Instructions for candidates:

1. Answer all exam questions.
2. This exam consists of 120 marks which contribute to 50% of the total assessment in this unit.
3. This unit has a hurdle requirement. You need to achieve at least 50% of the marks available on the examination.
4. It is expected that you spend approximately 1 minute of examination time to gain 1 mark.

Business Scenario: Travelairex

A company publishing reviews of airlines would like to change its current system of collecting passenger in-flight experiences and their analysis. In the current form, they collect information about the passenger (name, country and type of traveller) and the completed air travel (airline, aircraft, cabin class, route and date). They ask passengers to fill in a brief questionnaire rating their experience in terms of quality of cabin staff, food and beverages, ground services, inflight entertainment, seat comfort, Wi-Fi connectivity and the value for money (all as 0..5), the overall rating (1..10) and airline recommendation (0..1). Passengers also provide a free text description of their travel experience.

Objectives

To get more accurate analysis of the air travel experience, the company employed you to develop a set of predictive models to analyse the collected data and most importantly rely on the text of included reviews, which includes many unanticipated insights of relevance to the customer experience and airline recommendations.

	airline_name	link	title	author	author_country	content	date	aircraft	type_traveller	cabin_flown	route	overall	seat_cabin_	food_inflight	ground_	wifi_c	value_r	recommended
1	airline_name	link	title	author	author_country	content	date	aircraft	type_traveller	cabin_flown	route	overall	seat_cabin_	food_inflight	ground_	wifi_c	value_r	recommended
50	adria-airways	/airline-re	Adria Airv	David Grey		Round trip LGW - LJU in Economy Class bc	15/09/2005											0
51	adria-airways	/airline-re	Adria Airv	Markus Lattner		I have used their VIE-FRA-VIE service sev	3/12/2004											0
52	adria-airways	/airline-re	Adria Airv	Alex Stare		Flew march 12/14 on JP from ZRH-LJU-ZRI	18/03/2004											0
53	aegean-airlines	/airline-re	Aegean Al	N Sunder	United States	Brand new A320, interior spotless. Crew	21/07/2015	A320	Solo Leisure	Economy	Mykonos to	8	4	4			4	1
54	aegean-airlines	/airline-re	Aegean Al	Dimitrios	Greece	I booked a premium economy ticket so I	21/07/2015	Dash 8 Q4	Solo Leisure	Economy	Athens to S	8	4	5	4		2	1
55	aegean-airlines	/airline-re	Aegean Al	Kim Buhrr	United States	2 flights - HER to ATH (short 50 minutes) a	12/07/2015		FamilyLeisure	Economy	ATH to MXP	10	4	5	4		5	3
56	adria-airways	/airline-re	Adria Airv	Alan McLean		Round trip LGW - LJU in Business Class. Cl	1/11/2003											1
57	adria-airways	/airline-re	Adria Airv	Andrew Proud		In July 2003 I flew London (Gatwick)-Ljub	31/07/2003											0
58	adria-airways	/airline-re	Adria Airv	G Chew		I have just returned from a long weekend	6/12/2002											0
59	adria-airways	/airline-re	Adria Airv	Francisco Camino		I have flown Adria recently on a LH code	25/08/2002											0

Figure 1 (Part A): Data

Data

The provided data includes 41,396 reviews of airline passengers, which includes information about them (e.g. their name), their flight (such as date, airlines, aircraft and country of their flight), the text of their review (in plain English), and their rating of the flight's various aspects (such as seat comfort, cabin staff, food and beverages, inflight entertainment, ground services, WIFI connectivity, value for money, the overall rating of the flight and the passenger recommendation of the airline).

Airline dataset total samples: 41396

Type and the number of unique values per attribute:

- (nominal) airline_name: 41396
- (nominal) link: 41396
- (nominal) title: 41396

- (nominal) author: 41396
- (nominal) author_country: 39805
- (date) date: 41396
- (text) content: 41396
- (nominal) aircraft: 1278
- (nominal) type_traveller: 2378
- (nominal) cabin_flown: 38520
- (nominal) route: 2341
- (integer) overall_rating: 36861
- (integer) seat_comfort_rating: 33706
- (integer) cabin_staff_rating: 33708
- (integer) food_beverages_rating: 33264
- (integer) inflight_entertainment_rating: 31114
- (integer) ground_service_rating: 2203
- (integer) wifi_connectivity_rating: 565
- (integer) value_money_rating: 39723
- (binomial) recommended: 41396

The final recommendation is one of the possible labels. However, any of the ratings could also be predicted from both the structured and unstructured data of the customer feedback.

