Bank Marketing Dataset (CIND 119)

Prof: Syed Shariyar Murtaza, Ph.D.
December 15, 2020

Members

Ilhak Park, ilhak.park@ryerson.ca
Jaekang KIM, j20kim@ryerson.ca

Workload Distribution

Member Name	List of Tasks Performed
Ilhak Park	Equally performed
Jaekang KIM	Equally performed

Abstract

This study analyzes the customers that will subscribe to long-term deposit accounts, with a portugese bank seeking out an effective telemarketing strategy to target these customers. The bank has approached our team to help in devising a strategy by analyzing the dataset that was collected. Data preparation results include discretizing numerical to categorical attributes (eg. X bins to have Age groups), and normalizing attributes with skewed distribution to lower the impact of higher range of scales. In addition, the dataset has been balanced as an imbalance existed in class distribution. After performing these transformations on the original dataset, TP of the class y, for 'yes', greatly improves. Using the selected features from data preparation, two classification models (Decision Tree and Naïve Bayes) were applied. Based on the confusion matrix, Decision Tree is the better model with higher TP rate and accuracy. Call duration (duration) and previous marketing outcome (poutcome) are two attributes that provide the highest information gain. Thus, after eliminating attributes that provide weak information, an effective strategy would include contacting customers who were already contacted previously with a longer duration of the call.

Data Preparation

The dataset is collected from a Portuguese bank and the bank wants to have an effective telemarketing strategy to sell long-term deposit accounts (e.g., bonds, saving accounts, etc.). These marketing campaigns were based on phone calls and multiple contacts were often needed to determine whether a customer would subscribe to a long-term deposit account.

Definition of Variables

Variable	Description	Data Type
Age	Age of the customer	Numeric
Job	Type of job	Qualitative
Marital	Marital status	Qualitative

Education	Education of the customer	Qualitative
Default	Shows whether the customer has credit in default or not	Qualitative
Balance	Average yearly balance in Euros	Numeric
Housing	Shows whether the customer has housing loan or not	Qualitative
Loan	Shows whether the customer has personal loan or not	Qualitative
Contact	Shows how the last contact for marketing campaign has been made	Qualitative
Day	Shows on which day of the month last time customer was contacted	Numeric
Month	Shows on which month of the year last time customer was contacted	Qualitative
Duration	Shows the last contact duration in seconds	Numeric
Campaign	Number of contacts performed during the marketing campaign and for this customer	Numeric
Pdays	Number of days that passed by after the client was last contacted from a previous campaign, -1 means client was not previously contacted	Numeric
Previous	Number of contacts performed before this campaign and for this client	Numeric
Poutcome	Outcome of the previous marketing campaign	Qualitative
Y	Class attribute showing whether the client has subscribed a term deposit or not	Binary: “yes”, “no”

Summary of Numeric Variables

	Age	Balance (Euro)	Day	Duration
Attribute	Quantitative	Quantitative	Ordinal	Quantitative
Min	19	-3,313	1	4
Max	87	71,188	31	3,025
Mean	41.17	1,422.66	15.92	263.96
Std.Dev.	10.58	3,009.64	8.25	259.86

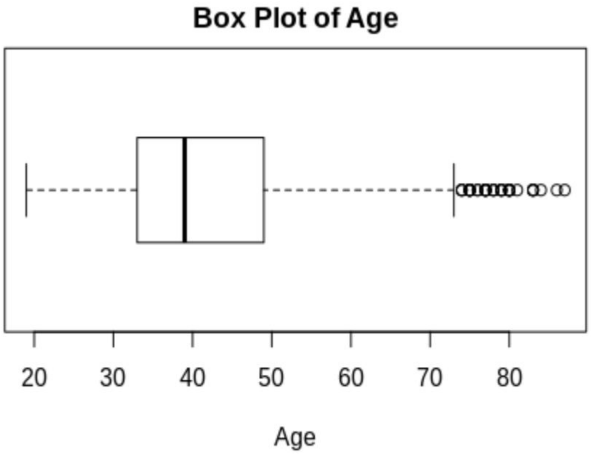
	Campaign	Pdays	Previous
Attribute	Quantitative	Quantitative	Quantitative
Min	1	-1	0
Max	50	871	25
Mean	2.79	39.77	0.54
Std.Dev.	3.11	100.12	1.69

