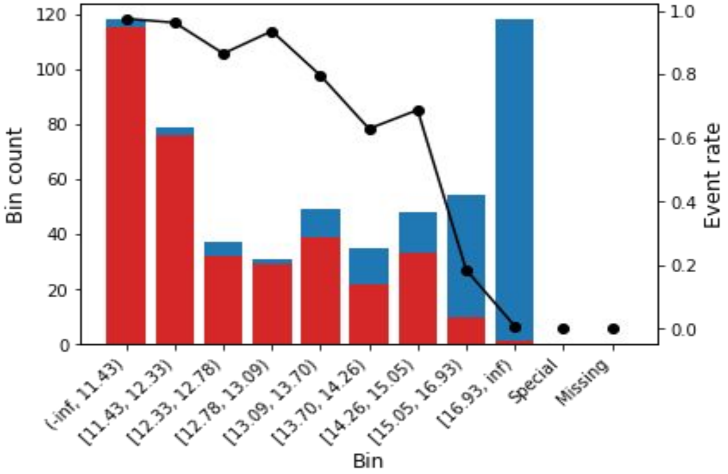
Optimal binning using Python

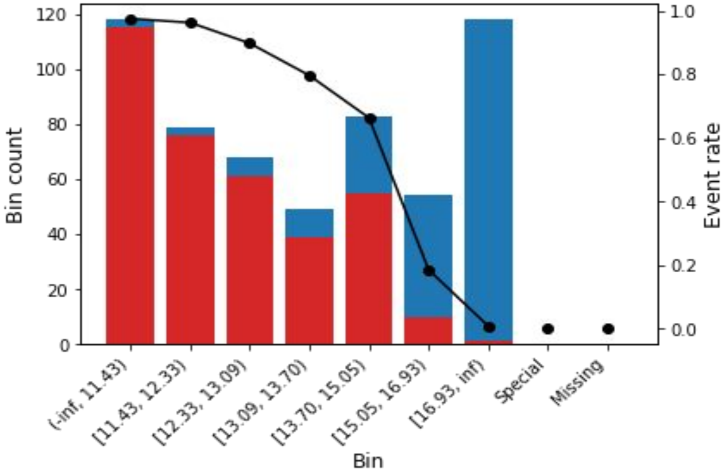
PyDay 2022

Guillermo Navas Palencia

What’s optimal binning? applications?

