

# Table Detection with ruling lines

Max Tiedl  
February 20, 2024

# Goal

- Find/extract tables with ruling lines
- Fintabnet<sup>(1)</sup> dataset

	Years Ended October 31,		
	2005	2004	2003
		(Restated)	
		(in millions)	
Net revenue . . . . .	\$1,796	\$2,021	\$1,588
Costs and expenses . . . . .	1,570	1,779	1,594
Income (loss) from discontinued operations . . . . .	226	242	(6)
Other income (expense), net . . . . .	6	—	3
Income (loss) from discontinued operations before taxes . . .	232	242	(3)
Provision (benefit) for taxes . . . . .	46	—	10
Net income (loss) from discontinued operations . . . . .	\$ 186	\$ 242	\$ (13)

(1) [https://dax-cdn.cdn.appdomain.cloud/dax-fintabnet/1.0.0/data\\_preview/index.html](https://dax-cdn.cdn.appdomain.cloud/dax-fintabnet/1.0.0/data_preview/index.html)

# Tasks

- Table Detection: find the table
- Layout Detection:
  - Find rows and columns
  - Check table boundaries
  - Find header
- Table Extraction:
  - Combine table and layout detection
  - Postprocessing

# Table detection — Rule based approach

- Extract table lines
  - Get lines with pdfplumber<sup>(1)</sup>
  - Concatenate lines
  - Get special lines: rectangles, lines of dots
- Detect page layout (one/two column)
- Find a table for each line
- Merge tables if they overlap

	Shares (in thousands)	Weighted Average Grant Price
Non-vested at October 31, 2008 . . . . .	3,640	\$34
Granted . . . . .	807	\$19
Vested . . . . .	(1,426)	\$28
Forfeited . . . . .	(90)	\$32
FY2006 LTTP Incremental Issuance . . . . .	161	\$19
Non-vested at October 31, 2009 . . . . .	3,092	\$30

acquired near the end of 2010, its financial  
and were included in "Eliminations, corporate  
The following tables present revenues,  
ating expenses as a percent of revenue,  
in millions), operating margin and selected  
sands, except yield amounts) for the years

	2017	Percent of Revenue
	\$ 7,401	100.0%
enefits	2,077	28.1
1	3,049	41.2
	353	4.8
zation	239	3.2
	225	3.1
s	143	1.9
	17	0.2
	1,214	16.4
es	7,317	98.9%
	\$ 84	
	1.1%	
	1,022	
old	\$ 24.77	

and other" in that period. The following tables present revenues,  
operating expenses, operating expenses as a percent of revenue,  
operating income (dollars in millions), operating margin and selected  
package statistics (in thousands, except yield amounts) for the years  
ended May 31:

	2017	Percent of Revenue
Revenues	\$ 7,401	100.0%
Operating expenses:		
Salaries and employee benefits	2,077	28.1
Purchased transportation	3,049	41.2
Rentals	353	4.8
Depreciation and amortization	239	3.2
Fuel	225	3.1
Maintenance and repairs	143	1.9
Intercompany charges	17	0.2
Other	1,214	16.4
Total operating expenses	7,317	98.9%
Operating income	\$ 84	
Operating margin	1.1%	
Package:		
Average daily packages	1,022	
Revenue per package (yield)	\$ 24.77	
Freight:		
Average daily pounds	3,608	
Revenue per pound (yield)	\$ 0.56	

(1) <https://github.com/jsvine/pdfplumber>

# Table detection — Model based approach

- Rule based approach does not detect tables reliable
- 2 different machine learning models
  - Yolov8s table extraction<sup>(1)</sup>
  - Microsoft table detection<sup>(2)</sup>
    - better results when changing the threshold
    - Overall better results
- Post processing to not cut off characters
  - Extend tables within threshold

(1) <https://huggingface.co/keremberke/yolov8s-table-extraction>

(2) <https://github.com/microsoft/table-transformer>

# Layout detection

- Table extraction with pdfplumber
  - Requires a bounding box to work at all
  - Tweaking the settings does not help
- Custom approach
  - Find rows based on the line-spacing
  - Find columns based on character spacing
  - Use rows/columns with pdfplumbers table extraction