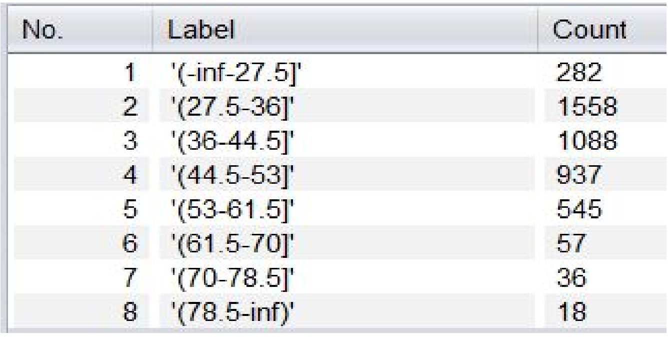
Summary of Qualitative Variables

	Job	Marital	Education	Default	Housing
Attribute	Nominal	Nominal	Ordinal	Nominal	Nominal
MostFreq	management	married	secondary	no	yes
CountFreq	969	2797	2306	4445	2559
UniqueItems	12	3	4	2	2

	Loan	Contact	Month	Poutcome	y
Attribute	Nominal	Nominal	Ordinal	Nominal	Nominal
MostFreq	no	cellular	may	unknown	no
CountFreq	3830	2896	1398	3705	4000
UniqueItems	2	3	12	4	2

The above tables are generated to have a better understanding of the data. For the numeric variables, min, max, mean and standard deviation are examined. On the other hand, the qualitative variables are divided into groups, and the variables' unique groups and its frequencies are examined for those variables.

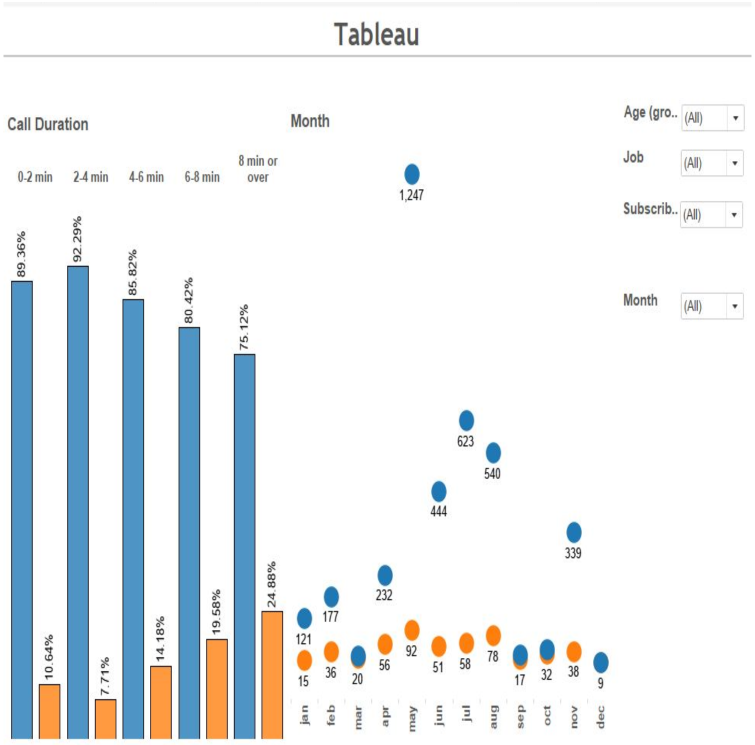




No.	Label	Count
1	'(-inf-27.5]'	282
2	'(27.5-36]'	1558
3	'(36-44.5]'	1088
4	'(44.5-53]'	937
5	'(53-61.5]'	545
6	'(61.5-70]'	57
7	'(70-78.5]'	36
8	'(78.5-inf)'	18

