BIDM Project

Predicting the contract type for IT/ITES outsourcing contracts

The authors believe that data modelling can be used to predict if an outsourcing contract will end up as a Fixed price contract or not.

Nandini Govindarajan
(61210556)

D Ajay Mohan Rao
(61210184)

Delfin de las Heras Rico
(61210214)

Harsh Pandey
(61210541)

Suraj Amonkar
(61210378)

Executive summary

As the focus of Information Technology outsourcing (ITO) and business process outsourcing (BPO) evolves from purely-cost-arbitrage and becomes a strategic business decision, the processes of vendor selection and contract negotiation is becoming increasingly complex, time consuming and expensive. Based on theories of contracts and procedural coordination[1] and the growth in the IT-BPO outsourcing industry[2] , we believe that the costs associated with contract negotiation will only increase in future. To offset these costs the practitioners will be well served with a model which helps predict the best suited type of contract – **Fixed price or not**, given certain deal-parameters.

This report outlines the design and working of one such model.

Our model uses "See Exhibit 3" as input variables (deal-parameters) – and predicts with 70.47% accuracy and 89.73% sensitivity the best suited type of contract for any given outsourcing deal.

Problem description

Per the NASSCOM numbers[3] the IT-BPO sector in India is estimated to aggregate revenues of USD 88.1 billion in FY2011. Software and services revenues (excluding Hardware), comprising over 86 per cent of the total industry revenues, and are expected to post USD 76.1 billion in FY2011; estimated growth of about 19.1 per cent over FY2010. This phenomenon of growth is not limited to India alone, but across the globe.

With the increasing value of each contract and increasing number of contracts the importance of choosing the right contract mechanism is also growing in importance. Anecdotal evidence indicates that both the client and vendor firms are spending increasingly high man-hours on pre-contract discussions, consultations, due-diligence, and decision-making. These spends exclude the monies spent on legal consultations.

The primary reason for such trends is that each contract is different – even the vendor firms, in many cases (and client firms in some cases), which have been through the process of contract negotiation multiple times lack the expertise and wherewithal to deal with the ambiguity and risk associated with contract terms and conditions.

A prediction model which articulates the 'when to –do what' scenarios using robust theory will be invaluable in such contract negotiations. Insights from such a model can provide contracting parties the de-facto guidelines for choosing the type of contract. This, naturally, will help the firms save considerable negotiation time and money.

---

[1] http://www.entrepreneurship.ethz.ch/education/lectures/Alliance_Advantage/Strategy_Nielsen_2010.pdf, http://oss.sagepub.com/content/19/4/585.full.pdf, http://www.sciencedirect.com/science/article/pii/S0378720606001340

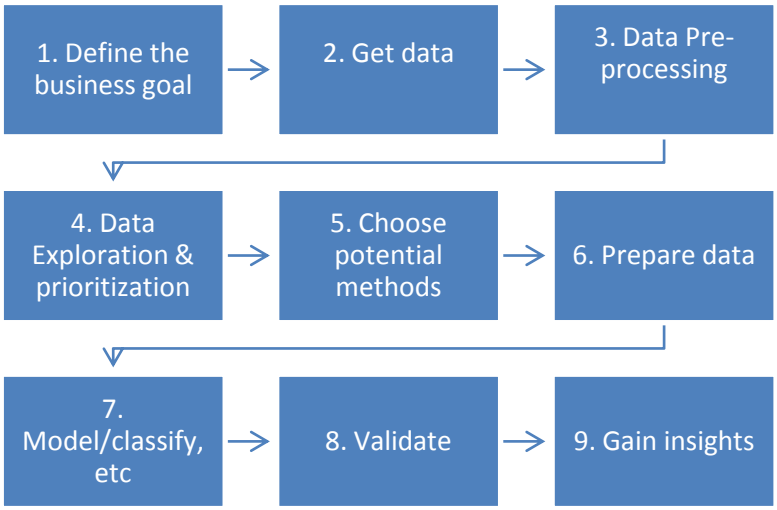
[2] http://www.nasscom.org/indian-itbpo-industry

[3] http://www.nasscom.org/indian-itbpo-industry

Methodology

Why predict "Pricing methodology"?

A fixed price vs. "Time and material"/others can make a sizeable difference in the margins of a vendor. Typically, in a fixed price deal, the number of people working on the project does not matter, as long as the project is delivered according to a certain schedule, whereas time and material projects require that each person working on a project to "log time". This opens up different options for a vendor to staff/ train resources and offers different arbitrage opportunities as well as risks. Typically, fixed price contracts tend to be on the riskier side as the vendor has signed up for a whole project, not just pieces of it. In a T&M project, there is a little more flexibility of adding more resources as the complexity changes. However, from a revenue perspective, an FP contract may be preferred by a vendor as it eases the billing cycle and involves lesser overhead. There are various other implications of a Fixed price vs. Other pricing methods and hence if a reasonable prediction model could be built, this can be extremely useful to plan prioritize which projects a vendor should bid for, and which are more likely to result in a sub-optimal margin.



1. Business goal

To reduce the ambiguity in, and time spent on contract negotiation while outsourcing the design/maintenance/management of IT related products and services. More specifically, the business goal is to predict whether a certain deal will be a fixed price deal or not.

