

Automated Data Collection

PRICE Systems, LLC





© 2016 PRICE Systems, LLC All Rights Reserved | Decades of Cost Management Excellence



Agenda

- Introduction and Motivation
- Data Collection Challenges
- Introduction to RapidMiner Capabilities
- RapidMiner Crawling the Web
- Implementation
- Future Directions

Introduction and Motivation



- Data Collection is a necessary evil for cost estimators
 - To support the creation of Cost Estimating Relationships (CERs)
 - To support estimating by analogy
 - To support selection of input values for cost estimating models
- Data collection is hard
 - Data is often hard to find
 - Even when data is found it is often hard to mine as the process is tedious and time consuming
 - Data is often very noisy making it hard to understand and work with

Introduction and Motivation



- Original motivation for this work arose from efforts around an overhaul of the TruePlanning [®] Information Technology Services Cost Model
- Many of the models we developed for IT Services required commodity pricing information
- Commodity pricing is hard to estimate because
 - Prices are constantly changing
 - Many companies have negotiated agreements with specific vendors
 - There are many things that drive commodity pricing outside of the scope of a typical cost estimating relationship

Data Collection Challenges For Commodity Prices



- Finding the right data
 - Accurate pricing data
 - Significant technical and specification information
 - Normalization across multiple vendors
- Keeping the data up to date
 - Commodity prices change frequently based on market factors, supply and demand, etc.
 - Good pricing data from last quarter is unlikely to relevant in this quarter
- Need a solution that is
 - Repeatable
 - Consistent
 - Can be accomplished quickly with the push of a button
 - Can be updated regularly (monthly, quarterly, bi-annually) without extensive time investment

RapidMiner



- Open Source application
- RapidMiner 5.3 and 7.2 currently available for download from <u>http://rapidminer.software.informer.com/5.3/</u> <u>https://my.rapidminer.com/nexus/index/html#downloads</u>
- Licensed under GMU Affero General Public License version 3
- User friendly, graphical user interface that allows for data collection and analysis

Summary of RapidMiner Capability



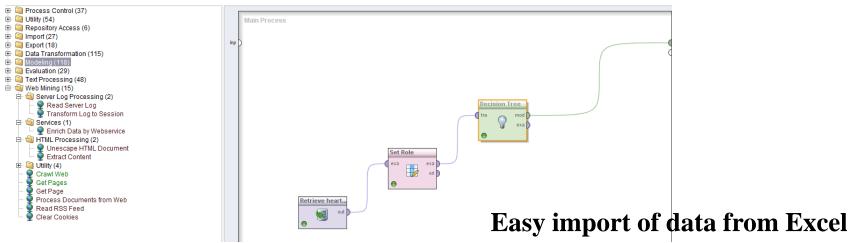
- Data Transformation
 - Filtering
 - Sorting
 - Replacing Missing values in a data set
 - Aggregation, etc.
- Modeling
 - Classification
 - Clustering
 - Correlation, etc.
- Evaluation
 - Validation
 - Regression
 - Significance, etc.

- Text Processing
 - Extraction
 - Transformation
 - Extraction
 - Tokenization, etc.
- Web Mining
 - Server Log Processing
 - HTML Processing
 - Web Crawling
 - Read RSS Feed, etc.

Introduction to RapidMiner Capabilities



Drag and Drop Interface to Build Processes



Data View O Meta Data View	Plot View A	dvanced Charts 🔘 Ann	otations
0 0	0		

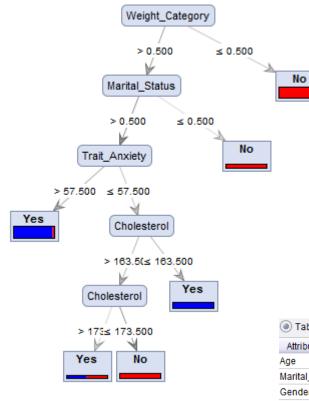
ExampleSet (138 examples, 0 special attributes, 8 regular attributes)

