Predicting Customer Purchase to Improve Bank Marketing Effectiveness

Group 6
Sandy Wu
Andy Hsu
Wei-Zhu Chen
Samantha Chien

Instructor : Galit Shmueli
National Tsing Hua University

Executive Summary

A bank marketing dataset from UCI Machine Learning Repository was adopted for this project (https://archive.ics.uci.edu/ml/datasets/bank+marketing). The dataset is about a Portuguese banking institution with records of direct marketing campaign phone calls, and the final outcomes indicating whether success campaigns are also included in a binary format (yes/no). A success campaign indicates the customer has finally subscribed a term deposit at the end of the campaign. Our business goal here is to identify the elements for a success campaign, from which we can improve marketing effectiveness by targeting the right customers; that is, we want to find what kind of campaign strategies/history combined with what kinds of customers will bring about high propensities for subscribing term deposits. Via building predictive models outputting the propensities of subscribing, we can therefore increase revenues and lower labour costs by having more efficient marketing strategies without harming customer relationship.

The dataset includes 41,188 campaign records, in which demographic, credit, current/previous campaign, and social/economic data are included (a total of 19 columns for each customer). Campaign data may include contact types, times of contacts, contact durations, etc. The data mining goal here is to predict the last column in the dataset which is the outcome of the current campaign (yes/no). It's a predictive classification problem, and supervised models were built to solve this problem. Furthermore, propensities of subscribing output from our predictive models can be also utilized to do ranking of customers, which facilitates making most revenues out of performing direct campaigns on only a small portion of customers (this comes with a cost down on labouring).

Naïve Bayes, logistic regression, decision tree, and random forest were included as our predictive models, and naïve rule that output propensities of success by simply calculating the proportion of the success records in the training set was chosen as the benchmark. Naïve Bayes models can produce robust predictions if the predictors have small correlations, even with a simple architecture. Logistic regression models are simple but equipped with decent capabilities for predictions. Decision trees are easy to interpret and are capable of giving insights about the important features; random forest is an improved version of decision tree, which can produce really good and robust predictions. Models mentioned above were implemented and their performances were compared based on lift curves. Lift curves can be constructed by accumulating the recall of true success (identified by the final column) starting from the customers with high propensities of subscribing where the propensities were collected from our predictive models. If there is no clear distinction by observing lift curves, the model with the highest sensitivity (the ability to identify true subscribers) will be chosen as our final model (perhaps also with less implementation efforts needed). Random forest was the final model adapted. It produced the best result in terms of lift curve, and an accuracy of 78.96% was achieved with 0.64 in sensitivity. This model includes 75% of the true subscribers with only contacting the top 40% of the total customers in terms of subscribing propensity.

Last but not least, this dataset contains many categorical columns and most of them have really low correlation between each other. If more ordinal/numerical columns were included, we can have better results and can lead to more cost down on labour costs and much more precision for conducting direct campaigns, lowering the chances of harming precious customer relationships.

I. Problem Description

Bussiness Goal

Our dataset came from a Portuguese bank which conducted direct marketing campaigns to promote term deposits to their customers. The common problem could be that the bank keeps re-calling the ‘’wrong’’ customers (the ones don’t need term deposits) and it may cause high labor costs and be susceptible of harming customer relationship. So, our business goal is to improve marketing effectiveness by targeting the right customers.

In this project, the stakeholders are the bank marketing team, bank employees, and customers. The bank would benefit from this solution where lower marketing costs and less probability of wrong marketing targets can be achieved without the risk harming customer relationship. The customers of the bank will receive more precise solutions to their needs. However, this solution may have a cold start problem where new customers have no previous marketing records and thus reduce the size of available features. Nonetheless, the success of this solution will bring about layoffs of the bank employees since they can spend less human labor on conducting marketing due to the decreased size of candidate customers for marketing, which is a direct consequence of the solution that will filter out a small portion of customers who really need the term deposits. Also, it also sets up barriers between customers with different credit scores, widening the poor and rich disparity and harming the mispredicted ones.

Data Mining Goal

The data mining goal is that we want to use customer information (demographic information, records of last marketing event and social/economic context attributes) to predict whether a certain customer will subscribe a term deposit. It is a predictive, forward-looking, and supervised task since the model is trained on the previous records and predict the probability of customers’ subscribing a term deposit when we have a new customer coming in.

The method we used are Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, and we utilized Synthetic Minority Over-sampling TEchnique (SMOTE) to deal with imbalanced data (the number of the non-subscribers is about seven times more than the subscribers’). For performance evaluation, we used mainly lift curves/charts along with Receiver Operating Characteristic (ROC) curves and sensitivity to evaluate our results.

II. Data Description

The target dataset was downloaded from UCI Machine Learning Repository, containing 41,188 rows and 21 columns with customer demographic information, credit status, current/previous campaign records, and social/economic indices. The outcome variable is column “y”, which indicated whether the customer subscribed or not in the current campaign. The imbalance ratio defined by the ratio of number of no and yes of column “y” is 790.27%, which will severely impacts the prediction capability without over-sampling. The figure below is a screenshot of the dataset for only a couple of samples.