Name	Type	Missing
Cluster airline_name	Polynominal	0
overall_rating	Real	0
seat_comfort_rating	Real	0
cabin_staff_rating	Real	0
food_beverages_rating	Real	15
inflight_entertainment_rating	Real	0
ground_service_rating	Real	9736
wifi_connectivity_rating	Real	10111
value_money_rating	Real	0

Figure 1 (Part B): Missing values in the attributes of the selected survey ratings

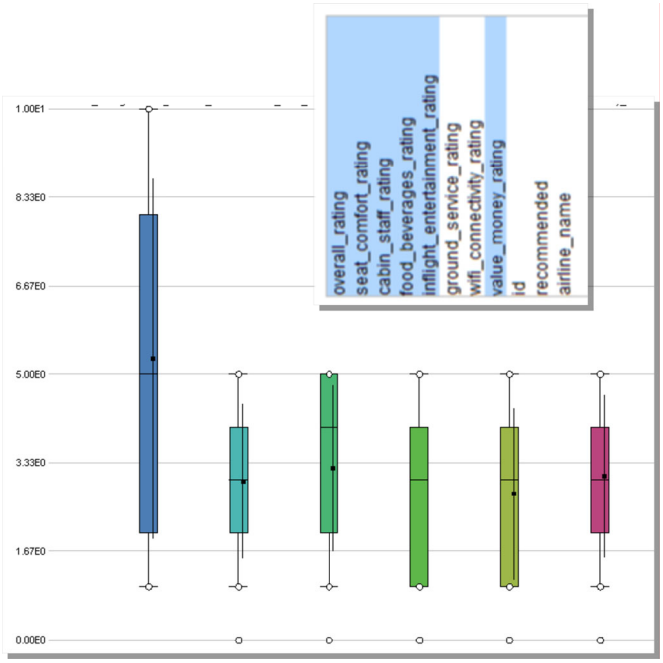

Figure 1 (Part C): Comparison of the distribution of the selected survey ratings

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
type_traveller = FamilyLeisure	-0.116	0.069	-0.005	0.997	-1.688	0.091	*
type_traveller = Couple Leisure	-0.074	0.070	-0.003	0.997	-1.060	0.289	
type_traveller = Solo Leisure	-0.163	0.061	-0.007	0.999	-2.669	0.008	***
type_traveller = Business	-0.214	0.089	-0.006	0.998	-2.391	0.017	**
cabin_flown = Economy	0.965	0.040	0.135	0.995	24.336	0	****
cabin_flown = Business Class	0.956	0.045	0.111	0.990	21.453	0	****
cabin_flown = Premium Economy	0.893	0.059	0.054	0.998	15.243	0	****
cabin_flown = First Class	1.029	0.069	0.048	1.000	14.912	0	****
seat_comfort_rating	0.319	0.010	0.124	0.483	31.804	0	****
cabin_staff_rating	0.549	0.010	0.233	0.468	57.215	0	****
food_beverages_rating	0.135	0.009	0.058	0.589	15.576	0	****
inflight_entertainment_rating	0.056	0.007	0.026	0.826	8.369	0.000	****
ground_service_rating	0.205	0.024	0.025	0.943	8.457	0	****
wifi_connectivity_rating	-0.076	0.047	-0.005	0.990	-1.623	0.105	
value_money_rating	1.179	0.009	0.524	0.430	128.238	0	****
date	-0.000	0.000	-0.058	1.000	-19.530	0	****
(Intercept)	2.557	0.278	?	?	9.185	0	****

