



11Ants AnalyticsTM

WHITE PAPER

11ANTS CUSTOMER CHURN ANALYZER OUTPERFORMS 85% OF SUBMISSIONS IN INTERNATIONAL PREDICTIVE ANALYTICS CONTEST

INTRODUCTION

11Ants Analytics Customer Churn Analyzer beat 85% of submissions in the world's most prestigious international data mining contest – with only 55 minutes of human work involved. This white paper has been written to enable the experiment to be repeated by others, and to enable companies evaluating various predictive analytics solutions to benchmark performance of other solutions both in terms of time taken and efficacy of models.

To our knowledge there is no other solution that can build superior models to 11Ants Customer Churn Analyzer with so little human effort required. This applies whether you are an experienced data scientist or a relative novice to the discipline and offers material time and quality advantages to both constituencies.

WHY THIS IS IMPORTANT

Too often automated model building solutions are dismissed as unsuitable for experts, or compromised. While this may have been true of previous generation technologies, 11Ants predictive analytics technologies are capable of routinely beating humans at complex predictive analytics problems, and – equally importantly – with a fraction of the time and energy expended. This white paper has been prepared to demonstrate the speed and efficacy of models on a tangible real world problem.

THE COMPETITION

[KDD Cup](#) is the annual Data Mining and Knowledge Discovery competition organized by [ACM Special Interest Group on Knowledge Discovery and Data Mining](#), the leading professional organization of data miners.

The [KDD Cup 2009](#) focused on customer relationship predictions. The European telecommunications ORANGE provided customer data for analysis. The data set consisted of 50,000 customer records with 216 variables per customer. The tasks were to build three separate propensity models for each customer record as outlined below in the task description. The models were:

- 1) Churn Propensity
- 2) Appetency Propensity
- 3) Up-sell Propensity

TASK DESCRIPTION

The competition task was to estimate the churn, appetency and up-selling probability of customers, hence there are three target values to be predicted.

- **Churn** (wikipedia definition): Churn rate is also sometimes called attrition rate. It is one of two primary factors that determine the steady-state level of customers a business will support. In its broadest sense, churn rate is a measure of the number of individuals or items moving into or out of a collection over a specific period of time. The term is used in many contexts, but is most widely applied in business with respect to a contractual customer base. For instance, it is an important factor for any business

with a subscriber-based service model, including mobile telephone networks and pay TV operators. The term is also used to refer to participant turnover in peer-to-peer networks.

- **Appetency:** In our context, the appetency is the propensity to buy a service or a product.
- **Up-selling** (wikipedia definition): Up-selling is a sales technique whereby a salesman attempts to have the customer purchase more expensive items, upgrades, or other add-ons in an attempt to make a more profitable sale. Up-selling usually involves marketing more profitable services or products, but up-selling can also be simply exposing the customer to other options he or she may not have considered previously. Up-selling can imply selling something additional, or selling something that is more profitable or otherwise preferable for the seller instead of the original sale.

EVALUATION

The performances are evaluated according to the arithmetic mean of the AUC for the three tasks (churn, appetency. and up-selling).

Sensitivity and specificity

The main objective of the challenge is to make good predictions of the target variables. The prediction of each target variable is thought of as a separate classification problem. The results of classification, obtained by thresholding the prediction score, may be represented in a confusion matrix, where tp (true positive), fn (false negative), tn (true negative) and fp (false positive) represent the number of examples falling into each possible outcome:

		Prediction	
		Class +1	Class -1
Truth	Class +1	tp	fn
	Class -1	fp	tn

Any sort of numeric prediction score is allowed, larger numerical values indicating higher confidence in positive class membership.

We define the sensitivity (also called true positive rate or hit rate) and the specificity (true negative rate) as:

Sensitivity = tp/pos

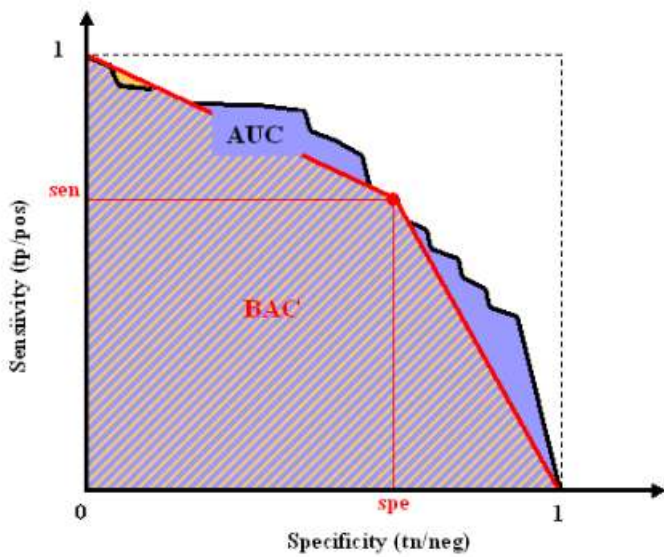
Specificity = tn/neg

where $pos=tp+fn$ is the total number of positive examples and $neg=tn+fp$ the total number of negative examples.

AUC

The results will be evaluated with the so-called Area Under Curve (AUC). It corresponds to the area under the curve obtained by plotting sensitivity against specificity by varying a threshold on the prediction values to determine the classification result. The AUC is related to the area under the lift curve and the Gini index used in marketing ($Gini=2 AUC -1$). The AUC is calculated using the trapezoid method. In the case when binary

scores are supplied for the classification instead of discriminant values, the curve is given by $\{(0,1), (tn/(tn+fp), tp/(tp+fn)), (1,0)\}$ and the AUC is just the Balanced Accuracy BAC.