age	job	marital	education	default	housing	loan	contact	month	DOW	duration	campaign	pdays	previous	poutcome
60	technician	married	basic.4y	no	yes	no	cellular	oct	tue	595	2	6	1	success
70	retired	married	basic.4y	no	yes	no	cellular	oct	tue	76	2	999	0	nonexistent
59	housemaid	married	basic.4y	no	yes	no	cellular	oct	tue	518	2	6	1	success
53	entrepreneu	married	basic.4y	no	yes	no	telephone	oct	tue	251	2	999	0	nonexistent
30	technician	married	professiona	no	yes	no	cellular	oct	tue	427	1	999	0	nonexistent
24	managemer	single	university.d	no	yes	no	cellular	oct	tue	676	2	999	1	failure

emp.var.ra	cons.price	cons.conf	euribor3m	nr.employ	y
-3.4	92.431	-26.9	0.737	5017.5	yes
-3.4	92.431	-26.9	0.737	5017.5	no
-3.4	92.431	-26.9	0.737	5017.5	yes
-3.4	92.431	-26.9	0.737	5017.5	no
-3.4	92.431	-26.9	0.737	5017.5	yes
-3.4	92.431	-26.9	0.737	5017.5	yes

Figure 1. Screenshot of bank marketing data. This is the raw data and will be further processed in the following steps.

III. Breif Data Preperation Details

First, we explored the dataset to see whether there are missing values and found that there were no missing values, which is illustrated in **Figure A1**.

There was a column we cannot use: duration. duration represents the time duration (in seconds) of the last contact, and it should not be used in building prediction model since if duration is 0, then y = “no” for sure. As a result, we excluded this column from building our models. On the other hand, categorical columns such as 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week' , 'poutcome' were transformed into dummy variables, and the number of columns increased from 21 to 64.

Observing the distribution of values of pdays, we found that only 3.6% of the customers got contacted recently with the last contact occured in the past 27 days, and 96.3% of them was contacted rather a long time ago (more than 27 days). Hence, pdays was binned to two categories where 1 means that the customer was contacted recently and 0 otherwise.

All columns were normalized (subtracting their mean and divided by their standard deviations) before feeding into the model for training (resulting columns have means and standard deviations 0 and 1, respectively). Through normalization, we put all column on the same position for comparison (or columns such as income will dominate the predictions).

Data Visualization

We have done several visualization/exploratory analysis before really going into building our models. The figures are included in the Appendix.

IV. Data Mining Solution

We split the dataset into 70% for training set and 30% for test set. We first build logistic regression model and found the sensitivity to be only 0.22. This is because the positive samples (subscribers) are much less compared to the non-subscribers. We then tried SMOTE algorithm and trained the model again, and the sensitivity increased to 0.63. Various models were also trained and their results were compared in **Table 1**. Models we adopted were naïve Bayes, logistic regression, decision tree and random forest.

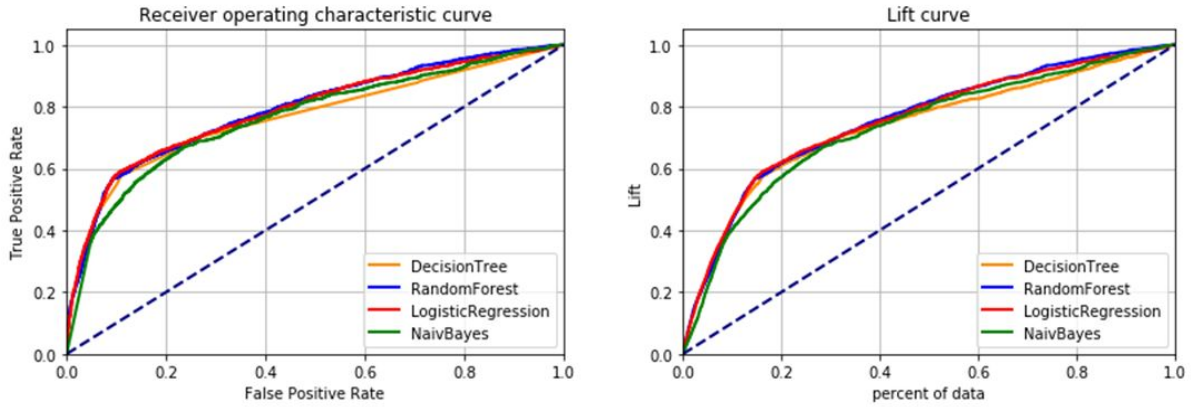

Figure 2. ROC curves (left) and lift curves (right) of four models: decision tree, random forest, logistic regression, and naïve Bayes. Those curves were very close, so we resorted to the model with the least sensitivity and chose as the best model.

Methods	Accuracy	Sensitivity	Specificity	AUC	F1
Logistic Regression	81.57%	0.63	0.84	0.79	0.87
Decision Tree	85.60%	0.58	0.84	0.77	0.87
Random Forest	78.96%	0.64	0.81	0.79	0.87
Naïve Bayes	63.45%	0.75	0.62	0.76	0.80
Naïve (Benchmark)	88.73%	0	1		

Table 1. Model performances. Naïve Bayes produced the highest sensitivity of 0.75 and was selected as the best model.

V. Conclusions

In this study, we applied different data mining methods to filter out the most likely subscribers. After evaluating our solutions with lift curves, we suggested that random forest showed the best results. The bank can recall 75% of the subscribers by contacting only top 40% customers in terms of propensity of subscribing. The followings are some operational and data collecting recommendations:

1. The outcome of current campaign is strongly influenced by previous campaigns. We suggested that the bank can provide more information between them.
2. Subscribers tend to be older than non-subscribers. Therefore, the bank can launch different promotions according to customer ages.
3. The result might be better if we can have a bigger data size to train with. Also, there seems to have low correlations among columns, and most correlated columns are from the campaign data. In addition to campaign data, we suggest that including more personal financial records such as income, credit card bills, payments etc. may bring about improvements in predictions.
4. Our result was mostly trained from dummy columns. We suggest that including ordinal/numerical columns may bring about improvement in predictions.