2. Get data

The data set is an extract of 20,000+ records from The Services Contracts Database, or SCD, published by the market intelligence firm IDC. This is a database of Information Technology contracts that span different types of deals – Application development, maintenance, etc and the contract details.

Each record contains more than 40 fields. Data in the file can be viewed as belonging to three types:

a) Contract specific data – example: signing date, contract value, contract length, confidence factor
b) Customer details – example: name, description, revenue, parent company,
c) Vendor details – example: Name, role, relationship with the customer(existing, new, etc)

Another way of looking at the data would be the data available ex-ante and ex-post.

- Ex-ante: Contract value (assuming a certain degree of approximation), customer name, customer revenue, engagement type
- Ex-post: Start date, End date, Losing bidders, contract status, confidence factor (how accurate the contract value is)

3. Data pre-processing

a) Removal of missing data value rows – this was done especially for the main variables we were interested in exploring more
b) Using dummies for list attributes – this needs to be updated
c) Using transformations/binning:

   For numeric data – customer revenue, services contract value and service run-rate, we tried multiple transformations as they had a wide range of information and a certain pattern of clustering ***(See exhibit 1)***

   Log()

   Log() + Binning

   Results were roughly the same with these options as well as original data, hence eventually stuck to the original form so as to not lose any information.
d) Converting nominal to ordinal data: This was done for vendor role, existing relationship, customer industry, etc. This is done using the data-conversion toolkit. The list of all the variables converted to Ord is denoted by "_ord" ***(See Exhibit 2)***

4. Data exploration and prioritizing:

Multiple visualization techniques were used in the exploration of the data (see exhibit < >)

Findings from Spotfire scatter plots:

a) Price methodology has about 25% "unknowns" – this became important for us later, as we selected this variable for classification

b) Many of the data variables have a large % of unknowns (eg: winning factors) – ignore these columns for any analysis

c) Pricing methodology did not seem to vary based on just customer revenue (i.e., rich customers don't (on an average) go for an FP contract)

d) **Government sector:** The total number of contracts showed a steady increase from 2001 until 2009. The Government sector leads all other industries not just by number of contracts but also by a HUGE margin in terms of contract value. Also, we can see that though the number of contracts gradually increased, the value of Govt contracts have been quite high since 2005 and they also show a seasonality – the contract signing drops in Q4 of every year and picks up during the rest of the year.***(See exhibit 3).*** Also, typical contract length varies by sector, with the Govt. Sector being the longest. ***(See exhibit 4)***

The final data elements used for the analysis are:

- **To be classified**: Price Methodology: Fixed price (FP) or non-Fixed price. The original data set has many other options such as "Time & Material", "Combined". However, it was decided to classify Fixed price vs. All other categories, as (a) the percentage of deals in the database with Fixed price was significantly higher (> 60%) of all the deals. (b) All other methods have common characteristics that are different from fixed price contracts.
- **Variables used for classification:** The final predictors used are as shown in exhibit 2. This was a combination of ordinal and numeric data.

5. Choosing the methods:

Since the variable to be predicted is a categorical one (Fixed Price contract vs. Not), we tried all the methods below:

i) Naive Bayes
ii) K-nearest neighbours
iii) Classification tree
iv) Logistic regression
v) C 4.5
vi) SVM – support Vector Machines
vii) Random Forest (Ensemble method)

The tool we used for the same is 'Orange', an open source data mining software. (http://orange.biolab.si/). We chose this over XL Miner due to technical ease and better visualization techniques available in Orange.

A graphical representation **(schema)** of the methods used and the entire process flow is shown ***in exhibit 5***

6. Preparing the data:

Out of the 20,000+ records that were in the database, we completely removed the ones where the price methodology was unknown, resulting in about 16,000+ rows. The stratified partitioning was done as 70% for the training (cross-validation) and 30% for the blinded hold-out set with random sampling.

7. Running the methods (on Training data)

Each of the methods and the results are explained below. In summary, most of the methods gave consistent accuracy in the range of 65 – 70% but the sensitivity was better in some methods, as explained below. A comparison of the accuracy and other measures is given in ***exhibit 6.***

i) Data driven methods:

a. **Naive Bayes**: Using Naive Bayes and a cut-off value of 0.5, an accuracy of 65.7% was achieved and a sensitivity of 72.96 %.

See ***exhibit 7*** for the input, output, ROC, confusion matrix and lift charts

b. **K-nearest neighbours:** K-NN was used with Euclidean distances and the number of neighbours chosen was 5 with the option to normalize continuous data.
The training data resulted in 67% accuracy with 79% sensitivity.

See ***exhibit 8*** for the input, output, ROC, confusion matrix and lift charts

ii) Logistic regression:

Logistic regression showed a better sensitivity (91.89%). This means that the model is better at predicting Fixed price than at predicting "Not Fixed Price" and would be more useful for someone who has a greater interest in predicting Fixed price contracts correctly.

See ***exhibit 9*** for the input, output, ROC, confusion matrix and lift charts

iii) Trees

a. Classification trees:

The most important predictors were found to be "Services Contract value", "market" and "signing region". The accuracy was around 68% here with a sensitivity of around 79%.