More details can be found at:

<http://www.kddcup-orange.com/evaluation.php>

http://en.wikipedia.org/wiki/Receiver_operating_characteristic

An introduction to ROC analysis, <http://www.sciencedirect.com/science/article/pii/S016786550500303X>

THE PROCESS

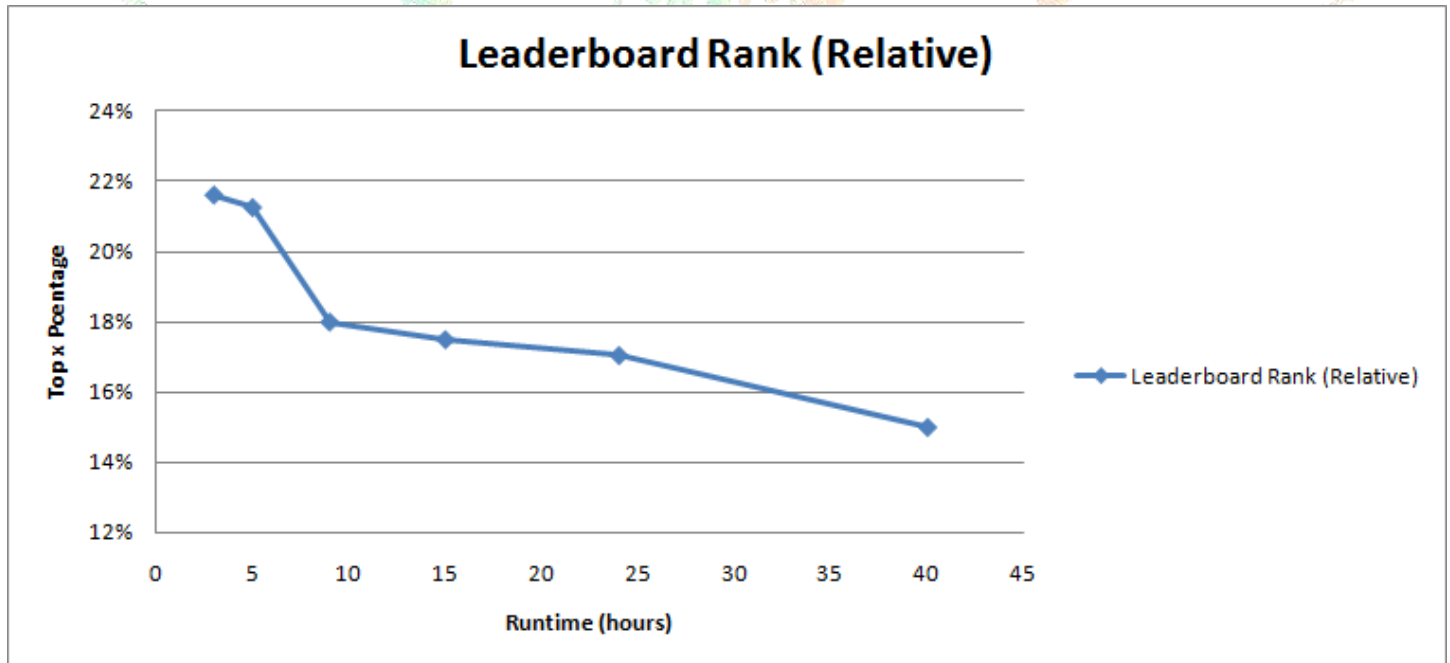
A detailed description of the process for those wishing to repeat the experiment can be found in Appendix A. This provides an overview of the process. Entire time committed from start to finish 35 minutes.

1. Download data from contest site.
2. Open data in Microsoft Excel.
3. Split data into analysis set and test set.
4. Press 'Analyze Data' button.
5. Leave running for a few hours.
6. Score completion supplied (blind) test set.
7. Upload results back to competition site for ranking.

THE RESULTS

The models generated by 11Ants Customer Churn Analyzer out-performed models built by 85% of the competition submissions. This was after 40 hours of processing time. A snapshot of the general performance of the models at hours 3, 5, 9, 15, 24 and 40 are shown below. You will note that even at hour three the model outperformed 78% of the submissions.

It should be noted that though computer processing time was up to 40 hours, actual human time did not exceed 30 minutes total.



Runtime (Hours)	Leader Board Rank (Relative)
3	22%
5	21%
9	18%
15	18%
24	17%
40	15%

APPENDIX A – DETAILED DESCRIPTION OF PROCESS

This description has been provided to enable anyone to emulate experiment. To obtain an evaluation license of 11Ants Customer Churn Analyzer contact sales@11AntsAnalytics.com.

Register an account at: <http://www.kddcup-orange.com/register.php>

(Task Time: Approximately 2 Minutes)

Download Orange Data

Challenge Datasets

Training and test data matrices and practice target values

The large dataset archives are available since the onset of the challenge. The small dataset will be made available at the end of the fast challenge. Both training and test sets contain 30,000 examples. The data are split similarly for the small and large versions, but the samples are ordered differently within the training and within the test sets. Both small and large datasets have numerical and categorical variables. For the large dataset, the first 14,740 variables are numerical and the last 260 are categorical. For the small dataset, the first 190 variables are numerical and the last 40 are categorical. Toy target values are available only for practice purpose. The prediction of the toy target values will not be part of the final evaluation.

WARNING: Due to the success of the challenge, we are experiencing higher volumes of data download than expected. If you have problems downloading, try to download the files one by one. You may also want to try the data mirror nearest to you (these mirrors are made available as a courtesy of our friends whom we warmly thank; they cannot be held responsible for any problem you may encounter; please contact the organizers if you have questions):
 Text files only: Switzerland, Florida, Tennessee, USA
 Text + Matlab files: Germany, Mexico, Austria, California
 The files are password protected with:
 login: KDDcup09
 password: letmein
 Please do not distribute this password to non registered users.






Official release of Orange

Small version (23,000 var.)	Large version (15,000 var.)	zip of text files	Toy targets (large)
orange_small_train.data.zip (8.2 Mbytes) orange_small_test.data.zip (8.2 Mbytes)	orange_large_train.data.chunk1 (52.7 Mbytes) orange_large_train.data.chunk2 (52.7 Mbytes) orange_large_train.data.chunk3 (52.6 Mbytes) orange_large_train.data.chunk4 (52.5 Mbytes) orange_large_train.data.chunk5 (52.6 Mbytes)	orange_large_train_toy.labels	
	orange_large_test.data.chunk1 (52.8 Mbytes) orange_large_test.data.chunk2 (52.5 Mbytes) orange_large_test.data.chunk3 (52.6 Mbytes) orange_large_test.data.chunk4 (52.6 Mbytes) orange_large_test.data.chunk5 (52.6 Mbytes)		

The data sets are available at: <http://www.kddcup-orange.com/data.php>

(Task Time: Approximately 5 Minutes)

Data files

Name	Date modified	Type	Size
 orange_small_test.data	11/02/2009 11:37 ...	DATA File	25,607 KB
 orange_small_train.data	11/02/2009 11:37 ...	DATA File	25,592 KB
 orange_small_train_appetency.labels	11/02/2011 2:15 p...	LABELS File	195 KB
 orange_small_train_churn.labels	11/02/2011 2:15 p...	LABELS File	192 KB
 orange_small_train_upselling.labels	11/02/2011 2:15 p...	LABELS File	192 KB

