HASKAYNE
SCHOOL OF BUSINESS

UNIVERSITY OF CALGARY
HASKAYNE SCHOOL OF BUSINESS

Predictive Models in Business Analytics – Assignment 3

Due: Monday November 28, 2022 at 11:00am

Group Number:	**01**

Last Name *(in alphabetical order)*	**First Name**
Jovanovic	Daniel
Sehgal	Rupin

QUESTION 1 – BIKE RENTALS

For this question, split the data where 70% is used for training, and 30% is used for test data. Use shuffled sampling with local random seed 1992. Note that we will be using a validation dataset later in the question.

a. Create a linear regression model where you include all attributes except the actual date. Save your RapidMiner file as "A3_Q1a.rmp" in the processes folder of your Local Repository.

Regression model (screenshot of the full equation is sufficient, make sure it is readable when printed):

```
  148.436 * Hour = 0
+ 63.463 * Hour = 1
- 31.971 * Hour = 2
- 114.085 * Hour = 3
- 189.362 * Hour = 4
- 176.114 * Hour = 5
- 17.899 * Hour = 6
+ 290.804 * Hour = 7
+ 679.149 * Hour = 8
+ 198.820 * Hour = 9
- 27.517 * Hour = 10
- 36.138 * Hour = 11
- 1.935 * Hour = 13
- 19.845 * Hour = 14
+ 87.455 * Hour = 15
+ 181.336 * Hour = 16
+ 475.291 * Hour = 17
+ 874.700 * Hour = 18
+ 660.507 * Hour = 19
+ 576.045 * Hour = 20
+ 572.306 * Hour = 21
+ 460.593 * Hour = 22
+ 266.054 * Hour = 23
+ 51.770 * Day of Week = Friday
+ 4.088 * Day of Week = Saturday
- 35.414 * Day of Week = Sunday
+ 23.757 * Day of Week = Tuesday
+ 35.552 * Day of Week = Wednesday
+ 25.233 * Day of Week = Thursday
+ 31.312 * Weekday = Weekday
+ 70.746 * Month = Dec
- 48.814 * Month = Feb
- 61.647 * Month = Mar
+ 14.966 * Month = Apr
+ 172.787 * Month = May
+ 311.667 * Month = Jun
- 23.982 * Month = Jul
- 208.384 * Month = Aug
+ 1.027 * Month = Sep
+ 150.247 * Month = Oct
+ 81.661 * Month = Nov
```

```
+ 126.100 * Seasons = Spring
+ 79.390 * Seasons = Summer
+ 232.877 * Seasons = Autumn
+ 147.168 * Holiday = No Holiday
+ 945.093 * Functioning Day = Yes
+ 14.386 * Temperature(°C)
- 9.843 * Humidity(%)
+ 1.667 * Wind speed (m/s)
+ 0.048 * Visibility (10m)
+ 14.027 * Dew point temperature(°C)
+ 52.203 * Solar Radiation (MJ/m2)
- 54.038 * Rainfall(mm)
+ 25.603 * Snowfall (cm)
- 502.904
```

	Training Performance	Test Performance
Root mean squared error	349.086 (+/- 0.000)	361.410 (+/- 0.000)
Absolute error	264.809 (+/- 227.459)	274.332 (+/- 235.285)
Relative error	161.05% (+/- 616.11%)	168.93% (+/- 751.98%)
Correlation	0.841	0.828
Squared correlation	0.707	0.685

b. Save the file of Question 1a as "A3_Q1b.rmp" in the process folder of your Local Repository. Update the file such that you include M5 prime as feature selection method and exclude collinear features.

Regression model (screenshot of the full equation is sufficient, make sure it is readable when printed):

```
  148.436 * Hour = 0
+ 63.463 * Hour = 1
- 31.971 * Hour = 2
- 114.085 * Hour = 3
- 189.362 * Hour = 4
- 176.114 * Hour = 5
- 17.899 * Hour = 6
+ 290.804 * Hour = 7
+ 679.149 * Hour = 8
+ 198.820 * Hour = 9
- 27.517 * Hour = 10
- 36.138 * Hour = 11
- 1.935 * Hour = 13
- 19.845 * Hour = 14
+ 87.455 * Hour = 15
+ 181.336 * Hour = 16
+ 475.291 * Hour = 17
+ 874.700 * Hour = 18
+ 660.507 * Hour = 19
+ 576.045 * Hour = 20
+ 572.306 * Hour = 21
+ 460.593 * Hour = 22
+ 266.054 * Hour = 23
+ 51.770 * Day of Week = Friday
+ 4.088 * Day of Week = Saturday
- 35.414 * Day of Week = Sunday
+ 23.757 * Day of Week = Tuesday
+ 35.552 * Day of Week = Wednesday
+ 25.233 * Day of Week = Thursday
+ 31.312 * Weekday = Weekday
+ 70.746 * Month = Dec
- 48.814 * Month = Feb
- 61.643 * Month = Mar
+ 14.971 * Month = Apr
+ 172.791 * Month = May
+ 311.670 * Month = Jun
- 23.978 * Month = Jul
- 208.381 * Month = Aug
+ 149.220 * Month = Oct
+ 80.634 * Month = Nov
```

```
+ 126.096 * Seasons = Spring
+ 79.387 * Seasons = Summer
+ 233.904 * Seasons = Autumn
+ 147.168 * Holiday = No Holiday
+ 945.093 * Functioning Day = Yes
+ 14.386 * Temperature(°C)
- 9.843 * Humidity(%)
+ 1.667 * Wind speed (m/s)
+ 0.048 * Visibility (10m)
+ 14.027 * Dew point temperature(°C)
+ 52.203 * Solar Radiation (MJ/m2)
- 54.038 * Rainfall(mm)
+ 25.603 * Snowfall (cm)
- 502.904
```

	Training Performance	Test Performance
Root mean squared error	349.086 (+/- 0.000)	361.410 (+/- 0.000)
Absolute error	264.809 (+/- 227.459)	274.332 (+/- 235.285)
Relative error	161.05% (+/- 616.11%)	168.93% (+/- 751.98%)
Correlation	0.841	0.828
Squared correlation	0.707	0.685

c. Save the file of Question 1a as "A3_Q1c.rmp" in the process folder of your Local Repository. Instead of M5 prime as feature selection (as in the previous question), only use forward selection as feature selection method. Stop selecting new attributes to include when no more improvements (in terms of reducing the root mean squared error) can be made. For the forward selection process itself, use 5-fold cross validation with shuffled sampling and local random seed 1992.

