Hands-on Exercise: Decision Tree with CreditSet Data

Loan Default Prediction

In this exercise, you will work with an excel data set that consist of information about individuals who have applied for a loan. The information included are client identification, income, age, loan amount they have applied for, and whether or not they have defaulted with a loan payment in the last 10 years. The process you go through and your final output is supposed to help you create a model that would be able to identify individuals who are most likely to pose a credit risk to the loan company.

We will use the RapidMiner Data mining software. Make sure you install the correct version of RpaidMiner Studio on your laptop. Choose either a 32 bit or a 64 bit based on the specifications of the operating system on your computer. Include as many screenshot (of relevant steps) as possible.

Particularly, you would use the Decision tree algorithm in the RapidMiner Machin learning software to build a model.

	A	в	С	D	E	F
1	clientid	income	age	Ioan	LTI	default_tenYear
2	1	66155.93	59.01702	8106.532	0.122537	0
з	2	34415.15	48.11715	6564.745	0.190752	0
4	3	57317.17	63.10805	8020.953	0.13994	0
5	4	42709.53	45.75197	6103.642	0.142911	0
6	5	66952.69	18.58434	8770.099	0.13099	1
7	6	24904.06	57.47161	15.4986	0.000622	0
8	7	48430.36	26.80913	5722.582	0.118161	0
9	8	24500.14	32.89755	2971.003	0.121265	1
10	9	40654.89	55.49685	4755.825	0.11698	0
11	10	25075.87	39.77638	1409.23	0.056199	0
12	11	64131.42	25.67958	4351.029	0.067846	0
13	12	59436.85	60.47194	9254.245	0.155699	0
14	13	61050.35	26.35504	5893.265	0.096531	0

Figure 1: Sample of Dataset

<u>Exercise</u>

The following are the main steps you would follow in building the Decision Tree (DT) model. (NB: check the DT videos uploaded on PILOT for more information and details for building a DT model).

Some steps have accompanying questions. Make sure your report includes a discussion of those questions as well. Include relevant snapshots and accompanying narratives in your report.

1. Create a new process in RapidMiner as shown in figure. Save your work at this point and remember to do it often throughout the process.

Loan Default Prediction Exercise

Eile Edit Process View Connections Set	ings Extensions <u>H</u> elp										
	Views:	Design	Results	Turbo Prep	Auto Model	Deployments	More 💌	Fin	d data, operatorsetc	2	All Studio 🔻
Repository ×	Process								Parameters	×	
🕒 Import Data 🛛 = 🔻	Process					100% 🔎 🍃	s 📮 🔍	ø 🖸	Process		
Training Resources (connected)	Process								logverbosity	init	•
Samples									loofile		
Community Samples (connected)) inp							res (
DB (Legacy)									resultfile		b
Local Repository (w055dxa)									random seed	2001	Œ
									send mail	never	•
									encodina	SYSTEM	•
Operators ×									Ide advanced	parameters	
road ave									Change compa	tibility (9.4.00	<u>)1)</u>
Data Access (4)									Help ×		
Read (3)									Process		
📩 Read CSV 📃									RapidMiner	Studio Core	
📩 Read Excel									Synopsis		
📩 Read Excel with Format									The root operator	which is the	e outer
• Write (1)	Recommended Operators	b						~	most operator of	every proce	55.
We found "Spreadsheet Table Extraction" in the Marketplace. Show me!	C Retrieve	2 ² 13	2% Select	Attributes	2² 6%	🚺 Set Role		2² 5%	Description		



 Import the excel Creditset data set to be used for building the model. Perform this task with a *Read Excel* operator. Type Read Excel into the Operator search box in the top-left corner. Drag the operator into the process area.



Figure 3: Read Excel operator

- 3. In the parameters area, click on "*Import Configuration Wizard*" button and navigate to the location of the CreditSet data on your computer. Follow the following steps to load the data correctly.
 - 3.1 Cells selected. Click on Next.

