DECISION SCIENCES INSTITUTE

Using Predictive Analytics to Gain Insights of Bank Direct Marketing Data

Manying Qiu
Virginia State University
Email:mqiu@vsu.edu

Shuting Xu
Georgia Gwinnett College
Email:sxu@ggc.edu

Yaquan Xu
Georgia Gwinnett College
Email:yxu@ggc.edu

ABSTRACT

The rapid advancement of data analytics technologies make it easier for business practitioners to apply data mining techniques. The visualization capability of data analytics tools is very helpful for businesses to understand the big data and evaluate models perforce. This paper analyzes the 2012 and 2014 bank direct marketing datasets to illustrate how the predictive analytics tool can efficiently and effectively provide more, better business insights.

KEYWORDS: Data analytics, Data mining, Predictive model, Visualization, Business insights

INTRODUCTION

This paper uses a predictive analytics tool to gain insights of two sets of data related bank direct marketing. Moro et al (2011, 2014) collected the first dataset (labeled as “2012 bank direct marketing”) from a Portuguese bank’s 17 term deposit campaigns from May 2008 to November 2010; the second dataset (labeled as “2014 bank direct marketing) was collected from 2008 to 2013. The 2012 dataset was mainly focused on the client information, while the 2014 dataset included some economic and financial factors because the authors felt the financial crisis during the data collection period affected client decisions. Moro et al. (2011, 2014) used data mining techniques to build models that can explain the success of a telemarketing call i.e., the contacted client subscribed the deposit.

For decades, academic and business researchers have been interested in business intelligence (BI) and data mining (DM) to analyze data, make better decisions and improve business performance. The foundation of predictive analytics is statistics and data mining. Due to the success achieved in businesses and advance of technology, predictive analytics continues to be an exciting area of research (Chen et al., 2012).

Lund et al. (2013) addressed the potential of big data analytics to raise productivity is one of the five opportunities for US economic growth. Advanced predictive analytics capabilities may provide a critical solution to help companies meet productivity challenge, improve decision making and gain valuable insights to market share and reduce costs. Shmueli and Koppius (2011) stated that predictive analytics not only assisted in creating practically useful models, but also played an important role alongside explanatory modeling in theory building and theory testing. U.S. Direct Marketing Association (DMA) reported that business linked to marketing activities generated revenues grew 35% from $156 billion in 2012 to $202 billion in 2014 and created 650K U.S. jobs (Urbanski, 2016).

According to Experian Data Quality, a global consulting firm, the challenges with respect to big data and data analytics are: gain insight quickly; get enough data; and maintain accurate data (Experian, white paper). Current business analytics tools can automatically select maximum contributory variables of predictive models, apply data mining algorithms for modeling and visualize the characteristics of the data and evaluate performance of the models generated. The predictive analytics tools can quickly identify the contributions of the variables in the dataset and the significance of each category. This information has profound impact on business strategies and actions. Next we will compare the analytical results of the 2012 and 2014 bank direct marketing datasets to show the importance of using sufficient data to build predictive models.

RESEARCH METHODOLOGY

We use SAP Business Objects Predictive Analytics to analyze the two sets of the bank direct marketing data (Moro, 2014). We want to investigate if this predictive analytics tool is capable of gaining insights of the data efficiently and revealing more valuable, actionable clues. First, we use the predictive analytics tool to automatically select the most relevant variables and create a logistic recreation model. Then we apply two data mining classification algorithms, namely data tree (R-CNR Tree) and neural network (R-NNET Neural network). The purpose of this research is to figure out which features or explanatory variables will provide maximum insights of the marketing campaign. Therefore, in this case, interpretation of the campaign performance is of the most interest, the accuracy of the predictive model is of the secondary interest.

RESEARCH RESULTS AND DISCUSSION

The SAP analytics tool generate the following overview of the two sets of bank campaign data:

Table 1: Overview of 2012 and 2014 bank campaign data

Features	2012 data	2014 data
number of explanatory variables	16	20
Number of records	45,211	41,188
Target key	Yes	Yes
No-frequency	88.27%	88.7%
Yes – frequency	11.73%	11.3%

