Predictive Models of Customer Participation in Bank Campaign

Data Analytics Final Project
CIDM 6308 | Professor Dr. Liang Chen

Team members

Rahmat Begum
Bimal Dawadi
Bacari Herring
Catherine McGovern
Nathan Nauman
Sanjay Srivastava
Michael Tsapos

Executive Summary

Goal and Motivation

In the banking system, one of the ways to increase revenue is to offer additional services to its existing customer base. The bank's marketing department usually runs a campaign to provide a new service to target customers and encourage them to enroll in it. Despite the marketing team's effort to run a campaign, the lower rate of campaign conversion remains a concern. The main goal of this project is to find attributes and data mining models to predict campaign conversion. Based on our findings, banks will be able to create new campaigns to target a group of customers who are more likely to get a new service. Our results will also guide what factors in the campaign operation can influence conversion and adjust future campaign planning accordingly.

Method and Data

Data mining technology and business analytics can be used on this set to offer insight on a granular level into the effectiveness of the marketing efforts and the variables that can most impact a customer's conversion. The data set we've chosen is comprehensive, spanning a number of years and containing over 21 different variables. The data set was selected from the UCI Machine Learning Repository at archive.ics.uci.edu/ml/datasets.php. We cleaned up that dada and then after analyzing co-relation between attributes, we decided to use 16 variables for our analysis. We then split data into two subsets with 70/30 ratio as training set and prediction set. We proceeded to build two models- Decision Tree and Logistic Regression with our training set and then feed the prediction set into the models to get prediction results for campaign conversion.

Key Findings

The Decision Tree model performed slightly well compared to the other Logistic Regression model with an accuracy of 89% over 86%. Among all 16 attributes, Duration (the last contact duration) becomes the most weighted attribute followed by Marital Status in the Decision Tree model. The longer the duration the more likely the customer is likely to enroll into long Term CD. Married people are more likely to convert than divorced or single ones. The Logistic Regression Model showed multiple attributes with varying effects. So, we decided to use The Decision Tree model for future analysis. Overall, both models showed that average higher confidence level (0.753 and 0.886) indicating customers are more likely to not enroll.

Actionable Business Recommendations

We intend to use the findings generated by the research study to improve the campaign conversion rate. The result indicated that the way current campaigns are conducted is not resulting in great success. The call center needs to focus on increasing call duration for customers to improve conversion rate. The target demographic should include more married people to have better campaign performance.

Introduction

Industry background: Banking, Telemarketing

In the past, telemarketing has not been a significant method of direct marketing for banks seeking to sell their products. This is significant for our group because methods of direct marketing have become increasingly attractive as banking institutions face increased competition. These days, many banks offer similar products, hence marketing efforts have become a key aspect of cultivating new business (Ł. Piotr et al, 2019). Our data set, created by a Portuguese bank, records the efforts of the bank associates making phone calls to customers to persuade them to make long term deposits with the bank (Moro, 2014). A long-term deposit, also known as a CD (certificate of deposit), requires the buyer to invest a sum of money over a certain amount (i.e. $500) for a period (long term deposits typically remain in the account untouched for a year) (Fernando, 2021). Hence these are products that would only be attractive only for certain customers. Direct marketing allows businesses to efficiently reach customers most likely to convert. Selling additional products and services to existing customers is known as "cross-selling." A 2013 report from Deloitte discusses the challenges of banks attempting to cross-sell additional products to their existing customers. Most customers have only one or two products with a particular bank (i.e. a savings and a checking account), but tend to go outside of that bank to purchase additional products.

Business Question/Problem:

How can we build predictive models to offer insight on a granular level into the effectiveness of the marketing efforts and the variables that can most impact a customer's conversion?

This question is important because marketing efforts play a significant role in a bank's business activities. Furthermore, determining customers to target in the campaign helps to maximize the return on the bank's marketing investments—to manage time and costs incurred by the campaign itself. In other words, how can the bank achieve maximum revenue with minimal phone calls (Ladyzýnski et al, 2019).

Based on our research we've identified a variety of attributes that help us classify a customer likely to convert including campaign duration, job type, marital status, education and days since last contact. Job type, marital status and education are of particular note; Deloitte points out the importance of significant life events and how they can impact financial activity and investment (Deloitte, 2013). Additionally it is well established in the financial industry that demographic information is an important way to have a more complete picture of customer behavior. It these variable can also assist in clustering. But demographic information isn't enough (Alis, et al, 2000). Moro et al. confirmed the idea that the duration of the call can make a notable impact on the outcome of a customer contact. (Moro, 2014). The examination of attributes in the context of data mining is referred to as "Feature Engineering" (Domingos, 2012). In our research, we will make predictions using both a decision tree as well as logistic regression, two data mining methods that use classification to make predictions.

Thoughts on Decision Trees and Logistic Regression

Decision trees have been used in the past to uncover connections between attributes that can provide classification guidelines (Danning, 2012). Scholars have noted that the simplicity, er simplification tendencies, tend to be an issue with decision trees and that multiple decision trees themselves might be a necessity during the analysis process (Osei-Bryson, 2004). Indeed, our group explored multiple inputs (minimums for branch split and node size) in our decision tree formation. Indeed, fine tuning these inputs are essential to prevent over-fitting with our decision tree. Decision Trees can parse out complex relationships, however Logistic Regression relies on linear relationships for insights. (Neale, 2020). Logistic Regression as a classification model can be viewed analytically as a means of managing heterogeneity and homogeneity within a data set. Seemingly unrelated consumers are examined in terms of similarities (Anitpov, Et al, 2010).

Scholars agree in the utility and profitability that is possible by examining data like this from multiple perspectives (Sandhya, etc all, 2012). Indeed, data mining research has shown that multiple models and data mining methods must be combined as the various means of customer segmentation and grouping–via classification and clustering have their own limitations. (Danning, 2012).

