“**Customer Behavior Analysis for Term Deposit Acceptance in Banking**”

MSBA 5314 - Predictive Analytics and ML

Achalshail Khadka

Babah Sesay

Pravallika Vasantham

Smit Patel

December 7th,2023

University of Central Oklahoma

Table Of Contents:

1. Abstract …………………………………………………………………………………….3
2. Introduction …………………………………………………………………………………4
3. Literature Review………………………………………………………………………….
4. Methodology ……………………………………………………………………………….

4.1 Data Understanding ………………………………………………………………….

4.2 Transformation ………………………………………………………………………..

4.3 Statistical Analysis …………………………………………………………………….

4.3.1 T-Test ……………………………………………………………………………..

4.3.2 Chi-Square ………………………………………………………………………..

4.3.3 ANOVA……………………………………………………………………………

4.3.4 Correlation Matrix ……………………………………………………………

4.4 Statistical Analysis and Predictive Analysis ………………………………………..

4.5 Cluster Analysis ………………………………………………………………………..

1. Selection of Champion Model…………………………………………………………….
2. Discussion and Conclusion ………………………………………………………………..
3. References……………………………………………………………………………………
4. Appendix ……………………………………………………………………………………..

**Abstract:**

This paper examines the Almighty Portuguese Bank, a key institution in Portugal's banking market, and it's falling acceptance of term deposits among potential clients. The issue, which is being exacerbated by changing customer financial habits and growing market competition, has a direct impact on the bank's profitability and strategic posture. The project's goal is to use data analytics to create prediction models and consumer segmentation techniques that will improve targeted marketing and customer acquisition.

Motivated by the bank's reputation for strong customer service and cutting-edge financial solutions, the project strives to solve the issue of dwindling term deposit subscriptions by identifying and addressing root reasons. The study of research underlines the role of decision tree prediction models and ensemble learning algorithms in influencing long-term deposit subscriptions.

The project's originality comes from taking into account the second major outcome—marketing campaigns—along with demographic characteristics to anticipate term deposit subscriptions. Existing solutions that rely on manual approaches or broad marketing initiatives have been found to be inefficient and costly.

The proposed method is developing a predictive model based on previous customer data, specifically from May 2008 to November 2010. The bank's goal is to reduce call center costs by targeting specific consumers who are likely to sign up for term deposits. With 45,211 observations and seventeen variables, the clean dataset gives useful insights into customer demographics and behavior.

Key statistics on variables such as job, marital status, education, housing, loan status, contact methods, month, and outcome categories are revealed in the analysis. The research intends to propose a prediction model appropriate to the current age, when emails and apps outnumber phone calls, by taking into account not only the top but also the second or third most essential characteristics.

Through predictive analytics, the project offers to improve the bank's marketing strategy by giving targeted promotions, better resource use, and higher conversion rates. The findings are intended to provide useful insights for optimizing communication channels, scheduling, and overall efficacy of marketing initiatives, hence increasing term deposit growth for the Almighty Portuguese Bank.

**Introduction** :

The Almighty Portuguese Bank, a key institution in Portugal's banking sector, is facing a critical crisis due to a decline in term deposit acceptance among potential clients. As a financial institution, the bank uses a variety of tactics to reach out to customers, including advertisements, email marketing, phone marketing, and digital marketing. While some bank accounts allow users to receive interest on their deposits, the bank makes money by lending money to individuals for automobiles, homes, and businesses at greater interest rates than those offered to customers.

The specific business issue at hand is the declining acceptability of term deposits, which is being driven by changing customer financial behavior, increased market rivalry, and a shift toward alternative investment possibilities. This drop in term deposit subscriptions has a direct influence on the bank's revenue and market position, emphasizing the importance of identifying and addressing the underlying causes.

Motivated by the need to revert this tendency, our research will use data analytics to create prediction models and customer segmentation techniques. These solutions will improve targeted marketing and customer acquisition by giving actionable information into communication channels and timing for marketing campaigns and customer engagement initiatives.

The problem is significant since it has a direct influence on the bank's profitability and strategic positioning. Increasing term deposit subscriptions efficiently is critical for keeping a strong financial portfolio and a competitive edge in the banking industry. The review of research emphasizes the importance of decision tree prediction models and ensemble learning algorithms, with critical criteria such as employee number, length, and month identified as influential elements influencing long-term deposit subscriptions.

Our project distinguishes itself from previous alternatives by taking into account the second most significant outcome—marketing campaigns—along with demographic characteristics. Many past studies have neglected the impact of marketing campaigns on term deposit subscriptions, therefore our technique tries to address a vacuum in the study. Predictive analytics, which includes modeling approaches like decision trees, logistic regression, neural networks, and ensemble methods, is becoming the main point for improving targeting precision, tailoring messaging, and optimizing resource allocation.

Existing methods, which rely on manual approaches or broad marketing campaigns, are deemed inefficient and expensive. Our proposed method entails developing a predictive model based on historical customer data gathered via marketing initiatives. The purpose is to accurately identify consumers who are likely to sign up for term deposits, contributing to the bank's overall growth.

Finally, the context of the organization and the addressed problem contextualizes the dataset collected from May 2008 to November 2010, a time when phone conversations and mailing letters were the primary ways of communication. The dataset, obtained from Kaggle.com, has been verified as clean and suitable for analysis. The research tries to develop a prediction model applicable to the modern period, where emails and apps outnumber traditional phone calls, by taking into account not only the top but also second or third most essential elements.

**Literature Review**:

Previous research has highlighted decision tree prediction models as useful tools for gaining a better understanding of the elements that influence long-term deposit subscriptions. Notably, the study discovered that the number of employees, length, and month were the most relevant factors to the prediction model. The most influential element was the number of employees, followed by duration and month, suggesting the importance of these criteria in shaping client decisions about term deposits.

Another pertinent study investigated the efficacy of decision tree-based ensemble learning algorithms such as random forest, gradient boosting machine, extreme gradient boosting, and light gradient boosting machine. The findings showed these models' remarkable performance, particularly with tabular data. Ensemble approaches captured complicated relationships within datasets successfully, demonstrating their potential for accurate predictions in the context of term deposit subscriptions.

The study distinguishes itself from prior studies by stressing the importance of marketing campaigns as a secondary critical outcome in addition to demographic characteristics. According to the research, predictive analytics plays a critical role in optimising banks marketing strategies. It detects high-potential term deposit candidates, allowing for focused promotions, more efficient resource utilisation, and higher conversion rates. Predictive analytics tools' user-friendly interface enables quick decision-making and adapts strategies to market developments, assuring regulatory compliance and helping to the campaign's efficacy in increasing term deposits.

A variety of modelling strategies for generating predictive models are suggested in the literature. As acceptable methodologies for evaluating the presented dataset, decision trees, logistic regression, neural networks, and ensemble methods are mentioned. These strategies provide a comprehensive toolkit for extracting useful insights from data, improving forecast accuracy, and informing marketing strategy decisions.

Existing banking solutions, as addressed in the literature, frequently rely on manual procedures or broad marketing campaigns. Manual approaches for prioritizing new clients have scalability and data use limits. Without individualized methods, generalized marketing campaigns targeting a broad range of clients result in inefficiency and expensive marketing expenses, emphasizing the need for more complex prediction models.

In conclusion, the literature study gives a thorough grasp of the important elements impacting long-term deposit subscriptions, the efficacy of ensemble learning algorithms, the significance of marketing efforts, and the utility of various modeling methodologies. The highlighted deficiencies in previous literature, especially the underemphasis on marketing campaign outcomes, contribute to the project's distinctive methodology.

**Methodology :**

**Data Understanding :**

The data for our project was collected from Kaggle which is available in this link

<https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets/data>. The dataset contains 45211 observations with seventeen variables. In the dataset 5289 observations exhibit a positive outcome (Fig: 2). Variable job has twelve categories, blue -collar being the highest frequency 21.53%, followed by management, technician (Fig.3). Variable “marital” has three categories namely divorced (11.52%), married (60.19%) and single(28.29%) (Fig.4). Variable “education” has four categories namely primary (15.15%), secondary (51.32%),tertiary(29.42%) (Fig:5). Variable housing has 55.58% yes and 44.42% no categories (Fig7.) . Variable Loan has 16.02% of yes category (Fig:8). Contact variable has three categories, namely cellular(64.77%),telephone(6.43%)(Fig.9). Variable month has twelve months as twelve categories where May has highest 30.45% of observations whereas December has lowest 0.47% of observations. (Fig.10). Variable outcome has four categories with failure category having 10.84% of observations and success having 3.34% of observations.

**Data Dictionary:**

| Variable | Description | Measurement Level | Role |
| --- | --- | --- | --- |
| Age | Age of the client | Numeric | Input |
| Job | Type of job | Categorical | Input |
| Marital | Marital Status | Categorical | Input |
| Education | Education level | Categorical | Input |
| Default | Has credit in default | Binary | Input |
| Balance | Average yearly balance, in euros | Numeric | Input |
| Housing | Has housing loan | Binary | Input |
| Loan | Has personal loan | Binary | Input |
| Contact | Contact communication type | Categorical | Input |
| Day | Last contact day of the month | Numeric | Input |
| Month | Last contact month of the year | Categorical | Input |
| Duration | Last contact duration, in seconds | Numeric | Input |
| Campaign | Number of contacts performed during this campaign and for this client (includes last contact) | Numeric | Input |
| pdays | The number of days that have passed since the client was last contacted from a prior campaign (-1 indicates that the client has not previously been reached). | Numeric | Input |
| Previous | Number of contacts made prior to this campaign and for this customer | Numeric | Input |
| poutcome | The result of the preceding marketing campaign | Categorical | Input |
| y | Is the client a term deposit subscriber? | Binary | Target |

**Statistical Analysis:**

| **Variable** | **Role** | **Mean** | **Standard Deviation** | **Non Missing** | **Missing** | **Minimum** | **Median** | **Maximum** | **Skewness** | **Kurtosis** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | INPUT | 40.93621 | 10.61876 | 45211 | 0 | 18 | 39 | 95 | 0.684818 | 0.31957 |
| balance | INPUT | 1362.272 | 3044.766 | 45211 | 0 | -8019 | 448 | 102127 | 8.360308 | 140.7515 |
| campaign | INPUT | 2.763841 | 3.098021 | 45211 | 0 | 1 | 2 | 63 | 4.89865 | 39.24965 |
| day | INPUT | 15.80642 | 8.322476 | 45211 | 0 | 1 | 16 | 31 | 0.093079 | -1.0599 |
| duration | INPUT | 258.1631 | 257.5278 | 45211 | 0 | 0 | 180 | 4918 | 3.144318 | 18.15392 |
| pdays | INPUT | 40.19783 | 100.1287 | 45211 | 0 | -1 | -1 | 871 | 2.615715 | 6.935195 |
| previous | INPUT | 0.580323 | 2.303441 | 45211 | 0 | 0 | 0 | 275 | 41.84645 | 4506.861 |

**Fig: 1 Summary statistics for numerical variables**

The statistical summary provides insights into several key variables. Starting with 'age', the average client is approximately 41 years old, and the distribution is moderately right-skewed, indicating that a larger number of clients are below the mean age. The 'balance' shows significant variability with a high standard deviation, suggesting a wide range in clients' average yearly balances. Its skewness and kurtosis indicate that the majority have lower balances, with a few clients holding much higher amounts.

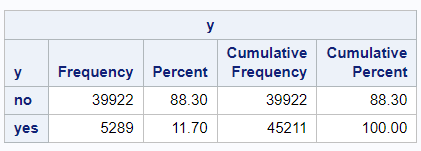
The 'day' variable, with a near-zero skewness, suggests a fairly uniform distribution across the month for last contact day. The 'duration' of the last contact averages around 4.3 minutes, with a wide range due to a few lengthy interactions, as indicated by its skewness and high maximum value.

'Campaign' shows that, on average, a client is contacted around 2.76 times during a campaign, but some clients have been contacted up to 63 times, which is unusually high and could indicate outliers or a particularly aggressive marketing strategy.

The 'pdays' and 'previous' variables, representing the number of days that have passed since the client was last contacted from a previous campaign and the number of contacts performed before this campaign, show high right skewness and kurtosis, which means most clients have not been contacted previously or it has been a long time since the last contact, with a small number having more frequent contacts.

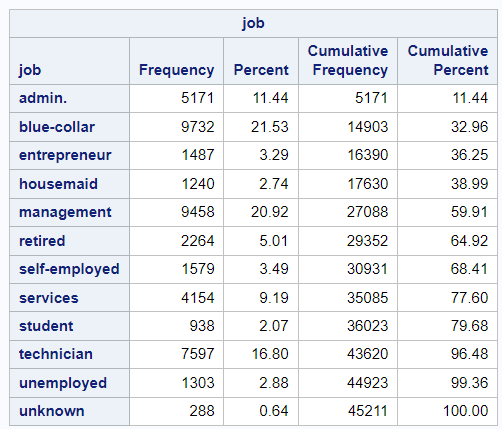
This exploratory analysis suggests that the bank's clients have a broad range of characteristics. For predictive modelling, understanding the impact of these variables on term deposit acceptance is key. Variables with high skewness and kurtosis will need to be treated carefully, as the outliers could influence the model's performance. The 'balance' variable, in particular, might require transformation or segmentation into categories for a more nuanced analysis.

**Frequency table for categorical variables**



**Fig. 2**

"No" and "Yes" indicate the two category variables under examination. The count or number of occurrences for each category is indicated. For example, the frequency for "No" is 39,922, while the frequency for "Yes" is 5,289. Represents the frequency of each category as a proportion of the total number of observations. In this situation, "No" has an 88.30% percentage, whereas "Yes" has an 11.20% percentage. As you progress down the table, this represents the running total of percentages. For "No," the cumulative percentage is the same as the percentage (88.30%), and for "Yes," it is the sum of the percentages for both categories (88.30% + 11.20% = 100.00%). These columns appear to give an alternative manner of representing category names in order to improve alignment. "Y - No" represents the "No" category, whereas "Y - Yes" represents the "Yes" category. This frequency table provides an in-depth look at the distribution of a categorical variable with two levels. It displays the number of observations in each category, the percentage of the total that each category represents, and the cumulative counts and percentages as you travel through the categories. The variable being summarised appears to be binary in your specific case, with categories "No" and "Yes," and the table provides a full breakdown of the distribution and cumulative distribution of these categories in your dataset.