See ***exhibit 10*** for the input, output, ROC, confusion matrix and lift charts

b. C 4.5 (uses entropy)

C4.5 is a method developed from the ID3 method that uses a different pruning mechanism and uses slightly different rules for the tree-building process. This method uses information gain for splitting the data at every node.

See ***exhibit 11*** for the input, output, ROC, confusion matrix and lift charts

iv) Ensemble methods

Random Forest: The random forest method, which is an ensemble method, shows the highest accuracy of all, at 70% and quite a high sensitivity factor of 89%. The high accuracy is expected as this is an ensemble method

See ***exhibit 12*** for the input, output, ROC, confusion matrix and lift charts

v) SVM – Support Vector Machine (* Limitation : May not work very well with lot of data)

This method is a kernel based method that transforms the data to a different dimensional space where it tries to find the optimal partitioning (hyper-plane) for the data. It is a computationally heavy method and runs slowly on the orange platform.

See ***exhibit 13*** for the input, output, ROC, confusion matrix and lift charts

Note that we also tried using k-means clustering and hierarchical clustering to evaluate the interaction within the input-features. This was done to understand any trends in the input-data. We did not observe anything interesting from these unsupervised methods.

8. Cross-validation

When the same methods are run on the validation data, very similar accuracy/sensitivity measures and confusion matrices are seen.

We used cross-validation to evaluate the performance of the classifiers as we believe this method is more robust. The process involves partitioning the data into n equal buckets and using one bucket for training and the other (n-1) buckets for testing in each of the n-folds (iterations) of the evaluation. Since every data-row is used in the training and the testing sets this evaluation is less prone to biases in the training data.

9. Performance on hold-out (test data)

We used the test data to predict the performance of this model. The classification accuracy was as follows:

Random forest	Naïve Bayes	Logistic regression	KNN	C4.5	Classification Tree
0.69458	0.643439	0.672779	0.675632	0.692747	0.667278

A screenshot of the predictions for the test (hold-out data) are shown ***in exhibit 14***

Overall Conclusion

The comparative performance of the classification methods with the degree of accuracy that was achieved (approx. 65-70%) is better than a random coin toss method, and we believe that this represents a fair way to pre-assess the pricing methodology of an outsourcing deal. In our model, we assume that there is a fair idea of the contract value, although the exact deal value is finally determined via deal negotiations. It also assumes that the vendor knows if this is a competitive bid or not (via "Bid Type"). These, based on our experience is the IT industry, are mostly true.

Applications of this model include:

- **Pre-assess a potential deal and decide whether to bid for it**. There is a substantial amount of effort that goes into deal making, especially when the market is looking up, and the number of potential deals to finalize are many. The "fixed price vs. other" decision is often one of the primary debated parts of a deal negotiation that can make a difference in the margin of a vendor, and thus useful in making strategic choices. A 70% chance of correctly predicting the method could save several man-months in terms of pre-sales time (Needs assessment, due diligence, proposal preparation) that can utilized in the deals that are likely to result in better ROI.
- **Consultants** can use this to advise their clients on the right approach for outsourcing deals and how to derive the best value from the same.
- **Applications** (after refining the model) in other industries for a vendor/ supplier contract
- **Gain competitive advantage** Since this model takes into account the "competitor" factor (whether or not the bid was open to others), it could provide insights (combined with other factors that were missing in the data (see below))

Potential for refinement: The model could be refined if further data was available in terms of:

- **Existing relationship:** Intuitively, this would have been a major factor in predicting the type of pricing method, but the dataset that we had contained a lot of "unknowns" (14,000+ rows out of 20,000) that rendered this field useless.
- **Winning factors** (If prior data was available on when an outsourcing was successful, what were the success factors). This was mostly "unknown" in the database

Other predictions that could be tried with this dataset:

- **Contract value** based on pricing method, contract length, industry, etc
- **Contract length** based on existing relationship, pricing method, industry, etc

Overall, this report presents an opportunity to predict the methodology to be adopted by a vendor in an outsourcing deal and this has a lot more potential to be tapped.

