­­

Business Intelligence

**Term Project**

**Submitted to: Submitted by:**

**Prof.Himadri Das PGDM, Group 2**

**Aniruddha Singh Chauhan (08pgdm70)**

**Bastine Kottackal (08pgdm80)**

**Bhavya Upadhyay (08pgdm81)**

**Raman Chaddha (08pgdm105)**

**Predictive Modeling – Insurance Buying**

**Problem Definition**

The task is to predict who buys mobile home (caravan) insurance in a particular region. The approach to be followed is to build the most accurate predictive data mining model which for a given set of attributes of a potential customer would predict whether the customer would actually buy the insurance or not.

The problem and the data set are authentic and have been taken from the web archives.

**Business Benefit**

Direct mailings to a company’s potential customers can be a very effective way for to market a product or service. However, much of this junk mail is really of no interest to the majority of the people that receive it. Most of it ends up thrown away, not only wasting the money that the company spent on it, but also filling up landfill waste sites or needing to be recycled.For the prediction task, the underlying problem is to the find the subset of customers with a probability of having a insurance policy above some boundary probability. The known policyholders can then be removed and the rest receives a mailing. The boundary depends on the costs and benefits such as of the costs of mailing and benefit of selling insurance policies.

The purpose is to give a clear insight to why customers have insurance policy and how these customers are different from other customers. It is useful and actionable for a marketing professional with no prior knowledge of computational learning technology to develop marketing strategy.

**Attributes**

There are 44 socio-demographic attributes derived via the customer's ZIP area code and 43 product ownership attributes about insurance of other policies. The data is provided for 13,335 people and is historic in nature, i.e., the distribution between buyers and non-buyers for this dataset is already known.

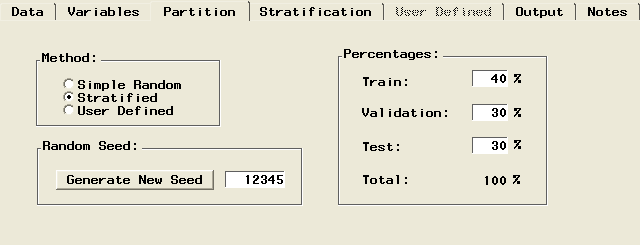
These data variables are listed in the appendix of the report. Most of the interval and nominal data is in a coded format, the codes for which are also specified. For example for the attribute ‘Roman Catholic’, suppose the customer value is 5, it implies that 50-62 % of the people living in the neighborhood of that person are Roman Catholic. Similarly suppose the value is 6 for “Roman Catholic” means 63-75% people living in the neighborhood of that person are Roman Catholic. Please see the Appendix for the details.

**Techniques Used**

The problem statement demanded building an accurate predictive model. Using the Enterprise Miner, we have compared two modeling techniques, which are Decision Trees and Neural Networks.

**Project Methodology**

The first step was to partition the data. This was done using the Data Partition node. The ratios are depicted in the following snapshot:



These were the default values (40% training data, 30% validation data and 30% test data). Sampling method used was Stratified Sampling and the data was sampled based on the target variable, that is, ‘no. of mobile home policies (0-1).’ This was done to use proper stratified data for all the three data segments.

We experimented with both Decision Trees and Neural Networks. In decision trees, we first modeled using binary split, then we modeled three way split and the four way split to find out which split gives more homogenous sets at each division.

Similarly in the case of neural networks we modeled using 1 hidden node, 2 hidden nodes, 3 hidden nodes etc. Based on the misclassification error we chose which one is better suitable for prediction.