**Fig:3**

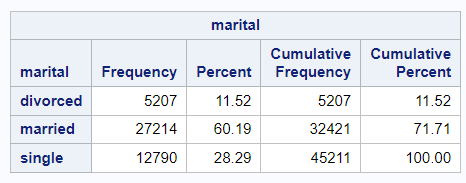
The 'Job' category is a crucial indicator in predicting customer behavior regarding term deposit acceptance. The frequency distribution of the job categories shows significant variance, which suggests that employment type may have an influence on a customer's decision to accept a term deposit offer.

Prevalence of Job Types: The most common job type among the clients is 'blue-collar', constituting approximately 21.53% of the dataset, followed closely by 'management' at 20.92%. These two categories together comprise over 42% of the client base. This heavy representation could indicate these groups are particularly targeted or more receptive to term deposit offers.

Minor Representations: On the other end, 'unknown', 'student', and 'housemaid' categories represent a smaller portion of the client base, totaling around 6.35%. The 'unknown' category may need further investigation to classify these entries correctly, as they could represent a potential source of bias or misclassification in predictive modeling.

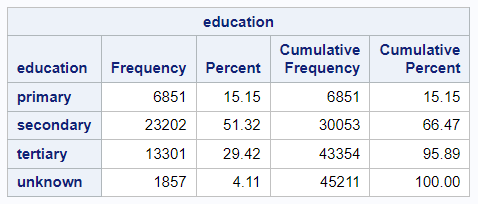
Potential Target Segments: 'Retired' individuals represent a significant 5.01% of the clientele. Given their potentially fixed income situation, they might be more inclined to invest in term deposits as a safe investment vehicle, which could make them a key demographic for targeted marketing campaigns.

Cumulative Insights: The cumulative percent column indicates a steady accumulation of frequencies across job types, which is expected in a variable with many categories. However, it is important to note that the first five job categories already make up nearly 40% of the data, which might suggest a Pareto effect where a relatively small number of job categories account for a large portion of term deposit acceptances.



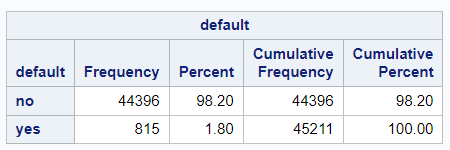
**Fig:4**

The dataset reveals a predominance of married clients at 60.19%, with singles and divorced clients representing 28.29% and 11.52%, respectively. This skew towards married individuals suggests that marital status is a key demographic factor that could influence the likelihood of accepting term deposits. As a married cohort, these clients might possess a more conservative risk profile, potentially making term deposits a more appealing investment option aligned with their financial planning and stability goals.



**Fig:5**

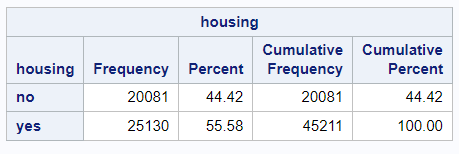
The education variable in our dataset shows a dominant percentage of clients with secondary education at 51.32%, followed by those with tertiary education at 29.42%, and primary education at 15.15%. A small fraction, 4.11%, is categorized as unknown. This distribution indicates that over half of the clients have at least completed secondary education, which may correlate with a certain level of financial literacy and openness to banking products like term deposits.



**Fig:6**

The 'default' variable reflects credit history, with a vast majority of my clients, 98.20%, having no default history, and a small minority, 1.80%, having defaulted. This skewness towards a 'no default' history is indicative of a client base with a generally good credit standing, which is typically associated with more favorable banking and investment behaviors.

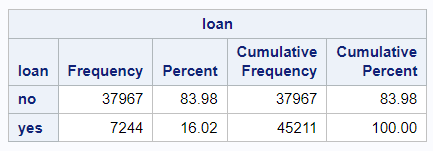
In our analysis, the 'default' attribute is critical as it may significantly influence a client's eligibility and propensity to accept term deposit offers. Given the high percentage of clients without a default history, we'd expect a relatively lower risk profile for the client base, which could correlate with a higher likelihood of term deposit subscription.



**Fig:7**

Reviewing the 'housing' variable shows that a majority of the bank’s clients, 55.58%, have a housing loan, while the remaining 44.42% do not. This information is essential as it gives insight into the financial commitments of the clients, which can affect their decision to invest in term deposits.

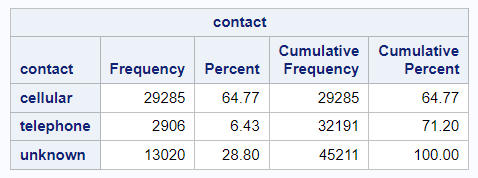
Having a housing loan could indicate a reduced disposable income, which may impact the ability or willingness to commit to a term deposit, typically considered a less liquid investment. Conversely, those without such a loan might have more financial flexibility to invest in term deposits.



**Fig:8**

The distribution of the 'loan' variable indicates that the vast majority of the bank’s client base, 83.98%, does not have a personal loan, while a smaller fraction, 16.02%, does. This suggests that most clients are not encumbered by the financial strain that personal loans can often represent, potentially leaving them with greater liquidity that could be directed towards term deposits.

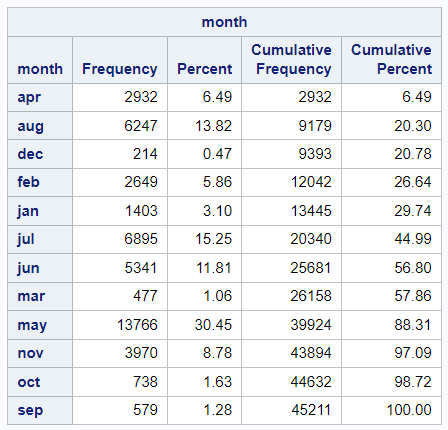
In analysing this data, we'd particularly focus on the impact of personal loans on the propensity to subscribe to term deposits. Clients without personal loans may have more financial freedom to invest in savings products. This aspect of their financial profile is a significant indicator for potential term deposit marketing strategies.



**Fig:9**

The 'contact' variable shows that most of the bank’s clients, 64.77%, were contacted via cellular phone, with only 6.43% contacted through a telephone and a substantial 28.80% marked as unknown. This suggests that mobile communication is the primary channel used for client outreach, which aligns with modern communication trends and likely reflects higher accessibility and convenience for clients.

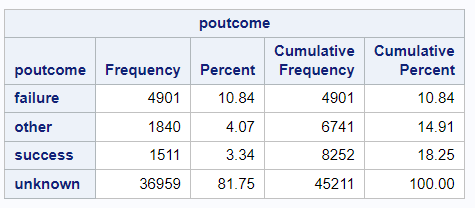
The significant unknown category could potentially indicate missing data or a lack of information on how these clients were contacted. This gap presents a challenge for analysis as it could mask important patterns in communication effectiveness.



**Fig:10**

The 'month' variable in the dataset indicates the timing of client contact, with May and July showing the highest frequencies of contact at 30.45% and 15.25%, respectively. This could suggest a seasonal strategy in the bank's marketing campaigns, possibly aiming to coincide with periods where clients might be more likely to invest, such as after receiving annual bonuses or before the summer vacation period.

December has the lowest frequency, which could reflect the holiday season when clients are less likely to focus on financial decisions. Understanding this seasonality is important for scheduling marketing efforts and could be crucial for optimizing the timing of campaigns to increase term deposit subscription rates.

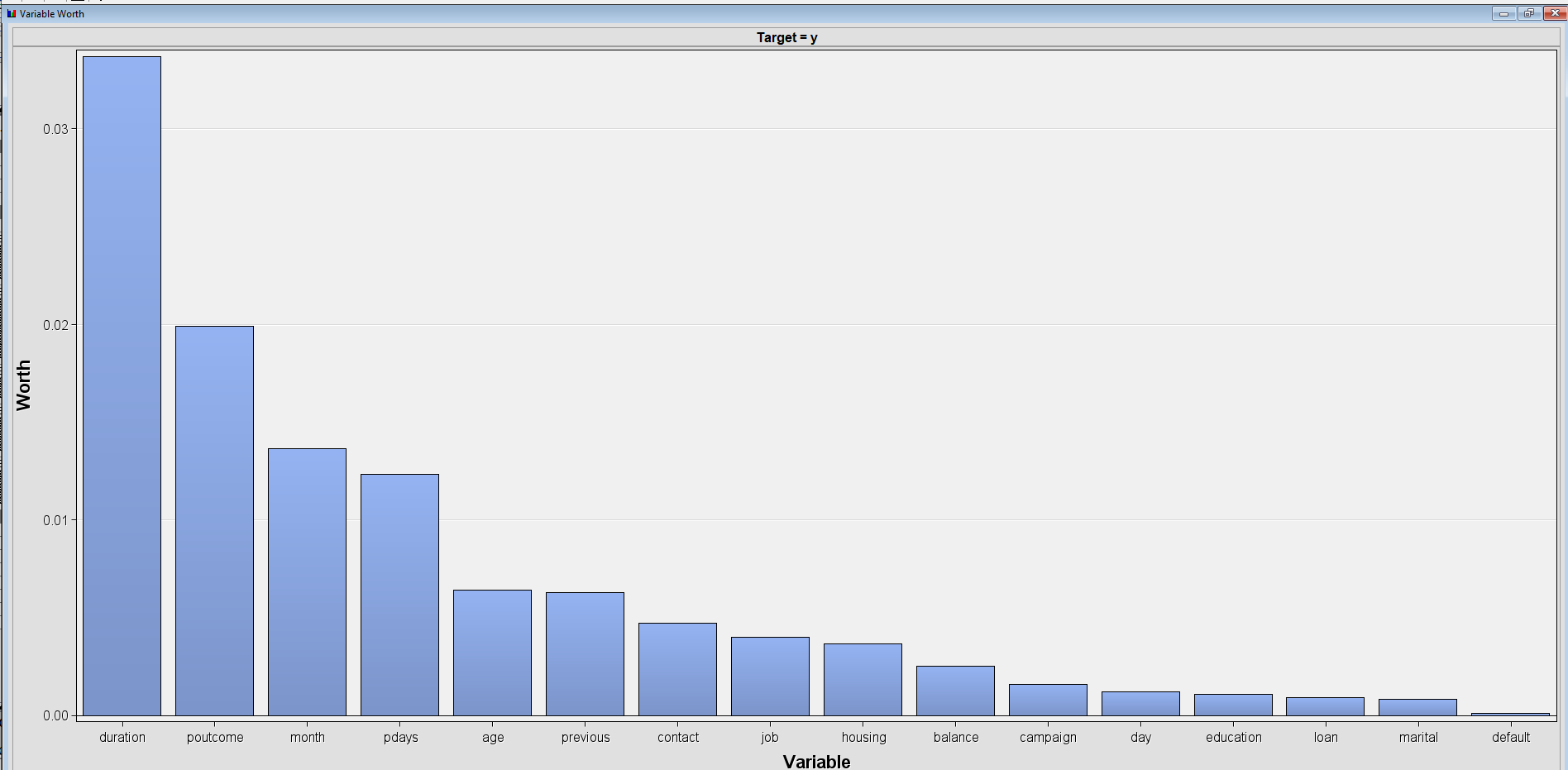


**Fig:11**

"failure," "other," "success," and "unknown" represent the different categories or levels of the categorical variable under consideration. The count or number of occurrences for each category is indicated. The frequency of "failure," for example, is 4,901. For "other," the frequency is 1,840. The frequency of "success" is 1,511. For "unknown," the frequency is 36,959. Represents the frequency of each category as a proportion of the total number of observations. In this situation, the percentage for "failure" is 10.84%. In the case of "other," the percentage is 4.07%.In the case of "success," the percentage is 3.34%. In the case of "unknown," the percentage is 81.75%.As you travel down the table, the running total of frequencies is displayed. As an example:The cumulative frequency of "failure" equals the frequency (4,901).The total of the frequencies for "failure" and "other" (4,901 + 1,840 = 6,741) is the cumulative frequency for "other." The sum of the frequencies for "failure," "other," and "success" (6,741 + 1,511 = 8,252) is the cumulative frequency for "success."The total cumulative frequency for "unknown" is 45,211 (8,252 + 36,959 = 45,211).As you progress down the table, this represents the running total of percentages. For example, the cumulative percent for "failure" is the percent (10.84%).The sum of the percentages for "failure" and "other" (10.84% + 4.07% = 14.91%) is the cumulative percent for "other." The total of the percentages for "failure," "other," and "success" (14.91% + 3.34% = 18.25%) is the cumulative proportion for "success." The cumulative percentage for "unknown" is 100.00%, which is the overall proportion of the dataset.

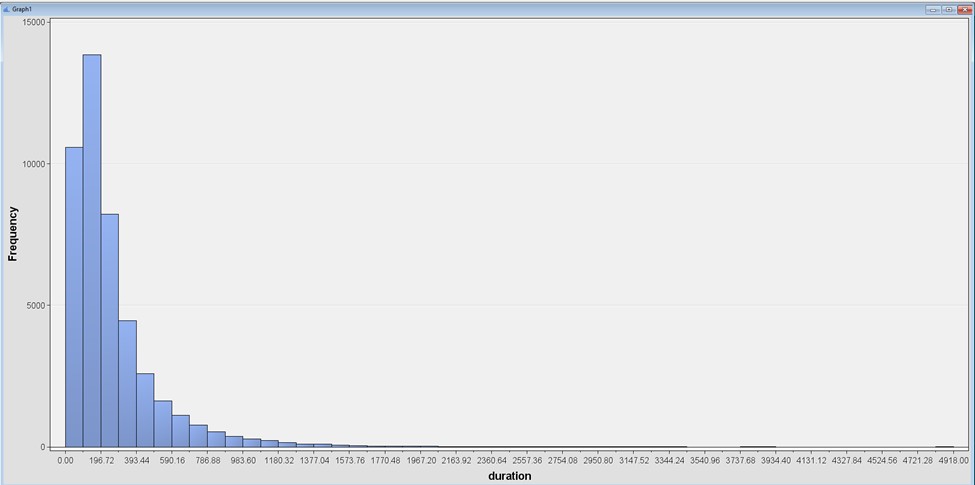
The frequency table below provides a detailed analysis of the distribution and cumulative distribution of the variable "poutcome." It shows you how many observations fall into each category, how much of a percentage each category represents, and the cumulative counts and percentages as you travel through the categories. In this dataset, the "unknown" category appears to be the most common.