Appendix

```
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
age               41188 non-null int64
job               41188 non-null object
marital           41188 non-null object
education         41188 non-null object
default           41188 non-null object
housing           41188 non-null object
loan              41188 non-null object
contact           41188 non-null object
month             41188 non-null object
day_of_week       41188 non-null object
duration          41188 non-null int64
campaign          41188 non-null int64
pdays             41188 non-null int64
previous          41188 non-null int64
poutcome          41188 non-null object
emp.var.rate      41188 non-null float64
cons.price.idx    41188 non-null float64
cons.conf.idx     41188 non-null float64
euribor3m         41188 non-null float64
nr.employed       41188 non-null float64
y                 41188 non-null object
```

Figure A1. No missing values in the dataset.

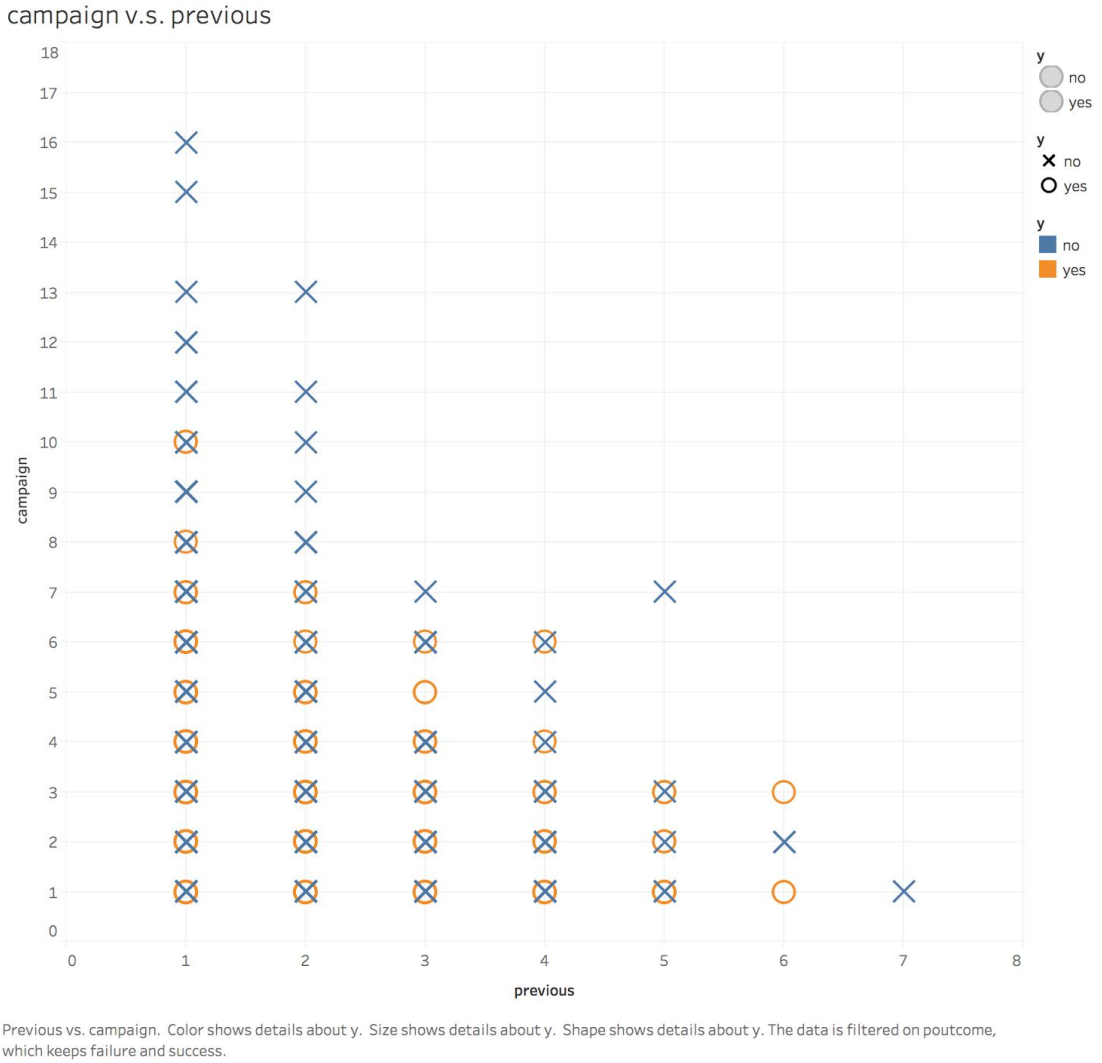

Figure A2. Circle view of campaign with respect to previous. campaign and previous are both essential for predictions: a pattern can be observed (reciprocals) and we can tell that customers with more contacts before may make a decision sooner.