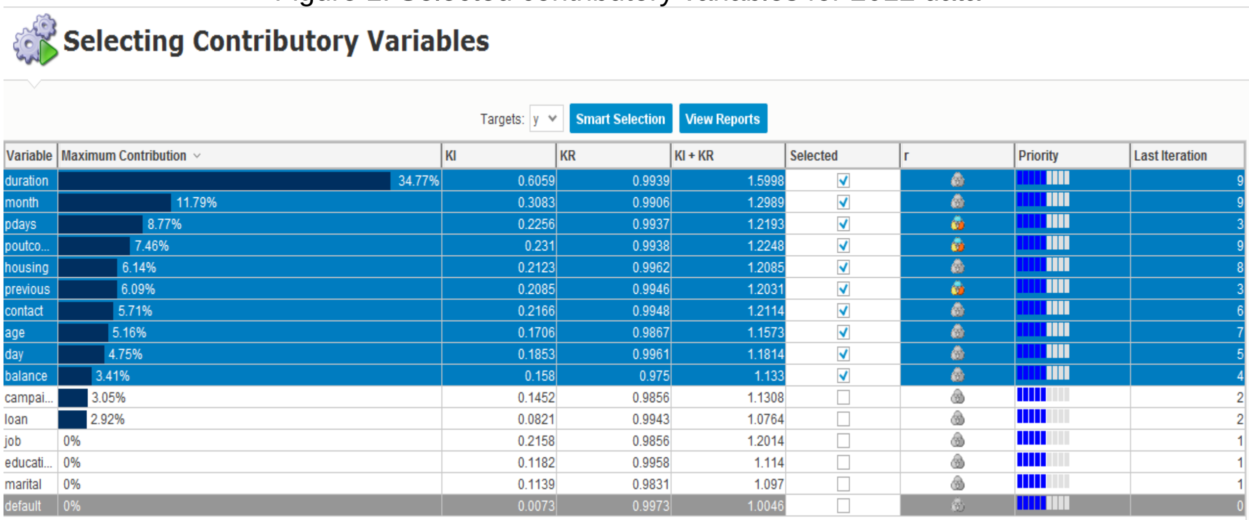
Comparing these two datasets, the 2012 dataset has about 4,000 more records, while the 2014 dataset has 4 more variables. The business goal of the direct marketing campaign is a client responded “Yes” and subscribed the term deposit after receiving the phone call. The success rate in 2012 dataset is 0.4% higher than that in 2012. Should the bank use more or less data variables?

Models for 2012 Data

<u>Statistical Logistic Model</u>

We use SAP predictive analytics tool to automatically create statistical logistic model for the 2012 bank direct marketing dataset. The selected most relevant contributory variables are the same as Moro (2011) results.

Figure 1: Selected contributory variables for 2012 data


Selecting Contributory Variables

Targets: y Smart Selection View Reports

Variable	Maximum Contribution	KI	KR	KI + KR	Selected	r	Priority	Last Iteration
duration	34.77%	0.6059	0.9939	1.5998	☑			9
month	11.79%	0.3083	0.9906	1.2989	☑			9
pdays	8.77%	0.2256	0.9937	1.2193	☑			3
poutco...	7.46%	0.231	0.9938	1.2248	☑			9
housing	6.14%	0.2123	0.9962	1.2085	☑			8
previous	6.09%	0.2085	0.9946	1.2031	☑			3
contact	5.71%	0.2166	0.9948	1.2114	☑			6
age	5.16%	0.1706	0.9867	1.1573	☑			7
day	4.75%	0.1853	0.9961	1.1814	☑			5
balance	3.41%	0.158	0.975	1.133	☑			4
campai...	3.05%	0.1452	0.9856	1.1308	☐			2
loan	2.92%	0.0821	0.9943	1.0764	☐			2
job	0%	0.2158	0.9856	1.2014	☐			1
educati...	0%	0.1182	0.9958	1.114	☐			1
marital	0%	0.1139	0.9831	1.097	☐			1
default	0%	0.0073	0.9973	1.0046	☐			0

Independent or Explanatory variables in the order of their contribution to the prediction are listed below:

1. duration: last contact duration, (in seconds)
2. month: last contact month of year
3. pdays: number of days that passed by after the client was last contacted from a previous
4. poutcome: outcome of the previous marketing campaign
5. housing: has housing loan or not
6. previous: number of contacts performed before this campaign and for this client
7. contact: contact communication type
8. age: age of the customer
9. day: last contact day of the month
10. balance: average yearly balance, in euros

The SAP predictive analytics tool also evaluates the significance of each explanatory variable. For example, as shown in Figure 2, the most effective telemarketing contact duration is around 12 to 19 minutes. If the contact duration is 1 minute or less, the effectiveness is below average. Figure 3 shows that the most effective months contacting clients are March, October and September. If clients are contacted in May, the success rate is below average.

