# Predictive Modelling II – Logistic Regression

**Scenario:**

ABC Telecom is one of the many telco in the country. With the liberalization of telco service providers in the market and the competitive nature of the business, ABC is facing the problem of customer churn.

Customer churn is described as the situation where customers stop using the company’s products or services, choosing to take up the products or services offered by the company’s competitors instead.

Like many other telco, ABC offers more than mobile phone service, it also offers Internet and streaming TV services. Hence, there is no shortage of data about customers, and ABC Telecom hopes that these data can help it predict and prevent future customer churn.

## Theory

**Learning Activity 1 – How does logistic regression work**

State whether the following statement is true or false.

|  |  |
| --- | --- |
| **Statement** | **True / False** |
| Linear Regression is not suitable for scenario where target is binary value (e.g. 0, 1) | True |
| Logistic Regression does not use the Y = mx + c formula | False |
| Categorical inputs needs to be prepared first before we can use them for Logistic Regression | True |
| Logistic Regression use the Least Square Criterion to find the best fit line | False |

**Takeaway**:

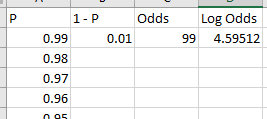
|  |
| --- |
| Linear Regression and Logistic Regression are similar in many ways.  Both works only with numeric inputs. We can transform the categorical inputs to numeric.  Both make use of the y = mx + c formula to learn from historical data.  Logistic Regression use a function to transform the predicted outcome to be between 0 and 1.  Logistic Regression uses **Maximum Likelihood Estimate** to find the best S-shape curve |

**Learning Activity 2 – Understanding Log Odds**

1. Open the Excel file “Understanding Log Odd.xlsx”

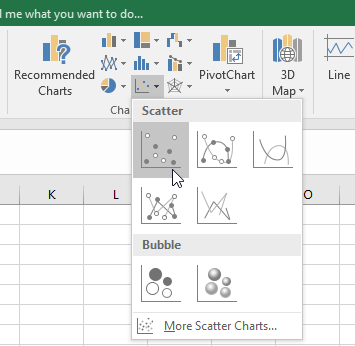
This file contains 4 columns

* Column 1: Probability of event, p
* Column 2: Probability of non-event (event not happening)
* Column 3: Odds
* Column 4: Natural log of the odds



1. Copy the formula for columns 2 – 4, paste it into cells B3 to D100.
2. Produce a scatter diagram of P vs Log Odds.

Select column 1 and column 4, click on Insert > Chart > Scatter



Notice the shape of the resulting graph.

* X-axis = P, Y-axis = Log of Odds
* Shape of graph is S-shape, with asymptotes at X = 0 and X = 1.
* Graph cuts X-axis at 0.5

**Takeaway**:

|  |
| --- |
| Natural Log of Odds gives us a function that ‘almost’ serves our purpose.  We have a regression formula that ranges from 0 to 1 for X. |

**Learning Activity 3 – Inversing the Log Odds**

1. Go to <https://www.wolframalpha.com/>
2. In the search bar, enter “inverse of log of (p/(1-p))



The inverse of Log Odds is

**Takeaway**:

|  |
| --- |
| The inverse of Log Odds gives us the S-shape curve that we need for Logistic Regression.  Y ranges from 0 to 1 with asymptotes at Y = 0 and Y = 1.  Regression formula is used to estimate probably of event, p |

**Learning Activity 4 – Measuring Accuracy of Logistic Regression**

1. We trained a model using logistic regression and obtained the performance shown in the coincidence matrix below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Churn  ‘Yes’ | Predicted Churn  ‘No’ |  |
| Actual Churn  ‘Yes’ | 85 | 30 | **115** |
| Actual Churn  ‘No’ | 25 | 70 | **95** |
|  | **110** | **100** | **210** |

|  |  |
| --- | --- |
| Calculate the misclassification rate: |  |
| Calculate the precision: |  |
| Calculate the accuracy: |  |

**Takeaway**:

|  |
| --- |
| There are various metrics which can be used to determine model performance.  There is no ‘best’ metric, it depends on the context of the analysis. |