Research Method

We are performing the data analysis on data from a Portuguese banking institution obtained via UCI's Machine Learning Archive. Two different sets of data (70% for training set and 30% for test set) would be prepared after performing a data cleanup on the obtained data. We would conduct the predictive analytics using Decision tree model and Logistic tree model to make predictions in the marketing domain and will conduct an evaluation using the Confusion Matrix. These predictions will help increase campaign success rates, and company productivity. They will also lead to increased customer satisfaction and customer engagement. In this research project we are using different data extraction techniques and analysis on several key attributes. On top of that we will be also analyzing the participation of the people based upon their job using the Tableau visualization.

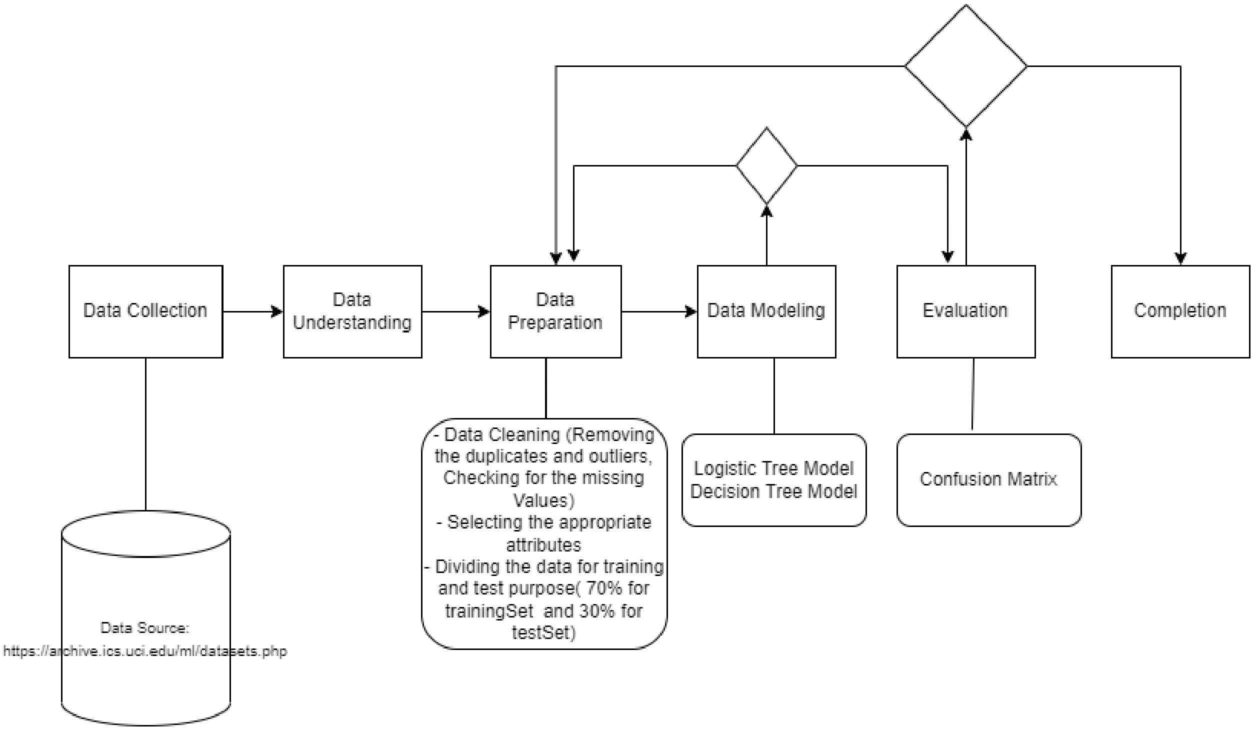

Data Collection
Data Understanding
Data Preparation
Data Modeling
Evaluation
Completion
- Data Cleaning (Removing the duplicates and outliers, Checking for the missing Values)
- Selecting the appropriate attributes
- Dividing the data for training and test purpose(70% for trainingSet and 30% for testSet)
Logistic Tree Model
Decision Tree Model
Confusion Matrix
Data Source:
https://archive.ics.uci.edu/ml/datasets.php

Figure: Research Methodology Process

Data Description

Attributes and Their Description

Category	Attribute Name	Description	Values
Input variables			
# bank client data:	Age	Age of contacts	numeric
	Job	type of job	categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown'
	Marital	marital status	categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed
	Education	education status	categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown'
	Default	has credit in default?	categorical: 'no','yes','unknown'
	Housing	has housing loan?	categorical: 'no','yes','unknown'
	Personal Loan	has personal loan?	categorical: 'no','yes','unknown'
# related with the last contact of the current campaign:	Contact Method	contact communication type	categorical: 'cellular','telephone'
	Month	last contact month of year	categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec'
	Day of Week	last contact day of the week	categorical: 'mon','tue','wed','thu','fri'
	Duration	last contact duration, in seconds	numeric
# other attributes:	Current Campaign Contacts	campaign: number of contacts performed during this campaign and for this client	numeric, includes last contact
	Days Since Previous Contact	pdays: number of days that passed by after the client was last contacted from a previous campaign	numeric; 999 means client was not previously contacted

	Previous Number of Contacts	previous: number of contacts performed before this campaign and for this client	numeric
	Previous Outcome	poutcome: outcome of the previous marketing campaign	categorical: 'failure','nonexistent','success'
# social and economic context attributes	Employment Variation Rate	emp.var.rate: employment variation rate - quarterly indicator	numeric
	Consumer Price Index	cons.price.idx: consumer price index - monthly indicator	numeric
	Consumer Confidence Index	cons.conf.idx: consumer confidence index - monthly indicator	numeric
	Euro Interbank Month Offered Rate	euribor3m: euribor 3 month rate - daily indicator	numeric
	Number of Employees	nr.employed: number of employees - quarterly indicator	numeric
Output variable (desired target)	Campaign Participation	y - has the client subscribed a term deposit?	binary: 'yes','no'

