IST 707 FINAL PROJECT REPORT

Group 1

Harper He

Chenyan Huang

Jingxian Sun

Chenying Jiang

CONTENT

INTRODUCTION

This dataset came from UCI Machine Learning Repository, a collection of databases, domain theories, and data generators that were used by the machine learning community for the empirical analysis of machine learning algorithms. The name of this data set was “Bank Marketing Data Set“. There were 45211 examples and 17 variables in the data set. Target Variable was variable “y”, a variables describing if the client would subscribe (yes/no) a term deposit.

OBJECTIVE AND BUSINESS QUESTIONS

We proposed to use data mining algorithms to build models, aiming to predict if the client will subscribe (yes/no) a term deposit. The business question we would like to answer was whether or not the customer opted term deposit based on their demographic information and bank’s marketing activities.

Our objective was to find the best model to help bank to predict a new customer’s subscription behavior based on customer’s demographic information, and to seek the most helpful attributes, so that bank could pay more attention and put more effort on customers who consist with those factors.

DATA DESCRIPTION

The data set was in the form of CSV. Below are the details of all the variables.

Variables		Type	Description	Data Sample
Bank Client Data	1 - age	numeric		
	2 - job	categorical	type of job	"admin.","unknown","unemployed","management","housemaid","entrepreneur","student","blue-collar","self-employed","retired","technician","services"
	3 - marital	categorical	marital status	"Married", "divorced", "single" ("divorced" means divorced or widowed)
	4 - education	categorical		"unknown", "secondary", "primary", "tertiary"

	5 - default	binary	has credit in default?	"yes", "no"
	6 - balance	numeric	average yearly balance, in euros	
	7 - housing	binary	has housing loan?	"yes", "no"
	8 - loan	binary	has personal loan?	"yes", "no"
Related with The Last Contact of The Current Campaign	9 - contact	categorical	contact communication type	"unknown", "telephone", "cellular"
	10 - day	numeric	last contact day of the month	
	11 - month	categorical	last contact month of year	"jan", "feb", "mar", ..., "nov", "dec"
	12 - duration	numeric	last contact duration, in seconds	
Other Attributes	13 - campaign	numeric	number of contacts performed during this campaign and for this client (includes last contact)	

	14 - pdays	numeric	number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted)	
	15 - previous	numeric	number of contacts performed before this campaign and for this client	
	16 - poutcome	categorical	outcome of the previous marketing campaign	"unknown", "other", "failure", "success"
Output Variable (desired target)	17 - y	binary	has the client subscribed a term deposit?	"yes", "no"

DATA PRE-PROCESSING AND DESCRIPTIVE ANALYSIS

After understanding the attributes, we cleaned the dataset with the respective algorithms in mind. And the descriptive analysis were conducted on data set after data cleaning works.

DATA CLEANING

We calculated the info gain rates for all variables. Following picture shows the rates.

```
> infoGain
       default        marital      education           loan
"0.0004244884" "0.0030314170" "0.0037483854" "0.0037949345"
      campaign            day        balance            job
"0.0061513095" "0.0081193541" "0.0082252816" "0.0119228893"
       housing            age       previous        contact
"0.0139277448" "0.0173396802" "0.0178693026" "0.0196593713"
         month          pdays       poutcome       duration
"0.0351313059" "0.0371739348" "0.0424112545" "0.1037161265"
```

We removed some variables that had low info gain rates – "day", "default", "education", "marital", "loan", "campaign". Afterwards, we discretised some remaining attributes and also deleted some attributes for future analysis. Our actions for left variables see below.

For the variable "Age", we turned numbers into categorical groups based on the most widely used age segments in marketing area, which were "18-24", "25-34", "35-44", "45-54","55-64", "65+".

For the variable "Balance", since it had both positive and negative values, this variable was not capable of log transformation. After comparing the number of instances in different balance ranges, we decided to cut "Balance" into groups evenly by 10 percentiles.