۵			<new proces<="" th=""><th>s*> – RapidMin</th><th>er Studio Trial 9</th><th>0.002 @ P-2880</th><th>23F98D74</th><th></th><th></th><th>- 8</th></new>	s*> – RapidMin	er Studio Trial 9	0.002 @ P-2880	23F98D74			- 8	
File Edit Process View Connections Clou	<u>d</u> <u>S</u> ettir	ngs Extensions	<u>H</u> elp								
					6 J				d data, operatorsetc	🔎 All Studio	
				Import Data	- Select the cells	to import.					
Repository ×									Parameters	×	
			Read Excel								
🕒 Import Data 📃 🔻	1.1										
F Training Resources (connected)	Sheet	creditset v	Cell range: A'G		Select All	Define h	eader row: 1		Import Confi	iguration Wizard	
Samples	0.000		ourrange. ree	shoot solaction	choot number 💌						
Community Samples (connected)		A	в	с	D	E	F	G	Sheet Selection	sileet lumber	
🕨 📕 DB	1	clientid	income	900	loan	-	default tenVe	dofault TonV	sheet number	1	
Local Repository (w055dxa)	2	1	66155 025	60.017	0106 522	0.122		truo	-		
Cloud Repository (disconnected)	2	0	00100.920	40.447	0100.002	0.123	0	true	imported cell range	A1	
	3	2	57017 17	40.117	0004.740	0.131	0	true			
	4	3	5/31/.1/	03.100	6402.903	0.442	0	true	encoding	SYSTEM *	
	0	4	42709.034	40.702	0103.042	0.434	0	true			
	0	0	00902.089	18.084	8770.099	0.131	1	laise	 first row as name 	ies 💙	
Operators ×	1	0	24904.004	57.472	10.499	0.001	0	true	R Hide advanced	parameters	
perfor ×	8	1	48430.30	20.809	5722.582	0.118	0	true	4.0		
4	9	8	24500.142	32.898	2971.003	0.121	8	Taise	Change compa	<u>tioliity (9.0.002)</u>	
Serformance (Classification)	10	9	40654.893	55.497	4/55.825	0.11/	0	true			
% Performance (Binominal Clas	11	10	25075.873	39.776	1409.23	0.056	0	true	Help X		
% Performance (Regression)	12	11	64131.415	25.68	4351.029	0.068	0	true	Pood Ex	col	
% Performance (Costs)	13	12	59436.847	60.472	9254.245	0.156	0	true	RapidMiner	Studio Core	
% Performance (Ranking)	14	13	61050.346	26.355	5893.265	0.097	0	true	 Tags: Load. Import. 	Read. Data. Files. XIs.	
Sector Support Vector			22022005		1760 700			avt 🖌 Canad	XIsx, Microsoft, Spr	eadsheets, Datasets	
We found "Model Visualization									Synopsis		
Extension" and "Model Management" in the Marketplace. <u>Show me!</u>	▲ C* Retrieve Select Attributes Select Attributes							8 31	This operator reads an ExampleSet from the specified Excel file.		

Figure 4.1: Variable selection

3.2 Click on drop down arrow at	the top right of default_TenYear and	d choose "change Role"
---------------------------------	--------------------------------------	------------------------

				Im	nport Data - Form	at your columns.			×	
Repository ×					Parameters ×					
Training Resources (connected) Samples Community Samples (connected) DB	Date format MMM d. yyy h:mm:ss az Replace errors with missing values income age v loan v LTI v default_ten & v default_ten & v default_ten & v polynominal polynominal									Import Configuration Wizard sheet selection sheet number
Local Repository (w055dxa) Oloud Repository (disconnected)	1 2 3	66155.925 34415.154 57317.170		59.017 48.117 63.108	8106.532 6564.745 8020.953	0.123 0.191 0.140	0 0 0 0	true Cl true R true E	nange enam clude	Role e column - cell range A1 Opens a dialog to change the role.
Operators X	4 5 6	42709.534 66952.689 24904.064		45.752 18.584 57.472	6103.642 8770.099 15.499	0.143 0.131 0.001	0 1 0	true false true		first row as names V
perfor X	7 8 9	48430.360 24500.142 40654.893		26.809 32.898 55.497	5722.582 2971.003 4755.825	0.118 0.121 0.117	0 1 0	true faise true		 Hide advanced parameters Change compatibility (9.0.002)
 Performance (Binominal Clas Performance (Regression) Performance (Costs) 	10 11 12	25075.873 64131.415 59436.847		39.776 25.680 60.472	1409.230 4351.029 9254.245	0.056 0.068 0.156	0 0 0	true true true	~	Help × Read Excel
% Performance (Ranking) % Performance (Support Vector ∨ < >> We found "Nodel Visualization Extension" and "Model Nanagamers" in	<	(•	<u>Previous</u>	> on problem	s.	RapidMiner Studio Core Tags: Load, Import, Read, Data, Files, Xis, Xisx, Microsoft, Spreadsheets, Datasets Synopsis
the Marketplace. Show me!	🗋 Re	trieve		2 59%	Set Role	28 32%	E Select Attributes	2 ² 319	6	This operator reads an ExampleSet from the specified Excel file.

