on average prices for cars with higher cylinder volume tends to be slightly higher. As anticipated, there is a very high correlation between the price of a car and the year in which it was manufactured. New cars have higher prices. Among the factors considered, we can see that year has the most impact. Since all the three makes that we have considered are Japanese manufacturers, we are assuming this will not impact on the price.

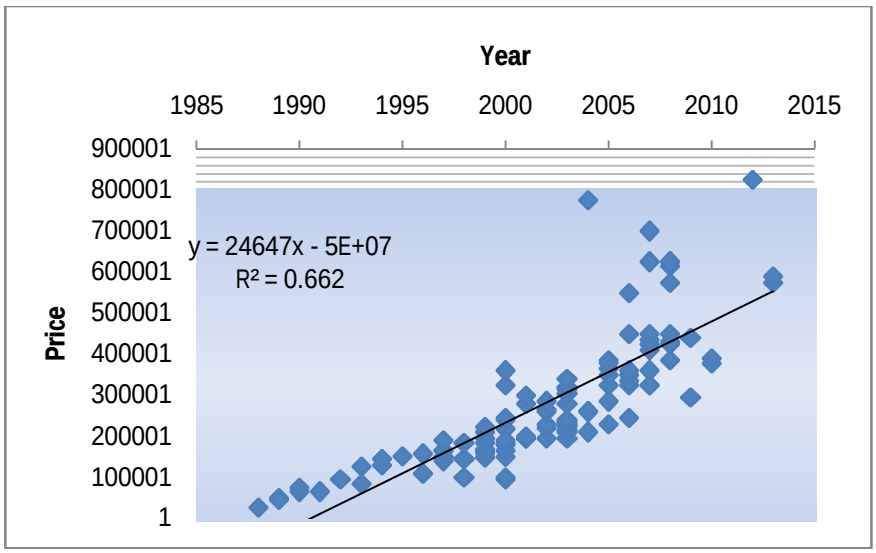

Figure 1. *Relationship between year of manufacture and price*

The graph shows that the relationship between year and price is not fully linear. There are many ups and down. In an attempt to increase the linearity, some outliers were removed. These were cars whose age were 20 and above or less than 4 years. The regression coefficient changed slightly from 0.814 to 0.819. Another variation was to use the logarithmic value of price instead of the actual values. In this case, the regression coefficient showed a significant increase from 0.814 to 0.851. In the first case, the mean deviation from the predicted price was Rs57468 while in the logarithmic scenario, it was at Rs51, 084. We can thus conclude that while logarithmic regression is slightly better than simple linear regression, linear regression on one variable alone is not sufficient to predict an accurate price for used cars. We also found that the regression coefficient is much higher for Nissan cars (0.917) than for toyota cars (0.803).

4.2. K-Nearest Neighbours (kNN)

K-nearest neighbour (IBk in Weka [11]) is a machine learning technique in which the new (unknown) data is compared to all the existing records in order to locate the best match(es) [12]. Despite its apparent simplicity, a lot of take has to be taken in pre-processing the data otherwise we can easily go off-track. Only three attributes were considered namely the make, year and cylinder volume. However, the data set was split into different sets, one containing only Toyota cars and the other only Nissan cars. This was done because most software cannot handle nominal values appropriate but this allowed us to compare the performance on each make. In general, it was also found that Toyota cars of the same age and cylinder are more expensive the Nissan cars with the same features. The data for year and cylinder had to be normalise to prevent large values (from one feature) from over-shadowing smaller values (from another feature). Thus, the following formulae were applied for normalising the data.

There were two options for normalising the year.

1. Normalised value for Year = (Year of Manufacture – 2014)/19 + 1
2. Normalised value for Year = (Year of Manufacture – 2010)/15 + 1

The second option was chosen because most used cars are more than 4 years old. Used cars which are less than 4 years old are very likely to be outliers and may significantly affect the prediction performance. The divisor is the span of years from the newest car to the oldest car in the database. One (1) is then added to the quotient to bring the value between 0 and 1 and to make sure that newer cars have higher normalised values than older cars.

