**BANK MARKETING CAMPAIGN**

Ashima Gupta

Gowtham Janga

Himesh Raste

Ruchika Jain

Sahibpreet Singh

MS BAPM, University of Connecticut

OPIM-5604: Predictive Modeling

Pankaj Prakash

20th April 2021

Table of Contents

1. Executive Summary3

2. Business Understanding4

2.1 Business Problem4

2.2 Business Value4

3. Methodology5

3.1 Dataset Overview5-6

3.2 Exploratory Data Analysis & Preprocessing7-12

3.3 Nominal Logistic Model 12-14

3.4 Lasso 14

3.5 Ridge 15

3.6 Decision Trees 15

3.7 Bootstrap Forest 16

3.8 Boosted Trees 16

3.9 Neural Nets 16-17

3.10 Naïve Bayes 17

4. Evaluation18-20

5. Deployment 21

6. References22

1. **Executive Summary**

Term deposits are a major source of income for a bank. A term deposit is a cash investment held at a financial institution. Customer’s money is invested for an agreed rate of interest over a fixed amount of time, or term. The bank has various outreach plans to sell term deposits to their customers such as email marketing, advertisements, telephonic marketing, and digital marketing.

Telephonic marketing campaigns remain one of the most effective way to reach out to people. However, they require huge investment as large call centers are hired to execute these campaigns. Hence, it is crucial to identify the customers most likely to convert beforehand so that they can be specifically targeted via call. The data is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe to a term deposit (variable y).

Our dataset has class imbalance. So instead of focusing on accuracy of the model we chose Recall as the metric for determining the final model. we executed different predictive models and chose Naïve Bayes as our final model because it assumption of independent nature of the variables. So, it captured almost twice the number of most likely subscribers when compared with other models. We were able to capture 62% of the clients saying “Yes” to the term deposit while observing only 20% of the overall database. we also recommended bank to focus more on some of the important predictors: **employment variation rate, consumer price index, month, contact, poutcome** when targeting the customers for term deposits subscription.

**2. Business Understanding**

**2.1 Business Problem**

The bank needs to get their target audience to subscribe to their Term Deposit which is a crucial source of income for the bank. They perform Digital, on-the-ground, and Telephonic campaigns to reach their audience with Telephonic Calling proven as the most effective so far. But, unfortunately reaching everyone on cell is very expensive for their campaigns as **randomly calling customers requires a lot of investment** in terms of resources, time, and money. Since the **conversion rate is very low** so there is a need to perform data analytics on the current datasets to target the eligible and interested audience only.

**2.2 Business Value**

* Term deposits are a major source of income for a bank. A term deposit is a cash investment held at a financial institution.
* The customer’s money is invested for an agreed rate of interest over a fixed amount of time, or term.
* The more the people subscribe to the term deposit, the more the monetary benefit to the bank.
* Performing Data Analytics would minimize the investment of the bank by narrowing the targeted population.

1. **Methodology**

**3.1 Dataset Overview**

The dataset was picked from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. There are four datasets out of which we choose bank-additional-full.csv (41188 rows and 21 columns) and bank-full.csv (45211 rows and 17 columns, its an older version of bank-additional-full.csv dataset). These columns have been bucketed into their respective categories which you can see below:

**Variables related to bank client data:**

Age: Client’s age (numeric).

Job: type of job (categorical: ‘admin.’, ’blue-collar’, ’entrepreneur’, etc.)

Marital: marital status (categorical: ‘divorced’, ’married’, ’single’, ’unknown’)

Education (categorical: ‘basic.4y’, ’basic.6y’, ’basic.9y’, ’high.school’, etc.)

Default: has credit in default? (categorical: ‘no’, ’yes’, ’unknown’)

Housing: has housing loan? (categorical: ‘no’, ’yes’, ’unknown’)

Loan: has personal loan? (categorical: ‘no’, ’yes’, ’unknown’)

Balance: Average yearly balance of the client (Numeric)

**Variables related to last contact of the current marketing campaign:**

Contact: 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 week (categorical: ‘mon’, ’tue’, ’wed’, ’thu’, ’fri’).

Day: Last contact day of month (numeric).

Duration: Last contact duration in seconds (numeric).

**Other attributes:**

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 performed before this campaign and for this client (numeric)

Poutcome: outcome of the previous marketing campaign (categorical: ‘failure’, ‘nonexistent’, ‘success’)

**Social and economic context attributes:**

Emp.var.rate: employment variation rate - quarterly indicator (numeric)

Cons.price.idx: consumer price index - monthly indicator (numeric)

Cons.conf.idx: consumer confidence index - monthly indicator (numeric)

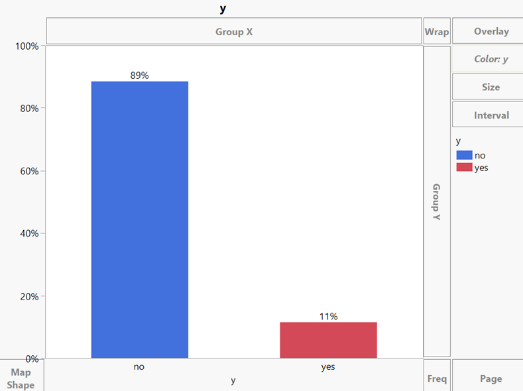
Euribor3m: euribor 3-month rate - daily indicator (numeric)

Nr.employed: number of employees - quarterly indicator (numeric)

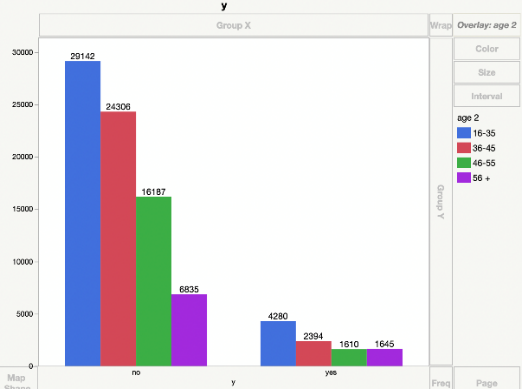
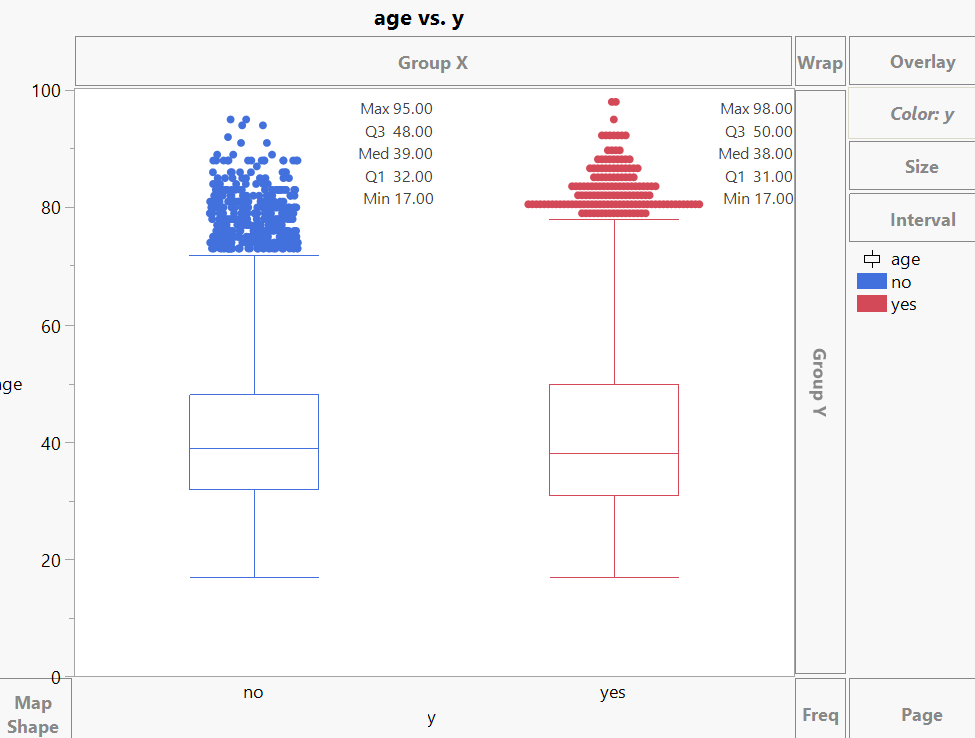
**Output variable (desired target):**

y - has the client subscribed a term deposit? (binary: ‘yes’, ‘no’)

**3.2 Exploratory Data Analysis & Preprocessing**

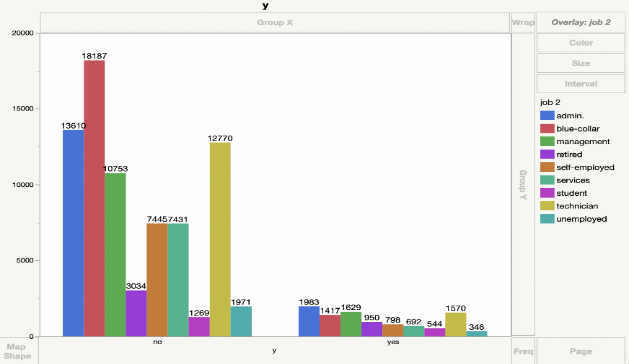
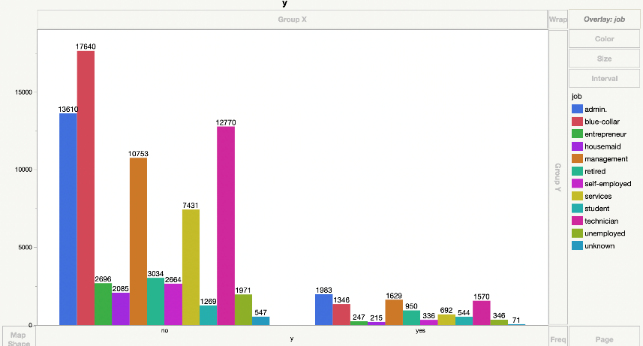
**y:** we can see the data is highly imbalanced. No’s are almost 8 times to Yes’s.

