# **Predicting Success of Bank Direct Marketing By MSAS**

ID:773950464

NAME: Yong Wan

UPI: wany991

## 1.Business and Situation understanding.

“First is developing an understanding of the application domain and the relevant prior knowledge and identifying the goal of the KDD process from the customer’s viewpoint.” (Fayyad et al., 1996)

Banks are a data-intensive industry and data is a crucial factor of competition, whether it is an offline business or online business. For the banking industry, it is not the most critical to choose the platform or technology, the key factor is to understand the demand of business. Mining the rules in the data banks had is a key competitive ability in the market. Citibank has been a good example. It has signed an agreement with IBM to explore the potential uses of IBM Watson in 2012. IBM Watson was able to be used to process a large amount of customer data and analyze customer needs.

It Has made people have no difference in distance in communication by the rapid development of the technology. It also enables consumers to change their roles e easier, from passive roles in tradition to active creators. Therefore, how to find new profit growing points, needs of customers and comprehensive personalized services is a big problem all bank face today. It is lucky that the development of information technology will be able to convert passive customer service to proactive customer care. telemarketing has become a very important marketing event.

Telemarketing first appeared in the United States in 1970s. It grew quickly With the popularity of mobile phones. Enterprises have begun to adopt this new marketing method, constantly expanding the size of customers, tapping the potential value of customers, forming a customer-centered market.Generally speaking, centralized customers are managed by call center by remote operation. The center communicates with customers through various channels. In the family, telephone (mobile phone or fixed telephone) is the most widely used. We would think that telemarketing means marketers are selling through Hawking. In fact, the definition of telemarketing will be more extensive. According to the difference between sender and receiver, telemarketing can be divided into two categories: exhaled telephone and access telephone.

1. call out

The reason for outgoing calls is active marketing because the originator is a company. Main business of exhaling business Including: Market Research and commodity sales.

1. call in

The founder of marketing is customer, so telephone is also known as response to telemarketing. The boss of the telephone company. Including: air ticket booking and dismantling.

Telephone as a marketing channel has its own advantages. Of course, there are also some shortcomings.

The advantages of marketing include:

1. providing instant feedback from customers is conducive to the interaction of marketers in marketing activities.
2. this is a fruitful marketing channel.
3. it is a relatively sensitive marketing channel. Through the timely feedback from customers, telephone salesmen can use electricity.
4. Adjust and change the content of the word accordingly.

The main disadvantages of telemarketing are as follows:

1. respondents often feel insulted or offended. Because telemarketing usually does not require prior approval.

The consent of the interviewer will make the client feel offended and may have an impact on marketing activities.

1. you can't see each other. You can't see each other's performance except voice.

The data set from Portugal has 41188 telephone contact records. The goal is to sell time deposits through telephone. So it is the main task to predict whether customers subscribe to time deposits (variable y). According to the characteristics of the data set and the goal to be achieved, the data is pre-processed and different DM models are selected for data mining. Then the model is evaluated, and the test set is used to forecast, providing bank telephone marketing related suggestions. Choose the best prediction model for landing application.And put forward some suggestions for customer service center telephone marketing.

### Base on the data, there were only 10% of success in the past three years, it means there is a massive gap between success and unsuccess. Also, we should do something to improve it to decrease the phone call times due to our customers may be annoyed and increase the profit for banks due to the phone agents get more success by same phone outs as before. So we aim to decrease 40% of the total amount of phoning out and at max decrease 5% of the total of successful calls. This means we will get at least 20% of success of phoning ,double than before. Also, it means we should get at least 90% correct prediction on customers who's original answer is Yes and nearly 40% accurate prediction on the customers who's original answer is No.

### This objective is to predict the success of phoning a customer then decided whether he was worth to call.

### In conclusion, the main tasks includes:

### 1.Describe the methods for model assessment (for example, accuracy, performance, etc.).

### 2.Define benchmarks for evaluating success. Provide specific numbers.

### 3.Define subjective measurements as best you can and determine the arbiter of success.

### 4.Consider whether the successful deployment of model results is part of data mining success. Start planning now for deployment.

### Because this project is so complicated, we have to make a good plan and arrangement for banking business.

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Time | Resources | Risk |
| Business understanding | 0.5 weeks | Analysis specialist and Advance manager of users | The bias of demand of banking business |
| Data understanding | 2 weeks | All analysts and all users | The misconceive of data |
| Data preparation | 4 weeks | Data managers of users and analysts | Missing key data |
| Data transformation | 1 week | Data transformation specialists | Losing key data |
| Data-mining method(s) selection | 3.5 weeks | Analysis specialist and Advance manager of users | Missing more efficiency Method |
| Data-mining algorithm(s) selection | 2 weeks | Analysis specialists and Advance manager of users | Choosing wrong algorithms |
| Data Mining | 2 weeks | Analysis specialists and All users | Meeting unpredictable things |
| Interpretation | 1 week | All analysts and key users | Misunderstanding conception |

## 2. Data understanding.

This dataset comes from the Bank Marketing Data Set (https://archive.ics.uci.edu/ml/datasets/bank+marketing) in the UCI Machine Learning Library [Moro et al., 2014]. These data relate to direct marketing activities of Portuguese banking institutions. These direct marketing campaigns are based on the phone. In general, the banking agency's customer service staff needs to contact the customer at least once to find out if the customer will subscribe to the bank's products (time deposits). Therefore, the task corresponding to the data set is a classification task, and the classification target is to predict whether the customer is (yes) or not (no) subscribes to the time deposit (variable y). We chose one Csv file named bank-additional-full.csv to mining : it contains all the samples (41188) and all feature inputs (20), sorted by time (from May 2008 to September 2010). Actually, we can not do more in collecting data phrase, we only able to use it in the most possible.

From the website we know that there were more than 150 properties in the beginning. We actually got the data which was performed manual feature selection from the telemarketing data obtained on the machine learning website. Some bank managers (domain expert) to define a set of related attributes, choosing from the original 150 features. Each record includes an output target, a contact result ("failed", "success"), and a candidate input variable. Input variables include telemarketing attributes (such as age, job type, contact type, previous marketing results, etc.). These records (such as the characteristics of social and economic impacts, the rate of change in unemployment rates) are collected through the Portuguese Statistics website data sources for a large number of potentially useful external data.

There are 20 input variables: age,job ,marital ,education ,default(has credit in default）, housing(has housing loan), loan(has personal loan),contact,month, day\_of\_week, duration(last contact duration, in seconds ), campaign, pdays,previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx,euribor3m,nr.employed

Classified, the results are shown in Table 3-1:

|  |  |
| --- | --- |
| bank client data | age,job ,marital ,education ,default, housing:, loan |
| elated with the last contact of the current campaign | loan,contact,month, day\_of\_week, duration |
| other attributes | campaign, pdays,previous, poutcome |
| social and economic context attributes | emp.var.rate, cons.price.idx, cons.conf.idx,euribor3m,nr.employed |

And there is only one output variable (desired target): y - has the client subscribed a term deposit? (binary: 'yes','no')

As we can see from Figure ,compared with young people, older people are more likely to order deposits. Blue-collar workers are the least likely to subscript deposits. Retirees are the most likely to have deposits. Singles or divorcees are more likely to script deposits than married people. The education degree is also an important factor. People who have more education is more likely to order deposits. Those with credit default records basically do not order deposits. When there are houses and loans, it seems that there is no impact on whether to order deposits.



In Figure ,basic on the data from last contacting, calling by cellular is more efficiency than by telephone,and contacting in second half year is more productive than in first half year.Constant, phone in which day in a week,it seems it does not matter,every day is almost same.But the duration in every call is very important, if a client have more time on the conversation,they are more easily to order deposits,however, we could not know how long the client will be on line before we predict the success rate.So we could not use this field to predict, but we can speak more with every customer as we know it is very useful to enhance client order our product.

## 

From Figure ,we may know that the possibility of purchasing in customers who were contacted by the same salesperson is far greater than those were called by different salespersons. As for the day gap of two times contacting a customer, there is no impact on whether to order the product. It is worth noting that there are abnormal data in the interval days, is 999, according to our further analysis, 999 means that we have never contacted this customer before.

The likelihood of a customer ordering a deposit increases as the number of contacts with the customer increases during the sales activity, and customers who have previously subscribed to a deposit are more likely to re-order than other customers.

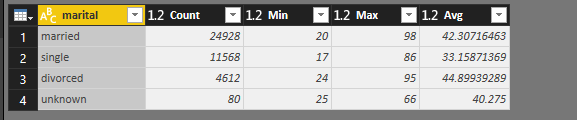


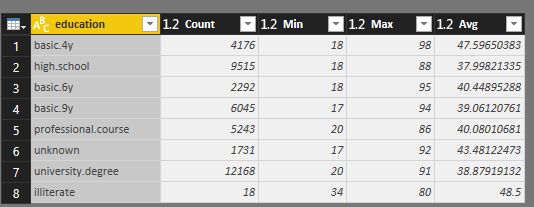
From Figure ,there are five fields to indicate social and economic context attributes:employment variation rate idx, consumer price index,consumer confidence index,euribor 3 month rate,number of employees. They are both very important for predicting weath a customer will subscript a deposit.