Regression model (screenshot of the full equation is sufficient, make sure it is readable when printed):

```
  14.415 * Temperature(°C)
- 9.801 * Humidity(%)
+ 940.644 * Functioning Day = Yes
+ 895.595 * Hour = 18
+ 698.169 * Hour = 8
+ 682.420 * Hour = 19
- 518.616 * Month = Aug
+ 593.998 * Hour = 21
+ 597.615 * Hour = 20
+ 481.626 * Hour = 22
+ 494.518 * Hour = 17
- 335.897 * Month = Jul
- 53.650 * Rainfall(mm)
+ 310.256 * Hour = 7
+ 149.140 * Month = Oct
+ 82.140 * Month = Nov
+ 287.672 * Hour = 23
+ 216.680 * Hour = 9
+ 83.149 * Weekday = Weekday
+ 388.712 * Seasons = Summer
- 169.127 * Hour = 4
- 155.188 * Hour = 5
+ 297.416 * Month = May
+ 0.048 * Visibility (10m)
+ 147.416 * Holiday = No Holiday
- 94.066 * Hour = 3
+ 198.544 * Hour = 16
- 49.149 * Month = Feb
+ 169.535 * Hour = 0
+ 57.127 * Solar Radiation (MJ/m2)
+ 231.652 * Seasons = Autumn
+ 103.598 * Hour = 15
+ 83.934 * Hour = 1
+ 13.947 * Dew point temperature(°C)
+ 138.660 * Month = Apr
+ 35.057 * Day of Week = Friday
+ 26.027 * Snowfall (cm)
+ 39.215 * Day of Week = Saturday
+ 70.892 * Month = Dec
+ 63.512 * Month = Mar
+ 18.784 * Day of Week = Wednesday
- 21.325 * Hour = 11
- 555.052
```

	Training Performance	Test Performance
Root mean squared error	349.214 (+/- 0.000)	361.651 (+/- 0.000)
Absolute error	264.932 (+/- 227.511)	274.718 (+/- 235.206)
Relative error	160.36% (+/- 611.55%)	168.55% (+/- 749.79%)

Correlation	0.841	0.828
Squared correlation	0.707	0.685

Compare the performance of the two feature selection methods (parts Q1b and Q1c) to the performance of the linear regression model without feature selection (part Q1a):
The linear regression model in Q1a and Q1b were identical across all of the relevant metrics, and they performed slightly better regarding Root mean squared error and Absolute error compared to the model created in Q1c. The model from Q1c performed better only for relative error, while the correlation and squared correlation were the same in all 3 models. The linear regression model in Q1a performs better or equally well as the models in Q1b and Q1c in every performance metric except one (relative error).

In the remain questions, we will ONLY consider the numerical attributes as predictor variables. Note that Hour is a categorical attribute (as already indicated in Question 1a).

d. Save the file of Question 1a as "A3_Q1d.rmp" in the process folder of your Local Repository. Update the file such that only the numerical attributes are considered.

Report the correlation matrix and covariance matrix for the attributes of your training dataset (screenshots are sufficient, make sure it is readable when printed):
Correlation Matrix:

Attributes	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
Temperature(°C)	1	0.167	-0.056	0.031	0.912	0.349	0.045	-0.222
Humidity(%)	0.167	1	-0.344	-0.545	0.545	-0.467	0.234	0.100
Wind speed (m/s)	-0.056	-0.344	1	0.178	-0.195	0.326	-0.016	0.006
Visibility (10m)	0.031	-0.545	0.178	1	-0.181	0.151	-0.165	-0.107
Dew point temperature(°C)	0.912	0.545	-0.195	-0.181	1	0.086	0.120	-0.156
Solar Radiation (MJ/m2)	0.349	-0.467	0.326	0.151	0.086	1	-0.074	-0.071
Rainfall(mm)	0.045	0.234	-0.016	-0.165	0.120	-0.074	1	0.016
Snowfall (cm)	-0.222	0.100	0.006	-0.107	-0.156	-0.071	0.016	1

Covariance Matrix:

Attributes	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
Temperature(°C)	1.000	0.167	-0.056	0.031	0.912	0.349	0.045	-0.222
Humidity(%)	0.167	1.000	-0.344	-0.545	0.545	-0.467	0.234	0.100
Wind speed (m/s)	-0.056	-0.344	1.000	0.178	-0.195	0.326	-0.016	0.006
Visibility (10m)	0.031	-0.545	0.178	1.000	-0.181	0.151	-0.165	-0.107
Dew point temperature(°C)	0.912	0.545	-0.195	-0.181	1.000	0.086	0.120	-0.156
Solar Radiation (MJ/m2)	0.349	-0.467	0.326	0.151	0.086	1.000	-0.074	-0.071
Rainfall(mm)	0.045	0.234	-0.016	-0.165	0.120	-0.074	1.000	0.016
Snowfall (cm)	-0.222	0.100	0.006	-0.107	-0.156	-0.071	0.016	1.000

What can be concluded from these two matrices?

Dew point temperature (C) has the highest correlation with Temperature (C) at 0.912 meaning that they both react in unison to the predictor variables. Visibility (10m) and Humidity have the highest negative correlation at -0.545. With both these correlations, it becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently because the independent variables tend to change in unison.

Regression model (screenshot of the full equation is sufficient, make sure it is readable when printed):

```
   421.683 * Temperature(°C)
- 237.237 * Humidity(%)
+ 49.955 * Wind speed (m/s)
- 9.564 * Visibility (10m)
+ 21.162 * Dew point temperature(°C)
- 106.996 * Solar Radiation (MJ/m2)
- 53.903 * Rainfall(mm)
+ 19.139 * Snowfall (cm)
+ 705.469
```

	Training Performance	Test Performance
Root mean squared error	494.886 (+/- 0.000)	513.798 (+/- 0.000)
Absolute error	365.930 (+/- 333.178)	378.455 (+/- 347.506)
Relative error	140.43% (+/- 381.38%)	146.41% (+/- 468.48%)
Correlation	0.642	0.605
Squared correlation	0.412	0.366

e. Save the file of Question 1d as "A3_Q1e.rmp" in the process folder of your Local Repository. Include a proper implementation of Principal Components Analysis (PCA) to make sure that the assumption of independence between the predictor variables is verified.

Report the correlation matrix and covariance matrix for the attributes of your training dataset after applying PCA (screenshots are sufficient, make sure it is readable when printed):
Correlation Matrix:

Attributes	pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7	pc_8
pc_1	1	0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000
pc_2	0.000	1	0.000	0.000	0.000	-0.000	0.000	-0.000
pc_3	-0.000	0.000	1	0.000	0.000	0.000	-0.000	0.000
pc_4	0.000	0.000	0.000	1	0.000	0.000	-0.000	0.000
pc_5	-0.000	0.000	0.000	0.000	1	-0.000	-0.000	-0.000
pc_6	0.000	-0.000	0.000	0.000	-0.000	1	0.000	0.000
pc_7	-0.000	0.000	-0.000	-0.000	-0.000	0.000	1	0.000
pc_8	-0.000	-0.000	0.000	0.000	-0.000	0.000	0.000	1

Covariance Matrix:

Attributes	pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7	pc_8
pc_1	2.431	0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000
pc_2	0.000	1.939	0.000	0.000	0.000	-0.000	0.000	-0.000
pc_3	-0.000	0.000	1.066	0.000	0.000	0.000	-0.000	0
pc_4	0.000	0.000	0.000	0.928	0.000	0.000	-0.000	0.000
pc_5	-0.000	0.000	0.000	0.000	0.766	-0.000	-0.000	-0.000
pc_6	0.000	-0.000	0.000	0.000	-0.000	0.644	0.000	0.000
pc_7	-0.000	0.000	-0.000	-0.000	-0.000	0.000	0.221	0.000
pc_8	-0.000	-0.000	0	0.000	-0.000	0.000	0.000	0.004

What can be concluded from these two matrices?