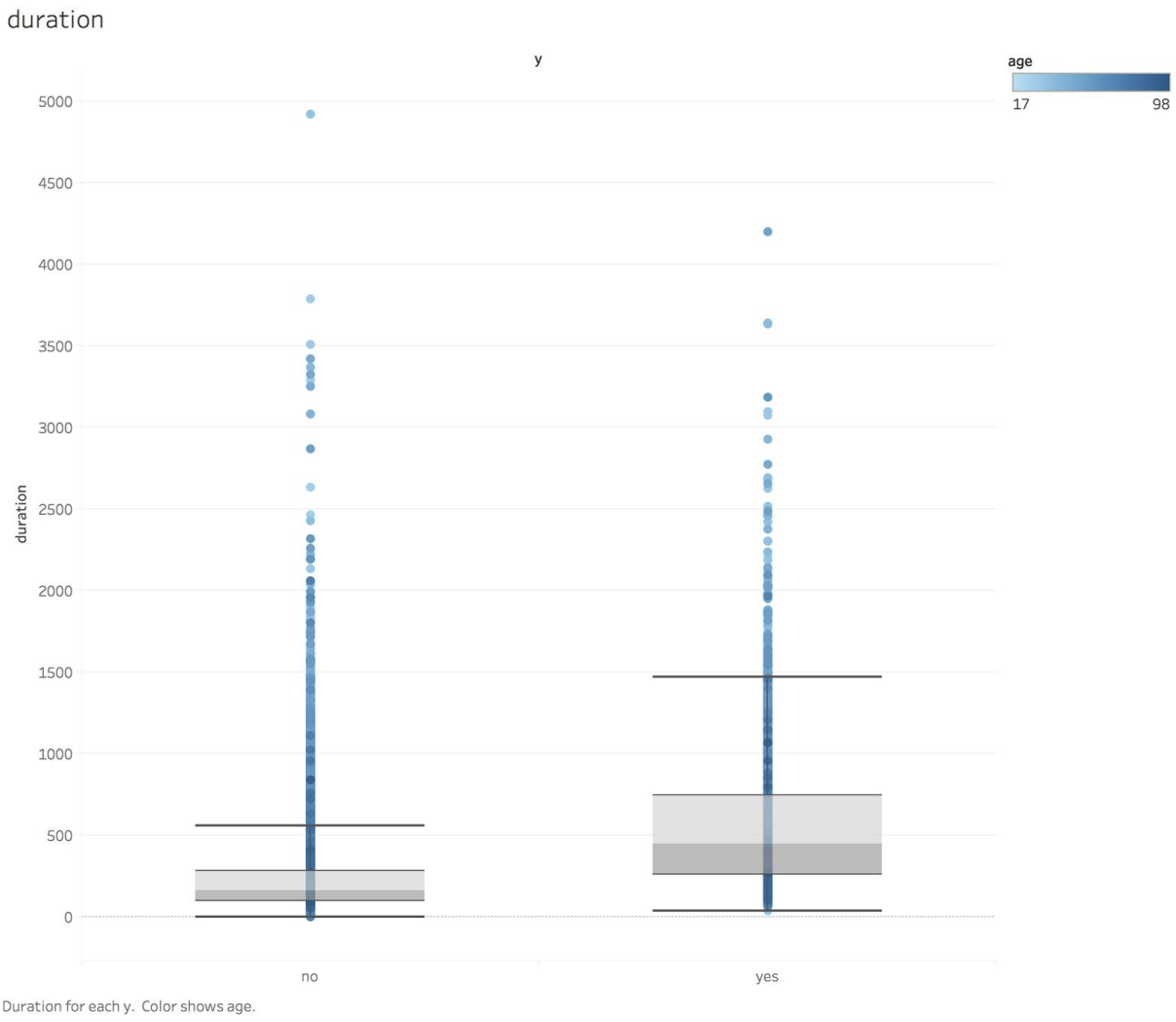

Figure A3. Boxplot of duration with respect to outcome.

