Fordham University

Bank Marketing

Decision Tree, ANN, Logistics, KNN, Ensemble

Jaqi Zhang

Siying Wang

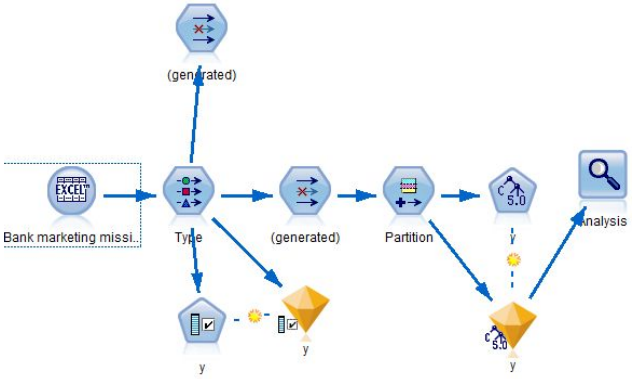
Nicolas Soria

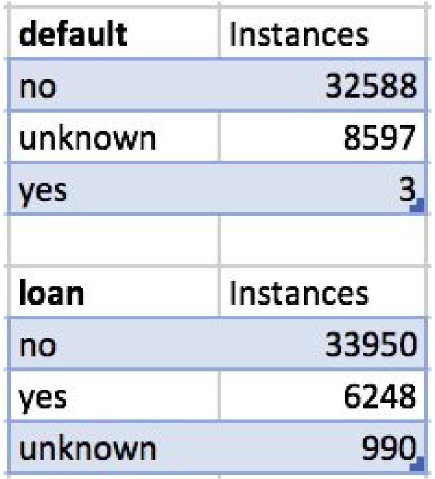
Report

We did our project on the dataset called “Bank Marketing”. The dataset contains 41,189 entries with each entry includes 21 attributes. The last attribute named “y” is our target variable with two classes, Yes and No. Our topic is that we want to predict the success of telemarketing calls for selling bank term deposits. The method we used is classification in which we ran five models: decision tree, ANN, KNN, Logistic and ensemble groups. This dataset is particularly interesting because predicting the effectiveness of an advertising approach and identifying key demographic patterns can be applied to various fields in a more macro level. Additionally, this dataset gives us a chance to work with three different types of attributes: demographics attributes, impact of marketing campaign and economics and social variables. Our problem statement is to build and choose the best model which can predict if either a potential customer or consumers is going to respond to the term deposit marketing. An accurate model should not only yield a high precision on predicting those who are going to say Yes, but also help us gain insight on which attributes significantly affect the outcome.

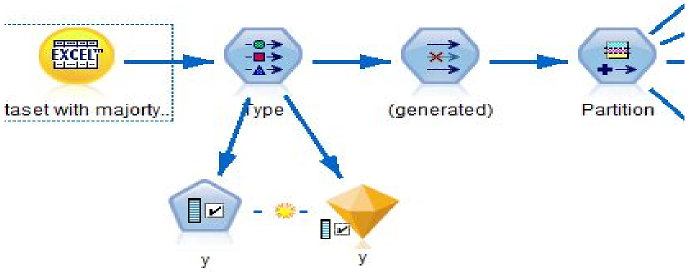
Preprocessing: Missing Data

Our initial dataset had two major problems that had to be addressed in the preprocessing stage: missing data values, and an imbalance in our target classes between yes or no. Our analysis of the dataset found there were 12,718 cells within the dataset marked as “unknown” within the categories of job, marital status, education, loan, and housing. We experimented with leaving the missing values in, deleting all records with an unknown value, and applying a majority rules solution (replacing unknowns with the highest mean within the attributes).





default	Instances
no	32588
unknown	8597
yes	3
loan	**Instances**
no	33950
yes	6248
unknown	990



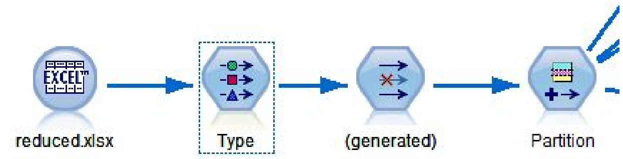
Ultimately, we found that our dataset with majority rules was performing the best in initial tests for models, as removing records with any missing values caused our imbalance in our target classes of yes and no to be even worse. After performing a count analysis on the dataset, we found that the bulk of the missing cells came in the default category, a category that initially had 32588 “no” responses, 8597 unknown, and only 3 yes responses. By applying majority rules to this column we simultaneously account for 2/3s of our unknown cells and essentially make the default attribute have no impact, which we are okay with because we enough attributes to still build a solid model. Analysis of the remaining categories lead us to believe that there was enough of a pattern to show that Majority Rule would not harm our data set and in fact improved our data set, so we applied majority rule to our entire dataset.