## Hands-on

About the data:

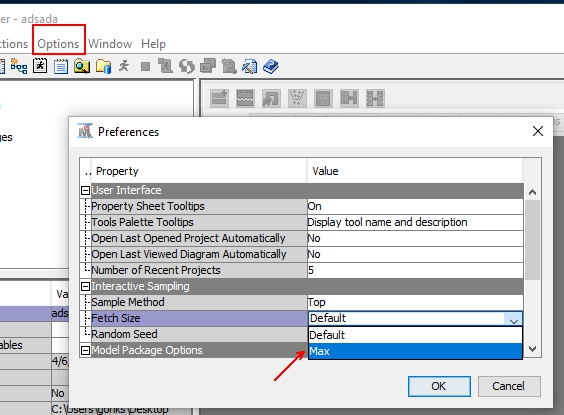
We are using a fictitious dataset containing customers of ABC Telco. This dataset was made available by IBM Watson, some modifications were made for the purpose of this practical. The dataset consists of over 3000 rows (each row represents a customer) and 33 columns.

Information about customers, such as their gender, marital status, whether or not they have dependents and whether they are senior citizens were captured.

In additional, information such as the tenure (how long customer has been with ABC Telco), whether they subscribe Internet and/or Streaming TV services in addition to the mobile phone service. Their call minutes

1. We will create a new diagram (not project).

* Log in to the virtual machine, click on SAS Enterprise Miner Workstation 14.1
* Copy and paste your project folder from the previous practical into the virtual machine.
* Select Open Project, browse to your pasted project folder and click Open.
* We will make a modification to the project preferences as our dataset for this practical is larger. Go to Options and change the Fetch Size for Interactive Sampling to Max.



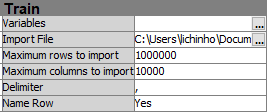
1. Create a new SAS EM Diagram and read the file *Telco\_Churn.csv* into SAS EM using the File Import Node

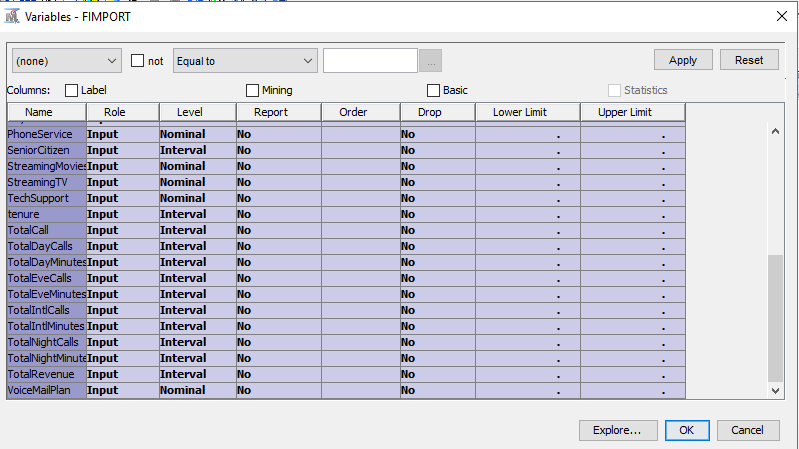
* File > New Diagram (name it *Logistic Regression*)
* Add File Import node (found in Sample group) into your diagram
  + Delimiter = comma (,)
  + Name Row = Yes
  + Table Role = Train
* Click on the ellipsis on Train > Variables and set the Data role / level of measurement as shown below:

|  |  |  |
| --- | --- | --- |
| **Data Field** | **Role** | **Level of Measurement** |
| customerID | ID | Nominal |
| Churn | Target | Binary |
| All other fields | Input | Leave as default |