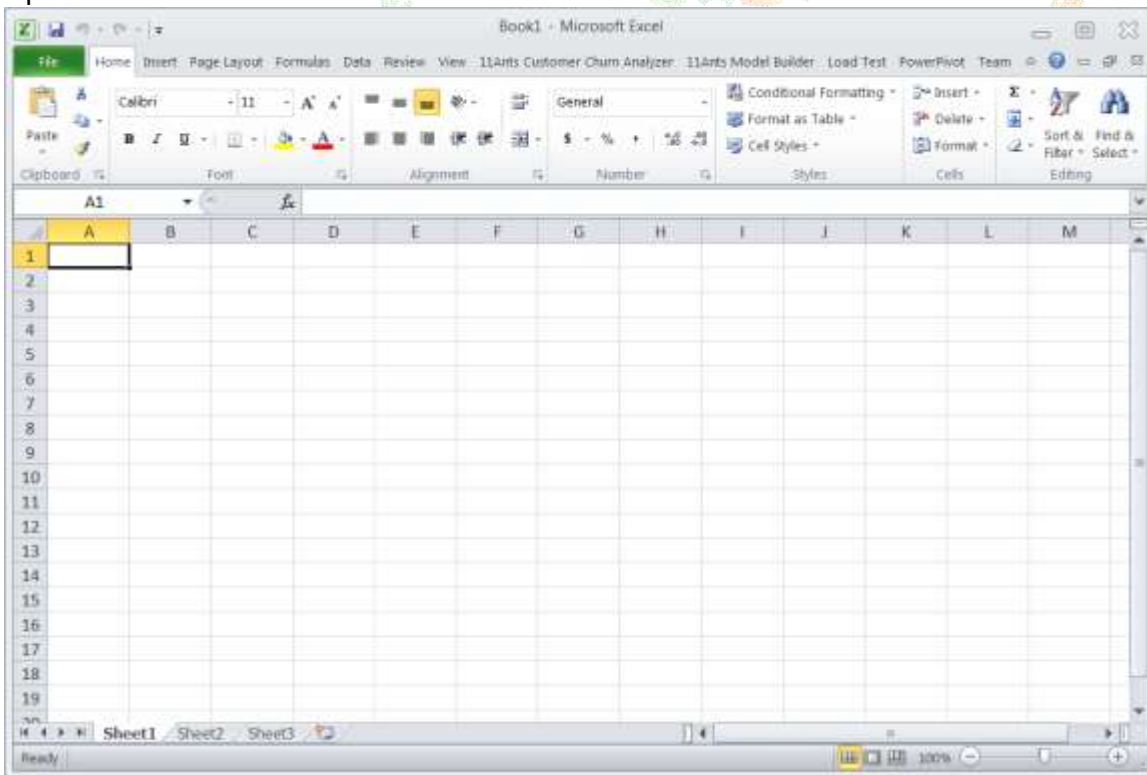
Orange_small_train.data, is the training set

Orange_small_test.data, is the test set

*.labels, are target labels

Prepare the churn data set for 11ants CCA

Open a new Excel window



Drag and Drop Orange_small_train.data into the Excel window. Excel will take a few seconds to load the data.

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15	Var16	Var17	Var18	Var19	Var20	Var21	Var22	Var23	Var24		
1	104																							454	500	
2	525																								168	210
3	3218																								1211	1313
4																										0
5	1039																								64	80
6	658																								224	200
7	5980																								308	365
8	75																								32	40
9	1118																								300	250
10	1143																								200	200
11	400																								92	113
12	708																								236	205
13	585																								0	0
14	3708																								480	600
15	3613																								148	185
16	259																								8	10
17	5152																								584	730
18	1449																								568	210
19	574																								12	15
20	608																								104	108
21	245																								598	210
22	14																								0	25
23	818																								100	240
24	864																								0	65
25	5718																								216	270
26	5043																								152	190
27																										
28	1513																								331	415
29	101																								0	0
30	62																								312	390
31	914																								112	140
32	408																								28	35
33	813																								152	165
34	1008																								160	200
35	2108																								512	640
36																										
37																										

TM

Now, the training data set (orange_small_train.data) is loaded.

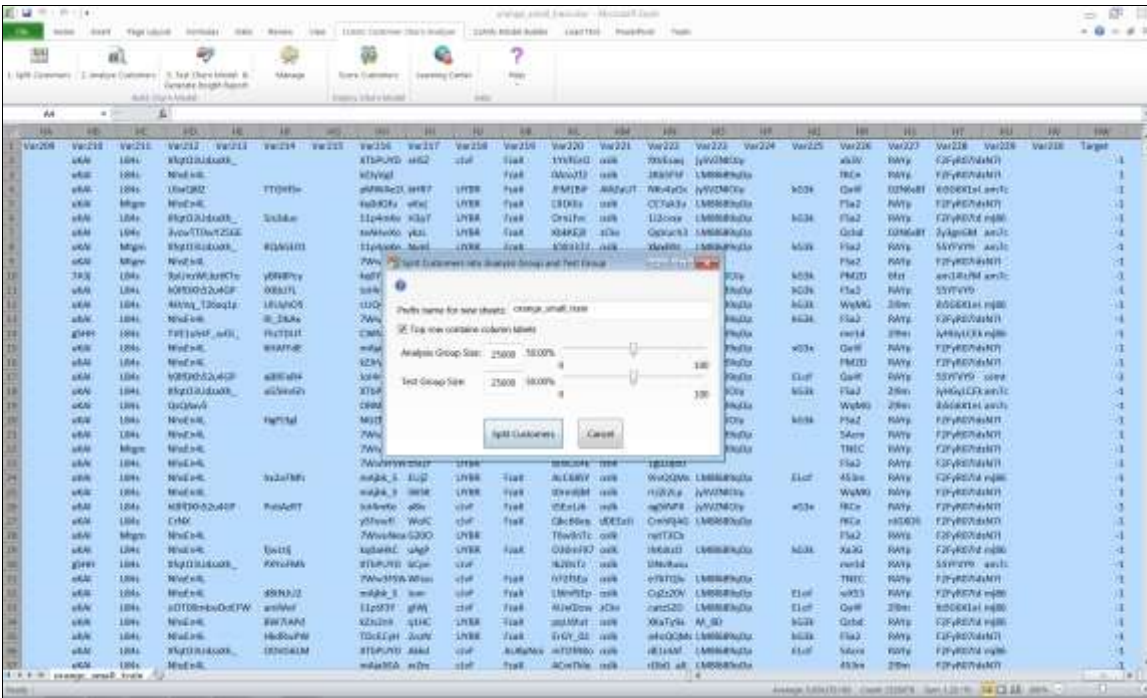
Use a Text editor to open the label file, orange_small_train_churn.labels, and then Copy and Paste the values (1s, -1s) into the rightmost column (column HW) of the Excel window we have opened before.