Data Preparation

The first step in the data preparation process (after data understanding was completed) on this data set was to update attribute names from the source csv file to make them more understandable and user friendly. The attribute “Loan” was updated to “Personal Loan”, “Contact” was updated to “Contact Method”, “Campaign” was updated to “Current Campaign Contacts”, “PDays” was updated to “Days Since Previous Contact”, “Previous” was updated to “Previous Number of Contacts”, “Poutcome” was updated to “Previous Outcome”, “Emp.Var.Rate” was updated to “Employee Variation Rate”, “Cons.Price.Idx” was updated to “Consumer Price Index”, “Cons.Conf.Idx” was updated to “Consumer Confidence Index”, “Euribor3m” was updated to “Euro Interbank Month Offered Rate”, “Nr.Employed” was updated to “Number of Employees”, “Y” was updated to “Campaign Participation” (Note this is the target/responding variable). As stated above the purpose of this was to make each attribute more distinct, clear and easy to understand exactly what it means. Second was to identify and remove any unknown / missing values as well as outliers. The following attribute sections all had outliers, unknown and missing attribute values that were removed, Age, Marital, Education, Default, and Housing. The next step was to remove attributes that were irrelevant to investigating our research question. Since we are looking at the decision of individual customers, the environmental factors of Consumer Confidence Index, Consumer Price Index, Employment Variation Rate, and Euro Interbank Month Offered Rate were all not used for analysis. Another non environmental attribute that was filtered out due to irrelevance was Number of Employees. Lastly the dataset was randomly split into two different data sets, the training set and test set. This was done following the normal A-B testing procedure, of 70% of the data is used in the training set and then the remaining 30% of the data is used in the test set. The training set is used to create our prediction models, and then those models are tested for accuracy on the testing set. The entire data preparation process helped create a more accurate, and relevant prediction.

Results

Looking at Figure 1 of the appendix, the results show that people in the administrative profession had the most campaign participation among customers. With almost 30% of total customers being in this profession 4.33% of them participated in the campaign, which is much higher than any other profession. The closest being blue- collar professionals, which accounts for 18% of total customers and 1.43% of them participating in the campaign. With that being said, we could draw the conclusion that customers in the administrative profession are more likely to participate in the campaign than anyone in a different profession.

Based on figure 2 of the appendix we can see that the bulk of the customers that participated did so in the first 1,000 days (about 2 and a half years) of the campaign. With the duration of 0 – 952 days, all having over 100 customers participate in the campaign. Taking a closer look, most of the participation was recorded in the first 714 days with a total of 2,409 customers participating within this time frame. Based on the above data it would seem as though participation correlates with the number of total customers. As the customer count goes down, and so does the participation.

Figure 3 of the appendix shows the results of participation based on the time of previous contact made to the customer. With the time frame 7 days within contact seeming to be the most influential on participation in the campaign with 3-7 days being the most effective in participation numbers with 603 customers participating during this period after last contact and 146 customers participating after the 7-day mark. In further detail 81% of participating customer's participated within the first 7 days after contact.

Figure 4 of the appendix shows the marital status of the customers, with married customers having the most customers who participated in the campaign. This graph would also show the correlation between the number of customers and participation as married customers had the highest number of total customers as well as the highest number of participating customers. Single customers account for the second greatest number of total customers as well as participating customers.

Based on the p-values in figure 5 of the appendix, all of which would be considered high. It would appear that none of the listed attributes would have much of an effect on customer participation.

In figure 6 of the appendix, the prediction outcome is that no is the most likely. Meaning that customers will be more likely not to participate in the campaign if it keeps going. Based on this information it may be helpful to discontinue the campaign or start a new campaign to yield the desired results.

The graph in figure 7 of the appendix shows the weight of the attributes, with duration being the attribute that has the most effect on participation accounting for over half of customer participation according to the weight results. With 0.216 Marital status is the second most influential attribute for campaign participation. Days since previous contact had the third highest weight and seems to play a significant part in participation. Lastly, current campaign contacts and age, have the lowers weight which shows they may play a minor part in participation compared to the other 3 attributes.

In figure 8 of the appendix, the decision tree model shows a heavy emphasis on days since previous contact. Which means that days since previous contact is an essential piece to getting customer to participate in the campaign.

Based on figure 9 of the appendix, the confidence level shows no is the more likely answer to whether customer will participate in the campaign. With this information we can assume that we will see an abundance of no over yes when it comes to campaign participation.

The confusion matrix for the Decision Tree model:

DT confusion matrix	Actual		
DT prediction	**yes**	**no**	**Grand Total**
yes	73	71	144
no	890	7966	8856
Grand Total	**963**	**8037**	**9000**

The accuracy of the model is

Accuracy=73+79669000=0.8932≈89.32%

The confusion matrix for the Logistic Regression model:

LogR confusion matrix	Actual		
LogR prediction	**yes**	**no**	**Grand Total**
yes	502	739	1241
no	461	7298	7759
Grand Total	**963**	**8037**	**9000**

The accuracy of the model is

Accuracy=502+72989000=0.8667≈86.67%

Discussion

The achieved results and models are now ready to be presented to the bank's marketing departments. Even though the built models have shown their high accuracy when run on the test dataset, there are still a few details to consider before their implementation on the "real" data.

The dataset used can be viewed as outdated, since it was collected in 2010 and may not represent the current situation in the company. Moreover, during the data preparation stage, large quantity of entries was eliminated due to insufficiency, which could also affect the models' accuracy. One of the difficulties we came across when building the models was the complexity of the results. Due to the large number of attributes, and especially the large number of numerical variables, the initial decision tree with default settings ended up being too hard to comprehend. Therefore, it was later adjusted to give a more compact and easy to follow model, sacrificing some prediction accuracy.