1. Run the File Import node.

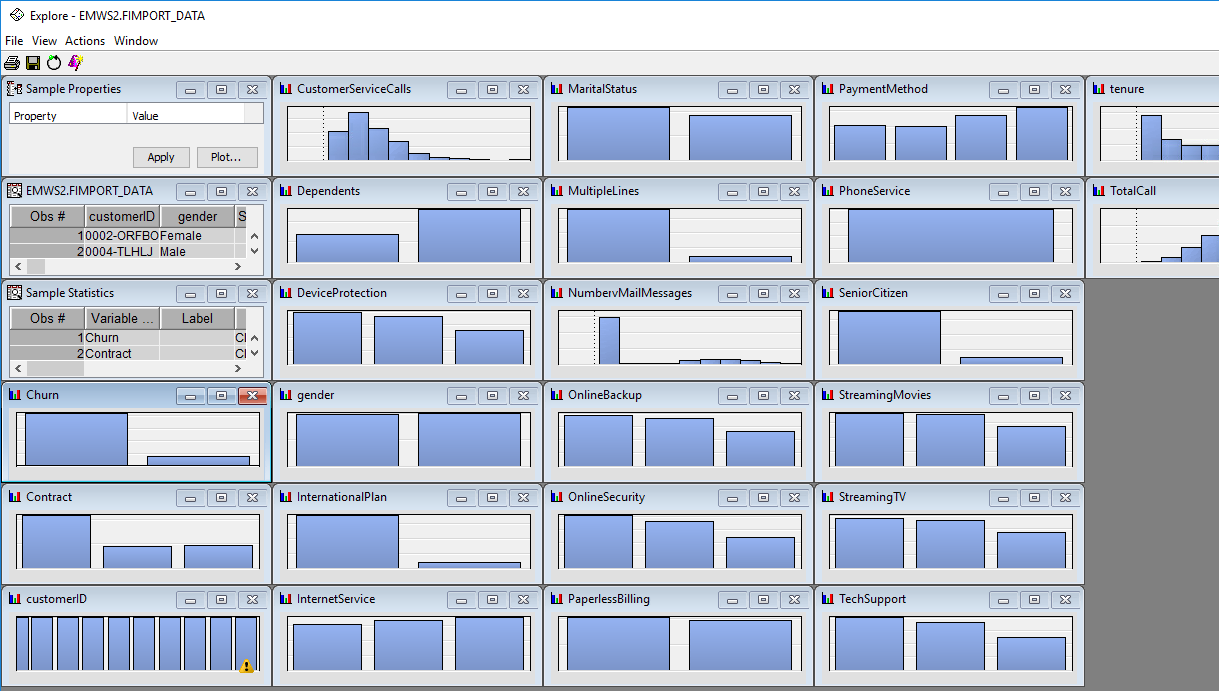
Explore the imported data by going to the properties pane, locating the Train group and clicking on the ellipsis next to it.





Select ALL

In the resulting window, hold down the shift key to select all fields and click on the Explore button.

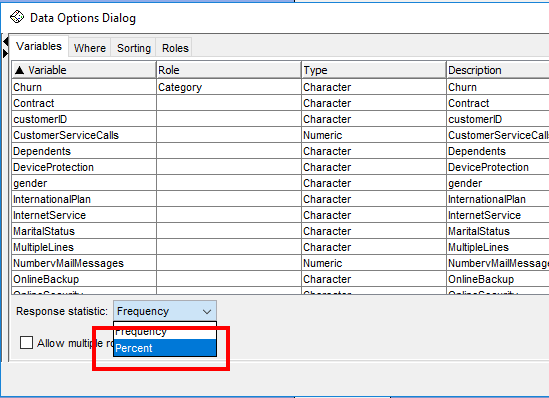


Locate the histogram for the Churn field and maximize the window.

By default, the bar is showing us the raw frequency.

Right-click on either of the bar, select Data Option

In the Data Option Dialog window, change the Response statistic to Percent.



What percentage of customer have Churn = Yes?

|  |
| --- |
| 14.5% |

1. Add a Replacement node (found in Modify group) downstream from the File Import node.

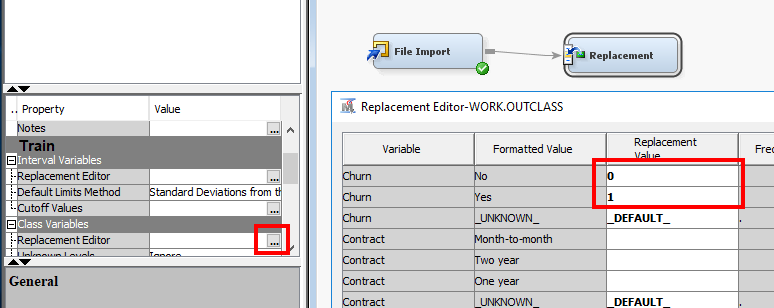
We will see how we can use it to recode the values in the Churn field from “Yes”, “No” to 1 and 0.