Figure 4.2: Change variable parameters

3.3 In the next screen, change role of default_TenYear to label

4				<new process<="" th=""><th>*> - </th><th>RapidMine</th><th>er Studio</th><th>Trial 9.0</th><th>).002 @ P·</th><th>-288023F98D74</th><th></th><th></th><th></th><th>- ć</th><th>3 ×</th></new>	*> -	RapidMine	er Studio	Trial 9.0).002 @ P·	-288023F98D74				- ć	3 ×
Eile Edit Process View Connections Clou	i <u>d S</u> ettii	ngs Extension	is <u>H</u> e	lp											
	_				In	nport Data	- Form	at your c	olumns.			×	d data, operatorsetc	🔎 All S	Studio 🔻
Repository X						Format	your	columi	ıs.				Parameters	×	
Training Resources (connected)	D	ate format MM	M d, yyy	y h:mm:ss a z		•		Replace	errors with n	nissing values ①			Import Confi sheet selection	guration Wizard	
Community Samples (connected)		income	۰.	age	\$ v	loan	\$ v	LTI	\$ v	default_ten \$ 🔻	default_Te ¢	•			
DB		real	- h	real		real		real		inteaer	polynominal		sheet number	1	I) (I)
Local Repository (w055dxa)	1	66155 025	9				Change	role			110	~			
Cloud Repository (disconnected)	2	34415.154		Pleas	e ente	er the new ro	le:				rue	-	imported cell range	A1	٢
	3	57317.170									rue		anading	OVOTEN	
	4	42709.534									rue		encoung	STOTEM	
	5	66952.689	Ent	er value						•	alse		🥖 first row as nam	ac 💙	(1)
Onerstern V	6	24904.064	id	1							rue			••••	
operators	7	48430.360	weig	pht							rue		Inde advanced	parameters	
perfor X	8	24500.142									alse		Change compa	Nhility (0.0.002)	
Reformance (Classification) ^	9	40654.893		55.497		4755.825		0.117		0	true			1011(1 (3.0.002)	
Performance (Binominal Classes)	10	25075.873		39.776		1409.230		0.056		0	true				
% Performance (Regression)	11	64131.415		25.680		4351.029		0.068		0	true		пеір 🗡		
% Performance (Costs)	12	59436.847		60.472		9254.245		0.156		0	true	\vee	🔸 Read Ex	cel	í
% Performance (Ranking)		<										>	RapidMiner	Studio Core	
% Performance (Support Vector V											Sono proble	ems.	Tags: Load, Import,	Read, Data, Files,	XIs,
<									+	- Previous	Einish 🗙 Car	cel	XIsx, Microsoft, Spre	adsheets, Datas	.ets
We found "Model Visualization													Synopsis		
Extension" and "Model Management" in the Marketplace. <u>Show me!</u>	<u>C</u> R	etrieve		2² 5 9%		Set Role			2² 32%	Elect Attributes	22	31%	This operator read from the specified	s an ExampleSet Excel file.	t,