Row No.	Age	Marital_Stat	Gender	Weight_Cat	Cholesterol	Stress_Man	. Trait_Anxiety	2nd_Heart
1	60	2	0	1	150	1	50	Yes
2	69	2	1	1	170	0	60	Yes
3	52	1	0	0	174	1	35	No
4	66	2	1	1	169	0	60	Yes
5	70	3	0	1	237	0	65	Yes
6	52	1	0	0	174	1	35	No
7	58	2	1	0	140	0	45	No
8	59	2	1	0	143	0	45	Yes
9	60	2	0	0	139	0	45	No
10	51	1	1	0	174	1	40	No
11	52	1	0	0	189	1	65	No
12	70	2	1	1	147	1	50	Yes
13	52	2	1	2	160	0	40	Yes
14	74	3	1	2	178	0	75	Yes
15	64	2	1	2	236	1	80	Yes
16	69	2	0	1	146	1	50	Yes
17	58	2	0	0	141	0	45	No
18	68	1	0	0	172	0	60	No

Introduction to RapidMiner Capabilities



Countless ways to visualize your data



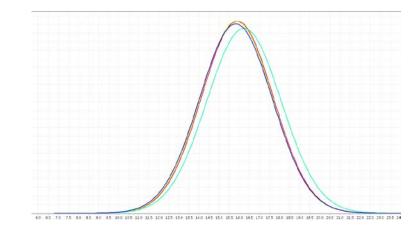


Table Month	O Delevie - Tekte		A
Table View	Pairwise Table	O Plot View	Annotations

Attributes	Age	Marital_Stat	Gender	Weight_Cat	Cholesterol	Stress_Man	Trait_Anxiety	2nd_Heart
Age	1	0.427	0.076	0.402	0.395	-0.359	0.667	-0.499
Marital_Statu	0.427	1	-0.033	0.065	0.274	-0.292	0.238	-0.365
Gender	0.076	-0.033	1	0.449	0.191	-0.241	0.082	-0.318
Weight_Cate	0.402	0.065	0.449	1	0.398	-0.347	0.501	-0.731
Cholesterol	0.395	0.274	0.191	0.398	1	-0.406	0.579	-0.507
Stress_Man;	-0.359	-0.292	-0.241	-0.347	-0.406	1	-0.321	0.439
Trait_Anxiety	0.667	0.238	0.082	0.501	0.579	-0.321	1	-0.483
2nd_Heart_/	-0.499	-0.365	-0.318	-0.731	-0.507	0.439	-0.483	1

Introduction to RapidMiner Capabilities



Multiple Regression Models, Weighting and Analysis

Attribute	Coefficient	Std. Error	Std. Coeffici	Tolerance	t-Stat	p-Value	Code
attribute_1	2.053	1.677	1.618	0.904	1.224	0.307	
attribute_2	2.007	1.531	1.721	0.930	1.310	0.257	
attribute_3	-5.631	1.425	-4.936	0.886	-3.951	0.000	****
attribute_4	3.253	0.874	2.808	0.953	3.720	0.000	****
attribute_8	-1.191	0.532	-0.753	0.910	-2.241	0.030	**
attribute_9	0.943	0.637	0.627	0.850	1.479	0.181	
attribute_10	-0.381	0.586	-0.246	0.802	-0.649	0.523	
attribute_11	0.247	0.665	0.139	0.701	0.371	0.714	
attribute_12	1.027	0.538	0.575	0.777	1.908	0.069	*
attribute_13	0.422	0.536	0.218	0.851	0.788	0.439	
attribute_14	-0.361	0.485	-0.200	0.955	-0.745	0.464	
attribute_15	0.257	0.494	0.165	0.996	0.521	0.608	
attribute_16	-0.256	0.480	-0.157	1.000	-0.532	0.600	
attribute_17	-0.696	0.502	-0.441	1.000	-1.387	0.219	
attribute_18	1.007	0.469	0.582	1.000	2.149	0.038	**
attribute_19	-0.732	0.404	-0.374	0.961	-1.811	0.087	*
attribute_20	0.615	0.473	0.287	0.925	1.300	0.263	
attribute_21	-0.623	0.464	-0.264	0.896	-1.341	0.242	
attribute_22	0.647	0.468	0.265	0.941	1.384	0.221	
attribute_23	-0.506	0.479	-0.196	0.970	-1.057	0.428	
attribute_24	1.156	0.490	0.411	0.995	2.358	0.022	**