Since both models identified a set of the most influential attributes, our first recommendation would be to encourage further collection of the corresponding data. As for the practical implementation: unfortunately, only a little number of attribute combinations are likely to lead to the customers' enrollment. In particular, the most successful (by number and positive result) outcome is when there was less than 17 days since last contact, the call duration was longer than 147.5 seconds and there were less than 7 current campaign contacts. Unfortunately, other combinations were not so successful. Therefore, our suggestion is to maintain contact with the client after the end of the campaign, trying to achieve less though longer calls.

Moreover, we would suggest additional changes to the attributes if further analysis takes place. In particular, a few attributes could be changed from numerical to polynomial, such as age and days since the last contact. That may help divide the customers into additional clusters that are easier to comprehend.

References

Alis, O., Karakurt, E. & Piero, M. (2000) Data Mining for Database Marketing at Garanti Bank. *Data Mining II, 94-107.*

Antipov, E., Pokryshevskaya, E. (2010) Applying CHAID for logistic regression diagnostics and classification accuracy improvement. *Journal of Targeting, Measurement, and Analysis for Marketing* 18**,** 109–117.

Danning, H., (2021) A Study on the Application of Decision Tree Algorithm in Mobile Marketing. *Journal of Physics: Conference Series*., Retrieved April 17, 2022, from pdf (iop.org)

Deloitte (2013), Kicking it up a notch Taking retail bank cross-selling to the next level.

Deloitte Development LLC. Retrieved April 17, 2022, from https://www.google.com/urla=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwijkLDE36b3AhVtmWoFHVoIBq4QFnoECCQQAQ&url=https%3A%2F%2Fwww2.deloitte.com%2Fcontent%2Fdam%2FDeloitte%2Fus%2FDocuments%2Ffinancial-services%2Fus-kickingitupanotch-092614.pdf&usg=AOvVaw3o85X1GSAXi-CHYswilqY

Gao, J. (2020) P-values – a chronic conundrum. *BMC Medical Research Methodology*. **20,** 167 . Retrieved April 17, 2022, from ://doi.org/10.1186/s12874-020-01051-6

Fernando, J. (2021) Certificate of Deposit, Investopedia, Retrieved April 19, 2022 from https://www.investopedia.com/terms/c/certificateofdeposit.asp

Pedro Domingos. (2012) A few useful things to know about machine learning. *Communications of the ACM*. ACM, 55(10):78–87.

Han, S. H., Lu, S. X., & Leung, S. C. H. (2012). Segmentation of telecom customers based on customer value by decision tree model. *Expert Systems with Applications*, *39*(4), 3964-3973.

Interpret the key results for binary logistic regression. Minitab Express. (n.d.). Retrieved April 17, 2022, from https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/interpret-the-results/key-results/

Lessmann, S., Stefan, V., (2009) A reference model for customer-centric data mining with support vector machines, *European Journal of Operational Research*, Volume 199, Issue 2,, Pages 520-530.

Ladyzýnski, P., ˙Zbikowski, K.(2), Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learningand random forests *Expert Systems with Applications* Volume 134, 15 November , Pages 28-35. Retrieved April 17, 2022, from https://doi.org/10.1016/j.eswa.2019.05.020

Moro, S. Cortez, P. and P. Rita. (2014) A Data-Driven Approach to Predict the Success of Bank Telemarketing. *Decision Support Systems*, Elsevier, 62:22-31, June

Neale, D. (2020). Linear regression vs decision trees. ML Corner. Retrieved April 18, 2022, from https://mlcorner.com/linear-regression-vs-decision-trees/.

Osei-Bryson K. (2004) Evaluation of decision trees: a multi-criteria approach. *Computers & Operations Research*. Volume 31, Issue 11. Pages 1933-1945

Ł. Piotr, Ż. Kamil., G. Piotr, (2019) Direct marketing campaigns in retail banking with the use of deep learning and random forests, *Expert Systems with Application*s, Volume 134, Pages 28-35.