Figure 4.3: Change variable parameters

File Edit Process View Connections Clou	ıd Settir	ngs Extension	ns He	new process	*> - F	RapidMiner	Studio	Trial 9.0.0	02 @ P-	288023F98D74				-	ð	
			-		In	port Data	- Forma	at your colu	umns.			×	d data, operatorsetc	AI	l Studio 🔻	•
Repository ×						Format y	your o	columns	•				Parameters	×		^
Training Resources (connected) Samples	₽	ate format MM	M d, yyy	/ h:mm:ss a z		•		Replace erro	ors with m	issing values 🕕			🎾 Import Confi	guration Wizard		
Community Samples (connected)		income	¢ •	age	¢ •	loan	¢ .	LTI	¢ •	default ten	🔻 default Te 🌣 🔻		sheet selection	sheet number	• •	=
DB		real		real		real		real		integer	Change Type 🔸		sheet number	1	٩	
Local Repository (w055dxa)	1	66155.925		59.017		8106.532		0.123		0	Change Role	^				
Cloud Repository (disconnected)	2	34415.154		48.117		6564.745		0.191		0	Rename column		imported cell range	A1	0	
	3	57317.170		63.108		8020.953		0.140		0	Exclude column		opcoding	OVOTEM	• 0	
	4	42709.534		45.752		6103.642		0.143		0	true The selected co	lumn	will be excluded from	the import		
	5	66952.689		18.584		8770.099		0.131		1	false		first row as nam	es 💙	1	
Operators ×	6	24904.064		57.472		15.499		0.001		0	true					\sim
	7	48430.360		26.809		5722.582		0.118		0	true		Readvanced	parameters		
perfor ×	8	24500.142		32.898		2971.003		0.121		1	false		Change compared	ibility (9.0.002)		
% Performance (Classification)	9	40654.893		55.497		4755.825		0.117		0	true					
1 Performance (Binominal Clas	10	25075.873		39.776		1409.230		0.056		0	true		Heln X			
% Performance (Regression)	11	64131.415		25.680		4351.029		0.068		0	true					~
% Performance (Costs)	12	59436.847	_	60.472		9254.245		0.156		0	true	~	🛃 Read Ex	cel		
1 Performance (Ranking)		<										,	RapidMiner	Studio Core		
Serformance (Support Vector												2.	Tags: Load, Import, XIsx, Microsoft, Spre	Read, Data, File adsheets, Data	s, <u>XIs</u> , <u>isets</u>	
We found "Model Visualization									+	- Previous	Einish Cancel		Synopsis			
Extension" and "Model Management" in the Marketplace. <u>Show mel</u>	C R	etrieve		2² 59%		Set Role		2	32%	Select Attribu	tes 🙎 31%	6	This operator read from the specified	s an ExampleS Excel file.	et	~

3.4 For subsequent steps, exclude default_tenYear, clientid and LTI attributes.

Figure 4.4: Change variable parameters

٩		<new proces<="" th=""><th>s*> – RapidMiner</th><th>Studio Trial 9.0.002 @</th><th>P-288023F98D74</th><th></th><th></th><th>- 0 ×</th></new>	s*> – RapidMiner	Studio Trial 9.0.002 @	P-288023F98D74			- 0 ×
<u>File Edit Process View Connections Clou</u>	u <u>d S</u> ettings E <u>x</u> ten	sions <u>H</u> elp						
	_		Increase Desta	F			d data, operatorsetc	🔎 All Studio 🔻
			Import Data	- Format your columns				
Repository ×			Parameters	×				
Training Resources (connected)	Date format	MMM d, yyyy h:mm:ss a z	🎾 Import Confi	guration Wizard				
 Samples Community Samples (connected) 	clientid	o → income	≱ ▼ age	¢ ▼ Ioan	LTI 0 -	default_ten 🛊 🔻 d	sheet selection	sheet number 🔻 🛈 📰
🕨 📕 DB	integer	Change Type 🔸	real	real	real	integer p	sheet number	1
Local Repository (w055dxa)	1 1	Change Role	59.017	8106.532	0.123	0 t	^	
Cloud Repository (disconnected)	2 2	Rename column	48.117	6564.745	0.191	0 t	imported cell range	A1 ①
	3 3	Exclude column	63.108	8020.953	0.140	0 t	oncoding	SYSTEM T
	4 4	42700.524	45 752 red column will be ex	6102 642 cluded from the import	0.143	0 t	encounty	STOTEM . U
	5 5	66952.689	18.584	8770.099	0.131	1 f	🖌 first row as nam	ies 💙 🕕
Operators ×	6 6	24904.064	57.472	15.499	0.001	0 t		×
	7 7	48430.360	26.809	5722.582	0.118	0 t	Hide advanced	parameters
perfor	8 8	24500.142	32.898	2971.003	0.121	1 f	Change compa	tibility (9.0.002)
% Performance (Classification) ^	9 9	40654.893	55.497	4755.825	0.117	0 t		
🐒 Performance (Binominal Clas	10 10	25075.873	39.776	1409.230	0.056	0 t	Help X	
% Performance (Regression)	11 11	64131.415	25.680	4351.029	0.068	0 t		^
% Performance (Costs)	12 12	59436.847	60.472	9254.245	0.156	0 t	v 🛛 🛃 Read Ex	cel
% Performance (Ranking)	<					> no problem	RapidMiner	Studio Core
Serformance (Support Vector V					← <u>P</u> revious	Einish Cance	 Tags: Load, Import, XIsx, Microsoft, Spring 	<u>Read, Data, Files, XIs,</u> <u>eadsheets, Datasets</u>
 We found "Model visualization Extension" and "Model Management" in the Marketplace. <u>Show mel</u> 	🗋 Retrieve	2 8 59%	📝 Set Role	2 ⁶ 329	Select Attribut	es 🔒 315	This operator read	ls an ExampleSet Excel file. ✓

