Individual Assignment 8

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

YOUR NAME (in English): Zoe Cheng

STUDENT ID: 102072243

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:

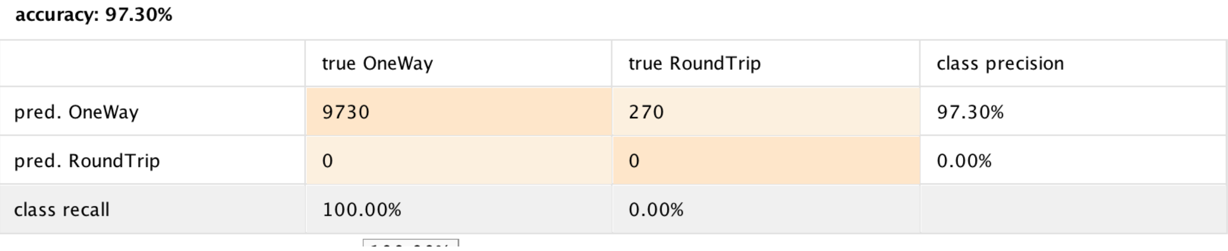
We continue to use the dataset from Assignment 7, in *HubwayTripsByRegistered.xlsx (or csv)*. Reminder: the file has a derived column **trip_type** which says whether the trip started and ended in the same station ("RoundTrip") or not ("OneWay").

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).

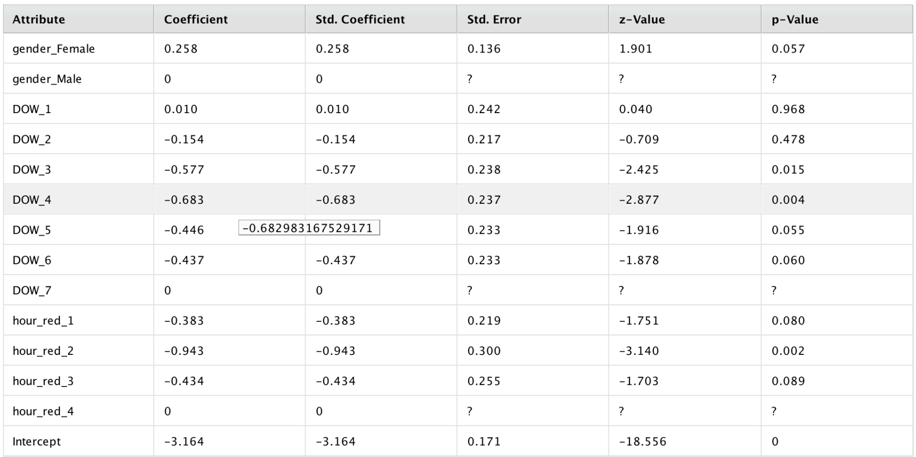
(a) How well does the model with all predictors perform in terms of *classification*? Include a screenshot of the Validation confusion matrix, sensitivity, and specificity.



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%	

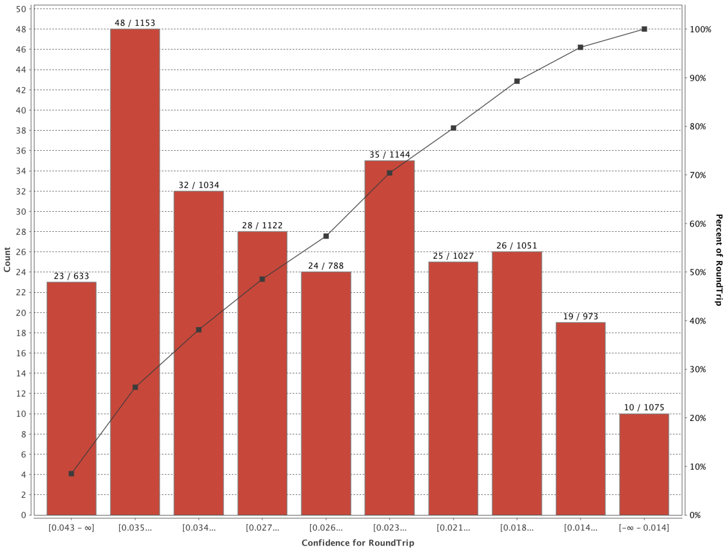
(b) Include a screenshot of the model coefficient information.