After running the statexplore in the SAS enterprise miner, Top five variables based on the variable worth are duration, poutcome, month, pdays and age. (Fig.12)

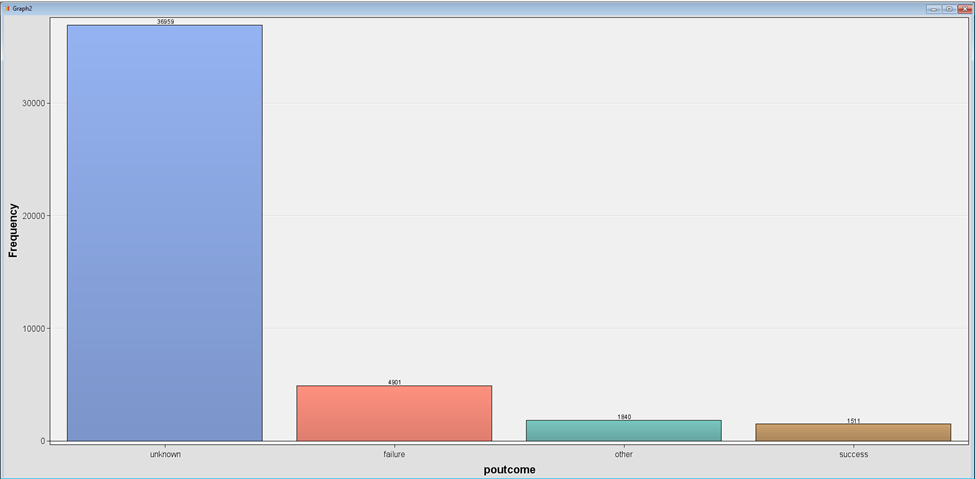


**Fig.12** Variable Worth

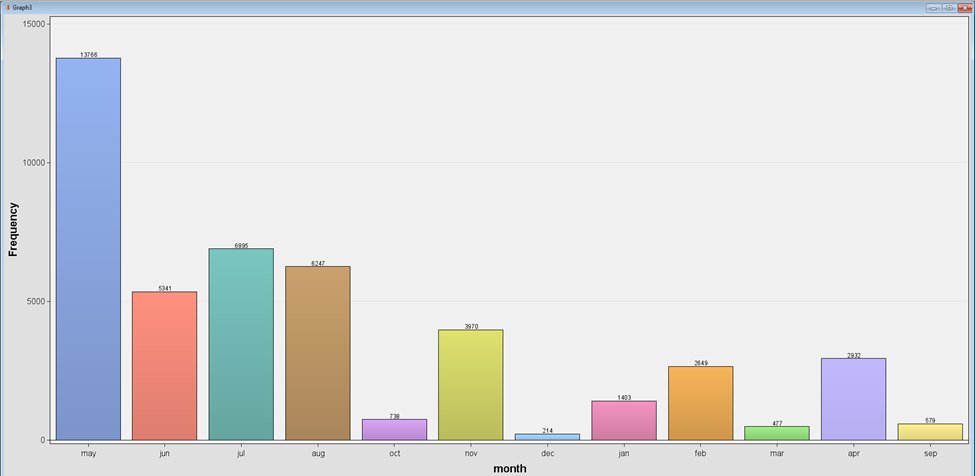
**Plot distribution of key variables (Top 5)**



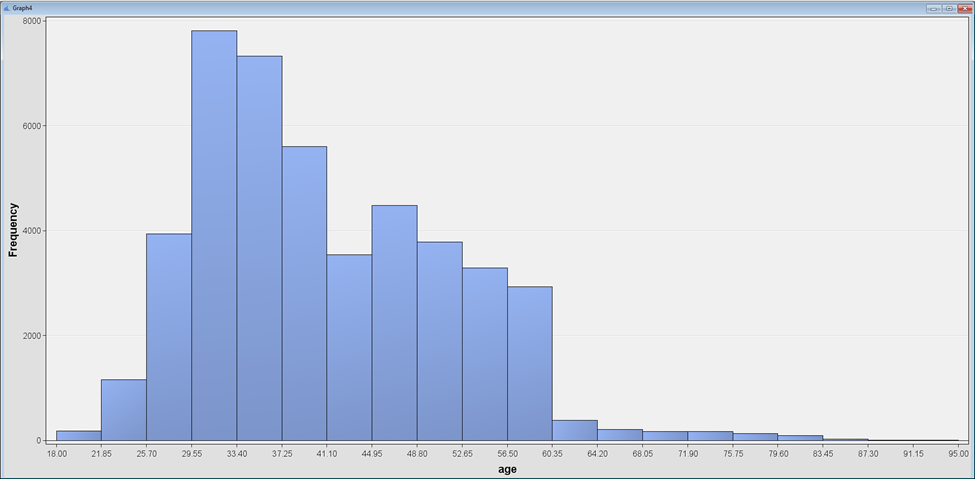
**Fig: 13** Histogram of variable Duration



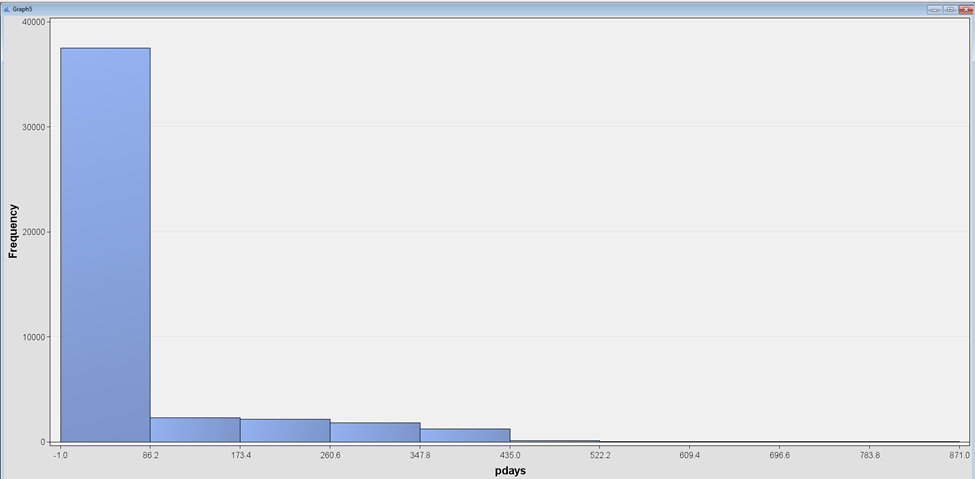
**Fig:14** Bar chart of variable poutcome



**Fig:15** Bar chart of variable month

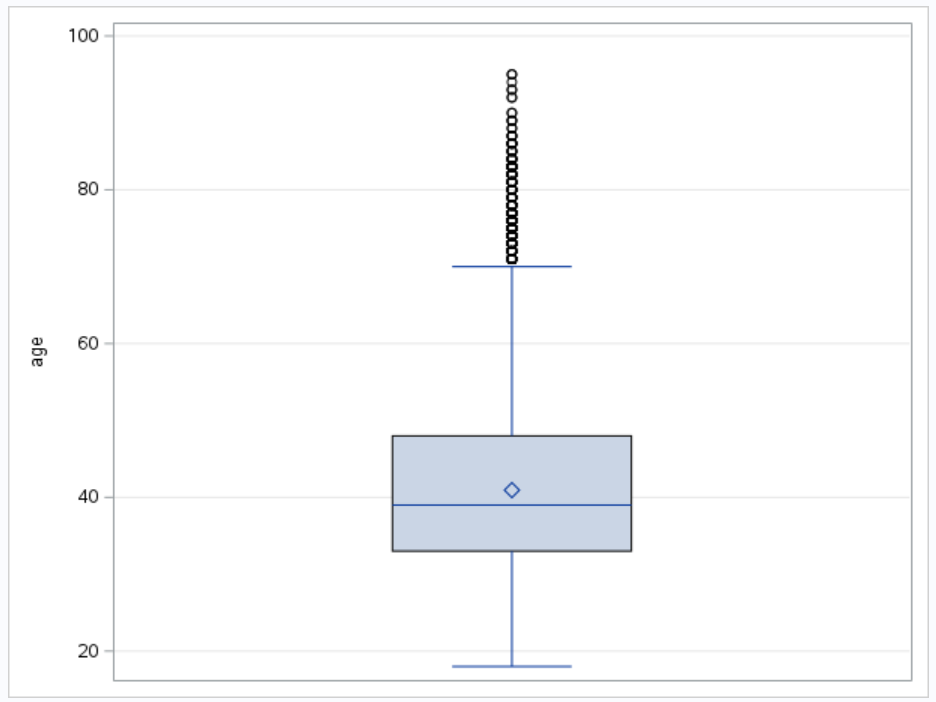


**Fig:16** Histogram of variable age

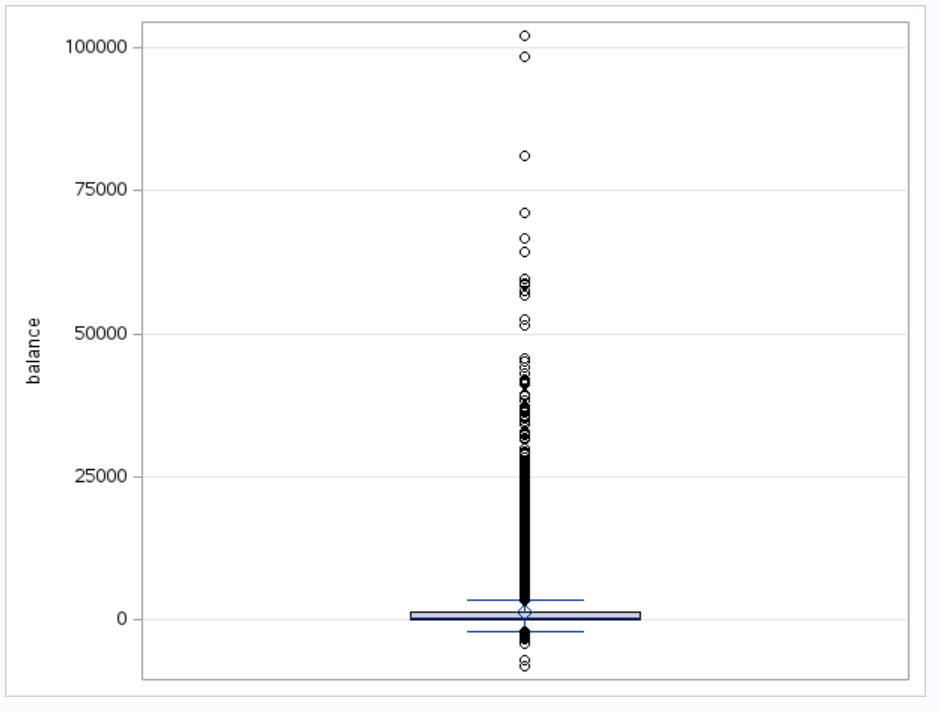


**Fig:17** Histogram of variable pdays

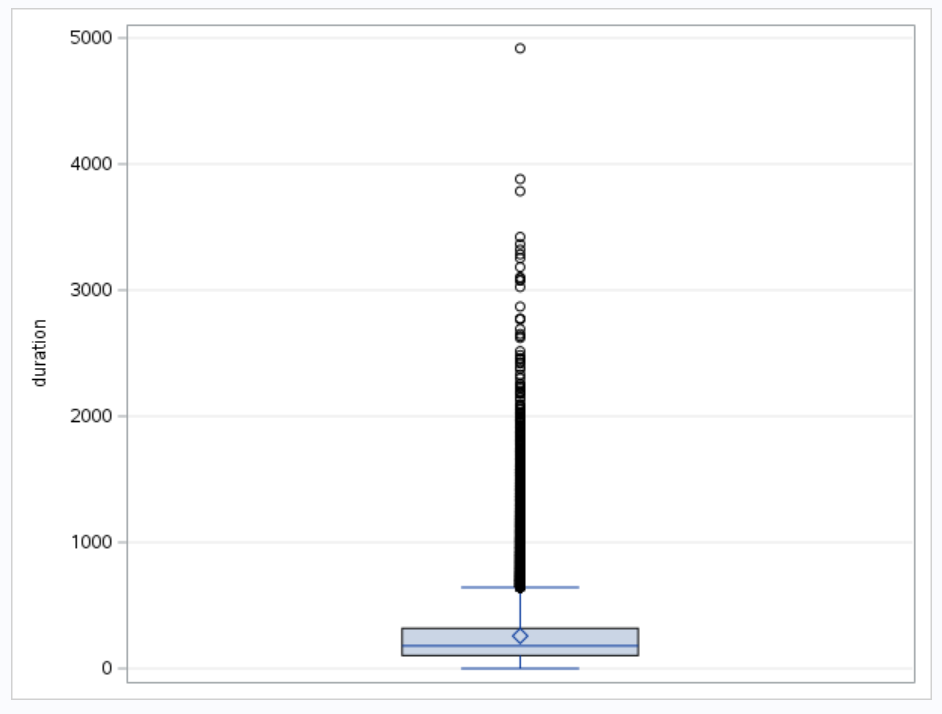
In the box plot of age there are outliers as it is clearly seen they are above the maximum range (approximately above 70 years) (Fig:18). It means most of the customers are between 35 to 45 years of age. The average yearly balance in euros is close to zero but positive because there are many customers who have negative bank balances. There are also outliers of values about 4000 and above(Fig:19). Most of the values for duration are between 0 and 750 seconds, there are few with values greater than 750 which are outliers (Fig:20). The average of the total number of times contacted the customer is about 2-3 from the box plot of the campaign. The outliers are of people who were contacted more than 5 times (Fig:21). In SAS Enterprise Miner we will manipulate outliers using transformation methods such as Winsorization or log transformation. We can also use powerful statistical methods or clustering to identify and remove outliers. Threshold adjustment can improve model performance and prevent overfitting of results.