**Results**

**Decision Tree**

1) **2 way split** (Binary split)

The nodes which is able to get a homogenous outputs give the results

IF 0.5 <= CONTRIBUTION\_BYCLE\_POLICIES

AND 3.5 <= CONTRIBUTION\_FIRE\_POLICIES < 4.5

AND INCOME\_\_\_30 < 2.5

AND CONTRIBUTION\_CAR\_POLICIES < 5.5

THEN

NODE : 35

N : 11

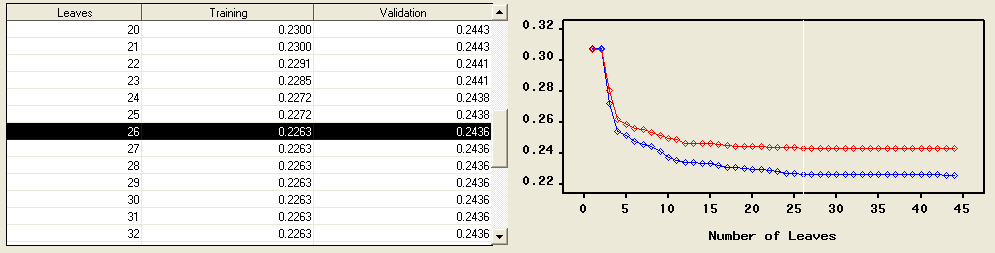
0 : 0.0%

1. : 100.0%

Which means that the customers whose contribution towards bicycle policies is greater than 0 and contribution to fire policies is from Rs 100 to 500 and the customers who lives in the neighborhood where 23% of people have income less than Rs30000 and contribution toward the car policies is between Rs 500 to Rs 900 .( Please refer the appendix attached with the project since for some variables we need to use the consumer contribution and for other variables we need to use the %of people who lives in his neighborhood with specific criteria(eg income) )

Similarly we can see that node nodes 67, 83, 87 are able to split the data into homogenous groups. After analyzing these data we can see that contribution toward car policy, contribution towards the fire policy and the income level in the neighborhood are the major factors which determine the purchase of insurance.

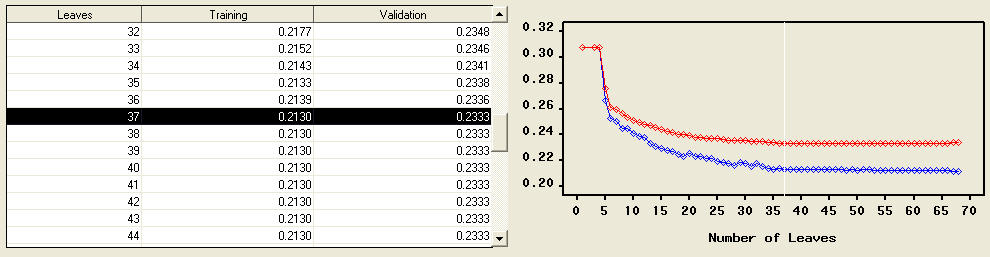
**2-way split**



Here we can see that after 26 leaves the training value is constant at 0.2263.

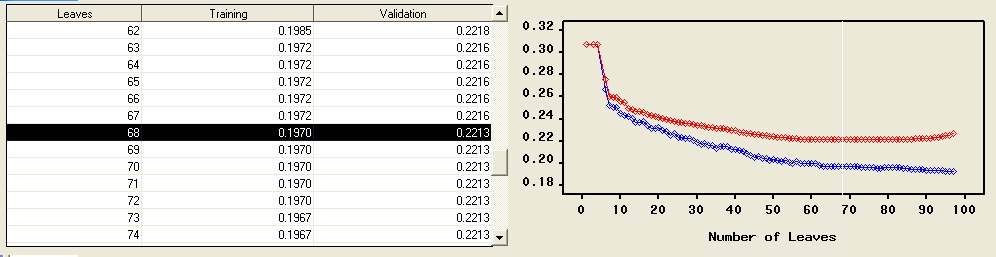
Similarly we have taken 3-way split and 4 way split of trees

ii) **3 way split**



Here we can see that the training error value is constant after 37 nodes at a value of 0.2130

iii) 4-way split



Here we can see that after 68 nodes the error value is constant at 0.1970.

