Insurance Company

.

A. Model Creation and Principal Component in SAS Enterprise Miner

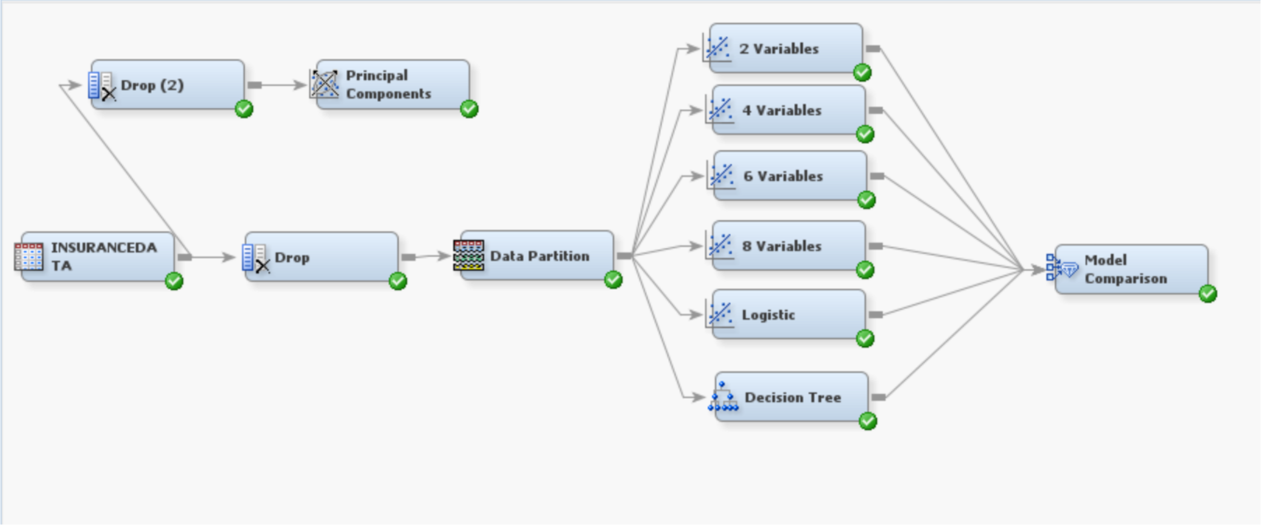
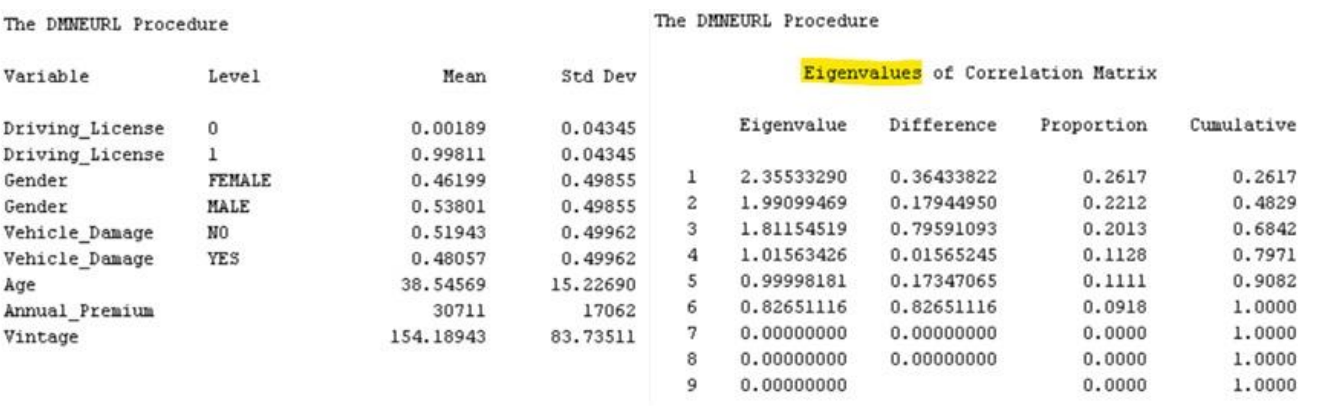