Exhibits

Exhibit 1 – service contract vs. Customer revenue

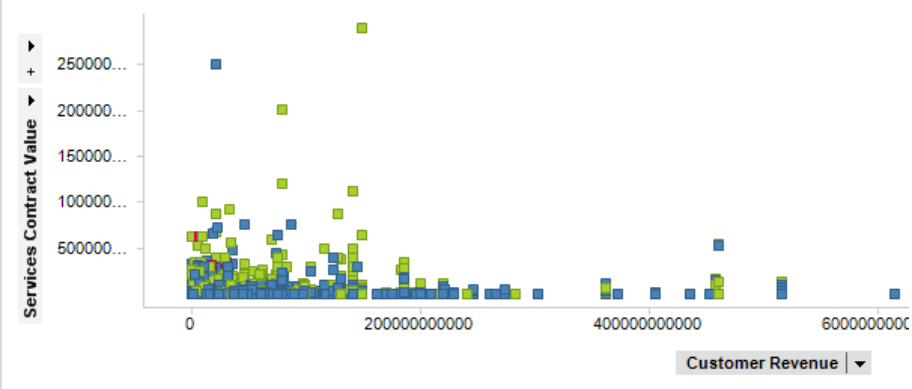


Exhibit 2 – final predictors used

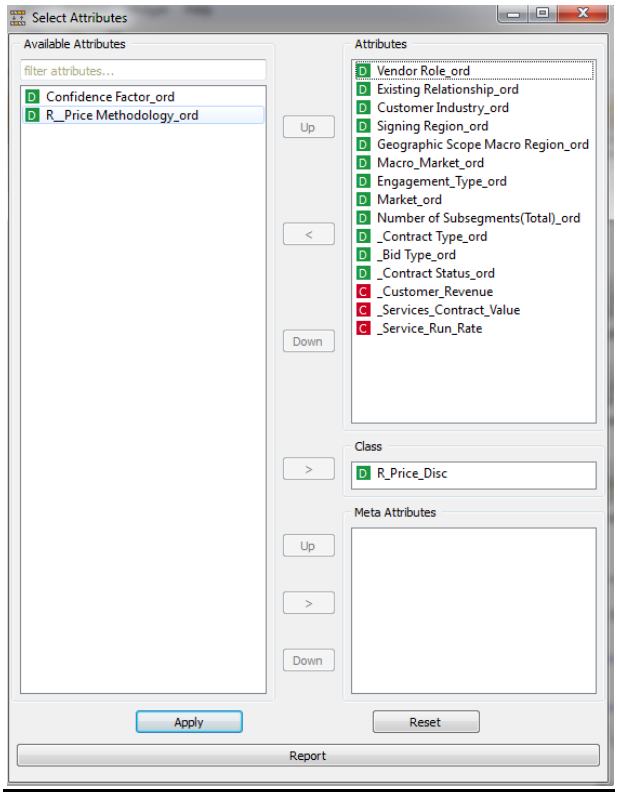
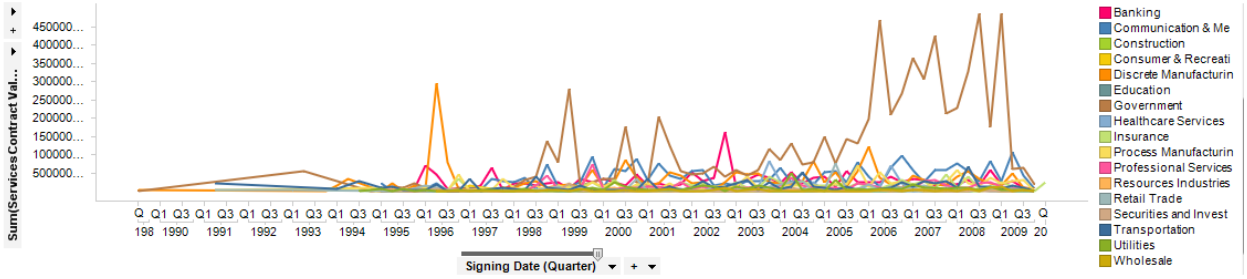