Figure 4.5: Change variable parameters

Image: Displaying the second	<u>d S</u> ettings E <u>x</u> tensions	<new process*=""> Help</new>	 RapidMiner Stud 	dio Trial 9.0.002 @	P-288023F98	3D74		- 1	× נ
			Import Data - For	mat your columns.		×	d data, operatorsetc	P All s	Studio 🔻
Repository X		Parameters X							
Training Resources (connected) Samples	Date format MMM d								
Community Samples (connected)	clientid 🔹	- income 🔹 -	age 🔹 🔹	Ioan 🔹 🔻	LTI	🔅 🔻 default_ten 🕸 🔻 d	sheet selection	sheet number	•
🕨 📕 DB	integer	real	real	real	real	Change Type	sheet number	1	1
Local Repository (w055dxa)	1 1	66155.925	59.017	8106.532	0.123	Change Role t			
Cloud Repository (disconnected)	2 2	34415.154	48.117	6564.745	0.191	Rename column t	imported cell range	A1	1
	3 3	57317.170	63.108	8020.953	0.140	Exclude column t			
	4 4	42709.534	45.752	6103.642	0.143	The selected column wil	encoding	mont	
	5 5	66952.689	18.584	8770.099	0.131	1 1	✓ first row as name	es 💙	(I)
Operators ×	6 6	24904.064	57.472	15.499	0.001	0 t			
	7 7	48430.360	26.809	5722.582	0.118	0t	Hide advanced	parameters	
perfor ×	8 8	24500.142	32.898	2971.003	0.121	1 f	Change compa	tibility (9.0.002)	
Serformance (Classification)	9 9	40654.893	55.497	4755.825	0.117	0 t			
% Performance (Binominal Clas	10 10	25075.873	39.776	1409.230	0.056	0 t	Help X		
% Performance (Regression)	11 11	64131.415	25.680	4351.029	0.068	0 t	line to		^
% Performance (Costs)	12 12	59436.847	60.472	9254.245	0.156	0 tv	🛃 🛃 Read Ex	cel	
1 Performance (Ranking)	<					no problems	RapidMiner	Studio Core	
Serformance (Support Vector ∨ <					← <u>P</u> revious	Einish Cancel	Tags: <u>Load</u> , <u>Import</u> , <u>XIsx</u> , <u>Microsoft</u> , <u>Spr</u>	<u>Read, Data, Files,</u> adsheets, Datas	<u>Xls</u> , <u>ets</u>
We found "Model Visualization				L			Synopsis		
the Marketplace. Show me!	C Retrieve	2² 59%	🚺 Set Role	2² 32%	E Select	Attributes 2 31%	This operator read from the specified	s an ExampleSe Excel file.	t v

Figure 4.6: Change variable parameters

4. Search and drag the **Split Validation** operator into the process area. Use a relative split and a split ratio of 0.7 in the parameters areas. Explain the meaning of **Split Ratio**. Discuss how it impacts data validation in data mining? Connect the operators together as shown in figure 5.



Figure 5: Validation operator

- Double-click to open the Split Validation nested operator. Two sub-processes are presented, a Training process and a Testing process. At this point, you will build your model in the Training process and test it in the Testing process.
- 6. At this point, you are ready to build your DT model. Search and drag the Decision Tree, Apply Model and Performance operators into the process areas. Connect the operators as shown figure 6. In the DT operator, choose a maximal depth of 10 and gain_ratio criterion. Leave all other DT parameters default. Mention and explain 1 other criterion parameter.



Figure 6: Validation operator sub-processes

- 7. Click on the Run button.
- 8. Include a snapshot of the **Accuracy criterion** (in the Performance tab) in your report. Explain the parameters and how it relates to prediction of individuals who are likely to default on their loan payments.
- 9. Include a snapshot of the **DT graph** and **Description** in your report. Explain the graph and how it relates to prediction of individuals who are likely to default on their loan payments.
- 10. Assume you would want to prune the Tree to a depth of 5. Edit the necessary parameter(s) in the DT operator and run the model again.
- 11. Include a snapshot of your new DT graph and explain how it relates to prediction of individuals who are likely to default on their loan payments.