Figure 2 (Part A) Table of coefficients

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 1.641 +/- 0.015 (micro average: 1.641 +/- 0.000)
absolute_error: 1.186 +/- 0.009 (micro average: 1.186 +/- 1.135)
correlation: 0.860 +/- 0.002 (micro average: 0.860)
squared_correlation: 0.739 +/- 0.004 (micro average: 0.739)
prediction_average: 6.038 +/- 0.014 (micro average: 6.038 +/- 3.212)
```

Figure 2 (Part B): Performance vector

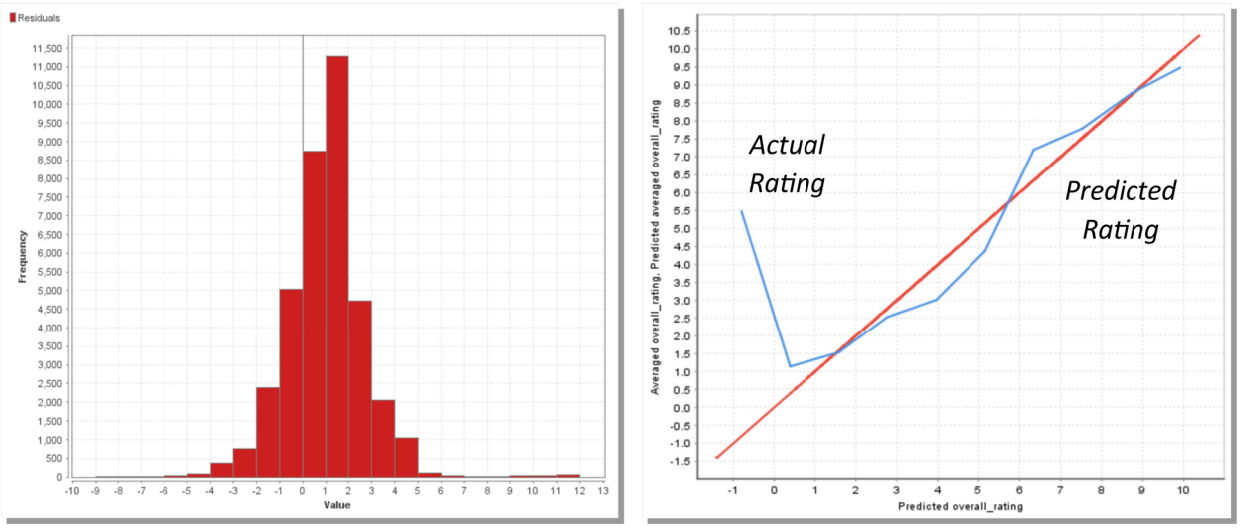

Figure 2 (Part C): Residuals (left) and Actuals (blue) vs prediction (red) (right)

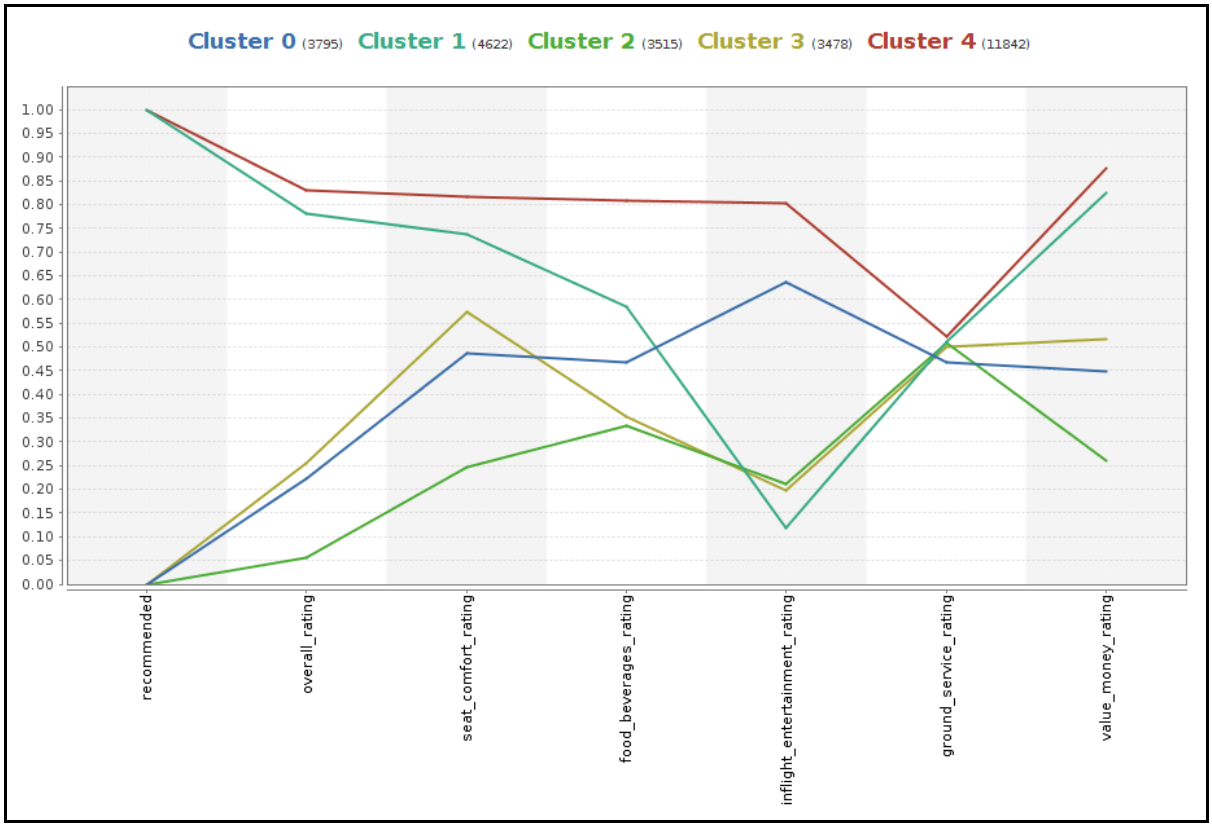

Figure 3 (Part A): k-Means centroid chart