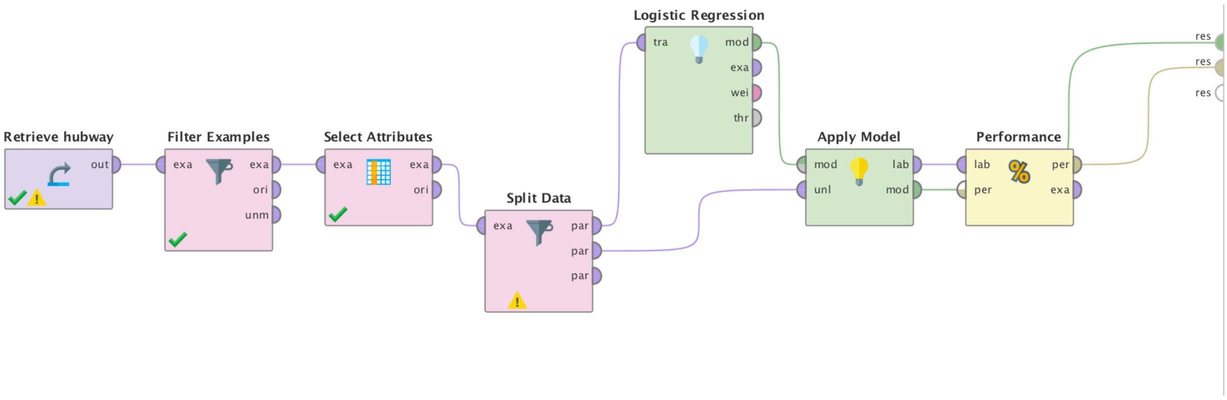
Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
gender_Female	0.258	0.258	0.136	1.901	0.057
gender_Male	0	0	?	?	?
DOW_1	0.010	0.010	0.242	0.040	0.968
DOW_2	-0.154	-0.154	0.217	-0.709	0.478
DOW_3	-0.577	-0.577	0.238	-2.425	0.015
DOW_4	-0.683	-0.683	0.237	-2.877	0.004
DOW_5	-0.446	-0.682983167529171	0.233	-1.916	0.055
DOW_6	-0.437	-0.437	0.233	-1.878	0.060
DOW_7	0	0	?	?	?
hour_red_1	-0.383	-0.383	0.219	-1.751	0.080
hour_red_2	-0.943	-0.943	0.300	-3.140	0.002
hour_red_3	-0.434	-0.434	0.255	-1.703	0.089
hour_red_4	0	0	?	?	?
Intercept	-3.164	-3.164	0.171	-18.556	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:

Better than random



(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.

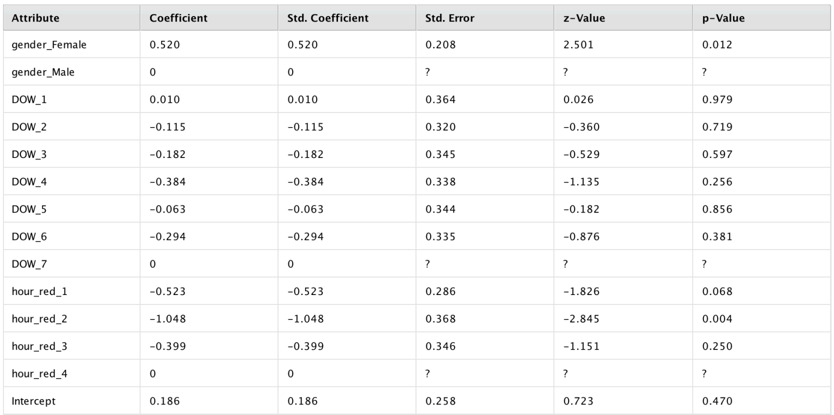
(e) Fill in the table to show your over-sampling procedure:

	# records	**% *ReturnTrip***
Training	540	0.5
Validation	10000	0.027
Test	10000	0.027

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

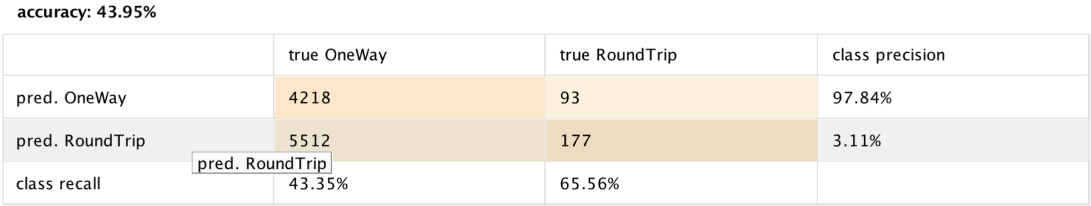
(f) Compare the results to the model which was run on the random partitioning:

- The coefficients in the two models are different – include a screenshot of your model coefficients



Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
gender_Female	0.520	0.520	0.208	2.501	0.012
gender_Male	0	0	?	?	?
DOW_1	0.010	0.010	0.364	0.026	0.979
DOW_2	-0.115	-0.115	0.320	-0.360	0.719
DOW_3	-0.182	-0.182	0.345	-0.529	0.597
DOW_4	-0.384	-0.384	0.338	-1.135	0.256
DOW_5	-0.063	-0.063	0.344	-0.182	0.856
DOW_6	-0.294	-0.294	0.335	-0.876	0.381
DOW_7	0	0	?	?	?
hour_red_1	-0.523	-0.523	0.286	-1.826	0.068
hour_red_2	-1.048	-1.048	0.368	-2.845	0.004
hour_red_3	-0.399	-0.399	0.346	-1.151	0.250
hour_red_4	0	0	?	?	?
Intercept	0.186	0.186	0.258	0.723	0.470

