BANK MARKETING DATASET ANALYSIS

SYNOPSIS

A Portuguese bank had been struggling with its telemarketing strategy of promoting subscription for its long-term deposit accounts for example saving accounts, bonds etc. Our team was asked to build an effective telemarketing strategy to promote these products to its potential subscribers based on the data the bank had previously collected. To serve this purpose, firstly we identified the important characteristics of these customers who are willing to subscribe to these accounts. Secondly, we built two classification models to compare and reach at the better model. The classification models include J48 Decision Tree and Naïve Bayes classification models. Post-predictive analysis proved that Naïve Bayes was a better model based on True Positive (TP), False Positive (FP) and F1 Score. Naïve Bayes classification model suggests that candidates with median age of 39 years, who stay on the call for longer duration and were contacted for some previous marketing campaign are most inclined on signing up for a new long-term deposit account.

1. DATA PREPARATION

We scrutinized the Bank Marketing dataset and compiled basic statistical measures for all the attributes. Based on the descriptive statistics and histogram of all attributes, we concluded none of them are normally distributed. No missing value was observed in the given dataset.

We loaded the dataset into Weka GUI for more further analysis. We used the Select Attributes function to evaluate the effect of each attribute on the class attribute y. In the next step, we used Gain Ratio evaluation; it evaluates the worth of an attribute by measuring Gain ratio evaluator with respect to the class. We also used CfsSubsetEval - Attribute subset evaluator in order to confirm our claims about attribute relation.

Based on the information provided from these evaluators we gathered some facts worth highlighting: The top three attributes are *Age, Duration and Poutcome* whereas, the attributes contributing the least were Education, Default, and Day.

- **Age** shows how old the customer is
- **Duration** measures the last contact duration in seconds
- **Poutcome** depicts the outcome of the previous marketing campaign
- **Education** is the highest education level completed by the customer
- **Default** indicates whether the customer has credit in default or not
- **Day** indicates on which day of the month last time customer was contacted

Gain ratio showed that the *age* is crucial in terms of determining the likelihood of customers who are willing to subscribe to the long-term deposit accounts. The least influential factor is on which day of the month the customer was contacted.

We noticed some outliers in the different attributes in dataset, but we found its effect is limited or insignificant; therefore, we did not remove any outlier from the given dataset.

We then started classifying the whole dataset with all the attributes. In our first classification model, we received high percentage accuracy in our model, however, in the prediction many customers were misclassified as "no", which caused a distortion in our model. Further, looking closely at the class attribute y, we saw there were 4000 *"no"* and only 521 *"yes"*. The variable y isn't represented equally in the given dataset which leads to the problem of *imbalanced data*. When we took a closer look at the accuracy measurement of our initial classifier, we observed that because there were significantly higher number of "no" than "yes" in the class attribution, the model automatically predicted "no". All the classifier algorithms are biased in such situation and tend to predict the majority class and consider the minority class as noise. Machine Learning models used for classification are based on the assumption that, there are equal number of instances in each class and therefore, imbalanced classification will pose a challenge for predictive modelling for bank marketing dataset. Our goal would be to correctly classify class "yes". Thus, in order to boost the performance of classification algorithms we decided to balance the dataset.

We also removed the attributes which were not relevant and had low attribute ranking from our

classification model. So, now we received our filtered dataset having three attributes namely *Age*, *Duration* and *Poutcome*.

Identify the Research Questions:

By building a predictive modeling (Decision Tree or Naïve Bayes), we can help the bank to find an effective telemarketing strategy to promote its long-term deposit accounts to its customer.

2. PREDICTIVE MODELLING/CLASSIFICATION

For predictive modelling we applied two classification algorithms to the raw dataset consisting 17 attributes. We split dataset into two parts 66% training dataset and 34% as test dataset. We ran J48 Decision tree and Naïve Bayes on raw dataset, and we obtained a high number of correctly classified instances (above 85%), but the True Positive rate for "yes" class under both models were below 40% or less. This substantially validates our data preparation conclusion that our dataset is indeed imbalanced. We then selected the important attributes and deleted the non-important attributes as mentioned in data preparation.

Later, to tackle the imbalance data, we decided to use ClassBalancer filter in Weka GUI to reweight the instances in the data so that, each class in attribute y has the equal weight for each class. As per theory, we balanced only the training dataset and while the test data remained intact. Later, using the balanced filtered training dataset and filtered test dataset, we again constructed two classification algorithms: J48 Decision Tree and Naïve Bayes. The results of the filtered dataset were compared with raw dataset and the observations are recorded below.