**Analysis of different Splits of trees**

After analyzing the different ways of tree split, 4 way split gives the minimum level of error, and when we used 5 way split there is no significant improvement, we have concluded that 4-way split is the ideal.

Results of 4 –way split

Here Nodes 89 and 114 is the only two nodes which split the data into 100% customers who buy insurances.

IF NO\_\_OF\_MOPED\_POLICIES < 0.5

AND 0.5 <= NO\_CAR < 3.5

AND SKILLED\_LABOURERS < 2.5

AND CONTRIBUTION\_FIRE\_POLICIES < 1.5

AND 5.5 <= CONTRIBUTION\_CAR\_POLICIES < 6.5

THEN

NODE : 88

N : 289

0 : 66.1%

1 : 33.9%

It can be interpreted as no moped policies is less than 1 and no car is between 0 to 4 and the percentage of skilled labors is less than 23% and contribution to car policies 500 to 6999 are the people who most likely to buy insurance.

IF 6.5 <= SKILLED\_LABOURERS

AND 1.5 <= RENTED\_HOUSE

AND NO\_\_OF\_MOPED\_POLICIES < 0.5

AND CONTRIBUTION\_FIRE\_POLICIES < 0.5

AND INCOME\_\_\_30 < 2.5

AND CONTRIBUTION\_CAR\_POLICIES < 5.5

THEN

NODE : 114

N : 5

0 : 0.0%

1. : 100.0%

It can be interpreted as the percent of skilled labors in the neighborhood is less than 75% and the percentage of people in the neighborhood living in the rented house is greater than 10% and the number of moped policies is zero and the contribution to fire polices is zero and percentage of people in the neighborhood with income less than 30 % is less than 36% and contribution toward car policies is less than Rs999 are the people who buy insurance.

When we combined the results of those two nodes we can find that the major factors influencing the buying are

1. CONTRIBUTION\_CAR\_POLICIES
2. CONTRIBUTION\_FIRE\_POLICIES
3. NO\_\_OF\_MOPED\_POLICIES
4. SKILLED\_LABOURERS in the neighborhood

No of moped policies is zero and contribution toward file policies are less Rs50 are the two common values in those nodes. And contribution toward car polices is less than 4999 is the union set for those attribute. But %of skilled neighbors show a different range for different nodes.

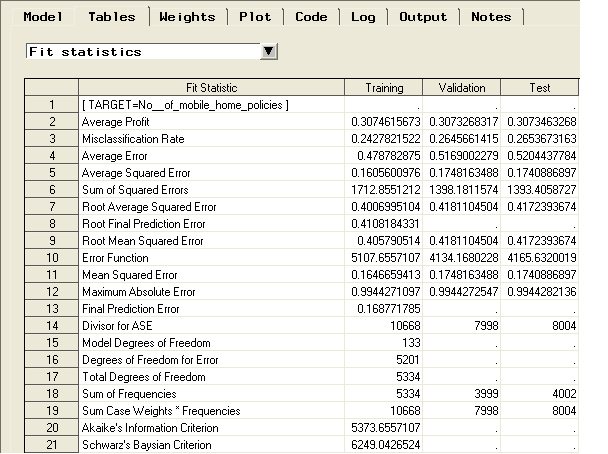
**Based on this we can conclude that the people who most likely to buy insurance are:**

1. No of moped policies is zero
2. Contribution toward file policies are less Rs50
3. Contribution toward car polices is less than 4999