Figure 2: Variable value analysis for duration

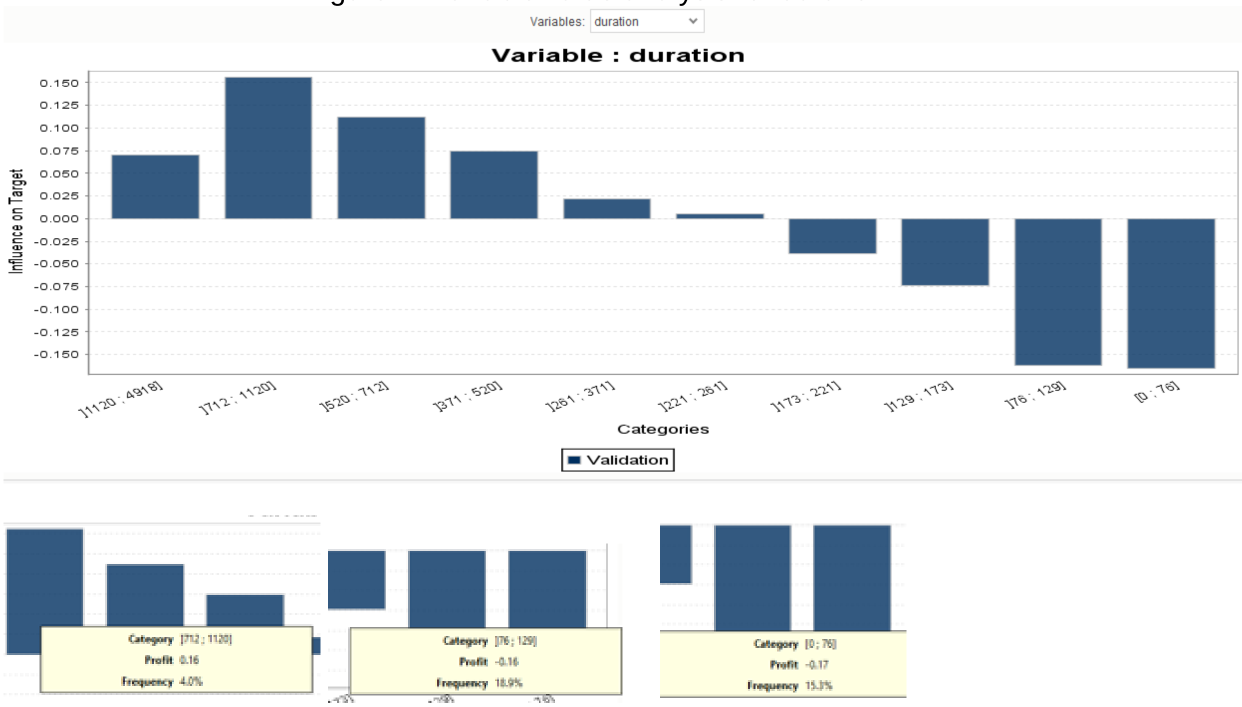
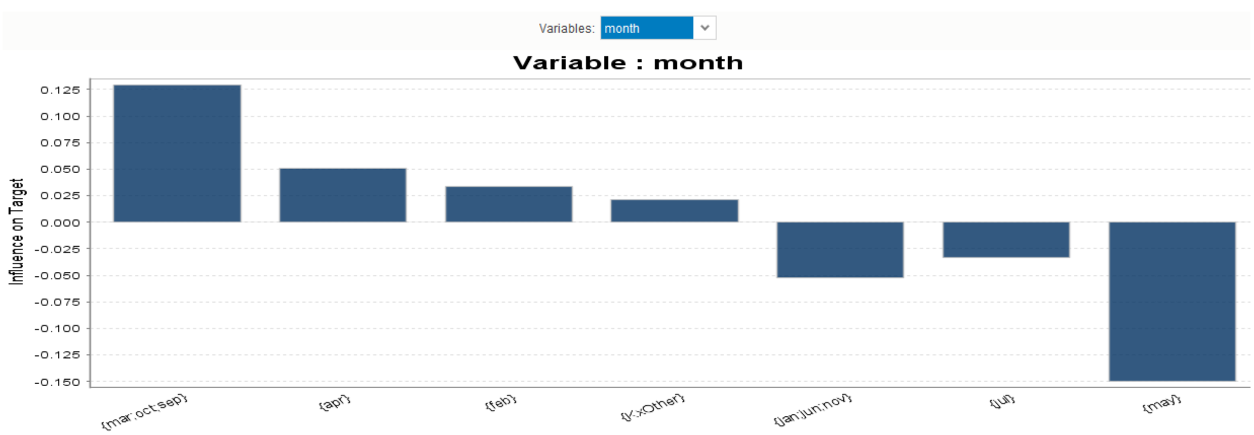