**Age: Age2:**



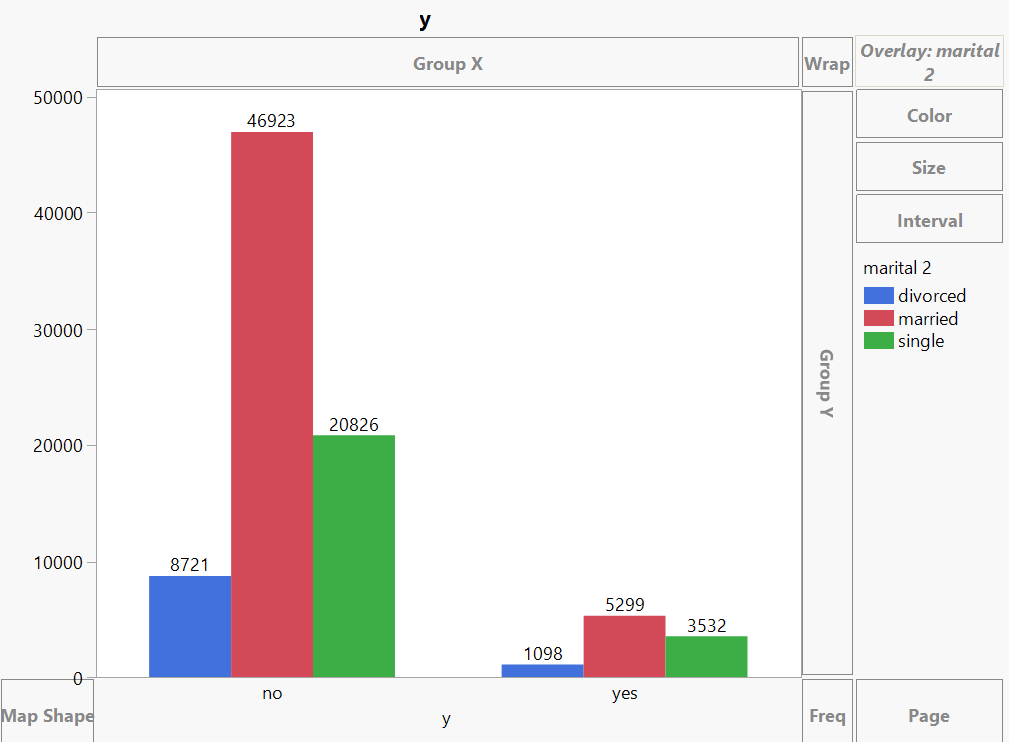
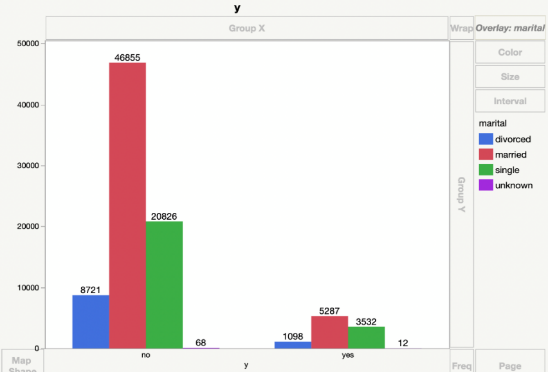
Boxplot shows median age of both subscribed & unsubscribed clients overlaps indicating it is not significant predictor. Recoded the age variable to age2, so that we know which range of age groups are more likely to subscribe and can target them.

**Job: Job2:**



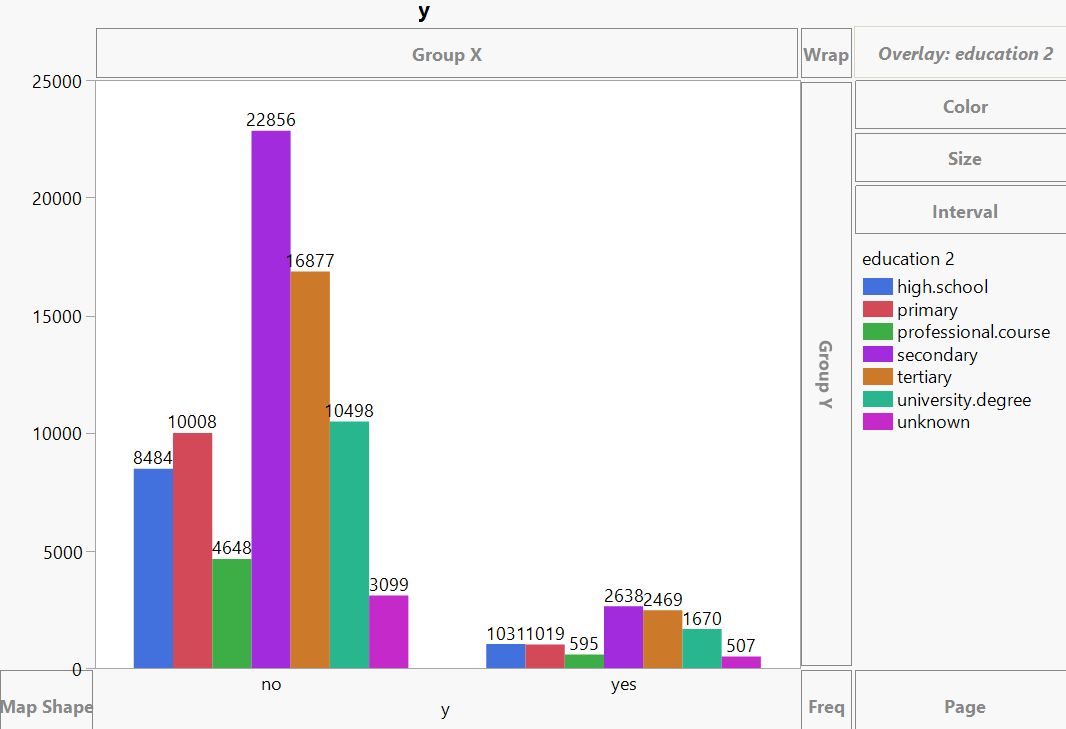
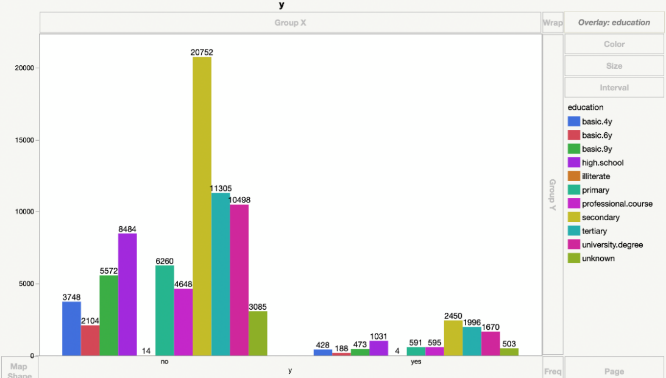
Recoded ‘job’ to ‘job2’ by grouping self-employed, entrepreneur, and housemaid to self-employed as they all represent the same. Grouped job category unknown (observations are <1% of total) with blue-collar (has highest observations). Admin, management and technicians are the most likely to subscribe the deposit.

**Marital: Marital2:**

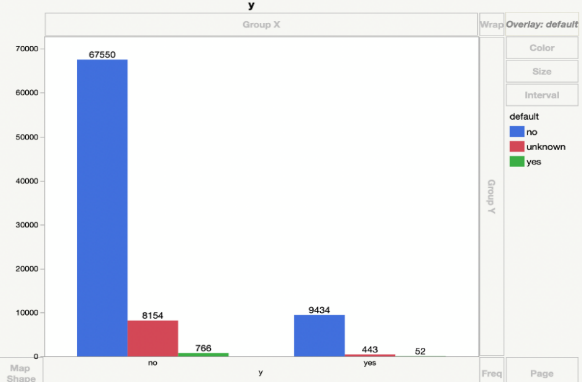


Grouped unknown (observations are <1% of total) with married (have highest observations) and recoded as Marital2. Married ones are the most likely to subscribe the deposit.

**Education: Education2:**

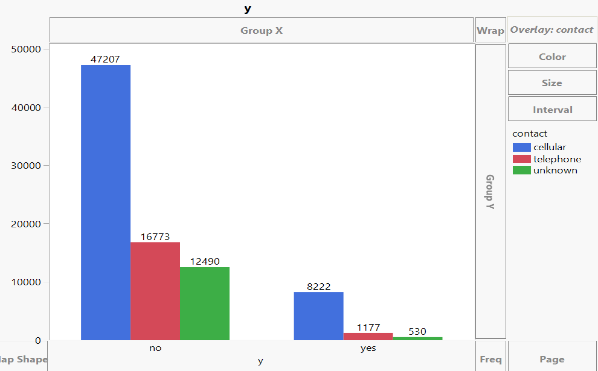


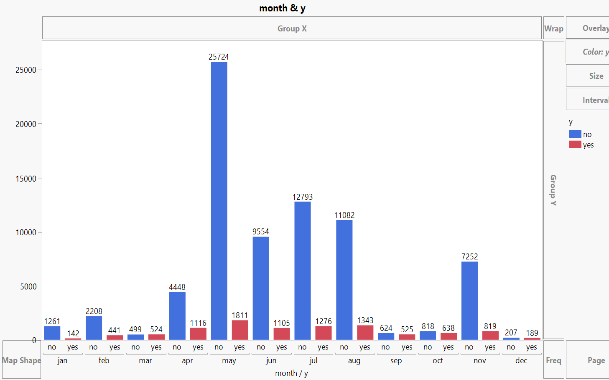
Grouped basic4y, basic6y, and basic 9y with primary, secondary, and tertiary respectively as they all represent the same and recoded as Education2. Didn’t group unknown category as observations are approx. 5% of total and will have significant impact on prediction. Secondary, tertiary, and university degree holders are most likely to subscribe the deposit.

**Default:**

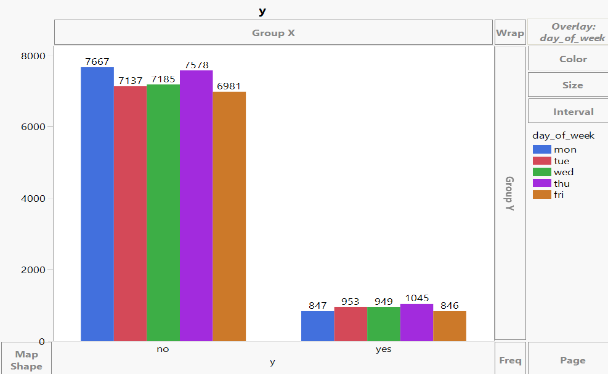
Observations having default yes is very negligible. So, we conclude most of the clients are not defaulters. Didn’t group unknown category as observations are approx. 11% of total and will have impact on prediction.

**Contact:**

Clients with a cellular contact are most likely to subscribe the deposit. Didn’t group unknown category as observations are approx. 15% of total and will have impact on prediction.

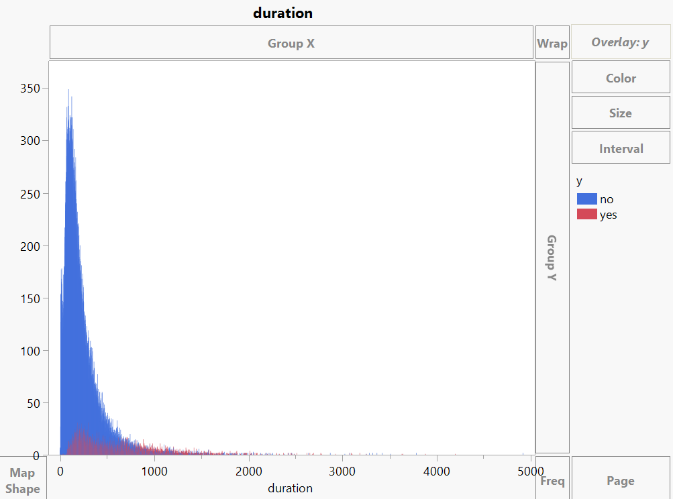
**Month:**

Clients were contacted the most in May, June, July and August. However, the % of people who said yes to the term deposit subscription were very less as compared to the ratio of people in the months of March, September, October and December.