As we browsed these data more and more, we found some data which may be was wrong,for example,in Figure, Somebody aged 23 is retired，We think it is impossible,so we should check it again and there is a kind job named ‘admin.’ that we think it is input error, it should be ‘admin’.

## 

Also we may know before,there are some data which labeled unknown in different fields,from Figure...,these data which were labeled unknown is less than 5 percentage in our whole data set, in the next step,we can simply to discard them.

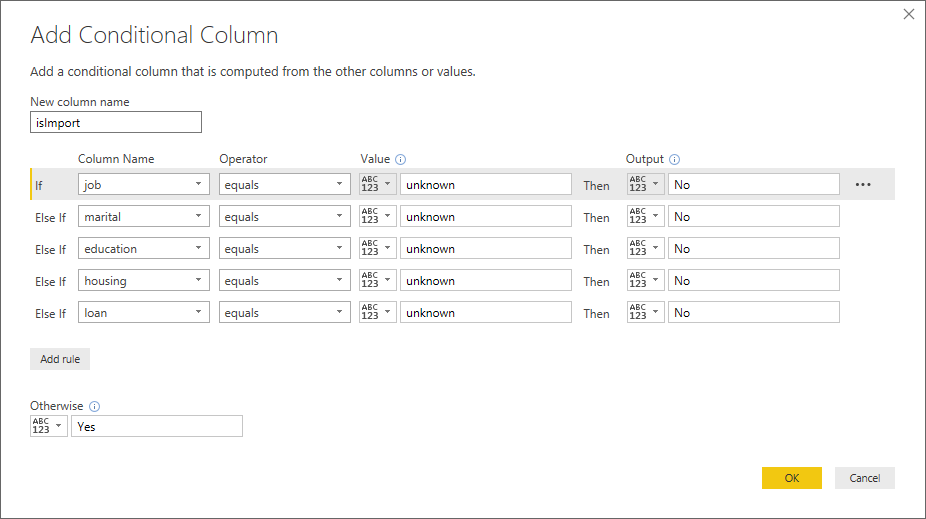


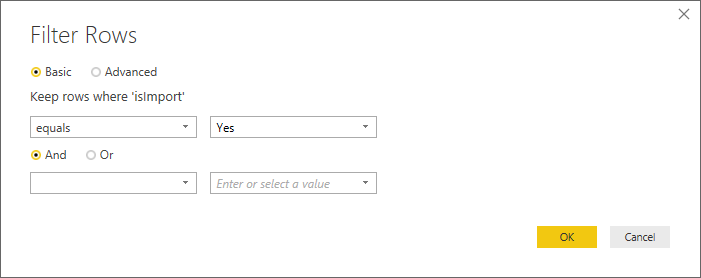


## 3.Data preparation

“Third is data cleaning and pre-processing. Basic operations include removing noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, and accounting for time-sequence information and known changes” (Fayyad et al., 1996)

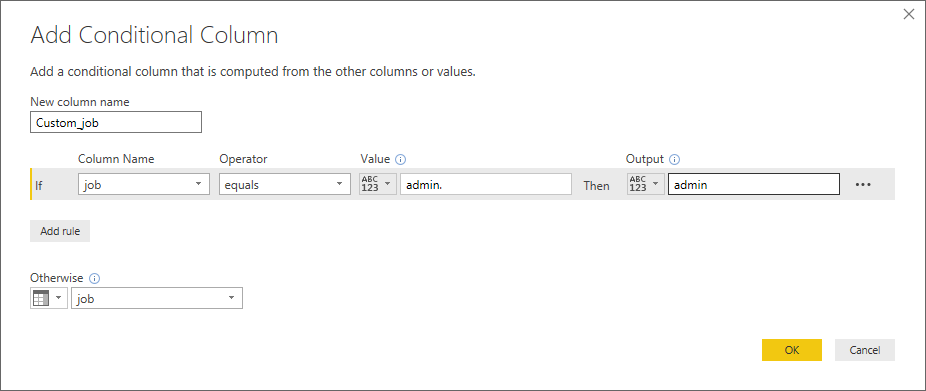
As we know there are some data which were labeled ‘unknown’, we want to discard them,so at first, we create a new column named ‘isImport’ by condition ‘if job=’unknown’ or marital=’unknown’ or education=’unknown’ then ‘No’ else ‘Yes’ in Figure . Then we filter rows by keeping rows where isImport=’Yes’ in Figure .





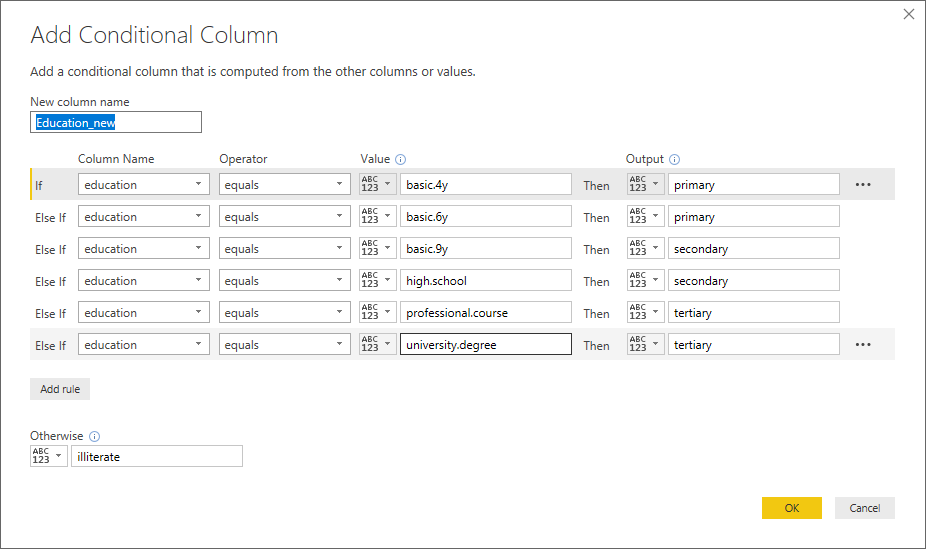
(Figure 2)

As we found there are some error data in field job in step 2, we change ‘admin.’ to ‘admin’ by three steps including add a field “custom\_job” and remove original field “job”,then rename the new field “customer\_job” to “job”,these all show in Figure .





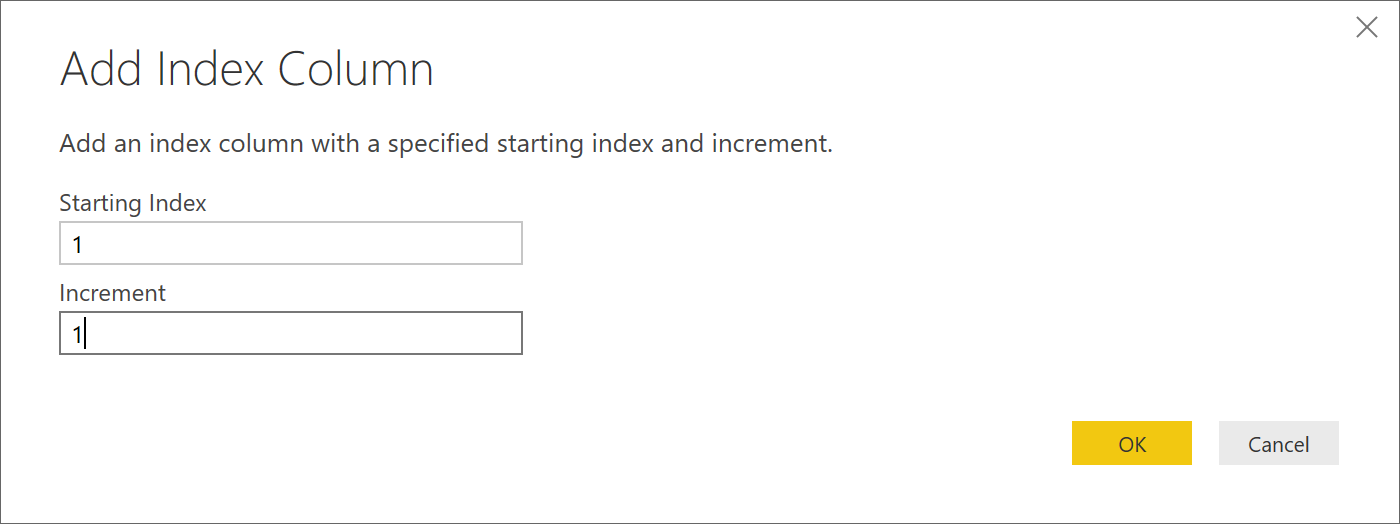
For education,there are too complicated , it is best if we merge them to three different values (‘illiterate’,‘primary’,’secondary’,’tertiary’) like most of people classify it.