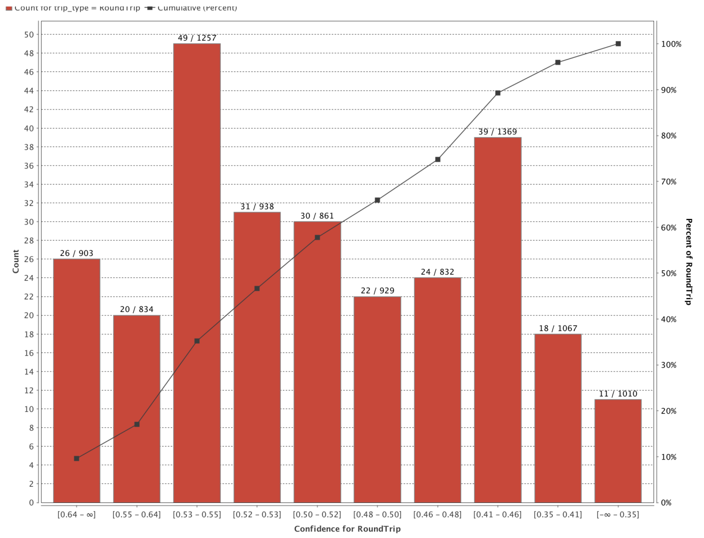
- The validation overall error rate is equal to **56.05%**. It is **higher** with oversampling.



accuracy: 43.95%

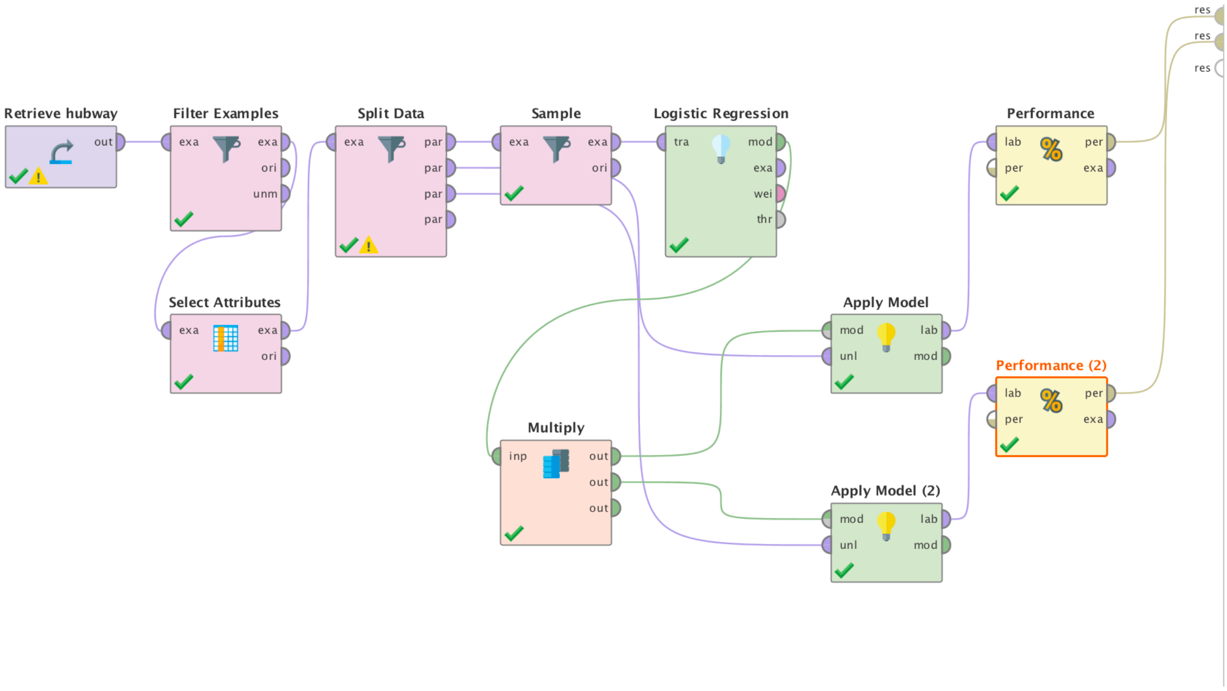
	true OneWay	true RoundTrip	class precision
pred. OneWay	4218	93	97.84%
pred. RoundTrip	5512	177	3.11%
class recall	43.35%	65.56%	

- The validation sensitivity to *ReturnTrip* is __**65.56%**__. It is **higher** with oversampling.
- The validation lift, first decile, is **higher** with oversampling – include a screenshot of your lift chart.



-

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


Retrieve hubway
out
Filter Examples
exa
exa
ori
unm
Split Data
exa
par
par
par
par
Sample
exa
exa
ori
Logistic Regression
tra
mod
exa
wei
thr
Select Attributes
exa
exa
ori
Multiply
inp
out
out
out
Apply Model
mod
lab
unl
mod
Apply Model (2)
mod
lab
unl
mod
Performance
lab
per
per
exa
Performance (2)
lab
per
per
exa
res
res
res