PCA is technique that increases interpretability but at the same time minimizes information loss by creating new uncorrelated variables (pc1, pc2…pcx). As such there is extremely small correlation between our new variables. Furthermore, there is very little variance between our attributes as well, other than when compared with themselves. This variance reduces relatively uniformly as with each new variable (pc1 v pc1 is highest and pc8 v pc8 is lowest).

Cumulative variance for each of the principal components:

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.559	0.304	0.304
PC 2	1.393	0.242	0.546
PC 3	1.032	0.133	0.680
PC 4	0.963	0.116	0.796
PC 5	0.875	0.096	0.891
PC 6	0.803	0.081	0.972
PC 7	0.470	0.028	0.999
PC 8	0.066	0.001	1.000

What can be concluded from the cumulative variances?

As the Cumulative variance increases with each new variable (lowest at pc1 and highest at pc8) the standard deviation decreases (highest at pc1 and lowest at pc8) and Proportion of Variance also decreases (highest at pc1 and lowest at pc8).

Regression model (screenshot of the full equation is sufficient, make sure it is readable when printed):

```
- 43.644 * pc_1
- 253.605 * pc_2
- 42.899 * pc_3
+ 66.690 * pc_4
+ 63.735 * pc_5
+ 56.229 * pc_6
- 372.909 * pc_7
+ 181.823 * pc_8
+ 705.469
```

	Training Performance	Test Performance
Root mean squared error	494.886 (+/- 0.000)	513.798 (+/- 0.000)
Absolute error	365.930 (+/- 333.178)	378.455 (+/- 347.506)
Relative error	140.43% (+/- 381.38%)	146.41% (+/- 468.48%)
Correlation	0.642	0.605
Squared correlation	0.412	0.366

Compare the performance of the linear regression model based on principal components (part Q1e) to the performance of the linear regression model without principal components (part Q1d):

There is no difference between the model from Q1d and the one from Q1e with regard to any of the relevant performance measures. More specifically, these two models generate the same values for the root mean squared error, absolute error, relative error, correlation, and squared correlation.

f. Save the file of Question 1e as "A3_Q1f.rmp" in the process folder of your Local Repository. Instead of including all principal components, change the PCA analysis of Question 1e such that only those principal components are included that represent at least 95% of the total variance.

Regression model (screenshot of the full equation is sufficient, make sure it is readable when printed):

```
- 43.644 * pc_1
- 253.605 * pc_2
- 42.899 * pc_3
+ 66.690 * pc_4
+ 63.735 * pc_5
+ 56.229 * pc_6
+ 705.469
```

	Training Performance	Test Performance
Root mean squared error	525.171 (+/- 0.000)	536.942 (+/- 0.000)
Absolute error	387.158 (+/- 354.843)	393.196 (+/- 365.655)
Relative error	162.77% (+/- 509.92%)	167.12% (+/- 625.72%)
Correlation	0.581	0.553
Squared correlation	0.338	0.306

Other than linear regression, many of the other prediction techniques can be used for this regression task. In the remainder of this question, we will use neural networks, where only the numerical predictor variables should be included (similar to Q1d-f).

g. Save the file of Question 1d as "A3_Q1g.rmp" in the process folder of your Local Repository. Create a new classification model with the use of neural network.

Equations for the nodes of the first layer where you include all numerical attributes as predictor variables (screenshot are NOT sufficient, as no equations are provided by RapidMiner):

Node 1: 1 / 1 + e^[-3.123 + (3.354 * Temperature) + (-6.043 * Humidity) + (2.263 * Wind Speed) + (-1.110 * Visibility) + (2.536 * Dew point temperature) + (-3.301 * Solar Radiation) + (0.064 * Rainfall) + (1.925 * Snowfall)]

Node 2: 1 / 1 + e^[-3.258 + (4.178 * Temperature) + (-7.609 * Humidity) + (3.284 * Wind Speed) + (-1.010 * Visibility) + (3.412 * Dew point temperature) + (-4.277 * Solar Radiation) + (-0.295 * Rainfall) + (1.821 * Snowfall)]

Node 3: 1 / 1 + e^[-3.791 + (-0.705 * Temperature) + (-3.034 * Humidity) + (-0.372 * Wind Speed) + (0.160 * Visibility) + (4.481 * Dew point temperature) + (0.858 * Solar Radiation) + (-3.919 * Rainfall) + (0.426 * Snowfall)]

	Training Performance	Test Performance
Root mean squared error	471.536 (+/- 0.000)	496.858 (+/- 0.000)
Absolute error	333.535 (+/- 333.317)	350.250 (+/- 352.409)
Relative error	120.24% (+/- 323.23%)	122.04% (+/- 365.74%)
Correlation	0.683	0.639
Squared correlation	0.466	0.409

h. Save the file of Question 1g as "A3_Q1h.rmp" in the process folder of your Local Repository. Instead of using all numerical predictor variables, we can also use forward selection as feature selection method for neural networks. Stop selecting new attributes to include when there is an increase in the squared correlation. Include 5 speculative rounds once terminated. For the forward selection process itself, use 5-fold cross validation with shuffled sampling and local random seed 1992.

Equations for the nodes of the first layer where you include all numerical attributes as predictor variables (screenshot are NOT sufficient, as no equations are provided by RapidMiner):

Node 1: 1 / 1 + e^[-4.251 + (-0.632 * Temperature) + (-3.274 * Humidity) + (0.789 * Solar Radiation) + (-0.400 * Wind Speed) + (-4.026 * Rainfall) + (4.579 * Dew point temperature)]

Node 2: 1 / 1 + e^[-4.377 + (3.662 * Temperature) + (-6.648 * Humidity) + (-3.897 * Solar Radiation) + (3.463 * Wind Speed) + (0.489 * Rainfall) + (3.344 * Dew point temperature)]

Node 3: 1 / 1 + e^[-4.410 + (2.751 * Temperature) + (-5.014 * Humidity) + (-2.762 * Solar Radiation) + (2.295 * Wind Speed) + (0.884 * Rainfall) + (2.444 * Dew point temperature)]

	Training Performance	Test Performance
Root mean squared error	473.496 (+/- 0.000)	497.161 (+/- 0.000)
Absolute error	333.984 (+/- 335.639)	349.728 (+/- 353.354)
Relative error	118.26% (+/- 318.03%)	120.20% (+/- 357.06%)
Correlation	0.679	0.638
Squared correlation	0.462	0.408

i. Save the file of Question 1h as "A3_Q1i.rmp" in the process folder of your Local Repository. Repeat the process of Question 1h but do no include any speculative rounds. Leave everything else in the process the same as in Question 1h.