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15	Var16	Var17	Var18	Var19	Var20	Var21	Var22	Var23	Target	
1	104																								1
2	525																								-1
3	3218																								-1
4																									-1
5	1039																								-1
6	658																								-1
7	5980																								-1
8	75																								-1
9	1118																								-1
10	1143																								-1
11	400																								-1
12	708																								-1
13	585																								-1
14	3708																								-1
15	3613																								-1
16	259																								-1
17	5152																								-1
18	1449																								-1
19	574																								-1
20	608																								-1
21	245																								-1
22	14																								-1
23	818																								-1
24	864																								-1
25	5718																								-1
26	5043																								-1
27																									-1
28	1513																								-1
29	101																								-1
30	62																								-1
31	914																								-1
32	408																								-1
33	813																								-1
34	1008																								-1
35	2108																								-1
36																									-1
37																									-1

Save the Excel file as a XLSX file, for example, orange_small_train.xlsx.

Now, we have the training set ready for 11ants CCA. (Task Time to prepare data: Approximately 3 Minutes)

Use 11ants CCA to build a churn model
Open orange_small_train.xlsx if it has been closed.



First select all (Ctrl+A), and then Click on <1. Split Customer> under the <11ants Customer Churn Analyzer> tab.

Click on the <Split Customer> button.

23				680	0					
24				721	7					
25				1652	7					
26				889	7					8
27			0							
28				1897	21					42

orange_small_train | orange_small_train_analyze | orange_small_train_test

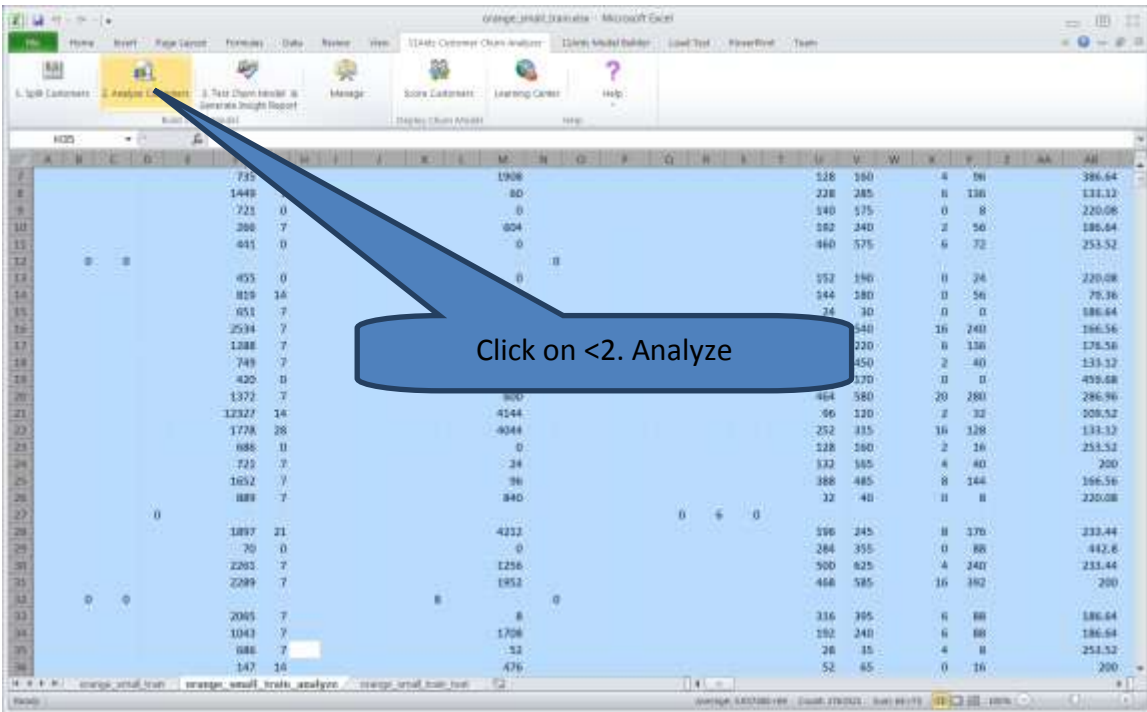
Ready | Average: 5.93748E+69 | Count: 1762521

Now, we have three sheets.

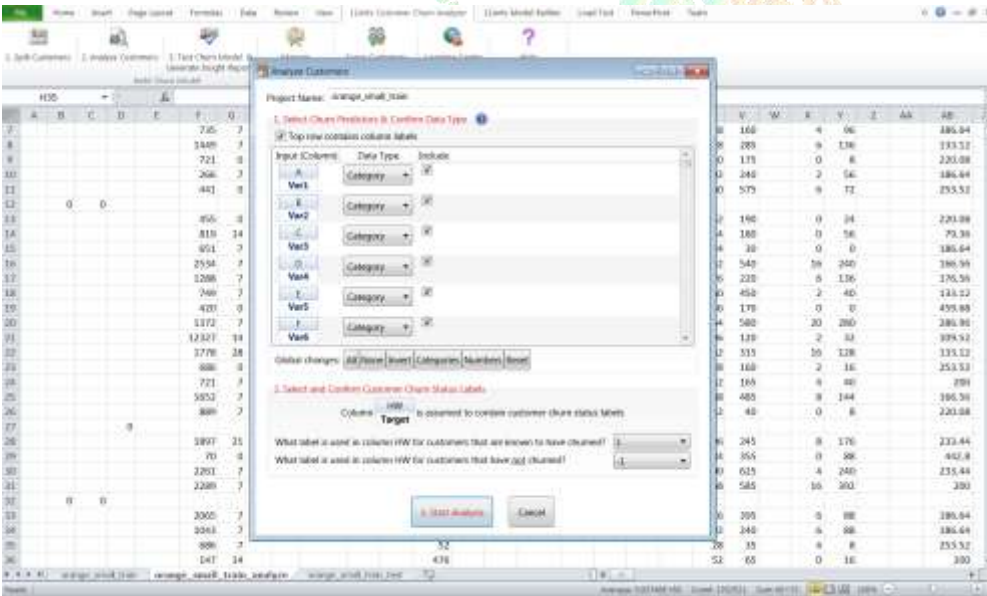
- 1) Orange_small_train, the full training set
- 2) Orange_small_train_analyze, we will be using this data set to build the churn model
- 3) Orange_small_train_test, we will test the churn model on this data set

Save the Excel file.