2.1 DECISION TREE – J48

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The advantage of using Decision Trees in classifying the data is that they are simple to understand and interpret

Raw data: By running decision tree - J48 on raw dataset we observed the size of the tree as 146 and the number of leaves are 104. It was very large and over-fitted tree. While evaluating test set it was observed that number of correctly classified instance was around 88%. But True positive rate for "yes" was 33%. It can be clearly seen from above observation that model is predicting more "no" instead of "yes", which is our target. Even precision and recall were quite low.

Balanced Filtered data: By running decision tree - J48 on balanced filtered dataset we observed the size of the tree reduced to 155 and the number of leaves to 80. But True positive rate for "yes" was increased to 67%. We can now say that model is predicting more "yes" instead of "no", which is our target. We got a lower False Positive rate of 16%. Even precision and recall modified to approximately 36.89% and 67.01% respectively.

2.2 NAÏVE BAYES

Naïve Bayesian classifiers assumes that there are no dependencies amongst attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, hence is called "naive". Bayes classifiers uses a very intuitive technique.

Raw data: By running decision Naïve Bayes on raw dataset and evaluating test set it was observed that number of correctly classified instances was approximately 88.48%. But True Positive rate for "yes" was 39%. It can be clearly observed that model is predicting more "no" instead of "yes" and giving us biased classification. Even precision and recall were low. Thereby, to discard this biasedness from the bank marketing dataset filtering out of few irrelevant attributes was essential.

Balanced Filtered data: By running Naïve Bayes on balanced filtered dataset and evaluating test set it was observed that number of correctly classified instance was 84.85%. But True Positive rate for "yes" increased to 58%. We can now say that model is predicting more "yes" instead of "no", which is our target. At the same time even, the False Positive for "yes" was seen as 11.4%.

Based on the various confusion matrices received we tried calculating Accuracy, Recall, Precision for both pre-processed and processed datasets for various classification algorithms. F1 score may be considered as a better measure to seek balance between Precision and Recall and when the class

distribution is uneven. Hence, we shall compare and F1 scores and notice that *Naïve Bayes* classification model for the processed dataset yields highest F1 score.

3. CONCLUSIONS AND RECOMMENDATIONS

The classification model is a technique where we categorize data into a given number of classes. The main objective of classification algorithms is to predict the class labels/categories for the new data. For our current machine learning project, we were supposed to classify the dataset in two classification algorithms and compare the performance of different trained models by using the selected metrics. Of all the classification techniques we used, Naïve Bayes classification trained on selected attributes(balanced) and whose performance was calculated on validation set using same attributes manifests as the best classification algorithm. We understand that the previous marketing campaign was unsuccessful as they targeted wide range of candidates. We identified high-potential clients through a series of predictive modeling techniques. The selected model suggests that candidates who spent longer duration on the phone and have been contacted previously should be targeted for higher success rate. In conclusion, by focusing on customers with a median age of 39, contacted in prior campaigns, and who spend more duration on call during the marketing drive will show a relatively higher tendency of opening new long-term deposit accounts with this Portuguese bank.

