2015

CRIM Analytics Competition Case Study

ARUSHI CHAWLA: 660164856
ANSHUL SHRESTHA: 665457383

UNIVERSITY OF ILLINOIS | Chicago

Table of Contents

Overview of the Solution:

The problem statement is to predict the customers based on the historical data who are most likely to take loan from the company, we used Rapid Miner to first explore the historical data and then do the further analysis. The exploratory steps include having a look at the variables and filter them based on their need and importance. Few variables have missing values which needs to be taken care of with some strategy and few need some data transformation. Few data descriptive models like Decision tree helps us in determining this. Once we are decided with our variables we then we apply various classification and predictive models like Logistic regression, Naïve Bayes, K-NN and J-48 decision trees after splitting our data into training and validation dataset.

After comparing the outcome from all these based on the threshold value, split ratio, confidence threshold, minimum size for split, minimum leaf size, maximum depth and pruning we decided our final model for the predictive modelling which is giving us the best result.

The same model is then applied to the scoring data and came up with the predicted values for the respective customers of the company.

Data Preparation:

The first step in preparing our data is to decide which variables are to be considered in the analysis process. We checked for the missing values in the data which were luckily not present. In all this process we also used Decision Tree as they are very strong classification model and have powerful approach in knowledge discovery and data mining to see which variables are most important by looking at the level of that variable. We are not considering Profit in training the data as it is not important for any prediction. Further the decision tree also showed it as a least important variable.

While using Logistic Regression we created the dummy variables for categorical variables like education, housing and job. Also we transformed yes to 1 and no to 0 for variables like loan, default and housing for its binary dummy conversions as Logistic Regression only works on the numeric data. But while using Weka J-48 Decision Tree we used the categorical data as J-48 works with categorical data. Also we discretized the variable age and created three categories as youth, adult and child to make good use of that variable which otherwise did not have much impact.

Model Building techniques and strategy:

Once we are decided with the variables in our model, we applied many predictive modeling models and compared the result after changing various parameters specific to models and also by changing the split ratio between testing and validation data. We finally set the split ratio to be 70:30 and threshold value to be 0.5. Also we got the best result with Weka J-48 which gave us accuracy percentage of 91.22% and validation accuracy percentage as 89.91%. Furthermore the class recall i.e. the true predicted **yes** are also 53% for the training data and 50% for the validation data which is quite good as the model is hence able to predict a little more than half of the true prospects as the predicted prospects. J-48 is a strong classification and predictive model and therefore gave us good result for the supervised classification.

Moreover the model also give us a good accuracy percentage of 96.12% for training and 95.25% for validation data in predicting the true false. However, for the current project our focus is predicting the true **yes** along with the overall accuracy of the model as that tells us the customers whom the company should contact to get the maximum benefit from the loan those customer take from the company.

Applying this model to the new Scoring dataset predicted 464 customers out of 4523 for being considered as the prospects which give me a response rate of 10.28% which is expected as the historical dataset also had a response rate of 11.66%. Considering the percentage accuracy of the model, the true positive accuracy percentage and the response rate we can assume of achieving a good model which scored and predicted the data quite well.