Appendices

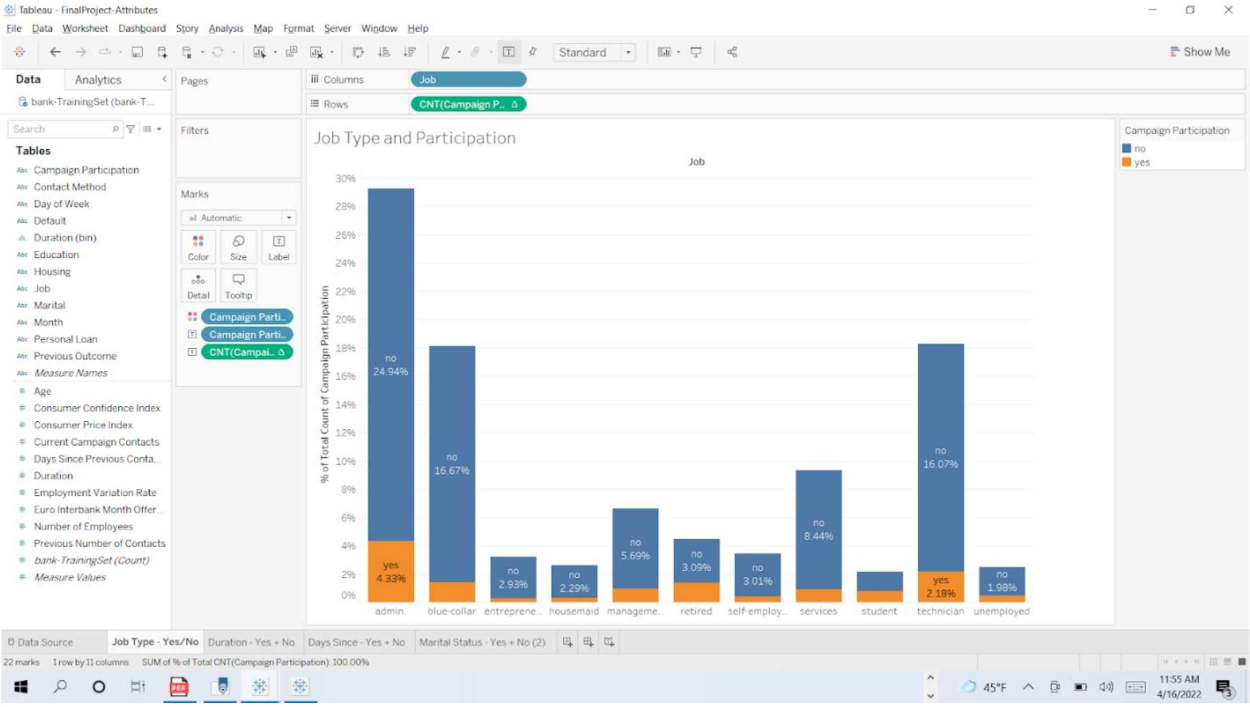


Figure 1

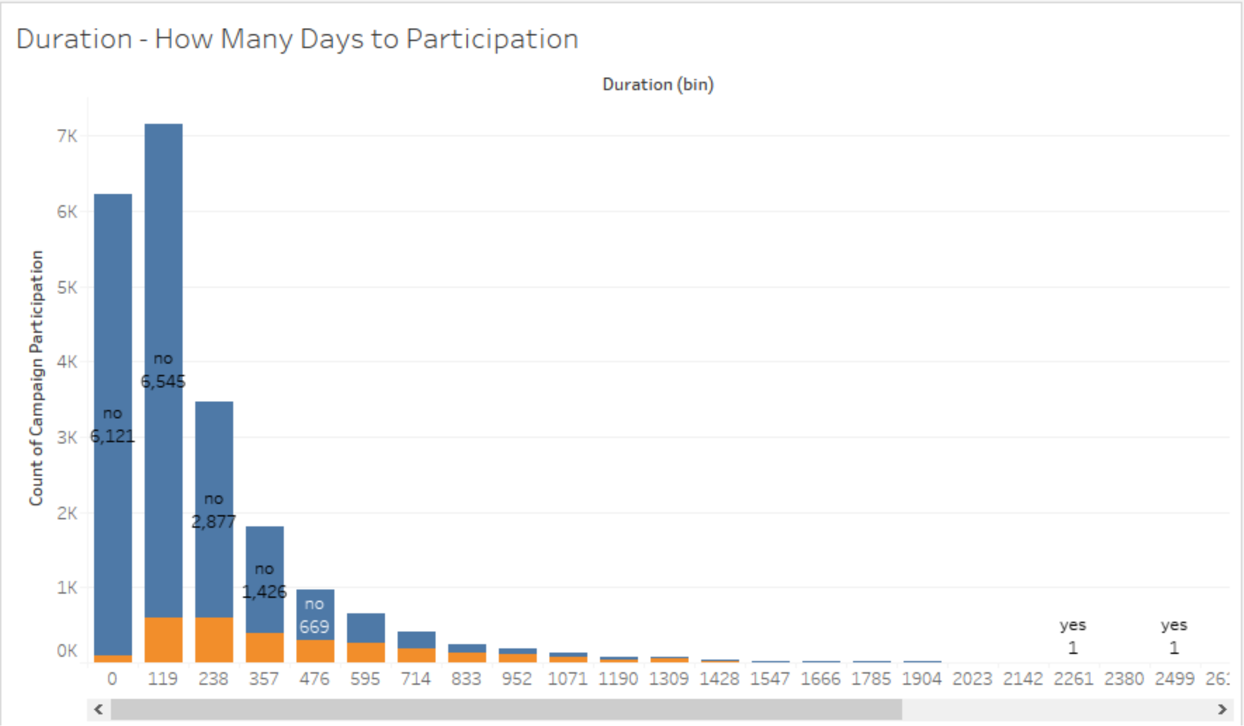


Figure 2

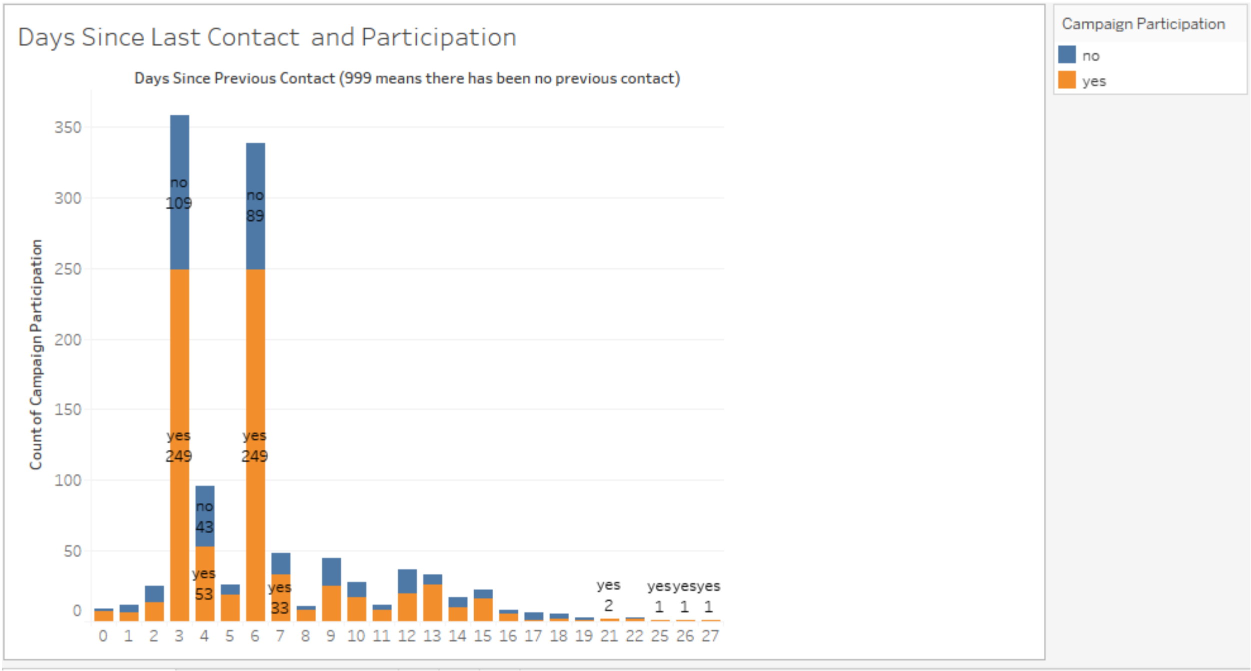