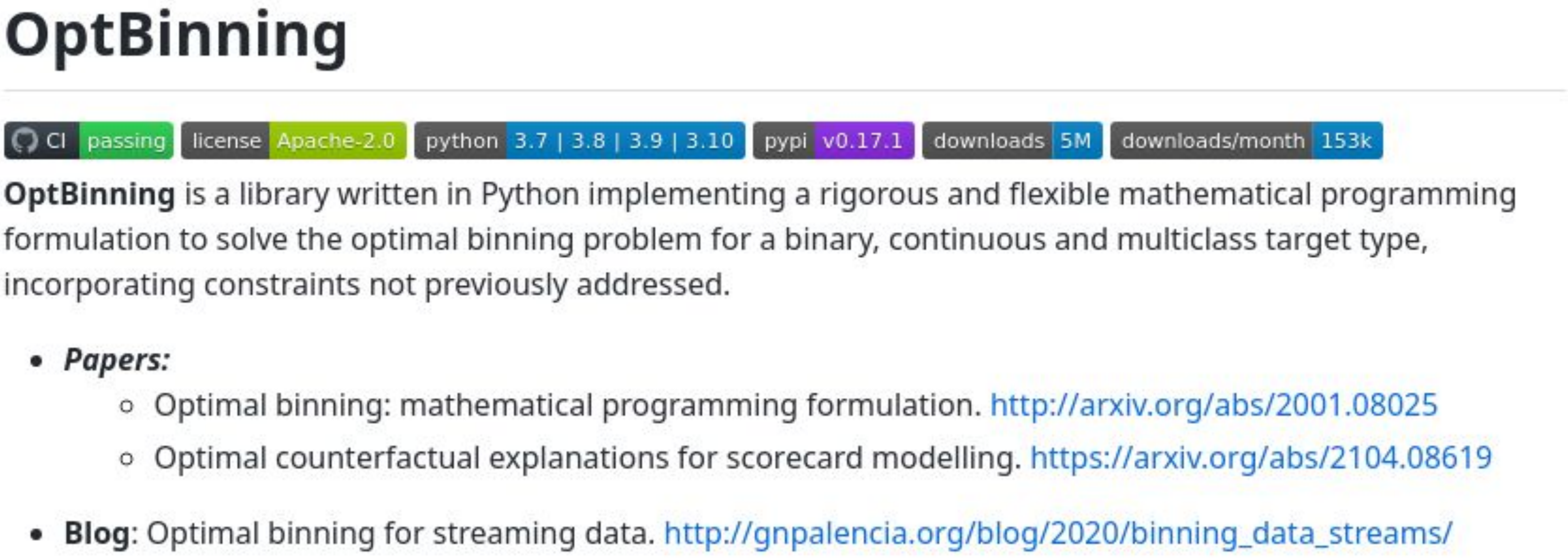


Optimization



- Mathematical optimization problem: impose constraints and maximize information value.
- Modelling non-linear relationships and preventing data issues.
- Technique to accelerate ML algorithms (Histogram-based GBM, e.g., LightGBM).
- Interpretability: widely used in finance and medical models (risk scoring models).

OptBinning: The Python Optimal Binning library



OptBinning

CI passing | license Apache-2.0 | python 3.7 | 3.8 | 3.9 | 3.10 | pypi v0.17.1 | downloads 5M | downloads/month 153k

OptBinning is a library written in Python implementing a rigorous and flexible mathematical programming formulation to solve the optimal binning problem for a binary, continuous and multiclass target type, incorporating constraints not previously addressed.

- ***Papers:***
    - Optimal binning: mathematical programming formulation. http://arxiv.org/abs/2001.08025
    - Optimal counterfactual explanations for scorecard modelling. https://arxiv.org/abs/2104.08619
- **Blog**: Optimal binning for streaming data. http://gnpalencia.org/blog/2020/binning_data_streams/