Equations for the nodes of the first layer where you include all numerical attributes as predictor variables (screenshot are NOT sufficient, as no equations are provided by RapidMiner):

Node 1: 1 / 1 + e^[-0.063 + (5.281 * Temperature) + (-2.051 * Humidity) + (1.036 * Solar Radiation)]

Node 2: 1 / 1 + e^[-4.874 + (4.412 * Temperature) + (-3.769 * Humidity) + (-2.492 * Solar Radiation)]

Node 3: 1 / 1 + e^[-4.383 + (3.453 * Temperature) + (-2.797 * Humidity) + (-1.966 * Solar Radiation)]

	Training Performance	Test Performance
Root mean squared error	485.417 (+/- 0.000)	509.077 (+/- 0.000)
Absolute error	345.547 (+/- 340.921)	361.158 (+/- 358.782)
Relative error	140.24% (+/- 461.42%)	142.04% (+/- 457.05%)
Correlation	0.659	0.615
Squared correlation	0.434	0.378

j. Save the file of Question 1h as "A3_Q1j.rmp" in the process folder of your Local Repository. Repeat the process of Question 1h but stop selecting new attributes when the squared correlation doesn't increase by at least 1%. Leave everything else in the process the same as in Question 1h.

Equations for the nodes of the first layer where you include all numerical attributes as predictor variables (screenshot are NOT sufficient, as no equations are provided by RapidMiner):

Node 1: 1 / 1 + e^[-0.063 + (5.281 * Temperature) + (-2.051 * Humidity) + (1.036 * Solar Radiation)]

Node 2: 1 / 1 + e^[-4.874 + (4.412 * Temperature) + (-3.769 * Humidity) + (-2.492 * Solar Radiation)]

Node 3: 1 / 1 + e^[-4.383 + (3.453 * Temperature) + (-2.797 * Humidity) + (-1.966 * Solar Radiation)]

	Training Performance	Test Performance
Root mean squared error	485.417 (+/- 0.000)	509.077 (+/- 0.000)
Absolute error	345.547 (+/- 340.921)	361.158 (+/- 358.782)
Relative error	140.24% (+/- 461.42%)	142.04% (+/- 457.05%)
Correlation	0.659	0.615
Squared correlation	0.434	0.378

k. Save the file of Question 1h as "A3_Q1k.rmp" in the process folder of your Local Repository. Repeat the process of Question 1h until there is no significant increase in performance (i.e., in the squared correlation), where the level of significance (i.e., alpha) is set to 5%, when including attributes in the attribute-selection process. Leave everything else in the process the same as in Question 1h.

Equations for the nodes of the first layer where you include all numerical attributes as predictor variables (screenshot are NOT sufficient, as no equations are provided by RapidMiner):

Node 1: 1 / 1 + e^[-4.155 + (-0.788 * Temperature) + (-2.877 * Humidity) + (0.914 * Solar Radiation) + (-0.401 * Wind Speed) + (-3.880 * Rainfall) + (4.386 * Dew point temperature) + (0.172 * Visibility)]

Node 2: 1 / 1 + e^[-4.392 + (3.948 * Temperature) + (-7.165 * Humidity) + (-4.015 * Solar Radiation) + (3.054 * Wind Speed) + (0.278 * Rainfall) + (3.194 * Dew point temperature) + (-0.954 * Visibility)]

Node 3: 1 / 1 + e^[-4.314 + (3.464 * Temperature) + (-6.287 * Humidity) + (-3.459 * Solar Radiation) + (2.463 * Wind Speed) + (0.457 * Rainfall) + (2.711 * Dew point temperature) + (-0.990 * Visibility)]

	Training Performance	**Test Performance**
Root mean squared error	472.263 (+/- 0.000)	497.783 (+/- 0.000)
Absolute error	337.168 (+/- 330.681)	354.127 (+/- 349.831)
Relative error	128.71% (+/- 353.46%)	130.70% (+/- 397.22%)
Correlation	0.683	0.639
Squared correlation	0.466	0.408

l. Compare the performance of the different feature selection methods.

Compare the performance of the different feature selection methods (parts Q1h until Q1k) to the performance of the neural network model without feature selection (part Q1g). Explain the impact of the different feature selection methods (if possible):

Comparison between Q1g and Q1h:

The model in Q1g performs better than Q1h regarding root mean squared error, slightly better with absolute error, worse in terms of relative error, and has a higher correlation and squared correlation for the training performance. In terms of the test performance the model in Q1g performed slightly better for root mean squared error, worse for absolute error and relative error, it had a minimally higher correlation and squared correlation.

Comparison between Q1g and Q1i:

The model in Q1g performs better than Q1i regarding root mean squared error, absolute error, and relative error, and had a higher correlation and squared correlation for the training performance. For the test performance the model in Q1g performed better in terms of the root mean squared error, absolute error, relative error, and had a higher correlation and squared correlation.

Comparison between Q1g and Q1j:

The model in Q1g performs better than Q1i regarding root mean squared error, absolute error, and relative error, and had a higher correlation and squared correlation for the training performance. For the test performance the model in Q1g performed better in terms of the root mean squared error, absolute error, relative error, and had a higher correlation and squared correlation. It had the same performance differences as between Q1g and Q1i.

Comparison between Q1g and Q1k:

The model in Q1g performs slightly better than Q1k regarding root mean squared error, absolute error, and relative error, and the same correlation and squared correlation for the training performance. For the test performance the model in Q1g performed slightly better in terms of the root mean squared error, absolute error, relative error, had the same correlation and slightly higher squared correlation.

m. Save the file of Question 1g as "A3_Q1k.rmp" in the process folder of your Local Repository. Include a proper implementation of Principal Components Analysis (PCA) to make sure that the assumption of independence between the predictor variables is verified. Be careful where to locate the PCA operator (i.e., which dataset(s) should be used for this process?).

Cumulative variance for each of the principal components:

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.559	0.304	0.304
PC 2	1.393	0.242	0.546
PC 3	1.032	0.133	0.680
PC 4	0.963	0.116	0.796
PC 5	0.875	0.096	0.891
PC 6	0.803	0.081	0.972
PC 7	0.470	0.028	0.999
PC 8	0.066	0.001	1.000

What can be concluded from the cumulative variances?

As the Cumulative variance increases with each new variable (lowest at pc1 and highest at pc8) the standard deviation decreases (highest at pc1 and lowest at pc8) and Proportion of Variance also decreases (highest at pc1 and lowest at pc8).