Figure 3

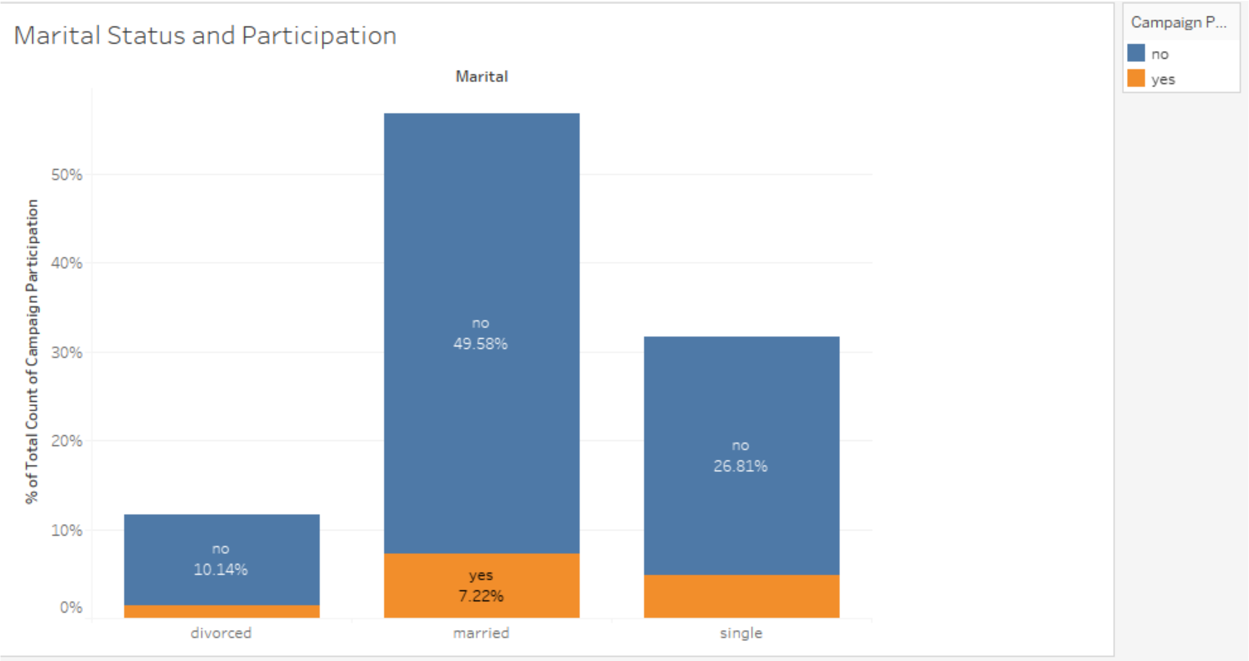
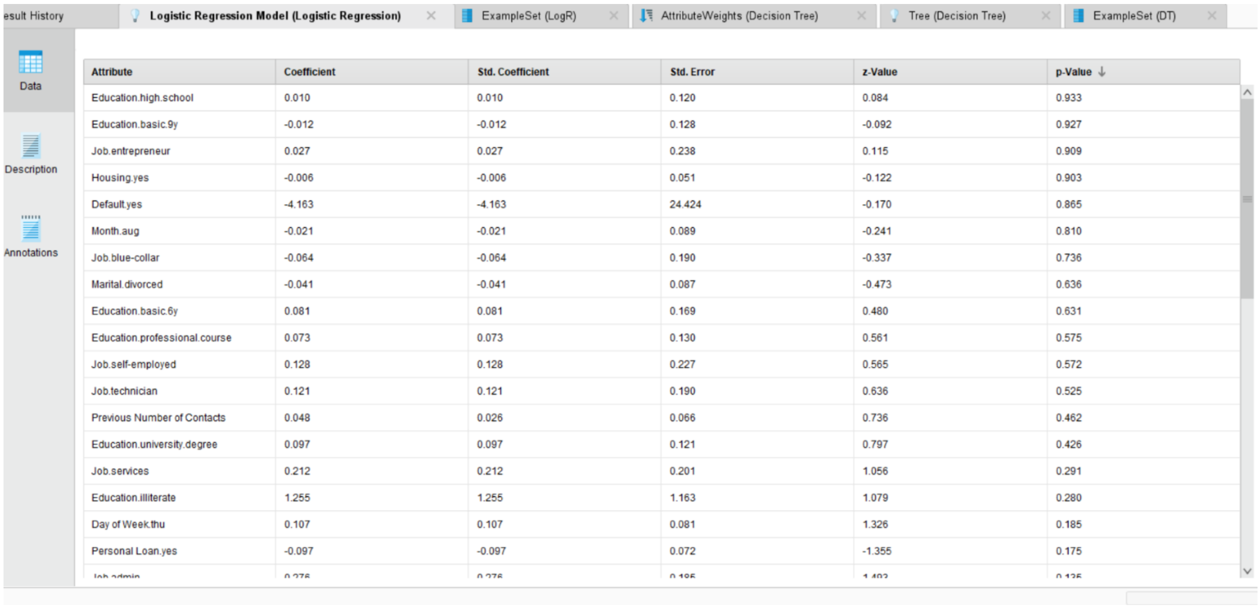
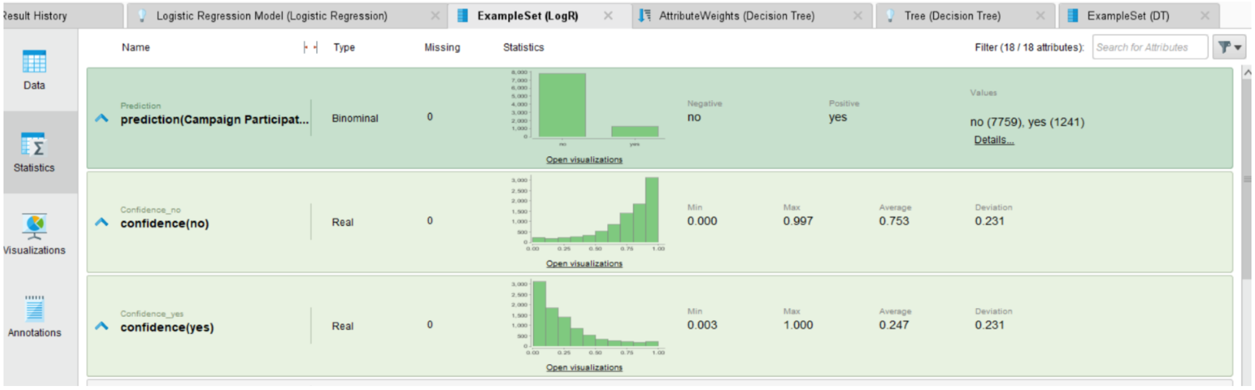