attribute	
duration	0
pension	0.062
education-allowance	0.064
shift-differential	0.185
bereavement-assistance	0.185
working-hours	0.227
col-adj	0.258
vacation	0.271
wage-inc-3rd	0.298
standby-pay	0.412
contrib-to-dental-plan	0.424
longterm-disability-assistance	0.535
wage-inc-2nd	0.547
contrib-to-health-plan	0.676
statutory-holidays	0.685
wage-inc-1st	1

PolynomialRegression

25.320 * a1 ^ 1.000 + 3.697 * a2 ^ 2.000 + 1.814 * a3 ^ 2.000 - 2.192 * a4 ^ 1.000 - 0.673 * a5 ^ 2.000 - 83.828

RapidMiner for Web Crawling

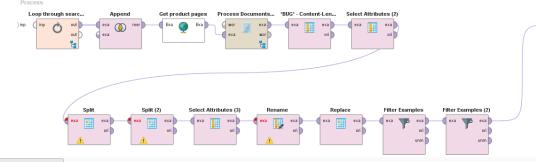


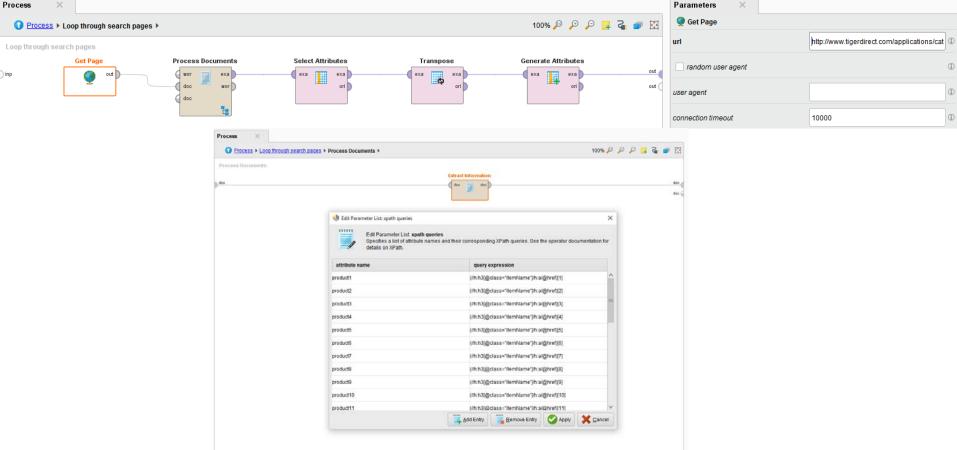
Starting with a webpage that is setup where you can view list of products:

- Get Page
 - Sends RapidMiner out to product page URL
 - Example: <u>http://www.tigerdirect.com/applications/category/category_slc.asp?page=1&Nav=|c:4935|&Sort=3&Recs=10</u>
- Process Documents
 - Uses Xpaths to grab product links
- Get Pages
 - Sends RapidMiner out to each product link that it grabbed in Process Documents
- Process Documents from Data
 - Uses Xpaths to grab product specifications on each product page