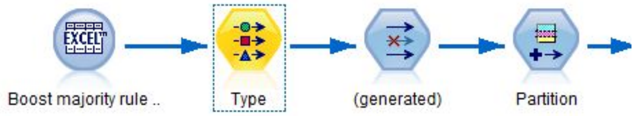
Preprocessing: Imbalance in target classes

In our initial dataset, our audit found that within our target variable which measures yes or no for agreeing to a loan, only 4640 records were marked as Yes out of 41,189 total records, which is only 11.26%. This was an issue, as the modeler is inclined to target overall accuracy and therefore assign most records as no to hit a high accuracy without necessarily giving us information on the yes class.

Initial dataset:

Response	Count	%
Yes	4640	11.26%
No	36548	88.73%

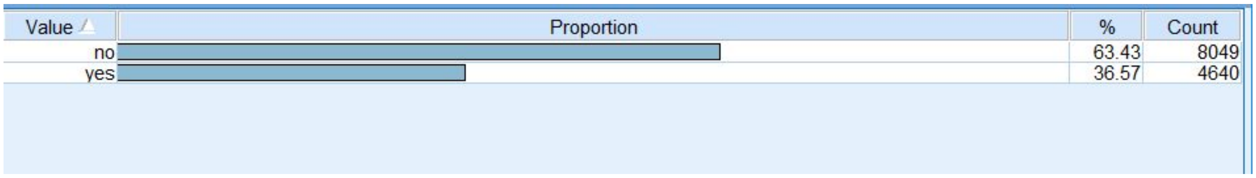




As with the imbalanced data, we ran two models with two different datasets- one boosting the Yes classes and one reducing from the No classes. We found that boosting actually not helpful to our analysis, as boosting lead to high accuracy in our training data but low accuracy in our dataset. This makes sense, as boosting does not add in new information but only duplicates existing records, and the proportion of initial yes/no records called for duplicating our 4,640 yes records too many times to reach a more balanced ratio.

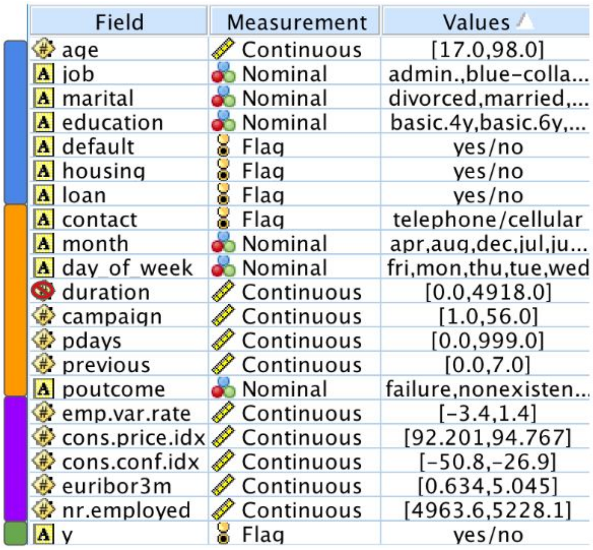
We found that reducing from the "no" class of our target variable actually improved our dataset and our ability to conduct analysis on the "yes" class records. While we know to be hesitant when reducing data, we felt justified in this dataset because boosting was impractical, and also because logically it makes sense: most of the population is going to respond “no” towards telemarketing calls about loans, so the goal isn’t classify if they will be yes or no, but to see if we can develop insights over who is answering yes to this telemarketing campaign. With that, we applied reduction to the no class, leaving our dataset to contain 8049 no responses and our initial 4640 yes responses to create a 63.4%-36.6 balance in our target class, as shown below.