Click the < orange_small_train_analyze > sheet to make it as the active sheet.

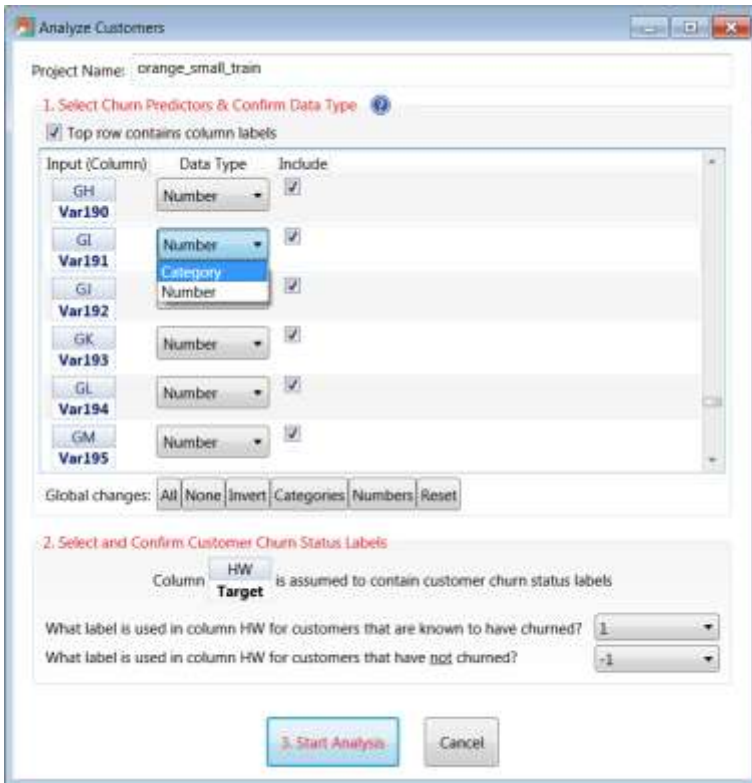


Select all (Ctrl+A), and then Click on <2. Analyze Customers>.

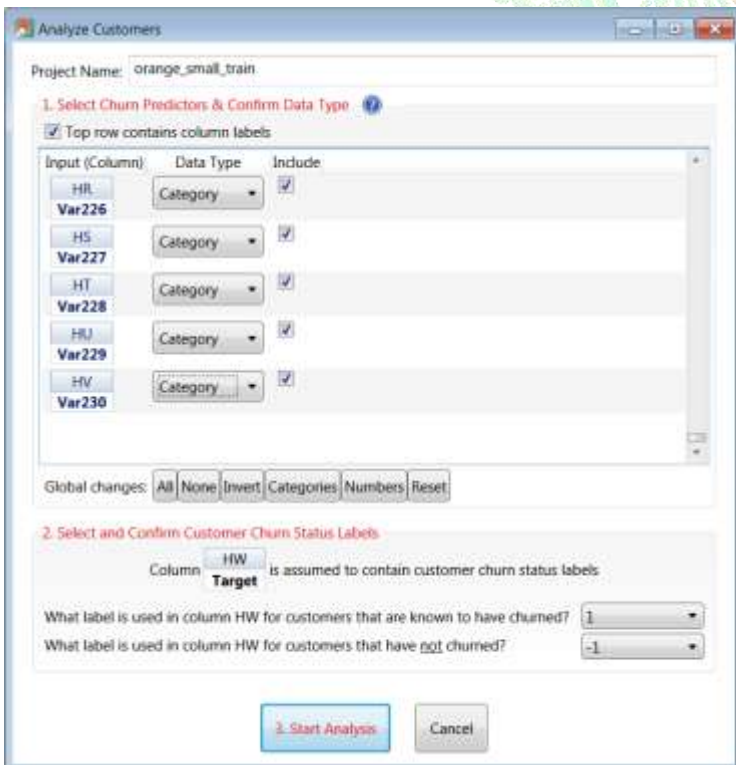


Tick the 'Top row contains column labels' option.

Click on the <All> button, and then Click on the <Numbers> button. By doing this, we set all column types to 'Numbers'. This data set has 230 columns (features). As described on the Orange website: "For the small dataset, the **first 190 variables are numerical** and the **last 40 are categorical**." Next, we need to manually set the last 40 columns as 'Category'.



Scroll down to Column 'GI Var191', select Category from the drop down box.

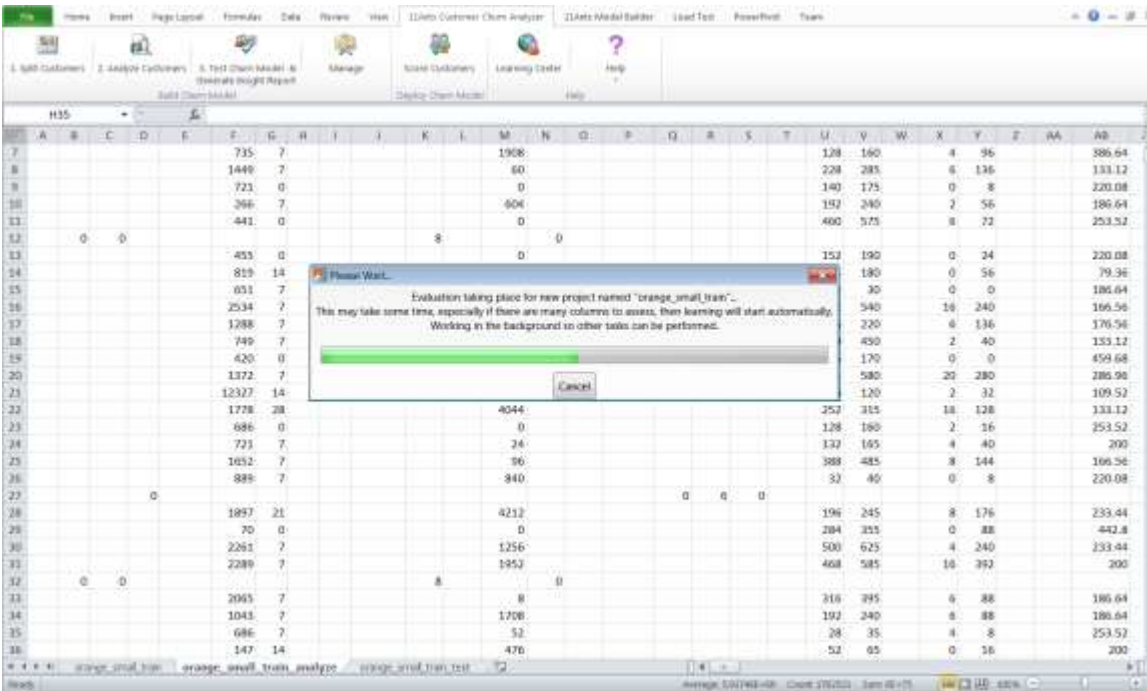


Do the same for the last 40 columns (to column HV Var230).