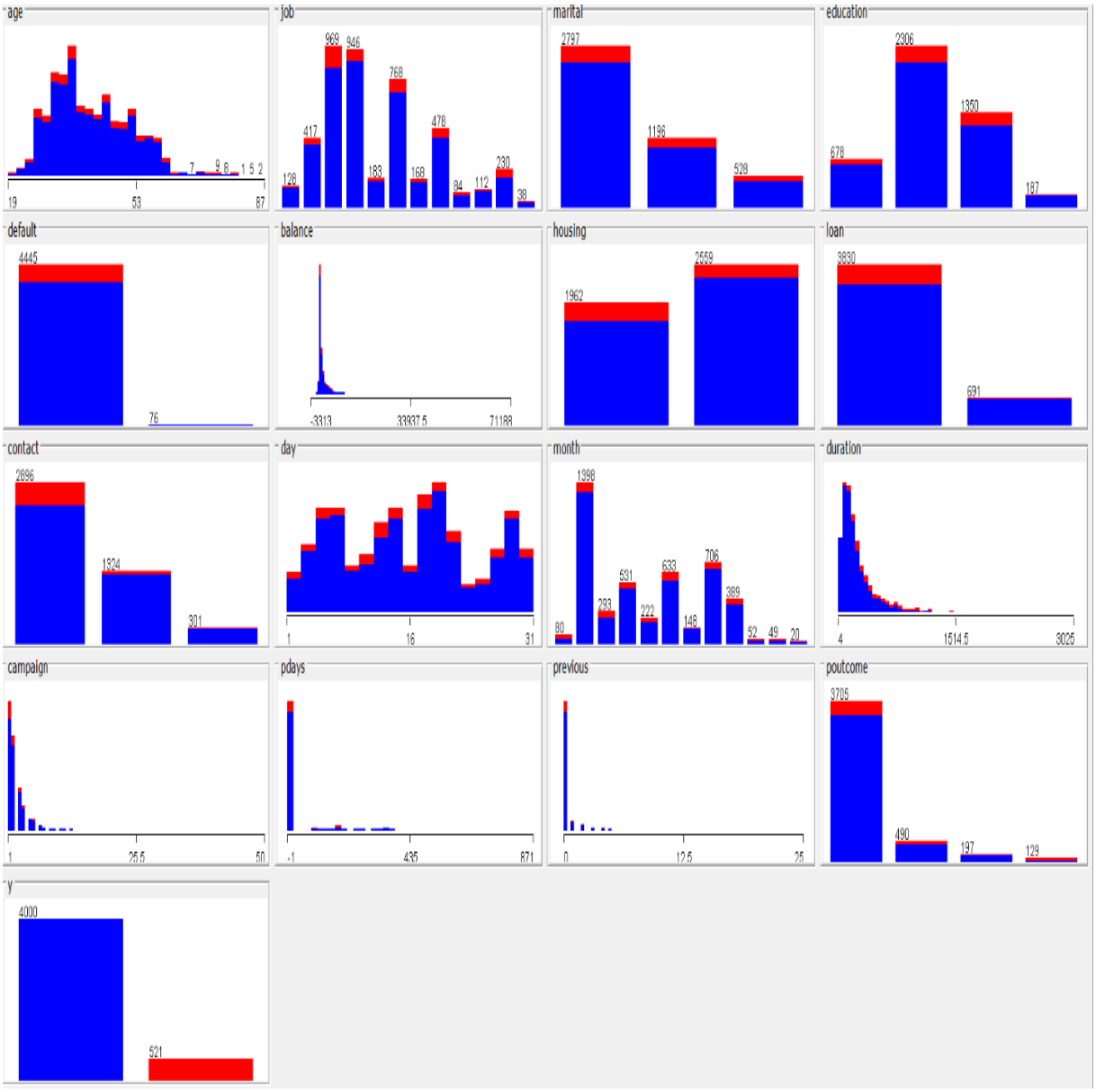
A. APPENDIX – DATA PREPARATION

A.1. Summary of Attributes

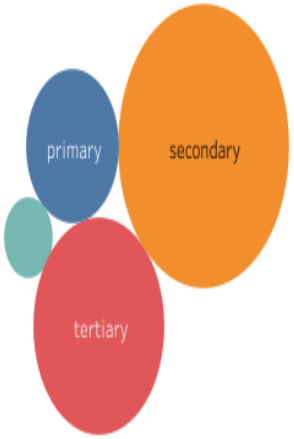
Column Number	Column Name	Type of Attribute	Min, Max	Mean, Std Dev
1	Age	Quantitative Attribute	(19, 87)	41.17 , 10.576
2	Job	Nominal Attribute	n/a	n/a
3	Marital	Nominal Attribute	n/a	n/a
4	Education	Ordinal Attribute	n/a	n/a

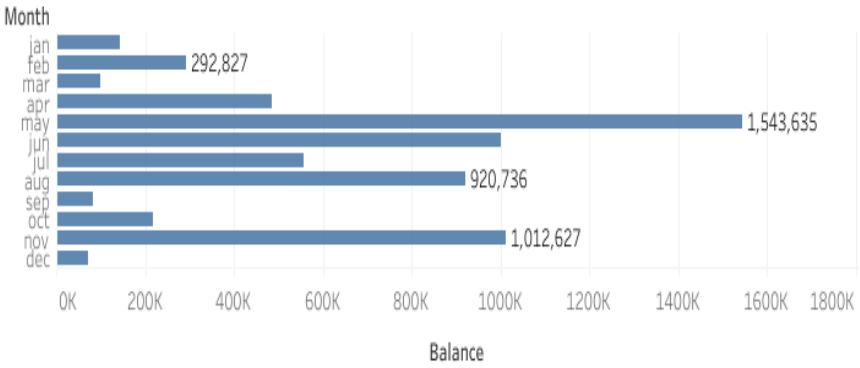
5	Default	Nominal Attribute	n/a	n/a
6	Balance	Quantitative Attribute	(-3313, 71188)	1422.658 , 3009.638
7	Housing	Nominal Attribute	n/a	n/a
8	Loan	Nominal Attribute	n/a	n/a
9	Contact	Nominal Attribute	n/a	n/a
10	Day	Ordinal Attribute	n/a	n/a
11	Month	Ordinal Attribute	n/a	n/a
12	Duration	Quantitative Attribute	(4, 3025)	263.961 , 259.857
13	Campaign	Quantitative Attribute	(1, 50)	2.794 , 3.11
14	Pdays	Quantitative Attribute	(-1, 871)	39.767 , 100.121
15	Previous	Quantitative Attribute	(0, 25)	0.543 , 1.694
16	Poutcome	Nominal Attribute	n/a	n/a
17	Y	Nominal Class	n/a	n/a