The formula for normalising cylinder volume is as follows:

Normalised value = Cylinder Volume/Maximum(Cylinder Volume)

A simple formula was used to normalised the data for cylinder volume. We have to ensure that cars with higher values for cylinder volume have higher normalised values than cars with lower values. A snapshot of the normalised data is shown in Table 4 below.

Table 4. *Normalised Data*

MAKE	MODEL	CYLINDER VOLUME (CC)	YEAR	PRICE (RS)
TOYOTA	STARLET	0.601852	0	152000
TOYOTA	STARLET	0.615741	0.066667	157000
NISSAN	B14	0.694444	0.266667	160000
NISSAN	K11	0.462963	0.266667	165000
NISSAN	N16	0.601852	0.333333	165000
TOYOTA	CORONA	0.734722	0.133333	165000
TOYOTA	CAMRY	0.694444	0.133333	165000
NISSAN	MARCH	0.462963	0.266667	168000
NISSAN	K11	0.509259	0.333333	180000
HONDA	CITY	0.462963	0.333333	185000
TOYOTA	CORONA	0.694444	0.2	185000
TOYOTA	STARLET	0.601852	0.266667	185000
HONDA	EK3	0.694444	0.133333	190000

Table 5. *Cross-validation with 10 folds*

k	Toyota	Nissan
	Mean Absolute Error (Rs)	
1	45, 189	27, 258
3	54, 584	27, 134
5	62, 670	25, 741
10	63, 837	29, 289

For both the Toyota and Nissan cars, we used cross-validation with 10 folds while the value of k was varied. The results are presented in Table 4. We can clearly see that kNN works significantly better for Nissan cars than for Toyota cars. This seems to suggest that prices for Nissan cars are more consistent than prices for Toyota cars. Also, we can see that for Toyota cars, the best value of k is 1 and the performance degrades for increasing values of k while for Nissan cars, the best value of k is 5. But performance is almost the same even for lower values of k. Performance starts to degrade for higher values of k.

4.3. Decision Trees

Only Nissan and Toyota cars were considered for building the decision tree. The prices were grouped into six nominal categories as most of the popular decision tree algorithms cannot handle numeric outputs [13]. There are many gaps in the ranges that have been defined because there were no cars within these ranges although it is certainly possible to get new data which fits within these zones. These large gaps have been useful in determining the boundaries for the classes. The defined ranges can certainly be extended when more data is available. The classes/categories are shown in Table 6 below.

Table 6. *Nominal Categories for Price of Cars*

#	Minimum	Maximum	Category
1	95000	168000	A
2	180000	245000	B
3	260000	325000	C
4	335000	410000	D
5	425000	450000	E
6	550000	625000	F

J48 is a free Java implementation of the C4.5 decision tree algorithm which is found in Weka. The following attributes were used to build the tree: make, model, cylinder volume, year and price. In one run, the model was not used and the results are compared and shown in Table 7.

Table 7. *J48 Decision Tree Classification Results*

Type of Training	% Success With Car Model	% Success Without Car Model
80% Percentage Split	59	64
Cross-validation with 10 folds	66	65
Training Set	84	82

From Table 7, we can see that the attribute car model do not have much influence on the classification results. If performance in terms of CPU cycles and storage space is an issue, then perhaps it can be left out. The random forest decision tree algorithm was also used to classify the car data and the results are as shown below in Table 8.