	2017	Percent of Revenue
	2017	2017
Revenues	\$ 7,401	100.0%
Operating expenses:		
Salaries and employee benefits	2,077	28.1
Purchased transportation	3,049	41.2
Rentals	353	4.8
Depreciation and amortization	239	3.2
Fuel	225	3.1
Maintenance and repairs	143	1.9
Intercompany charges	17	0.2
Other	1,214	16.4
Total operating expenses	7,317	98.9%
Operating income	\$ 84	
Operating margin	1.1%	
Package:		
Average daily packages	1,022	
Revenue per package (yield)	\$ 24.77	
Freight:		
Average daily pounds	3,608	
Revenue per pound (yield)	\$ 0.56	

# Layout detection — Find Rows

- New row: line spacing > threshold
  - Line spacing: y-distance between two characters
  - Default threshold=-0,3pt
  - Find cells covering multiple lines → postprocessing

Net revenues	\$ 3,103	\$ 2,886	\$ 3,128
Income from continuing operations before income tax provision	446	463	615
Income from continuing operations	380	352	476
Loss from discontinued operations, net of tax	—	—	—

# Layout detection — Check boundaries

- Boundaries at the top and bottom are not always perfect
- Check rows at the top and bottom:
  - Remove continuous text, listings (footnotes) and centered titles

Three fiscal years ended September 29, 2007	2007	2006	2005
Numerator (in millions):			
Net income .....	\$ 3,496	\$ 1,989	\$ 1,328
Denominator (in thousands):			
Weighted-average shares outstanding, excluding unvested restricted stock .....	864,595	844,058	808,439
Effect of dilutive securities .....	24,697	33,468	48,439
Denominator for diluted earnings per share .....	889,292	877,526	856,878
Basic earnings per share .....	\$ 4.04	\$ 2.36	\$ 1.64
Diluted earnings per share .....	\$ 3.93	\$ 2.27	\$ 1.55

Potentially dilutive securities representing 13.7 million, 3.9 million, and 12.7 million shares of common stock for the years ended September 29, 2007, September 30, 2006, and September 24, 2005, respectively, were excluded from the computation of diluted earnings per share for these periods because their effect would have been antidilutive. These potentially dilutive securities include stock options, unvested restricted stock, and RSUs.

Total assets .....	8,295,422	8,264,317	7,756,892	9,537,187	8,755,270
Minority interest .....	61,935	61,756	71,774	18,435	—
Long-term debt .....	1,085,000	500,000	—	—	—
Total stockholders' equity .....	4,818,081	5,904,290	5,733,763	N/A	N/A
Total invested equity .....	N/A	N/A	N/A	8,152,629	7,554,301

(1) Includes Hotels.com revenue amounts on a gross basis. Beginning January 1, 2004, we prospectively commenced reporting revenue for Hotels.com on a net basis.



# Layout detection — Find header

- Important for multi-header tables
- Ruling lines are not consistent

	October 31, 2009				October 31, 2008			
	Cost	Gross Unrealized Gains	Gross Unrealized Losses	Estimated Fair Value	Cost	Gross Unrealized Gains	Gross Unrealized Losses	Estimated Fair Value
(in millions)								
Debt securities . . . . .	\$36	\$ —	\$ —	\$36	\$101	\$ —	\$(5)	\$ 96
Equity securities . . . . .	4	5	—	9	4	5	—	9
	<u>\$40</u>	<u>\$ 5</u>	<u>\$ —</u>	<u>\$45</u>	<u>\$105</u>	<u>\$ 5</u>	<u>\$(5)</u>	<u>\$105</u>

	October 31, 2009				October 31, 2008			
	Cost	Gross Unrealized Gains	Gross Unrealized Losses	Estimated Fair Value	Cost	Gross Unrealized Gains	Gross Unrealized Losses	Estimated Fair Value
(in millions)								
Debt securities . . . . .	\$36	\$ —	\$ —	\$36	\$101	\$ —	\$(5)	\$ 96
Equity securities . . . . .	4	5	—	9	4	5	—	9
	<u>\$40</u>	<u>\$ 5</u>	<u>\$ —</u>	<u>\$45</u>	<u>\$105</u>	<u>\$ 5</u>	<u>\$(5)</u>	<u>\$105</u>

- Instead: first occurrence of a font-change
  - Only font-changes in the upper right part of the table
  - Fallback: widest ruling line

# Layout detection — Find columns

- Divide table into multiple horizontal segments
  - Header: each row is a separate segment
  - Body (everything below the header)
- New column: character spacing > threshold
  - Character spacing: x-distance between two characters
  - Default threshold=5pt
- Extend columns to the top if they do not intersect with characters or lines