**Day\_of\_week:**

All the days of week have similar distribution of records in terms of clients subscribing to the term deposit. So, we deduce that this variable won’t have significant impact on the predicted variable.

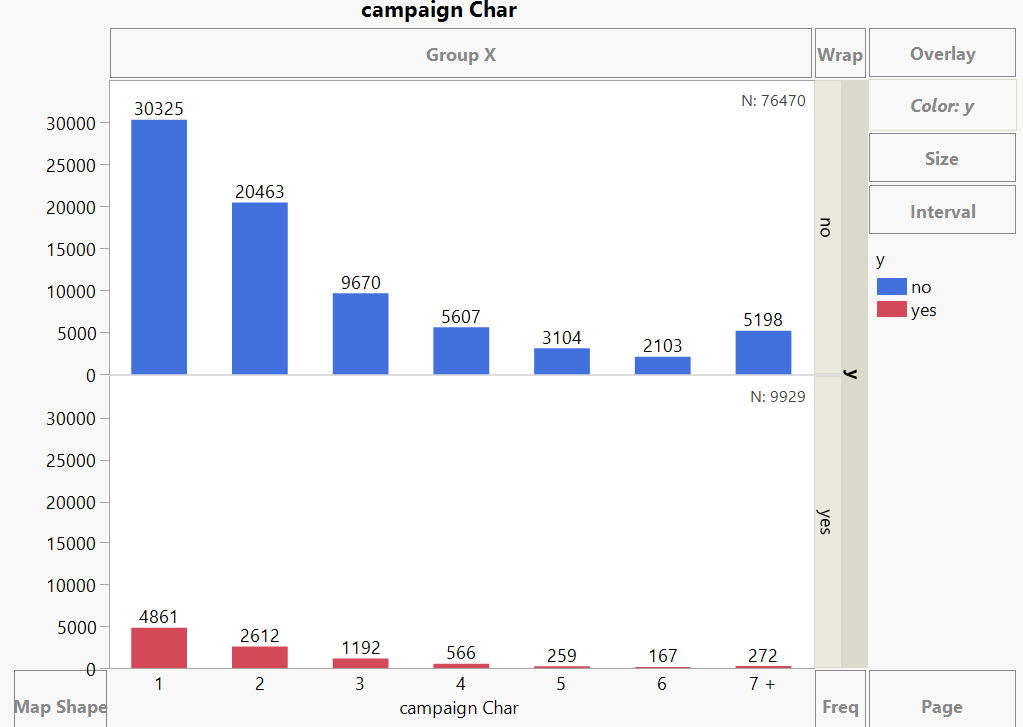
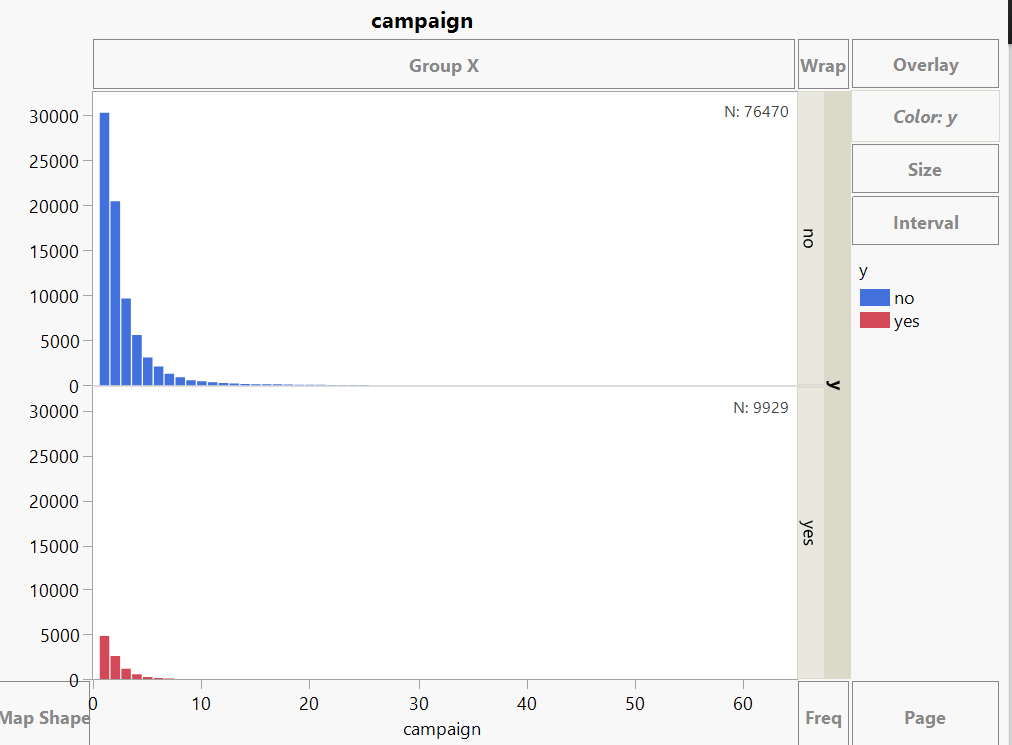
**Day:** Day also has similar distribution as Day\_of\_week in terms of clients subscribing to the term deposit. So, this variable also won’t have significant impact on the predicted variable.

**Duration:**

From the distribution we can infer, higher the call duration greater the number of subscriptions.

Also, its mentioned in dataset source that duration should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. But we want to see how duration impacts the prediction. So will be including this variable in the baseline model and based on the impact we will take a decision whether to include this variable or not in our final model.

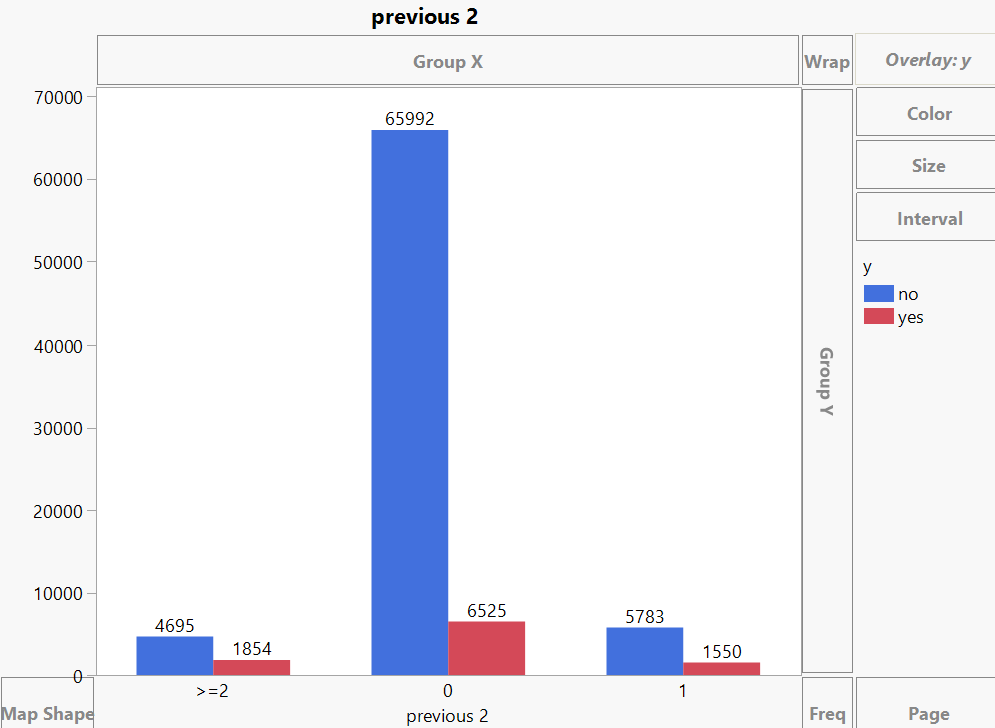
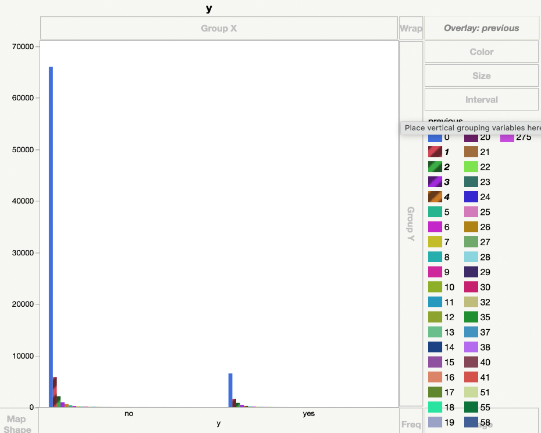
**Campaign: Campaign2:**



Grouped clients who have been contacted more than six times during this campaign to 7+ category (observations approx. 6% of total) and recoded as campaign2. From the plots we can observe approx. 95% of clients subscribed to deposits within 5 calls.

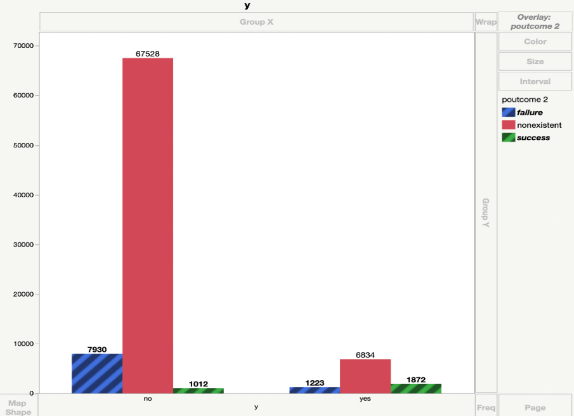
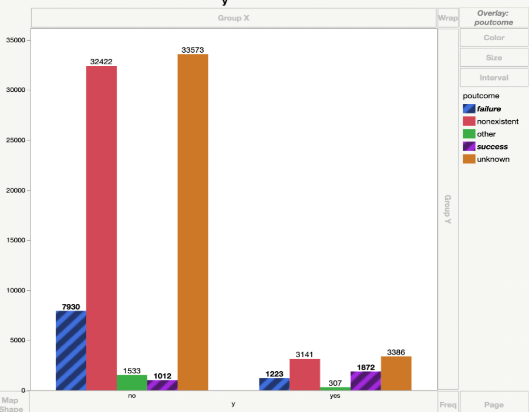
**Loan:** Similar to Marital,grouped unknown (observations are <1% of total) with No category and recoded as loan2. People who do not have a loan are the most likely to subscribe the deposit.

**Previous: Previous2:**



Previous is similar to campaign, but data of number of contacts performed before this campaign. Grouped clients who are contacted multiple times as >=2 category (observations approx. 7% of total) and recoded as Previous2. From the plots we can observe the probability of subscribing to the term deposit is increasing if the customer is contacted more.