https://github.com/guillermo-navas-palencia/optbinning

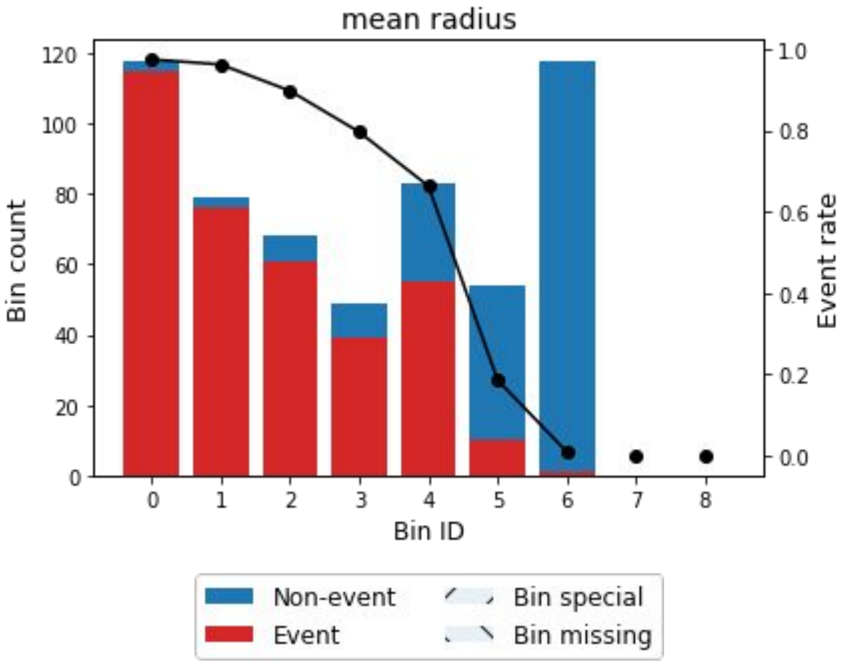
OptBinning features

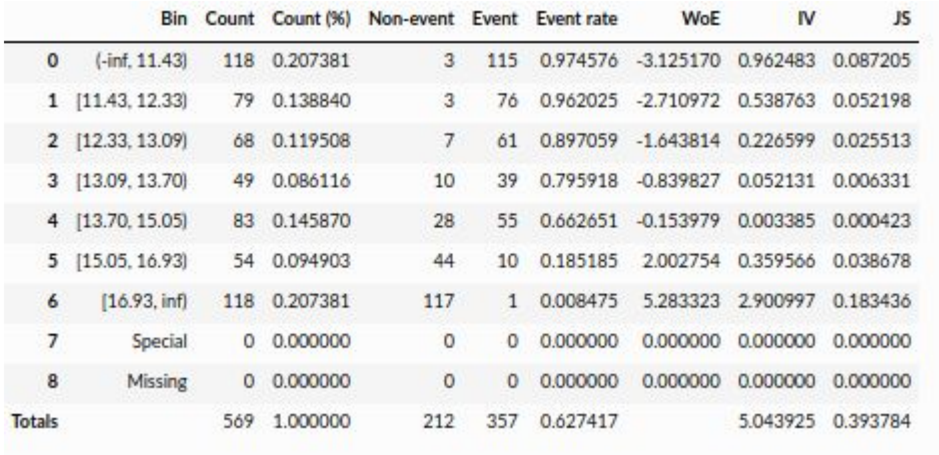
- General
  - Scikit-learn API
  - Google OR-Tools: Open-source optimization solvers

- Binning algorithms:
  - Binary, continuous and multiclass target.
  - Binning 1D/2D.
  - Piecewise polynomial binning.
  - Scenario-based binning.
  - Batch and stream binning.

- Scorecard modelling
  - Binary and continuous target.
  - Counterfactual explanations.
  - Combine 1D and 2D binning (coming soon).

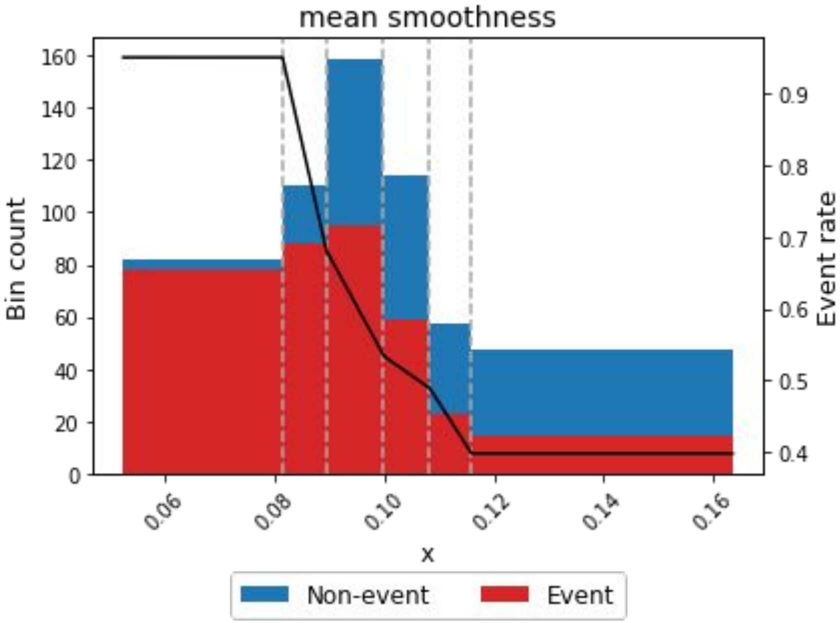
Examples (binary target)