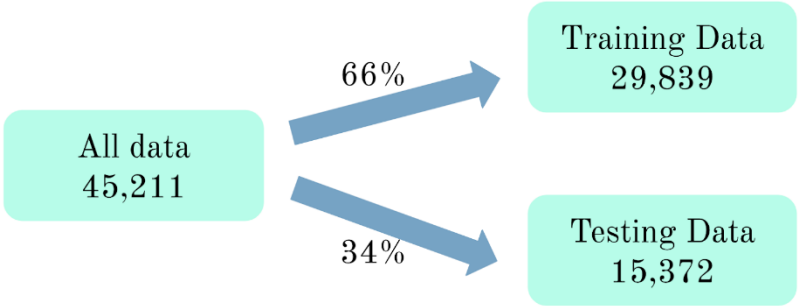
For the variable "Duration", it was highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration was not known before a call is performed. Also, after the end of the call "y" was obviously known. Thus, this input was discarded since our intention was to have a realistic predictive model.

For the variable "Pdays", we divided it into 6 categories, "not previously contacted", "contacted in last 1 month", "contacted in last 3 months", "contacted in last 6 months", "contacted in last 1 year", and "contacted 1 year before".

For the variable "Previous", we split it into 4 categories, "contacted 0 times", "contacted 1-3 times", "contacted 3-10 times", and "contacted more than 10 times"

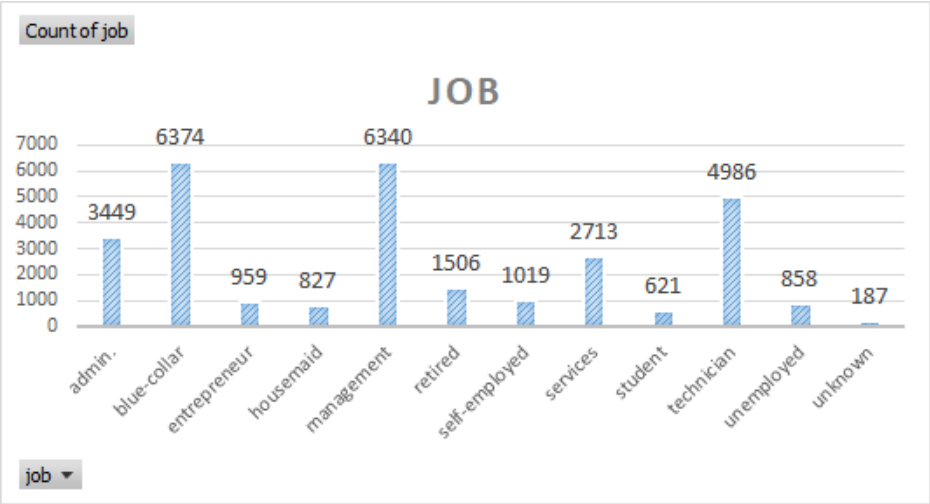
For the variable "Contact" and "Poutcome", there were to many unknown values in these 2 variables, so we decided to remove them.

In the end, we divided the whole dataset into training dataset (66% of the whole set) to build models and testing dataset (34% of the whole set) to predict the target variable and test overfitting. See the picture below.