Equations for the nodes of the first layer where you include all numerical attributes as predictor variables (screenshot are NOT sufficient, as no equations are provided by RapidMiner):

Node 1: 1 / 1 + e^[-1.029 + (0.386 * pc_1) + (-2.099 * pc_2) + (-3.847 * pc_3) + (4.555 * pc_4) + (-1.500 * pc_5) + (1.279 * pc_6) + (-0.843 * pc_7) + (-2.761 * pc_8)]

Node 2: 1 / 1 + e^[-3.792 + (1.606 * pc_1) + (-0.718 * pc_2) + (-2.638 * pc_3) + (0.819 * pc_4) + (-1.327 * pc_5) + (2.505 * pc_6) + (-6.149 * pc_7) + (-2.350 * pc_8)]

Node 3: 1 / 1 + e^[0.544 + (1.021 * pc_1) + (-3.411 * pc_2) + (-0.923 * pc_3) + (2.498 * pc_4) + (-2.435 * pc_5) + (-1.340 * pc_6) + (2.185 * pc_7) + (0.147 * pc_8)]

	Training Performance	Test Performance
Root mean squared error	437.446 (+/- 0.000)	474.102 (+/- 0.000)
Absolute error	306.334 (+/- 312.281)	330.676 (+/- 339.744)
Relative error	94.21% (+/- 251.94%)	97.68% (+/- 297.80%)
Correlation	0.735	0.680
Squared correlation	0.540	0.462

Compare the performance of the neural network model based on principal components (part Q1m) to the performance of the neural network model without principal components (part Q1g):

For both the training and test datasets, the model created in Q1m performed better than the model from Q1g with regard to the root mean squared error, absolute error, and relative error. Furthermore, the Q1m model had a higher correlation and squared correlation compared to the Q1g model across both the training and test datasets.

QUESTION 2 – CHURN

For this question, split the data where 75% is used for training, and 25% is used for test data. Use shuffled sampling with local random seed 1992. Note that we will be using a validation dataset later in the question.

a. Create a logistic regression model with the L_BFGS solver, where you include all attributes except the customer's phone number and residing state. Any examples with missing attribute values should be removed. Save your RapidMiner file as "A3_Q2a.rmp" in the processes folder of your Local Repository.

Regression model – where you eliminate collinear variables (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -8.642 + 0.057 (Area Code = A415) – 0.029 (Area Code = A510) + 1.831 (Inter Plan = yes) – 2.602 (VoiceMail Plan = yes) + 0.002 (Account Length) + 0.053 (No of Vmail Mesgs) + 0.007 (Total Day Min) + 0.004 (Total Day Calls) + 0.039 (Total Day Charge) + 0.004 (Total Evening Min) + 0.002 (Total Evening Calls) + 0.043 (Total Evening Charge) + 0.001 (Total Night Min) + 0.001 (Total Night Calls) + 0.033 (Total Night Charge) + 0.033 (Total Int Min) – 0.080 (Total Int Calls) + 0.133 (Total Int Charge) + 0.459 (No of Calls Customer Service)

	Training Performance	**Test Performance**
confusion matrix		

	False	True
Pred. False	2090	294
Pred. True	48	69

	False	True
Pred. False	702	93
Pred. True	11	27

accuracy	86.33%	87.52%
sensitivity (or recall)	19.01%	22.50%
specificity	97.75%	98.46%
lift	406.32%	493.22%
AUC	0.813	0.836

b. Note that the *Logistic Regression* operator in RapidMiner also optimizes a cut-off threshold value, given by the thr output port from this operator. Save the file of Question 2a as "A3_Q2c.rmp" in the process folder of your Local Repository.

	Training Performance	**Test Performance**
confusion matrix		

	False	True

	False	True
Pred. False	621	40

	Pred. False | 1811 | 143 Pred. True | 327 | 220	Pred. True | 92 | 80
accuracy	81.21%	84.15%
sensitivity (or recall)	60.61%	66.67%
specificity	84.71%	87.10%
lift	277.10%	322.87%
AUC	0.813	0.836

Threshold value: 0.2133294307213244

Explanation of the new performance:
Due to the inclusion of the threshold, the model from question 2b performed worse with respect to the accuracy, lift, and specificity when compared to the model in question 2a for the training dataset. However, the sensitivity was higher for the model in 2b, and the AUC was the same in both models. Likewise, in terms of the test performance, the model in 2a performed better regarding accuracy, specificity, and lift, while 2b's model performed better in regards to sensitivity, and once again the AUC was the same in both models.

The new performance can be explained by the fact that since we now have a lower threshold (i.e., compared to the default threshold of 0.5), this suggests that a higher propensity score is required for any given record to be classified as "true". As a result, the model is now more selective with regards to which records receive the positive classification. This is supported by the fact that the model from question 2b has a higher sensitivity score when compared to the model from 2a, which means that the model from 2b is able to predict "true" classifications with higher precision.

c. Save the file of Question 2a as "A3_Q2c.rmp" in the process folder of your Local Repository. Update the file to include backward elimination as feature selection method. Stop eliminating when no more improvements (in terms of increasing the accuracy) can be made. Include 5 speculative rounds once terminated. For the backward elimination process itself, split the data in 40% training and 60% validation dataset with shuffled sampling and local random seed 1992.

Regression model – where you eliminate collinear variables and use backward elimination (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -6.722 – 0.055 (Area Code = A510) + 1.680 (Inter Plan = yes) – 2.549 (VoiceMail Plan = yes) + 0.002 (Account Length) + 0.036 (Total Day Charge) + 0.039 (Total Evening Charge) + 0.002 (Total Night Calls) + 0.028 (Total Night Charge) – 0.074 (Total Int Calls) + 0.119 (Total Int Charge) + 0.006 (Total Day Min) + 0.003 (Total Evening Min) + 0.030 (Total Int Min) + 0.001 (Total Night Min) + 0.054 (No of Vmail Mesgs)

<table>
<tr><th></th><th>Training Performance</th><th>Test Performance</th></tr>
<tr><td>confusion matrix</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>2103</td><td>317</td></tr>
<tr><td>Pred. True</td><td>35</td><td>46</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>704</td><td>100</td></tr>
<tr><td>Pred. True</td><td>9</td><td>20</td></tr>
</table>
</td></tr>
<tr><td>accuracy</td><td>85.93%</td><td>86.91%</td></tr>
<tr><td>sensitivity (or recall)</td><td>12.67%</td><td>16.67%</td></tr>
<tr><td>specificity</td><td>98.36%</td><td>98.74%</td></tr>
<tr><td>lift</td><td>391.27%</td><td>478.74%</td></tr>
<tr><td>AUC</td><td>0.746</td><td>0.743</td></tr>
</table>

d. Save the file of Question 2c as "A3_Q2d.rmp" in the process folder of your Local Repository. Repeat the process of Question 2c but do no include any speculative rounds. Leave everything else in the process the same as in Question 2c.