From step 2,we know there are some error data related fields job and age,we should filter them;

Due to there is not a primary key in original data, we decide to add a field named RecordID by function @Index as a primary key as showing in Figure 4..

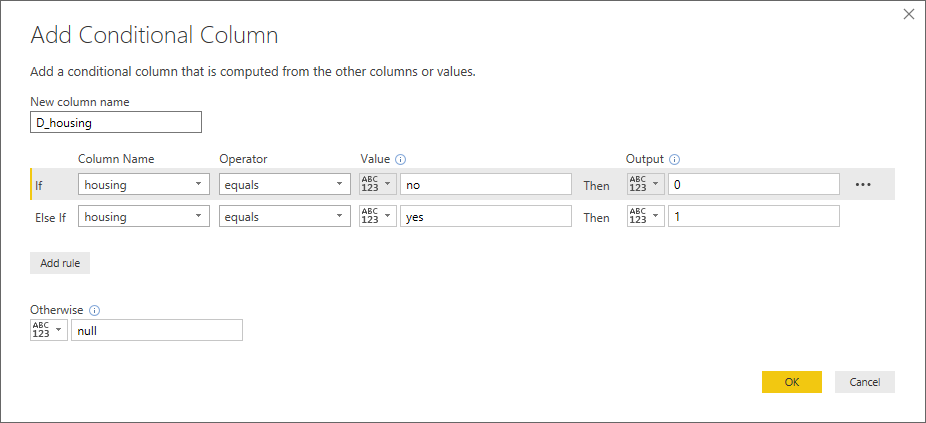
It will be convenient to do merging operation.



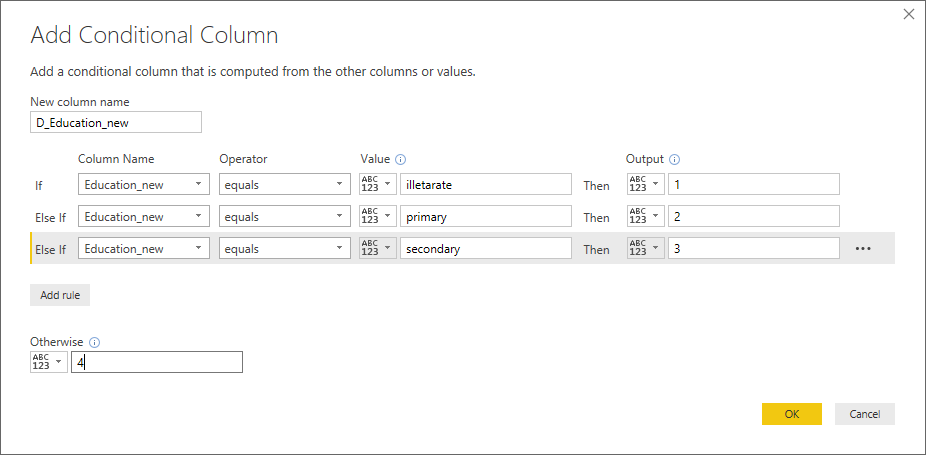
(Figure 4)

Some models only can be calculated with quantify variables,so we have to convert categorical variables to quantify variables. Actually, the categorical variables can be further divided into three kinds of quantify variables including two categorical variables,ordered categorical variables and unorder categorical variables.

In here,housing and loan are both two categorical variables,so we simpley convert them to (0=’No’,1=’Yes’,null=’unknown’)



The variable education\_new can be considered as an ordered categorical variable,we want to convert ‘illiterate’,’primary’,’secondary’,’tertiary’ to 1,2,3,4

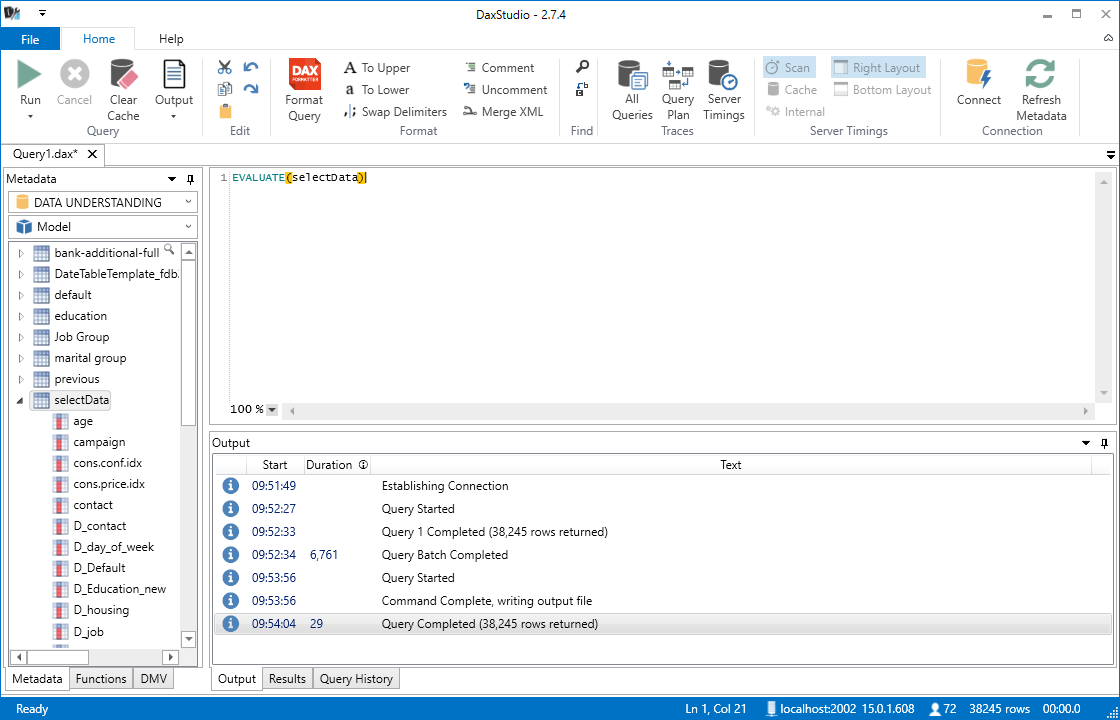


**In general, the variable marital can be thought of as an unordered categorical variable. It should be noted that encoding can be performed using dummy variables. N categories need to set n-1 dummy variables. The variable marital is divided into divorced, married, single, encoded using two dummy variables M1 and M2.**

|  |  |  |
| --- | --- | --- |
| marital | M1 | M2 |
| divorced | 0 | 0 |
| married | 1 | 0 |
| single | 0 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| FieldName | type | value | Value Remark |
| D\_default |  | 0,1,2 | 1:yes,0:no,2:unknown |
| D\_housing |  | 1,0 | 1:yes,0:no |
| D\_loan |  | 1,0 | 1:yes,0:no |
| D\_Education\_new |  | 1,2,3,4 | 1:illiterate,2:primary  ,3:secondary,4:tertiary |
| D\_job |  | 1,2,3,4,5,6,7,8,9,10 | 1:admin,2:blue-collar, 3:entrepreneur ,4:housemaid, 5:management,6:retired, 7:self-employed,8:services, 9:student,10:technician |
| D\_contact |  | 1,0 | 1:cellular,0:telephone |
| D\_month |  | 1,2,3,4,5,6,7,8,9,10,11,12 | 1:January,2:February,3:March,4:April,  5:May,6:June,7:July,8:August,  9:September,10:October,11:November,12:December |
| D\_day\_of\_week | 1 | 1,2,3,4,5 | 1:mon,2:tue,3:wed,4:thu,5:fri |
| D\_poutcome | 0 | 1,2,3 | 1:failure,2:nonexistent,3:success |

After data preparing,we use DaxStudio output “selectData” to “bank after data preparing.csv”.