Implementation to Support Commodity Pricing Data Collection







res

Implementation to Support Commodity Pricing Data Collection



Combining into one process

Process ×						Parameters ×	
Process >				100% 🔎	🗵 电 🍒 📮 🍳	Category URLs (Generate	Data by User Specification)
Process						attribute values	🗒 Edit List (8) 🛈
Dinp	Subprocess	Tiger Direct 1. Add your URL to Category URLs			res	set additional roles	🕞 Edit List (0) 🛈
		2. Use Make Selection to use choose which one you want to mine.			Edit Parameter List: attribute v		×
	ocuments	 Loop through search pages will grab all the product page links. You can change the iterations macro if you just want to test for a couple of pages. 		attribute n	attribute value	butes and their values in the sing	e example returned.
in to out out exa	exa wor		Т	Tablets	"http://www.tigerdirect.com/ap	plications/category/category_slc.a	sp?page=1&Na 🔳
	te	 Get product pages will retrieve each product page. XPaths are separate for each category. You should 	L	aptops	"http://www.tigerdirect.com/ap	plications/category/category_slc.a	sp?page=1&Na
		use the Select Subprocess and and place the Extract Information in the respective section. e.g Tablet is the	ν	Vorkstations	"http://www.tigerdirect.com/ap	plications/category/guidedSearch.	asp?CatId=6&s 📕
		first category in Category URLs so Extract Information should be the first Select Subprocess	S	Servers	"http://www.tigerdirect.com/ap	plications/category/category_slc.a	sp?page=1&Na
		5. Select Subprocess can be used for individual category	F	Printers	"http://www.tigerdirect.com/ap	plications/category/guidedSearch.	asp?Catld=218
		post-processing. 6. Use the "category" Macro to save the file.	By the first state of the stat	sp?page=1&Na 🔳			
		6. Ose the category macro to save the me.	F	Routers	"http://www.tigerdirect.com/ap	plications/category/category_slc.a	sp?page=1&Na 📕
			5	Switches	"http://www.tigerdirect.com/ap	plications/category/category_slc.a	sp?page=1&Na 🔳
	Select Subp	out exa exa exa ori ori ori			Edit Parameter List: attribute values This parameter defines the attributes and attribute value "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications "http://www.tigerdirect.com/applications		ply 🔀 Cancel



Process to find relevant pages when webpage has various product links to follow:

- Crawl Web
 - Start at the highest level URL for each product
 - Example: <u>http://www.dell.com/us/business/p/laptops</u>
 - Crawling rules will send RapidMiner out to any URL matching your search
- Process Documents from Data
 - Uses Xpaths to grab product specifications on each product page

Implementation to Support Commodity Pricing Data Collection



Main Process			9	Crawl Web					
CrawtWeb	Process Docu., Replace (2) Replace (3)	uri		http://www.dell.com/us/business/pilaptops					
. *		Cas or	awling rules	Edit List (2).					
	• • • •	C] write pages into files						
			add pages as attribute						
	Replace (4) Filter Examples Rename Reorder Attri. Write Excel	m	ax pages	100					
			ax depth						
	Edit Parameter List: crawling rules	x	main	server *					
		de	łay	5000					
	Edit Parameter List crawling rules Specifies a set of rules that determine, which links to follow and which pages to process.	m	ax threads	4					
	rule application rule value	mi	ax page size	1000					
		ā, 10	er agent	[rapid-miner-crawler					
	store_with_matching_urt * .*\?ref=PD_OC	à					Extract	Information	
			Fortune 4					XPath	
			doc			doc	dneivitybe		
			•			doc	attribute type	Nominal	
					_		xpath queries	🔀 Edit List (11)	
		Ø Edit Paramete	er List: xpath queries		×		namespaces	Edit List (0)	_
	📮 Add Entry 🖳 Bernove Entry 🧭 Apply 💥 Çan		it Parameter List: xpath queries						
			ails on XPath.	d their corresponding XPath queries. See the op	erator documentation for		🧹 ignore CDATA		
			attribute name	query expres	sion		✓ assume html		
		Name		normalize-space(//*[@id='masthea					
		Unit Purchase P	rice	//h:span[contains(.,'Starting at') and	@class='startText']//h:spa				
		Processor		string(//h:h5[contains(.,'Processor')	//following-sibling::h:div/h				
		Operating_Syste	m	string(//h:h5[contains(.,'Operating S	ystem')]//following-sibling				
		Memory		string(//h:h5[contains(.,'Memory')]//	following-sibling::h:div/h:sr				
		Hard_Drive		string(//h:h5[contains(.,'Hard Drive')	//following-sibling::h:div/h				
		Type of Device		string('laptop')					
		Memory_GB		string(//h:h5[contains(.,'Memory')]//	following-sibling::h:div/h:sr				
		Hard_Drive_GB		string(//h:h5[contains(.,'Hard Drive')]//following-sibling::h:div/h				
		Level		string('2')					
		Definition Name		string('End User Services New Pro	ects')				
							🔞 Help 🕺 🗎 Comment 🕱		
				Add Entry	Apply 🔀 Cancel		Extract Information		
							Synopsis		
Row No.	URL Name Definition Name Hard_Drive Hard_Drive Lev	-		g_System Processor Type of		Price			
1	http://www.dv Latitude 13 30 End User Services M.2 128GB 5 128 2	4GB (1x4G): 4		10 Pro 64t Pentium DC laptop	699				
2	http://www.du Latitude 12 7(End User Services M.2 128GB 5 128 2	4GB (1x4GB) 4	Windows	10 Pro, 64- 6th Generat laptop	1049				
3	http://www.di Latitude 13 7(End User Services M.2 128GB § 128 2	4GB LPDDR 4	Windows	7 Professie Intel® Core1 Iaptop	1299				
4	http://www.dv New Inspiron End User Services 500GB 5400 500 2	4GB Single (4	Windows	10 Home, Intel® Pentii Iaptop	499				
5	http://www.dv New Inspiron End User Services 500GB 5400 500 2	4GB Single (4	Windows	10 Home, Intel® Pentil Iaptop	449				
6	http://www.du New Inspiron End User Services 256GB Solid 256 2	8GB Dual Ct 8	Windows	10 Home € 6th Generat laptop	749				