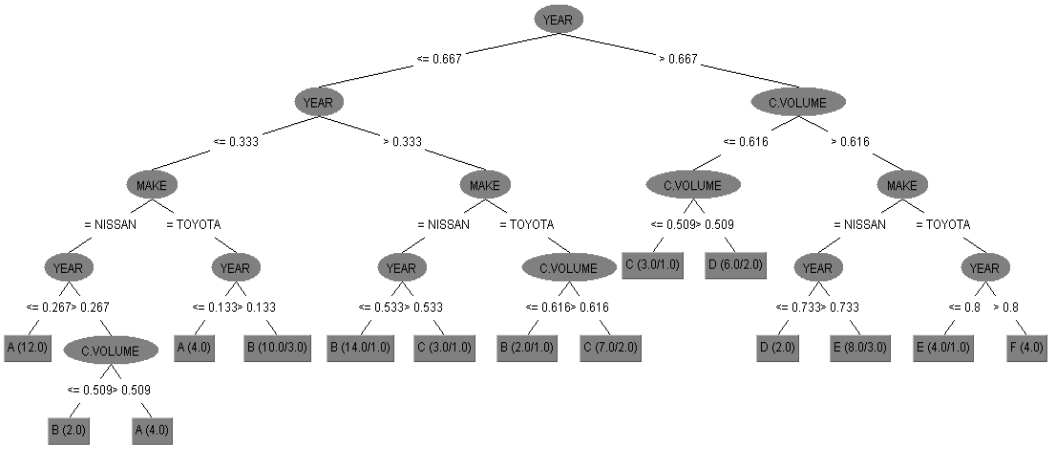

Figure 2. *J48 Decision Tree on Training Set Without Car Model*

The tree in Figure 2 shows that J48 has been able to identify the attribute YEAR as the most decisive feature. This tallies with the evidence found earlier by using regression analysis. For newer cars, the second most influential attribute was their cylinder volume followed by make. For older cars, J48 found that it was best to further split them into two categories according to their age. Following this line of reasoning, the next significant feature was make followed by either cylinder volume or year itself.

Table 8. *Random Forest Classification Results*

Type of Training	% Success With Car Model	% Success Without Car Model
80% Percentage Split	59	59
Cross-validation with 10 folds	62	70
Training Set	99	92

The Random Forest algorithm is very good at classifying the data based on the whole training set only however when the data is split between a training set and a testing set, the performance is comparable to that of J48.

4.4. Naïve Bayes

Naïve Bayes (NaiveBayes in Weka [8]) is one of the most useful machine learning technique [14]. There are two reasons for that. Firstly, it is very easy to implement in software and secondary the accuracy is usually as good as more complex algorithms. Two different experiments were conducted. In the first one, the original data was used, i.e., the attributes' values were not converted in any way. The same attributes were used as in decision trees. The results are shown in Table 9.

Table 9. *Naïve Bayes Classification Results on the Raw Data*

Type of Training	% Success With Car Model	% Success Without Car Model
80% Percentage Split	65	59
Cross-validation with 10 folds	60	59
Training Set	77	68

From Table 9, we can see that in all cases, it was better to use car model to make the prediction than not to use it but still the different in performance was not that marked. In the second experiment, all the numerical attributes such as cylinder volume and year were converted into nominal classes. The categorisation is shown in the Table 10 below.

Table 10. *Classification of Numerical Attributes into Nominal Classes*

Cylinder Volume (cc)		**Year of Manufacture**	
Less than 1000	Extra small	1995-1998	Very old
Between 1000 and 1200	Small	1999-2002	Old
Between 1201 and 1400	Medium	2003-2005	Not so old
Between 1401 and 1600	Large	2006-2007	Quite new
Above 1600	Extra Large	2008-2010	New

The results obtained from the second experiment were highly similar to those obtained in the first experiment. The detailed accuracy by class and the confusion matrix for the second experiment is shown below in Figure 3.