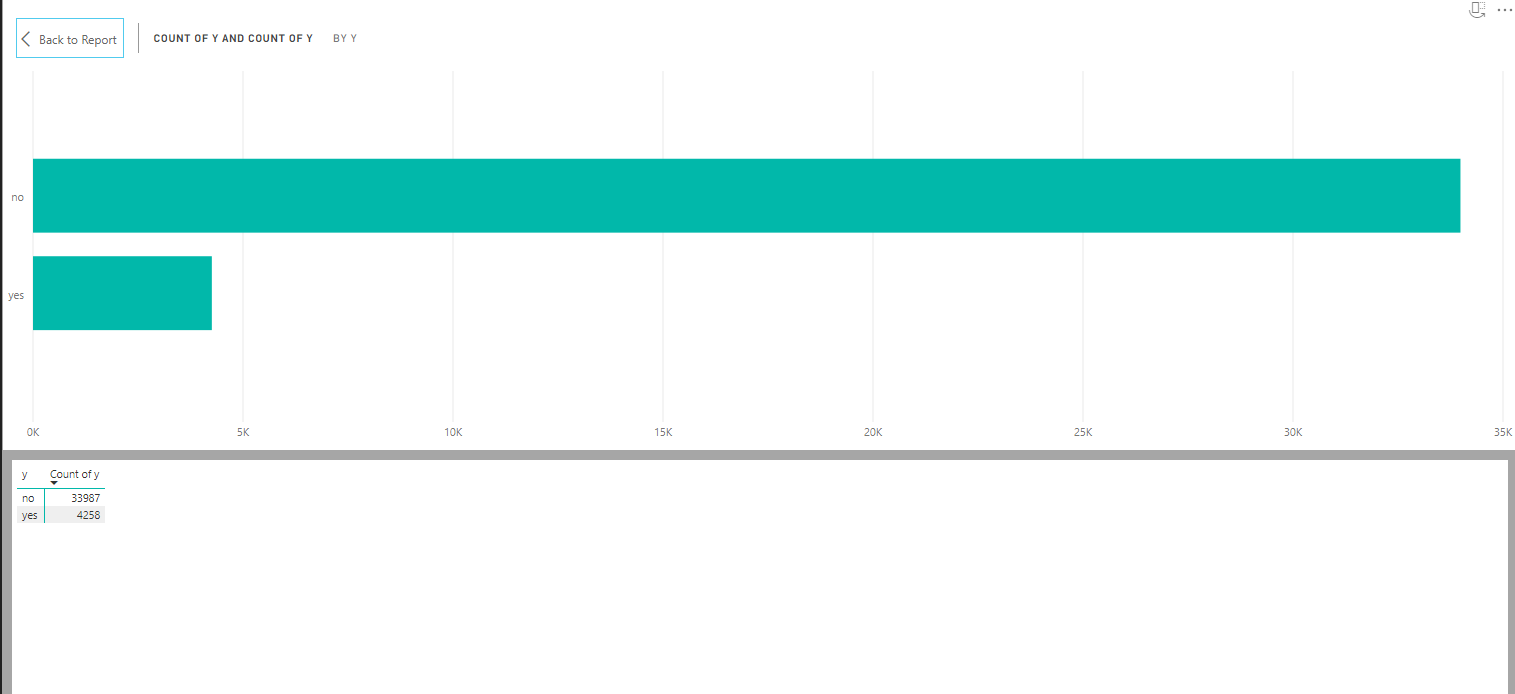
## 4.Data transformation

“Fourth is data reduction and projection: finding useful features to represent the data depending on the goal of the task. With dimensionality reduction or transformation methods, the effective number of variables under consideration can be reduced, or invariant representations for the data can be found.” (Fayyad et al., 1996)

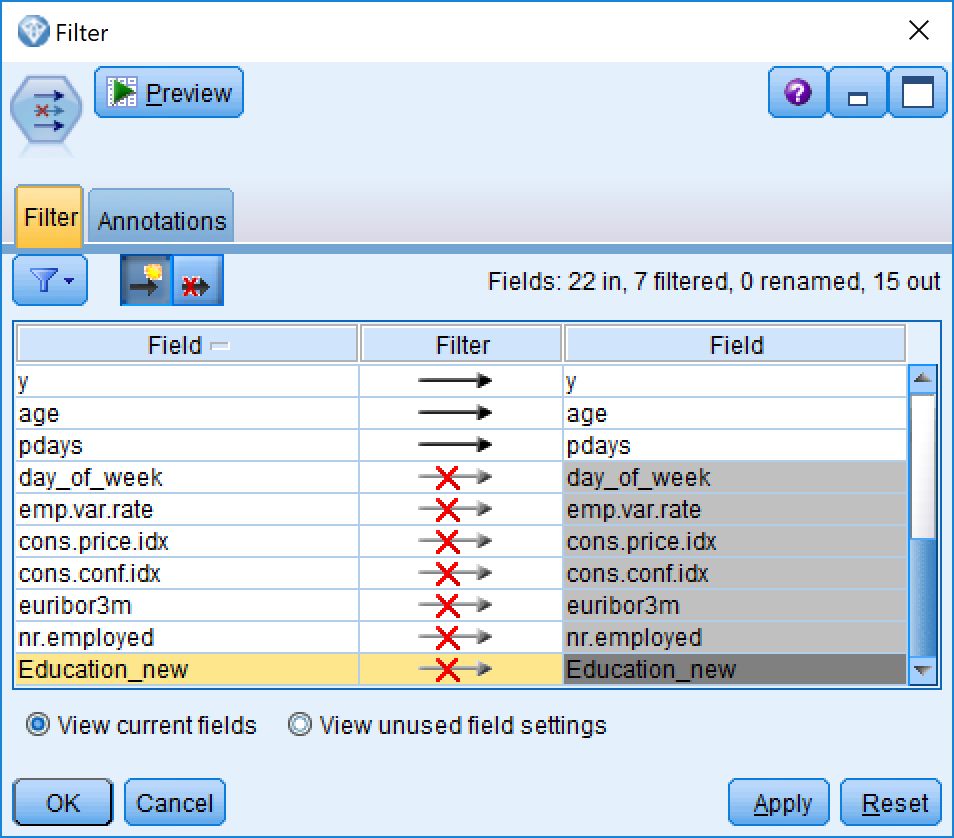
### 4.1 Reduce the data

When we generate a distribution graph of y, we saw in Figure 7 that the data is

skewed (12% for Y and 88% for No). If a model did nothing, just sample say all of them are No for, it will be got 88% accuracy, I think it is useless for prediction. To fix this, we use ”Balance Node (Reduce)” to make the data approximately 50/50 as showing in Figure 8.

(Figure 8)

As there are too many null data on some fields, such as day\_of\_week, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, we discard these six fields as we can see in Figure 9。



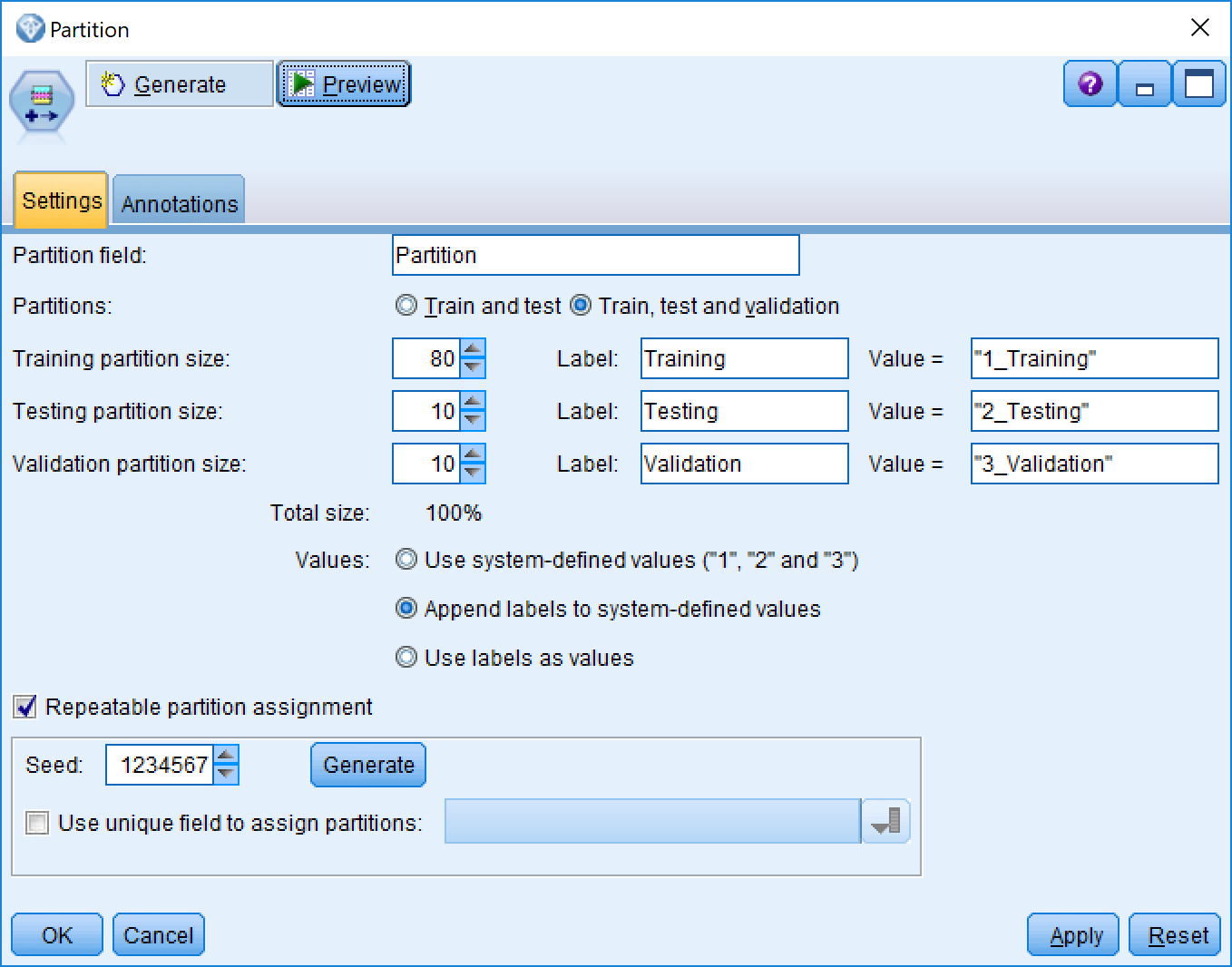
(Figure 9)