Exhibit 3 – Service contract value and counts over time

Line Chart



Line Chart

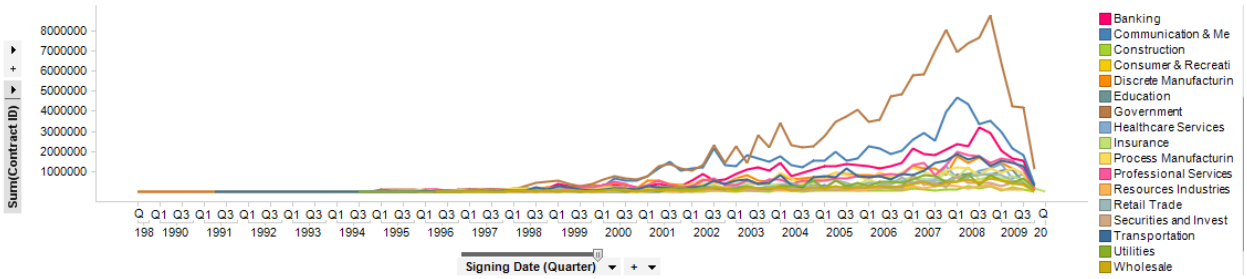


Exhibit 4 – Cust industry vs. Contract length

Scatter Plot

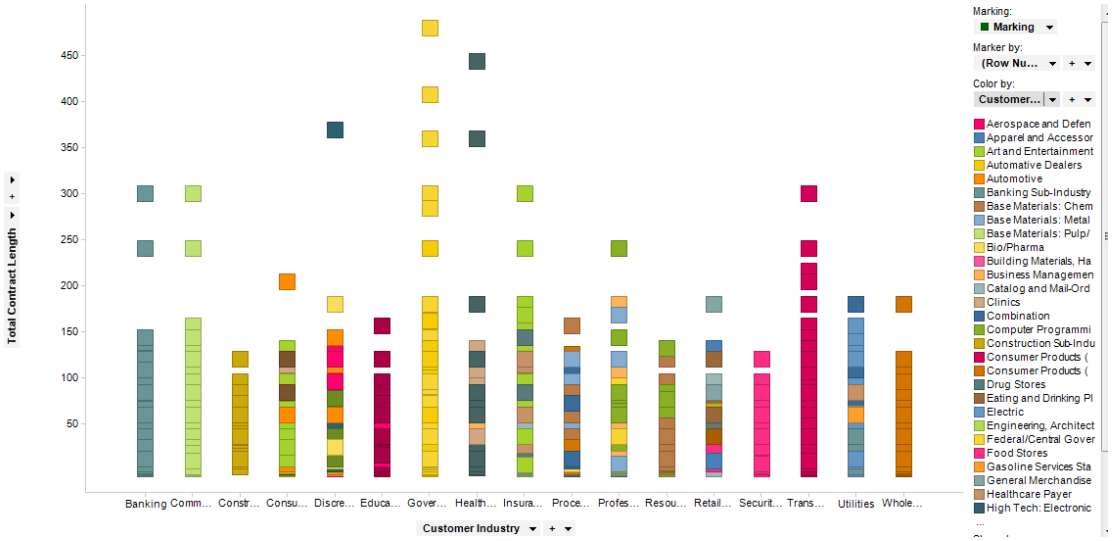


Exhibit 5 – Schema / process flow

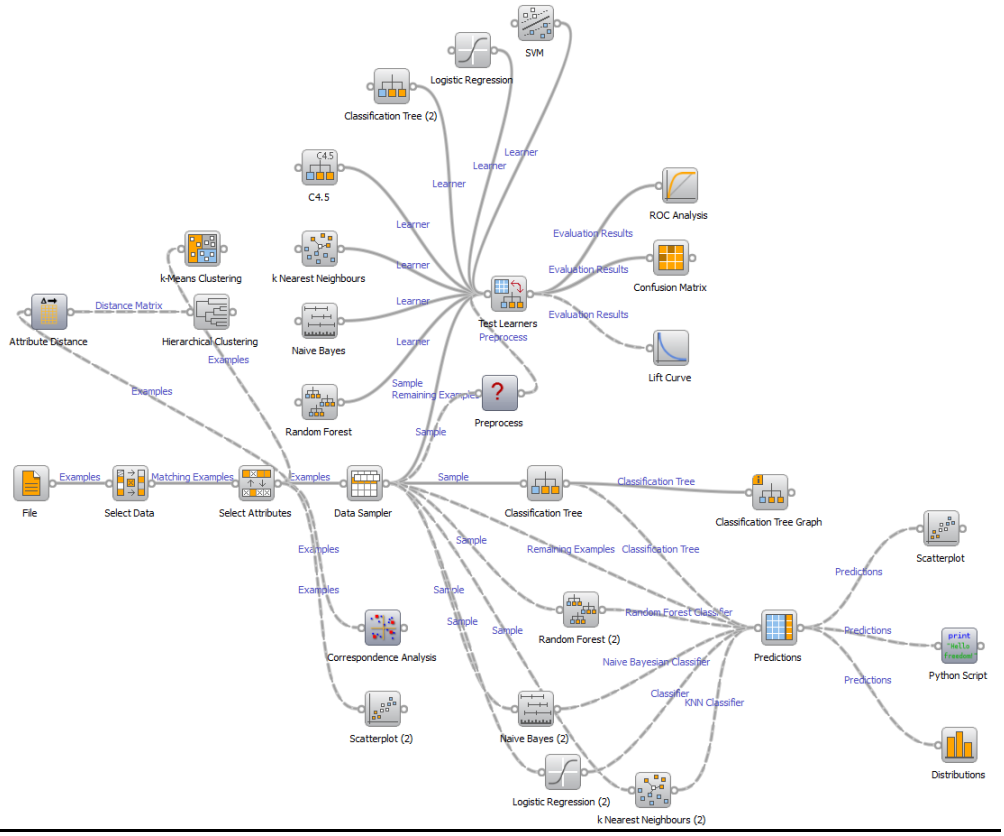
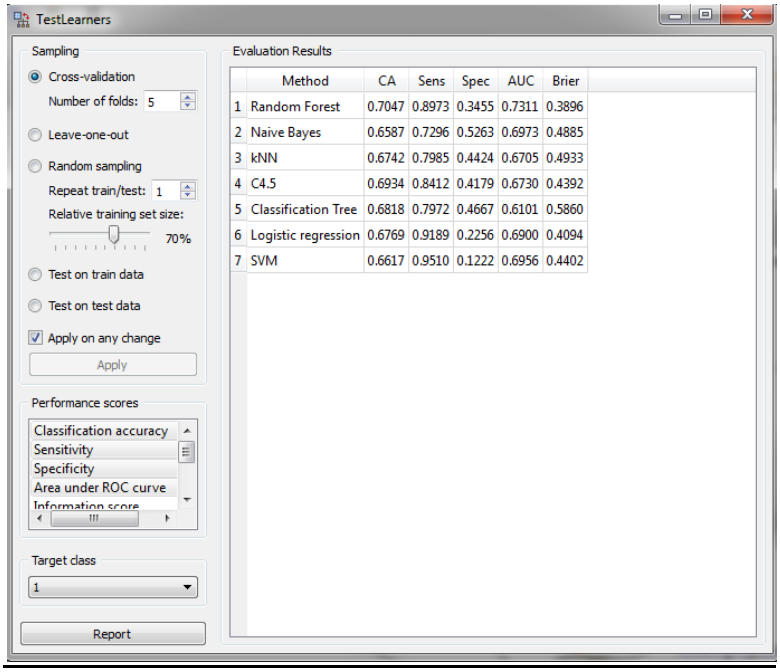
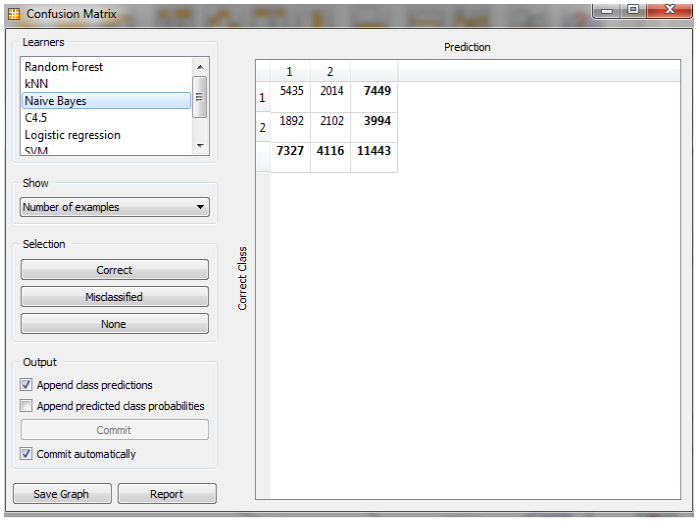