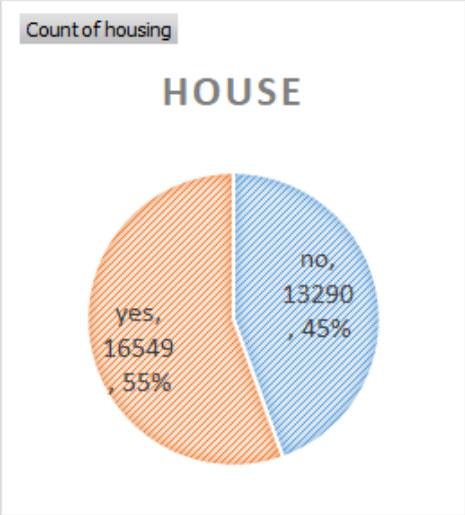


DESCRIPTIVE ANALYSIS

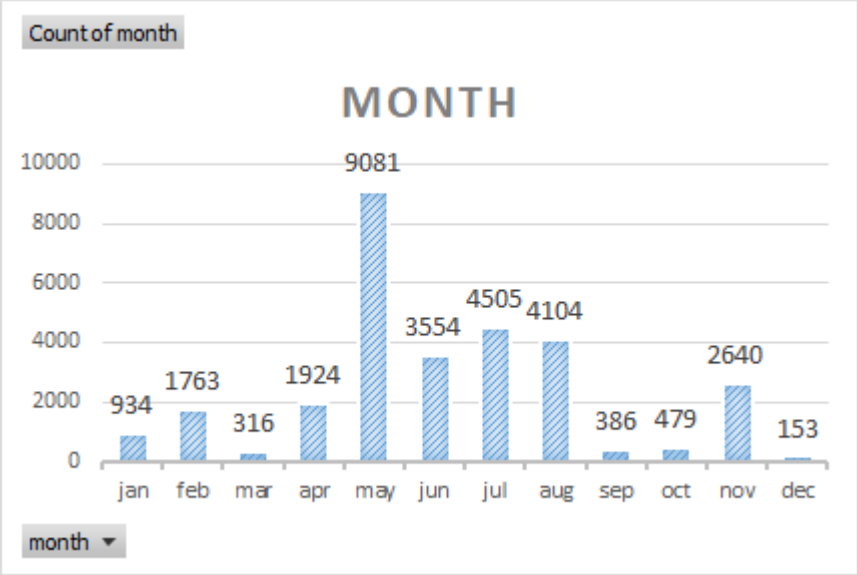
After data cleaning, we did descriptive analysis on our training data and obtained the distribution of remaining attributes.



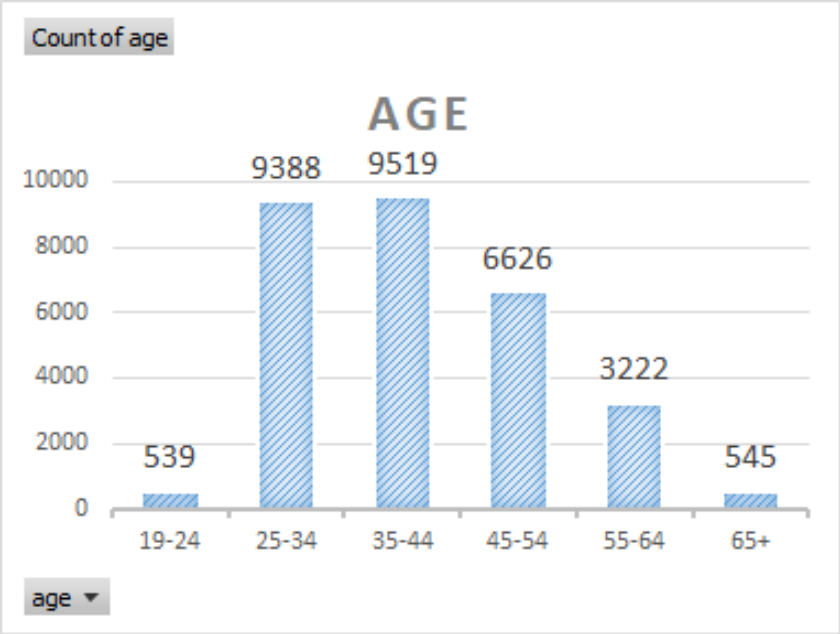
Above is the demographic information of clients. Regarding the type of job, “blue-collar”, “management” and “technician” are the jobs that most clients do.



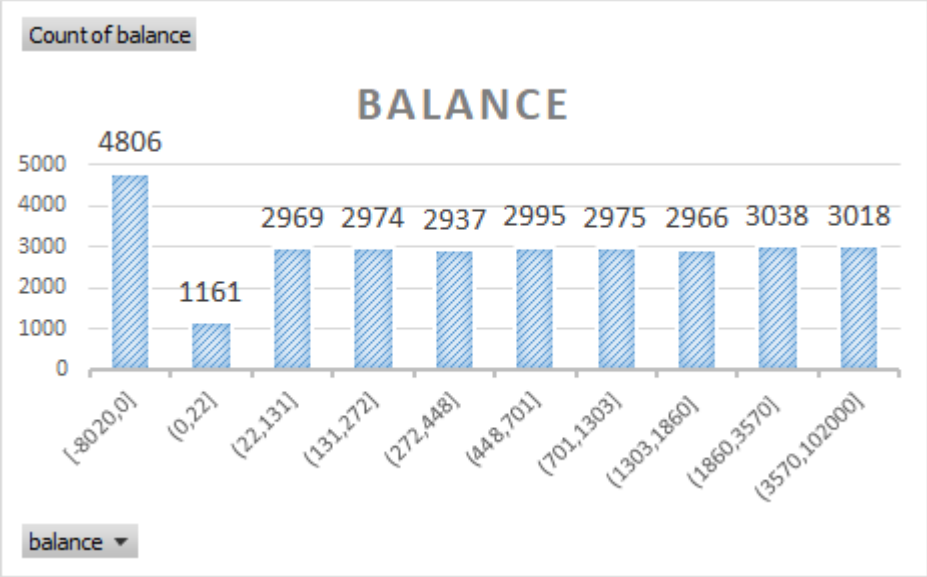
“House” indicates if a client has housing loan. We can see that 55% of all the clients have housing loan while 45% do not.