### 4.2 Project the data

Through “Feature Selection Model”, we decided to discard default field as it shows single category too large

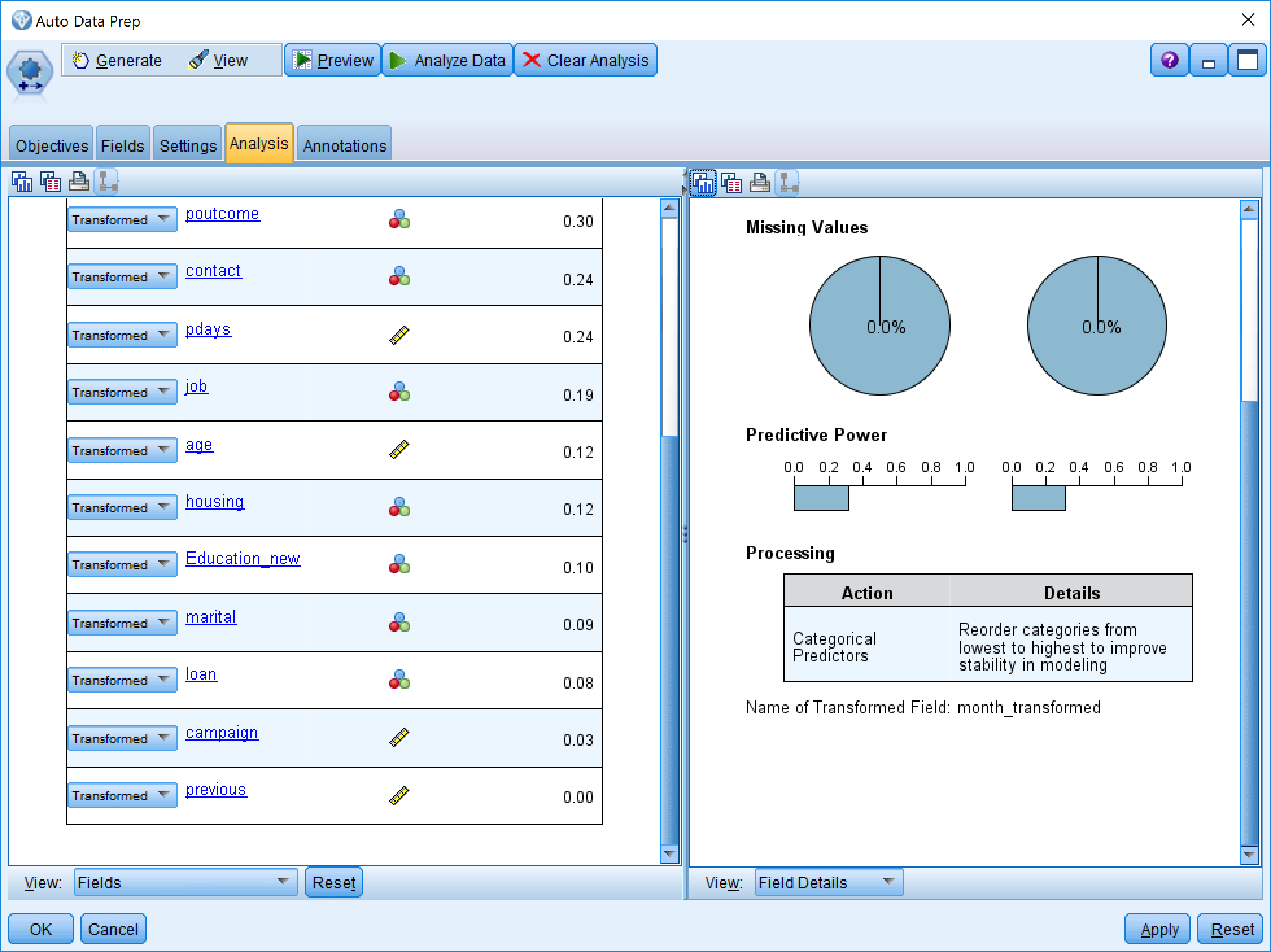
For the duration, this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should be discarded as we want to have a realistic predictive model.

In Figure 10, we use ”Partition” to random split the data to 80% for(training) ,10% to testing and 10% to validation.



(Figure 10)

After that, we used “Auto Data Prep” to convert all of categorical or nominal field to continuous data in Figure 11. It will be benefit on model building.



(Figure 11)

## 5.Data-mining method(s) selection

“Fifth is matching the goals of the KDD process (step 1) to a particular data-mining method. For example, summarization, classification, regression, clustering, and so on, are described later as well as in Fayyad, Piatetsky-Shapiro, and Smyth (1996).” (Fayyad et al., 1996)

the data mining methods we chose included

Python Model Group:Rondam Forest,Rondam Trees, Bayesian Network;