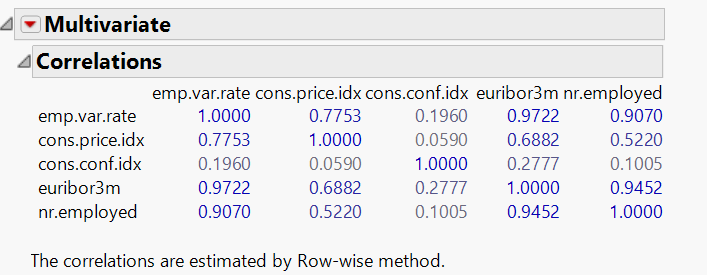
**Poutcome: Poutcome2:**



Grouped unknown, and other categories to non-existent as they are similar categories and recoded as Poutcomes2. From plot we can see, 65% of people who subscribed in previous campaign have subscribed for this campaign also. Targeting those clients should be the priority.

**Housing:** Similar to Marital,grouped unknown (observations are <1% of total) with YES category. People who do not have a housing loan are the most likely to subscribe the deposit.

The variables **Emp.var.rate, Cons.price.idx, Euribor3m, Cons.conf.idx** and **nr.employed** are social and economic indicators. We believe they are correlated in nature. So, to confirm if there is correlation between these variables, we are computing a correlation matrix:

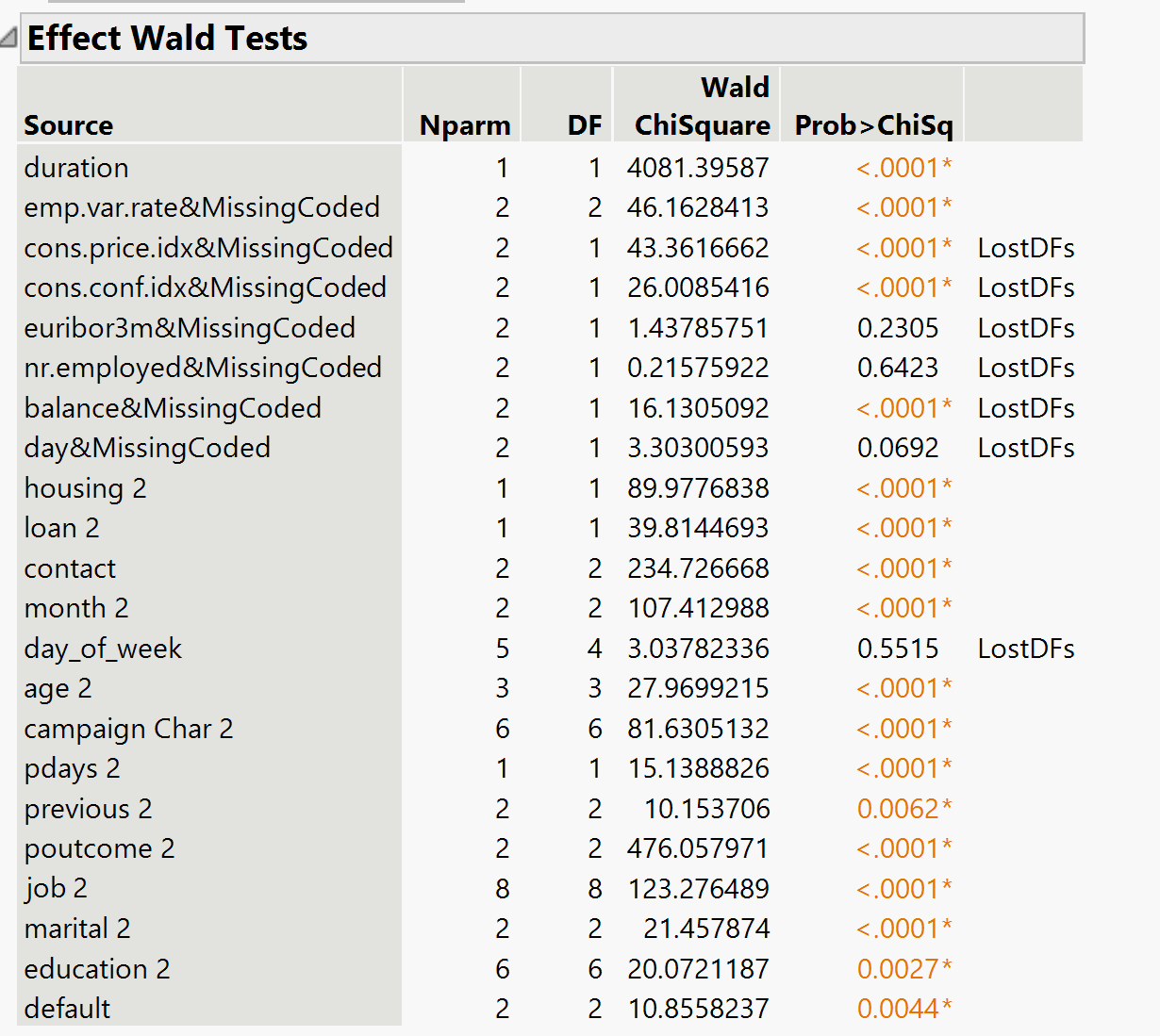
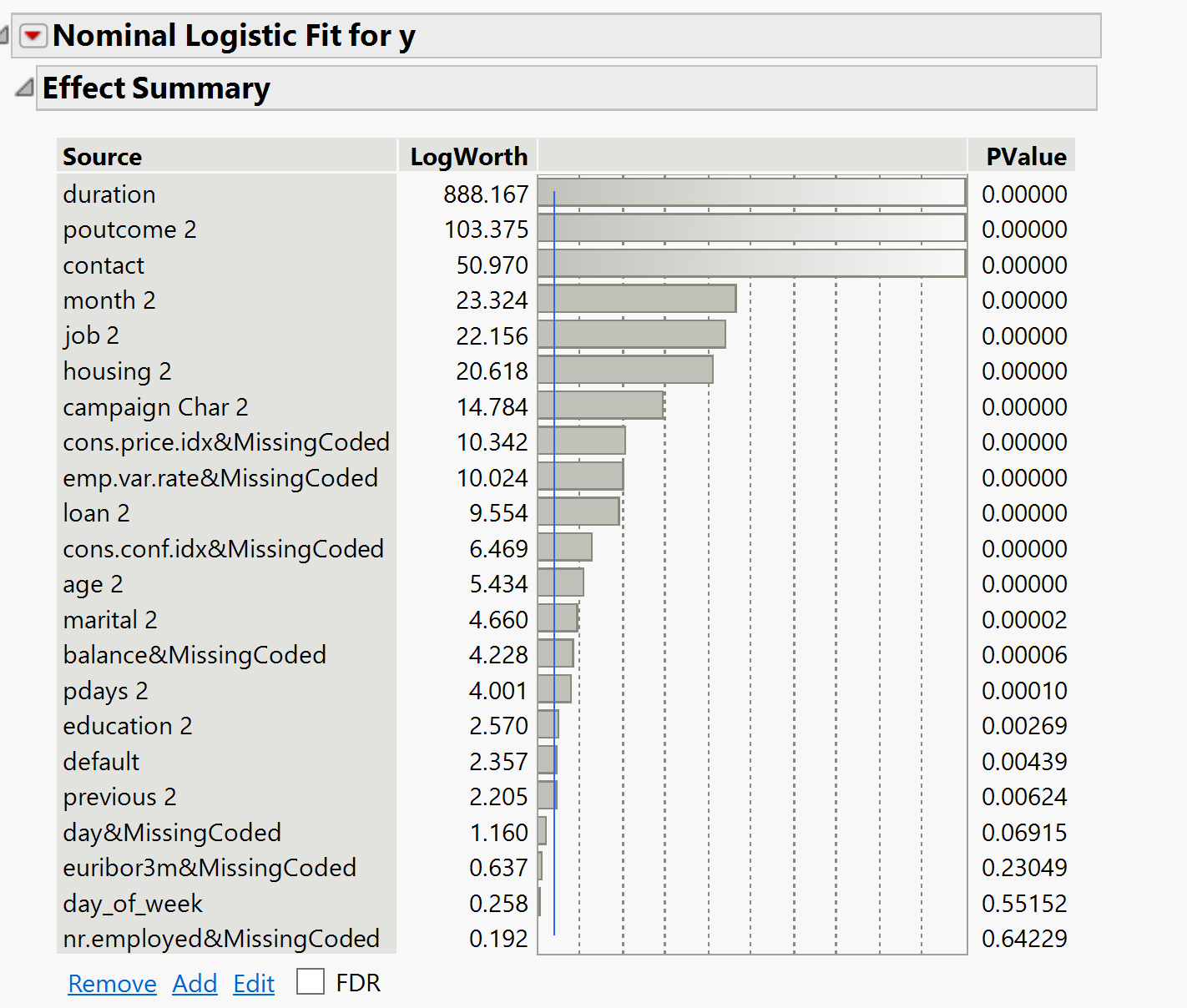
Based on correlation matrix, we observe that Emp.var.rate, Euribor3m and nr.employed are highly correlated (>0.85).

So, we will model these variables in two different ways: first one is by considering only one variable out of these three correlated variables, and the second one is by performing PCA on all these five variables and use the PCA components for modeling.

**3.3 Nominal Logistic Regression**

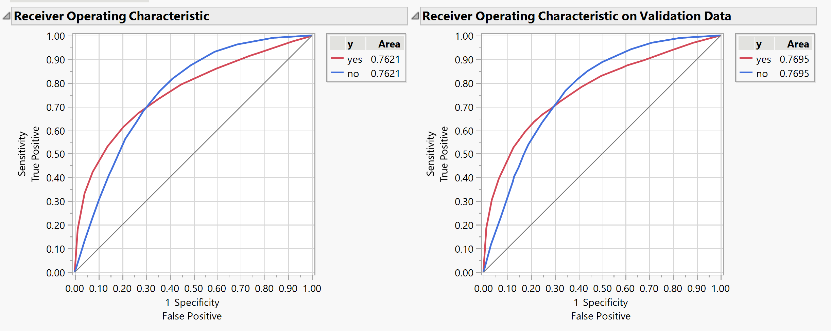
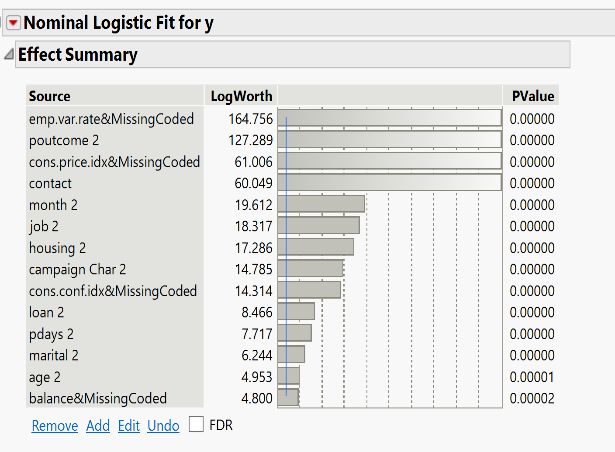
**For 86K:** Merged bank-full and bank additional full dataset. There were 5 unique variables that had missing values, so we used "Informative Missing" feature in JMP to deal with them.

