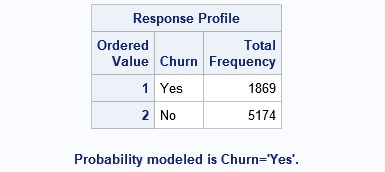
**Shortlisting the variables:**

We calculated the percentage difference between mean values of the churners and the non-churners. We checked for the correlation between the variables. There are no missing values in the dataset for the selected variables and there is no multicollinearity in the model.

The 10 variables which we have selected are:

SeniorCitizen OnlineSecurity TechSupport Dependents MonthlyCharges PaperlessBilling Contract PaymentMethod PhoneService tenure.

1. **Use SAS to develop a good logistic regression model to predict which customers are likely to leave (churn)?**



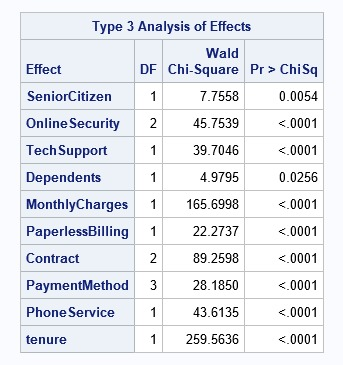
According to the given data, 1869 Customers churned in the last month which is about 26.53% of the total number of customers.

We developed a logistic regression model based on the probability that a customer will churn.

Log(Churn) is the dependent variable.

SeniorCitizen, OnlineSecurity, TechSupport, Dependents, MonthlyCharges, PaperlessBilling, Contract, PaymentMethod, PhoneService and tenure are the independent variables.

1. **Include a table of coefficients, t-values, and odds ratio. Interpret the logistic output explaining AIC/BIC, meaning of coefficients, significance, prediction accuracy (percent concordance), odds-ratios etc.**



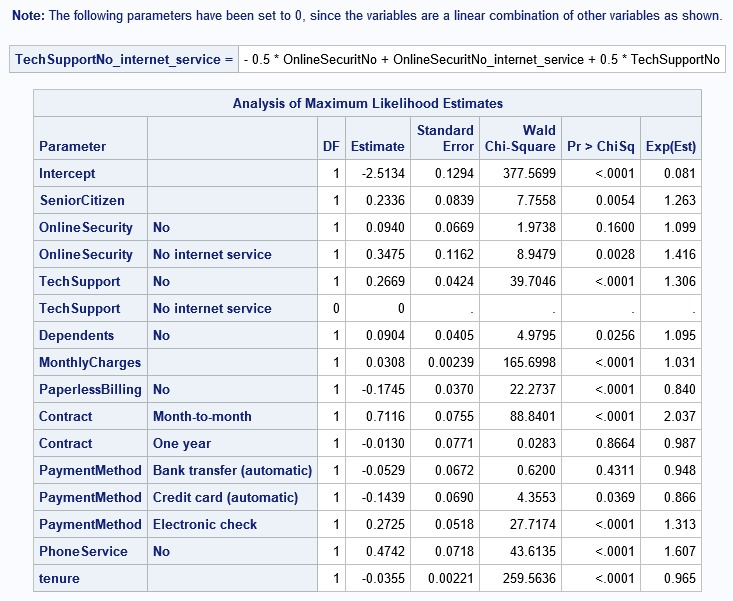
**Inference:**

The above table shows the hypothesis tests for each variable in the model. The chi-square test statistic and associated p-values shown in the table indicate that every variable in the model is significant.

For variables SeniorCitizen, TechSupport, Dependents, MonthlCherges, PaperBilling, Phone Service and tenure the test duplicates the test of the coefficients. However, for categorical variables OnlineSecurity, Contract and PaymentMethod, the table gives the multiple degree of freedom test for the overall effect of the variable.



This table displays the design variables used in the analysis by the logit model for the categorial variables.



The above table shows the coefficients (labeled Estimate), their standard errors (error), the Wald Chi-Square statistic, and associated p-values.

The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable

Since the TechSupport- No\_Internet\_Service category is a linear combination of other variables, its coefficient is reduced to zero. Hence, we ignored it in our analysis by considering it as equal to the No category.

**Inference:**

In the above table, the estimates of each variable is the utility value of the corresponding variable or the category.

The coefficients for SeniorCitizen, TechSupport, Dependents, MonthlyCherges, PaperBilling, PhoneService and tenure are statistically significant at 95% confidence interval.

For Class Variables OnlineSecurity = No internet service (versus the omitted category OnlineSecurity = Yes) is significant but OnlineSecurity = No (versus the omitted category OnlineSecurity = Yes) is not statistically significant at 95% confidence interval.

For Class variable Contract = Month-to- Month (versus the omitted category Contract = Two Years) is significant but Contract = One Year (versus the omitted category Contract = Two Years) is not statistically significant at 95% confidence interval.

For Class Variable PaymentMethod = Elctronic Check and Payment Method = Credit Card  (versus the omitted category PayemetMethod = Mailed Check) is statistically significant but PaymentMethod = Bank Transfer (versus the omitted category PayemetMethod = Mailed Check) is not statistically significant at 95% confidence interval.



