Business intelligence project final

This needed knowledge about clients' behavior towards deposits. They needed a marketing strategy to Increase their funds by attracting more deposits from existing as well as new clients. One such marketing strategy Is adopting a Tell- marketing campaign and taking vital business decisions from the results of the campaign to increase the deposit subscription rate and generate profits. Of course, this involves investment in running the campaign and if not properly administered would affect the cost structure of the bank. However, proper data mining methods can be used to minimize the cost and perform a successful marketing campaign.

This requires the knowledge of sleeplessness's to Tell-Marketing campaigns. In this project we are going to address the same situation which a Portuguese banking institution faced. DATASET DESCRIPTION The data which we are mining Is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone access if the product (bank term deposit) would be (eyes') or not (no') subscribed. There are 20 input variables. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y). ATTRIBUTE INFORMATION Variable Age Job Marital Type

Interval Nominal Education Default Housing Loan Contact Month Binary Description Age of the person type of Job marital status IANAL. Course', university. Degree', unknown' has credit in default has housing loan? Has personal loan? Contact communication type last contact month of year Missing Values 1731 8597 day_of_week Duration last contact day of the week last contact duration, in seconds Campaign number of contacts performed during this campaign and for this client Paydays number of days that passed by after the client was last contacted from a previous campaign Previous Epitome Categorical al MME. Vary. Rate cons. Price. DXL ones. Con. Ids ribosome Y(Target Variable) 3 number of contacts performed before this campaign and outcome of the previous marketing campaign employment variation rate – quarterly indicator consumer price index – monthly indicator consumer confidence index – monthly indicator ribbon 3 month rate – daily indicator number of employees – quarterly indicator has the client subscribed a term deposit DATA PRE-PROCESSING Among the input variables, following six variables have missing values, which are a result of error in data acquisition and hence considered as ‘ unknown’. Variables 330 Modal Class Admit Married University Degree

No Yes Frequency of 10422 27214 12168 32588 21576 33950 All the variables except Education and Default have negligible amount of unknown values. But for the variables ‘ Education’ and ‘ Default’ there is noticeable number of unknown values. The missing values in these variables need to be replaced and imputed in order to prevent Enterprise Miner from disregarding these attributes altogether, during the modeling process. 4 We also examined the data through the Stratosphere node. Following findings are obtained: Figure 1: Variable Worth Plot We found that out of the 20 input variables, 10 are Categorical/Nominal and 10 are of

Interval type. Also, our target variable, I. E. Y is binary (nominal) in nature. Variable Worth plot orders the variables in their worth in predicting the target variable. We see that ‘ duration’ has the highest order in determining the output variable y. This attribute highly affects the output target (e. G. , if duration-? O then Fan’). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Hence, e have rejected this variable for more accurate results. Data Partition Node: We have not transformed any of the variables. We know that practically, a distribution is considered normal if its keenness is less than 1. 5. Since the keenness for all variables in the dataset is less than 1, we conclude that the variables are practically normal. We partitioned our dataset into Training, Validation and Test Data as per following ratio – Training -60% Validation – 30 % -rest-10% 5 ** We did our analysis by trying different ratio combinations like 60: 20: 20 or 40: 30: 30, however, we got the best results for ratio 60: 30: 10.

Hence we considered this ratio as our final value of portioning data. Replacement Node: In the next step, we used the Replacement Node in order to make sure that Enterprise Miner does not reject the variables with a lot of UNKNOWN values. Replacement Node is employed in order to guide Enterprise Miner that ‘ unknown’ value signifies missing value for a particular input variable. Len the replacement editor for class variables, we specified Replacement Value as _MISSING_ for Variables which have Formatted Value as unknown.

Impute Node: Next, we impute the missing values using the Impute node. Initially, we envisioned hat we’ll use clustering and replace the missing value for a particular variable through the modal value of that variable in a particular cluster. But, the better The values of missing class variables, in our case, are imputed using predicted values from a decision tree. For each class variable, ASS Enterprise Miner builds a decision tree (in this case, potentially using surrogate splitting rules) with that variable as the target and the other input variables as predictors.

Interactive Binning: To better understand the effects of input variables on the target variable, we chose to bin two variables Job and education. The criteria for binning the two variables are shown below. Figure 2: Bins for ‘ Education’ Variable 6 Figure 3: Bins for Job’ Variable Transform Variables: We observe tattoo variables likelihood’s’ and ‘ campaign’ are positively skewed and one variable ‘ REP_paydays’ is negatively skewed. We chose to adopt log transformation for positively skewed variables and Square transformation for negatively skewed variables.

In order to do so, we take the help of Transform variable node. Figure 4: Keenness before and after transformation Applying log transformation, the keenness has reduced to a great extent. However, even after square transforming the ‘ paydays’, keenness remained the same. This is because more than 70% of observations have 999 as its value (which means that the respondents were never contacted before the campaign). This left us with 2 options, either to reject the variable or use the variable as is.

In order to get a better picture of advantages of transformation and binning, we applied Regression model twice, one without transformation and the other one with transformation. We observed that without transformation and binning, Intercept alone is the most dominating factor which is neglecting many important variable in determination of output variable y. Figure 6:

Effects Plot for Regression Output without transforming input variables In order to reduce the effects of over dependency, we have applied certain transformations and re binning, as mentioned earlier.

After transformation and re- binning, we observed that results are more realistic and of practical importance. As we can see below, the absolute coefficient value is determined correctly and also, the effect of each and every variable is shown. The blue columns denote a negative relation between the output variable and corresponding input variable whereas the red columns denote a positive relation between the output variable and responding input variable. Figure 7: Effects Plot after Transformation and Binning Figure 8: Regression Output with Transformation 8 9 The regression uses 35 independent variables to predict the dependent variables.

The independent variables contain dummy variables for both nominal and interval variables. Some of the variables are not significant individually, but they are significant Jointly with other variable. Although the intercept is still the most influential estimator, it is closely followed by the variables such as month, various macroeconomic indicators, number of days employed etc. The cumulative lift plot is shown below. The validation plot closely follows the training plot, which says that the model is a good estimator across different datasets. The curve is in a negative slope which is expected from a good estimator.

Based on the cumulative lift plot of Training and Validation data, we observed that the predictive model chosen ‘ LOGIC’ is a good estimator for target variable y. However, in order to find the best model, we employed

Decision trees and Neural Networks as well. NEURAL NETWORKS: The prediction formula adopted by Neural Networks to predict is similar to a aggression’s, but with an interesting and flexible addition. We employed the Neural Network Model with default stationmaster the Interactive binning Node in order to make sure, we don’t miss out any variable having missing values.

The results show the iteration plot for the average squared error versus optimization iteration. We can clearly see the divergence in training and validation average squared error occurring near iteration 15. Hence, this model could not be considered as the ideal model. 10 Figure 6 Iteration Plot for Average Square Error between Validation data and Training data To evaluate the data better, we applied Decision tree model with default settings. As we have already discussed, we provided the output of data Partition node to the Decision tree Node.