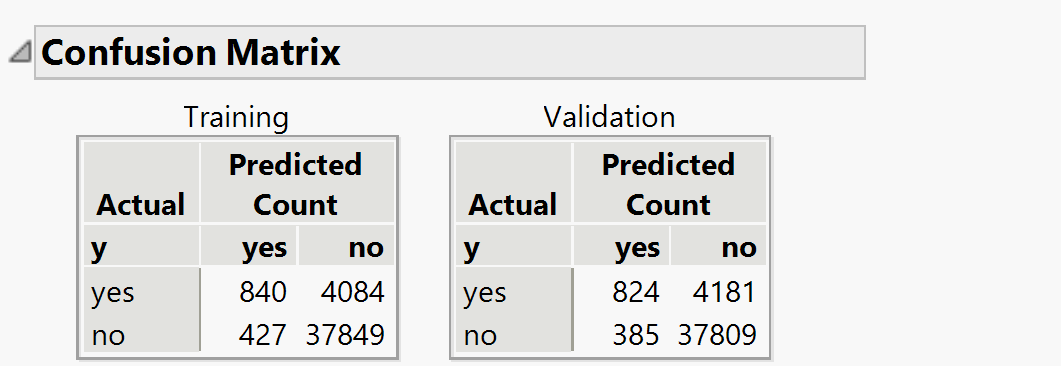
**Baseline Model:** Included all the variables.



Dropped variables in the following order:

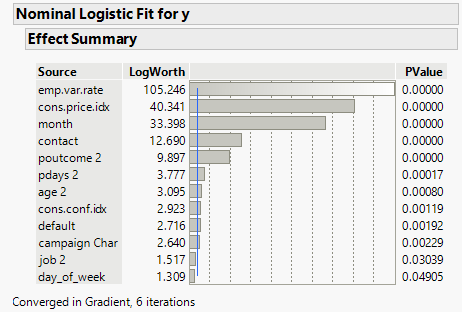
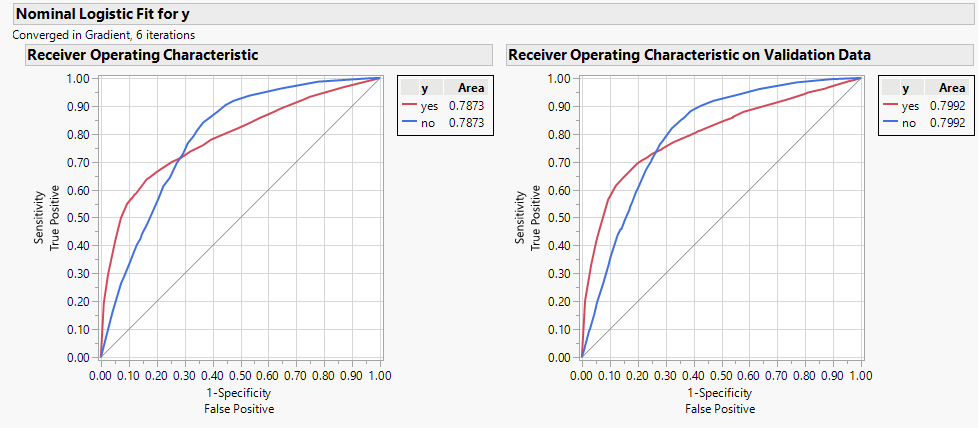
duration (target leakage); education2, previous2, nr.employed, nr.euribor3, default, day\_of\_week, day (p-value>0.05)

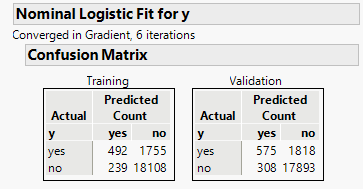
**Final Model:**



AUC: 0.7695, Accuracy: 0.8943, Precision: 0.6816, Recall: 0.1646

**For 41K:** Followed the same preprocessing techniques as of 86K. Included all the variables and ran the baseline model, based on its result dropped the insignificant variables in the following order: duration (target leakage); previous 2, housing 2, loan 2, marital 2, nr.employed, Euribor (p-value>0.05). Even though “day\_of\_week” has p-value>0.05, based on its EDA we felt the variable isn’t playing a significant role in prediction. So, we dropped that variable as well.

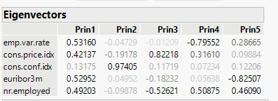
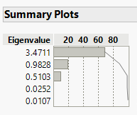
**Final Model:**

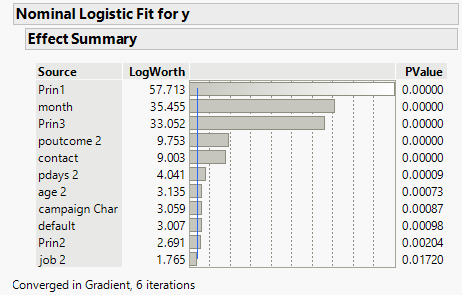
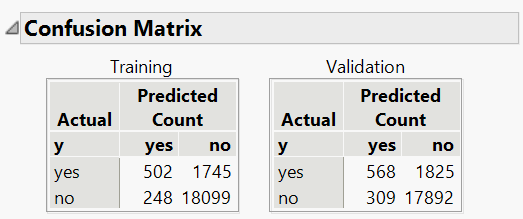


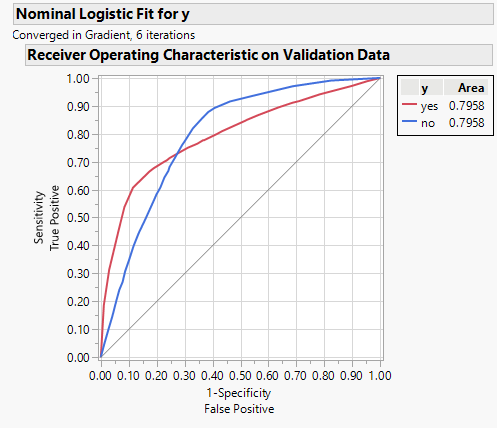
AUC: 0.7991, Accuracy: 0.8968, Precision: 0.6512, Recall: 0.2403

After comparing AUC & Recall of both the datasets, we decided to go ahead with 41K dataset.

**41K (with PCA):** Used the PCA columns instead of the five social economic indicators. Around 95% of variation was captured by the three PCA elements.

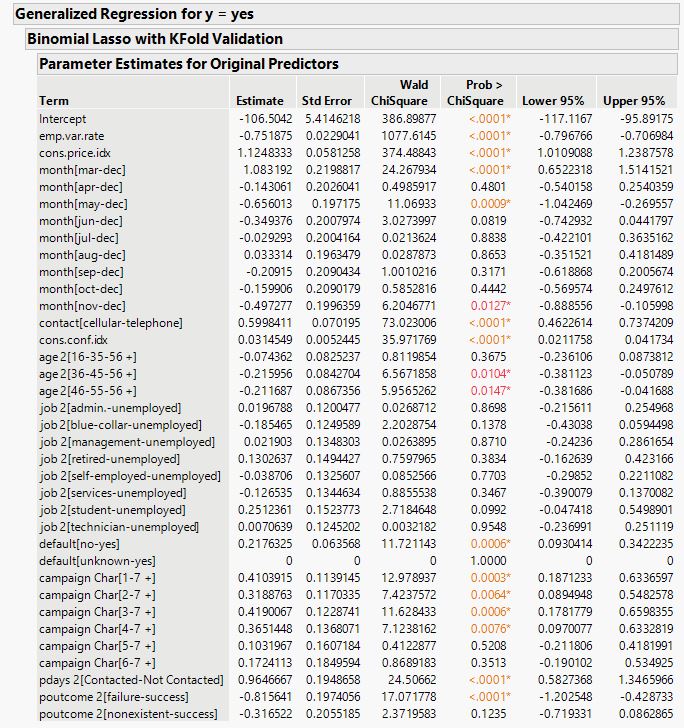


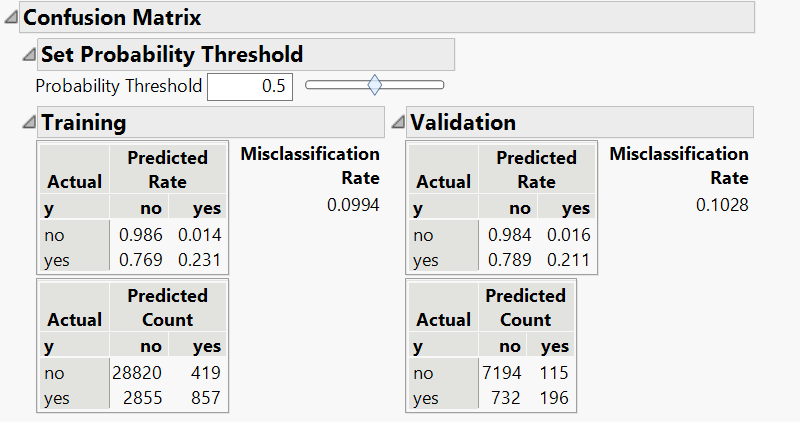
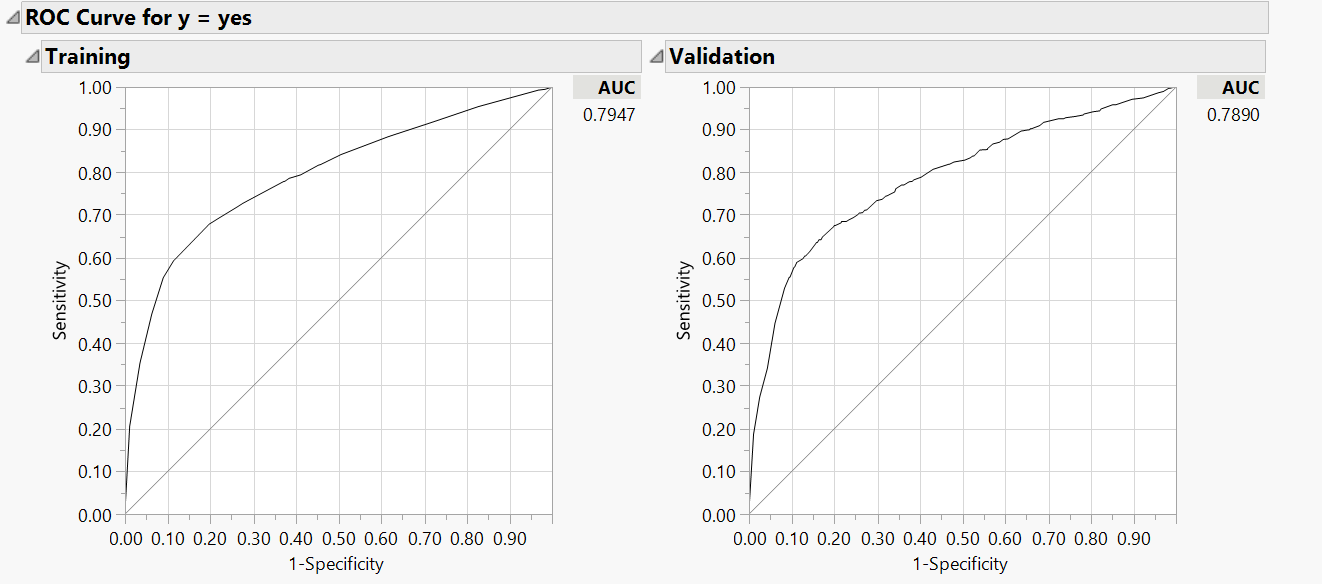
**Final Model:**