(Task Time: Approximately 2 Minutes)

Total time to this point: Approximately 15 Minutes).

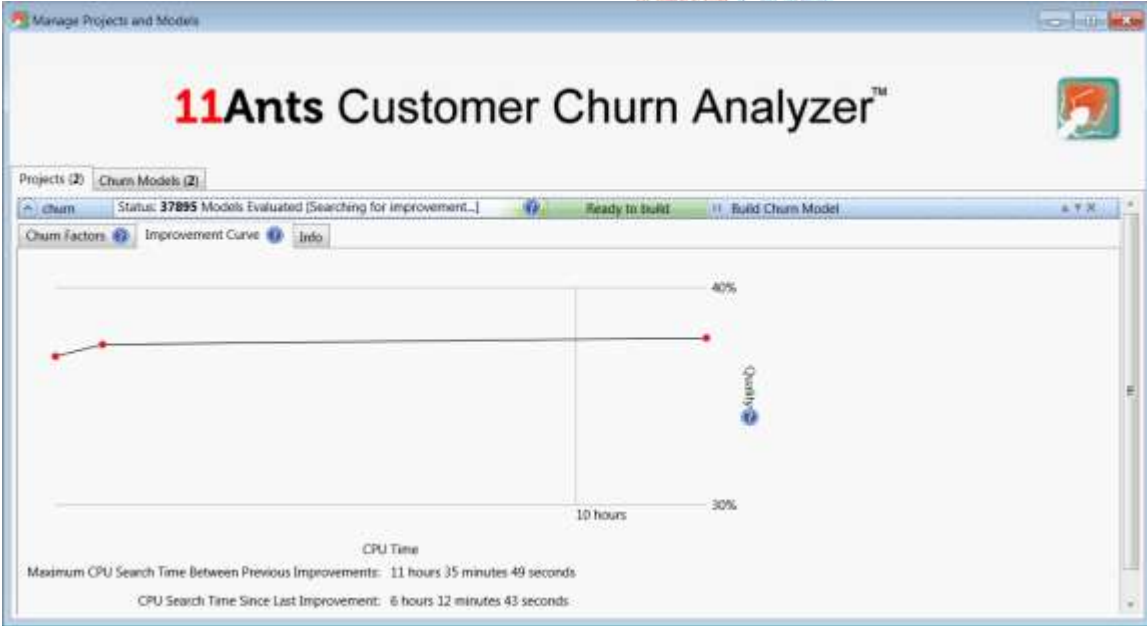
Click on the <3. Start Analysis> button. Then, 11ants CCA will start the model building process.



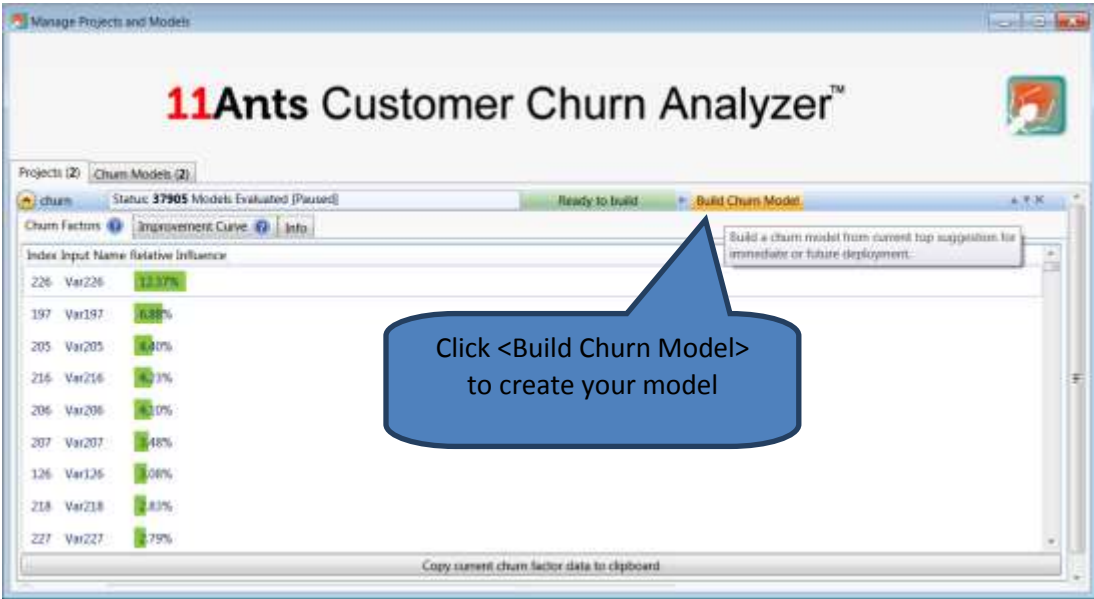
A dialog will pop up showing the status of constructing a new analysis project. The time required for this stage depends on how large the data set is. For our example, the data set has 25,000 rows and 230 columns, which are 5,750,000 cells.

This takes about 20 minutes.
 To this point total time required is approximately 35 minutes total.