**Neural Networks**

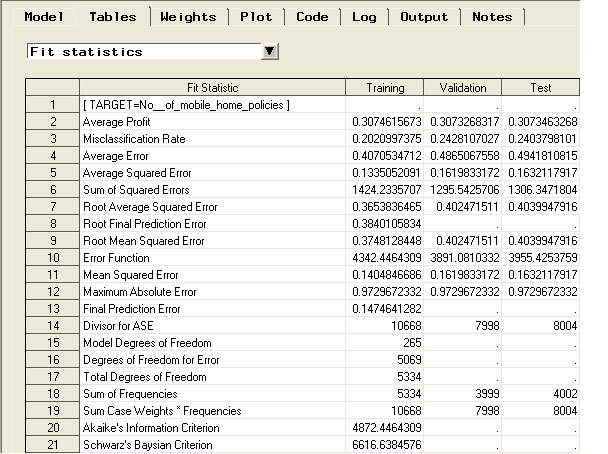
Here we tried neural networks with two hidden neurons, three hidden neurons, four hidden neurons etc, and based on the misclassification rates we chose the ideal neural network for prediction. We also used the help of lift chart to assess different models.

**1 hidden neuron decision tree**



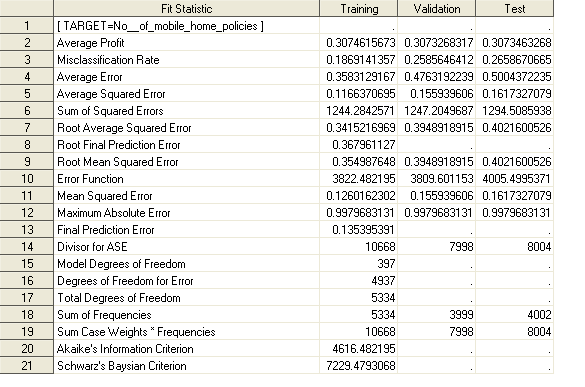
Here we can see that the misclassification rate is 26.53% for the test data.

**2 hidden neuron decision tree**

****

Here we can see that the misclassification rate 24.03% for the test data.

**3 hidden neuron decision tree**

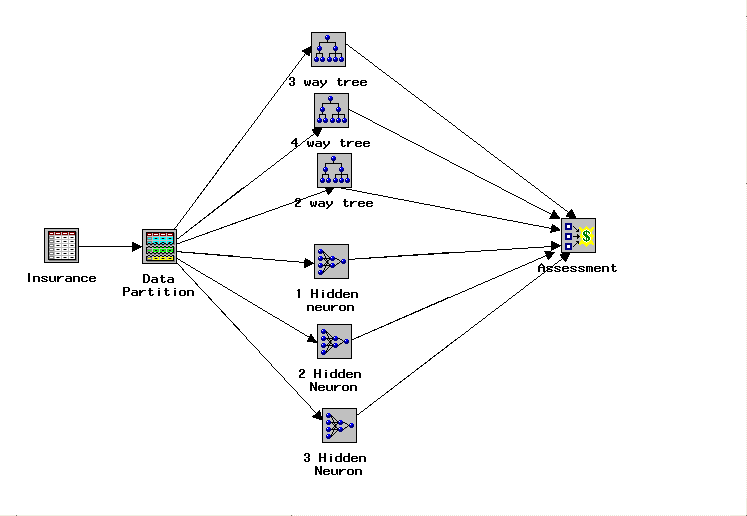
****

Here the misclassification rate for the test data is 26.58%.

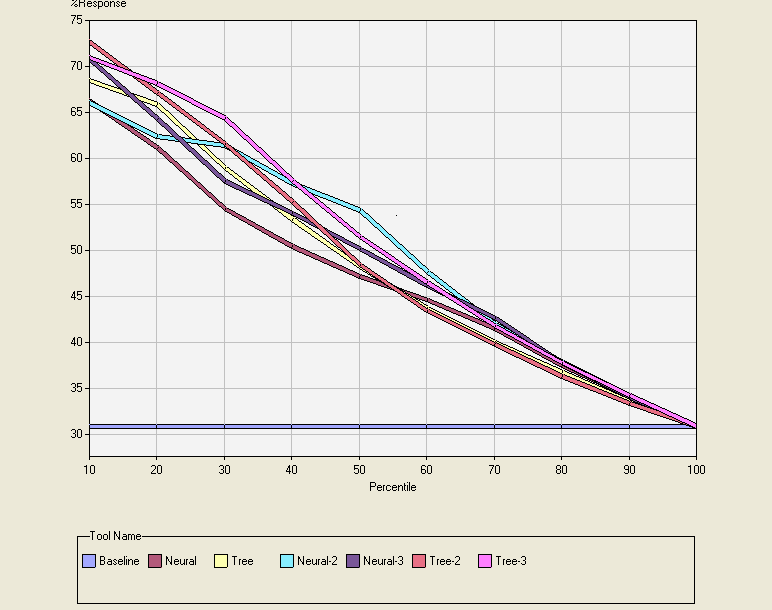
**Since the misclassification rate reduces from 1 hidden node to 2 hidden nodes, but then increases for the 3 hidden nodes’ model, 2 hidden nodes decision tree is the most accurate one.**