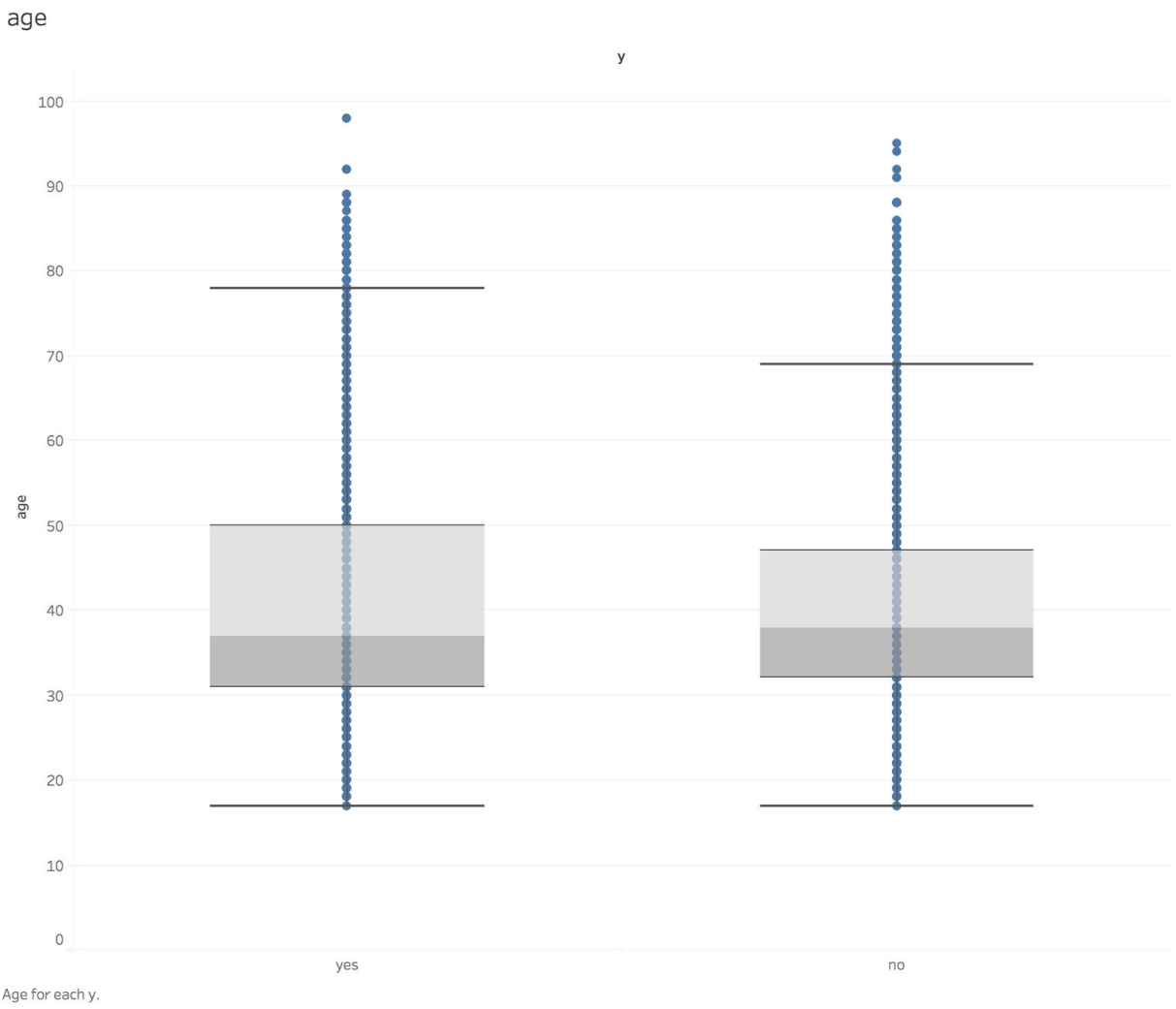

Figure A4. Boxplot of age for each outcome. This plot shows a significant distinction between subscribers and non-subscribers. Customers that subscribe have large deviations in their ages, and with a lot of portion being elder ones compared to the other group. This shows the fact that age may be a crucial predictor.

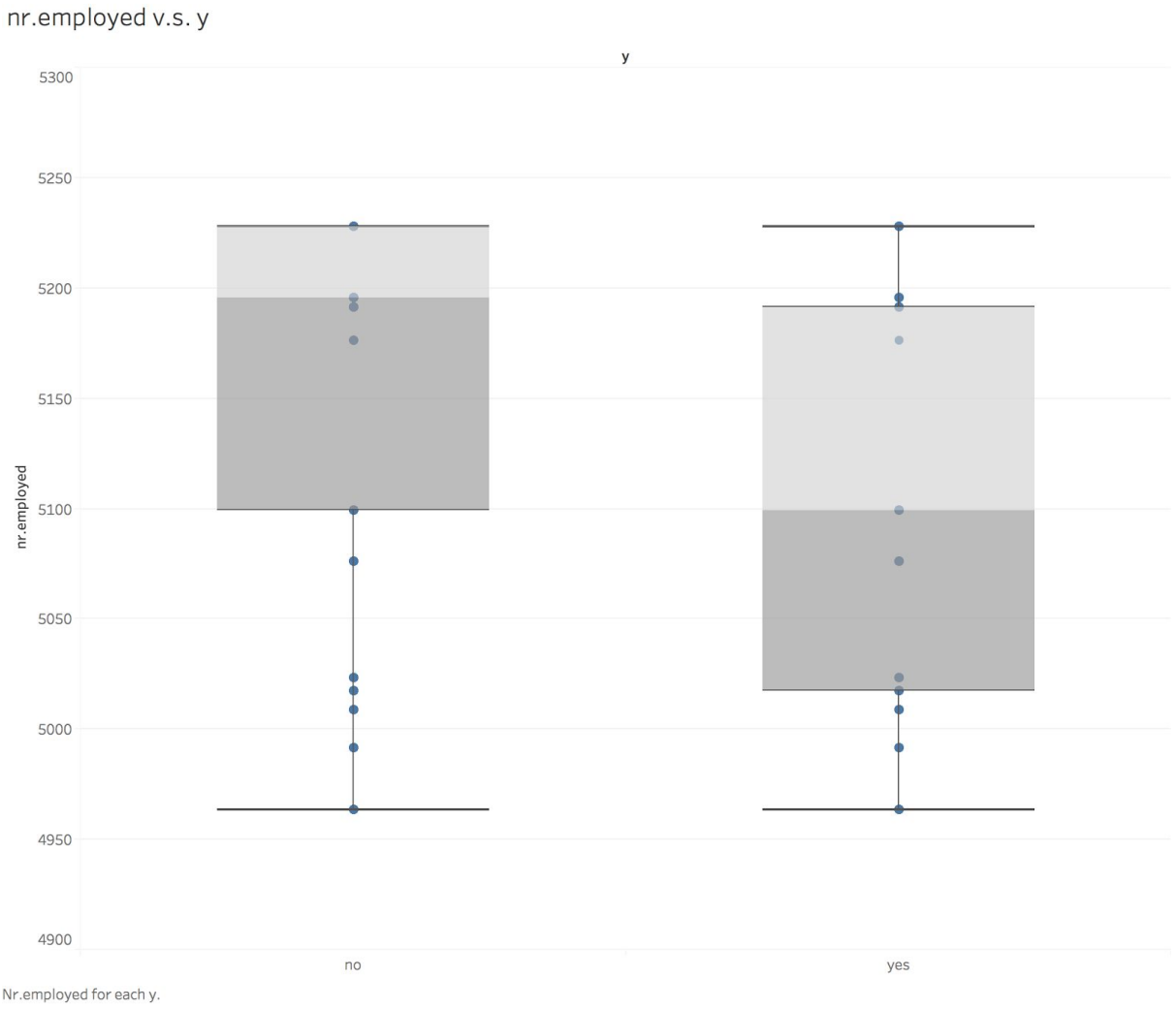

Figure A5. Boxplot of number of employees with respect to outcome. The boxplot here shows consistency with a known fact in macroeconomics that when a lot of people have jobs and they will be more willing to spend their money and vice versa. For the times with less employment rates, people tends to keep their money in the bank. This also indicates that nr.employed is an essential predictor.

