

## Individual Assignment 8

### Topic: Logistic Regression, Performance, and Over-Sampling in RapidMiner

**YOUR NAME (in English):** YUCHIH

**STUDENT ID:** 103034034

#### Submission instruction:

1. Upload a single PDF to Canvas. It should include all the screenshots and answers to (a)-(g).
2. Hand in the printed copy of your document in class.

#### Data Details and Goal:

We continue to use the dataset from Assignment 7, in *HubwayTripsByRegistered.xlsx (or csv)*.  
Reminder: the file has a derived column **trip\_type** which says whether the trip started and ended in the same station (“RoundTrip”) or not (“OneWay”).

Our goal is to classify the trip type of a new ride, right after the bike is taken from the station.  
Create a CSV file from your Assignment 7 file, that includes all dummies, after sampling 30,000 records.

In RapidMiner:

- Select only the relevant attributes (outcome and predictors) to be used in the logistic regression model.
- Partition the dataset into training, validation, and test sets (equal percentages). Use local random seed 1234.
- Run a logistic regression of *trip\_type* with predictors DOW, gender, and hour\_bin (all are dummies).

(a) How well does the model with all predictors perform in terms of *classification*? Include a screenshot of the Validation confusion matrix, sensitivity, and specificity.

accuracy: 97.30%

	true OneWay	true RoundTrip	class precision
pred. OneWay	9730	270	97.30%
pred. RoundTrip	0	0	0.00%
class recall	100.00%	0.00%	

(b) Include a screenshot of the model coefficient information.

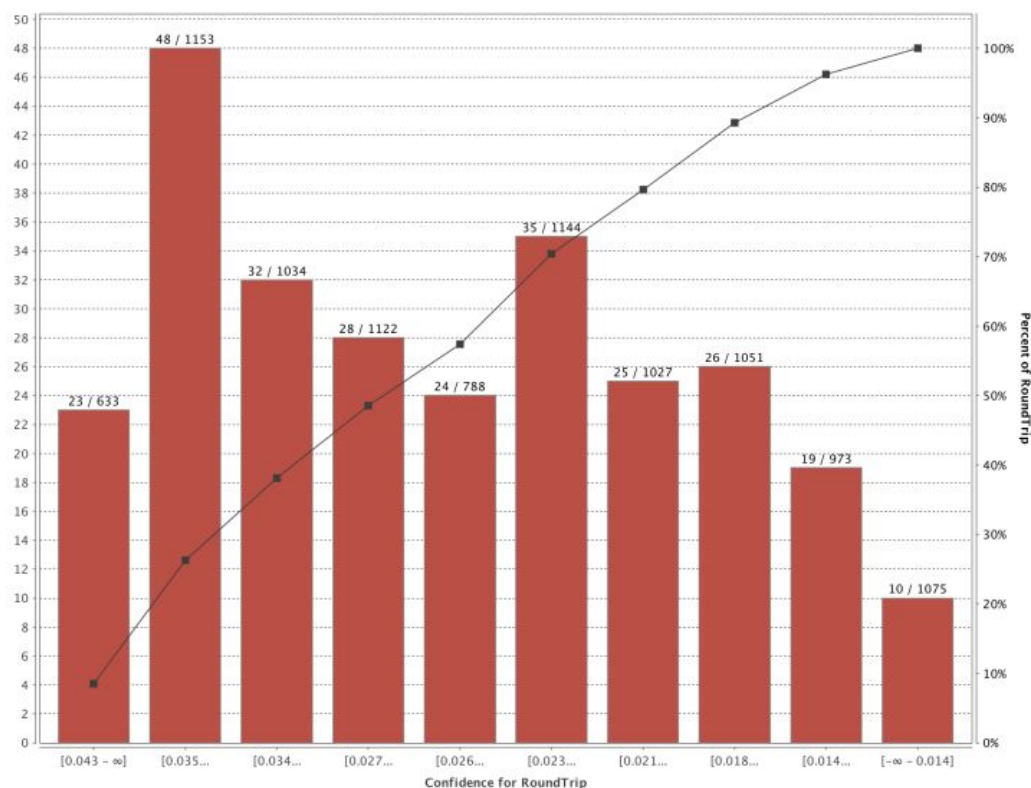
Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Male.0	0.110	0.110	0.139	0.792	0.428
DOW_1.1	0.188	0.188	0.241	0.780	0.435
DOW_2.1	-0.355	-0.355	0.231	-1.542	0.123
DOW_3.1	-0.430	-0.430	0.236	-1.823	0.068
DOW_4.0	0.497	0.497	0.237	2.094	0.036
DOW_5.1	-0.328	-0.328	0.231	-1.422	0.155
DOW_6.1	-0.248	-0.248	0.232	-1.069	0.285
bin_hour_1.0	0.482	0.482	0.399	1.207	0.228
bin_hour_2.1	-0.478	-0.478	0.163	-2.925	0.003
bin_hour_3.1	-0.247	-0.247	0.144	-1.718	0.086
Intercept	-4.047	-4.047	0.420	-9.635	0