The distribution of 'Age', for instance, is positively skewed and has upper outliers. In order to discretize and normalize such numerical attributes, 'Age' has been divided into eight groups as shown in the above table. Since the 'Previous' variable contains around 82% of value, '0', it has been grouped into '0 time', '1 time' and 'more than 2 times' for simplicity. Also, 'Balance' is a numerical attribute that has a heavy right skewed distribution with upper outliers. The variable has been discretized and divided into five different groups, including 'negative balance'. 'Campaign' also has a right skewed distribution, so it has been grouped into: '1 time' to '5 time' and '6 time +'. 'Duration' measures the last contact duration in seconds. For simplicity, it has been converted to minutes and binned into five groups with a range of two minutes each and have '8 min or over' for the last interval.

To obtain a better result, some of the meaningless attributes are removed prior to testing. More than 98% of 'Default' contains 'no', and this imbalance would not make this attribute work effectively when testing. In addition, the variable 'Pday' has been removed. 'Pdays' is supposed to be a numerical data, but it contains '-1', indicating the client was not previously contacted, which represents a nominal category, and around 82% of the data has this value. During the correlation analysis, 'Pdays' and 'Previous' are observed to have the highest correlation coefficient, therefore; 'Pdays' would not play an important role.



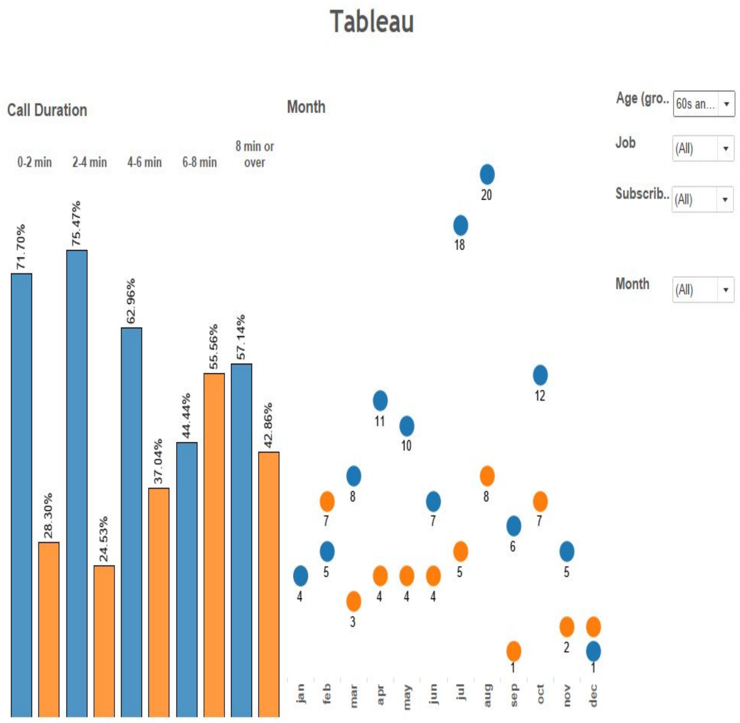
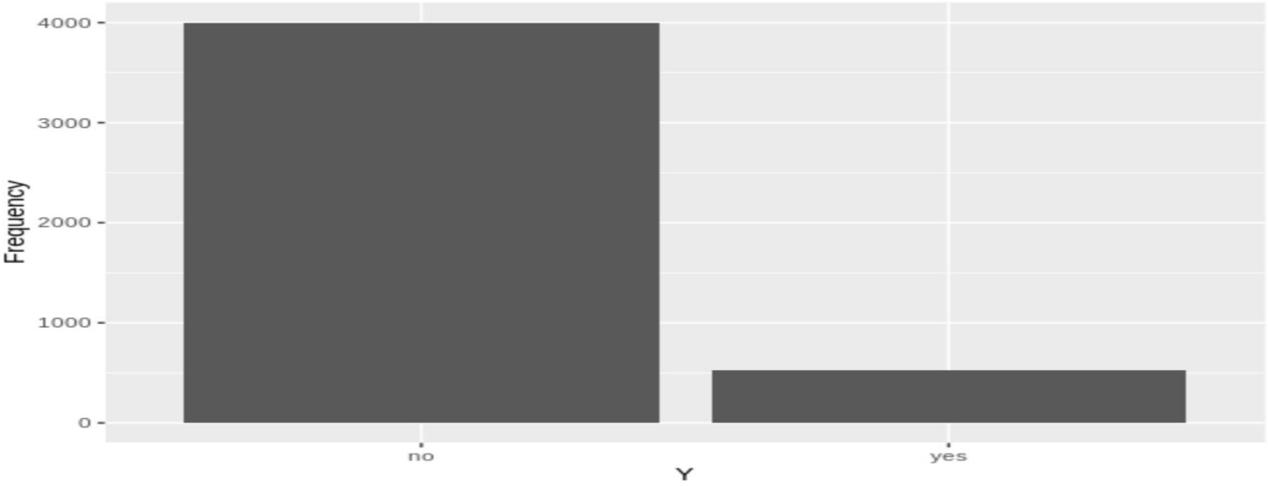
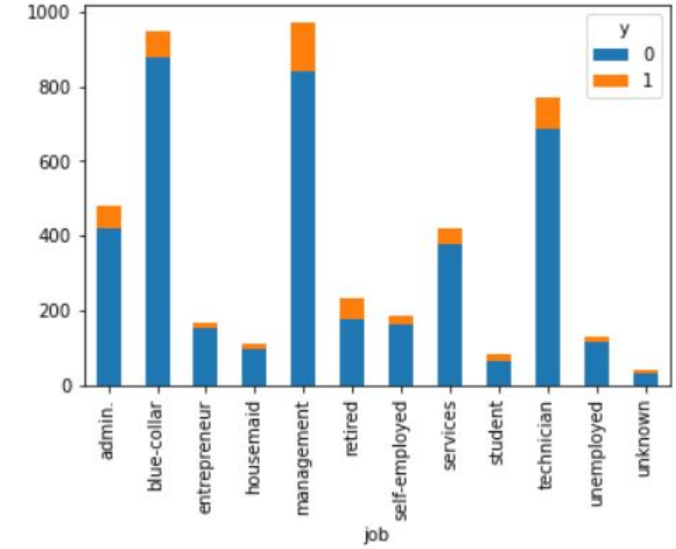