Exhibit 6 – Comparison of predictor efficacy across methods



	Method	CA	Sens	Spec	AUC	Brier
1	Random Forest	0.7047	0.8973	0.3455	0.7311	0.3896
2	Naive Bayes	0.6587	0.7296	0.5263	0.6973	0.4885
3	kNN	0.6742	0.7985	0.4424	0.6705	0.4933
4	C4.5	0.6934	0.8412	0.4179	0.6730	0.4392
5	Classification Tree	0.6818	0.7972	0.4667	0.6101	0.5860
6	Logistic regression	0.6769	0.9189	0.2256	0.6900	0.4094
7	SVM	0.6617	0.9510	0.1222	0.6956	0.4402

Exhibit 7: Naive Bayes



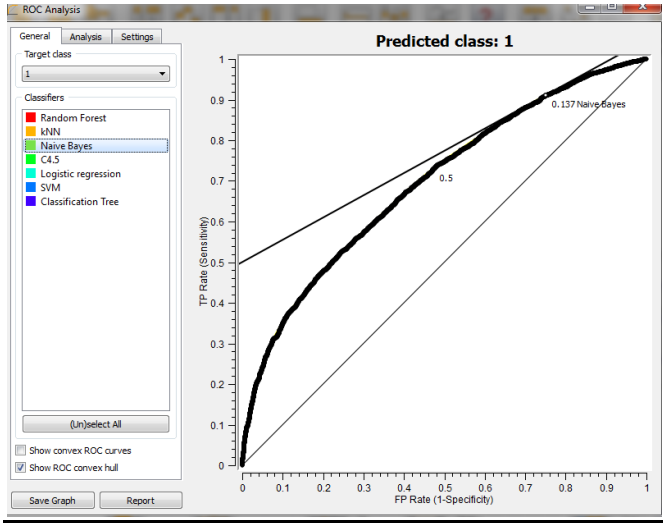
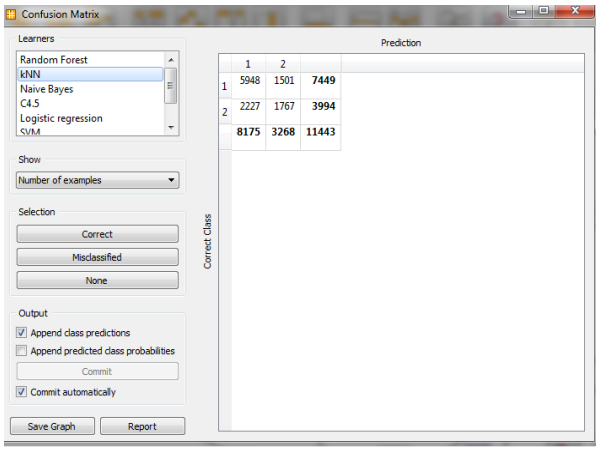