In the Property pane, under the **Class** Variables, look for Replacement Editor. Note the SAS EM refers to “categorical variables” as “class variables”.

*Click on the ellipsis. In the resulting window, look* for the variable Churn.

* Set replacement value for No to be 0
* Set replacement value for Yes to be 1

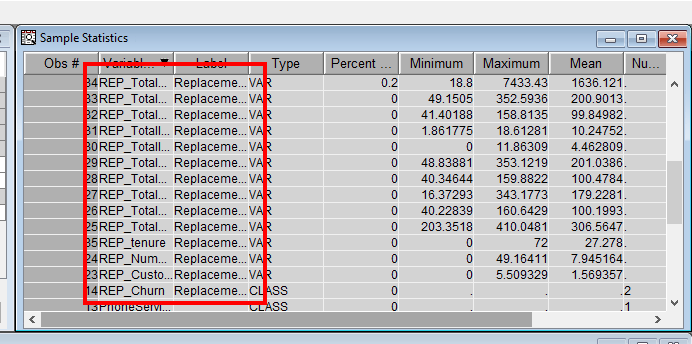
Leave the settings for Interval Variables unchanged.



**Execute the Replacement node.**

1. Still at the properties pane, search for **Exported Data** in the general group.

Explore the result of the Replacement node by clicking the ellipsis, selecting the TRAIN port and clicking on the Explore button.



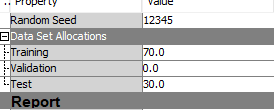
The Replacement node will produce several new fields, all with the prefix REP\_

In this instance, all the interval (numeric) fields were adjusted to fit within ±3 standard deviation from their mean values, this is an easy way to take care of outlying data points.

The selected class variable (in this case, Churn) is recoded based on our settings (Yes became 1 and No became 0)

1. Connect a **Data Partition** node (found in Sample group) downstream from the **Replacement** node.

In the properties pane, under the Data Set Allocations group, set Training to be 70 (70%) and Test to be 30 (30%).



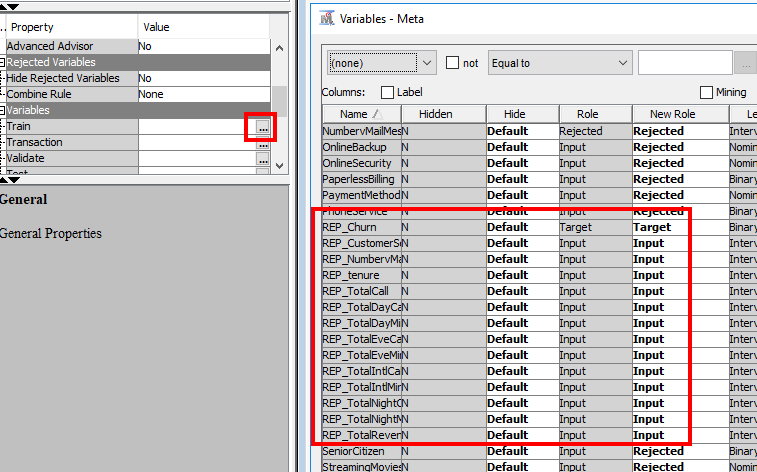
1. Connect a **Metadata** node (found in the Utilities group) downstream from the **Partition** node.

The **Metadata** node allows us to adjust details about our data fields such as their roles, levels etc.

In the properties pane for the **Metadata** node, look for the Variables group, click on the ellipsis next to Train.

In the resulting window, set the New Role as the following:

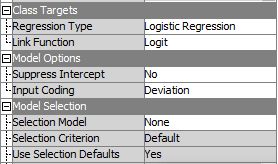
* REP\_Churn : Target
* All other fields beginning with REP\_ : Input
* All other fields : Rejected



Run the **Metadata** node.