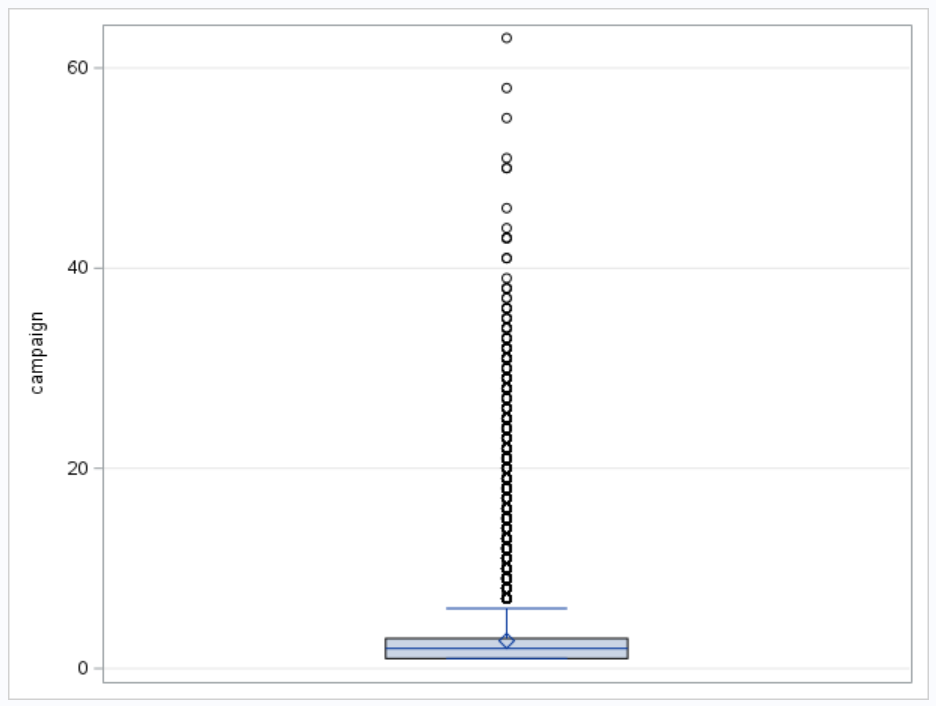
**Fig:18** Box plot of variable age



**Fig:19** Box plot of variable balance



**Fig:20** Box plot of variable duration



**Fig:21** Box plot of variable Campaign

**T-Test** :

Ho: The Duration does not play an important role in convincing customers for the term deposit

H1: The Duration plays an important role in convincing customers for the term deposit.

Assuming the duration has a normal distribution, From Fig:24 the p-value of the pooled is less than 0.05 meaning we reject the null hypothesis and can conclude that there is a significant relationship between Duration and convincing to term deposit.

| Variable : duration (duration)  y=no | | | | | |
| --- | --- | --- | --- | --- | --- |
| Test for Normality | | | | | |
| Test | | Statistic | | p-Value | |
| Kolmogorov-Smirmov | | D | 0.150476 | Pr>D | <0.0100 |
| Cramer-von Mises | | W-sq | 393.8067 | Pr>W-Sq | <0.0050 |
| Anderson Darling | | A-sq | 2233.782 | Pr>A-sq | <0.0050 |

| Variable : duration (duration)  y=yes | | | | | |
| --- | --- | --- | --- | --- | --- |
| Test for Normality | | | | | |
| Test | | Statistic | | p-Value | |
| Kolmogorov-Smirmov | | D | 0.115686 | Pr>D | <0.0100 |
| Cramer-von Mises | | W-sq | 28.17065 | Pr>W-Sq | <0.0050 |
| Anderson Darling | | A-sq | 169.6593 | Pr>A-sq | <0.0050 |

**Fig: 22**

Variable : duration (duration)

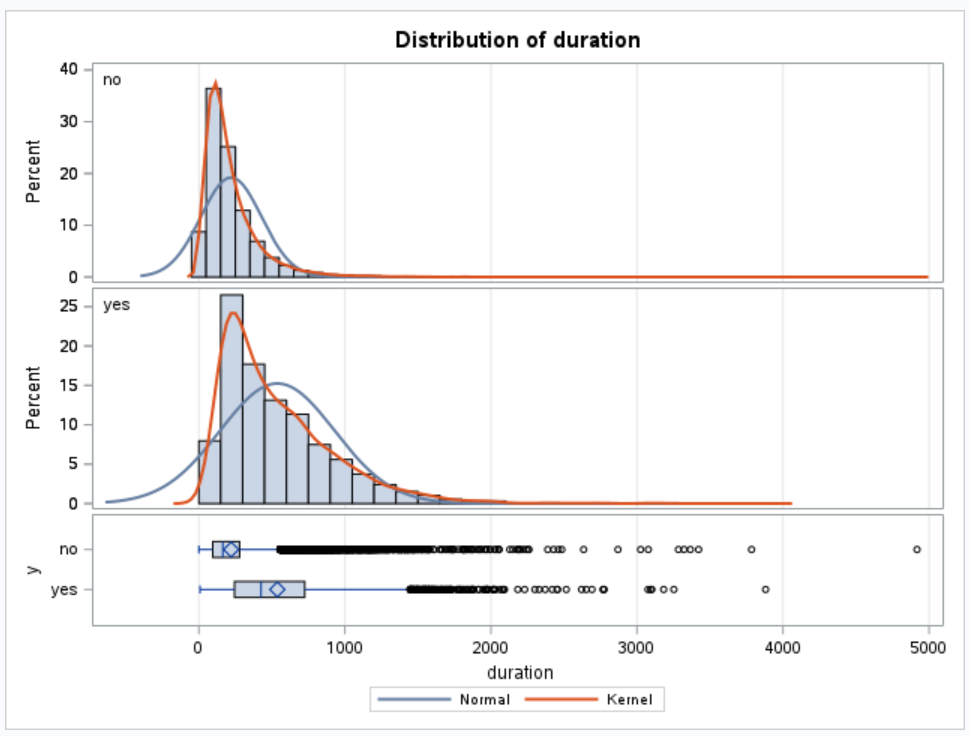
| y | method | N | Mean | Std Dev | Std Err | Minimum | Maximum |
| --- | --- | --- | --- | --- | --- | --- | --- |
| no |  | 39922 | 221.2 | 207.4 | 1.0379 | 0 | 4918.0 |
| yes |  | 5289 | 537.3 | 392.5 | 5.3974 | 8.0000 | 3881.0 |
| Diff(1-2) | Pooled |  | -316.1 | 236.6 | 3.4627 |  |  |
| Diff(1-2) | Satterthwalte |  | -316.1 |  | 5.4962 |  |  |

| y | method | Mean | Std Dev | Std Err | 95% CL | Std Dev |
| --- | --- | --- | --- | --- | --- | --- |
| no |  | 221.2 | 207.4 | 1.0379 | 206.0 | 208.8 |
| yes |  | 537.3 | 392.5 | 5.3974 | 385.2 | 400.2 |
| Diff(1-2) | Pooled | -316.1 | 236.6 | 3.4627 | 235.1 | 238.2 |
| Diff(1-2) | Satterthwalte | -316.1 |  | 5.4962 |  |  |

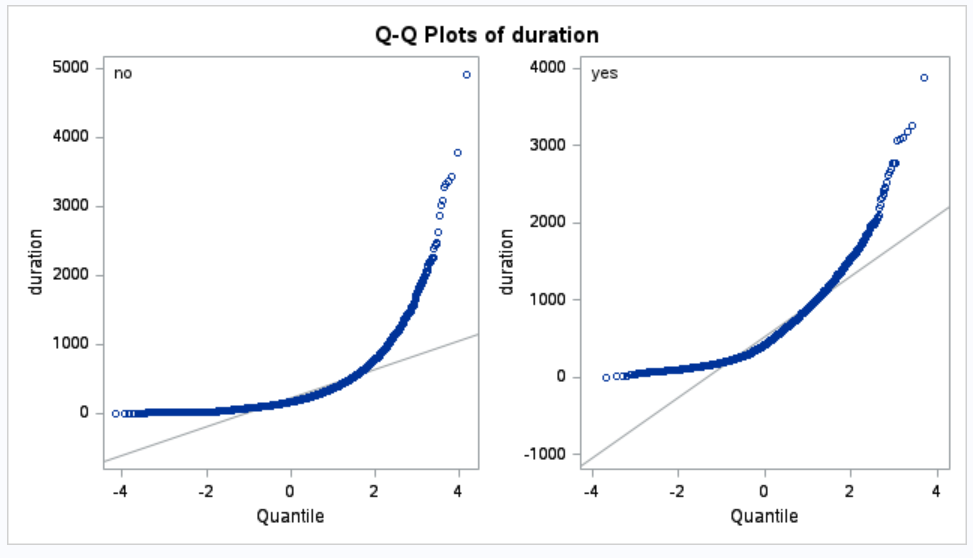
| Method | Variances | DF | t Value | Pr>[t] |
| --- | --- | --- | --- | --- |
| Pooled | Equal | 45209 | -91.29 | <0.0001 |
| Satterthwalte | Unequal | 5685.3 | -57.51 | <0.0001 |

| Equality of Variances | | | | |
| --- | --- | --- | --- | --- |
| Method | Num DF | Den DF | F value |  |
| Folded F | 5288 | 39921 | 3.58 | <0.0001 |

**Fig:23**



**Fig: 24**



**Fig: 25**

**Chi square test**

The measure of association of the variable or chi-square value is highest for the poutcome. The variable poutcome is the second important in the variance importance chart and a categorical variable.

| Chi-Square Statistics  (Maximum 500 Observations printed ) | | | |
| --- | --- | --- | --- |
|
| Data Role = TRAIN Target = y | | | |
| Input | Chi-Square | DL | Prob |
| poutcome | 4391.5066 | 3 | <0.0001 |
| month | 3061.8369 | 11 | <0.0001 |
| contact | 1035.7142 | 2 | <0.0001 |
| Housing | 875.6937 | 1 | <0.0001 |
| Job | 836.1055 | 11 | <0.0001 |
| education | 236.1949 | 3 | <0.0001 |
| loan | 210.1949 | 1 | <0.0001 |
| Marital | 196.4959 | 2 | <0.0001 |
| Default | 22.7235 | 1 | <0.0001 |

**Table : 1**

So far we have analysed the data thoroughly and found no missing, or any anomalies that could affect the predictive result because of the data imbalance. Since, the dataset was already analysed by few of the people, the data is clean and organised. Many people have used the duration as the main variable and analysed it but, for this project we are planning to use the second important variable which is poutcome, a categorical variable. The major challenge for us will be to find an alternative predictive model which is not done by other people in their paper. One of the models we are planning to do is comparing gini coefficients and see if there is any strong relation between dependent and independent variables.

**One way ANOVA :**

| Least Squares Means  Adjustment for Multiple Comparison : Tukey-Kramer | | |
| --- | --- | --- |
| poutcome | Duration LSMEAN | LSMEAN Number |
| failure | 244.185880 | 1 |
| Other | 255.715217 | 2 |
| success | 316.868961 | 3 |
| unknown | 257.738332 | 4 |

| Least Squares Means for effect outcome  Pr>[t] for H0:LSMean(i) = LSMean(j)  Dependent Variable: duration | | | | |
| --- | --- | --- | --- | --- |
| i/j | 1 | 2 | 3 | 4 |
| 1 |  | 0.3566 | <0.0001 | 0.0030 |
| 2 | 0.3566 |  | <0.0001 | 0.9877 |
| 3 | <0.0001 | <0.0001 |  | <0.0001 |
| 4 | 0.0030 | 0.9877 | <0.0001 |  |

**Table 2**

The ANOVA test is performed between two top variables, Duration and poutcome. From the table above, there is a significant difference between, Failure and Success, Failure and unknown, other and success and failure and unknown because the p-value is less than 0.05.

**Correlation matrix:**

| Pearson Correlation Coefficients , N = 45211 | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | age | Balance | day | duration | campaign | pdays | previous |
| Age  age | 1.00000 | 0.09778 | -0.00912 | -0.00465 | 0.00476 | -0.02376 | 0.00129 |
| Balance  balance | 0.09778 | 1.00000 | 0.00450 | 0.02156 | -0.01458 | 0.00344 | 0.01667 |
| Day  Day | -0.00912 | 0.00450 | 1.00000 | -0.03121 | 0.16249 | -0.09304 | -0.05171 |
| Duration  duration | -0.00465 | 0.02156 | -0.03021 | 1.00000 | -0.08457 | -0.00156 | 0.00120 |
| Campaign  campaign | 0.00476 | -0.01458 | 0.16249 | -0.08457 | 1.00000 | -0.08863 | -0.03286 |
| pdays  pdays | -0.02376 | 0.00344 | -0.09304 | -0.00156 | -0.08863 | 1.00000 | 0.45482 |
| Previous  previous | 0.00129 | 0.01667 | -0.05171 | 0.00120 | -0.03286 | 0.45482 | 1.00000 |

**Table : 3**

The correlation matrix provides insight into the relationships between numerical variables in the dataset. Notably, most variables have a weak correlation with each other, indicating that they contribute independently to the dataset's variability.

The correlation between 'age' and 'balance' is slightly positive (0.09778), suggesting that as clients get older, they may have slightly higher balances. However, the correlation is weak, which means 'age' is not a strong predictor of 'balance'.

The 'day' of the last contact does not show a significant correlation with other variables, which implies that the timing of the last contact in the month is not a strong predictor of other client characteristics.

'Duration' shows a very weak negative correlation with 'campaign' (-0.08457), indicating that as the number of contacts in a campaign increases, the duration of the calls might decrease slightly. This could suggest that repeated contacts do not necessarily equate to longer conversations, which may reflect efficiency in communication or client fatigue.

'Campaign' and 'pdays' are negatively correlated (-0.08863), hinting that more frequent contacts in the current campaign are associated with a longer time since the client was last contacted from a previous campaign.

The strongest correlation observed is between 'pdays' and 'previous' (0.45482), indicating that clients contacted previously are likely to have been contacted more recently. This could reflect a strategy of focusing on clients who have shown interest in the past.

While the correlations present are mostly weak, they do provide some indication of the relationships between variables. For instance, the strongest correlation between 'pdays' and 'previous' will be especially useful in predictive modeling, as it suggests that past contact is a good predictor of recent contact. The weak correlations elsewhere suggest that multicollinearity is unlikely to be a concern when these variables are included in a predictive model.

**Data partition output**:

| Training | | | | |
| --- | --- | --- | --- | --- |
| Variable | Formatted value | Frequency count | percent | Label |
| y | no | 19961 | 88.2996 | y |
| y | yes | 2645 | 11.7004 | y |

table(4)

| Validation | | | | |
| --- | --- | --- | --- | --- |
| Variable | Formatted value | Frequency count | percent | Label |
| y | no | 19961 | 88.3035 | y |
| y | yes | 2644 | 11.6965 | y |

table(5)

While handling the data partition node, we decided to divide the data as 50% for the training and 50% for the validation data.