AUC: 0.7958, Accuracy: 0.8964, Precision: 0.6477, Recall: 0.2374

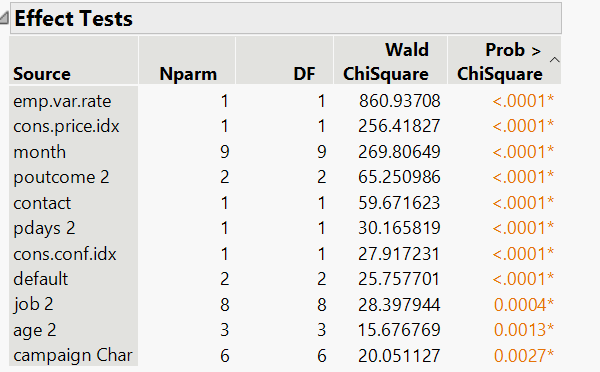
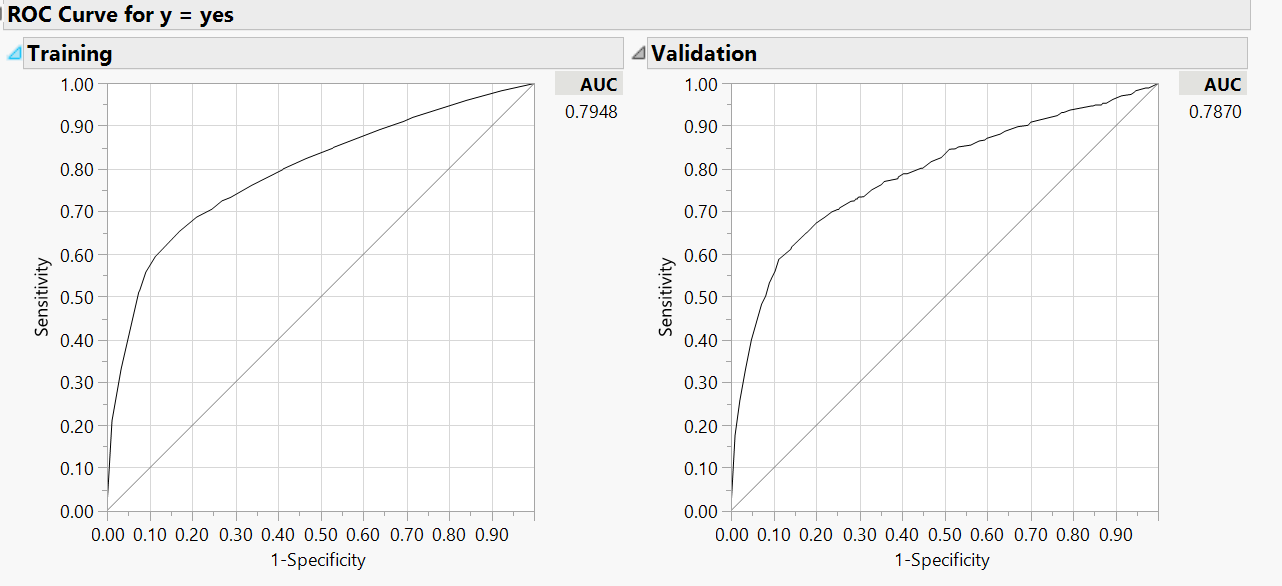
After comparing AUC & Recall of both 41K datasets, we decided to go ahead without PCA.

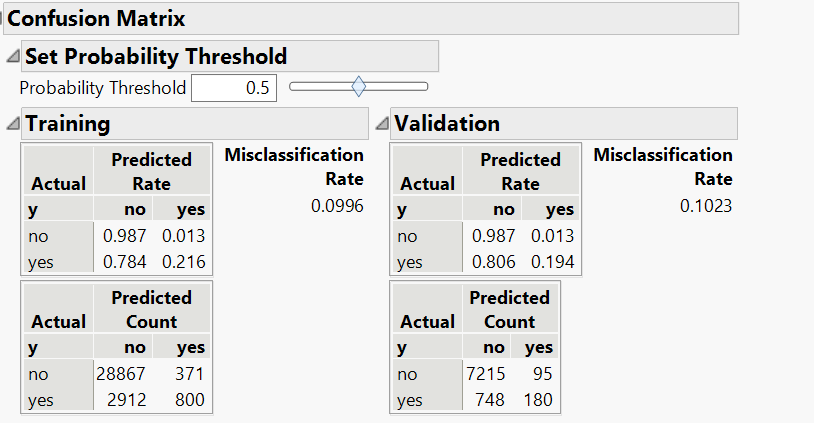
**3.4 Lasso:** Used the variables from 41K nominal logistic regression model.



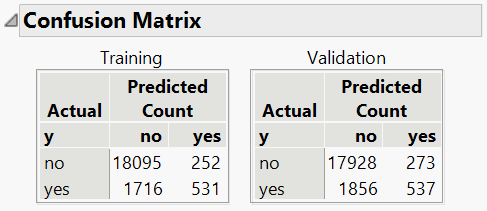
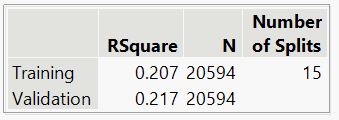
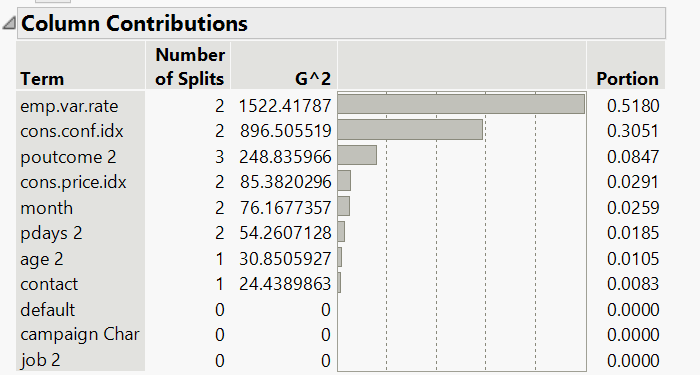
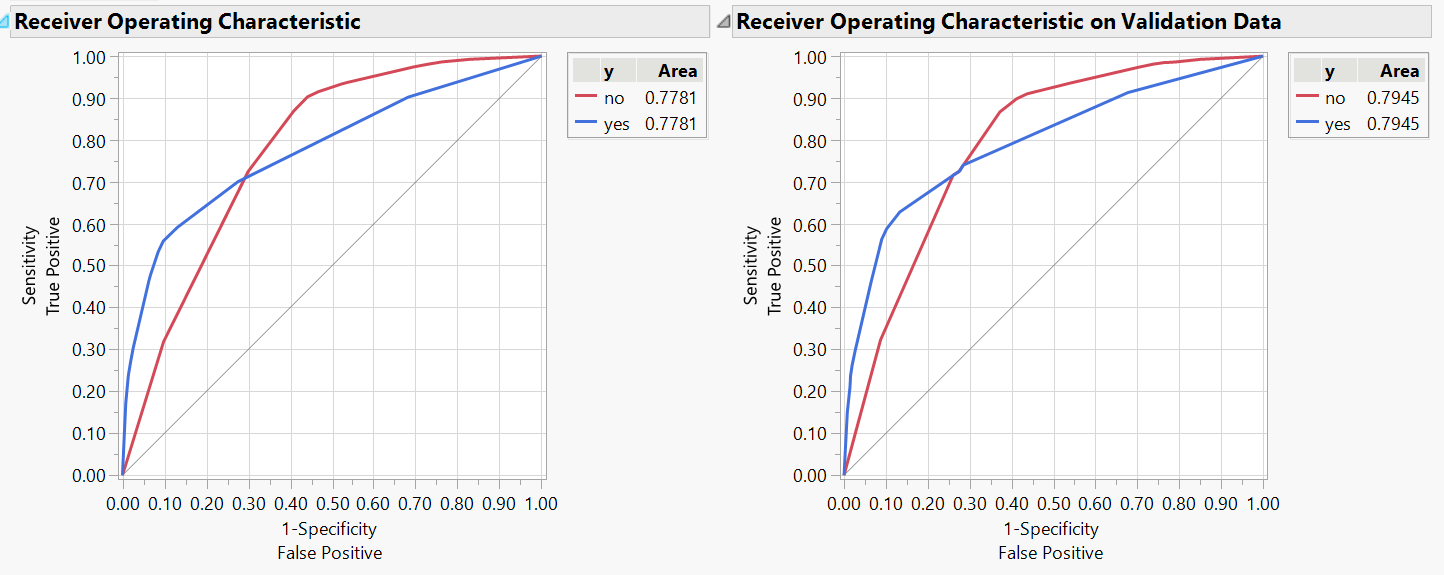
Lasso shrank the Estimate of default[unknown-yes] to zero.

AUC: 0.7890, Accuracy: 0.8972, Precision: 0.6302, Recall: 0.2112

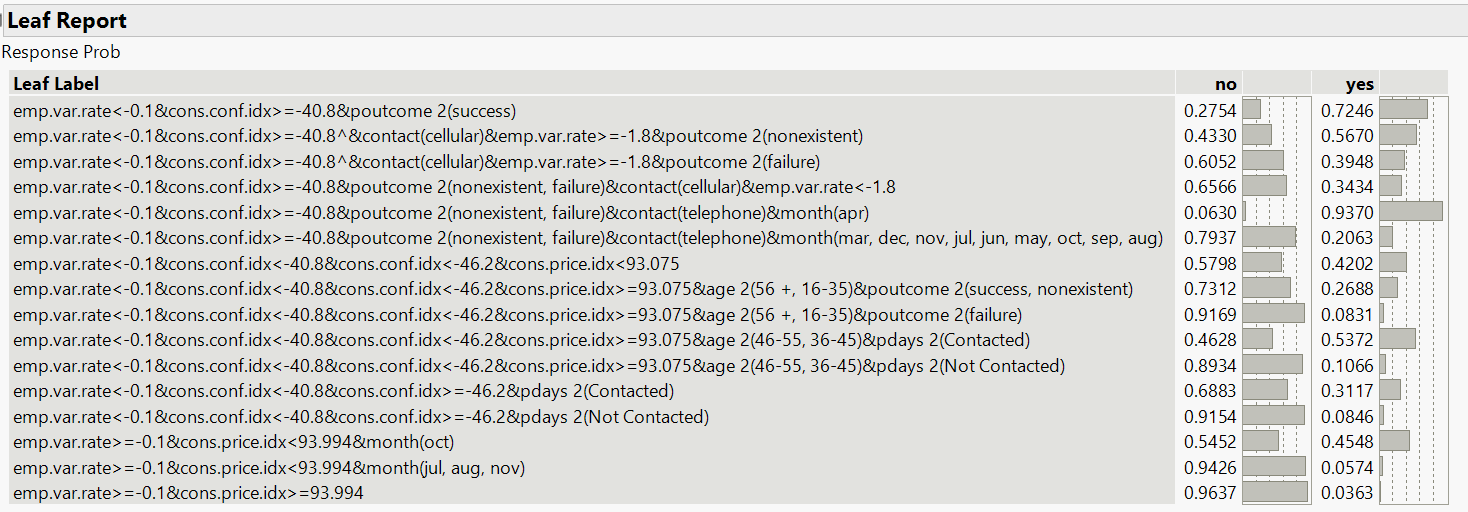
**3.5 Ridge:** Used the variables from 41K nominal logistic regression model.