Figure 4



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value ↓
Education.high.school	0.010	0.010	0.120	0.084	0.933
Education.basic.9y	-0.012	-0.012	0.128	-0.092	0.927
Job.entrepreneur	0.027	0.027	0.238	0.115	0.909
Housing.yes	-0.006	-0.006	0.051	-0.122	0.903
Default.yes	-4.163	-4.163	24.424	-0.170	0.865
Month.aug	-0.021	-0.021	0.089	-0.241	0.810
Job.blue-collar	-0.064	-0.064	0.190	-0.337	0.736
Marital.divorced	-0.041	-0.041	0.087	-0.473	0.636
Education.basic.6y	0.081	0.081	0.169	0.480	0.631
Education.professional.course	0.073	0.073	0.130	0.561	0.575
Job.self-employed	0.128	0.128	0.227	0.565	0.572
Job.technician	0.121	0.121	0.190	0.636	0.525
Previous Number of Contacts	0.048	0.026	0.066	0.736	0.462
Education.university.degree	0.097	0.097	0.121	0.797	0.426
Job.services	0.212	0.212	0.201	1.056	0.291
Education.illiterate	1.255	1.255	1.163	1.079	0.280
Day of Week.thu	0.107	0.107	0.081	1.326	0.185
Personal Loan.yes	-0.097	-0.097	0.072	-1.355	0.175
Job.admin	0.276	0.276	0.185	1.493	0.136