**Transformation of variable** :

From the transformation statistics tables the following three variables ( Table : 6 ) have been Transformed ( Table :7 ) because of the high skewness.

| Variable Name | Minimum | Maximum | Mean | Standard Deviation | Skewness | Kurtosis |
| --- | --- | --- | --- | --- | --- | --- |
| Balance | -4057 | 102127 | 1382.029 | 3126.159 | 8.970275 | 158.207 |
| Duration | 0 | 4918 | 258.5178 | 255.3082 | 3.014084 | 17.08493 |
| Previous | 0 | 58 | 0.585862 | 1.91783 | 7.500899 | 101.4746 |

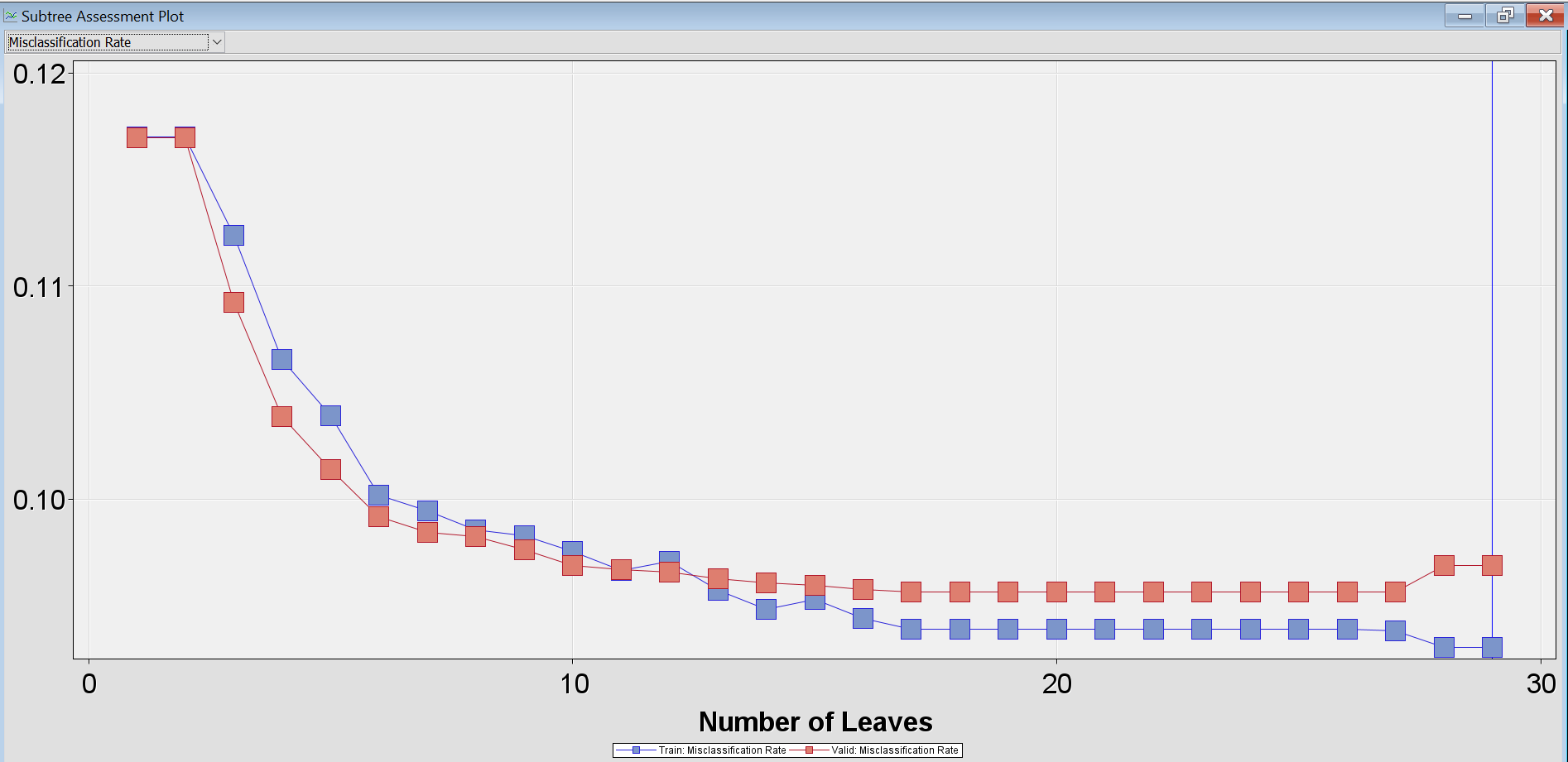
**Table(6)**

| Variable Name | Minimum | Maximum | Mean | Standard Deviation | Skewness | Kurtosis |
| --- | --- | --- | --- | --- | --- | --- |
| LOG Balance | 0 | 11.67294 | 8.530226 | 0.330039 | 1.443702 | 27.0098 |
| LOG Duration | 0 | 8.500861 | 5.174497 | 0.920817 | -0.45131 | 0.863861 |
| LOG Previous | 0 | 4.077537 | 0.230141 | 0.537769 | 2.469031 | 5.823371 |

**Table : (7)**

**Maximal tree**

The decision tree does not have a built-in feature for ignoring redundant inputs. The decision tree uses the logworth = -log(chi-squared p-value) for splitting of the leaf. When the split goes on till the end, it creates the maximal tree with the maximum number of the leaf possible. From the fig(26), we can see that the number of leaves for the maximal tree turns out to be 29 from the subtree assessment plot. The validation misclassification rate is 0.096881.



**Fig( 26 ) - subtree assessment plot for maximal tree**

The decision tree also gives a variable importance table where variables are ranked based on the logworth value as shown in table(8).

| Variable name | Importance |
| --- | --- |
| Duration | 1 |
| poutcome | 0.6926 |
| month | 0.4988 |
| age | 0.2303 |
| contact | 0.2098 |
| housing | 0.2045 |
| campaign | 0.1029 |

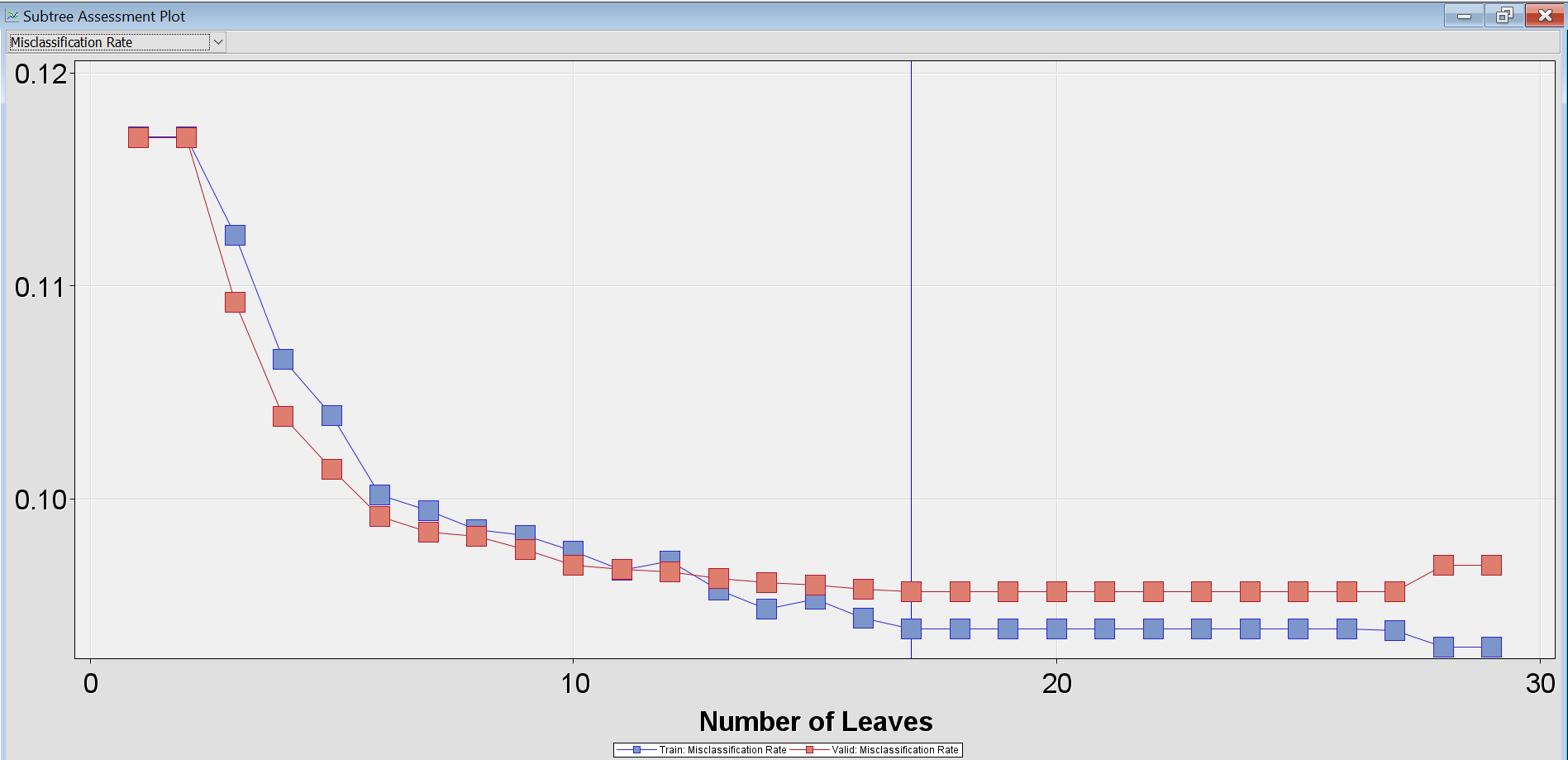
**Table(8) - variable importance for maximal tree**

| False negative | True negative | False postive | True positive |
| --- | --- | --- | --- |
| 1545 | 19344 | 617 | 1099 |

**Table(9) - validation sensitivity and specificity**

**Misclassification tree:**

The maximal tree cannot be used as a good model because it includes all the variables possible without any pruning. Therefore, the pruned decision also known as misclassification tree can was run using the default setting with assessment and misclassification as method in SAS enterprise miner. A tree with 17 leaves was formed as shown in fig(27). The validation misclassification rate is 0.095643.



**Fig( 27 ) - Subtree assessment plot of misclassification tree**

The variable importance table was also created based on the misclassification tree which is shown in table(10). Two variable age and campaign was not considered as important in the pruned tree.

| Variable name | Importance |
| --- | --- |
| Duration | 1 |
| poutcome | 0.6958 |
| month | 0.4948 |
| contact | 0.2157 |
| housing | 0.0699 |

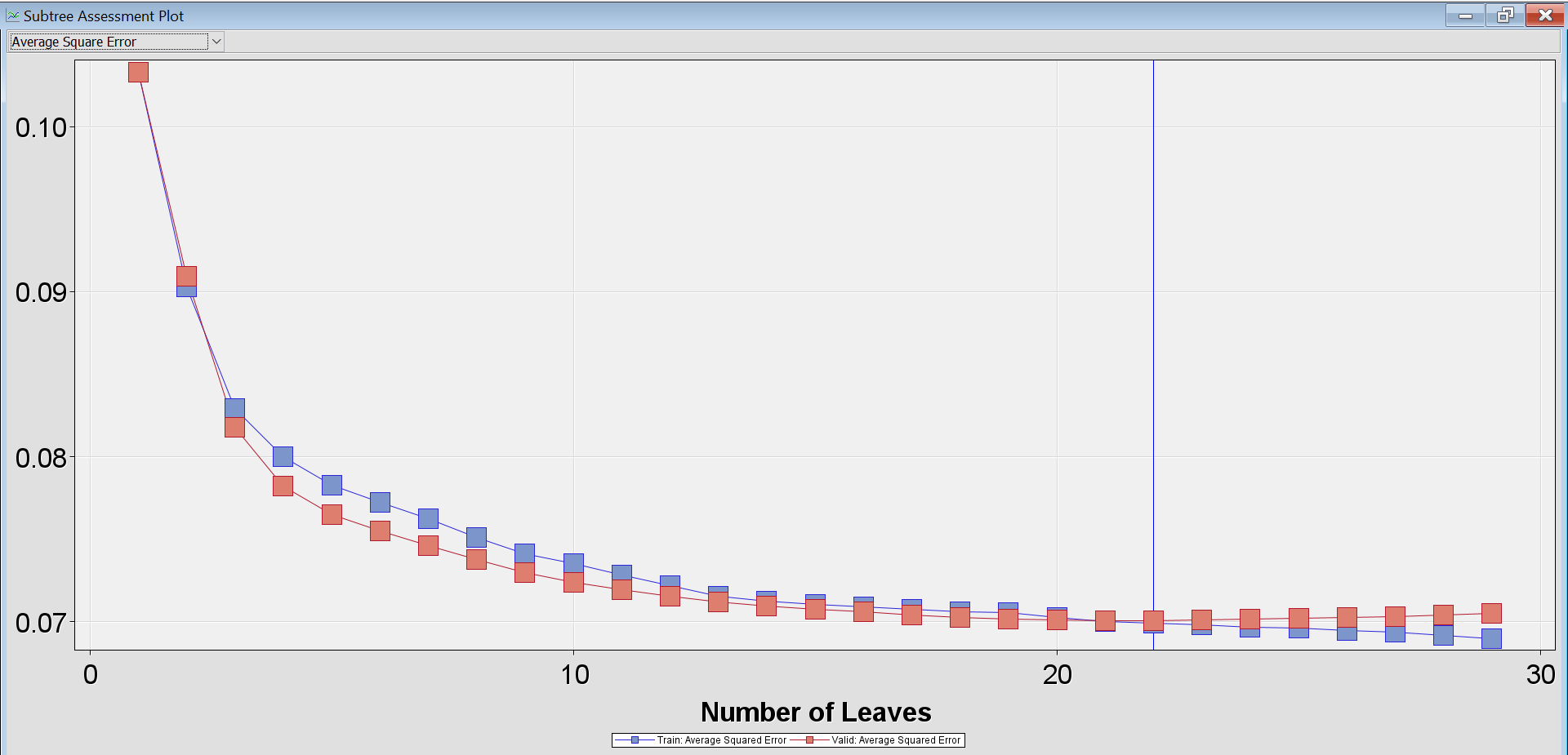
**table(10) - variable importance for misclassification tree**

| False negative | True negative | False positive | True positive |
| --- | --- | --- | --- |
| 1545 | 19344 | 617 | 1099 |

**Table(11) - Validation sensitivity and specificity**

**Probability tree:**

The probability tree is created using the assessment as method and assessment measure as average square error. The tree was created with 22 leaves a shown in fig(28). The validation misclassification rate is 0.095952.