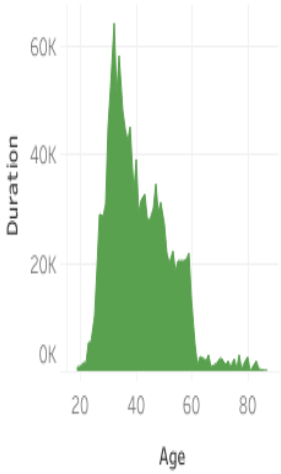
A.2. Histograms for all attributes of raw data

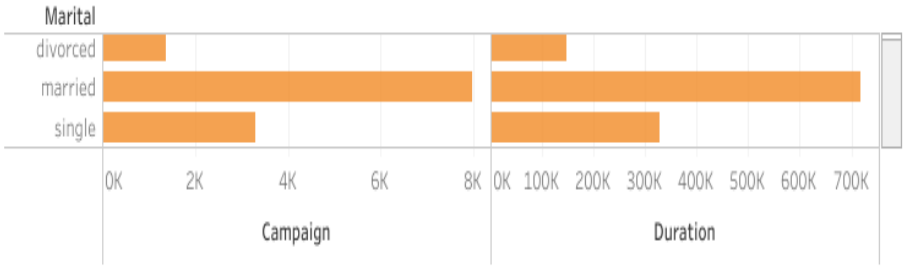

age
19
53
87
job
128
417
969
946
183
768
168
478
84
112
230
38
marital
2797
1196
528
education
678
2306
1360
187
default
4445
76
balance
-3313
33937.5
71188
housing
1962
2559
loan
3830
691
contact
2896
1824
301
day
1
16
31
month
80
1398
293
631
222
633
148
706
389
52
49
20
duration
4
1514.5
3025
campaign
1
25.5
50
pdays
-1
436
871
previous
0
12.5
25
poutcome
3705
490
197
129
y
4000
521

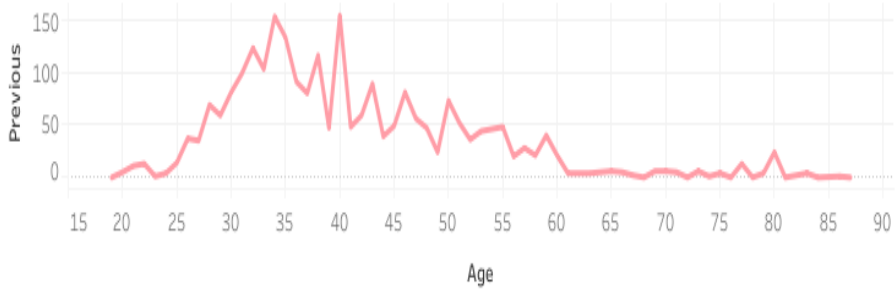
Graphical Representation of Bank Marketing Dataset



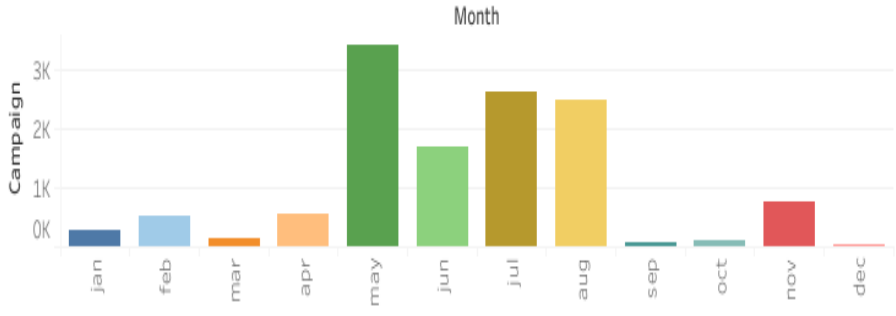








Poutcome	
failure	20,362
other	7,855
success	5,698
unknown	152,215



A.3. Gain ratio for attribute ranking

```
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 y):
        Gain Ratio feature evaluator

Ranked attributes:
 0.0525968    1 age
 0.05168956  12 duration
 0.04057678  16 poutcome
 0.03436486  14 pdays
 0.02382205  15 previous
 0.01371859   9 contact
 0.01024113  11 month
 0.00792732   7 housing
 0.00666696   8 loan
 0.0053377    6 balance
 0.00517803  13 campaign
 0.00325674   2 job
 0.00229025   3 marital
 0.00146323   4 education
 0.00000986   5 default
 0           10 day

Selected attributes: 1,12,16,14,15,9,11,7,8,6,13,2,3,4,5,10 : 16
```