(c) If our goal is *ranking* the top 20% of rides most likely to be RoundTrip, how well does our model perform compared to randomly choosing 20% of rides? Include a chart that supports your answer. See these two videos on ranking in RapidMiner:

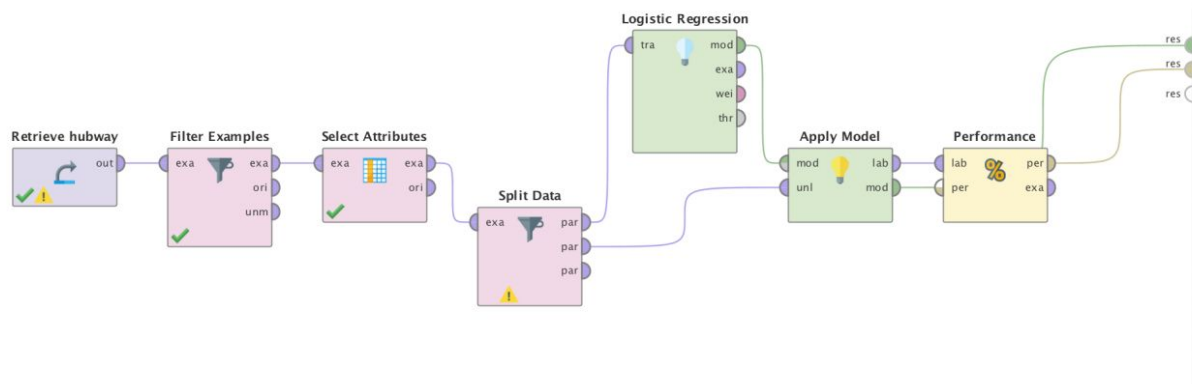
<https://www.youtube.com/watch?v=bsiG-xCoKP8>

<https://www.youtube.com/watch?v=j6Hxf5UtDKU>

It's a lift chart question. The performance of ranking the 20% with higher probability to be RoundTrip will out perform "Random Select" strategy.



(d) Include a screenshot of your **Main Process**



### Over-sampling:

Re-use your data partitioning. Now we want the training data to be over-sampled (50%-50%) , but not the validation and test sets!

In RapidMiner, apply *Sample* for the training data. In the options, click on “balance data” (it might be a hidden parameter – find it!). Create a sample that includes all the RoundTrip records and an equal number of OneWay records.