Fig**( 28 ) - Subtree assessment plot for probability tree**

| Variable name | Importance |
| --- | --- |
| Duration | 1 |
| poutcome | 0.6793 |
| month | 0.4887 |
| age | 0.2311 |
| contact | 0.2106 |

**Table (12) - Variable importance for probability tree**

| False negative | True negative | False postive | True positive |
| --- | --- | --- | --- |
| 1664 | 19456 | 505 | 980 |

**Table(13) - validation sensitivity and specificity**

**Stepwise regression:**

| Effect | DF | Wald Chi-Square | Pr > Chisq |
| --- | --- | --- | --- |
| LOG\_Duration | 1 | 2216.0574 | <0.0001 |
| month | 11 | 527.3169 | <0.0001 |
| poutcome | 3 | 426.4290 | <0.0001 |
| contact | 2 | 235.9481 | <0.0001 |
| housing | 1 | 105.2944 | <0.0001 |

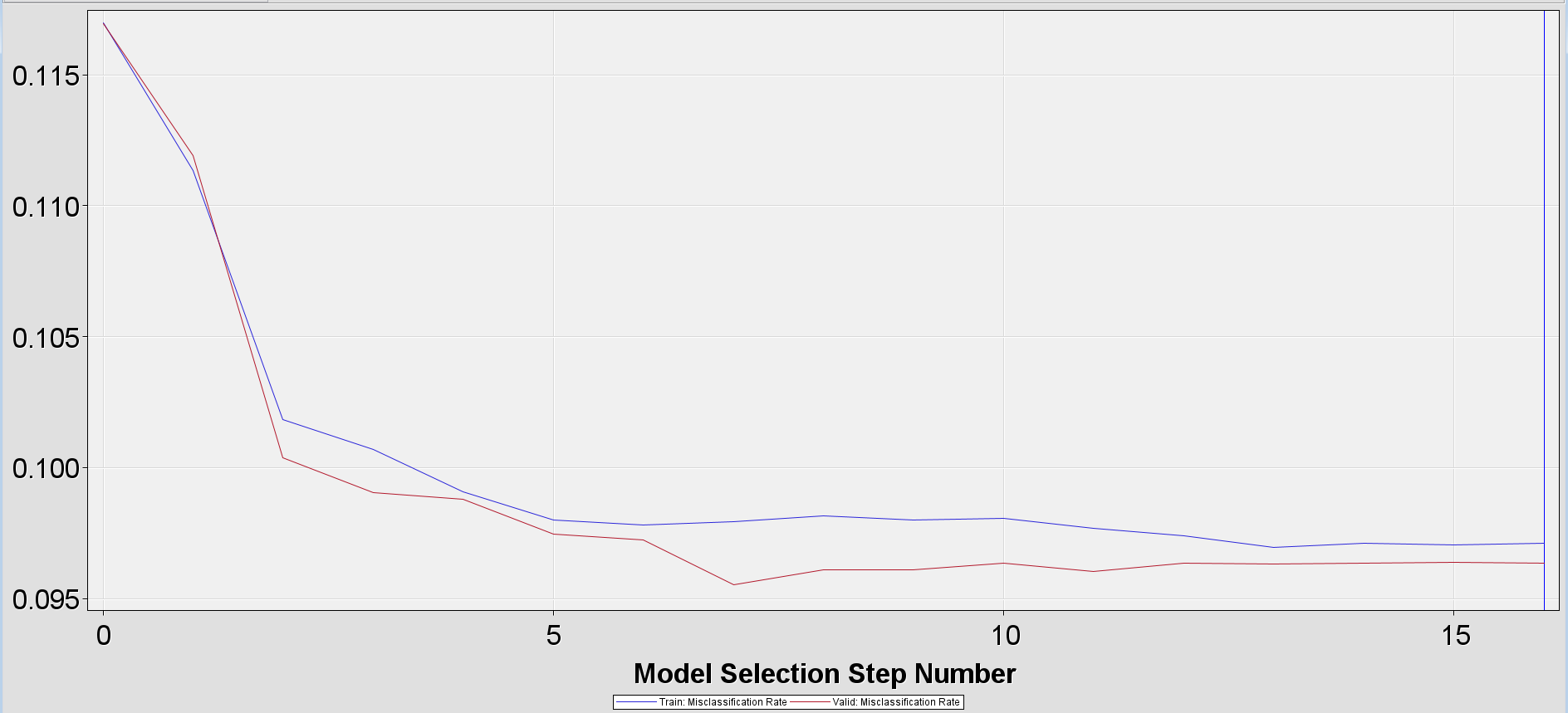
**Table(14)**

The stepwise regression was run using the regression node in the SAS enterprise miner. Since the model uses the binary variable as a target variable, we used the binary logistic regression to fit the model. After running it, the validation misclassification rate turned out to be 0.09635. The output also gave the most important variables in terms of Wald Chi-square. Top 5 variables are shown in the table(14).

**Odds ratio estimate analysis**

Log\_duration - For LOG\_duration, the odd ratio estimate equals 6.640, this means that for each additional duration, the odds of agreeing to term deposit change by a factor of 6.640, a 564% increase.

poutcome - For Outcome, the odds ratio for failure vs unknown estimate equals 1.037, this means that for cases with failure outcome, the odds of agreeing to term deposit is 1.037 times higher than the cases with unknown outcome.



**Fig( 29 ) - Iteration plot of the misclassification rate for stepwise regression**

The iteration plot shows that the model with the smallest misclassification rate occurs in step 7. But the model selection stopped at 16.

**Validation misclassification regression**

The validation misclassification regression was run using the regression node. Since the model uses the binary variable as a target variable, we used the binary logistic regression to fit the model. After running it, the validation misclassification rate turned out to be 0.0955. The output also gave the most important variables in terms of Wald Chi-square and top five variables are shown in the table(15).

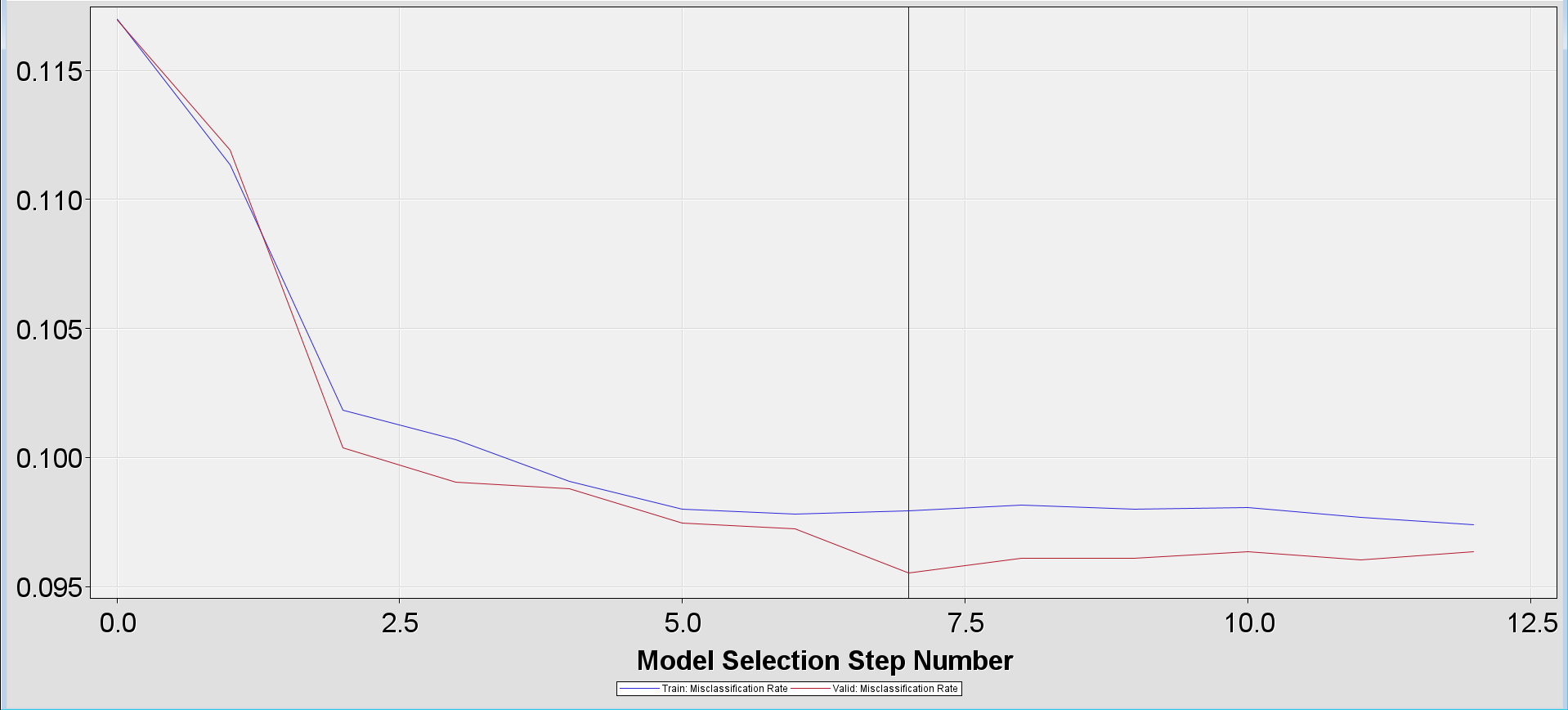
| Effect | DF | Wald Chi-Square | Pr > Chisq |
| --- | --- | --- | --- |
| LOG\_Duration | 1 | 2235.4115 | <0.0001 |
| poutcome | 3 | 595.6692 | <0.0001 |
| month | 11 | 570.0580 | <0.0001 |
| contact | 2 | 235.2660 | <0.0001 |
| housing | 1 | 121.9120 | <0.0001 |

**Table(15)**

**Odd ratio estimate analysis**

Log\_duration - For LOG\_duration, the odd ratio estimate equals 6.640, this means that for each additional duration, the odds of agreeing to term deposit change by a factor of 6.561, a 556.1% increase.

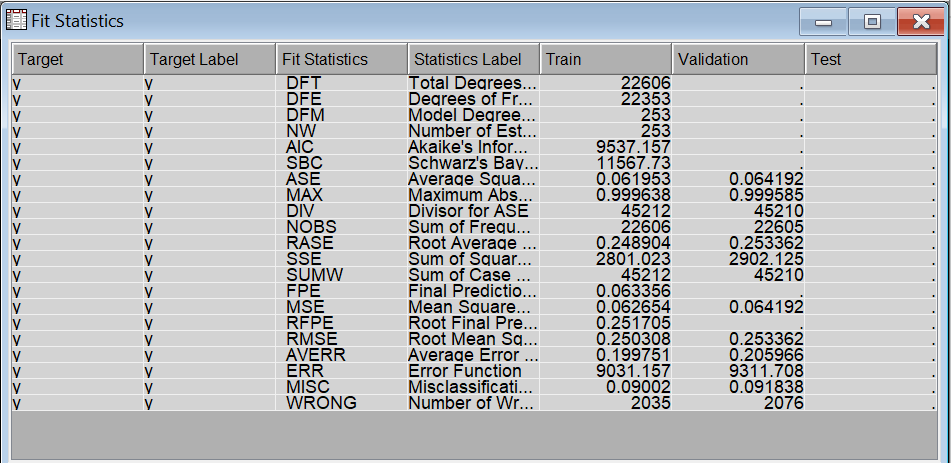
poutcome - For Outcome, the odds ratio for success vs unknown estimate equals 10.991, this means that for cases with success outcome, the odds of agreeing to term deposit is 10.991 times higher than the cases with unknown response.



**Fig( 30 ) - Iteration plot of the misclassification rate for validation misclassification regression**

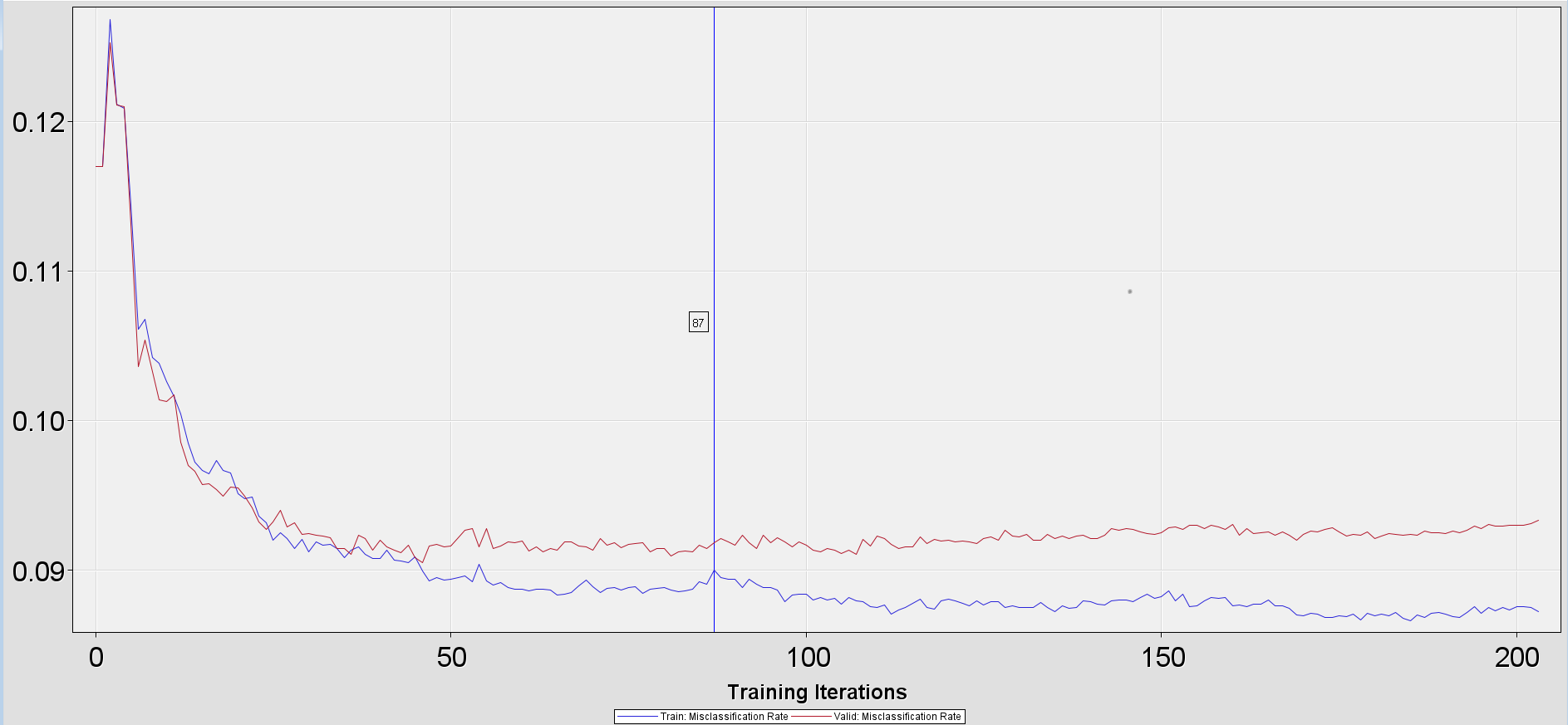
**Neural network using stepwise regression:**

The neural network was connected to the stepwise regression node with 6 hidden units and was run to get the output as shown in table(16).