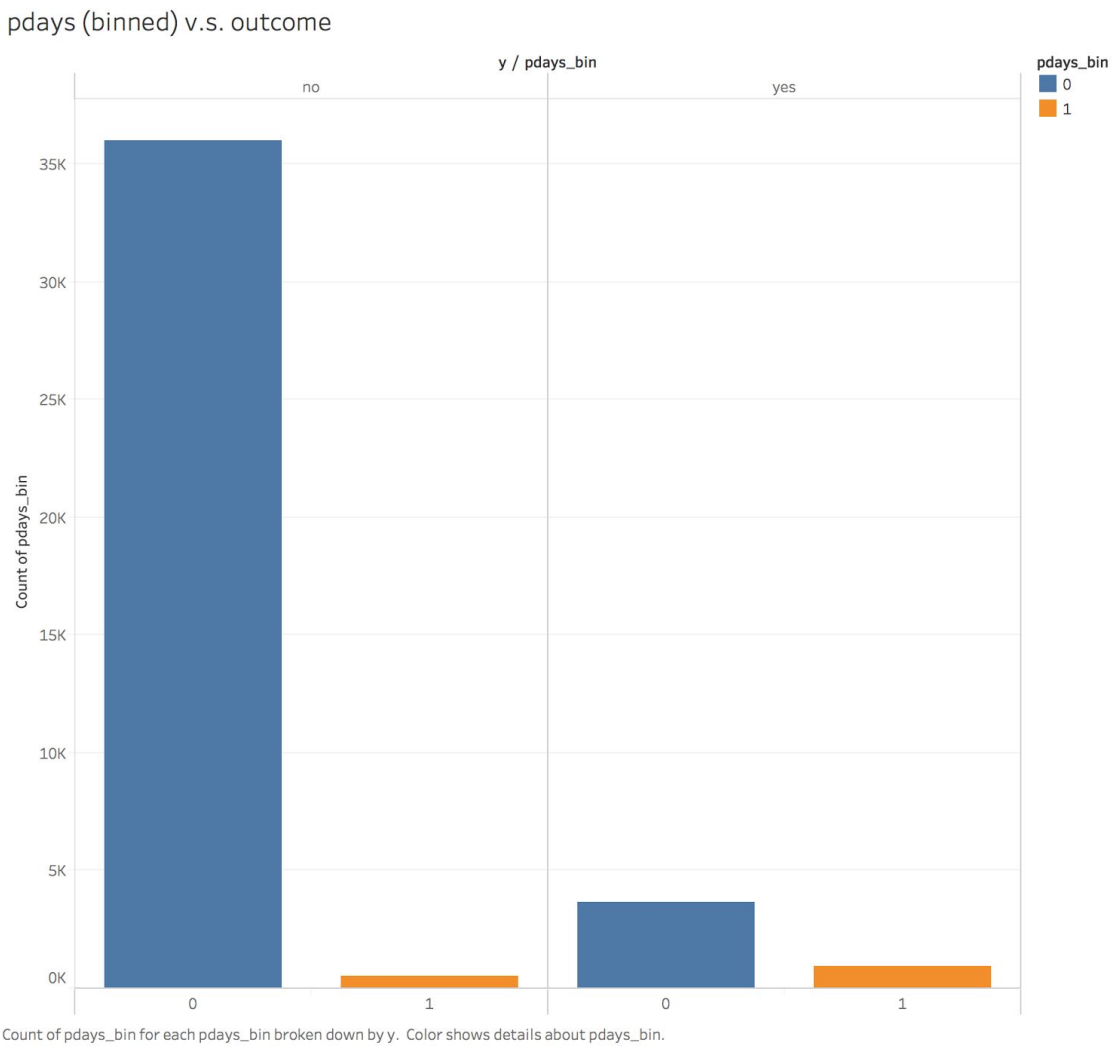

Figure A6. Side-by-side bar chart of pdays (binned) with respect to outcome. pdays has been binned into 0 (pdays == 999; contacted long time ago or even never contacted) and 1 (other cases; contacted currently). As we can observed from this figure, for the subscribers, there are still some customers that were not contacted before this campaign. Also, the ratio of number of 0's and 1's for binned pdays is larger for the non-subscriber group. pdays is crucial for building predictive models.