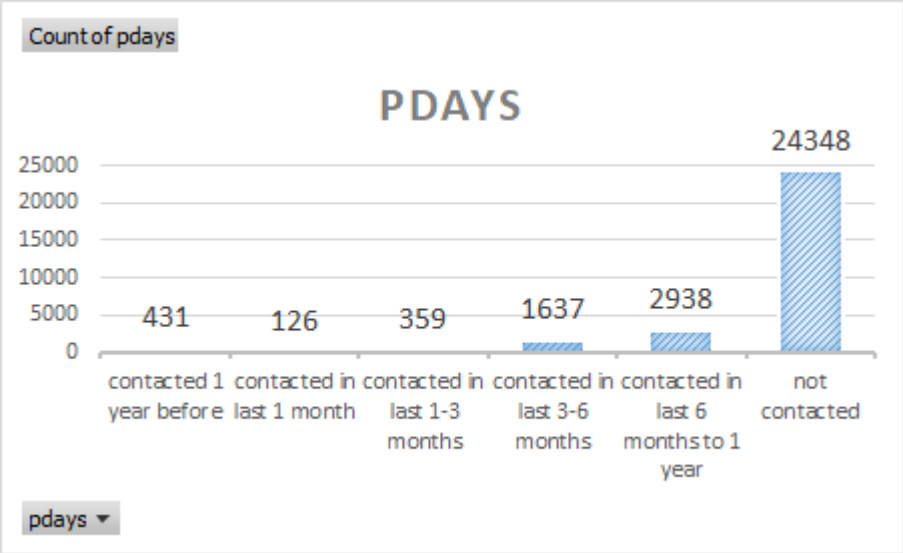
The variable “month” means “last contact month of year”. As we can see, most clients were contacted in May.



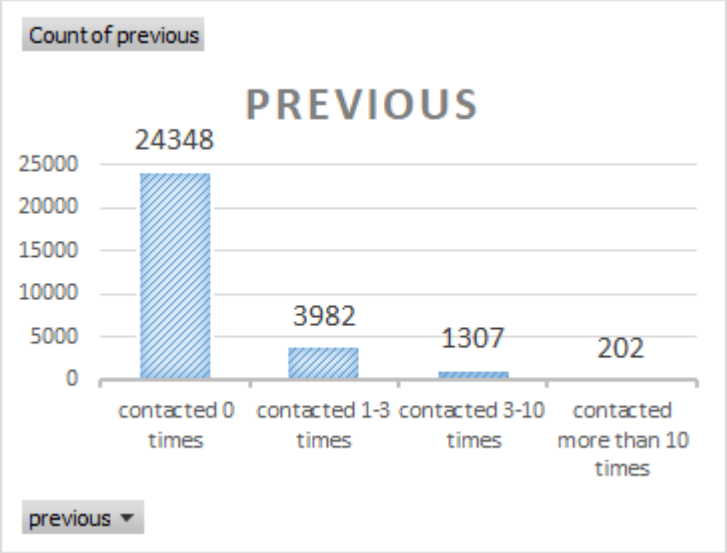
The distribution of age variable tells us that most our clients fall in the age range of 25 to 34, 35 to 44 and 45 to 54.



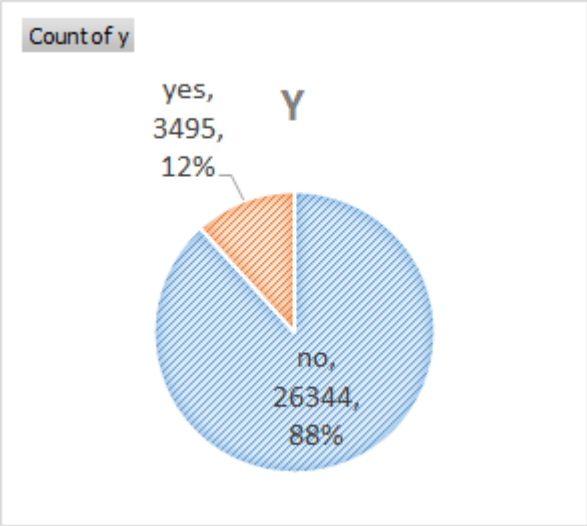
“Balance” is the average yearly balance of client, in euros. After discretion, we can see that more than 4800 clients are unable to make ends meet.