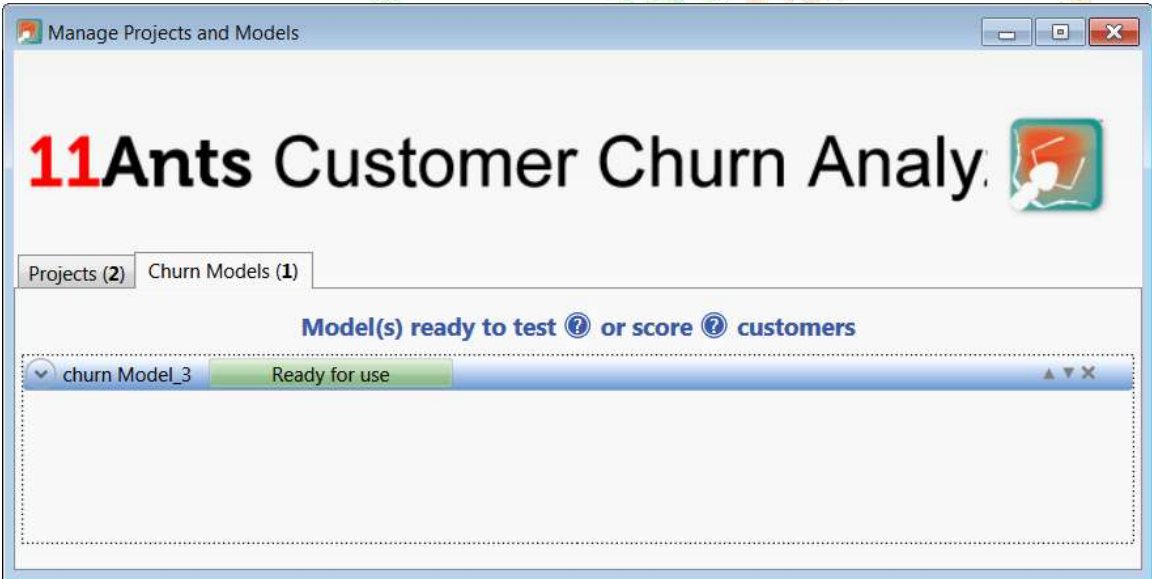
Then, 11ants CCA will start analyzing the data and searching for the best model



You can leave the CCA to run for a few hours.
 Create the Churn model



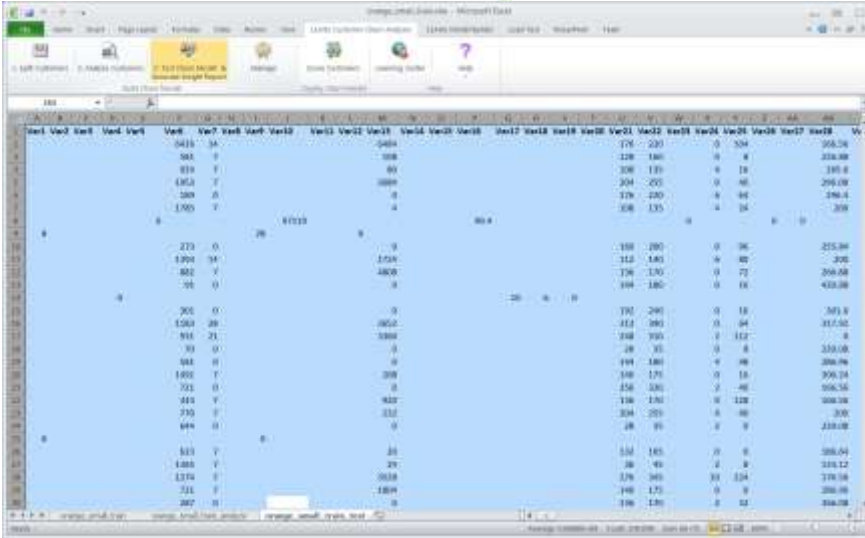
This takes about 20 seconds.



Now the churn model has been built and saved. Your model name might be different from this one.

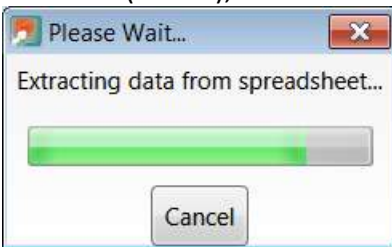
Test the churn model

Next, we will test our model on the data in the < orange_small_train_test > sheet.



The screenshot shows a Microsoft Excel spreadsheet with a data table. The columns are labeled 'Var1' through 'Var7'. The data consists of numerical values for each variable across multiple rows. The spreadsheet is titled 'orange_small_train_test'.

Select all (Ctrl+A), and Click on <3. Test Churn Model & Generate Insight Report>.

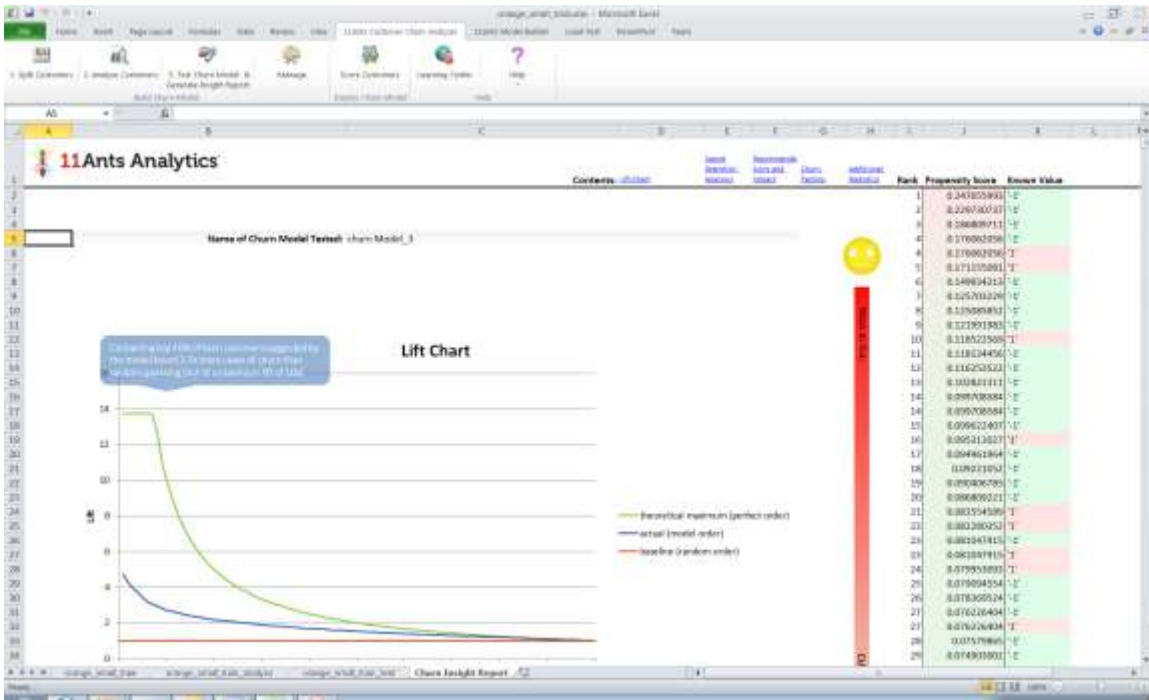


This takes about 5 seconds.