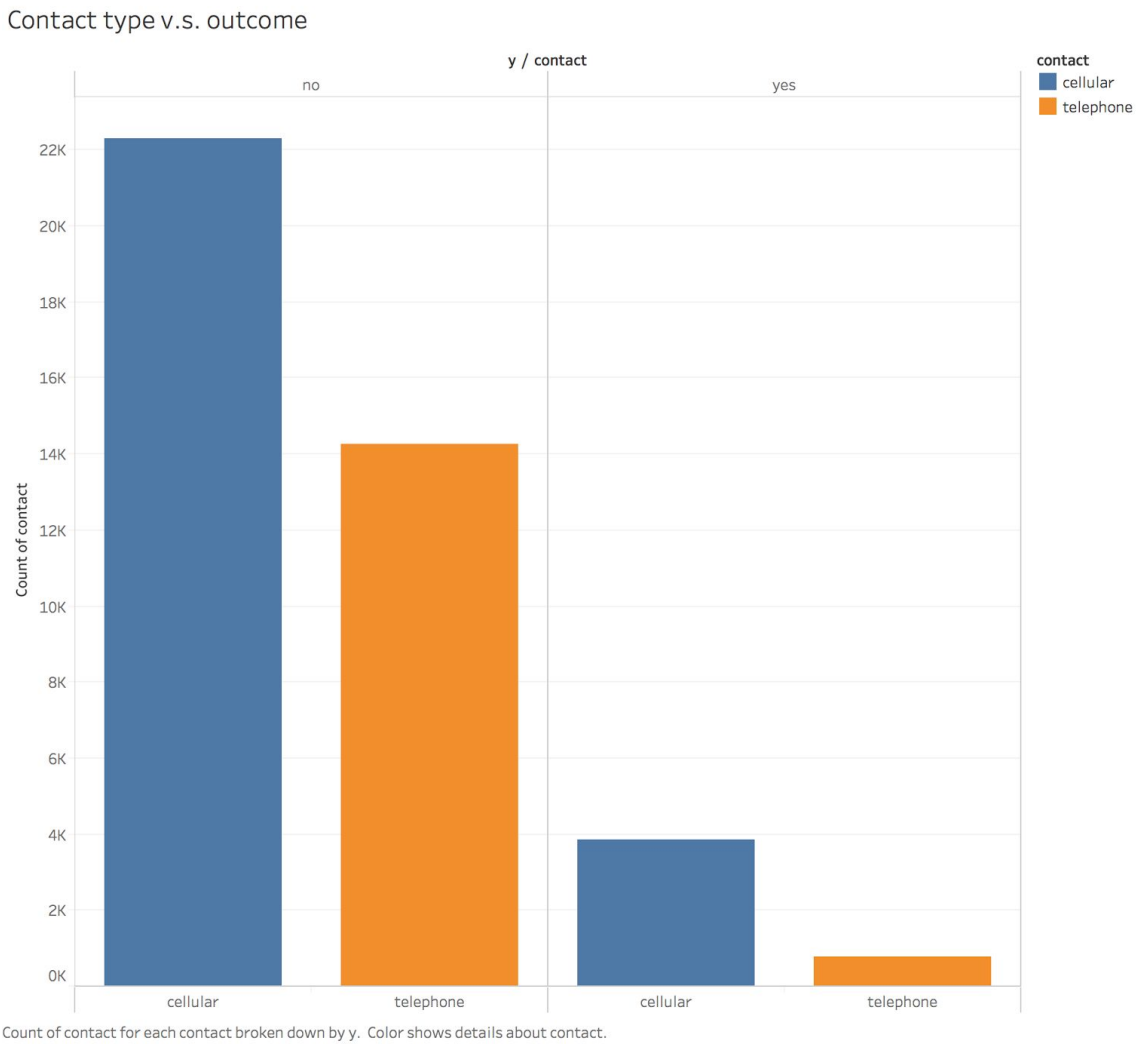

Figure A7. Side-by-side bar chart of contact and outcome. For the subscribers, larger portion of them were contacted via telephone compared to the other group. The predictor contact can be observed to be one of the important features in decision trees (branch node near top).