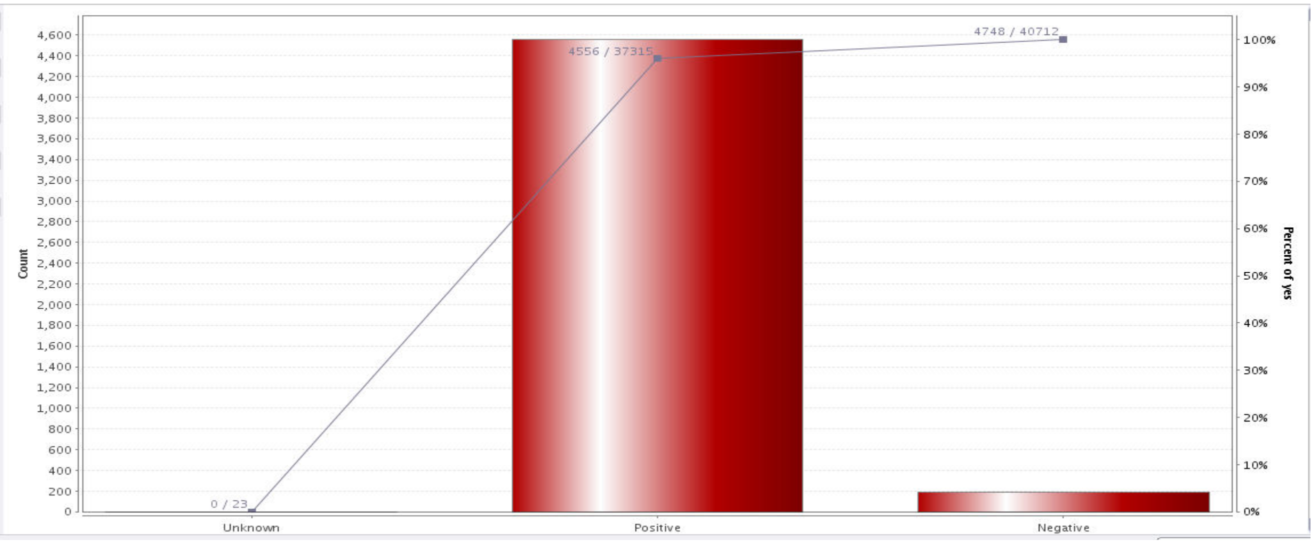
Click for confusion matrix

Interesting/unexpected finding:

We bucketed the **balance** attribute into following 3 categories:

1. Positive Balance
2. Negative Balance
3. Unknown : Garbage value 999999

The graph below shows the cumulative count of **yes** (the customers who were retained). The total no. of customers who were retained in the historical data was 4748 and out of these more than 95% (4556) of them had a positive balance.



Job: 28.4% of the customers who were student were retained by these campaigns. It is the highest retention rate for any of the job category. This made us look the next variable which was education. *Click for details*
Education: The no. of customers contacted having tertiary education, 15% of them were retained. *Click for details*
Marital: The customers who were single had the highest retention rate when contacted as compared to other categories of marital.
Also, 35.67% of the students without housing were retained by the campaigns.
Out of the 830 students contacted 236 were retained and all of them had a positive balance.
Click for details
From the above discussed attributes it can inferred that students who opt for tertiary education and do not have housing and have positive balance take the loans

Conclusion/business recommendations:

Applying this model to the new Scoring dataset predicted 464 customers out of 4523 for being considered as the prospects which give me a response rate of 10.28% which is expected as the historical dataset also had a response rate of 11.66%. Considering the percentage accuracy of the model, the true positive accuracy percentage and the response rate we can assume of achieving a good model which scored and predicted the data quite well.

Hence we can conclude that the company should go ahead and contact these 464 prospects as there is a high chance of them being getting converted into customers. Therefore there is no risk in spending 10 Euros on these prospects as there is a high chance of being benefitted by the loan amount from those prospects. Also they should target students who will be pursuing any bachelors or master degree, as according to our analysis their chance of converting from prospect to customers is the highest.

Considering the goal of the project is to get the highest profit from the loans minus the costs of calling customers, we can assume that profit from loan to be X. Then the cost of contacting these customers is 10 Euros per head which totaled to 4640 Euros. So the company will have a good profit of (464X-4640) Euros where X is the total loan amount and will definitely be greater than 10.

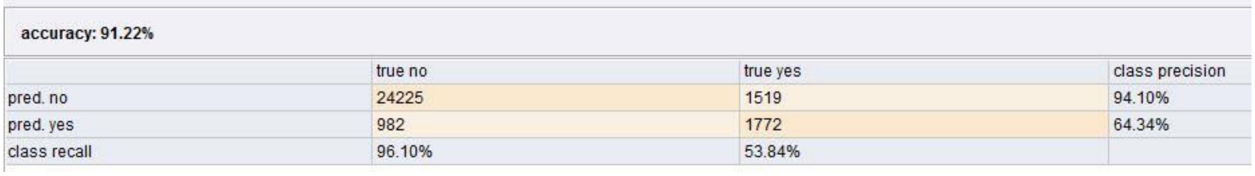
Areas of future research:

In future we can also calculate the proper cost matrix along with the lift analysis of the profit. We can see the lift which shows me the improvement of using a model when compared to no model and compare the profit company will have. With the help of the cost and lift curves we can getting a better idea of setting the threshold of the confidence(1) which in turn will help us in predicting the customers better.