Figure 3: Variable value analysis for month



We also use the SAP tool to evaluate the model performance. The business performance metrics are shown in two formats: curve chart and column chart (Figures 4 and 5). The charts indicate the prediction power of the generated model, compared to the random model and the perfect model. If we select 15% of the population, then the hypothetical perfect model will identify 100% of the target (positive response), while the generated predictive model may identify 65.3% of the targets. When 35% of the population are selected, the predictive model may identify over 90% of the targets. The model performance in column chart is easier to read than the traditional Curve chart.

Figure 4: Model performance in curve chart

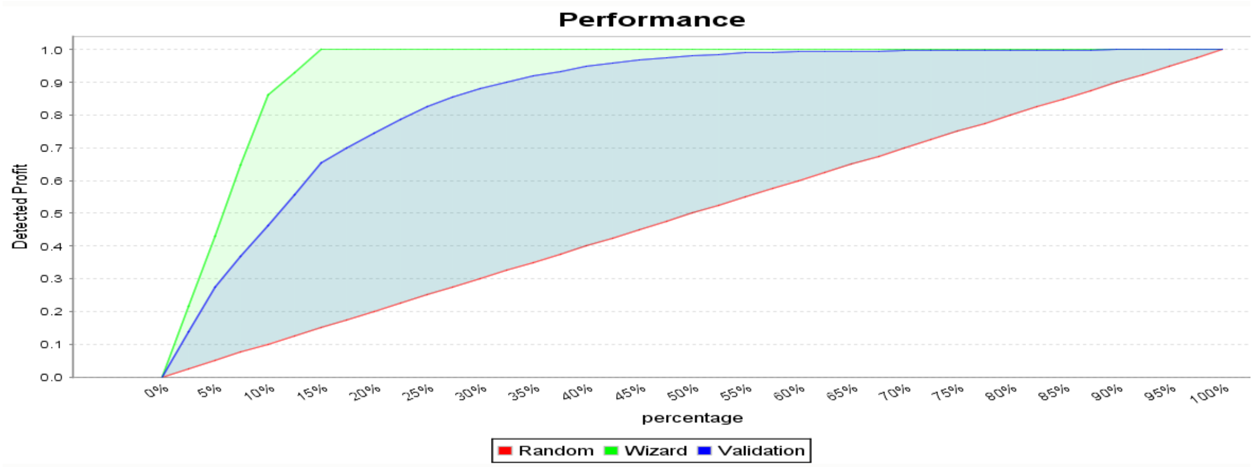
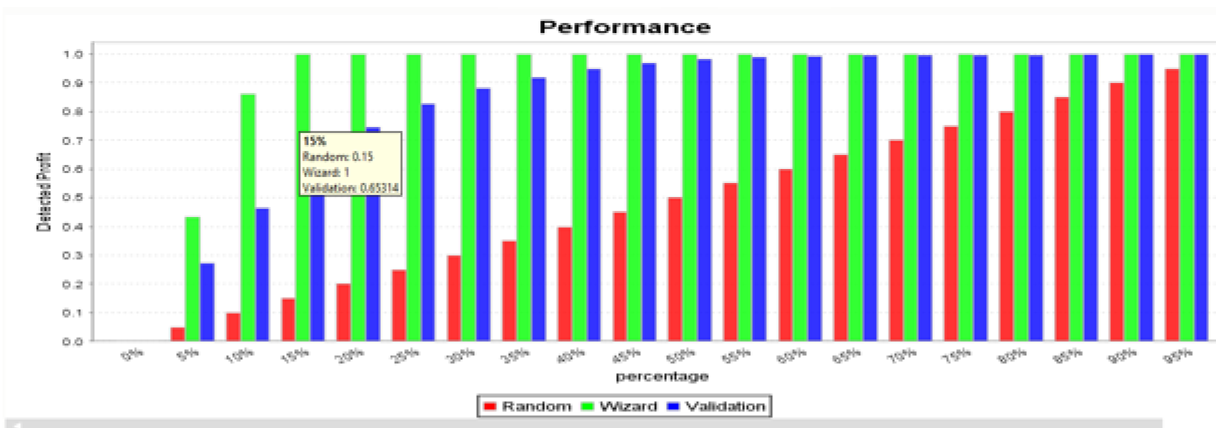