**Decision Tree v/s Neural Network**

All the models were compared for accuracy using the assessment node. The following snapshot shows the linkages:



On comparing the most accurate models of decision trees and neural networks, we see that 2 hidden nodes’ neural network is the most accurate since the misclassification rate is least in this case. This fact is re-emphasized by the following lift chart:



**Conclusions**

1. The strongest single predictor of having an insurance policy is having a single car insurance policy where the contribution is high (Rs 1000-4999), or having two car policies.

2) The other predictors that are most significant are:

- "purchasing power class" is high

- a private third party insurance policy

- a boat policy

- a social security insurance policy

- a single fire policy with higher contribution (Rs 200-499)

3) Intuitively, these predictors identify customers who have a car and are wealthier than average, and who in general carry more insurance coverage than average. It is not surprising that these are the people who are most likely to have insurance. A customer with a single car policy with a low premium is less likely than average to have insurance.

This finding is actionable and it was not obvious in advance. A customer of this type is presumably less wealthy or less risk-averse, so less likely to own an insurance or less likely to buy insurance for it.

1. The income level of the customer (or more specifically the customer’s neighborhood), is also important. The likelihood of buying the caravan insurance for those with very low purchasing power is low, and increases with purchasing power but only up to a point. At very high levels of purchasing power, the likelihood of buying the caravan insurance reduces somewhat, as can be seen from Decision Tree.
2. Variables relating to education level were also found to be significant. We attribute this to the fact that education correlates closely with income, and do not believe that education level should be considered to have an independent effect on caravan policy purchases.
3. The amount spent on fire policies was also found to be important. This result is less intuitive and may warrant further investigation. In some sense, the fact that a family buys any kind of insurance makes them more likely to buy a caravan policy, and surely purchasers of fire insurance are likely to be in the upper-middle-class target population.
4. Looking at customer main type, we find that the most likely buyers belong to the driven grower category, followed by successful hedonist. However, the living well category was unlikely to buy the policy, again suggesting that the target population is upper-middle-class, but not extremely wealthy. The categories of ‘cruising seniors’ and ‘retired and religious’ were unlikely to buy the caravan policy. This would seem like a group that would be likely to both buy a caravan, and to insure it. Farmers are extremely unlikely to buy caravan policies. Customer subtype is harder to analyze because of the small number of individuals in most categories.

Categories with strong likelihood of buying policies include, in order, ‘affluent young families’, ‘high income, expensive child’, and ‘high status seniors.’ Categories unlikely to buy policies include ‘young and rising’, ‘mixed rural’, ‘large family farms’, ‘young, low educated’, and ‘low income Catholics.’

**In conclusion, we suggest a partial customer profile based on the available input features: Established families, upper-middle to upper wealth level, at least two insured automobiles, with other insurance policies a plus. People to avoid include low-income, no cars, farmers and, surprisingly, senior citizens**.