Technical Appendix

Confusion Matrix

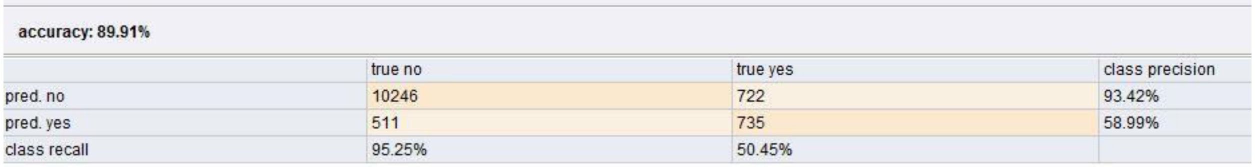
Training dataset (70% of historical data)



accuracy: 91.22%

	true no	true yes	class precision
pred. no	24225	1519	94.10%
pred. yes	982	1772	64.34%
class recall	96.10%	53.84%	

Validation dataset (30% of historical data)

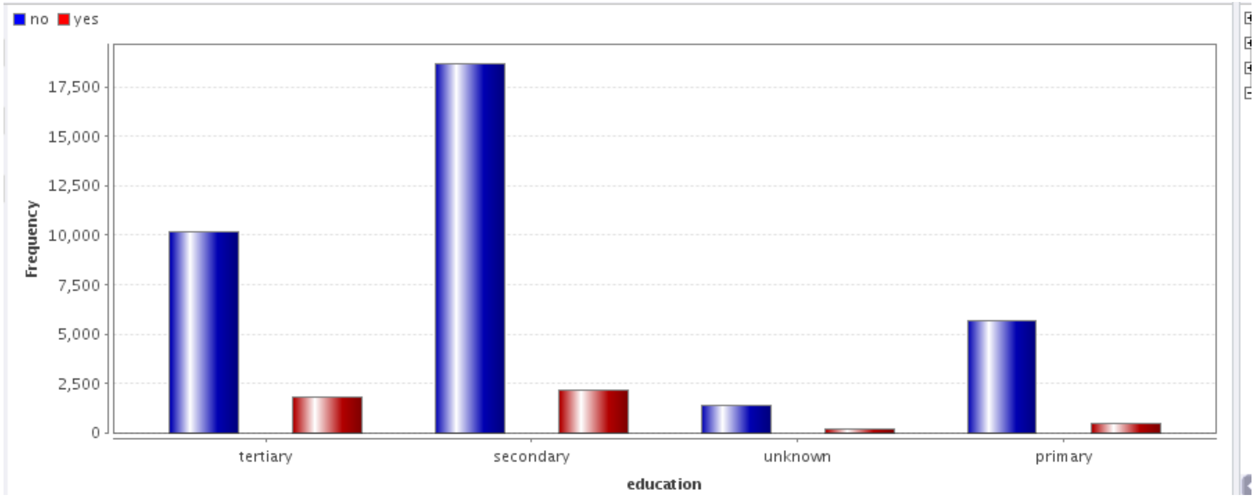


accuracy: 89.91%

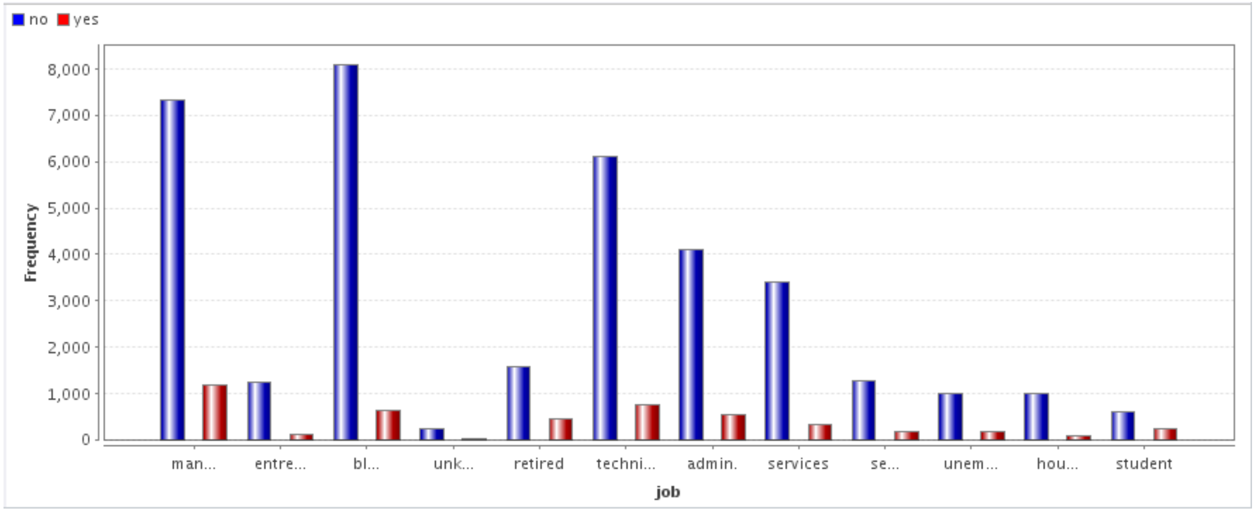