1. Connect a **Regression** node (found in Model group) downstream from the **Metadata** node.

In the properties pane, ensure that Logistic Regression as the Regression Type.



Run the **Regression** node and view the results.

Since we are training the model to predict a class (0 or 1) instead of predicting a continuous value, statistics like RMSE will not make sense.

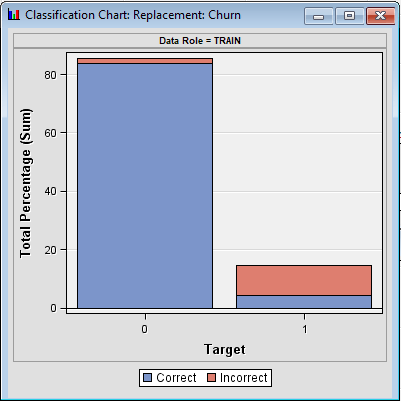
Instead, we should look at the **Misclassification Rate.**

Another useful way to look at the model performance is the Classification Chart.

Still in the Results window, click on View > Assessment > Classification Chart

You should get a chart similar to the one shown below.

The red portion is the percentage that was misclassified.

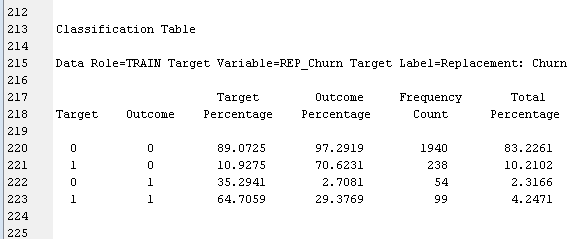


1. Expand the Output pane in the Results window.

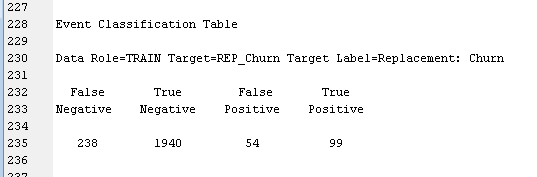
Scroll to row 210. You should see a section called **Classification Table**.

This shows you the equivalent of the confusion matrix.

Target = Actual, Outcome = Predicted



Scrolling down a little more, you should be able to find the count for False Negative, True Negative etc.



Fill in the FN, TN, FP and TP into the table below

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted Churn  ‘1’ (Yes) | Predicted Churn  ‘0’ (No) |  |
| Actual Churn  ‘1’ (Yes) | 99 | 238 | **337** |
| Actual Churn  ‘0’ (No) | 54 | 1940 | **1994** |
|  | **153** | **2178** | **2331** |

|  |  |
| --- | --- |
| Calculate the misclassification rate: |  |
| Calculate the precision: | = 64.7% |
| Calculate the accuracy: | 2039/2331 = 87.5% |
| Calculate Recall/Sensitivity | 99/337 = 87.4% |

Do you think the predictive model that we trained is performing well?

|  |
| --- |
| Given that ABC Telco is trying to predict customer churn, we should focus our attention on model predicting churn=1  We can see that among the cases where Churn = 1 (Target = 1), model wrongly predicts 70.6% of them to be Outcome = 0.    See Table, this is the FN / (TP+FN) = 238/337 = 70.6.  In other words, we can say the model is not sensitive to predict the correct outcome (Recall is low). The model performance is not very good. |

1. Still at the Results window for the **Regression** node, take a look at the Effects plot.

Look for the following fields and state whether they have positive or negative influence on the target field. Write down the interpretation of the sign in the context of our problem.