At December 31, 2010 (in millions)	Total Payments	Payments due by Period			
		2011	2012 - 2013	2014 - 2015	Thereafter
Borrowings <sup>(a)</sup>	\$ 82,862	\$ 10,323	\$ 16,031	\$ 9,223	\$ 47,285
FRBNY Credit Facility <sup>(b)</sup>	20,985	-	20,985	-	-
Interest payments on borrowings	51,940	4,531	9,532	5,963	31,914
Loss Reserves	91,151	20,235	25,157	14,074	31,685
Insurance and investment contract liabilities	462,496	18,743	32,916	30,706	380,131
Aircraft purchase commitments	13,533	282	1,742	3,523	7,986
Operating leases	2,054	429	657	422	546
Other long-term obligations <sup>(c)</sup>	365	61	95	80	129
Total <sup>(d)</sup>	\$ 725,386	\$ 54,604	\$ 107,115	\$ 63,991	\$ 499,676

# Table extraction

- Combine table and layout detection
- Post processing — Refine table
  - Merge underlying cells in the body if:
    - Cell is the only cell in the row and begins at the left of the table
    - Text does not end with a colon
    - Text cell does not end with a line sequence
    - The text in both cells has the same font

	High
Fiscal year ended December 31, 2006	
First quarter .....	\$ 42.70
Second quarter .....	36.08
Third quarter .....	27.90
Fourth quarter .....	25.69

OPERATING ACTIVITIES:	
Net income .....	
Adjustments to net income:	
Depreciation and amortization .....	

# Table extraction

- Post processing — Refine table
  - Merge underlying rows in the header if:
    - The structure of both rows is the same
    - No ruling line between both rows
  - Merge cells only containing dollar signs with the next cell

	Fiscal Year Ended		
	December 28, 2013	December 29, 2012	December 31, 2011
Net sales	100.0%	100.0%	100.0%
Cost of sales, including purchasing and warehousing costs	49.9	50.1	50.3
Gross profit	50.1	49.9	49.7
Selling, general and administrative expenses	39.9	39.3	39.0
Operating income	10.2	10.6	10.8
Interest expense	(0.6)	(0.5)	(0.5)
Other, net	0.0	0.0	0.0
Provision for income taxes	3.6	3.8	3.9
Net income	6.0%	6.2%	6.4%

# Evaluation

Metric	Custom approach	Microsoft table detection + custom layout detection	Microsoft table + layout detection
Number of tables found	10135	8366	8417
Precision	77.71 %	84.78 %	82.7 %
Recall	78.76 %	70.93 %	69.61 %
F1-Score	0.782	0.772	0.756
Cell Precision	70.57 %	75.19 %	40.06 %
Cell Recall	71.52 %	62.9 %	33.72 %
Cell F1-Score	0.71	0.685	0.366

# Problems

Domestic Life Insurance:			
Sales of fixed maturities	\$ (33)	\$ 65	\$ (4)
Sales of equity securities	17	18	7
Other:			
Foreign exchange transactions	(6)	11	—
Derivatives instruments	25	65	8
Other-than-temporary decline	(192)	(119)	(98)
Other	(26)	(5)	(33)
Total Domestic Life Insurance	\$ (215)	\$ 35	\$ (120)
Domestic Retirement Services:			
Sales of fixed maturities	\$ 1	\$ (106)	\$ 107
Sales of equity securities	31	115	30
Other:			
Foreign exchange transactions	(13)	—	—
Derivatives instruments	(33)	(12)	(14)
Other-than-temporary decline	(368)	(267)	(305)
Other	(22)	(7)	(25)
Total Domestic Retirement Services	\$ (404)	\$ (277)	\$ (207)
Foreign Life Insurance & Retirement Services:			
Sales of fixed maturities	\$ (209)	\$ 191	\$ 223
Sales of equity securities	459	281	295
Other:			
Foreign exchange transactions	106	40	(382)
Derivatives instruments	276	(599)	248
Other-than-temporary decline	(81)	(39)	(38)
Other*	156	210	26
Total Foreign Life Insurance & Retirement Services	\$ 707	\$ 84	\$ 372
Total	\$ 88	\$ (158)	\$ 45

as or and for the years ended December 31 (in millions, except per share data)

## Statement of earnings data

	2013	2012	2011
Net sales	\$18,790	\$18,380	\$17,444
Net earnings(a)	\$ 4,128	\$ 5,275	\$ 3,433
Basic earnings per share(a)	\$ 2.58	\$ 3.35	\$ 2.18

# Conclusion & Outlook

- Ruling lines very inconsistent
- More time → more rules
- Test with other datasets