“pdays” means number of days that passed by after the client was last contacted from a previous campaign. As we can see, more than half clients are not contacted for a previous campaign.



Previous is number of contacts performed before this campaign and for this client. Similar to the "pdays", most than half clients are not contacted for a previous campaign.



For our target variable y, the distribution is very imbalanced, 88% clients didn't subscribe the term deposit while only 12% subscribed.

From those graphs, we found that most of attributes in this data set had uneven distributions. So we would take this situation into account for later analysis and evaluation.

DATA ANALYSIS

We applied association rules mining and five classification algorithms on the training data to find the relations between attributes and attribute "y". We were going to find the best model of each algorithms. In addition, based on the models we built, we applied test data to do predictions. Since the test data came from the original data set, we had already known customer's answers. Compare the real answers and our prediction results, we could test the overfitting and determined the best model with highest accuracy rate and F1 score among five classification models.

1. CLASSIFICATION

In this part, five classification models (SVM, random forest, kNN, decision tree and Naive Bayes) were built on the training data in RStudio to predict whether customers would subscribe a term deposit. In order to measure the model, some indexes, like accuracy rate, precision, recall and F1 score, were calculated. Accuracy rates were calculated by 10-fold cross validation. In this scenario, around 88% of customers rejected to subscribe a term deposit. If all prediction results were "no", the accuracy would be 88%, which was a high accuracy rate literally. The accuracy rate was not the most convincible index in this case. Instead, F1 score evaluated each models more objectively. Hence models were evaluated by both accuracy rates and F1 scores. Following content tells details models mentioned before.

SVM

We built 11 SVM models, differentiating kernel types and misclassification rates (represented as C). Kernel types were rbfdot, polydot, tanhdot, vanilladot, laplacedot, besseldot and anovadot. Cost were either 3 or 5. Here are the results.

Accuracy		Kernel					
		polydot	tanhdot	vanilladot	laplacedot	besseldot	rbfdot
Cost	3	88.22%	79.47%	88.22%	88.37%	88.41%	88.38%
	5	88.15%	/	88.20%	**88.37%**	88.34%	88.15%

Precision		Kernel					
		polydot	tanhdot	vanilladot	laplacedot	besseldot	rbfdot
Cost	3	0.50	0.11	0.50	0.88	0.69	0.78
	5	0.50	/	0.50	0.92	0.72	0.80

Recall		Kernel					
		polydot	tanhdot	vanilladot	laplacedot	besseldot	rbfdot
Cost	3	0.05	0.11	0.05	0.24	0.15	0.20
	5	0.05	/	0.05	0.36	0.18	0.24

F1 score		Kernel					
		polydot	tanhdot	vanilladot	laplacedot	besseldot	rbfdot
Cost	3	0.08	0.11	0.08	0.38	0.24	0.31
	5	0.08	/	0.08	**0.52**	0.29	0.36

The model with "besseldot" kernel and a cost of 3 returned the highest accuracy, while the model with "laplacedot" kernel and a cost of 5 derived the highest F1 score. The accuracy of besseldot model is 88.41% and the F1 score of this model was 0.24. On the other hand, the accuracy of laplacedot model was 88.37% and the F1 score is 0.52. Considering the huge gap between two F1 scores, we thought that the model with "laplacedot" kernel and 5 of cost was the best one.

RANDOM FOREST

We built two random forest models. One model had a number of tress equalled to 100 while one equalled to 500. The following table shows the results.

		Accuracy	Precision	Recall	F1 score
Trees	100	88.25%	0.95	0.31	0.47
	500	**88.35%**	0.95	0.32	**0.48**

By comparing 100 tress and 500 trees, the 500 trees had better performance. The accuracy for 500 trees was 88.35%. And F1 score was 0.4815052.

KNN

We conducted two kNN models – one with k = 3 and on with k = 5. The following table shows the result.

		Accuracy	Precision	Recall	F1 score
K	3	**91.40%**	0.94	0.70	**0.80**
	5	87.65%	0.97	0.69	0.78

Comparing to k = 5, the k =3 model has better performance. When the K = 3, the kNN model achieved the highest accuracy as 91.3972%. The precision equalled to 0.9399 while the F1 score was 0.7996.