**Table(16) - Output statistics for neural network with 6 hidden units**

In the assessment of our neural network model, we examined crucial performance indicators. The Average Squared Error (ASE) for the training and validation sets were determined to be 0.0619 and 0.0642, respectively. Similarly, the Root Average Squared Error (RASE) displayed values of 0.250 and 0.253, signifying the overall accuracy of the model's predictions. The Misclassification Rate, indicative of the ratio of incorrectly classified instances, was recorded at 0.09002 and 0.091838 for the training and validation phases. Additionally, the Maximum Absolute Error reached 0.99, representing the maximum disparity between predicted and actual values. Collectively, these metrics suggest a commendable performance of our neural network model. Information criteria, specifically Akaike's Information Criterion (AIC) and Schwarz's Bayesian Criterion (SBC), were computed as 9688.261 and 10755.72, respectively, facilitating model comparisons. Lower values for these criteria indicate a more advantageous balance between goodness of fit and model complexity.

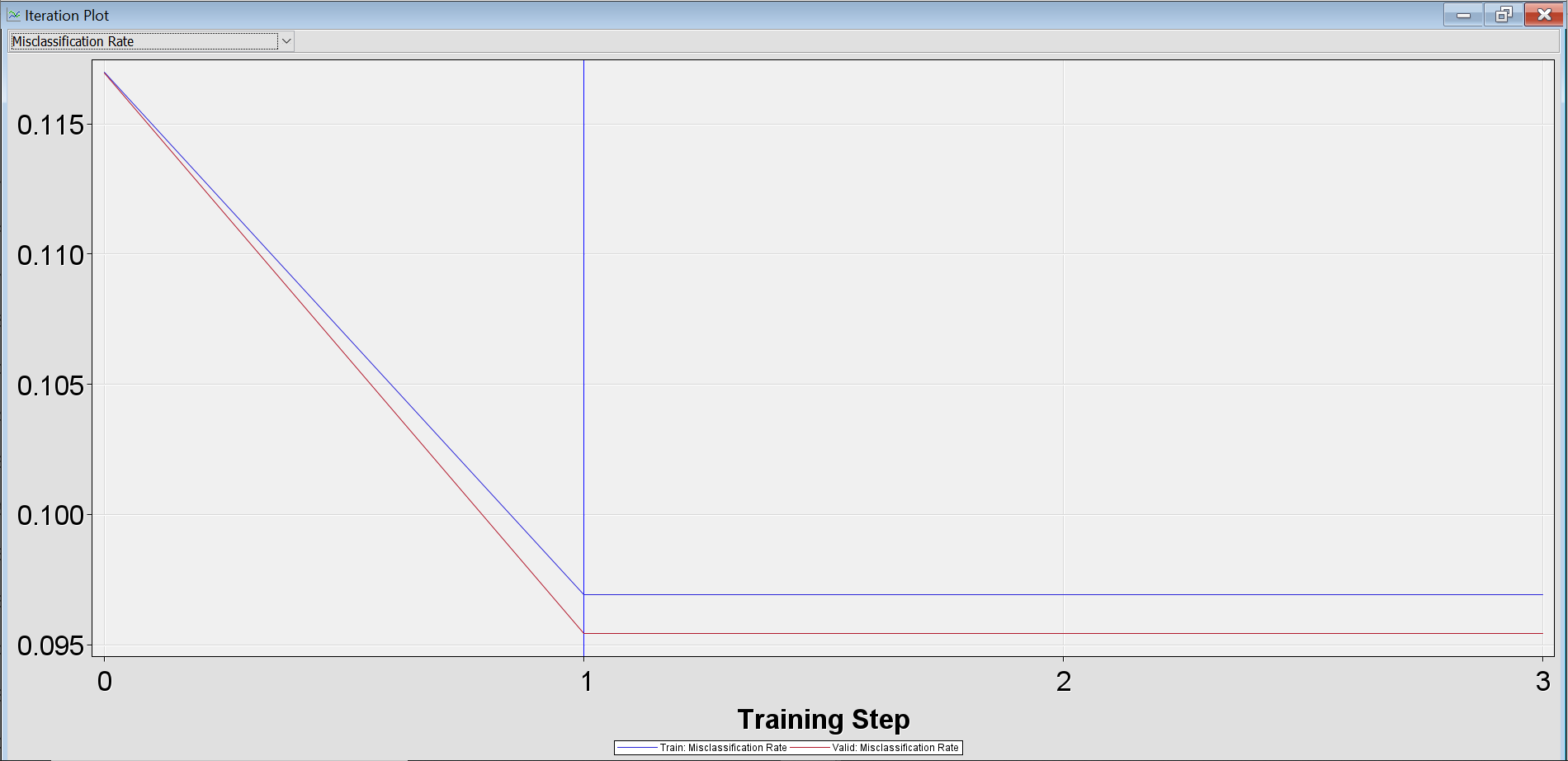


**Fig(31) - Iteration plot**

The neural network selects the modelling weights from iteration 87 for the final model as shown in **fig (31).**  After 87th iteration, the validation rate is staying constant or is increasing at one point.

**AutoNeural using stepwise regression :**

The autoneural network was connected to the stepwise regression and was run to get the misclassification rate of 0.095421. The number of hidden units was changed to 1 from 2. The iteration plot is shown as in fig(32).



**Fig 32**

**Partial least square regression**

| Number of Extracted Factors | X-variation accounted for by this factor | Total X variation accounted so far | Y variation accounted for by this factor | Total Y variation accounted for so far |
| --- | --- | --- | --- | --- |
| 1 | 6.4683 | 6.4683 | 18.0281 | 18.0281 |
| 2 | 4.1168 | 10.5851 | 6.5042 | 24.5323 |
| 3 | 5.0418 | 15.6269 | 0.7149 | 25.2472 |
| 4 | 3.9117 | 19.5386 | 0.2559 | 25.5031 |
| 5 | 3.1948 | 22.7334 | 0.0783 | 25.5813 |
| 6 | 2.6156 | 25.3490 | 0.0267 | 25.6081 |
| 7 | 2.8102 | 28.1592 | 0.0099 | 25.6179 |
| 8 | 2.5470 | 30.7062 | 0.0019 | 25.6198 |
| 9 | 3.3604 | 34.0666 | 0.0003 | 25.6201 |
| 10 | 2.7224 | 36.7890 | 0.0004 | 25.6205 |
| 11 | 2.0834 | 38.8724 | 0.0004 | 25.6209 |
| 12 | 2.6032 | 41.4756 | 0.0001 | 25.6210 |
| 13 | 2.4103 | 43.8858 | 0.0000 | 25.6210 |
| 14 | 2.1788 | 46.0646 | 0.0000 | 25.6210 |
| 15 | 1.7024 | 47.7670 | 0.0000 | 25.6210 |

**Table (17)**

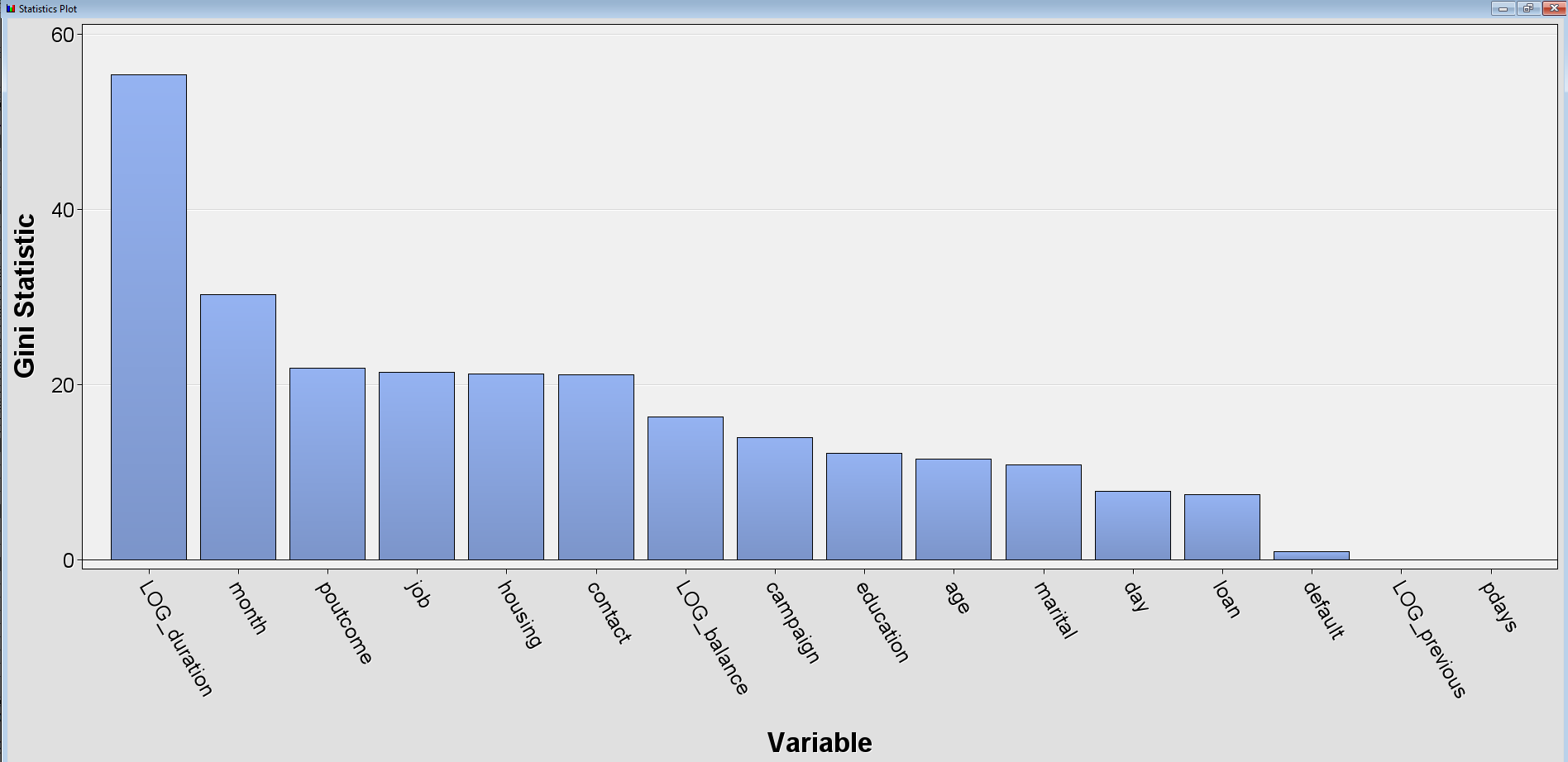
The partial least square toll extracts 15 latent factors from the training data set as shown in table (17). These factors account for 48% of the variation in the inputs and 5.2% of the variation in the target.

The PLS selected some variables as input and rejected the rest of it. The variable which was selected is mentioned in the table().

| Variable | Variable importance | role |
| --- | --- | --- |
| LOG duration | 1.926766 | Input |
| Poutcome SUCCESS | 1.450531 | Input |
| Poutcome UNKNOWN | 0.929698 | Input |
| LOG previous | 0.875388 | Input |
| Contact UNKNOWN | 0.836971 | Input |

**Table (18)**

**Interactive binning**

****

**Fig (33) - Variable importance based on gini statistic**

**Quantile binning with regression**

The quantile binning transforms the numerical variable to categorical variable by dividing it into different bins. To be exact, the quantile node partitioned each interval into four bins with equal sizes. This might help in analyzing complex models and improve the performance of it. The misclassification rate turns out to be 0.099978.

**Bucket binning with regression**

Similar to the quantile binning, the bucket node partitioned each interval input into four bins but with equal widths. The misclassification rate turns out to be 0.104711.

**Optimal binning with regression**

The optimal node partitions for each input variable using the decision tree methods. The misclassification rate turns out to be 0.096793.

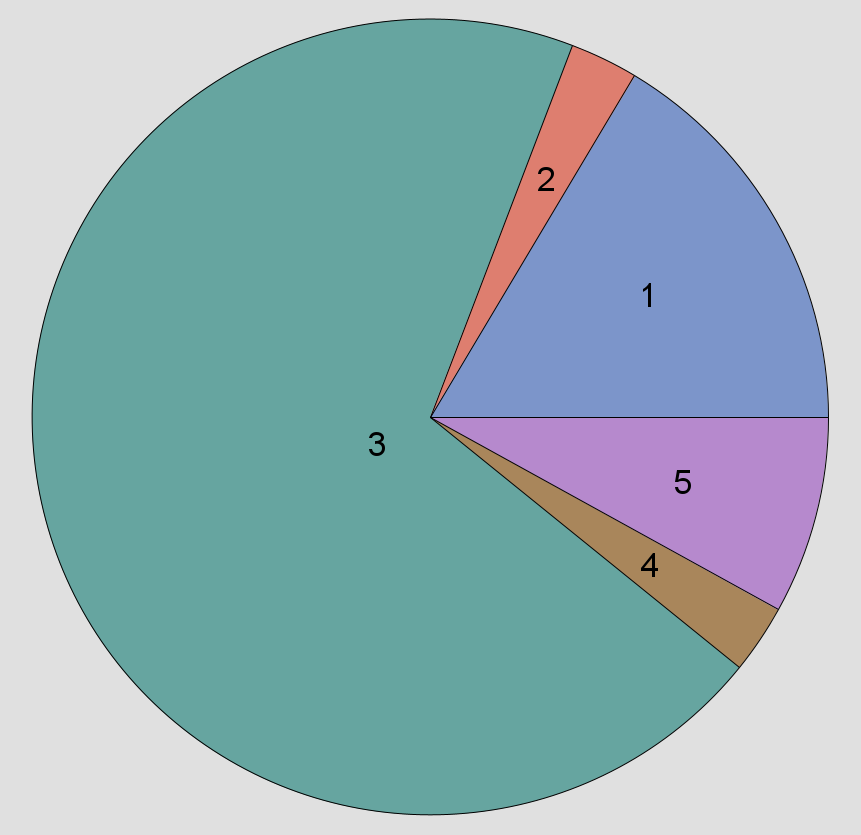
**Ensemble model**

The ensemble model is the combination of all the models which are connected to the control point. It creates a new model by combining the posterior probabilities for class targets or predicted values for interval targets. This model’s validation misclassification rate turns out to be 0.095687.

**Cluster analysis:**

| Variable name | Number of splitting rules | Number of surrogate rules | Importance |
| --- | --- | --- | --- |
| Pdays | 1 | 1 | 1.0000 |
| Duration | 1 | 1 | 0.95047 |
| poutcome | 2 | 1 | 0.78995 |

**Table(19) - Variable importance table of cluster analysis**



**Fig (34)**

We chose 3 variables for the cluster analysis which is duration, poutcome and pdays . We selected the number of clusters to be 5. The segment 3 contained the highest number of frequencies to be 27755 with lowest duration of 158 seconds. It looks like most of the failures, unknown or other categories are in segment 3.