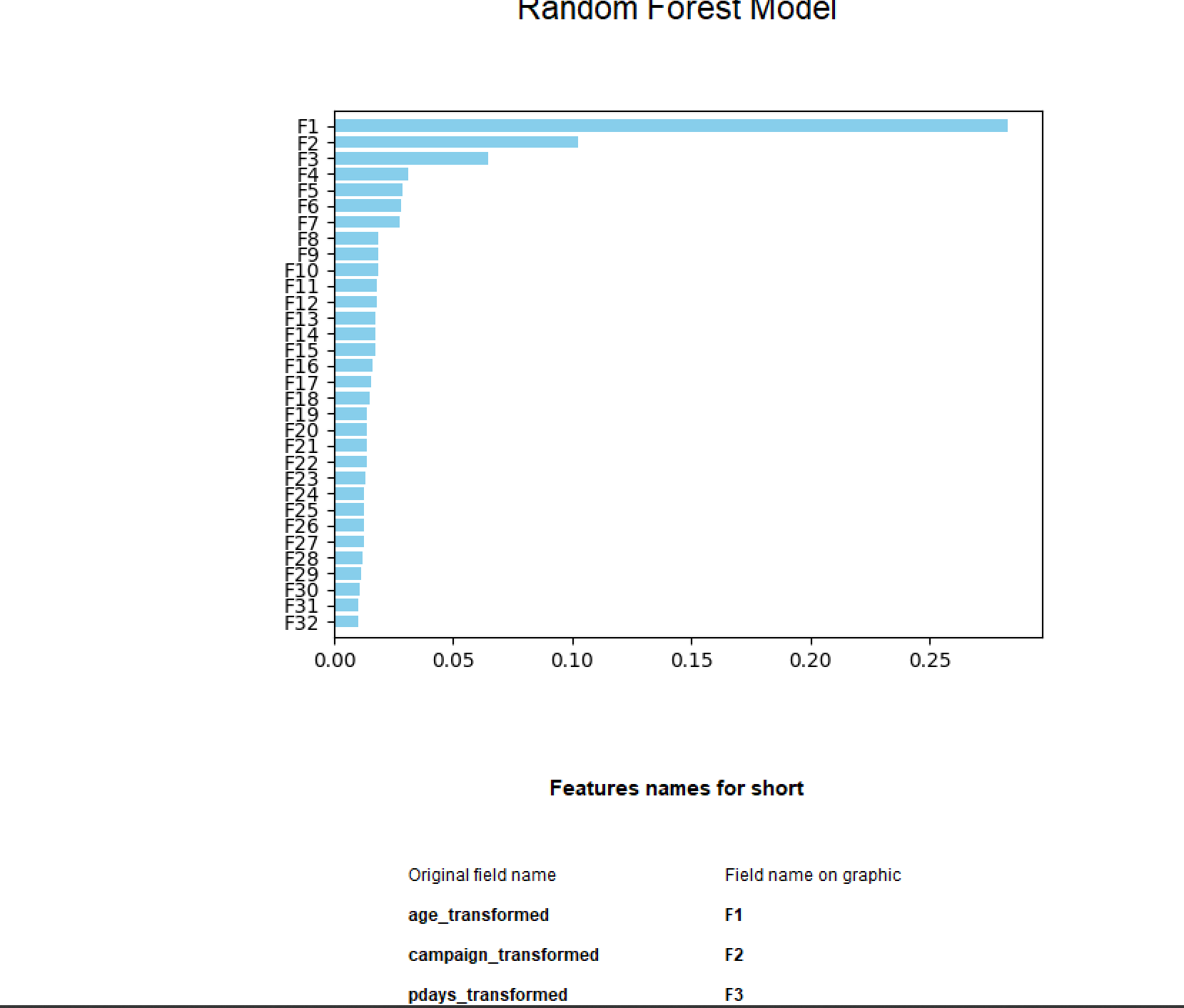
Decision Tree:C5.0 and C&R Tree

lassification and regression technique:Support Vector Machine (SVM),KNN

other mothods: XGBoost Tree,

### 5.1 Rondam Forest

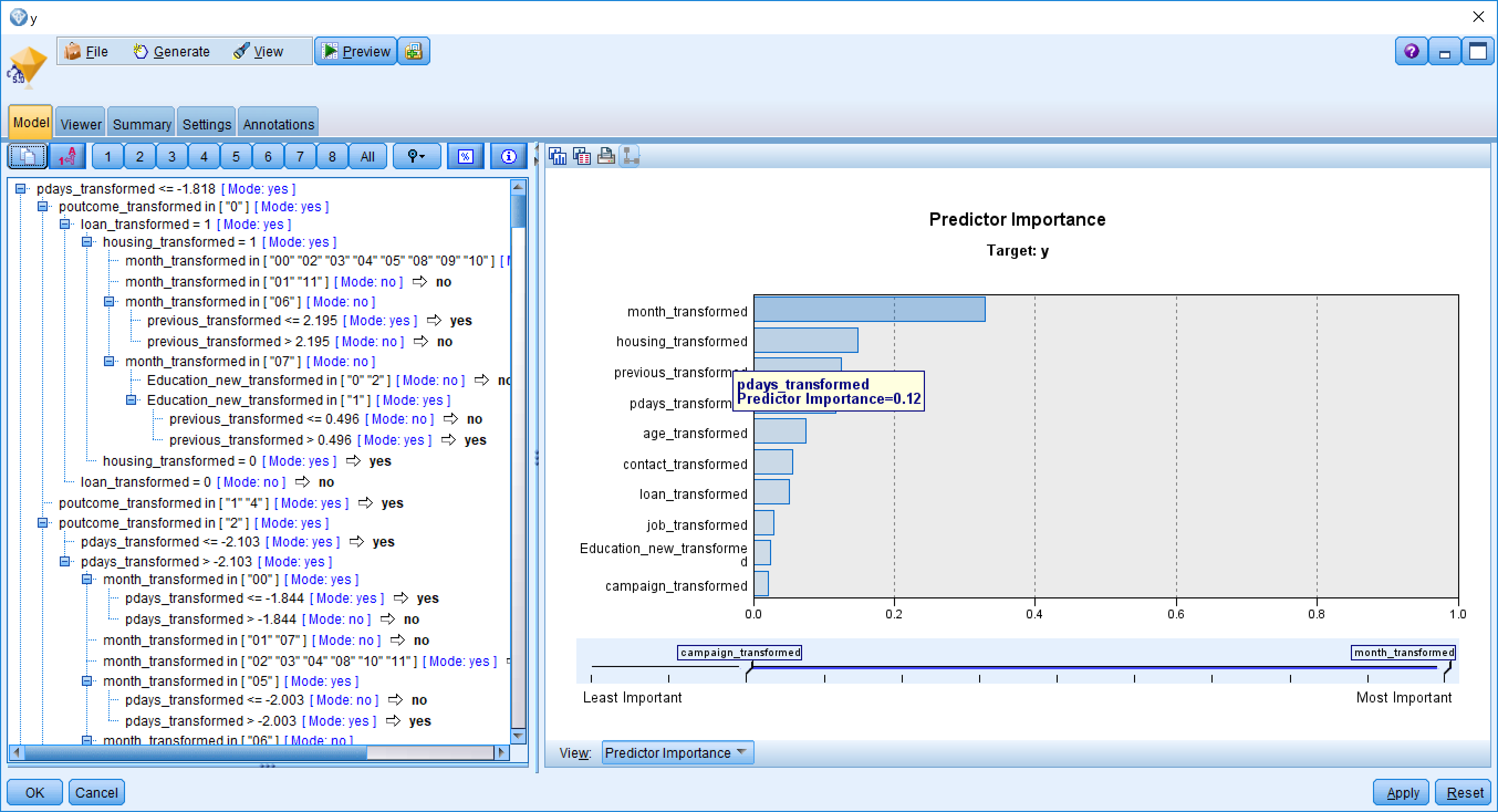
From Figure 12, Age is the most important field for Random Forest Model, Campaign and Pdays are the second and third most important fields.



(Figure 12)

### 5.2 C5.0

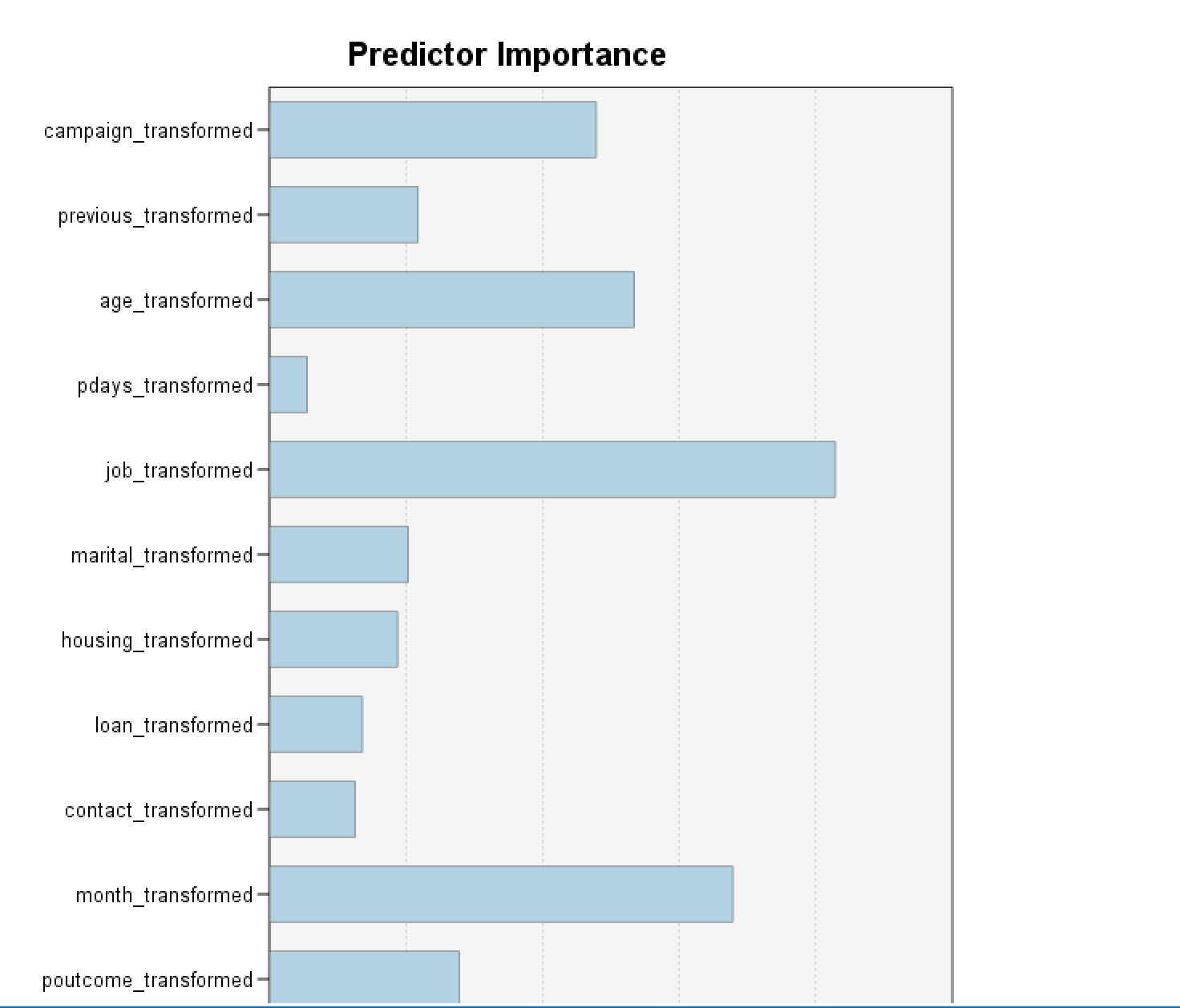
From Figure 13, Month, housing, previous are the most important three factors for C5.0.