DECISION TREE

Two decision tree models were built by “RWeka” package depended on different confidence rates. Here are the results.

		Accuracy	Precision	Recall	F1 score
Confidence	0.3	88.52%	0.65	0.11	0.19
	0.5	**88.17%**	0.74	0.26	**0.38**

The F1 score of model with 0.5 confidence was nearly two times of that of model with 0.3 confidence. In this case, model with 0.5 confidence had better performance. The best decision tree model was a pruned tree with 0.5 confidence. The accuracy of that model was 88.1698%. The F1 score was 0.380102.

NAIVE BAYES

We built Naïve Bayes model on default settings. The accuracy of model was 87.2616 %. And the result of F1 score was 0.349689.

SUMMARY

Here is a table comparing the best model of each algorithm.

		Accuracy	Precision	Recall	F1 score

Model	SVM	88.37%	0.92	0.36	0.52
	Random Forest	88.35%	0.95	0.32	0.48
	kNN	**91.40%**	**0.94**	**0.70**	**0.80**
	Decision Tree	88.17%	0.74	0.26	0.38
	Naïve Bayes	87.26%	0.44	0.29	0.35

Clearly, the kNN model with k = 3 was the best model among all models, because of its highest accuracy and F1 score.

2. ASSOCIATION RULE MINING

In this part, association rule mining was used to find the frequent pattern and rules that could predict the occurrence of subscription based on the occurrences of other items.

Firstly, we would like to get the combination of variables with highest association between them and successful subscription.

	lhs	rhs	support	confidence	lift	count
[1]	previous=contacted 0 times	{y=yes}	0.0748507	0.09157579	0.7827834	3384
[2]	pdays=not contacted	{y=yes}	0.0748507	0.09157579	0.7827834	3384
[3]	pdays=not contacted, previous=contacted 0 times	{y=yes}	0.0748507	0.09157579	0.7827834	3384
[4]	housing=no	{y=yes}	0.07418713	0.16702355	1.4277056	3354
[5]	housing=no, previous=contacted 0 times	{y=yes}	0.0464499	0.12377697	1.0580368	2100
[6]	housing=no, pdays=not contacted	{y=yes}	0.0464499	0.12377697	1.0580368	2100

From results above, the three most frequent pattern including y=yes(successful subscription) showed that customers who were not contacted before were most likely to subscribe. This conclusion was unreasonable because there were too many customers were not contacted before. So “not contacted” or “contacted 0 times” didn’t mean that customers met those criterions would subscribe.

	lhs	rhs	support	confidence	lift	count
[1]	previous=contacted 1-3 times	{y=yes}	0.02948463	0.2214286	1.892756	1333
[2]	housing=no	{y=yes}	0.07418713	0.1670236	1.427706	3354
[3]	job=management	{y=yes}	0.02877682	0.1375555	1.175815	1301
[4]	age=25-34	{y=yes}	0.03921699	0.1248240	1.066987	1773
[5]	housing=no, previous=contacted 0 times	{y=yes}	0.04644990	0.04644990	1.058037	2100
[6]	housing=no, pdays=not contacted	{y=yes}	0.04644990	0.12377697	1.058037	2100

From results above, people who had been contacted 1-3 times were most likely to subscribe. People who had no housing loan were on the second place to subscribe. And people whose job was management were the third most likely to subscribe.

	lhs	rhs	support	confidence	lift	count
[1]	previous=contacted 1-3 times	{y=yes}	0.0294846 3	0.2214286	1.892756	1333
[2]	housing=no	{y=yes}	0.0741871 3	0.1670236	1.427706	3354
[3]	job=management	{y=yes}	0.0287768 2	0.1375555	1.175815	1301
[4]	age=25-34	{y=yes}	0.0392169 9	0.1248240	1.066987	1773
[5]	housing=no, previous=contacted 0 times	{y=yes}	0.0464499 0	0.0464499 0	1.058037	2100
[6]	housing=no, pdays=not contacted	{y=yes}	0.0464499 0	0.1237769 7	1.058037	2100

The strongest rule was {previous=contacted 1-3 times} => {y=yes}, the second strongest rule was {housing=no} => {y=yes} and the third was {job=management} => {y=yes} .