**Appendix-1 Attributes**

|  |  |  |
| --- | --- | --- |
| Nr | Attributes | Description |
| 1 | Customer main type | see L2 |
| 2 | Customer Subtype | see L0 |
| 3 | Number of houses | 1 - 10 |
| 4 | Avg size household | 1 - 6 |
| 5 | Avg age | see L1 |
| 6 | Roman catholic | see L1 |
| 7 | Protestant ... | see L1 |
| 8 | Other religion | see L1 |
| 9 | No religion | see L1 |
| 10 | Married | see L1 |
| 11 | Living together | see L1 |
| 12 | Other relation | see L1 |
| 13 | Singles | see L1 |
| 14 | Household without children | see L1 |
| 15 | Household with children | see L1 |
| 16 | High level education | see L1 |
| 17 | Medium level education | see L1 |
| 18 | Lower level education | see L1 |
| 19 | High status | see L1 |
| 20 | Entrepreneur | see L1 |
| 21 | Farmer | see L1 |
| 22 | Middle management | see L1 |
| 23 | Skilled labourers | see L1 |
| 24 | Unskilled labourers | see L1 |
| 25 | Social class A | see L1 |
| 26 | Social class B1 | see L1 |
| 27 | Social class B2 | see L1 |
| 28 | Social class C | see L1 |
| 29 | Social class D | see L1 |
| 30 | Rented house | see L1 |
| 31 | Home owners | see L1 |
| 32 | 1 car | see L1 |
| 33 | 2 cars | see L1 |
| 34 | No car | see L1 |
| 35 | National Health Service | see L1 |
| 36 | Private health insurance | see L1 |
| 37 | Income < 30 | see L1 |
| 38 | Income 30-45.000 | see L1 |
| 39 | Income 45-75.000 | see L1 |
| 40 | Income 75-122.000 | see L1 |
| 41 | Income >123.000 | see L1 |
| 42 | Average income | see L1 |
| 43 | Purchasing power class | see L1 |
| 44 | pr\_num | 8 digit code |
| 45 | Contribution private third party insurance | see L4 |
| 46 | Contribution third party insurance (firms) | see L4 |
| 47 | Contribution third party insurane (agriculture) | see L4 |
| 48 | Contribution car policies | see L4 |
| 49 | Contribution delivery van policies | see L4 |
| 50 | Contribution motorcycle/scooter policies | see L4 |
| 51 | Contribution lorry policies | see L4 |
| 52 | Contribution trailer policies | see L4 |
| 53 | Contribution tractor policies | see L4 |
| 54 | Contribution agricultural machines policies | see L4 |
| 55 | Contribution moped policies | see L4 |
| 56 | Contribution life insurances | see L4 |
| 57 | Contribution private accident insurance policies | see L4 |
| 58 | Contribution family accidents insurance policies | see L4 |
| 59 | Contribution disability insurance policies | see L4 |
| 60 | Contribution fire policies | see L4 |
| 61 | Contribution surfboard policies | see L4 |
| 62 | Contribution boat policies | see L4 |
| 63 | Contribution bicycle policies | see L4 |
| 64 | Contribution property insurance policies | see L4 |
| 65 | Contribution social security insurance policies | see L4 |
| 66 | Number of private third party insurance | 1 - 12 |
| 67 | Number of third party insurance (firms) | 1 - 12 |
| 68 | Number of third party insurane (agriculture) | 1 - 12 |
| 69 | Number of car policies | 1 - 12 |
| 70 | Number of delivery van policies | 1 - 12 |
| 71 | Number of motorcycle/scooter policies | 1 - 12 |
| 72 | Number of lorry policies | 1 - 12 |
| 73 | Number of trailer policies | 1 - 12 |
| 74 | Number of tractor policies | 1 - 12 |
| 75 | Number of agricultural machines policies | 1 - 12 |
| 76 | Number of moped policies | 1 - 12 |
| 77 | Number of life insurances | 1 - 12 |
| 78 | Number of private accident insurance policies | 1 - 12 |
| 79 | Number of family accidents insurance policies | 1 - 12 |
| 80 | Number of disability insurance policies | 1 - 12 |
| 81 | Number of fire policies | 1 - 12 |
| 82 | Number of surfboard policies | 1 - 12 |
| 83 | Number of boat policies | 1 - 12 |
| 84 | Number of bicycle policies | 1 - 12 |
| 85 | Number of property insurance policies | 1 - 12 |
| 86 | Number of social security insurance policies | 1 - 12 |
| 87 | **Number of mobile home policies** | **0 or 1** |