**Model comparison:**

| Model | Validation misclassification rate | Validation ROC index |
| --- | --- | --- |
| SW neural network HU=6 | 0.091838 | 0.928 |
| Neural network | 0.092988 | 0.926 |
| SW AutoNeural | 0.095421 | 0.909 |
| Valid Misclassification regression | 0.095510 | 0.907 |
| Misclassification tree | 0.095643 | 0.813 |
| Ensemble | 0.095687 | 0.921 |
| Probability tree | 0.095952 | 0.885 |
| Polynomial regression | 0.096306 | 0.909 |
| Stepwise regression | 0.096350 | 0.909 |
| SW Optimal binning | 0.096793 | 0.908 |
| Variable selection chi square regression | 0.096881 | 0.906 |
| Maximal tree | 0.096881 | 0.885 |
| VS R square regression | 0.097722 | 0.905 |
| Partial least square regression | 0.099359 | 0.889 |
| SW Quantile binning | 0.099978 | 0.904 |
| SW bucket binning | 0.104711 | 0.787 |

**Table (20) - Model comparison**

The champion model for our analysis turns out to be the stepwise regression with a neural network of hidden units = 6 with the lowest validation misclassification rate of 0.091838 and with highest validation ROC index of 0.928 as shown intable (20).

|  | Predicted | | |
| --- | --- | --- | --- |
| Actual |  | positive | negative |
| positive | 1239 (TP) | 1405 (FN) |
| negative | 671 (FP) | 19290 (TN) |

**Table(21) - Confusion matrix of the champion model**

**Odd ratio estimate for Champion model**

For LOG\_duration, the odd ratio estimate equals 6.640, this means that for each additional duration, the odds of agreeing to term deposit change by a factor of 6.640, a 564% increase.

For Outcome, the odds ratio for failure vs unknown estimate equals 1.037, this means that for cases with failure outcome, the odds of agreeing to term deposit is 1.037 times higher than the cases with unknown outcome.

For Outcome, the odds ratio for other vs unknown estimate equals 1.272, this means that for cases with other outcome, the odds of agreeing to term deposit is 1.037 times higher than the cases with unknown outcome.

For Outcome, the odds ratio for success vs unknown estimate equals 10.074, this means that for cases with success outcome, the odds of agreeing to term deposit is 10.074 times higher than the cases with unknown response.

For contact, the odds ratio for cellular vs unknown estimate equals 4.750, this means that for cases with cellular contact, the odds of agreeing to term deposit is 4.750 times higher than the cases with unknown contact.

For contact, the odds ratio for telephone vs unknown estimate equals 4.910, this means that for cases with telephone contact, the odds of agreeing to term deposit is 4.910 times higher than the cases with unknown contact.

**Conclusion :**

The banking dataset was uploaded to the SAS enterprise miner to analyse the important variables and turns out that transformed variables of Duration which is LOG\_duration was the most important variable followed by month, poutcome, contact and housing. Based on this we can conclude that the longer the customer stayed on call, they were more than likely to subscribe to the term deposit. As this data set is from 2008 to 2010, that time all the marketing was done using calls. Furthermore, the clients were used to picking up the calls because that time there was such a thing which could block or stop the unnecessary calls. Based on the information we have from about 13 years ago, we can still use the data to target the customer which will agree to a term deposit today but, the only difference will be that this time the contact approach will be not only by call but also by emails, mail and internet advertisement. From the poutcome variable, we know which customers subscribed last time so, based on that for next term deposit we can only target those customers first. It will save time and money for the institution. The project utilizes predictive analytics to enhance the marketing strategy of the Almighty Portuguese Bank, aiming to deliver targeted promotions, optimize resource allocation, and achieve higher conversion rates. The objective is to generate valuable insights that can be applied to optimize communication channels, scheduling, and overall effectiveness of marketing initiatives, ultimately leading to increased growth in term deposits for the bank. The project utilizes predictive analytics to enhance the marketing strategy of the Almighty Portuguese Bank, aiming to deliver targeted promotions, optimize resource allocation, and achieve higher conversion rates. The objective is to generate valuable insights that can be applied to optimize communication channels, scheduling, and overall effectiveness of marketing initiatives, ultimately leading to increased growth in term deposits for the bank.

References:

James Chen. (2022, March 20). Term Deposit. Investopedia.

<https://www.investopedia.com/terms/t/termdeposit.asp>

Junfeng Guo; Handan Hou (January 2019), Statistical Decision Research of Long-Term Deposit Subscription in Banks Based on Decision Tree

<https://ieeexplore.ieee.org/abstract/document/8669650>

Kinan Morani; Esra Kaya Ayana; Şeref Naci Engin(October 2018), Development of Prediction in Clients’ Consent to a Bank Term Deposit Using Feature Selection

<https://ieeexplore.ieee.org/document/8751816>

Rony; Md. Mehedi Hassan; Eshtiak Ahmed ; Asif Karim; Sami Azam; D. S. A. Aashiqur Reza(December 2021), Identifying Long-Term Deposit Customers: A Machine Learning Approach Mohammad Abu Tareq

<https://ieeexplore.ieee.org/document/9672452>

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, Banking Dataset - Marketing Targets, Banking Dataset of different customers to predict if they will convert or not.

<https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets>

Zian Shin, Jihoon Moon, Seungmin Rho(November 2021) , A Comparative Analysis of Ensemble Learning-Based Classification Models for explainable Term Deposit Subscription Forecasting

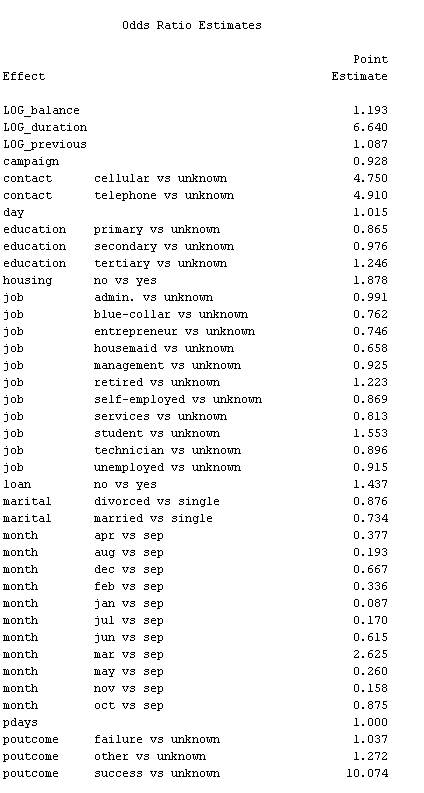
<http://calsec.or.kr/jsebs/index.php/jsebs/article/view/457>

Zarinabegam Kasimali Mundargi; Aditya Bodhankar; Ajinkya Mahajan; Diksha

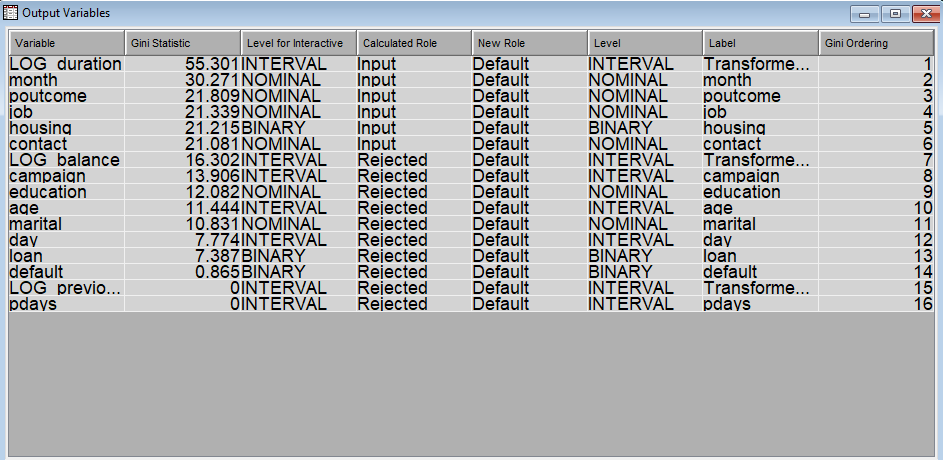
Prasad; Shivani Mahajan; Riya Dhakalkar(January 2023), Bank Fixed Term Deposit analysis using Bayesian Logistic Regression

<https://ieeexplore.ieee.org/document/10080695>

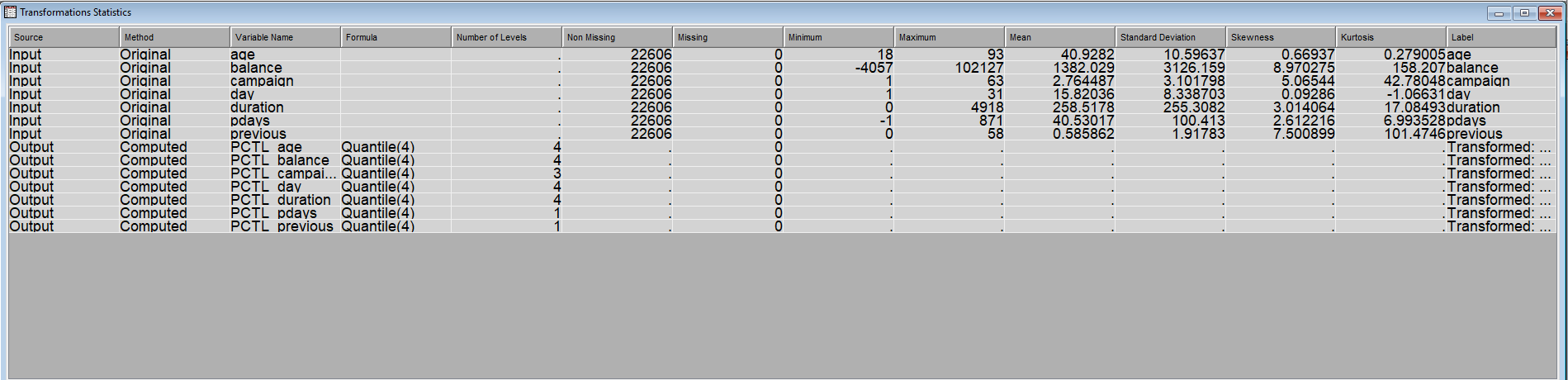
**Appendix :**

****

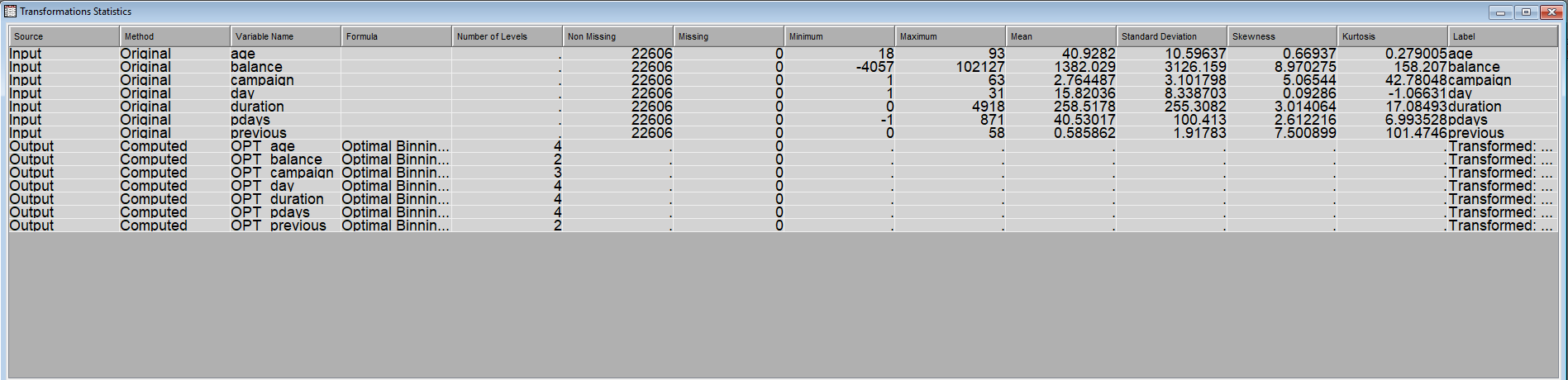
**A. Fig(1) - SW odd ratio estimate**

****

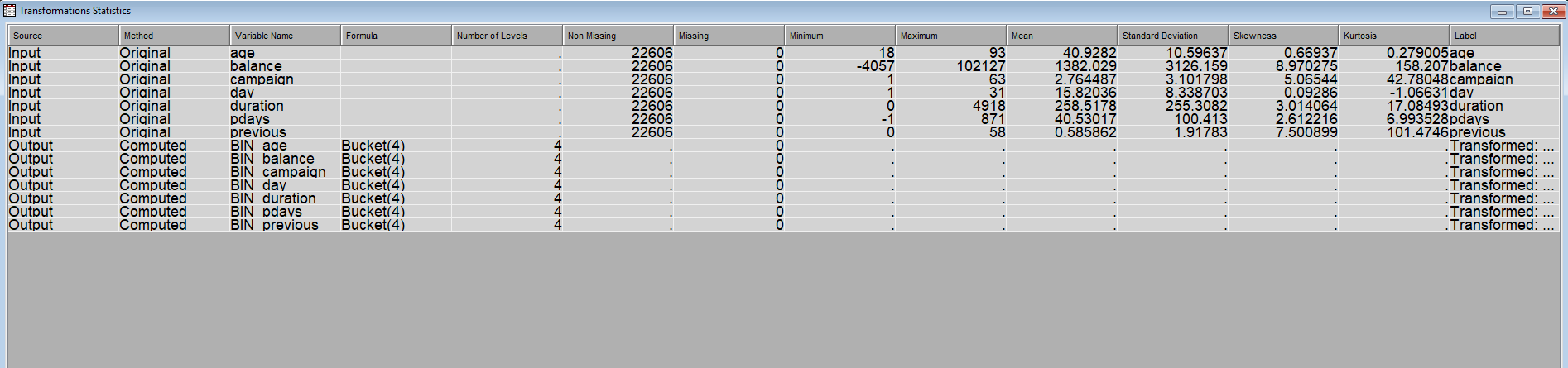
1. **Fig(2) - interactive binning**

****

1. **Fig(3) - Quantile binning**

****

1. **Fig(4) - Optimal binning**

****

1. **Fig(4) - Bucket binning**