Windows 10 Home 6 6th Generat laptop

Windows 7 Professie Intel® Core¹ laptop

Windows 7 Professi Intel® Core1 Iaptop

Windows 10 Home, 6th Generat laptop

749

999

1399

999

© 2016 PRICE Systems, LLC All Rights Reserved | Decades of Cost Management Excellence

http://www.d/ New Inspiron End User Services 256GB Solid 256

http://www.di Precision 15 (End User Services 500GB 2.5 ir 500

http://www.di Precision 15 ! End User Services 500GB 2.5 ir 500

http://www.di XPS 15 Lapto End User Services 500GB 7200 200

2

2

2

2

8GB Dual Ct 8

8GB (2x4GB 8

8GB (2x4GB 8

8GB (1x8G): 8

6

7

8

9

10

Implementation to Support Commodity Pricing Data Collection



- Processes have been created to crawl Dell, HP and TigerDirect for pricing and performance data for:
 - Laptops
 - Workstations
 - Tablets
 - Printers
 - Storage Devices
 - Servers
 - Other Supporting Hardware
- These processes create Excel files that are directly importable into the IT Hardware TrueFindings[®] database
- This database can be updated in several hours to support monthly or quarterly updates of the database

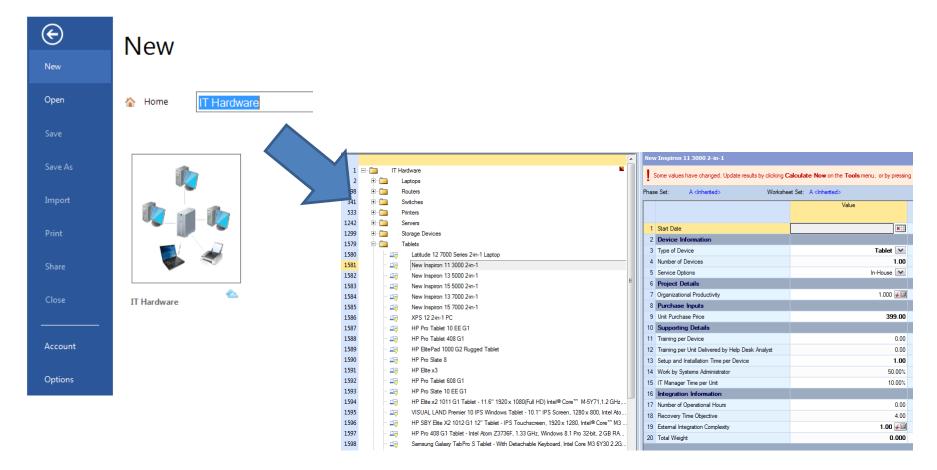
IT Hardware for TrueFindings[®] and TruePlanning[®]