Regression model – where you eliminate collinear variables and use backward elimination (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -7.204 + 0.044 (Area Code = A415) – 0.021 (Area Code = A510) + 1.681 (Inter Plan = yes) – 2.553 (VoiceMail Plan = yes) + 0.002 (Account Length) + 0.054 (No of Vmail Mesgs) + 0.006 (Total Day Min) + 0.002 (Total Day Calls) + 0.036 (Total Day Charge) + 0.003 (Total Evening Min) + 0.002 (Total Evening Calls) + 0.039 (Total Evening Charge) + 0.001 (Total Night Min) + 0.002 (Total Night Calls) + 0.028 (Total Night Charge) + 0.030 (Total Int Min) – 0.074 (Total Int Calls) + 0.118 (Total Int Charge)

<table>
<tr><th></th><th>Training Performance</th><th>Test Performance</th></tr>
<tr><td>confusion matrix</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>2103</td><td>317</td></tr>
<tr><td>Pred. True</td><td>35</td><td>46</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>702</td><td>101</td></tr>
<tr><td>Pred. True</td><td>11</td><td>19</td></tr>
</table>
</td></tr>
<tr><td>accuracy</td><td>85.93%</td><td>86.55%</td></tr>
<tr><td>sensitivity (or recall)</td><td>12.67%</td><td>15.83%</td></tr>
<tr><td>specificity</td><td>98.36%</td><td>98.46%</td></tr>
<tr><td>lift</td><td>391.27%</td><td>439.64%</td></tr>
<tr><td>AUC</td><td>0.747</td><td>0.743</td></tr>
</table>

e. Save the file of Question 2c as "A3_Q2e.rmp" in the process folder of your Local Repository. Update the stopping behavior such that you allow a relative decrease in the accuracy up to 10%

when eliminating attributes in the attribute-selection process. Leave everything else in the process the same as in Question 2c.

Regression model – where you eliminate collinear variables and use backward elimination (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -3.989 + 0.068 (Total Day Charge)

	Training Performance	**Test Performance**
confusion matrix		

	False	True
Pred. False	2138	361
Pred. True	0	2

	False	True
Pred. False	713	120
Pred. True	0	0

accuracy	85.57%	85.59%
sensitivity (or recall)	0.55%	0.00%
specificity	100.00%	100.00%
lift	688.98%	Unknown
AUC	0.641	0.635

f. Save the file of Question 2c as "A3_Q2f.rmp" in the process folder of your Local Repository. Update the file to include forward selection as feature selection method instead of backward elimination. Stop selecting new attributes when no more improvements (in terms of increasing the accuracy) can be made. Include 5 speculative rounds once terminated. For the forward selection process itself, split the data in 40% training and 60% validation dataset with shuffled sampling and local random seed 1992.

Regression model – where you eliminate collinear variables and use forward selection (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -5.262 + 0.012 (Total Day Min) – 0.832 (VoiceMail Plan = yes) + 0.006 (Total Evening Min) + 0.002 (Total Night Calls) – 0.065 (Total Int Calls) + 0.037 (Area Code = A415) + 0.002 (Total Evening Calls)

	Training Performance	**Test Performance**
confusion matrix		

	False	True
Pred. False	2138	349
Pred. True	0	14

	False	True
Pred. False	713	115
Pred. True	0	5

accuracy	86.05%	86.19%
sensitivity (or recall)	3.86%	4.17%
specificity	100.00%	100.00%
lift	688.98%	694.17%
AUC	0.672	0.641

g. Save the file of Question 2f as "A3_Q2g.rmp" in the process folder of your Local Repository. Repeat the process of Question 2f but do no include any speculative rounds. Leave everything else in the process the same as in Question 2f.

Regression model – where you eliminate collinear variables and use forward selection (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -5.069 + 0.012 (Total Day Min) – 0.831 (VoiceMail Plan = yes) + 0.006 (Total Evening Min) + 0.002 (Total Night Calls) – 0.064 (Total Int Calls)

<table>
<tr><th></th><th>Training Performance</th><th>Test Performance</th></tr>
<tr><td>confusion matrix</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>2138</td><td>350</td></tr>
<tr><td>Pred. True</td><td>0</td><td>13</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>713</td><td>116</td></tr>
<tr><td>Pred. True</td><td>0</td><td>4</td></tr>
</table>
</td></tr>
<tr><td>accuracy</td><td>86.01%</td><td>86.07%</td></tr>
<tr><td>sensitivity (or recall)</td><td>3.58%</td><td>3.33%</td></tr>
<tr><td>specificity</td><td>100.00%</td><td>100.00%</td></tr>
<tr><td>lift</td><td>688.98%</td><td>694.17%</td></tr>
<tr><td>AUC</td><td>0.672</td><td>0.642</td></tr>
</table>

h. Save the file of Question 2f as "A3_Q2h.rmp" in the process folder of your Local Repository. Update the stopping behavior such that you require a relative increase in the accuracy of at least 10% when selecting attributes to include in the regression model. Leave everything else in the process the same as in Question 2f.

Regression model – where you eliminate collinear variables and use forward selection (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -3.989 + 0.012 (Total Day Min)

	Training Performance	Test Performance		
confusion matrix			False	True

		False	True	Pred. False	713	120
	Pred. False	2138	361	Pred. True	0	0
	Pred. True	0	2			
accuracy	85.57%			85.59%		
sensitivity (or recall)	0.55%			0.00%		
specificity	100.00%			100.00%		
lift	688.98%			Unknown		
AUC	0.641			0.635		

In the remain questions, we will ONLY consider the numerical attributes as predictor variables.

i. Save the file of Question 2c as "A3_Q2i.rmp" in the process folder of your Local Repository (i.e., with the same backward elimination process). Update the file such that only the numerical attributes are considered.

Report the correlation matrix and covariance matrix for the attributes of your training dataset (screenshots are sufficient, make sure it is readable when printed):

Correlation Matrix:

Attributes	No of Vmail Mesgs	Total Day Charge	Total Evening Charge	Total Night Calls	Total Night Charge	Total Int Min	Total Int Calls	Total Int Charge	Total Day Min	Total Evening Min	Total Evening Calls	Account Length	Total Night Min
No of Vmail Mesgs	1	-0.004	0.002	0.011	0.022	0.011	0.019	0.011	-0.004	0.002	-0.006	-0.014	0.022
Total Day Charge	-0.004	1	-0.006	0.024	0.008	-0.013	0.005	-0.013	1.000	-0.006	-0.002	0.027	0.009
Total Evening Charge	0.002	-0.006	1	0.009	-0.003	-0.019	-0.017	-0.019	-0.006	1.000	-0.010	-0.018	-0.003
Total Night Calls	0.011	0.024	0.009	1	0.023	-0.019	-0.000	-0.019	0.024	0.009	0.001	0.003	0.023
Total Night Charge	0.022	0.008	-0.003	0.023	1	-0.002	0.008	-0.002	0.008	-0.003	-0.002	-0.005	1.000
Total Int Min	0.011	-0.013	-0.019	-0.019	-0.002	1	0.050	1.000	-0.013	-0.019	0.011	0.019	-0.002
Total Int Calls	0.019	0.005	-0.017	-0.000	0.008	0.050	1	0.050	0.005	-0.017	0.028	0.009	0.008
Total Int Charge	0.011	-0.013	-0.019	-0.019	-0.002	1.000	0.050	1	-0.013	-0.019	0.011	0.019	-0.002
Total Day Min	-0.004	1.000	-0.006	0.024	0.008	-0.013	0.005	-0.013	1	-0.006	-0.002	0.027	0.009
Total Evening Min	0.002	-0.006	1.000	0.009	-0.003	-0.019	-0.017	-0.019	-0.006	1	-0.010	-0.018	-0.003
Total Evening Calls	-0.006	-0.002	-0.010	0.001	-0.002	0.011	0.028	0.011	-0.002	-0.010	1	0.026	-0.002
Account Length	-0.014	0.027	-0.018	0.003	-0.005	0.019	0.009	0.019	0.027	-0.018	0.026	1	-0.005
Total Night Min	0.022	0.009	-0.003	0.023	1.000	-0.002	0.008	-0.002	0.009	-0.003	-0.002	-0.005	1