Figure 1. Shows Logistic Regression models in SAS

<u>Principal Components:</u>

A principal component analysis is a weighted linear combination of variables in which the weights are set to account for the greatest degree of variance in the data.



```
The DMNEURL Procedure

Variable            Level                 Mean       Std Dev

Driving_License     0                  0.00189       0.04345
Driving_License     1                  0.99811       0.04345
Gender              FEMALE             0.46199       0.49855
Gender              MALE               0.53801       0.49855
Vehicle_Damage      NO                 0.51943       0.49962
Vehicle_Damage      YES                0.48057       0.49962
Age                                   38.54569      15.22690
Annual_Premium                           30711         17062
Vintage                              154.18943      83.73511
```

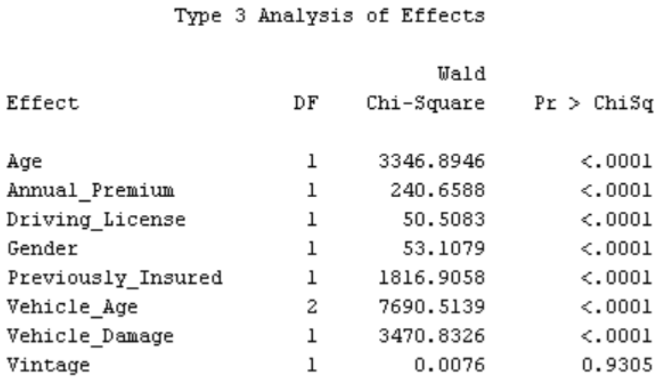
```
The DMNEURL Procedure

              Eigenvalues of Correlation Matrix

     Eigenvalue    Difference    Proportion    Cumulative

1    2.35533290    0.36433822        0.2617        0.2617
2    1.99099469    0.17944950        0.2212        0.4829
3    1.81154519    0.79591093        0.2013        0.6842
4    1.01563426    0.01565245        0.1128        0.7971
5    0.99998181    0.17347065        0.1111        0.9082
6    0.82651116    0.82651116        0.0918        1.0000
7    0.00000000    0.00000000        0.0000        1.0000
8    0.00000000    0.00000000        0.0000        1.0000
9    0.00000000                      0.0000        1.0000
```

Figure 2. Shows PCA: Summary and Eigen Values of Correlation Matrix

In the right side of the figure for Eigen Value for the first Principal Component, it says that 0.2617 Cumulative. The number of variables that were taken into account is the sum of Eigen value. If the

Eigen value of the Principal Components is above 1, they are included in the analysis for further study. In this case, the first four Eigen Values are considered for analysis.

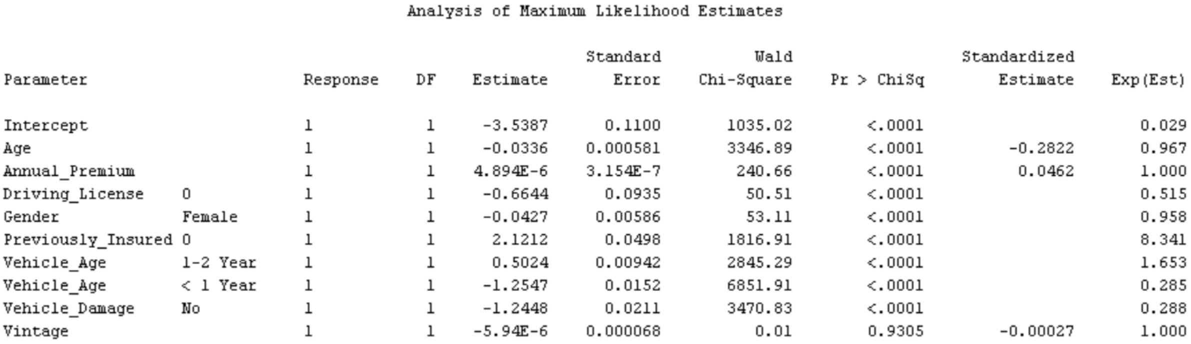
Analysing the Logistic Regression Model with 8 Variables



```
                 Type 3 Analysis of Effects

                                     Wald
Effect                  DF     Chi-Square    Pr > ChiSq

Age                      1      3346.8946        <.0001
Annual_Premium           1       240.6588        <.0001
Driving_License          1        50.5083        <.0001
Gender                   1        53.1079        <.0001
Previously_Insured       1      1816.9058        <.0001
Vehicle_Age              2      7690.5139        <.0001
Vehicle_Damage           1      3470.8326        <.0001
Vintage                  1         0.0076        0.9305
```