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.857     0.094      0.75      0.857     0.8        0.952    A
                 0.769     0.169      0.667     0.769     0.714      0.806    B
                 0.231     0.028      0.6       0.231     0.333      0.685    C
                 0.273     0.068      0.375     0.273     0.316      0.681    D
                 0.8       0.133      0.444     0.8       0.571      0.895    E
                 0         0          0         0         0          0.988    F
Weighted Avg.    0.612     0.104      0.582     0.612     0.575      0.827

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 18  3  0  0  0  0 |  a = A
  5 20  0  1  0  0 |  b = B
  0  6  3  2  2  0 |  c = C
  1  1  2  3  4  0 |  d = D
  0  0  0  2  8  0 |  e = E
  0  0  0  0  4  0 |  f = F
```

Figure 3. *Detailed Accuracy by Class and Confusion Matrix*

Although the overall accuracy was 61% on this training set, we can see from the True Positive (TP Rate) Rate that different classed had widely differing accuracy values. Class A, B and E has very high accuracy values while Class C, D and F had very low accuracy values. The confusion matrix very neatly and concisely shows where the misclassifications have occurred. For example, the matrix indicates that out of 21 instances belonging to class A, 18 have been classified correctly and the three misclassified instances have been placed in class B. Similarly, there are 4 instances belonging to class F but all of them have been incorrectly classified in class in E.

5. Evaluation and Conclusion

In this paper, four different machine learning techniques have been used to forecast the price of used cars in Mauritius. The mean error with linear regression was about Rs51, 000 while for kNN it was about Rs27, 000 for Nissan cars and about Rs45, 000 for Toyota cars. J48 and NaiveBayes accuracy dangled between 60-70% for different combinations of parameters. The main weakness of decision trees and naïve bayes is their inability to handle output classes with numeric values. Hence, the price attribute had to be classified into classes which contained a range of prices but this evidently introduced further grounds for inaccuracies. The main limitation of this study is the low number of records that have been used. As future work, we intend to collect more data and to use more advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices.

7. REFERENCES

[1] NATIONAL TRANSPORT AUTHORITY. 2014. Available from: http://nta.gov.mu/English/Statistics/Pages/Archive.aspx [Accessed 15 January 2014].

[2] MOTORS MEGA. 2014. Available from: http://motors.mega.mu/news/2013/12/17/auto-market-8-decrease-sales-new-cars/ [Accessed 17 January 2014].

[3] LISTIANI, M., 2009. *Support Vector Regression Analysis for Price Prediction in a Car Leasing Application*. Thesis (MSc). Hamburg University of Technology.

[4] RICHARDSON, M., 2009. *Determinants of Used Car Resale Value*. Thesis (BSc). The Colorado College.

[5] WU, J. D., HSU, C. C. AND CHEN, H. C., 2009. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. *Expert Systems with Applications. Vol. 36, Issue 4, pp. 7809-7817.*

[6] DU, J., XIE, L. AND SCHROEDER S., 2009. Practice Prize Paper - PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation and Genetic Algorithms to Used-Vehicle Distribution. *Marketing Science, Vol. 28, Issue 4, pp. 637-644.*

[7] GONGGI, S., 2011. New model for residual value prediction of used cars based on BP neural network and non-linear curve fit. *In: Proceedings of the 3rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.*

[8] LEXPRESS.MU ONLINE. 2014. Available from: http://www.lexpress.mu/ [Accessed 17 January 2014]

[9] LE DEFI MEDIA GROUP. 2014. Available from: http://www.defimedia.info/ [Accessed 17 January 2014]

[10] GELMAN, A. AND HILL, J., 2006. *Data Analysis Using Regression and Multilevel Hierarchical Models*. Cambridge University Press, New York, USA.

[11] WEKA 4: DATA MINING SOFTWARE IN JAVA. 2014. Available from: http://www.cs.waikato.ac.nz/ml/weka/index.html [Accessed 17 January 2014].
[12] LI, Y. H. AND JAIN, A. K., 1998. Classification of Text Documents. *The Computer Journal, Vol. 41, pp. 537-546.*
[13] QUINLAN, J. R., 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
[14] MITCHELL, T. M., 1997. *Machine Learning*. McGraw-Hill, Inc. New York, NY, USA.