Covariance Matrix:

Attributes	No of Vmail Mesgs	Total Day Charge	Total Evening Charge	Total Night Calls	Total Night Charge	Total Int Min	Total Int Calls	Total Int Charge	Total Day Min	Total Evening Min	Total Evening Calls	Account Length	Total Night Min
No of Vmail Mesgs	188.598	-0.539	0.093	3.056	0.677	0.419	0.632	0.113	-3.171	1.090	-1.746	-7.597	15.067
Total Day Charge	-0.539	85.858	-0.235	4.304	0.179	-0.345	0.121	-0.093	505.048	-2.758	-0.420	10.132	3.991
Total Evening Charge	0.093	-0.235	18.549	0.773	-0.033	-0.233	-0.179	-0.063	-1.384	218.224	-0.816	-3.107	-0.741
Total Night Calls	3.056	4.304	0.773	378.046	1.030	-1.031	-0.010	-0.278	25.315	9.084	0.292	2.295	22.922
Total Night Charge	0.677	0.179	-0.033	1.030	5.188	-0.016	0.042	-0.004	1.052	-0.390	-0.105	-0.464	115.295
Total Int Min	0.419	-0.345	-0.233	-1.031	-0.016	7.787	0.341	2.103	-2.030	-2.745	0.614	2.168	-0.349
Total Int Calls	0.632	0.121	-0.179	-0.010	0.042	0.341	5.870	0.092	0.710	-2.112	1.347	0.865	0.928
Total Int Charge	0.113	-0.093	-0.063	-0.278	-0.004	2.103	0.092	0.568	-0.545	-0.741	0.166	0.586	-0.092
Total Day Min	-3.171	505.048	-1.384	25.315	1.052	-2.030	0.710	-0.545	2970.862	-16.244	-2.477	59.601	23.464
Total Evening Min	1.090	-2.758	218.224	9.084	-0.390	-2.745	-2.112	-0.741	-16.244	2567.354	-9.608	-36.580	-8.695
Total Evening Calls	-1.746	-0.420	-0.816	0.292	-0.105	0.614	1.347	0.166	-2.477	-9.608	394.389	20.305	-2.357
Account Length	-7.597	10.132	-3.107	2.295	-0.464	2.168	0.865	0.586	59.601	-36.580	20.305	1594.585	-10.310
Total Night Min	15.067	3.991	-0.741	22.922	115.295	-0.349	0.928	-0.092	23.464	-8.695	-2.357	-10.310	2562.043

What can be concluded from these two matrices?

Correlation Matrix:

There is a number of highly correlated variables as highlighted above. For example, "Total Day Min" and "Total Day Charge" are perfectly positively correlated. The same relationship can be observed between the variables "Total Evening Min" and "Total Evening Charge", along with "Total Night Min" and "Total Night Charge". All of these should be removed because highly correlated models make it hard for the model to achieve optimal accuracy.

Covariance Matrix:
The highest covariance apparent in the matrix seems to be certain variables in comparison to themselves, including "Total Day Min", "Total Evening Min", "Total Evening Calls", "Account Length", and "Total Night Min". However, a relatively large covariance also exists between "Total Day Min" and "Total Day Charge", along with "Total Evening Min" and "Total Evening Charge".

Regression model – where you eliminate collinear variables (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -6.593 – 0.024 (No of Vmail Mesgs) + 0.035 (Total Day Charge) + 0.038 (Total Evening Charge) + 0.002 (Total Night Calls) + 0.021 (Total Night Charge) + 0.034 (Total Int Min) – 0.071 (Total Int Calls) + 0.125 (Total Int Charge) + 0.006 (Total Day Min) + 0.003 (Total Evening Min) + 0.002 (Total Evening Calls) + 0.002 (Account Length) + 0.001 (Total Night Min)

<table>
<tr><th></th><th>Training Performance</th><th>Test Performance</th></tr>
<tr><td>confusion matrix</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>2138</td><td>346</td></tr>
<tr><td>Pred. True</td><td>0</td><td>17</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>713</td><td>112</td></tr>
<tr><td>Pred. True</td><td>0</td><td>8</td></tr>
</table>
</td></tr>
<tr><td>accuracy</td><td>86.17%</td><td>86.55%</td></tr>
<tr><td>sensitivity (or recall)</td><td>4.68%</td><td>6.67%</td></tr>
<tr><td>specificity</td><td>100.00%</td><td>100.00%</td></tr>
<tr><td>lift</td><td>688.98%</td><td>694.17%</td></tr>
<tr><td>AUC</td><td>0.678</td><td>0.660</td></tr>
</table>

j. Save the file of Question 3i as "A3_Q2j.rmp" in the process folder of your Local Repository. Include a proper implementation of Principal Components Analysis (PCA) after backward elimination is performed to make sure that the assumption of independence between the predictor variables is verified.

Report the correlation matrix and covariance matrix for the attributes of your training dataset after applying PCA (screenshots are sufficient, make sure it is readable when printed):
Correlation Matrix:

Attributes	pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7	pc_8	pc_9	pc_10	pc_11	pc_12	pc_13
pc_1	1	0.000	-0	0.000	0.000	-0	-0.000	-0	-0	-0.000	0.000	-0.000	0.000
pc_2	0.000	1	0.000	0	-0.000	0	-0.000	-0	-0.000	0.000	0.000	-0.000	-0.000
pc_3	-0	0.000	1	-0	-0.000	-0	-0.000	0	-0.000	0.000	-0.000	-0.000	-0.000
pc_4	0.000	0	-0	1	-0.000	-0	-0	-0.000	0.000	0.000	0.000	-0.000	0.000
pc_5	0.000	-0.000	-0.000	-0.000	1	-0.000	-0.000	-0.000	-0.000	0.000	-0.000	-0.000	0.000
pc_6	-0	0	-0	-0	-0.000	1	0.000	-0.000	0.000	0.000	0.000	0.000	0.000
pc_7	-0.000	-0.000	-0.000	-0	-0.000	0.000	1	0.000	-0.000	0.000	0.000	0.000	0.000
pc_8	-0	-0	0	-0.000	-0.000	-0.000	0.000	1	-0.000	-0.000	0.000	0.000	0.000
pc_9	-0	-0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000	1	-0.000	-0.000	-0.000	0.000
pc_10	-0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.000	-0.000	1	0.000	-0.000	-0.000
pc_11	0.000	0.000	-0.000	0.000	-0.000	0.000	0.000	0.000	-0.000	0.000	1	-0.000	-0.000
pc_12	-0.000	-0.000	-0.000	-0.000	-0.000	0.000	0.000	0.000	-0.000	-0.000	-0.000	1	-0.000
pc_13	0.000	-0.000	-0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.000	-0.000	-0.000	1