Exhibit 8: K Nearest Neighbours



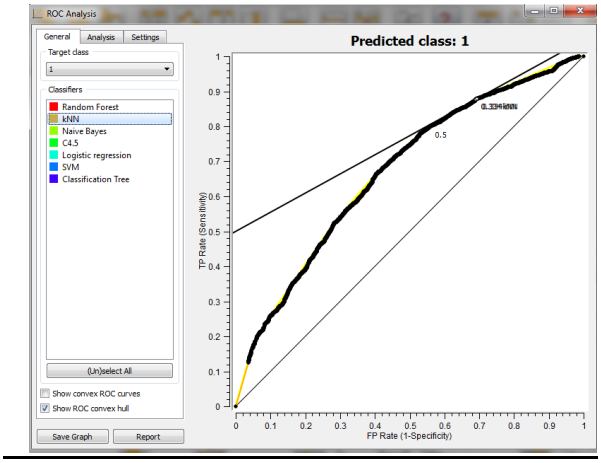
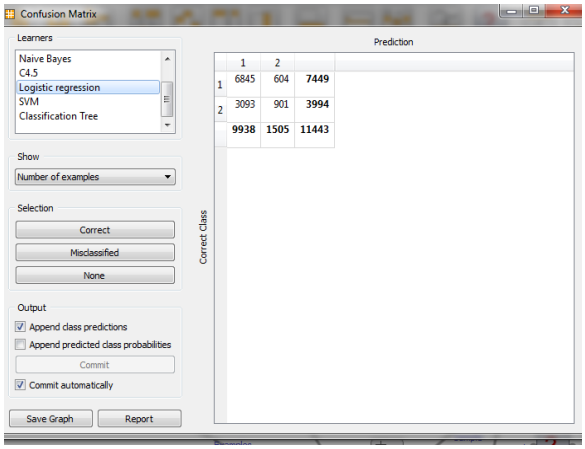


Exhibit 9: Logistic Regression



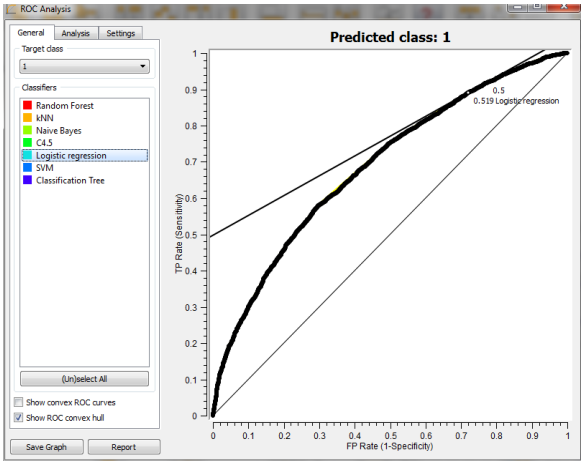
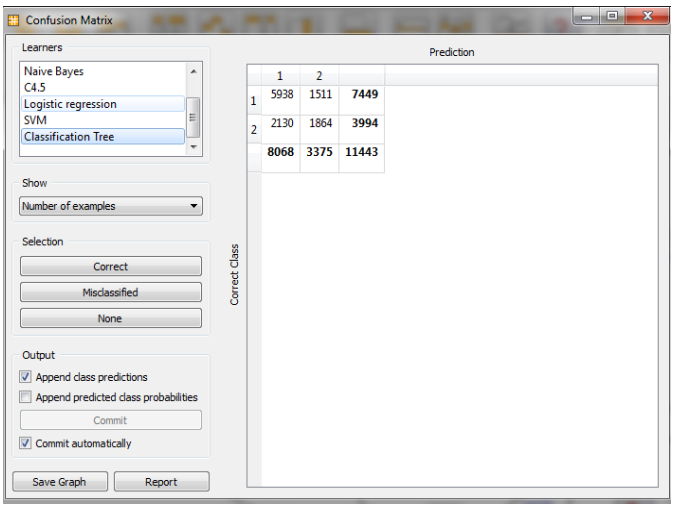


Exhibit 10: Classification Trees



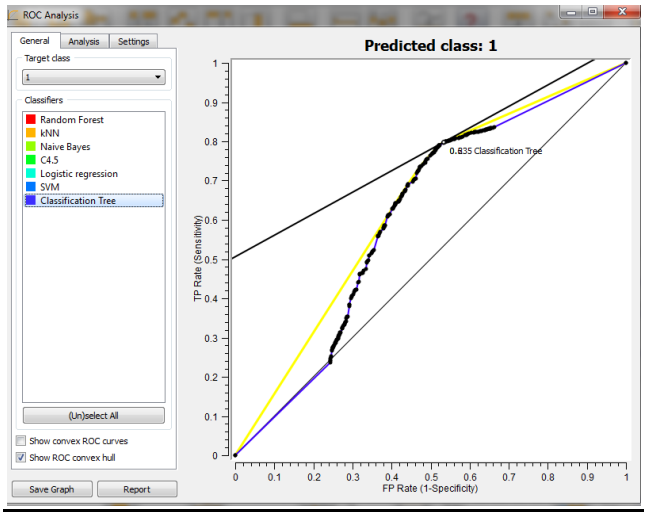
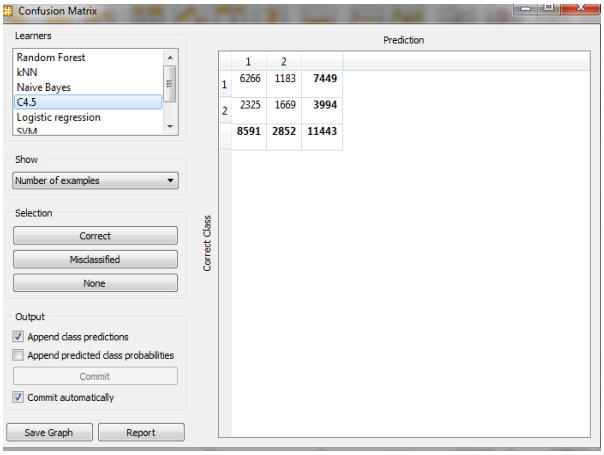