(e) Fill in the table to show your over-sampling procedure:

	# records	% <i>ReturnTrip</i>
<b>Training</b>	10000	0.5
<b>Validation</b>	10000	0.027
<b>Test</b>	10000	0.025

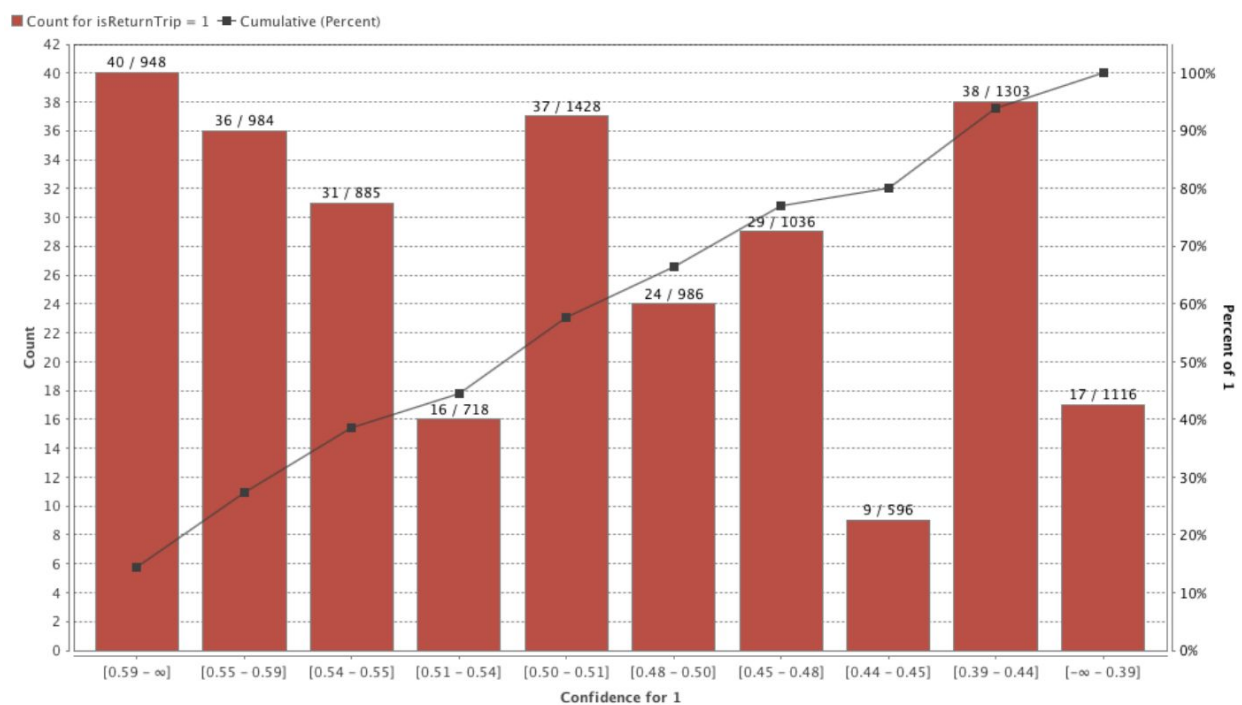
Re-run logistic regression on the over-sampled training dataset. Apply the model to the validation and test data.

(f) Compare the results to the model which was run on the random partitioning:

- The coefficients in the two models are **different** (choose one) – include a screenshot of your model coefficients

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Male.0	0.047	0.047	0.196	0.238	0.812
DOW_1.1	0.021	0.021	0.358	0.059	0.953
DOW_2.1	-0.417	-0.417	0.323	-1.292	0.196
DOW_3.1	-0.143	-0.143	0.338	-0.423	0.672
DOW_4.0	0.469	0.469	0.329	1.426	0.154
DOW_5.1	-0.106	-0.106	0.335	-0.317	0.751
DOW_6.1	0.020	0.020	0.339	0.058	0.954
bin_hour_1.0	0.660	0.660	0.523	1.260	0.208
bin_hour_2.1	-0.530	-0.530	0.232	-2.288	0.022
bin_hour_3.1	-0.300	-0.300	0.212	-1.416	0.157
Intercept	-0.671	-0.671	0.542	-1.239	0.216

- The validation overall error rate is equal to 48.84%. It is higher(choose one) with oversampling.
- The validation sensitivity to *ReturnTrip* is 3.19%. It is higher (choose one) with oversampling.
- The validation lift, first decile, is higher(choose one) with oversampling – include a screenshot of your lift chart.



(g) Include a screenshot of your **Main Process (Design)**