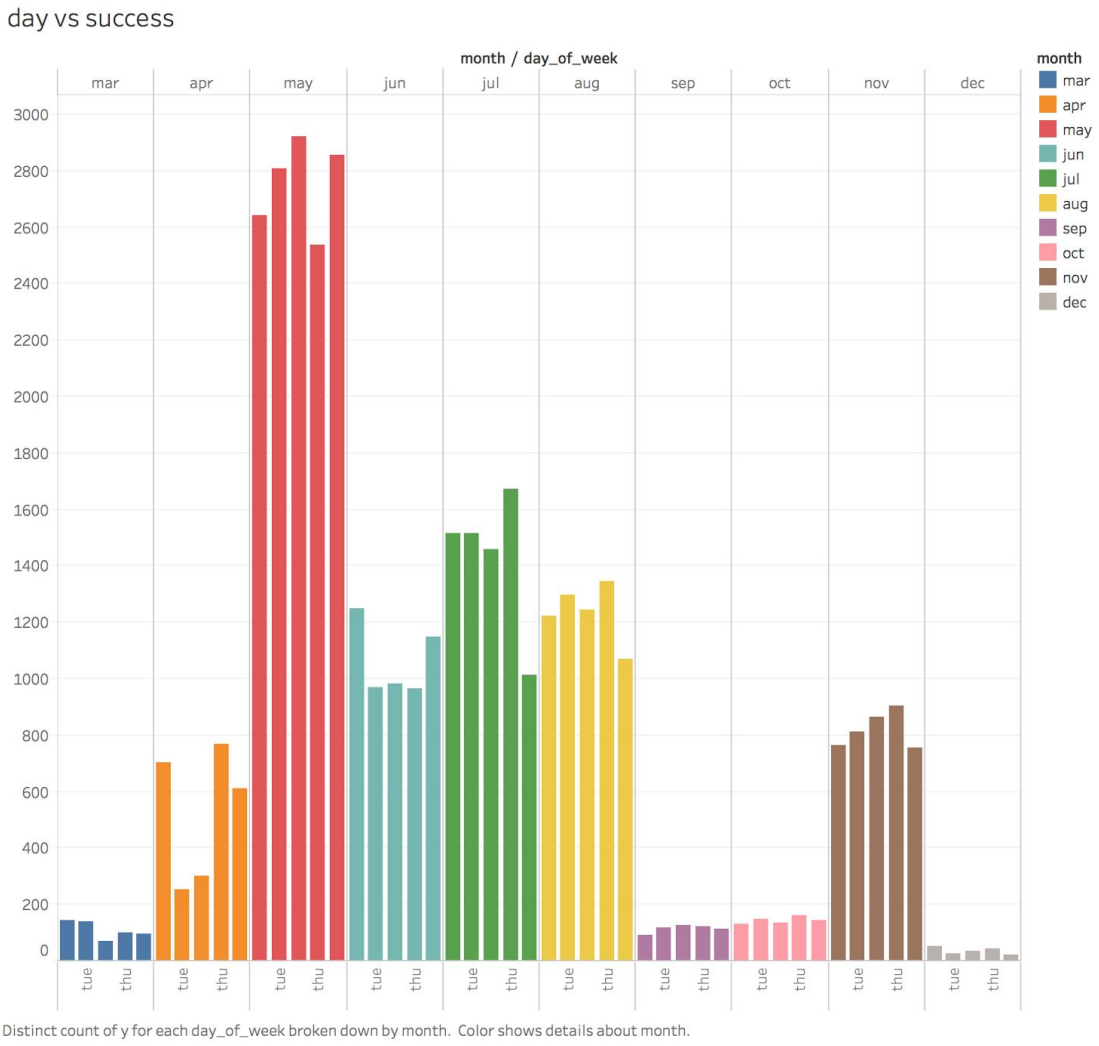

Figure A8. Bar chart for month and DOW. The largest portion of conducted campaigns are located in May. We can get some information of the campaign strategies of the bank from this figure.

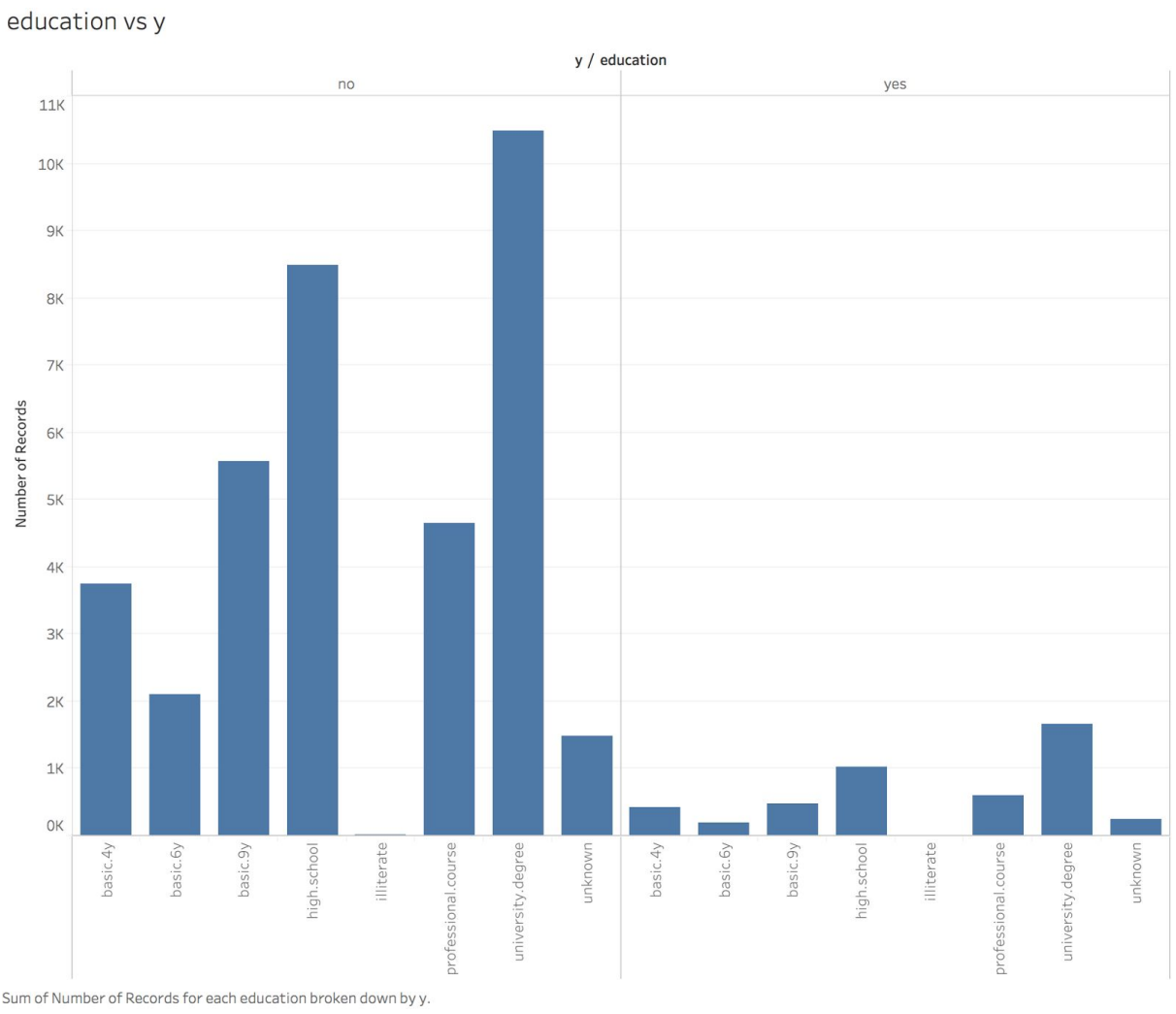

Figure A9. Bar chart of education and outcome. No prominent difference can be observed from this figure between two groups. Hence, education may not be an important feature.