Figure 5: Model performance in column chart



Data Mining Models

We use two data mining approaches, namely decision Tree with R-CNR Tree algorithm and neural network with NNET Neural Network algorithm to build classification models. Although our primary interest is to find the variables which contribute to the success of the campaign most, we are also interested in the accuracy of the predictive model. The decision tree model generated is easier to understand with a little higher accuracy compared with the neural network model. The three most contributory variables used for both algorithms are: duration, month, and pdate. Additional input variables improve the model accuracy slightly.

Figure 6: Decision tree model generated for 2012 data

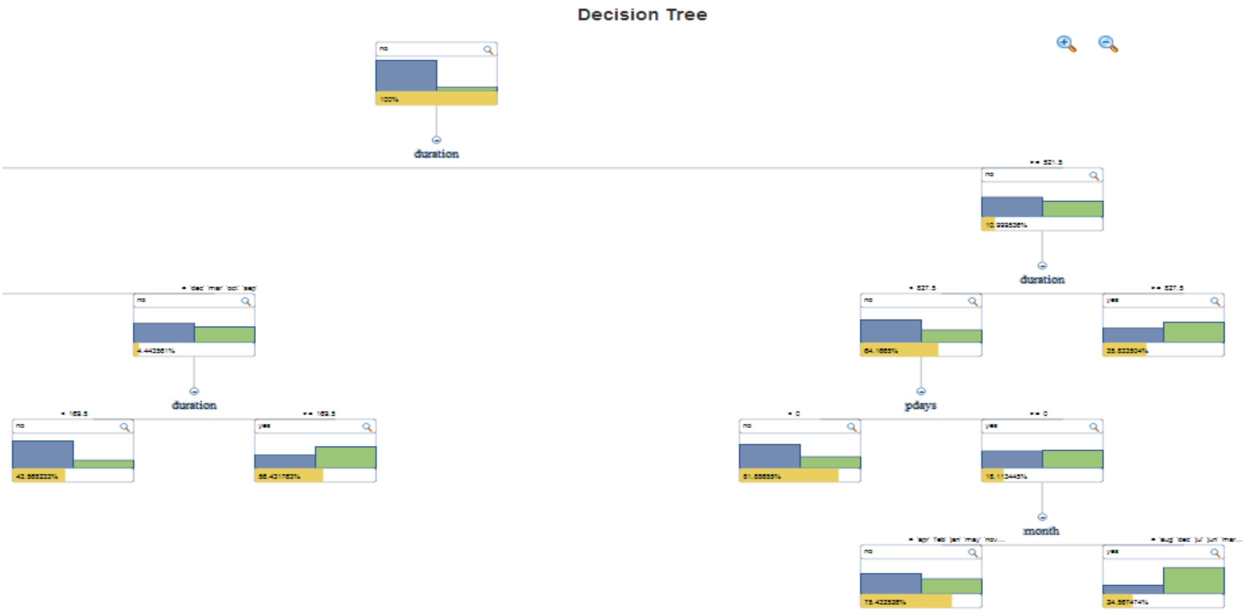


Figure 7 shows the confusion table generated to evaluate the performance of decision tree model. The calculation of the accuracy of the decision tree model uses the following formula:

$$\text{Model accuracy} = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}}$$

The accuracy of the decision tree model is: (2254+3580) / (2254+3580+3035+1342) = 0.9

Figure 7: Confusion matrix for decision tree model on 2012 data

Confusion Matrix

2 X 2 Matrix **Model Accuracy: 0.9**

Actual → Predicted ↓	no	yes
no	38580	3035
yes	1342	2254

Derivatives

Derivative → Class ↓	Precision	Sensitivity	Specificity	Negative Prediction
no	0.97	0.93	0.63	0.43
yes	0.43	0.63	0.93	0.97

Figure 8 shows the confusion table generated to evaluate the performance of neural network model. The accuracy of the neural network model is: (1360 + 38860) / (1360 + 38860 + 3929 + 11062) = 0.89