**L0:**Value Label

1 High Income, expensive child

2 Very Important Provincials

3 High status seniors

4 Affluent senior apartments

5 Mixed seniors

6 Career and childcare

7 Dinki's (double income no kids)

8 Middle class families

9 Modern, complete families

10 Stable family

11 Family starters

12 Affluent young families

13 Young all american family

14 Junior cosmopolitan

15 Senior cosmopolitans

16 Students in apartments

17 Fresh masters in the city

18 Single youth

19 Suburban youth

20 Etnically diverse

21 Young urban have-nots

22 Mixed apartment dwellers

23 Young and rising

24 Young, low educated

25 Young seniors in the city

26 Own home elderly

27 Seniors in apartments

28 Residential elderly

29 Porchless seniors: no front yard

30 Religious elderly singles

31 Low income catholics

32 Mixed seniors

33 Lower class large families

34 Large family, employed child

35 Village families

36 Couples with teens 'Married with children'

37 Mixed small town dwellers

38 Traditional families

39 Large religous families

40 Large family farms

41 Mixed rurals

**L1:**

1 20-30 years

2 30-40 years

3 40-50 years

4 50-60 years

5 60-70 years

6 70-80 years

**L2:**

1 Successful hedonists

2 Driven Growers

3 Average Family

4 Career Loners

5 Living well

6 Cruising Seniors

7 Retired and Religeous

8 Family with grown ups

9 Conservative families

10 Farmers

**L3:**

0 0%

1 1 - 10%

2 11 - 23%

3 24 - 36%

4 37 - 49%

5 50 - 62%

6 63 - 75%

7 76 - 88%

8 89 - 99%

9 100%

**L4:**

0 RS 0

1 Rs 1 – 49

2 RS 50 – 99

3 Rs 100 – 199

4 RS 200 – 499

5 RS 500 – 999

6 RS 1000 – 4999

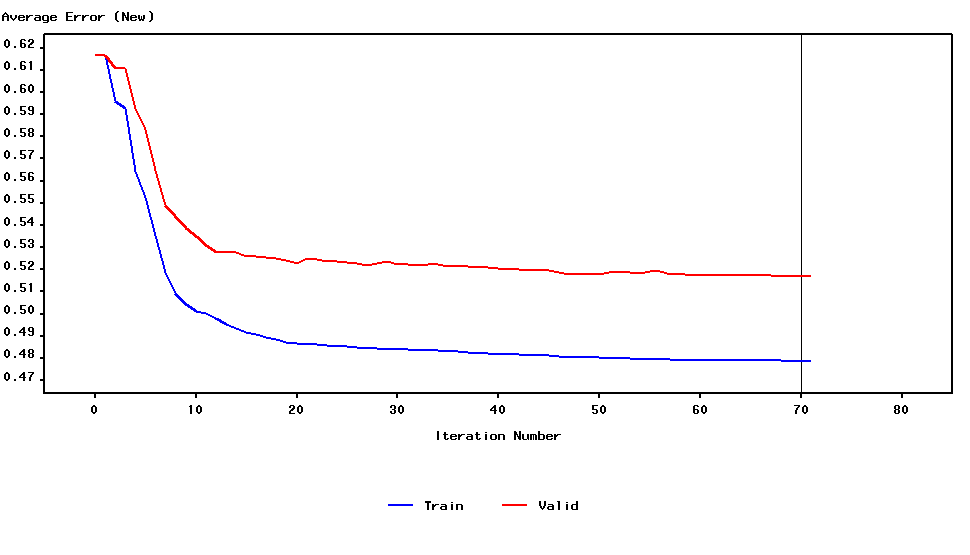
7 RS 5000 – 9999

8 RS 10000 - 19999

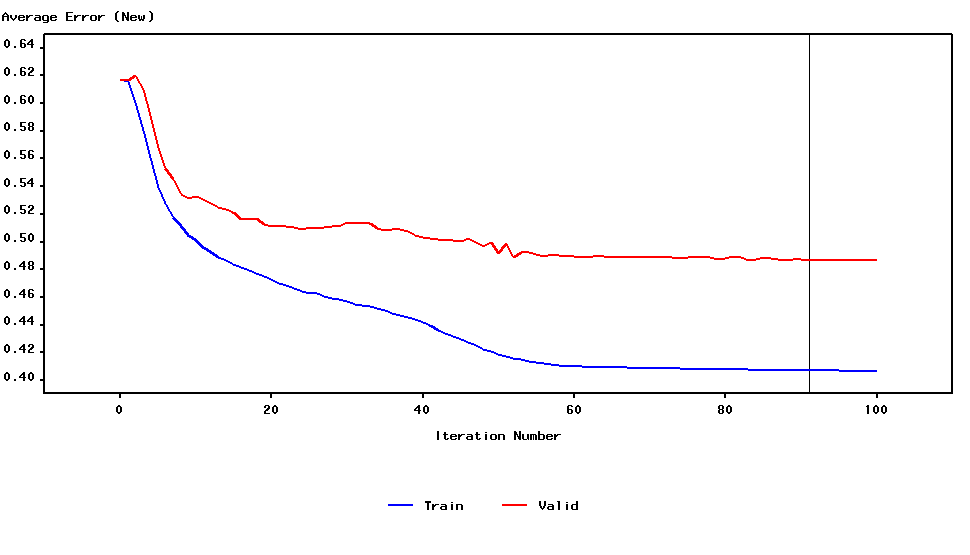
9 RS >20000

**Appendix 2 Neural Network Plots**

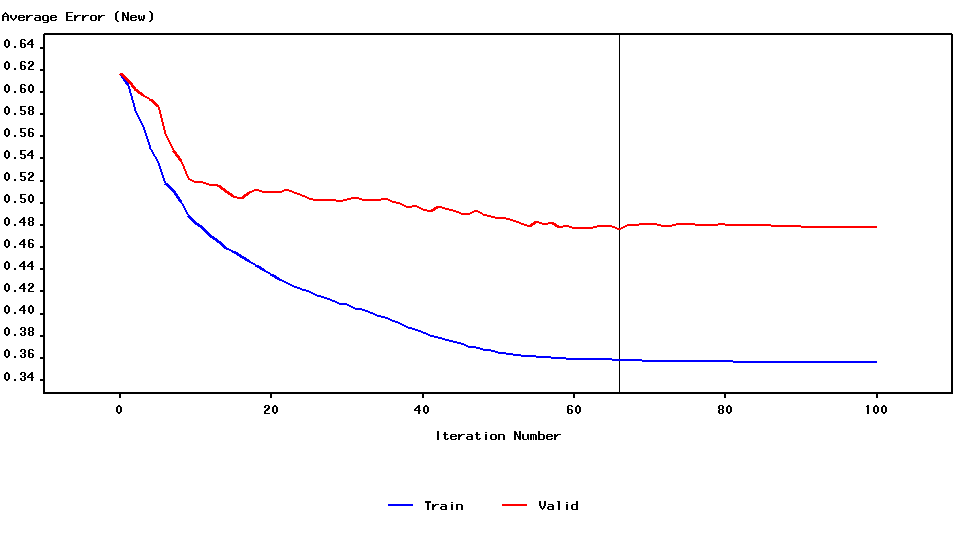
1. **hidden node:**

****

1. **hidden nodes:**

****

1. **hidden nodes:**

****