Final Dataset proportion within target classes:



Value	Proportion	%	Count
no		63.43	8049
yes		36.57	4640

Dataset/Attribute Analysis:



Field	Measurement	Values
age	Continuous	[17.0,98.0]
job	Nominal	admin.,blue-colla...
marital	Nominal	divorced,married,...
education	Nominal	basic.4y,basic.6y,...
default	Flag	yes/no
housing	Flag	yes/no
loan	Flag	yes/no
contact	Flag	telephone/cellular
month	Nominal	apr,aug,dec,jul,ju...
day_of_week	Nominal	fri,mon,thu,tue,wed
duration	Continuous	[0.0,4918.0]
campaign	Continuous	[1.0,56.0]
pdays	Continuous	[0.0,999.0]
previous	Continuous	[0.0,7.0]
poutcome	Nominal	failure,nonexisten...
emp.var.rate	Continuous	[-3.4,1.4]
cons.price.idx	Continuous	[92.201,94.767]
cons.conf.idx	Continuous	[-50.8,-26.9]
euribor3m	Continuous	[0.634,5.045]
nr.employed	Continuous	[4963.6,5228.1]
y	Flag	yes/no

With 20 attributes being used as input data and 1 attribute (“y”, marked green above) used as target data, we found if helpful to divide the input attributes into 3 categories: demographics attributes (marked blue above), marketing and campaign attributes (marked orange), and economic/social attributes (marked purple).

Demographic attributes give us background data about the specific consumer the company is trying to reach such as age or education, and is independent from what the company can control. Marketing and campaign attributes directly relate to actions the company has or can make, and therefore this is controllable by the company. For instance, when the consumer is contacted and how long until the follow up contact was made were decisions made by the company and could reasonably be optimized going forward. The final category of economic and social attributes are broad statistics that come from the country of ~10 million people at the time of the call. These statistics are important to note as it is reasonable to assume that economic indicators have an impact on purchase decisions, especially with regards to loans. For example, a consumer could have the ideal candidate demographically for a "yes" response and the company could make all of the correct marketing decisions, but if a recession is in full swing than the customer may decline anyway, which is what these stats should help us identify.

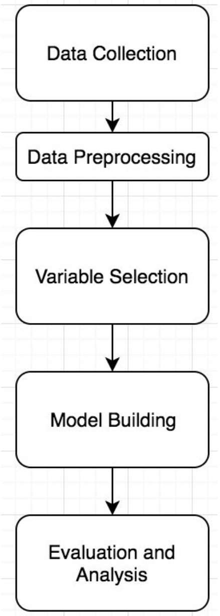
The only attribute we did not include in the dataset was duration, which measured the length of the phone call in which the bank loan was pitched. In our initial runs, this attribute was deemed far and away more important than the others, which overshadowed all other attributes. Logically it makes sense to remove it, as duration is only revealed once the customer already accepts or declines a loan as our target variable, and our research is supposed to identify potential customers before we make the call.

Methodology overview:

During Data collection, our initial dataset was collected from UCI depository.

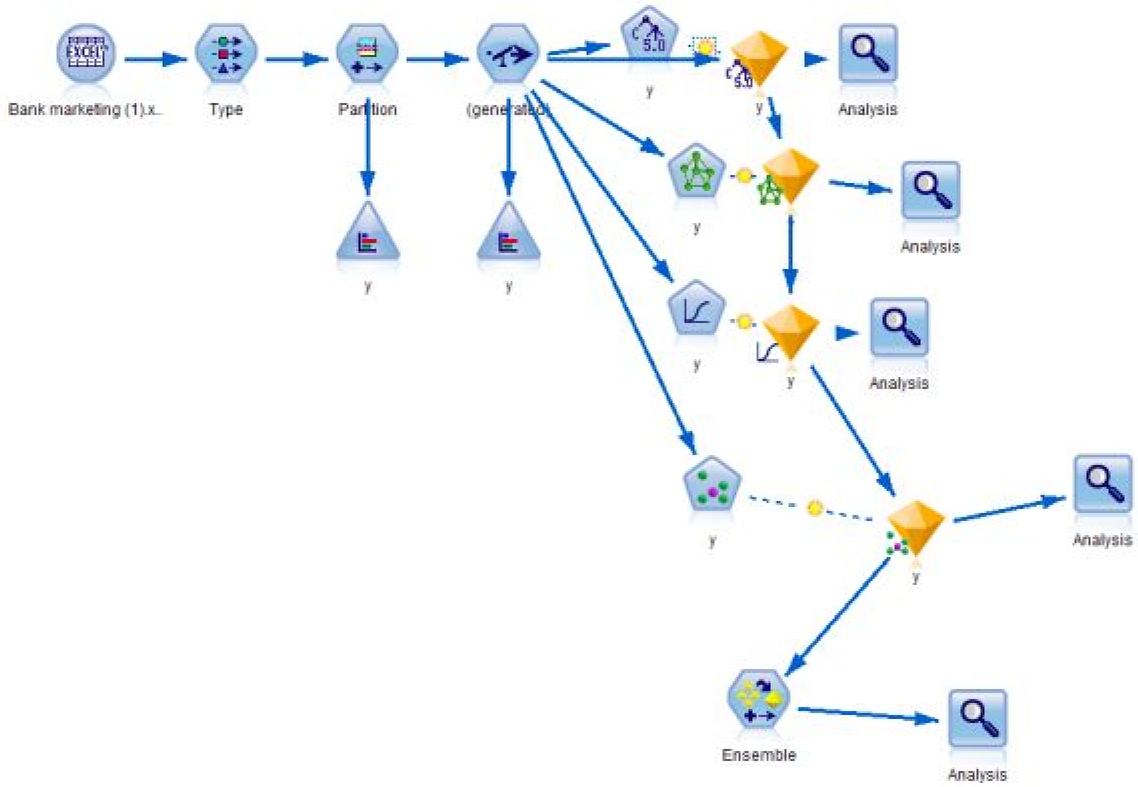
In our Data preprocessing phase, we reduced from the no classes and applied a majority rule to removing missing values/