AUC: 0.7870, Accuracy: 0.8977, Precision: 0.6545, Recall: 0.1940

**3.6 Decision Trees:** Used the variables from 41K nominal logistic regression model.

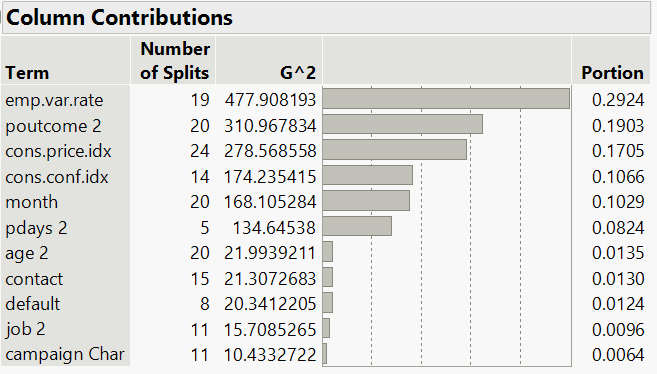
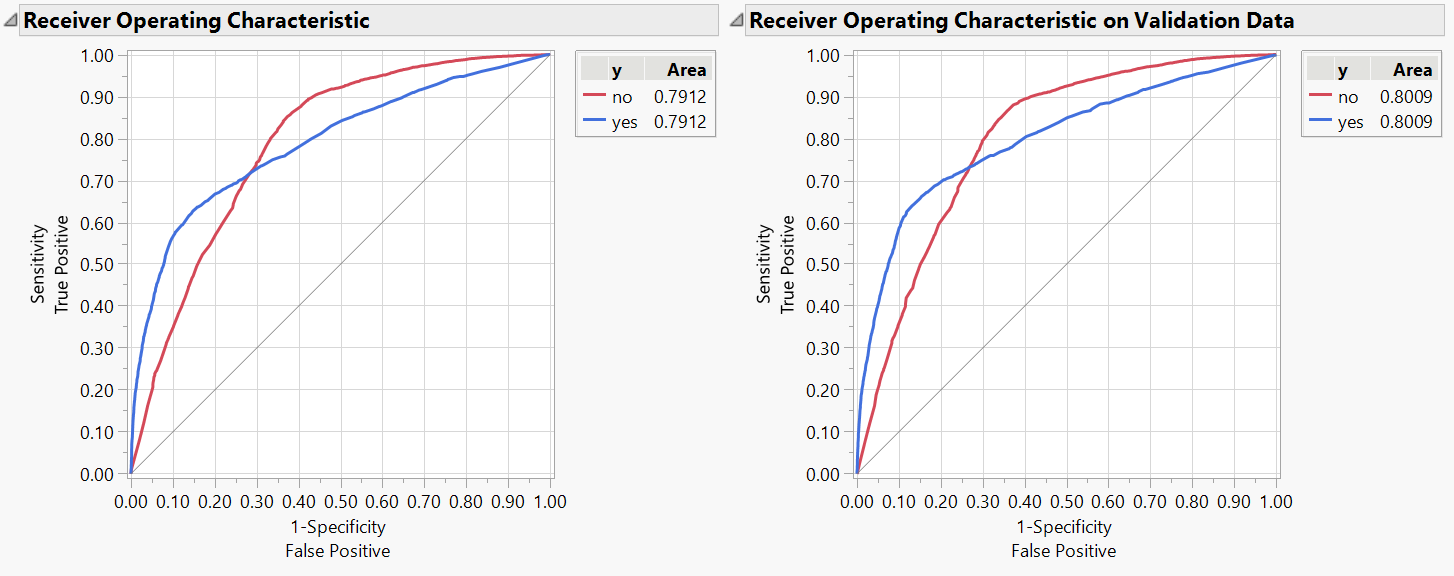
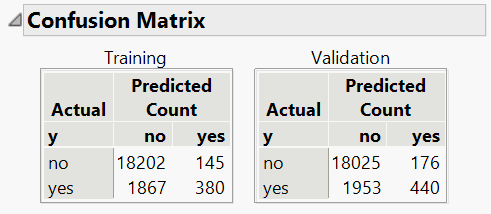
AUC: 0.7870, Accuracy: 0.8977,



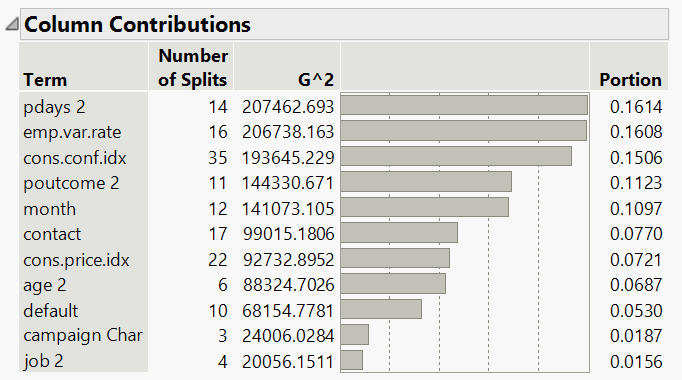
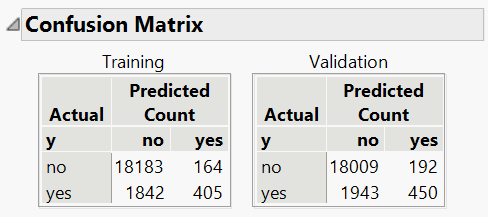
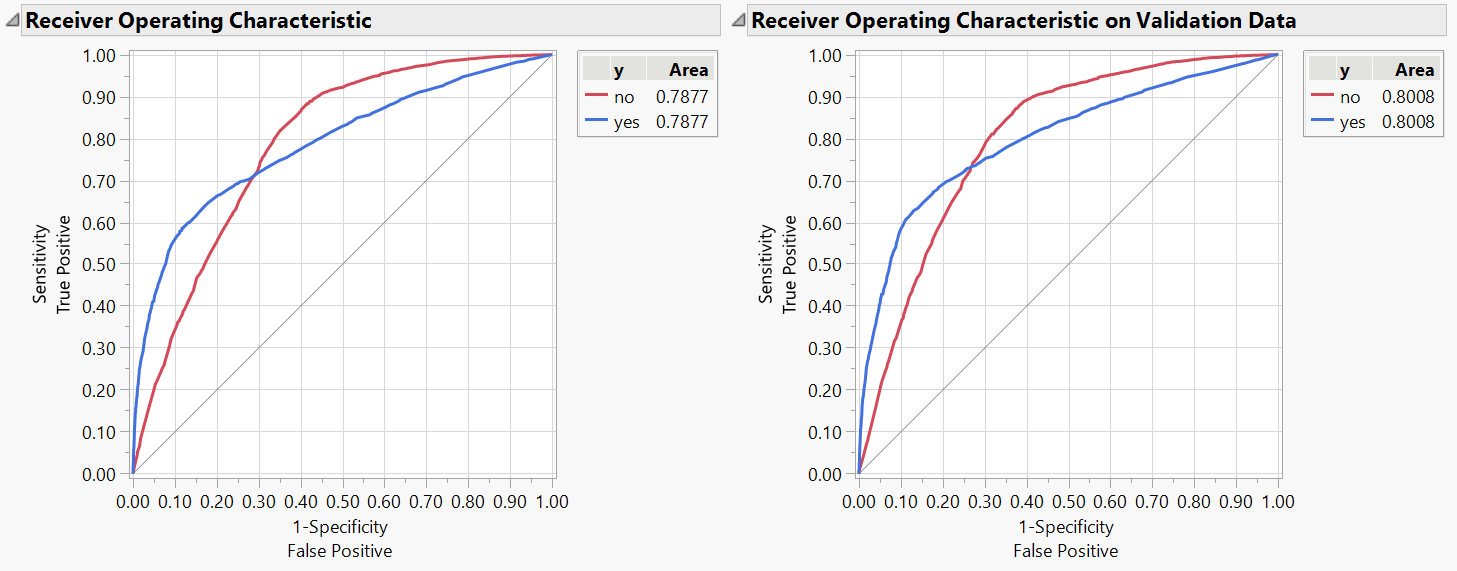
At 15th split, we got the best predicted model. From the leaf report, we can say there is a higher chance of subscribing the term deposit if they fall under the below category:



AUC: 0.7945, Accuracy: 0.8966, Precision: 0.6630, Recall: 0.2244

**3.7 Bootstrap Forest:** Used the variables from 41K nominal logistic regression model.

AUC: 0.8009, Accuracy: 0.8966, Precision: 0.7143, Recall: 0.1839

**3.8 Boosted Tree:** Used the variables from 41K nominal logistic regression model.

AUC: 0.8008, Accuracy: 0.8963, Precision: 0.7009, Recall: 0.188

* 1. **Neural Nets:** Used the variables from 41K nominal logistic regression model.

Implemented the neural net models using various hidden layers and various nodes with the TanH, Linear and Gaussian activation function and the squared penalty function in the fitting options.

Built various Neural Network Models:

1. TanH(3) – 5 seconds (Run Time)

2. TanH(3) TanH2(3) – 9 seconds

3. TanH(6) – 10 seconds

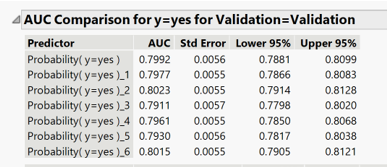
4. TanH(3) Linear(3) Gaussian(3) TanH2(3) Linear2(3) Gaussian2(3) – 10 seconds

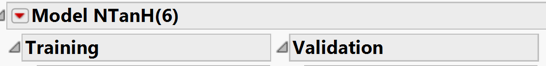
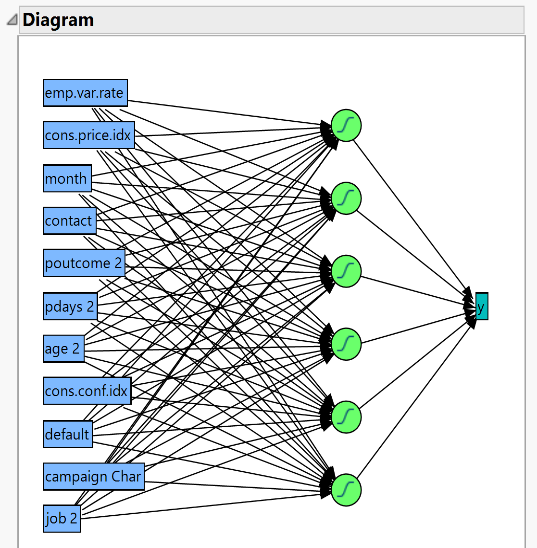
5. TanH(3) Linear(3) Gaussian(3) – 8 seconds