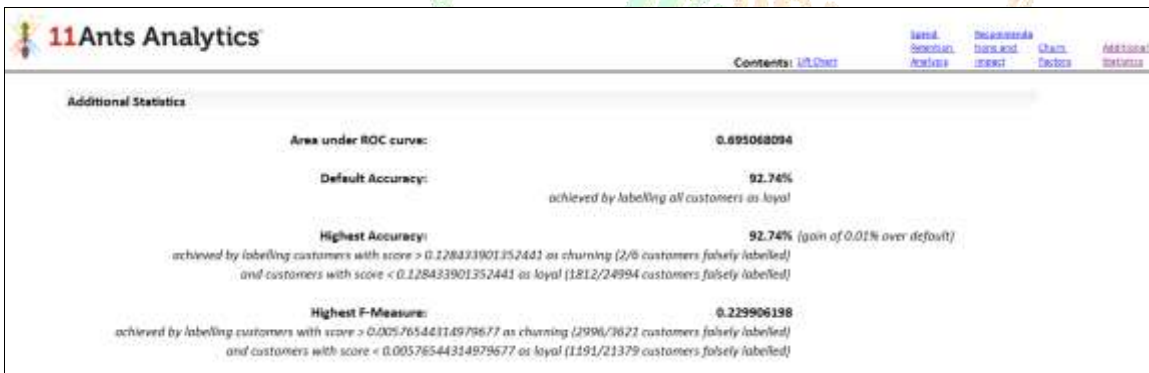


Select the model we just created.

Then, Click on <Test Now>.



This is the Insight Report generated by 11antst CCA.

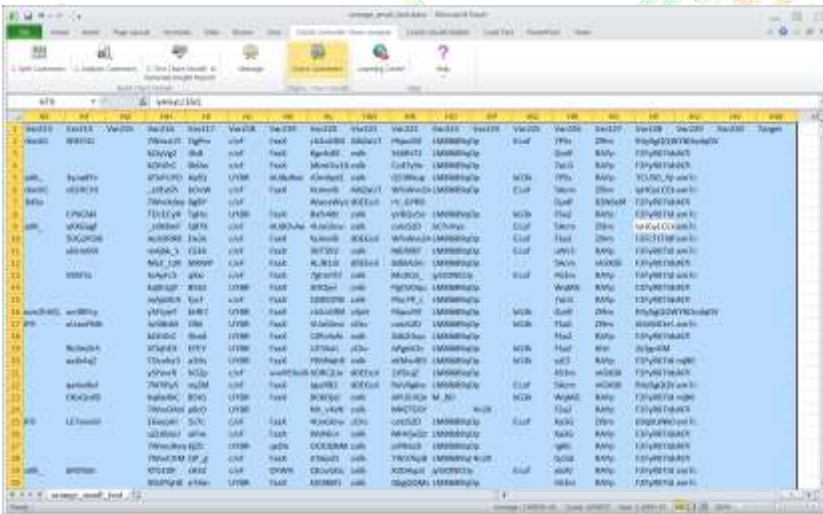


Click on <Additional Statistics>, this will bring us to the additional statistics section. The churn model has been evaluated based on different evaluations metrics.

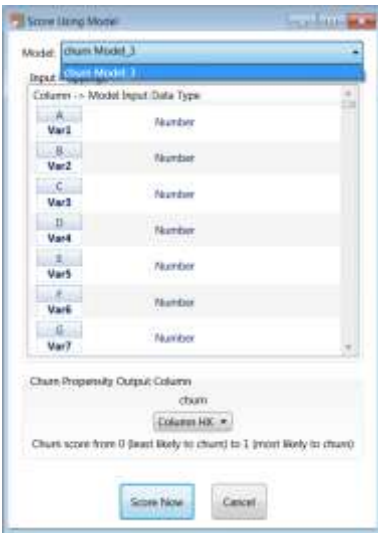
Submitting to the KDD website

Before submitting, we need to perform the same procedure as we have done for the training data set orange_small_train.xlsx, to generate the test set.

Open an empty Excel window.
Drag and Drop Orange_small_test.data into Excel.
Name column HW as 'Target'.



Click Ctrl+A to select all data, and then Click on <Score Customers>.
This takes about 30 seconds.



Choose the model we have created.
Click <Score Now>.

This procedure to be repeated for the two other tasks: the appetency and the upselling problems.







Save your prediction values to text files respectively:

orange_small_test_appetency.resu
orange_small_test_upselling.resu

Copy the three *.labels files, and name them as:

orange_small_train_appetency.resu
orange_small_train_churn.resu
orange_small_train_upselling.resu

Now, you have the following 6 result files.

 orange_small_test_appetency.resu
 orange_small_test_churn.resu
 orange_small_test_upselling.resu
 orange_small_train_appetency.resu
 orange_small_train_churn.resu
 orange_small_train_upselling.resu

Use WINZIP or WINRAR to zip these files into a single ZIP file, for example, my_submission.zip.

Challenge Submission Form

WARNING: Only team leaders should make submissions. Please limit yourself to 3 submissions daily. Submissions received before April 10, midnight (server time) will count towards the fast challenge. Report submission problems before April 9 to get support.

Method	<input type="text"/>
Method description	<div style="border: 1px solid gray; height: 150px;"></div>
Results file	<input type="text"/> <input type="button" value="Browse... (.zip, .tar.gz)"/>

(Please click only once !)

Finally, you can submit your result to the KDD website at:

<http://www.kddcup-orange.com/submit.php>

Total time for all three models should not exceed 55 minutes of human time.

APPENDIX B – REFERENCES

11Ants Customer Churn Analyzer, <http://www.11antsanalytics.com/>

KDD Cup 2009, <http://www.kddcup-orange.com/>

AUC Metric http://en.wikipedia.org/wiki/Receiver_operating_characteristic

Customer Churn, http://en.wikipedia.org/wiki/Customer_attrition

An Introduction to ROC Analysis, <http://www.sciencedirect.com/science/article/pii/S016786550500303X>

TM

APPENDIX C – ABOUT 11ANTS CUSTOMER CHURN ANALYZER

11Ants Customer Churn Analyzer is software purpose built for building customer churn propensity models. For further information about 11Ants Customer Churn Analyzer, please visit www.11AntsAnalytics.com or contact sales@11AntsAnalytics.com .