Figure 8: Confusion matrix for neural network model on 2012 data

Confusion Matrix

2 X 2 Matrix Model Accuracy: 0.89

Actual → Predicted ↓	no	yes
no	38860	3929
yes	1062	1360

Derivatives

Derivative → Class ↓	Precision	Sensitivity	Specificity	Negative Prediction
no	0.97	0.91	0.56	0.26
yes	0.26	0.56	0.91	0.97

Models for 2014 Data

<u>Statistical Logistic Model</u>

We use the SAP predictive analytics tool to automatically select the most contributory input variables depicted in Figure 9.

Figure 9: Selected contributory variables for 2012 data

Selecting Contributory Variables

Targets: y Smart Selection View Reports

Variable	Maximum Contribution	KI	KR	KI + KR	Selected	r	Priority	Last Iteration
nr.employed	28.84%	0.5454	0.9876	1.533	✓			7
duration	25.42%	0.6199	0.9853	1.6052	✓			8
euribor3m	14.68%	0.5423	0.9844	1.5267	✓			8
pdays	7.44%	0.1918	0.9978	1.1896	✓			4
emp.var.rate	6.23%	0.5314	0.9858	1.5172	✓			3
cons.price.i...	4.9%	0.5044	0.9843	1.4887	✓			6
cons.conf.idx	4.9%	0.4979	0.9915	1.4894	✓			5
previous	2.79%	0.2129	0.9962	1.2091				2
age	2.03%	0.2078	0.9893	1.1971				2
month	1.46%	0.3029	0.9914	1.2943				2
contact	1.31%	0.2355	0.98	1.2155				2
poutcome	0%	0.2295	0.9962	1.2257				1
job	0%	0.2131	0.9879	1.201				1
education	0%	0.1113	0.9852	1.0965				1
day_of_week	0%	0.0249	0.9969	1.0218				1
default	0%	0.1198	0.9894	1.1092				1
campaign	0%	0.1211	0.9749	1.096				1
marital	0%	0.0843	0.9901	1.0744				1
housing	0%	0	1	1				0
loan	0%	0	1	1				0

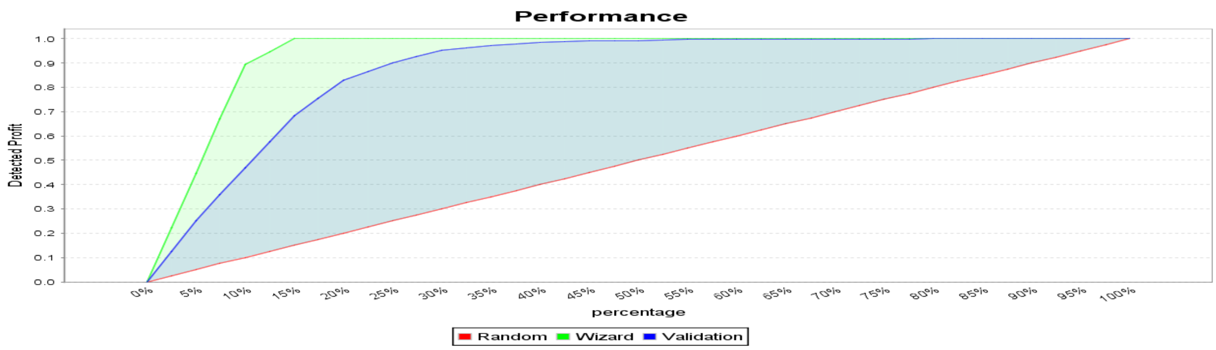
The 2014 dataset uses a few more social and market variables:

1. nr.employed: number of employees - quarterly indicator
2. euribor3m: euribor (Euro Interbank Offered) 3 month rate - daily indicator
3. emp.var.rate: employment variation rate - quarterly indicator
4. cons.price.idx: consumer price index - monthly indicator

5. cons.conf.idx: consumer confidence index - monthly indicator

These input variables turned out to be more important than the bank's client data. We notice there is a strong negative correlation of 0.96 between "number of employees" and "euribor 3 month.
Figure 10 shows the curve chart to evaluate the model performance. When 15% of the population are selected, the predictive model may identify 68.5% of the targets. The model may identify 90% of the targets when 25% of the population are selected.