Exhibit 11: C 4.5



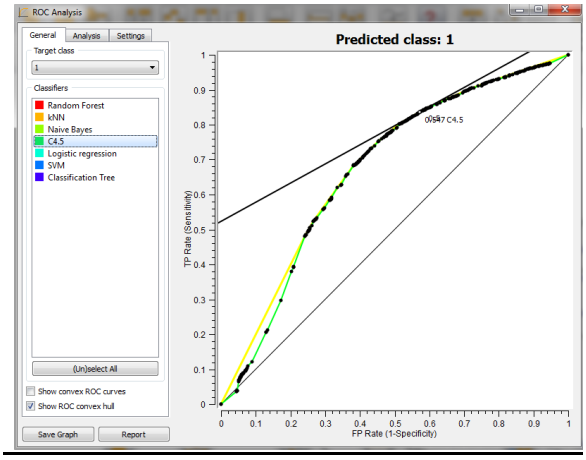
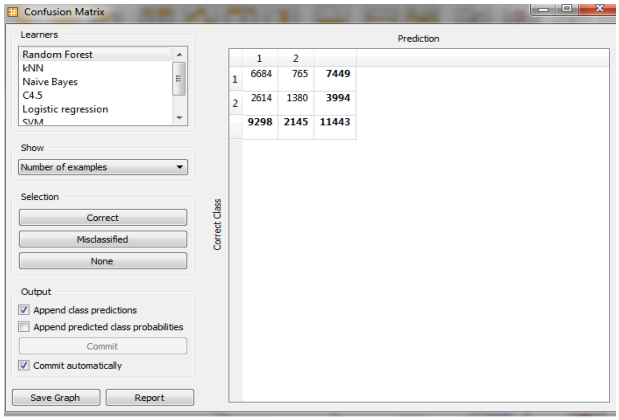


Exhibit 12: Random Forest



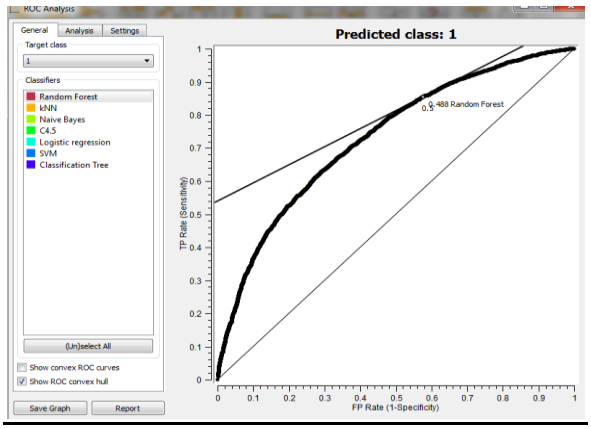


Exhibit 13: SVM (Support Vector Method)

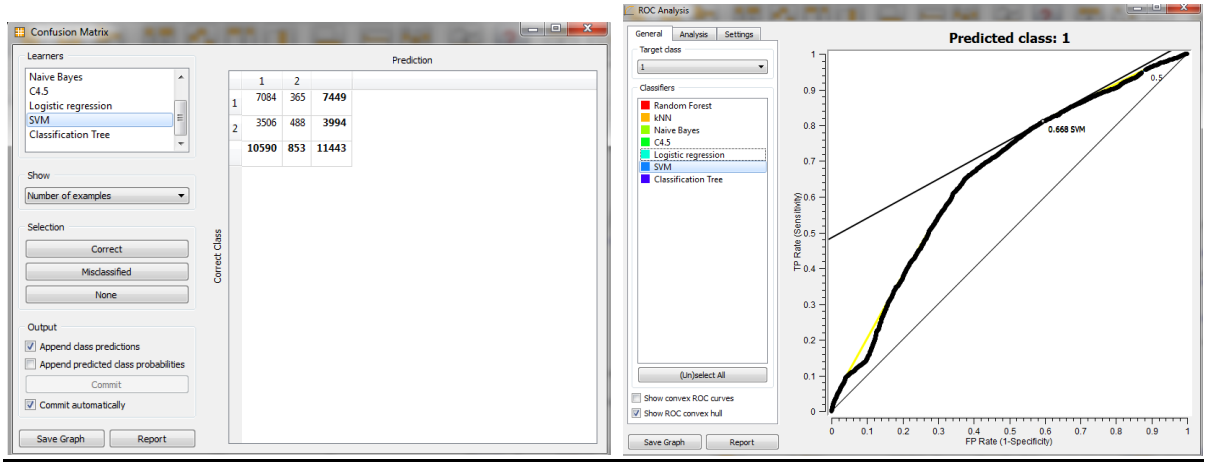


Exhibit 14: Test results