From the results of arules of yes-subscription, we found that people who had been contacted 1-3 times, or did not have housing loan or had a job as management were more likely to subscribe the term deposit. This finding was reasonable and close to reality. If a customer had already been contacted before, he or she would have a change to know information about the bank program and might willing to subscribe it. In addition, a customer without housing loan might have more disposable money to pay for a term deposit. Moreover, a customer, working as a manager, might earn more salary to pay for a term deposit.

In next step, we put y=no in the right hand side to find what factors would lead to rejection of subscribing.

	lhs	rhs	support	confidence	lift	count
[1]	previous=contacted 0 times	{y=no}	0.742512 7	0.9084242	1.02877 8	3356 9
[2]	pdays=not contacted	{y=no}	0.742512 7	0.9084242	1.02877 8	3356 9
[3]	pdays=not contacted, previous=contacted 0 times	{y=no}	0.742512 7	0.9084242	1.02877 8	3356 9
[4]	housing=yes	{y=no}	0.513028 1	0.9229973	1.04528 2	2319 4
[5]	housing=yes, previous=contacted 0 times	{y=no}	0.413691 7	0.9357582	1.05973 4	1870 3
[6]	housing=yes, pdays=not contacted	{y=no}	0.413691 7	0.9357582	1.05973 4	1870 3

In the results above, the three most frequent patterns shared same feature – the customers were never contacted before.

	lhs	rhs	support	confidence	lift	count
[1]	housing=yes, month=may	{y=no}	0.251935 4	0.9432712	1.06824 2	1139 0
[2]	housing=yes, pdays=not contacted	{y=no}	0.413691 7	0.9357582	1.05973 4	1870 3

[3]	housing=yes, previous=contacted 0 times	{y=no}	0.4136917	0.9357582	1.059734	18703
[4]	housing=yes, pdays=not contacted, previous=contacted 0 times	{y=no}	0.4136917	0.9357582	1.059734	18703
[5]	month=may	{y=no}	0.2840080	0.9328006	1.056384	12840
[6]	housing=yes	{y=no}	0.5130281	0.9229973	1.045282	23194

We found that people who had housing loan and were lastly contacted month in May were least likely to subscribe. People who had housing loan and had not been contacted were second least likely to subscribe.

	lhs	rhs	support	confidence	lift	count
[1]	housing=yes, month=may	{y=no}	0.2519354	0.9432712	1.068242	11390
[2]	housing=yes, pdays=not contacted	{y=no}	0.4136917	0.9357582	1.059734	18703
[3]	housing=yes, previous=contacted 0 times	{y=no}	0.4136917	0.9357582	1.059734	18703
[4]	housing=yes, pdays=not contacted, previous=contacted 0 times	{y=no}	0.4136917	0.9357582	1.059734	18703
[5]	month=may	{y=no}	0.2840080	0.9328006	1.056384	12840
[6]	housing=yes	{y=no}	0.5130281	0.9229973	1.045282	23194

In summary, the strongest rule was *{housing=yes, month=may} => {y=no}* and the second strongest rule was *{housing=yes, pdays=not contacted} => {y=no}* , a rule showing the same thing as the third pattern showed.

After analysing rules with y=no set in right hand side, we found that people who had not been contacted in the last campaign, or had housing loan or were contacted in May were more likely to not subscribe the term deposit. If a customer had not been contacted before, he would had less possibility of knowing the bank program so that he would not subscribe it. Moreover, a customer with housing loan might not have disposable money to pay for a term deposit.

3. PREDICTION

In this part, we used test data to do predictions. Then we compared prediction results with the real values so that we could calculate accuracy rates, precisions, recalls and F1 scores of all models. Next we would find the best model among all models. Following are the results.

DECISION TREE (CONFIDENCE = 0.5)

	Predictions		
Targets		No	Yes
	No	13294	1495
	Yes	284	299

We applied test data on the decision tree model which confidence was 0.5. The accuracy rate was 0.8842701. The precision was 0.5128645. The recall was 0.1666667. The F1 score was 0.2515776.

Compared indexes calculated in prediction part with those returned from decision tree model built on train data (The accuracy of model is 88.1698%. The precision is 0.7394541. The recall is 0.255794. The F1 score is 0.380102.), we determined that the F1 score of prediction was much lower than that of training model. Even the accuracy of both prediction and training model were nearly the same, we could not say that decision tree model was a ideal model in this case.