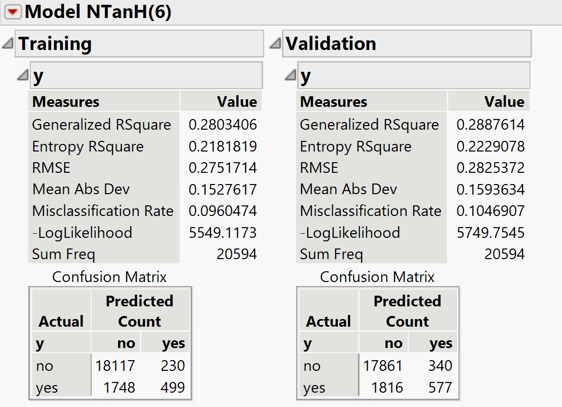
6. TanH(3) Linear(3) TanH2(3) Linear2(3) – 15 seconds

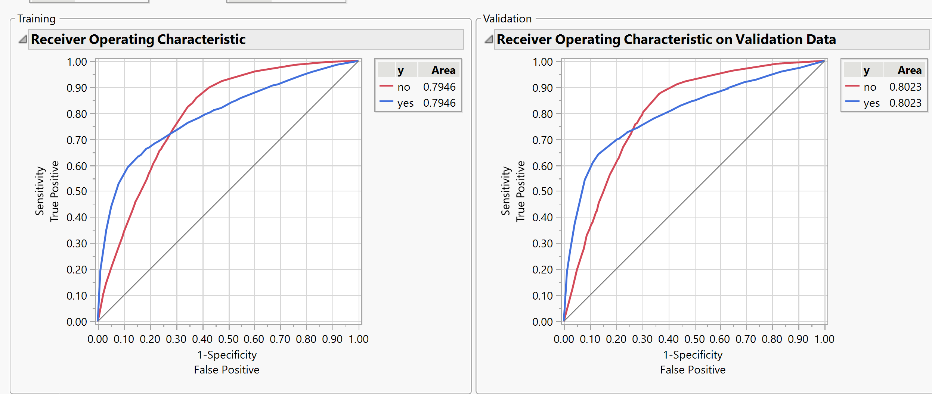
7. TanH(2) Linear(2) Gaussian(2) TanH2(2) Linear2(2) Gaussian2(2) – 12 seconds

**AUC Comparison of these 7 models:**

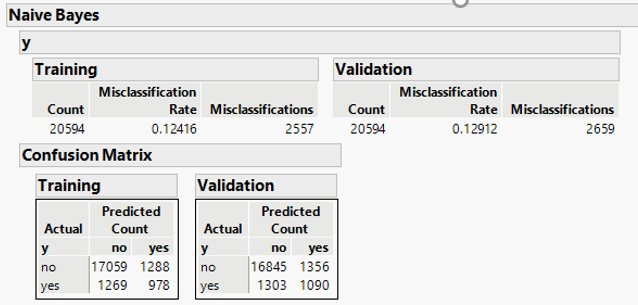
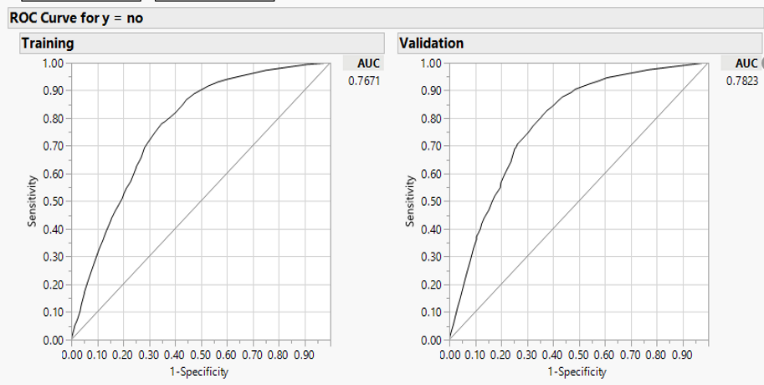


Best Neural Net model is 1 hidden layer, 6 Nodes with TanH activation function (AUC=0.8023).





AUC: 0.8023, Accuracy: 0.8953, Precision: 0.6292, Recall: 0.2411

**3.10 Naïve Bayes:** Used the variables from 41K nominal logistic regression model.

AUC: 0.7823, Accuracy: 0.87, Precision: 0.445, Recall: 0.455

**4. Evaluation**

The performance metric used for this case study is AUC because as we saw in Exploratory data analysis, the dataset we are working with is an imbalanced dataset with “no” being the majority class.

Along with AUC, we are considering Recall also as a measure for choosing the final model because our business case is to identify the clients who are most likely to subscribe to term deposit. The most important factor in our business case is False-Negative which means we predicated it as “No” but actually it was “Yes”. The metric which deals with it is Recall that’s why we chose Recall also as a measure for choosing the final model.

**Model Comparison:**

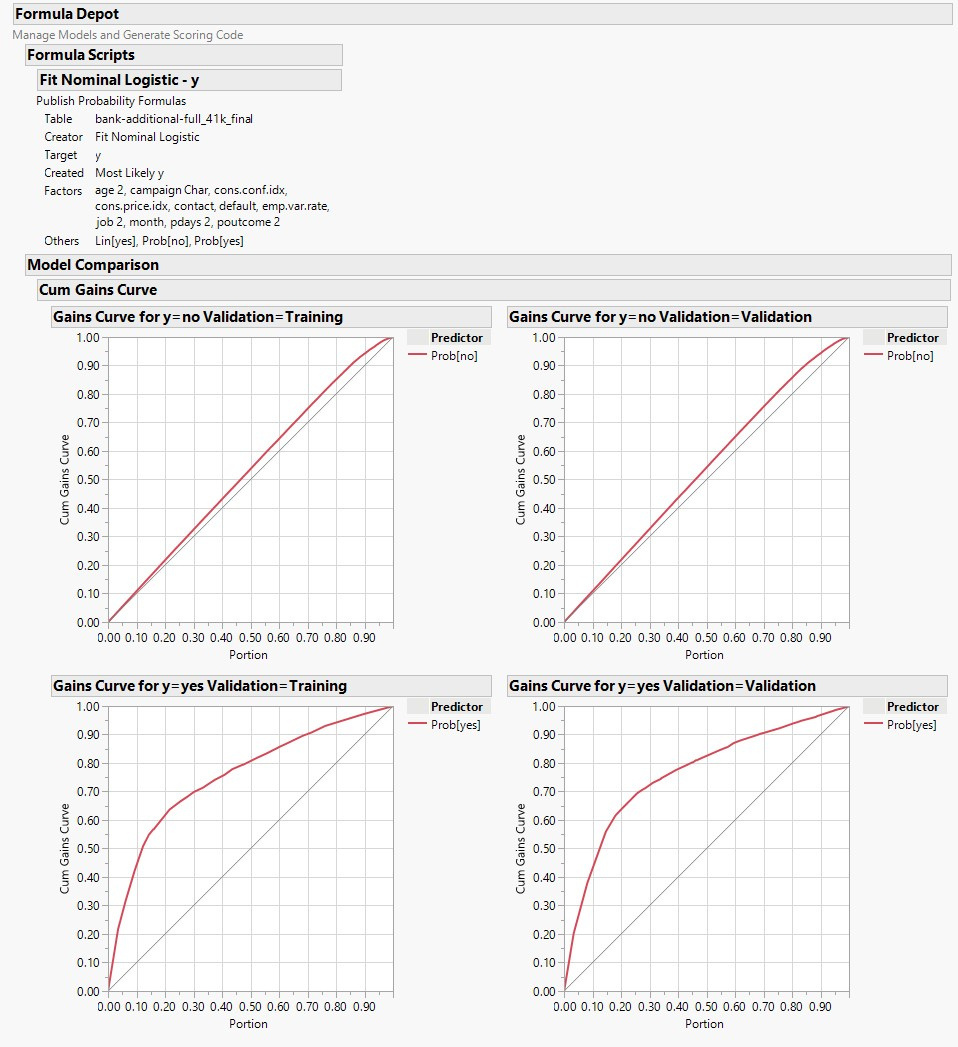
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AUC** | **Subscribers Captured** | **Accuracy** | **Precision** | **Recall** |
| Nominal Logistic Regression | 0.7991 | 575 | 0.8968 | 0.6512 | 0.2403 |
| Lasso | 0.7890 | 196 | 0.8972 | 0.6302 | 0.2112 |
| Ridge | 0.7870 | 180 | 0.8977 | 0.6545 | 0.1940 |
| Decision Trees | 0.7945 | 537 | 0.8966 | 0.6630 | 0.2244 |
| Bootstrap Forest | 0.8009 | 440 | 0.8966 | 0.7143 | 0.1839 |
| Boosted Trees | 0.8008 | 450 | 0.8963 | 0.7009 | 0.1880 |
| Neural Nets: 1 Hidden Layer,  6 Nodes – TanH function | 0.8023 | 577 | 0.8953 | 0.6292 | 0.2411 |
| Naïve Bayes | 0.7823 | 1090 | 0.87 | 0.445 | 0.455 |

As per AUC, Neural Nets with 1 Hidden layer, 6 Nodes – TanH function performed better than rest of the models. NNs (Neural Nets) are able to capture some relationships which other models didn’t capture properly, hence it has higher AUC. Naïve Bayes is the least performing model. Naïve Bayes low AUC might be because of its assumption of conditional independence for every feature which might not be in our case i.e., In our data, there might be some variables that are not independent.

As per Recall, Naïve Bayes captured highest probable subscribers of the term deposit. It’s recall is almost twice than the Neural Net model (highest AUC).

Among the eight, we choose Naïve Bayes to put into deployment because our business case is to capture as many term deposit subscribers as possible which Naïve Bayes clearly does. Also, we feel the Naïve Bayes AUC 0.7823 is approx. to Neural Net model (highest AUC - 0.8023).

By using Naïve Bayes, we captured approx. 62% of the clients saying “Yes” to term deposit while observing only 20% of the overall database. So, we believe we solved the business problem successfully.



After going through all our models, we found variables **employment variation rate, consumer price index, month, contact, poutcome** play a significant role in predicting “yes”. So, we are recommending the bank to focus more on these predictors when targeting the clients for term deposits subscription.

* "**Employment Variation Rate**" and "**Consumer Price Index**" tells the important economic and social status of the targeted lead. A stable employment rate denotes a stable economic environment. An increase in CPI increases the subscription.
* "**Month**" is an important factor determining the success of the campaign as it indicates the months with the highest conversion rate probability.
* Clients with cellular **contact** are most likely to subscribe to the term deposit as compared to telephone contact. The Comfort level of the leads is more on a cellphone.
* **Poutcome:** It is a categorical variable that specifies the outcome of the previous marketing campaign (‘failure’, ‘nonexistent’, ‘success’). People with the response of "Yes" in the previous campaign are the most likely audience for subscribing to the term deposit.

**5. Deployment**

Performing data analytics is the most important part of any industry with rising competition. Marketing analytics campaigns are used by the majority of business organizations. Our model can be used in the real world to help the bank get their targeted customers with minimum investment and get the maximum conversion rate on the leads predicted by the model.

**6. References**

Dataset source:

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

<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

Reference Papers:

* S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
* S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October 2011. EUROSIS. [bank.zip]