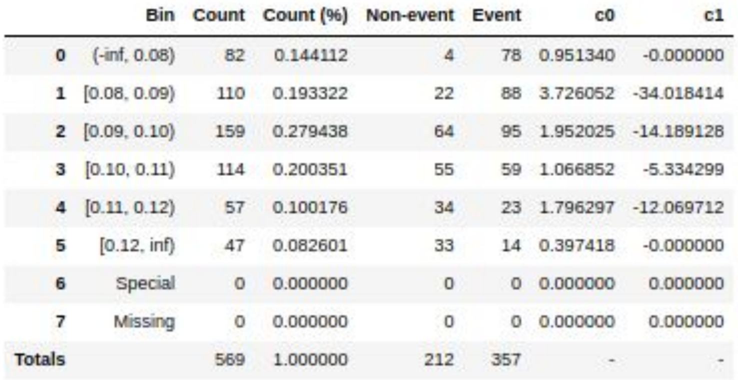




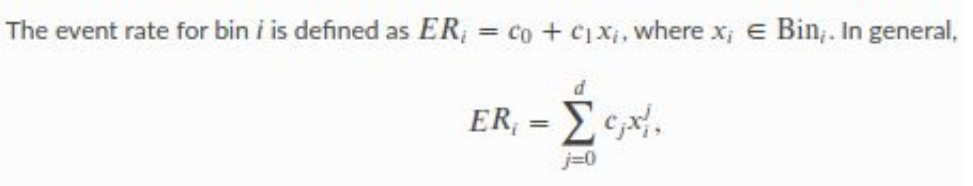
	Bin	Count	Count (%)	Non-event	Event	Event rate	WoE	IV	JS
0	(-inf, 11.43)	118	0.207381	3	115	0.974576	-3.125170	0.962483	0.087205
1	[11.43, 12.33)	79	0.138840	3	76	0.962025	-2.710972	0.538763	0.052198
2	[12.33, 13.09)	68	0.119508	7	61	0.897059	-1.643814	0.226599	0.025513
3	[13.09, 13.70)	49	0.086116	10	39	0.795918	-0.839827	0.052131	0.006331
4	[13.70, 15.05)	83	0.145870	28	55	0.662651	-0.153979	0.003385	0.000423
5	[15.05, 16.93)	54	0.094903	44	10	0.185185	2.002754	0.359566	0.038678
6	[16.93, inf)	118	0.207381	117	1	0.008475	5.283323	2.900997	0.183436
7	Special	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000
8	Missing	0	0.000000	0	0	0.000000	0.000000	0.000000	0.000000
Totals		569	1.000000	212	357	0.627417		5.043925	0.393784

Examples (binary target - piecewise polynomial)





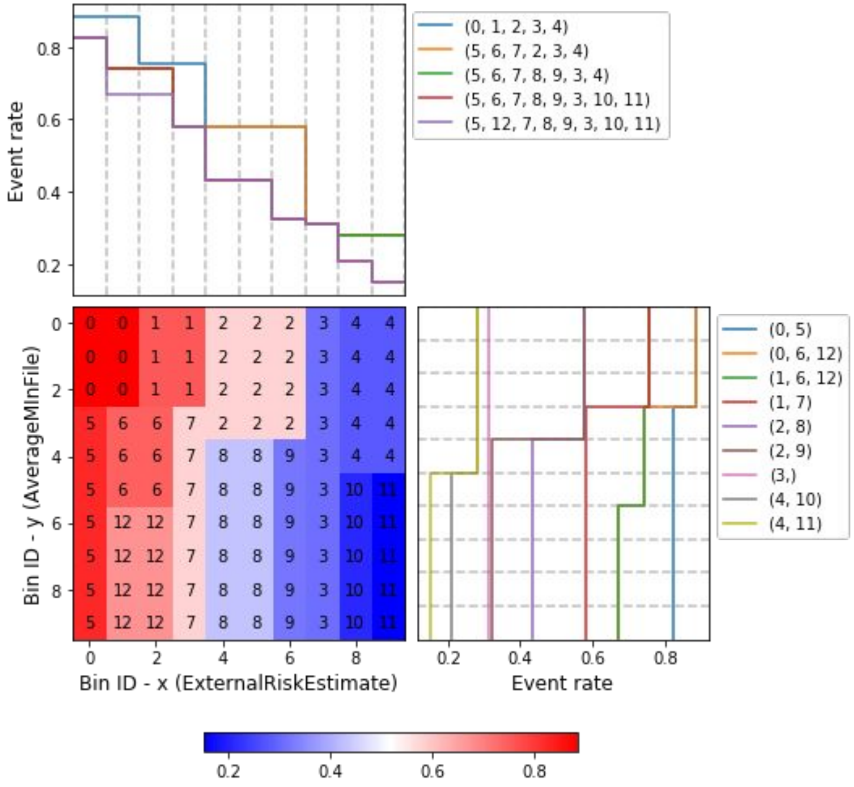
	Bin	Count	Count (%)	Non-event	Event	c0	c1
0	(-inf, 0.08)	82	0.144112	4	78	0.951340	-0.000000
1	[0.08, 0.09)	110	0.193322	22	88	3.726052	-34.018414
2	[0.09, 0.10)	159	0.279438	64	95	1.952025	-14.189128
3	[0.10, 0.11)	114	0.200351	55	59	1.066852	-5.334299
4	[0.11, 0.12)	57	0.100176	34	23	1.796297	-12.069712
5	[0.12, inf)	47	0.082601	33	14	0.397418	-0.000000
6	Special	0	0.000000	0	0	0.000000	0.000000
7	Missing	0	0.000000	0	0	0.000000	0.000000
Totals		569	1.000000	212	357	-	-