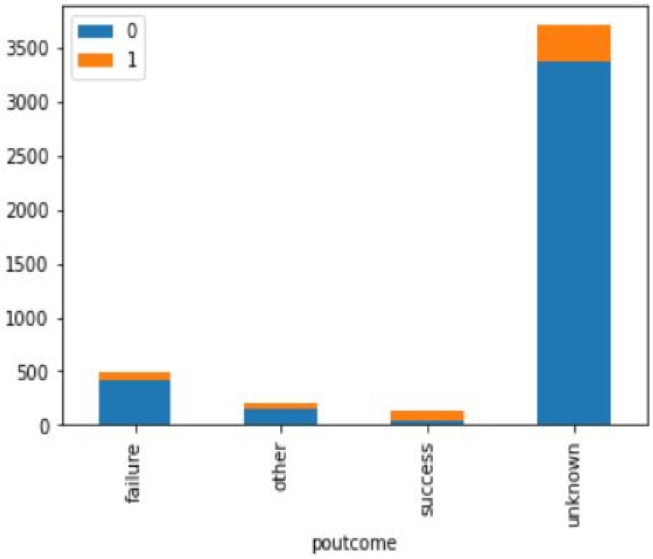
Tableau analysis shows that ‘Day’, which measures day of the month, could be removed. On the other hand, it shows that winter time has a high percentage of ‘yes’ compared to summer as shown in the left graph above. Therefore, the attribute, ‘month’ has not been removed in order to test out if certain months are more effective for the campaign. The above two graphs display a dramatic change in ‘Duration’ when the age groups of ‘61 or above’. This analysis demonstrates that old people are more convincible as the duration of conversation increases.