In Variable Selection, Input was the only variable we chose not to use as it was too strong of an indicator and did not help our analysis.

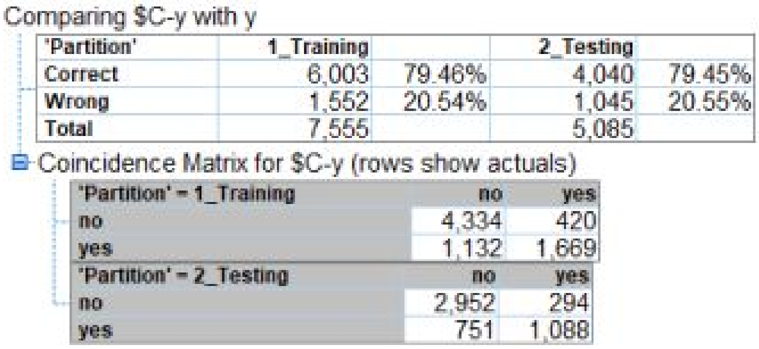


Results and Discussion:

For model building, we used decision trees, neural networks, logistic regression, KNN, and an Ensemble, show as figure below, and then conducted evaluation and analysis on the results.



Among these 5 models, we found that Ensemble model yields the highest overall accuracy rate, which is 80.7%. But Decision tree model has the highest precision rate on predicting the Yes class. The model predicts 1382 records to be Yes class, and 1088 of them are actually Yes class, which gives us precision rate of 80.7%. Since our task is to predict if a consumer is responding or saying Yes to the marketing campaign, we will focus on analyzing results from Decision tree model. Below shows the coincidence matrix of Decision tree model.



Comparing $C-y with y

'Partition'	1_Training		2_Testing	
Correct	6,003	79.46%	4,040	79.45%
Wrong	1,552	20.54%	1,045	20.55%
Total	7,555		5,085	

Coincidence Matrix for $C-y (rows show actuals)

'Partition' = 1_Training	no	yes
no	4,334	420
yes	1,132	1,669

'Partition' = 2_Testing	no	yes
no	2,952	294
yes	751	1,088

Insights:

The results from Decision tree model shows that Number Employed (national employment rate), Consumer price index and Pdays (number of days passed by since last time the consumer was contacted) are the top three predictors. We have 2 examples from the Decision tree model trend which shows Yes class and No class. If Number Employed <= 5078.650, Pdays<=15.5, Yes; If Number Employed > 5087.650, Consumer Confidence Index <= -46.65, Consumer Price Index> 92.959, Euribor3m > 1.378, Age <= 26.5 and Previous > 0.5, No.

We found that out of the top three predictors, two are economic indicators: Number Employed and Consumer Price index, this shows that economy has a high impact on if consumer are purchasing bank deposits. Take Consumer Price Index as an example, the model shows that with high Price index, consumers are less likely to purchase term deposit. This result makes sense if we use the 2008 Economic Recession as an example. During that time, both companies and individuals were losing a lot of money. With high Price index, which leads to high inflation, consumers do not have cash to purchase bank deposit.

Conclusion:

After exploring the problem statement, data processing by majority rule and dataset reduction, examining variables, building models and Evaluation and Analysis, we come to the conclusion that when predicting if consumers are responding to bank deposit marketing, economic indicator is a decisive factor.