A.4. CFS subset evaluator

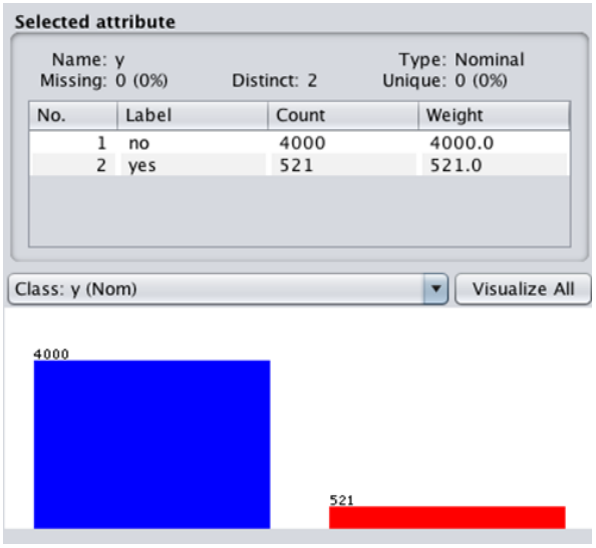
```
              previous
              poutcome
              y
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 97
        Merit of best subset found:    0.095

Attribute Subset Evaluator (supervised, Class (nominal): 17 y):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,12,16 : 3
                     age
                     duration
                     poutcome
```

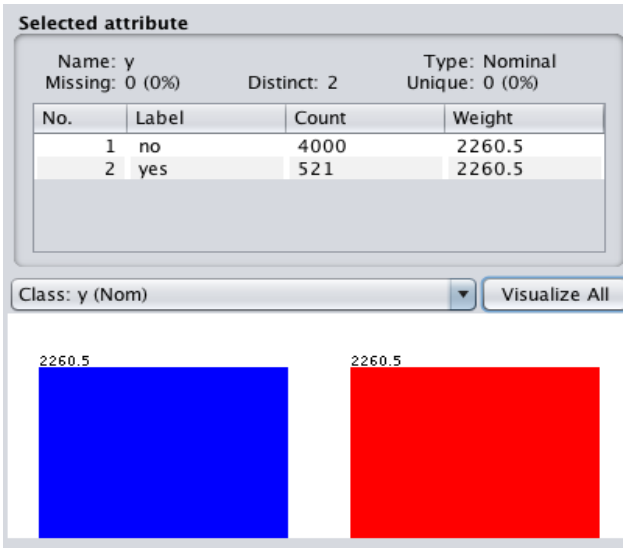
B. APPENDIX – Predictive Modeling

B.1. Initial Classification results (without ClassBalancer Tool)

Classifier	Correctly Classified Instances Rate	TP rate for "yes" Class	FP rate for "yes" Class
J48	88.5%	33.3%	4.2%
Naïve Bayes	87.6%	47.2%	7.1%

B.2. ClassBalancer Tool





We used **ClassBalancer** tool to balance the initial data set (left) using weights (right).

B.3. Classification results with ClassBalancer Tool

Classifier	Correctly Classified Instances Rate	TP Rate for "yes" Class	FP Rate for "yes" Class
J48	81.7%	67%	16.3%
Naïve Bayes	**84.9%**	58.1%	11.4%

B.4. Evaluation of confusion matrices

ACCURACY	Raw Data	Balanced Data (3 explanatory variables)
Naïve Bayes	88.48%	84.85%
Decision tree	87.57%	81.66%

RECALL	Raw Data	Balanced Data (3 explanatory variables)
Naïve Bayes	33.3%	58.1%
Decision tree	47.2%	67.01%

PRECISION	Raw Data	Balanced Data (3 explanatory variables)

Naïve Bayes	48.72%	42.05%
Decision tree	46.96%	36.89%

F1 SCORE	Raw Data	Balanced Data (3 explanatory variables)
Naïve Bayes	39.53%	**48.78%**
Decision tree	47.08%	47.58%