	true no	true yes	class precision
pred. no	10246	722	93.42%
pred. yes	511	735	58.99%
class recall	95.25%	50.45%	

Back to model building techniques and strategy

Education

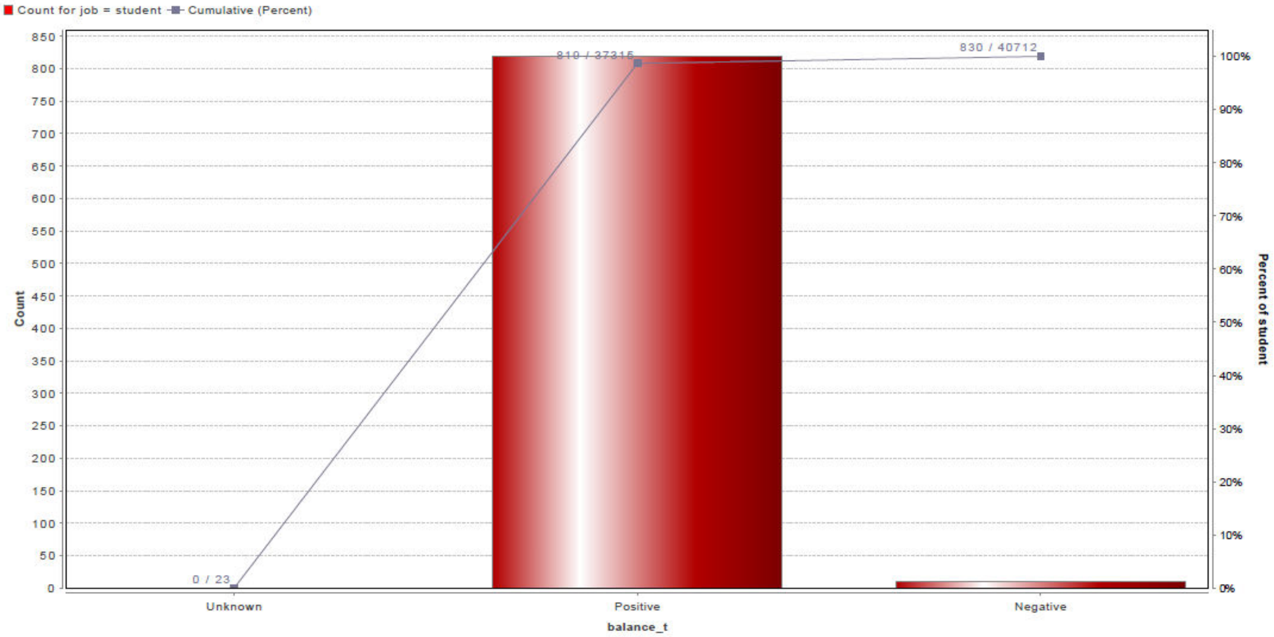


Jobs



Back to interesting/unexpected findings

Students vs balance(positive or negative)



Back to interesting/unexpected findings

Go to beginning of the report