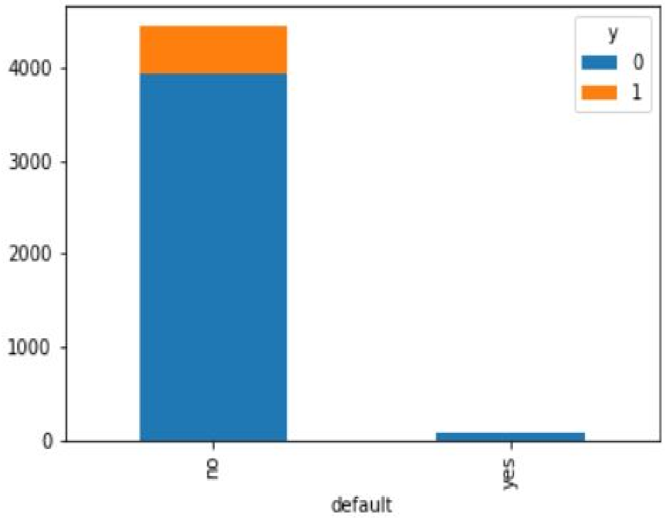


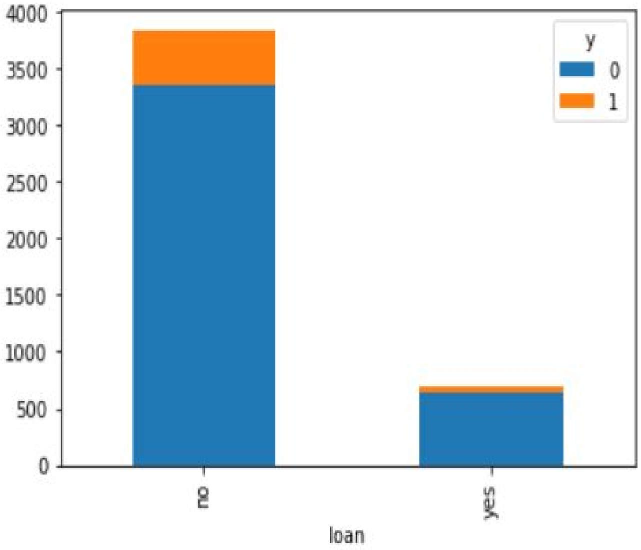
The above graph is a frequency distribution of the variable ‘Y’. The number of clients who have subscribed to a term deposit (‘Yes’) is 521 while 4,000 of the other clients have not (‘No’). In order for a model to generate higher balanced accuracy and balanced detection rates, the imbalanced data needs to be balanced.

The following images are visual representations of the ratio of 'Yes' (1) and 'No' (0) given attribute.









Predictive Modeling/Classification

For predictive modeling, the dataset is split by a train-test set (66% train). First, classification algorithms are applied to the original dataset so that comparisons can be made against the re-trained models with selected features. For the original dataset, TP rate for Decision Tree and Naïve Bayes models are 0.333 and 0.472 respectively despite the high accuracy. This rate tells us that an imbalance does exist in the dataset; thus, the task of balancing the training data during the data preparation stage (preprocess in weka) is indeed a good idea before retesting with selected features. For 4000 no and 521 yes, the weight has been reassigned evenly at half of 4521 which equals 2260.5.

In addition, two evaluators in weka, for correlation and information gain, were examined before running the classification algorithms. The results aligned very closely with our hypothesis in the data prep stage for attributes that can be eliminated. The decision was to use the ranked attribute evaluated from information gain and delete/eliminate the irrelevant attributes. Correlation ranking is not used in the classification algorithm that follows. Therefore, our new engineered dataset will be of the balanced dataset made of only the attributes that ranked the highest in information gain. These attributes include previous campaign outcome and economical situation. This new dataset is used to re-run the Decision Tree and Naïve Bayes algorithms.