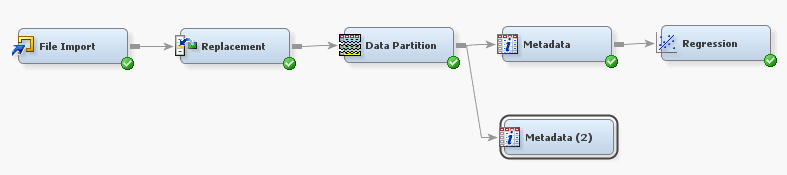
|  |  |  |
| --- | --- | --- |
| **Effect** | **Sign of influence (+ or -)** | **Interpretation** |
| REP\_CustomerServiceCalls | + | The more often customer calls customer service |
| REP\_tenure | - | The longer the tenure, |

In your opinion, does the sign of influence make sense?

|  |
| --- |
| Probably.  If we assume that customers only call customer service when they are having problems (e.g. connection, bill error etc). A customer that had to call customer service often is likely to be upset with the service he is receiving.  It also makes sense that the longer a customer is with the telco, the less likely he is to churn, customer is already familiar with the telco and may have some tied up services, payment arrangements. |

1. Connect a new **Metadata** node from the **Partition** node.

Your diagram should resemble the one shown below.



This time, we will use the following settings for the New Role:

* REP\_Churn : Target
* All other fields beginning with REP\_ : Input
* PaymentMethod : Input
* All other fields : Rejected

Connect a **Regression** node downstream from this **Metadata** node.

Ensure that the regression method is Logistic, run the **Regression** node.

Expand the Output window and scroll to line 150.

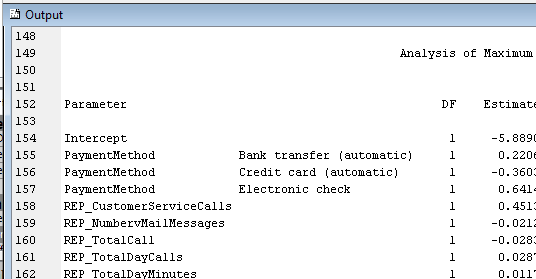
You can see that SAS created the dummy variable for PaymentMethod automatically.

Our original data comprises 4 payment methods:

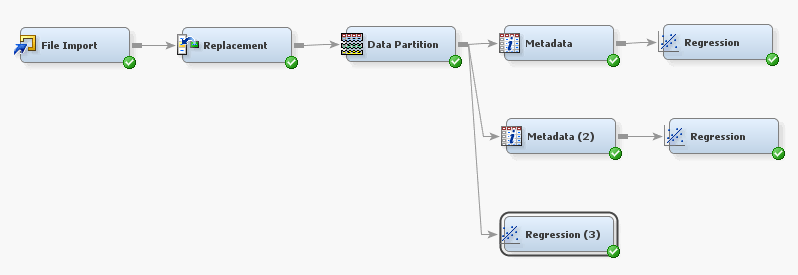
* Mailed check
* Electronic check
* Credit card (automatic)
* Bank transfer (automatic)

Three dummy fields were created:

* PaymentMethod Bank transfer (automatic)
* PaymentMethod Credit card (automatic)
* PaymentMethod Electronic check



1. Add another **Regression** node into your diagram (your third one). Connect the **Partition** node to it. Your diagram should resemble the one shown below.



Ensure that the regression method is set to Logistic.

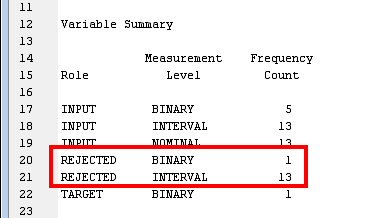
Set the Model Selection mode to be Backward.

Run the **Regression** node.

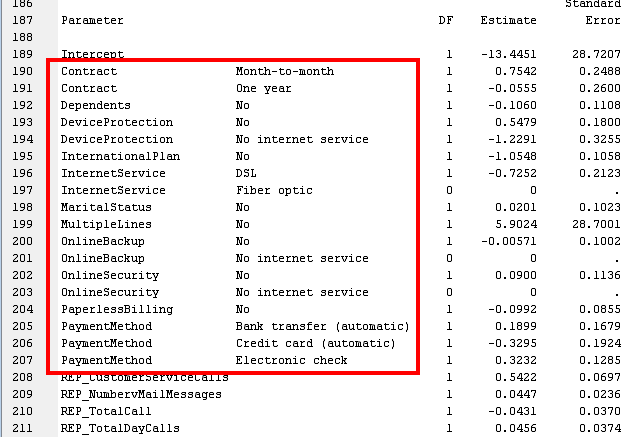
Expand the Output panel in the Results window.