Figure 10: Model performance in curve chart



Data Mining Models

We also use decision Tree with R-CNR Tree algorithm and neural network with NNET Neural Network algorithm to build classification models for 2014 dataset. The most contributory variables used for both algorithms are: nr.employeed, duration, eurubor3m and pdate. The model accuracy for the 2014 data is a little higher than the 2012 data, but the true positive rate is above 50%.

Figure 11 and 12 show the confusion tables generated to evaluate the decision tree model and the neural network model. From these tables we can calculate the accuracy of the decision tree model is.91 (vs. .9 for 2012 data), and the accuracy of neural network model is: .9 (vs. .89 for 2012 data). We also compare the decision tree and neural network models performance for the 2014 data with the contributory variables identified from the 2012 data. The model accuracies are the same: 91, but using nr.employeed, duration, eurubor3m and pdate as input variable achieved hither true desired outcome compared to using duration, month, pdays as input variables for both decision tree model and neural network model on 2014 data.

Figure 11: Confusion matrix for decision tree model on 2014 data

2 X 2 Matrix **Model Accuracy: 0.91**

Actual → Predicted ↓	no	yes
no	35068	2064
yes	1480	2576

Derivatives

Derivative → Class ↓	Precision	Sensitivity	Specificity	Negative Prediction
no	0.96	0.94	0.64	0.56
yes	0.56	0.64	0.94	0.96

Figure 12: Confusion matrix for neural network model on 2014 data

Confusion Matrix

2 X 2 Matrix **Model Accuracy: 0.91**

Actual → Predicted ↓	no	yes
no	35222	2272
yes	1326	2368

Derivatives

Derivative → Class ↓	Precision	Sensitivity	Specificity	Negative Prediction
no	0.96	0.94	0.64	0.51
yes	0.51	0.64	0.94	0.96

The 2014 bank direct marketing dataset selected four more explanatory variable data, such as, nr.employed, euribor3m, emp.var.rate, cons.price, cons.conf.idx, which contribute to build better, more accurate predictive model to hit the target. This indicates the importance of collect and select more relevant data to build the predictive model. Moro et al. (2014) saw the impact of financial crisis on the effectiveness of bank direct marketing campaign so they included more dependent variables for their data mining process. This further indicated that the bank's database should collect and store sufficient and accurate data for the current need but also prepare for any economic, financial and consumer behavior changes. The advancement of predictive analytics techniques can handle more data with higher speed and enhanced visualization.

CONCLUSION

Predictive analytics has drawn interests from academic and business researchers for decades. The advancement of computing power, especially graphics capabilities have made predictive analytics more applicable to business practitioners. The competitive business environment demands researchers quickly gain insights of the massive data and turn data into actions. With the help of the predictive analytics tools, this goal is more achievable. This paper shows a case of using SAP predictive analytics tools to predict clients' responses to a bank's term deposit campaigns. By comparing the data analytics results of the 2012 and 2014 bank direct marketing datasets, we proved the importance of collecting and including more relevant data variables in data mining processes. The modern data storage and execution capabilities can easily accomplish these tasks. To prepare for the rapid changes in the business environment, business should collect and store relevant data as much as possible, not just for current data analytics needs but also prepare for the future needs.

REFERENCES

Chen, Hsinchun and Roger H. L. Chiang (2012) "business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly*, 36 (4), pp. 1165-1188.

Experian data quality. "Maximizing personalization: How to improve data insight to better consumer connections," White paper (available at https://www.edq.com/globalassets/white-papers/maximizing-personalization.pdf)

Lund, S., Manyika, J., Nyquist, S., Mendonca, L., and Ramaswamy, S. 2013. "Game Changers: Five Opportunities for US Growth and Renewal," McKinsey Global Institute Report (available at http://www.mckinsey.com/global-themes/americas/us-gamechangers).

Moro, S., R. Laureano and P. Cortez. (2011) "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology," *Proceedings of the European Simulation andModelling Conference - ESM'2011, pp. 117-121.*

Moro, S. P. Cortez and P. Rita. (2014) "A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, Elsevier, 62:22-31.

Shmueli, G., and Koppius, O. R. 2011. “Predictive Analytics in Information Systems Research,” *MIS Quarterly* (35:3), pp.553-572.

Urbanski, Al (2016) “Data-driven marketing economy tops $200 billion,” available at http://www.dmnews.com/dataanalytics/data-driven-marketing-economy-tops-200-billion/article/471021/