	Value	Sh	ared	Project	Cost Object						
Start Date		N	ame				Value	Method	_	Туре	
Device Information										.76-	
Type of Device	Laptop 💌		_								`
Service Level	3.00 🖌 🗐	Ā	∃Lapto	ps - Unit Pu	urchase Price		710.031578947	Distributi	on	Mean	
Number of Deployments	Custom - Yearly	7	∃Printe	rs - Unit Pu	rchase Price		1552.25946327	Distributi	on	Mean	
Quantity Per Next Higher Level	1.00	5	Route	rs- Unit Pu	rchase Price		6406.29338842	Distributi	ion	Mean	
Purchase or Lease	Purchase 💌	5	Serve	rs - Unit Pu	rchase Price		1573.89285714	Distributi	ion	Mean	
Service Options	In-House 💌					Drice	1369.10394265	Distributi	ion	Mean	
Project Details				-	urchase Price	- riice					
Organizational Productivity	1.000						389.513089005			Mean	
Purchase Inputs					rchase Price		932.086206896	Distributi	on	Mean	
Unit Purchase Price	710.03	p 1	∃Works	stations - U	nit Purchase Pr	ice	783.866310160	Distributi	on	Mean	
Inventory	Selected Column	G C	ipboard 🕞	Import/Expor	t 🗟 Finding 🗟						
Unit Lifetime	Search TrueFindings Database		× Data	(95 rows)							
Supporting Details	Vour Search			Name	Definition Name	Type of I	Device Number of De	Unit Purcha	 Externa 	l Integr Total	l Weight
Software Price per Unit	✓ ★ + Type of Device		1	HP Chromebo	End User Servi	laptop	1	179	1	0	
Annual Training per End User			2	HP Chromebo	End User Servi	laptop	1	189	1	0	
Annual Support per End User			3		End User Servi	laptop	1	199	1	0	
Training per Unit Delivered by Help Desk Analyst			4	ASUS 11.6" Eee	e End User Servi P End User Servi	laptop	1	241 259	1	0	
maining per onic betweeeu by help besk Analysi			6	HP Chromebo		laptop laptop	1	259	1	0	
			7		End User Servi	laptop	1	302	1	0	
			8	Acer Aspire E5	End User Servi	laptop	1	302	1	0	
	► Keyword										
	Characteristics		Distri	bution Finder	Dependency Finder	Curve Fir	nder MultiCurve Finder	Findings			
	b Definition Name ♥ b Hard_Drive ♥			dent Variables		×				Statistics	
	Memory Q Operating_System Q		Deper	ident Variabl (Unit Purchase Price	•	'Unit Purchase Price			Chart Type:	Box F
	Processor Q						1875	1875 -		Statistics	Data
	▶ URL Q Performance									Min	179
	Number of Devices (95)					ice	1375	1375 -			1699
	 Unit Purchase Price (95) External Integration Complexity (95) 					e Pr				25%	476.5
	b Total Weight (95)					Unit Purchase Price	875		5th Percentile 5th Percentile		
	▷ Hard_Drive_GB (95) ▷ Memory_GB (95)					Pur		<u>a</u>	ediage/Mear	Mean	710.0315
	,					Unit	and	5 × 2	5th Percentile	Section 2	
							375	375 - 1	5th Percentile	Whiskers Perc	centiles:
							<i>6</i>	ā		Williskers Fere	ientites.

Implementation to Support Commodity Pricing Data Collection

 These data points can also be accessed via the File New Template Search in TruePlanning 16.0 for immediate drag and drop into a project file.



Future Directions



- PRICE intends to update and extend the IT Hardware Commodity database
 - Regular updates with the existing processes that have been developed
 - Developing new processes based on user requirements & suggestions
- We are currently investigating the feasibility of creating similar processes to support the IT Software pricing requirements
 - This may be problematic because many software applications require calls to the vendor for quotes – we're hopeful we may be able to find sources
- We are considering extending this project to include commodity prices for electronic components to support Microcircuit cost estimation
- We would like to apply this expertise to develop custom solutions for clients based on their specific purchasing processes