You can see that the **Regression** node automated many of the operations we carried out earlier using the **Metadata** node.

It automatically recognized the output from the **Replacement** node and rejected the original fields.



It also created the ‘dummy’ variables for nominal fields.



1. Add a **Model Comparison** node to your diagram, connect all 3 **Regression** nodes to it.

Leave all settings as default, run the node and determine the champion model.

1. Use **File Import** node, read the file “Telco\_Churn Score.csv” into your diagram.

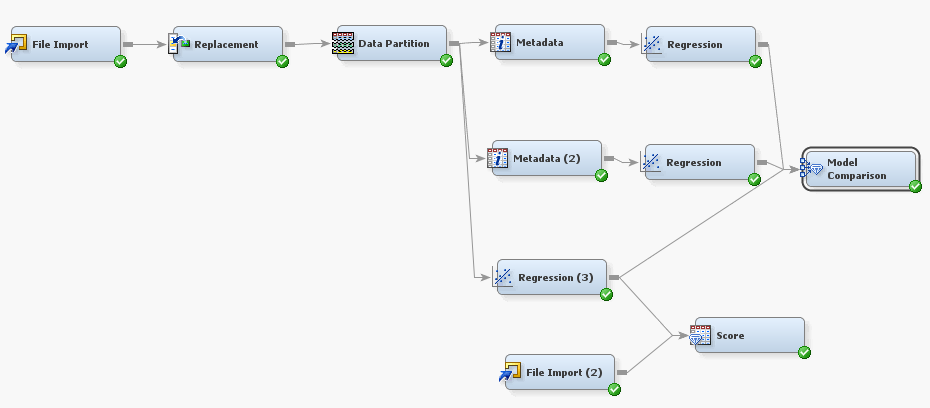
Set the data role as **Score**.

This file contains only 10 rows, each representing a current customer with ABC Telco.

Add a **Score** node (found in Assess group) to your diagram.

Connect the champion model and the **File Import** node to the **Score** node.

Your diagram should resemble the one shown below.



Run the **Score** node.

1. Once done, in the properties pane, look at the Exported Data under the General group.

Click on the ellipsis and then select the Score port, click Browse.

The right most column in the resulting window shows the predicted values for customer churn.

## Interpretation and Application

What actions would you recommend that ABC Telco take, in order to retain customers who are predicted to be likely to churn?

|  |
| --- |
| Answer:  Offer incentive for longer contracts  Solicit feedback, take action, update on outcome  Engage assign best people to communicate with group  Tailor suitable plans based on the group’s subscription. |

False Positive and False Negative both describe the scenario where model wrongly predicts the outcome. In your opinion, which metric do you think ABC Telco should pay more attention to in this context?

|  |
| --- |
| Answer: |

## Summary

1. Logistic regression allows us to build a predictive model for categorical outcomes
2. Logistic regression is still based on the Y = mx + c equation, albeit indirectly.
3. Logistic regression uses the Maximum Likelihood Estimation to find the best fit ‘line’.
4. Data preparation is required when we use categorical predictors. One way is to recode values into numbers, another way is to create dummy variables (one-hot encoding).

## Resources for further reading

1. StatQuest: Logistic Regression

[A YouTube video introducing how logistic regression works – 8 minutes]

<https://www.youtube.com/watch?v=yIYKR4sgzI8>

1. Logistic Regression Details Pt1: Coefficients

[A YouTube video explaining the math behind logistic regression – 22 minutes]

<https://www.youtube.com/watch?v=vN5cNN2-HWE>

You can continue on the other videos in the playlist if you are keen

1. Choosing the right Encoding method-Label vs OneHot Encoder

[Article explaining the difference between the two different strategies employed to prepare categorical data: Label recoding vs One-Hot encoding]

<https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b>

1. Customer churn: 12 ways to stop churn immediately

[Blog post on some tips to stop customer churn]

<https://www.superoffice.com/blog/reduce-customer-churn/>