Decision tree

The algorithm on the new dataset shows that the Decision Tree starts with the root of **duration** then branches out to the next relevant attribute of **poutcome**. Duration is the attribution with the maximum gain. Thus, Decision Tree suggests that customers are more likely to subscribe if already contacted previously with a longer duration of the call. There is a noticeable increase in the TP/precision/recall. Not only are the percentage of results more relevant, there is a high correct classification of relevant results as well. Since there is a trade-off between these two, we could look at maximizing f-measure which is a function of precision and recall. For our result, F-measure came to 0.808.

<table>
<tr><th></th><th>Decision Tree - original</th><th>Decision Tree - selected features</th></tr>
<tr><td><b>CCI</b></td><td>88.484%</td><td>80.054%</td></tr>
<tr><td><b>ICI</b></td><td>11.516%</td><td>19.947%</td></tr>
<tr><td><b>TP</b></td><td>0.333</td><td>0.833</td></tr>
<tr><td><b>FP</b></td><td>0.042</td><td>0.233</td></tr>
<tr><td><b>Precision</b></td><td>0.513</td><td>0.785</td></tr>
<tr><td><b>Recall</b></td><td>0.333</td><td>0.833</td></tr>
<tr><td><b>F-Measure</b></td><td>0.404</td><td>0.808</td></tr>
<tr><td colspan="3"></td></tr>
<tr><td colspan="3"><b>Confusion matrix</b></td></tr>
<tr><td>a = no
b= yes</td><td><table><tr><th>a</th><th>b</th></tr><tr><td>1300</td><td>57</td></tr><tr><td>120</td><td>60</td></tr></table></td><td><table><tr><th>a</th><th>b</th></tr><tr><td>588.3</td><td>178.58</td></tr><tr><td>130.16</td><td>650.82</td></tr></table></td></tr>
</table>

Naïve Bayes

Accuracy using Naïve Bayes is 74.898% (below) which is lower than accuracy of the Decision Tree above. Other results (TP, FP, etc) improved like the Decision Tree algorithm (vs the original dataset). However, in almost all cases, Decision Tree seem to have the better result.

	Naïve Bayes - original	**Naïve Bayes - selected features**
CCI	87.573%	74.898%
ICI	12.427%	25.101
TP	0.472	0.733

FP	0.071	0.235
Precision	0.470	0.761
Recall	0.472	0.733
F-Measure	0.471	0.747
Confusion matrix		

a = no
b= yes
	a b
	1261 96
	95 85
		a b
		586.6 180.27
		208.26 572.72

Based on the result above, Decision Tree is the better performing algorithm since it is a selective model as opposed to Naïve Bayes which is a general model. While the Decision Tree helps us pick the best features from the dataset, this is not really the case for Naïve Bayes. Especially in Naïve Bayes, selecting the features matter. In our bank dataset, our data is fairly large with over 4500 rows, and Decision Tree's ability to prune and exclude properties that don't matter can impact the result and the result can be seen as shown in the chart above.

Conclusions and Recommendations

The purpose of this study has been to identify the customers most likely to subscribe to a long term deposit by analyzing the bank marketing dataset. With Weka and R tools, data was organized and prepared for predictive modeling. Certain attributes are eliminated, and feature selection and balancing are applied in order to run the classification algorithm more efficiently (seen in the result of classification after rebalancing). With predictive modeling, it is confirmed that some attributes such as (but not limited to) **default** and **pdays** attributes are irrelevant (as found in job prep. stage) in determining whether or not a customer would subscribe to a deposit. On the other hand, using the previous history of calls or economical situation (eg. simply a presence of an individual loan) helped the most in predicting the target customers. Although clustering can be performed to give further insight into this analysis, our team would like to recommend to the bank that they focus on targeting the customers (in good financial standings) already contacted in previous campaigns for a longer phone call duration.