Figure 3. Shows Chi-Square test for Logistic Regression model with 8 variables

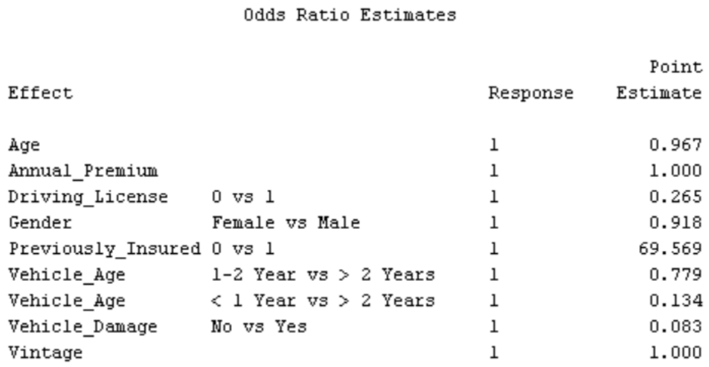
From the figure 3 it can be seen that Age, Annual_Premium, Driving_License, Gender, Previously_Insured, Vehicle_Age and Vehicle_Damage are significant as the value is less than 0.05. But Vintage is not significant as the value is 0.9305 which is above 0.05. It means that Vintage does not offer any explanation or dependency towards the target variable whether customer is interested in vehicle insurance or not.



```
                                        Analysis of Maximum Likelihood Estimates

                                                                Standard          Wald                       Standardized
Parameter                         Response    DF    Estimate       Error    Chi-Square    Pr > ChiSq        Estimate    Exp(Est)

Intercept                         1            1     -3.5387      0.1100       1035.02        <.0001                       0.029
Age                               1            1     -0.0336    0.000581       3346.89        <.0001         -0.2822       0.967
Annual_Premium                    1            1    4.894E-6    3.154E-7        240.66        <.0001          0.0462       1.000
Driving_License    0              1            1     -0.6644      0.0935         50.51        <.0001                       0.515
Gender             Female         1            1     -0.0427     0.00586         53.11        <.0001                       0.958
Previously_Insured 0              1            1      2.1212      0.0498       1816.91        <.0001                       8.341
Vehicle_Age        1-2 Year       1            1      0.5024     0.00942       2845.29        <.0001                       1.653
Vehicle_Age        < 1 Year       1            1     -1.2547      0.0152       6851.91        <.0001                       0.285
Vehicle_Damage     No             1            1     -1.2448      0.0211       3470.83        <.0001                       0.288
Vintage                           1            1    -5.94E-6    0.000068          0.01        0.9305        -0.00027       1.000
```

Figure 4. Shows Maximum Likelihood estimates for Logistic regression with 8 variables

If the variable are classification Yes/No or 1/0 then these values are calculated by taking Yes or 1 as the base value.



```
                              Odds Ratio Estimates

                                                                 Point
Effect                                              Response    Estimate

Age                                                 1              0.967
Annual_Premium                                      1              1.000
Driving_License    0 vs 1                           1              0.265
Gender             Female vs Male                   1              0.918
Previously_Insured 0 vs 1                           1             69.569
Vehicle_Age        1-2 Year vs > 2 Years            1              0.779
Vehicle_Age        < 1 Year vs > 2 Years            1              0.134
Vehicle_Damage     No vs Yes                        1              0.083
Vintage                                             1              1.000
```

Figure 5. Shows the Odds ratio estimates

From figure 5 it can be seen that the odds for customer's interest in vehicle insurance will increase at 0.083, if the Vehicle_Damage = No. It means that customers with Vehicle damage are more likely to get vehicle insurance. Similarly, for Previously_Insured, Driving_licence and Vehicle_Age above 2 years too.