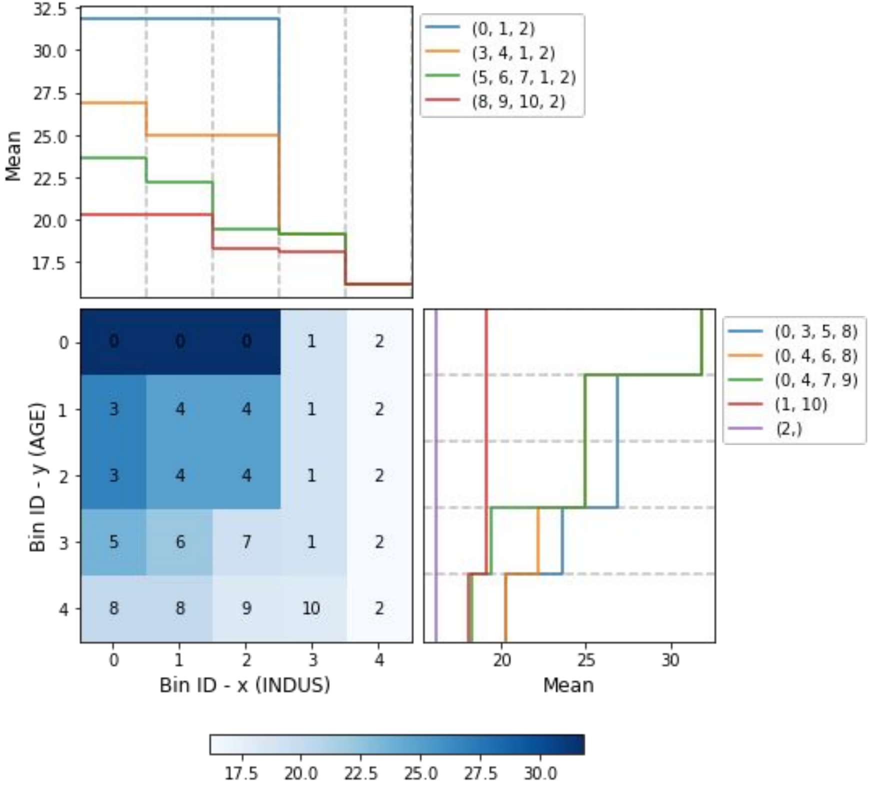


The event rate for bin i is defined as $ER_i = c_0 + c_1 x_i$, where $x_i \in \text{Bin}_i$. In general,

$$ER_i = \sum_{j=0}^{d} c_j x_i^j,$$

Examples (binning 2D binary and continuous target)





OptBinning: official users

- Fintech
    - Jeitto (BNPL - Brasil)
    - Bilendo (Credit Risk Software - Germany)
    - Aplazame (BNPL - Spain)
    - Praelexis Credit (Credit Risk Software - South Africa)
    - Risika (Credit Risk Software - Denmark)
    - Tamara (BNPL - Saudi Arabia)

- Software
    - Loginom (Low-code - Russia)

- Banks and financial institutions
    - ING
    - Morningstar
    - BBVA AI Factory
    - N26
    - +