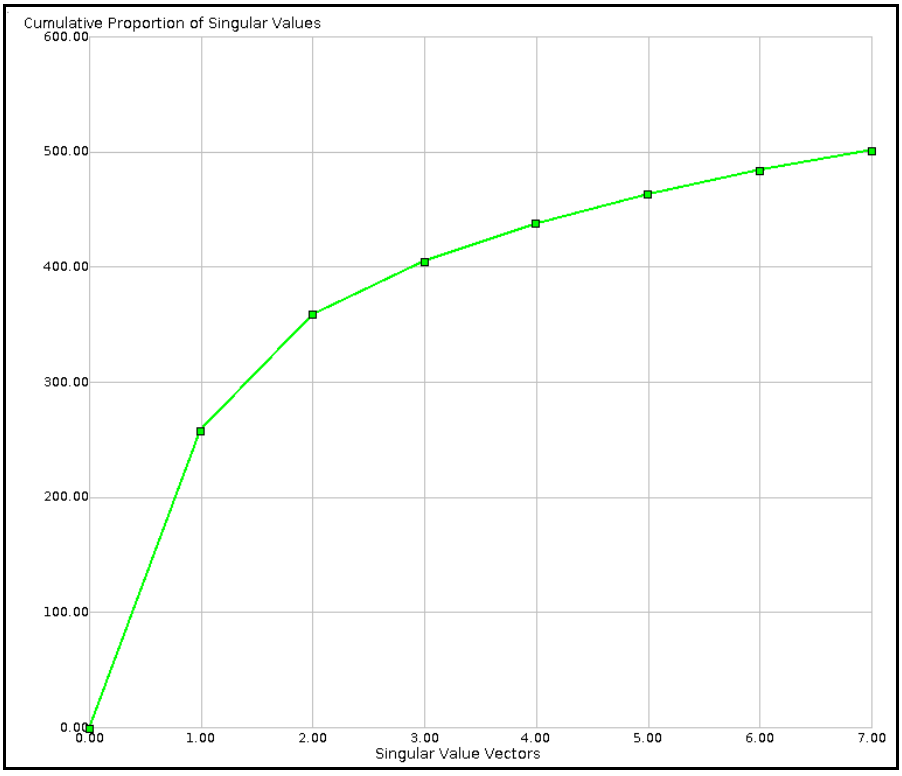

Figure 3 (Part B): SVD cumulative variance plot

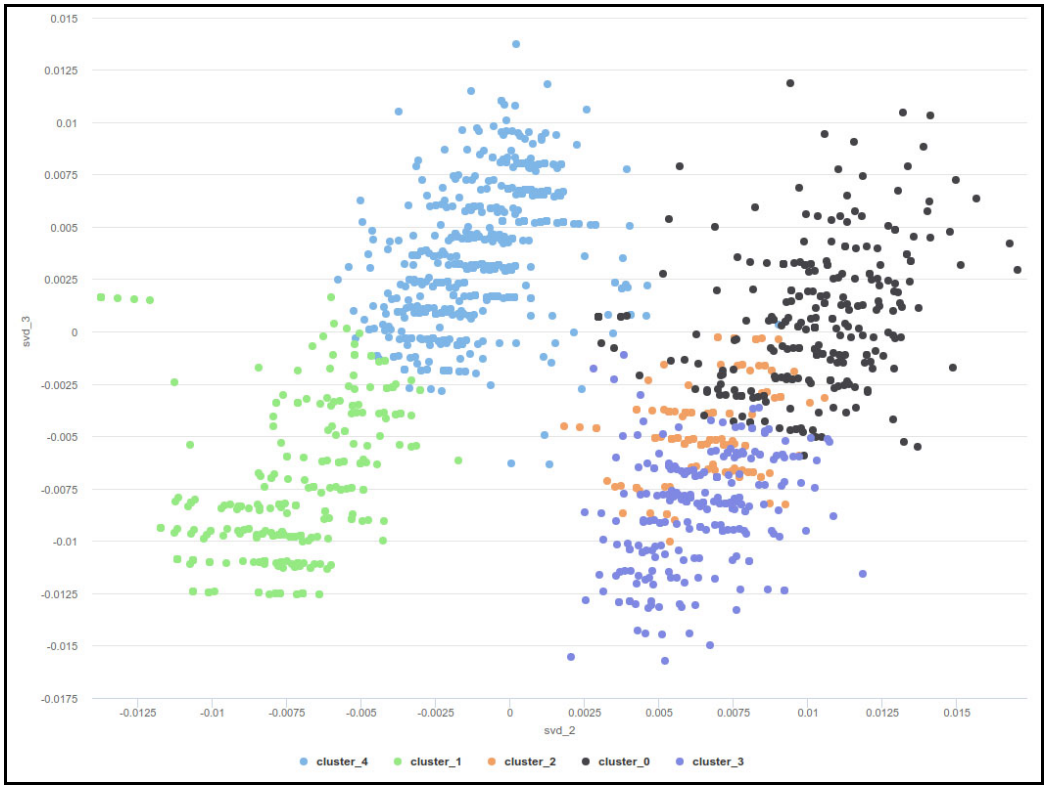

Figure 3 (Part C): Cluster scatter plot with SVD
(SVD2 and SVD3 provided for clarity)

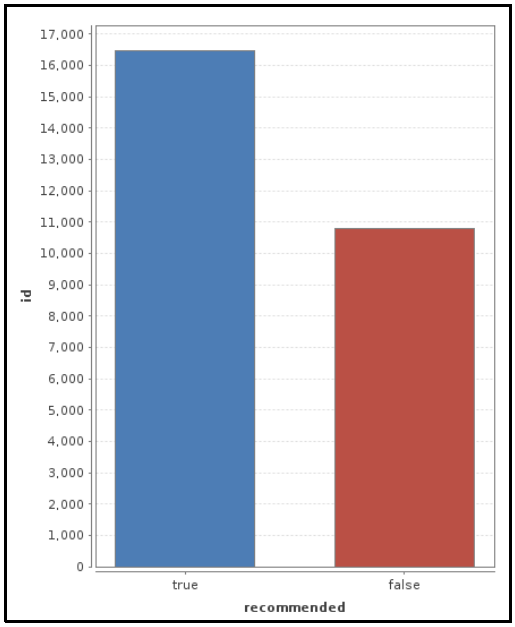

Figure 4 (Part A): Class distribution of the recommendation

PerformanceVector