B. Model Comparison

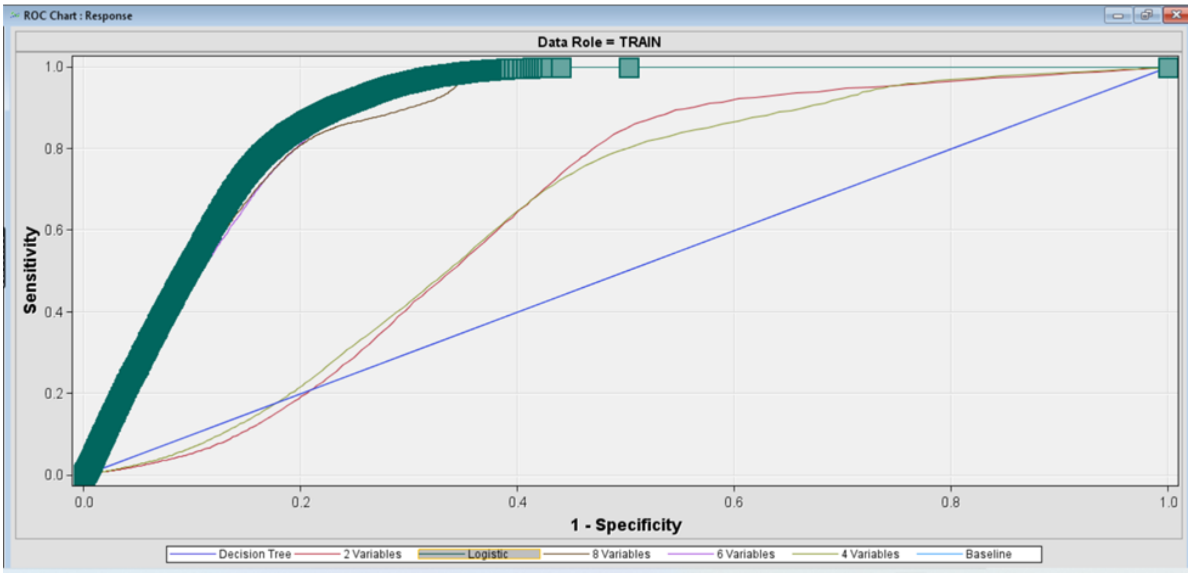
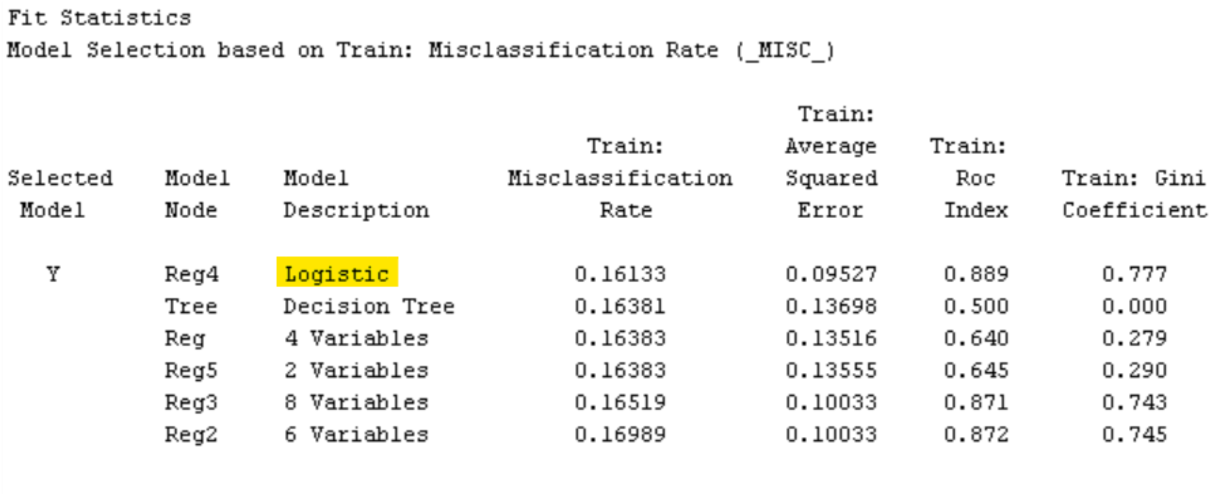