Covariance Matrix:

Attributes	pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7	pc_8	pc_9	pc_10	pc_11	pc_12	pc_13
pc_1	2.049	0.000	-0.000	0.000	0.000	-0	-0.000	-0.000	-0.000	-0.000	0	-0.000	0.000
pc_2	0.000	2.027	0.000	0	-0.000	0.000	-0.000	-0	-0.000	0.000	0.000	-0.000	-0.000
pc_3	-0.000	0.000	1.988	-0	-0.000	-0	-0.000	0	-0.000	0.000	-0.000	-0.000	-0.000
pc_4	0.000	0	-0	1.949	-0.000	-0	-0	-0.000	0.000	0.000	0	-0.000	0.000
pc_5	0.000	-0.000	-0.000	-0.000	1.037	-0.000	-0.000	-0.000	-0.000	0	-0	-0.000	0
pc_6	-0	0.000	-0	-0	-0.000	1.020	0.000	-0.000	0.000	0	0	0	0
pc_7	-0.000	-0.000	-0.000	-0	-0.000	0.000	0.996	0.000	-0.000	0	0	0	0
pc_8	-0.000	-0	0	-0.000	-0.000	-0.000	0.000	0.972	-0.000	-0.000	0	0	0.000
pc_9	-0.000	-0.000	-0.000	0.000	-0.000	0.000	-0.000	-0.000	0.961	-0	-0	-0.000	0
pc_10	-0.000	0.000	0.000	0.000	0	0	0	-0.000	-0	0.000	0	-0	-0
pc_11	0	0.000	-0.000	0	-0	0	0	0	-0	0	0.000	-0	-0
pc_12	-0.000	-0.000	-0.000	-0.000	-0.000	0	0	0	-0.000	-0	-0	0.000	-0
pc_13	0.000	-0.000	-0.000	0.000	0	0	0	0.000	0	-0	-0	-0	0.000

What can be concluded from these two matrices?

Due to the inclusion of the PCA operator there is minimal correlation between different variables, and also very low covariance between the different variables. There is relatively high covariance when comparing the same variable with itself, however the covariance decreases with each successive variable (highest at pc1 and lowest at the pc8).

Cumulative variance for each of the principal components:

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.432	0.158	0.158
PC 2	1.424	0.156	0.314
PC 3	1.410	0.153	0.467
PC 4	1.396	0.150	0.616
PC 5	1.019	0.080	0.696
PC 6	1.010	0.078	0.775
PC 7	0.998	0.077	0.851
PC 8	0.986	0.075	0.926
PC 9	0.980	0.074	1.000
PC 10	0.003	0.000	1.000
PC 11	0.001	0.000	1.000
PC 12	0.000	0.000	1.000
PC 13	0.000	0.000	1.000

What can be concluded from the cumulative variances?

As the cumulative variance increases with each new variable (lowest at pc1 and highest at pc9) the standard deviation decreases (highest at pc1 and lowest at pc12) and Proportion of Variance also decreases (highest at pc1 and lowest at pc10). The cumulative variance is the same from pc9 onwards, while the standard deviation for pc12 and pc13 is the same, and the proportion of variance is likewise identical from pc10 onwards.

Regression model – where you eliminate collinear variables (screenshot is NOT sufficient, as no equations are provided by RapidMiner):

Logit(Churn = true) = -2.005 + 0.193 (pc_1) – 0.228 (pc_2) – 0.112 (pc_3) + 0.428 (pc_4) – 0.004 (pc_5) + 0.351 (pc_6) + 0.128 (pc_7) – 0.101 (pc_8) + 0.000 (pc_9) + 4.212 (pc_10) – 5.337 (pc_11) + 68.158 (pc_12) + 34.484 (pc_13)

<table>
<tr><th></th><th>Training Performance</th><th>Test Performance</th></tr>
<tr><td>confusion matrix</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>2138</td><td>347</td></tr>
<tr><td>Pred. True</td><td>0</td><td>16</td></tr>
</table>
</td><td>
<table>
<tr><td></td><td>False</td><td>True</td></tr>
<tr><td>Pred. False</td><td>713</td><td>113</td></tr>
<tr><td>Pred. True</td><td>0</td><td>7</td></tr>
</table>
</td></tr>
<tr><td>accuracy</td><td>86.13%</td><td>86.43%</td></tr>
<tr><td>sensitivity (or recall)</td><td>4.41%</td><td>5.83%</td></tr>
<tr><td>specificity</td><td>100.00%</td><td>100.00%</td></tr>
<tr><td>lift</td><td>688.98%</td><td>694.17%</td></tr>
<tr><td>AUC</td><td>0.679</td><td>0.659</td></tr>
</table>

k. Let's say that you performed a process similar to the process that you created in part Q2j. Use the information from the assignment to find the propensity score for the new example to be classified as true.

Include your calculations here. If you use Excel, make sure to include the Excel file in your zip folder and to include (multiple) screenshots of your Excel calculations/numbers in this report.

2K) Features selected:

No. of voicemails → PC2
Total Day Min → PC1
Total Evening Min → PC3
Total Night Calls → P4
Total Int calls → PS

Normalization of example set $\left(\frac{x-\mu}{\beta}\right)$

PC1: $\frac{184.5 - 180.9992}{\sqrt{2970.8619}} = 0.0642$

PC2: $\frac{0 - 8.2843}{\sqrt{183.5979}} = -0.6032$

PC3 = $\frac{351.6 - 201.3928}{\sqrt{2567.3535}} = 2.9645$

PC4 = $\frac{90 - 100.1719}{\sqrt{378.0456}} = -0.5232$

PC5 = $\frac{4 - 4.4578}{\sqrt{5.8694}} = -0.1890$

Odds:

PC1 = $e^{0.044} = 1.044982$
PC2 = $e^{-0.562} = 0.570068$
PC3 = $e^{-0.363} = 0.695586$
P4 = $e^{-0.343} = 0.709638$
PS = $e^{0.265} = 1.303431$

$$\text{Odds(true)} = e^{-1.982} \times 1.044982^{PC1} \times 0.570068^{PC2} \times 0.695586^{PC3} \times 0.709638^{P4} \times 1.303431^{PC5}$$

$$= 0.137656 \times 1.044982^{0.0642} \times 0.570068^{(-0.6032)} \times 0.695586^{(2.9645)} \times 0.709638^{(-0.5232)} \times 1.303431^{-0.1890}$$

$$= 0.07517827$$

$$P = \frac{odds}{1+odds}$$

$$P = \frac{0.07517827}{1+0.07517827} = 0.0699$$

P(true) = 6.99.1.