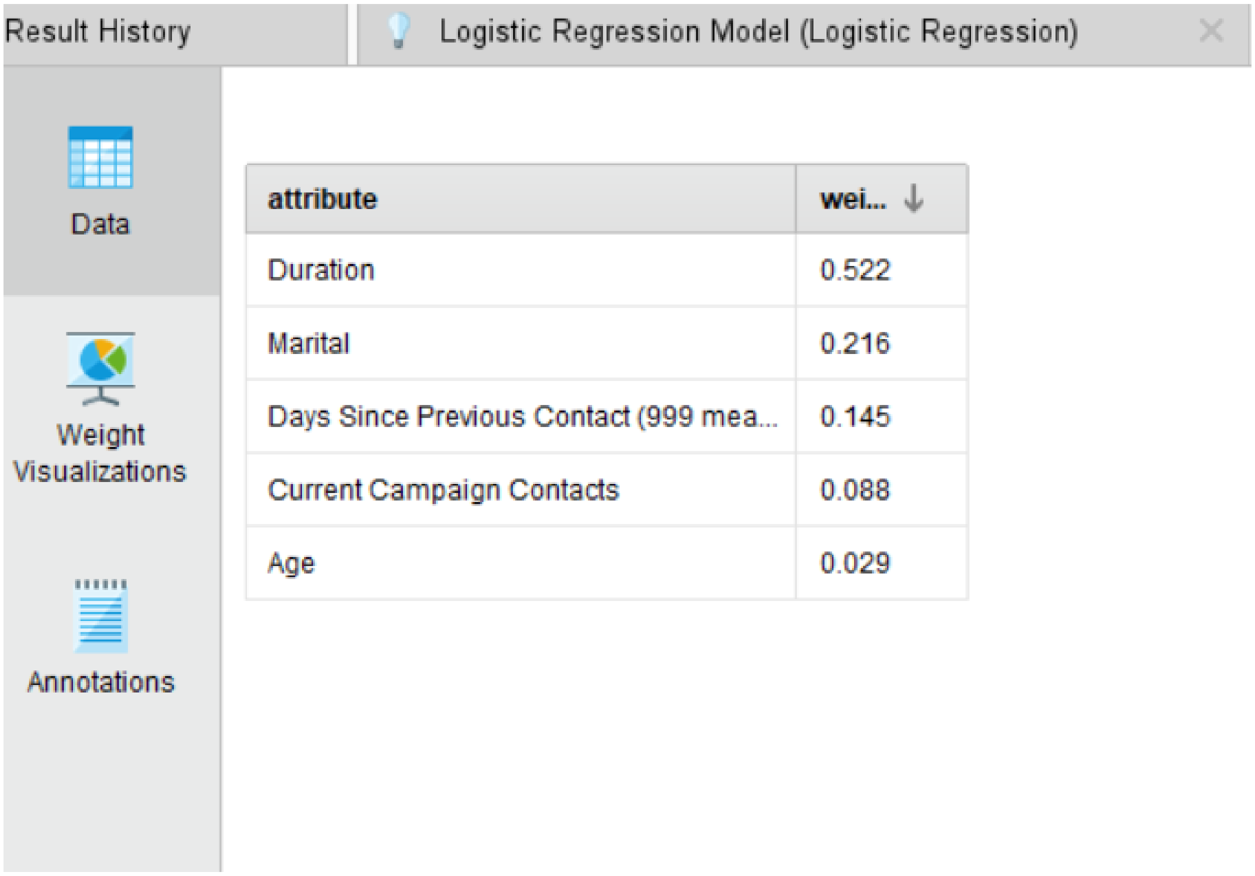
Figure 5



Filter (18 / 18 attributes):

Name	Type	Missing	Statistics			
Prediction prediction(Campaign Participat...	Binominal	0	Negative no	Positive yes	Values no (7759), yes (1241)	
Confidence_no confidence(no)	Real	0	Min 0.000	Max 0.997	Average 0.753	Deviation 0.231
Confidence_yes confidence(yes)	Real	0	Min 0.003	Max 1.000	Average 0.247	Deviation 0.231

Figure 6



attribute	wei...
Duration	0.522
Marital	0.216
Days Since Previous Contact (999 mea...	0.145
Current Campaign Contacts	0.088
Age	0.029

Figure 7

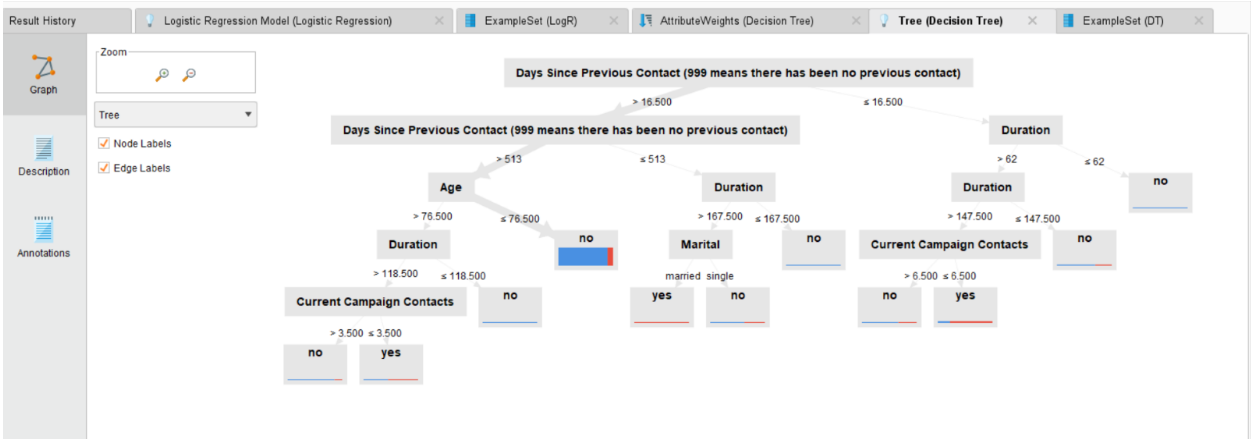
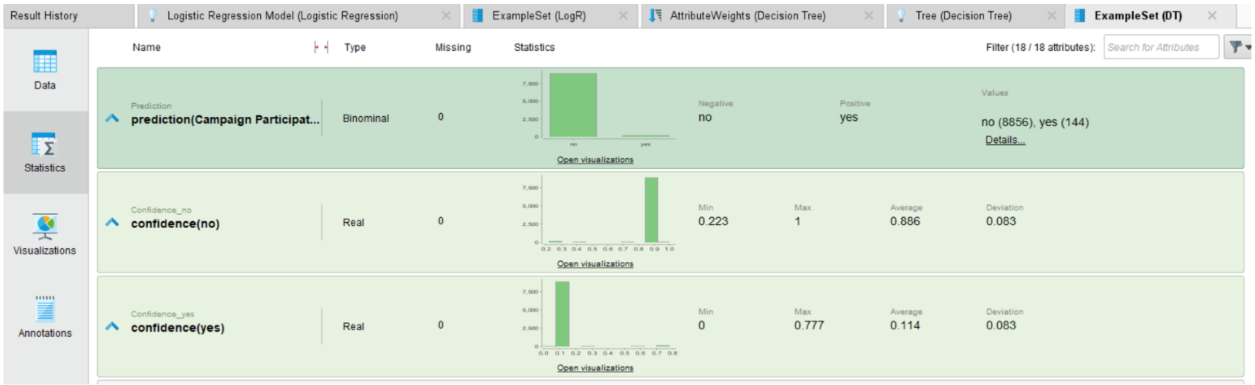


Figure 8



Name	Type	Missing	Statistics			
Prediction prediction(Campaign Participat...	Binominal	0	Negative no	Positive yes	Values no (8856), yes (144)	
Confidence_no confidence(no)	Real	0	Min 0.223	Max 1	Average 0.886	Deviation 0.083
Confidence_yes confidence(yes)	Real	0	Min 0	Max 0.777	Average 0.114	Deviation 0.083

Figure 9