Figure 6. Shows Receiver Operating Characteristic curve

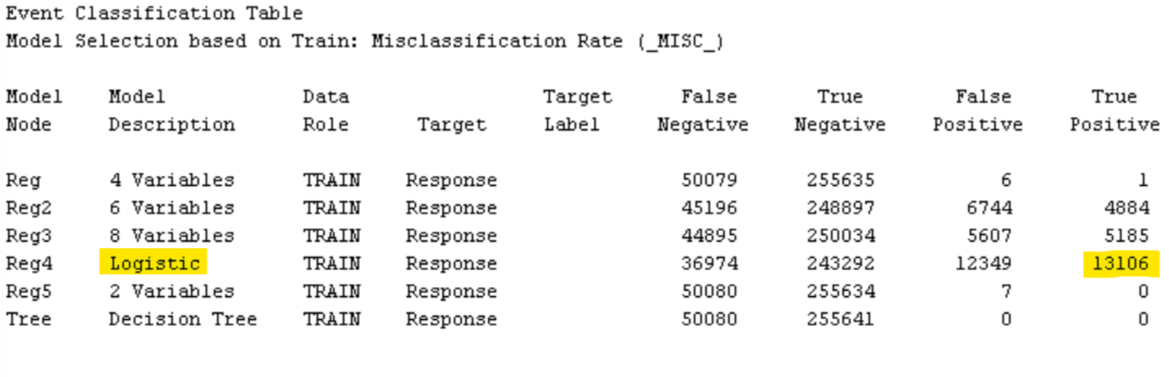
A ROC curve is made by comparing the true positive rate to the false positive rate. This is done by putting the two numbers together. It is the percentage of positive observations that were predicted to be positive that were correct.



```
Fit Statistics
Model Selection based on Train: Misclassification Rate (_MISC_)

                                                       Train:
                                       Train:          Average    Train:
Selected   Model    Model              Misclassification  Squared    Roc      Train: Gini
 Model     Node     Description        Rate            Error      Index    Coefficient

   Y       Reg4     Logistic           0.16133         0.09527    0.889      0.777
           Tree     Decision Tree      0.16381         0.13698    0.500      0.000
           Reg      4 Variables        0.16383         0.13516    0.640      0.279
           Reg5     2 Variables        0.16383         0.13555    0.645      0.290
           Reg3     8 Variables        0.16519         0.10033    0.871      0.743
           Reg2     6 Variables        0.16989         0.10033    0.872      0.745
```

Figure 7. Shows Fit Statistics indicating the best selected model

The best model is been selected on the basis of least misclassification rate. In this case model name ‘Logistic’ is selected because it has the lowest misclassification rate.



```
Event Classification Table
Model Selection based on Train: Misclassification Rate (_MISC_)

Model    Model            Data                  Target    False       True       False      True
Node     Description      Role      Target      Label     Negative    Negative   Positive   Positive

Reg      4 Variables      TRAIN     Response                50079      255635          6          1
Reg2     6 Variables      TRAIN     Response                45196      248897       6744       4884
Reg3     8 Variables      TRAIN     Response                44895      250034       5607       5185
Reg4     Logistic         TRAIN     Response                36974      243292      12349      13106
Reg5     2 Variables      TRAIN     Response                50080      255634          7          0
Tree     Decision Tree    TRAIN     Response                50080      255641          0          0
```

Figure 8. Shows Misclassification Rate for all the selected Models

From figure 8. It can be observed that the maximum True Positive values (13106) are for Reg4 (Logistic) model. Misclassification rate is used to figure out how many observations were wrongly predicted by some kind of classification model. It tells us how many of those observations were misclassified.

C. Conclusion

The best model selected to predict whether a customer is interested in to subscribe a vehicle insurance or not is the Logistic regression model consisting all the variables given in dataset except ID variable. The main group of customers to be targeted are the people who do not have vehicle insurance at present. It can also predicted that customers who do not have a driving licence might not have a car and they wouldn’t be needing any insurance. So, customers with driving licence must only be considered. Vehicle age above 2 years customer are more likely to get there vehicle insured.