(Figure 13)

### 5.3 Random Trees

From Figure 14, in Random Trees, the most important factor is Job.



(Figure 14)

For the whole result, scores list in the Table 1- Eight modelling result comparing above:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Order | Modeller Name | Total Training | Correct | Wrong | The Rate of Correct |
| 1 | KNN | 14293 | 10295 | 3998 | 72.03% |
| 2 | Random Tree | 14372 | 10235 | 4137 | 71.21% |
| 3 | C5.0 | 14235 | 10122 | 113 | 71.11% |
| 4 | Bayes Net | 14154 | 10003 | 4151 | 70.67% |
| 5 | Random Forest | 14418 | 10187 | 4231 | 70.65% |
| 6 | SVM | 14208 | 10006 | 4202 | 70.43% |
| 7 | CRT | 14238 | 9747 | 4491 | 68.46% |
| 8 | Neural Net | 14302 | 7516 | 6786 | 52.55% |

(Table 1-Eight modelling result comparing)

Select the appropriate data-mining method(s) based on discussion

Basing on Table 1- Eight modelling result comparing, obviously Neural Net got a very low result, it looks it totally did not match this task, so we decided to choose other seven methods including KNN,vand Random Tree,C5.0,Bayes Net, Random Forest and SVM , CRT to do the next step research.

## 6.Data-mining algorithm(s) selection

“Sixth is exploratory analysis and model and hypothesis selection: choosing the datamining algorithm(s) and selecting method(s) to be used for searching for data patterns. This process includes deciding which models and parameters might be appropriate (for example, models of categorical data are different than models of vectors over the reals) and matching a particular data-mining method with the overall criteria of the KDD process (for example, the end user might be more interested in understanding the model than its predictive capabilities).” (Fayyad et al., 1996)

For this task, as we know, there are only 10% of customers will be attracted

by marketing activities. So comparing to the correct rate of the whole data, the correct rate of customers who labelled success is more important. So we decide to look the different between these parts each modeling got.

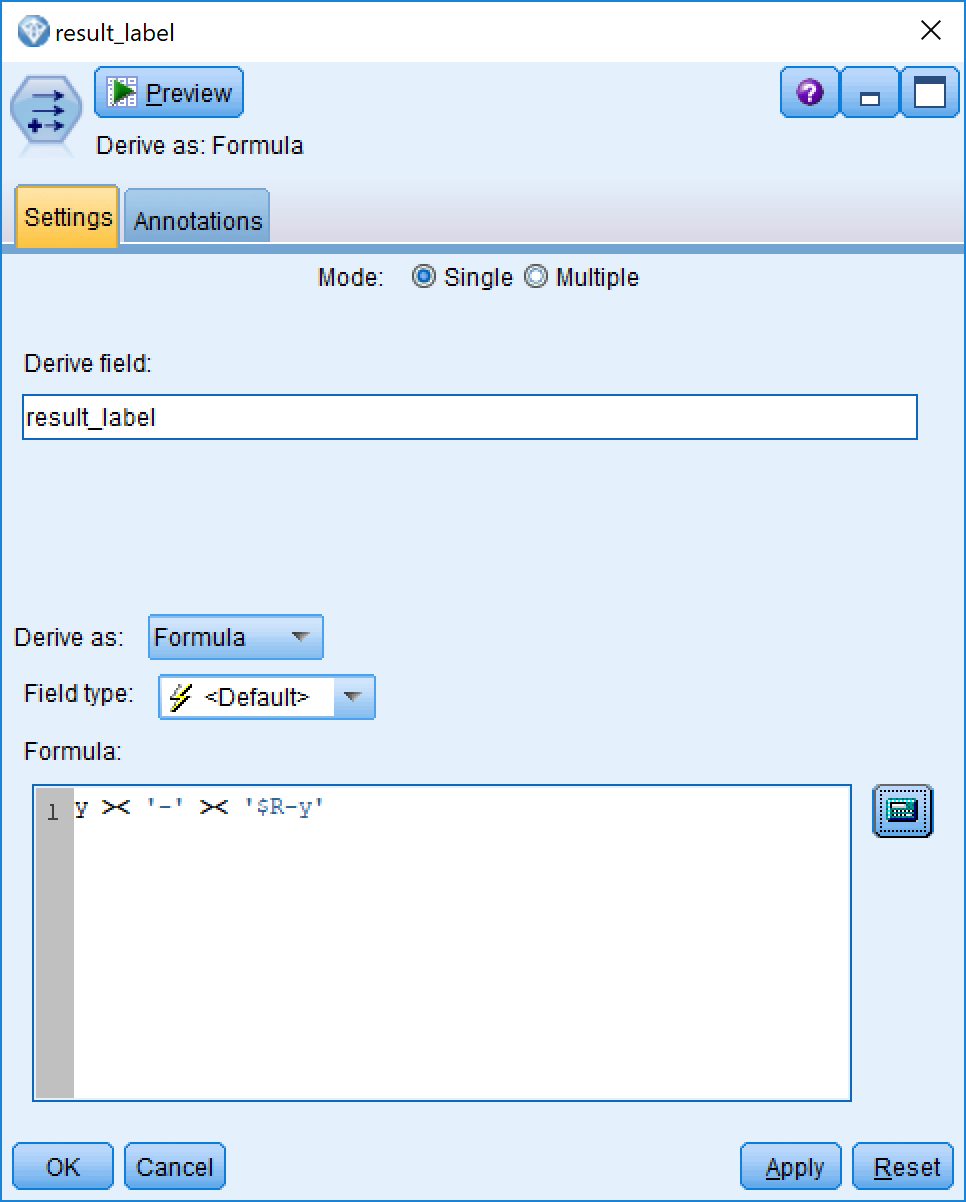
We create a new field result\_label in Figure 15, combined from the original exception field “Y” and the calculating exception field, which has labelled “No-No”,”No-Yes”,”Yes-Yes”,”Yes-No”,

“No-NO” means original is “NO”, and the calculation result also is “NO”

“Yes-Yes” means original is “Yes”, and the calculation result also is “Yes”

“No-Yes” means original is “NO”, and the calculation result is “Yes”

“Yes-No” means original is “Yes”, and the calculation result is “No”



(Figure 15)

From Table-2 Yes-No comparing,KNN ,Random Forest and SVM got the best three score on “YES”, they only miss 14.38%,15.99% and 17.29% respectively. So we decided these three methods will be able to enter the next step.

|  |  |  |  |
| --- | --- | --- | --- |
| Order | Modeller Name | Number | The Rate of incorrect |
| 1 | KNN | 2541 | 14.38% |
| 2 | Random Forest | 2849 | 15.99% |
| 3 | SVM | 3085 | 17.29% |
| 4 | C5.0 | 3246 | 18.25% |
| 5 | Bayes Net | 3597 | 20.23% |
| 6 | Random Tree | 3958 | 22.36% |
| 7 | CRT | 4439 | 25.02% |

Table-2 Yes-No Comparing

From Table-3 No-Yes comparing, CRT, Random Forest and Bayes Net got the best three score on original “NO”, they only miss 5.56%,7.14% and 8.94% respectively. It means If we add these prediction data to the waiting phoning pool, it will not increase the number of failures. So we decided these three methods also will be able to enter the next step.