NAIVE BAYES

	Predictions		
Targets		No	Yes
	No	12905	1323
	Yes	673	471

The accuracy rate of prediction was 87.01535%. Precision was 0.4117133. Recall was 0.2625418. F1 score was 0.3206263.

The accuracy rates, precisions, recalls and F1 scores of both prediction and training model (The accuracy of model was 87.2616 %. Precision for model was 0.4413432. Recall was 0.2895565. And F1 score was 0.349689) were close.

KNN (K=3)

	Predictions		
Targets		No	Yes
	No	13294	1495
	Yes	284	299

The accuracy of prediction was 97.72934%. Precision of prediction was 0.9870968. Recall was 0.6181818. F1 score was 0.7602484.

The performance of kNN training model was excellent with a accuracy of 91.3972%, a precision of 0.9399, a recall of 0.6958 and a F1 score of 0.7996. This model acted better while test data were applied to predict.

RANDOM FOREST (NTREE = 500)

	Predictions		
Targets		No	Yes
	No	13317	1541
	Yes	261	253

The accuracy of prediction was 88.28%. Precision of prediction results was 0.4922179. Recall was 0.1410256. F1 score was 0.2192374.

Recalling the F1 score – 0.4815052 – of random forest training model, we found that even the accuracy rates were nearly same – the accuracy of training model was 88.35%, the F1 score of

training model was almost two times of that of prediction results. Hence, there was a overfitting issue with training data.

SVM

	Predictions		
Targets		No	Yes
	No	13407	1607
	Yes	171	187

The accuracy was 88.43352%. Precision was 0.5223464. Recall was 0.1042363. F1 score was 0.1737918.

There was a huge gap between the F1 score of prediction results and that of training model. Although the accuracy rates were close of prediction results and of training model, SVM should not be considered as a optimal one.

THE BEST MODEL

Following is the table listing evaluation indexes of all prediction results.

		Accuracy	Precision	Recall	F1 score
Model	SVM	88.43%	0.52	0.10	0.17
	Random Forest	88.28%	0.49	0.14	0.22
	kNN	**97.73%**	**0.99**	**0.62**	**0.76**
	Decision Tree	88.43%	0.51	0.17	0.25
	Naïve Bayes	87.02%	0.41	0.26	0.32

Considering the highest accuracy and F1 score plus less overfitting, we believed that kNN was the best model overall.

CHALLENGES

The RWeka package had limited heap space, which stood in the way of processing data. RWeka could not help us to build models except Naïve Bayes and decision tree models. After searching other methods, we used some packages, such as “caret” and “kernlab”, to build models.

Setting parameters was also challenging. Because there was no specific directions to work on setting parameters, the solution was to focus on several models which had better initial performances and then dig deeper on parameters of these models in all aspects. For example, for SVM model, the project explored the aspects of kernel and then figured out the best one among 6 possible kernels. After a large amount of exploration, we finally got some satisfying results .

CONCLUSION

The business question we would like to answer is whether or not the customer opted term deposit based on their demographic information and bank’s marketing activities. What we have done is capable of predicting customer behaviour before the primary communication stage. By following our

suggestions, the bank should pay more attention to target customers who have higher probability to subscribe term deposit, and put less efforts on clients who have lower possibility to accept term deposit. Specific strategies are mentioned in the following paragraphs.

According to results returned from association rules, we suggests that the bank marketing team could focus on customers who have been contacted 1-3 times without housing loan or working as “management”. These two groups have higher possibility to subscribe the deposit service.

After comparing the accuracy, precision, recall and F1 on test set, we would like to recommend the kNN model to the bank for predicting future customers’ behaviour. The bank could focus and pay more attention on the customers who are predicted to answer “yes”. Moreover, the marketing department of the bank can take the right measures on existing customers while some demographic characteristics of them alter. For instance, if the occupation levels or loan status of clients, who were not classified as our potential customers, changed, the marketing department could promote the deposit program to those clients and specify customer demands.

However, despite on the existing demographic attributes, we think that the bank could provide more data related to marketing and communication methods. For example, the bank could specify how staffs contact customers. In this case, attribute “contact” includes the method of contacting customers. Nonetheless, there are too many missing values in this attributes that we cannot take any advantages on this attribute. If the bank could provide more details about communication methods, we could analyse those methods and reach efficiencies of all methods.