From the above table, the point estimate is the odds ratio for each variable. We can calculate the percentage increase in odds considering that

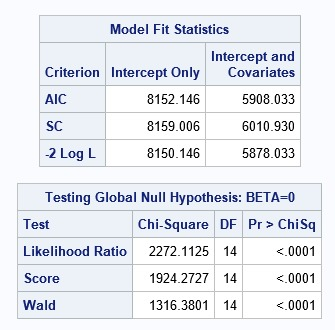
Percentage increase in odds = (point estimate – 1) \* 100

* If the customer is a **senior citizen**, the odds of Customer Churn **increases by 26.3%** compared to the customer not being a senior citizen, holding everything else constant.
* If the customer did not sign up for **Tech Support**, the odds of Customer Churn **increases by 70.5%** compared to the customer signing up for Tech Support, holding everything else constant.
* If the customer has **no Dependents**, the odds of Customer Churn **increases by 19.8%** against having dependents, holding everything else constant.
* For every **$1 increase in the Monthly Charges**, the odds of Customer Churn **increase by 3.1%**.
* If the customer **did not sign up for Paper billing**, the odds of Customer Churn **decreases by 29.5%** against signing up for Paper Billing, holding everything else constant.
* If the customer **did not sign up for Phone Service**, the odds of Customer Churn **increases by 158.1%** against signing up for Paper Billing, holding everything else constant.
* For every **1 unit increase of Tenure**, the odds of Customer Churn **decrease by 3.5%**, holding everything else constant.
* If the customer **didn’t sign up for the Online Support service**, the odds of Customer Churn **increases by 70.8%** against signing up for the Online Support Service.
* If the customer **didn’t sign up for the internet service**, the odds of Customer Churn **increases by 120.1%** against signing up for the Online Support Service.
* If the customer signed up for the **month-to-month contract**, the odds of Customer Churn **increases by 309.7%** against signing up for a two-year contract.
* If the customer signed up for **one-year contract**, the odds of Customer Churn **increases by 98.5%** against signing up for a two-year contract.
* If the customer signed up for the **payment method through bank transfer**, the odds of customer churn **increases by 2.3%** against signing up for the mailed check service.
* If the customer signed up for the **payment method through credit card**, the odds of customer churn **decreases by 6.6%** against signing up for the mailed check service.
* If the customer signed up for the **payment method through electronic check**, the odds of customer churn **increases by 41.6%** against signing up for the mailed check service.



**Inference:**

**Percent Concordant:** 84.50 % of pairs where the observation with the customer churning has a higher predicted probability than the customer not churning.



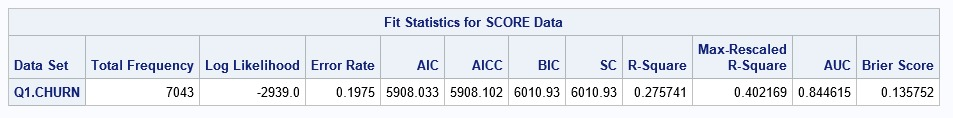
**Inferences:**

The table “Model Fit Statistics” describes and tests the overall fit of the model.

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) is used for the comparison of non-nested models on the same sample. Ultimately, the model with the smallest AIC is considered the best. The column “Intercept Only” refers to the respective criterion statistics for the null model. The column “Intercept and Covariates” corresponds to the respective criterion statistics for the fitted model which includes all independent variables and the intercept. When compared with the criteria corresponding Intercept Only value to assess model fit/significance

The -2 Log L (8150.146) can be used in comparisons of nested models

The Likelihood ratio chi-square of 2272.1125 with a p-value of less than 0.0001 tells us that our model as a whole, fits significantly better than an empty model. The Score and Wald tests are asymptotically equivalent tests of the same hypothesis tested by the likelihood ratio test, these tests also indicate that the model is statistically significant.



1. **Which are the top three factors that affect churn in your model.**

From the analysis in question 2, we calculated the percentage increase in odds of customer churning from the odds ratio values. By comparing the percentage increase in odds, we can infer that

Contract, Internet Service and Phone Service are the top 3 factors that affect the churn in the model.

1. **What other variables (that if collected) would help to improve the fit of the model.**

The other variables that can be used to improve the model fit are:

a) Frequency: How frequent the customer is using the Tech support services i.e. the number of interactions with the tech support department. The greater number of interactions can reflect that the customers interaction with the product or the service is more. And thus, the chances of the customer churning is less.

b) Topics of questions asked: Can used to identify how intensively the customer is using the services. Which can in turn be used to predict whether that customer will churn or not. If the questions are regarding the service feature then the possibility of customer retention are higher with respect to the questions related to the package cost.

c) Satisfaction ratings: The rating given by a customer after he/she talks to the customer support person can be used to identify the level of satisfaction of the customer.

d) The due amount payable by the customer can be used to identify the possibility of the customer churn. The higher the amount due the more possibility of customer churn.

e) Usage: Usage history can be used to identify how often customer uses our product or services if the usage is less then there are chances that the customer might switch to a cheaper service provider or cheaper product.

1. **Compute the hit ratio for your model. Hit ratio is defined as the percentage of correct predictions using the logit model. Use the model to predict 1 or 0 using the same data.**



**Inference:**

Using the logit model, we predicted the customers who are going to churn by considering

If prediction < 0.5 then churn =0 and if prediction > 0.5 then churn =1.

We compared these predictions to the actual values and calculated the hit ratio which is 80.25%.