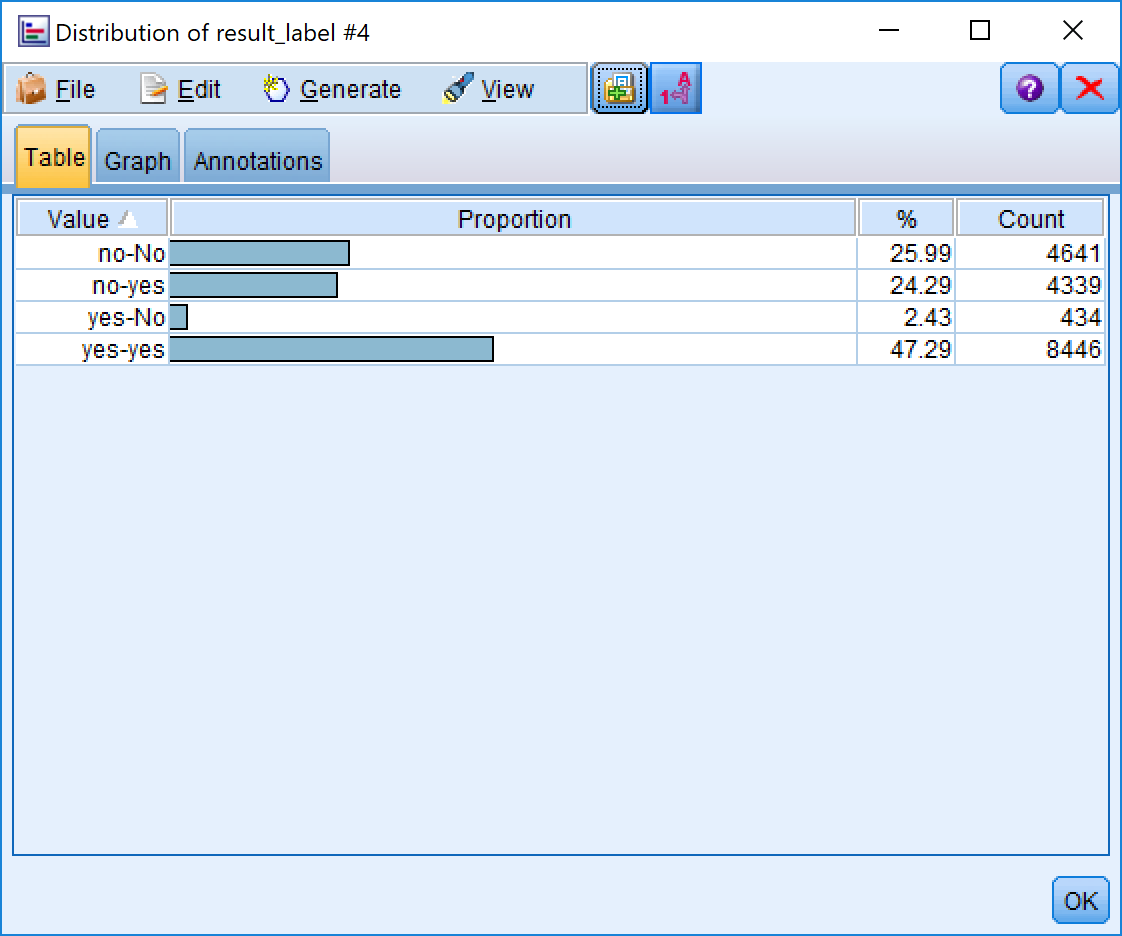
Table-3 No-Yes Comparing

|  |  |  |  |
| --- | --- | --- | --- |
| Order | Modeller Name | Number | The Rate of incorrect |
| 1 | CRT | 986 | 5.56% |
| 2 | Random Forest | 1263 | 7.14% |
| 3 | Bayes Net | 1590 | 8.94% |
| 4 | C5.0 | 1932 | 10.86% |
| 5 | SVM | 2205 | 12.36% |
| 6 | Random Tree | 2347 | 13.17 |
| 7 | KNN | 2437 | 13.69% |

So Totally, five Modeling are still in the pool we want to choose. They are KNN, Random Forest, SVM, CRT and Bayes Net. But there is also no any modelling which match the original target in the beginning of this report. Due to the accuracy on Yes is in 70%-80%,lower than our target, and the accuracy on No is also between 70-80%,higher than our target. So we could do something to improve the accuracy on Yes by decreasing the accuracy on No.

It is our final decision to combine the five models result, if any of these five models predict it is Yes, we predict it is “Yes”, in other words, if all of these models predict a customer will answer NO, we set it is No, otherwise it will be Yes.

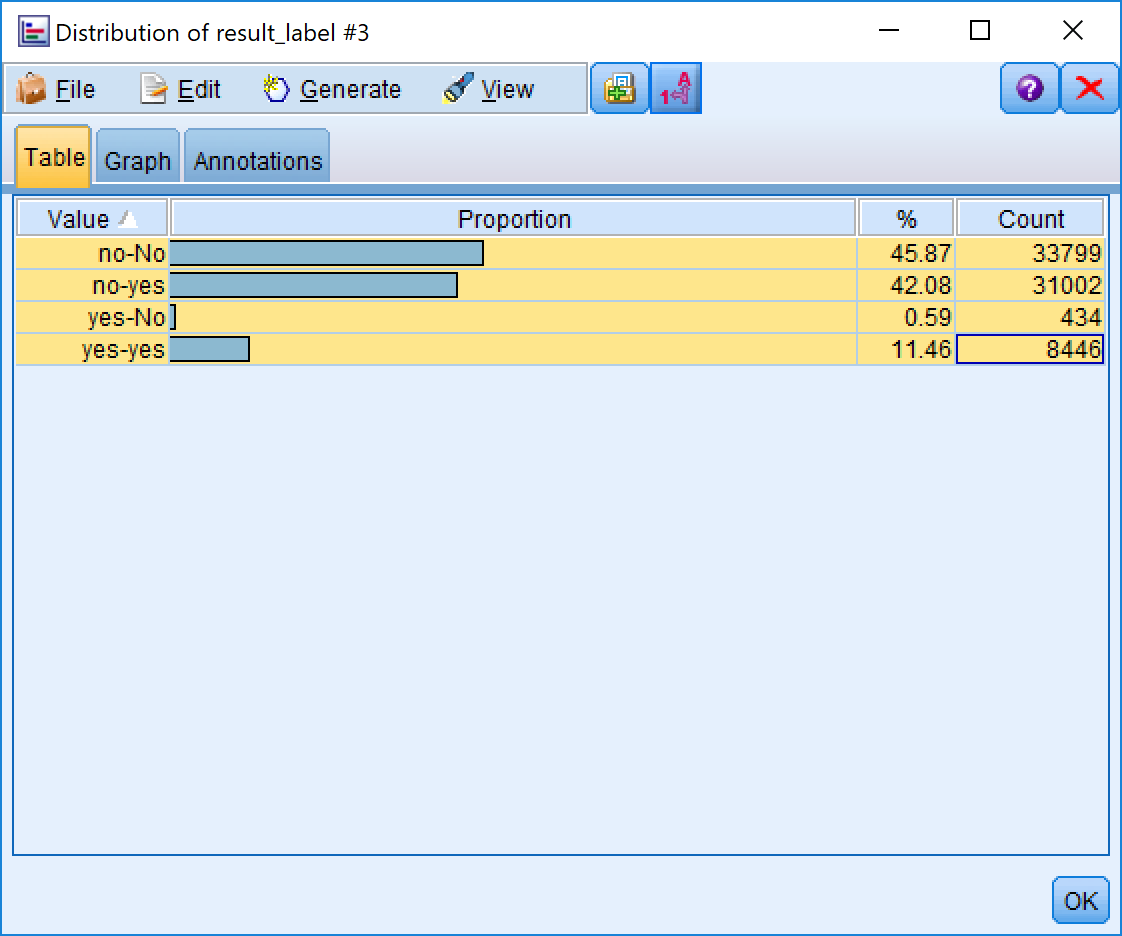
Let us have a look at the final data in Figure 16.



(Figure 16)

## 7.Data Mining

“Seventh is data mining: searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, and clustering. The user can significantly aid the data-mining method by correctly performing the preceding steps.” (Fayyad et al., 1996) This is, of course, the flashy part of data mining, where sophisticated analysis methods are used to extract information from the data.

As we can see from Figure 17, we get a very good result from putting the whole datasets into this data mining models. Data from 45% of the total is marked as no-no, indicating that the original data is No, and all five models are calculated as No. For this part of the customers, we do not need to make a phone call and then from the total of 0.59% of the data. Being marked as “YES-No” indicates that if we call out according to this forecast data, we will lose 0.59% of potential orders. In conclusion, we reduced our out-of-pocket workload by 46.46%, lost 0.59% accuracy, and increased the effective rate of exhalation from 12% to 26.32%, which is consistent with and greatly exceeds our default target, so there is no doubt this project was successful.

(Figure 17)

## 8.Interpretation

“Eighth is interpreting mined patterns, possibly returning to any of steps 1 through 7 for further iteration. This step can also involve visualization of the extracted patterns and models or visualization of the data given the extracted models.” (Fayyad et al., 1996) We assess and evaluate the models and the results and their reliability. “You are ready to evaluate how the data mining results can help you to achieve your objectives.” (SPSS, 2007)

From the whole project, there are four critical factors to predicate whether a customer would receipt marketing activity from phoning. They are Age, Job, Previous and Poutcome. We did not use the social and economic context attributes to prediction which included employment variation rate, consumer price index ,consumer confidence index, euribor 3 month rate and number of employees. It is not true we think they are not important. In conversely, we think they are very important factor for a bank marketing activity. It just because there are no these data in the dataset one (bank-full.cvs). In the future, we will consider how to use social and economic context attributes to extend our model.

## References

1. 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

2.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.

3. Langley, A., Mintzberg, H., Pitcher, P., Posada, E., & Saint-Macary, J.Opening up decision making: The view from the black stool. organization Science, 6(3), pp. 260-279, 1995.