```
PerformanceVector:
accuracy: 94.47% +/- 0.15% (mikro: 94.47%)
ConfusionMatrix:
True:    1        0
1:       32384    1341
0:       1834     21877
kappa: 0.886 +/- 0.003 (mikro: 0.886)
ConfusionMatrix:
True:    1        0
1:       32384    1341
0:       1834     21877
AUC: 0.982 +/- 0.001 (mikro: 0.982) (positive class: 0)
```

Random Forest

PerformanceVector

```
PerformanceVector:
accuracy: 89.82% +/- 0.59% (mikro: 89.82%)
ConfusionMatrix:
True:    1        0
1:       31922    3552
0:       2296     19666
kappa: 0.787 +/- 0.012 (mikro: 0.787)
ConfusionMatrix:
True:    1        0
1:       31922    3552
0:       2296     19666
AUC: 0.957 +/- 0.003 (mikro: 0.957) (positive class: 0)
```

Logistic Regression

PerformanceVector

```
PerformanceVector:
accuracy: 94.24% +/- 0.21% (mikro: 94.24%)
ConfusionMatrix:
True:    1        0
1:       32394    1485
0:       1824     21733
kappa: 0.881 +/- 0.004 (mikro: 0.881)
ConfusionMatrix:
True:    1        0
1:       32394    1485
0:       1824     21733
AUC: 0.984 +/- 0.001 (mikro: 0.984) (positive class: 0)
```

k-NN

Figure 4 (Part B): Performance of three classification models, i.e. Random Forest (top-left), Logistic Regression (top-right) and k-NN (bottom)

Figure 4 (Part C): ROC charts for three runs, i.e. Random Forest (top-left), Logistic Regression (top-right) and k-NN (bottom)

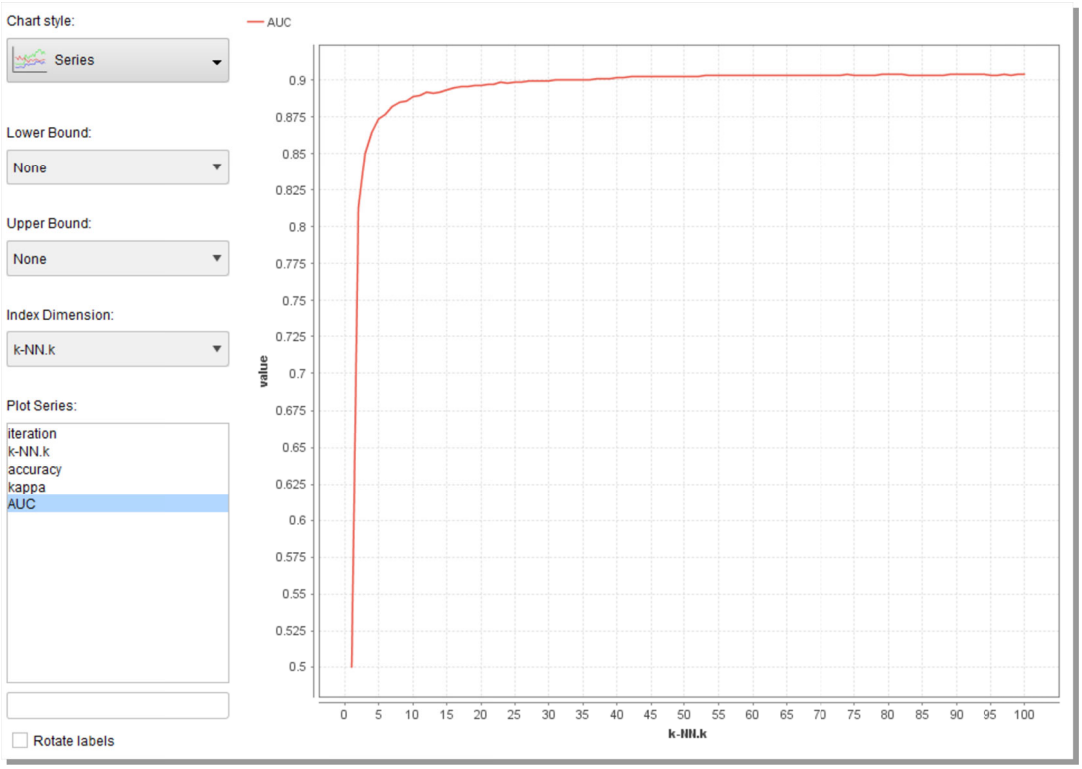

Figure 5 (Part A): Performance results k vs AUC

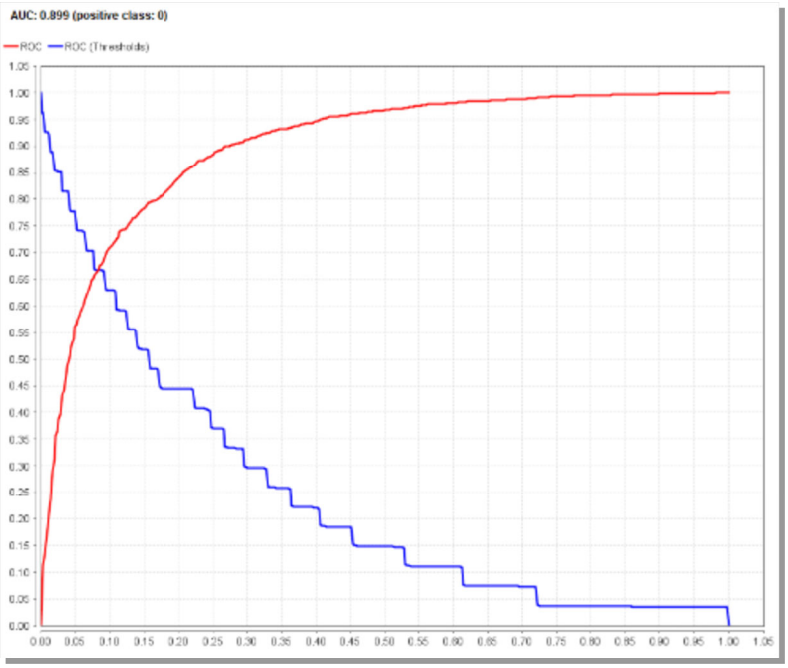

Figure 5 (Part B): ROC chart