Accurate, Data-Efficient Learning from Noisy, Choice-Based Labels for Inherent Risk Scoring

W. Ronny Huang Miguel A. Perez

7 December 2018

On a scale from 0 to 1, what is the risk of this to your health?

Methodology

Risk to health

	Max	Min
Green Beans	☐	☑
Cheesecake	☑	☐
Poutine	☐	☐

Methodology

Risk to health

	Max	Min
Green Beans	☐	☑
Cheesecake	☑	☐
Poutine	☐	☐

Risk to health

	Max	Min
Pecan Pie	☑	☐
Kale	☐	☑
Green Beans	☐	☐

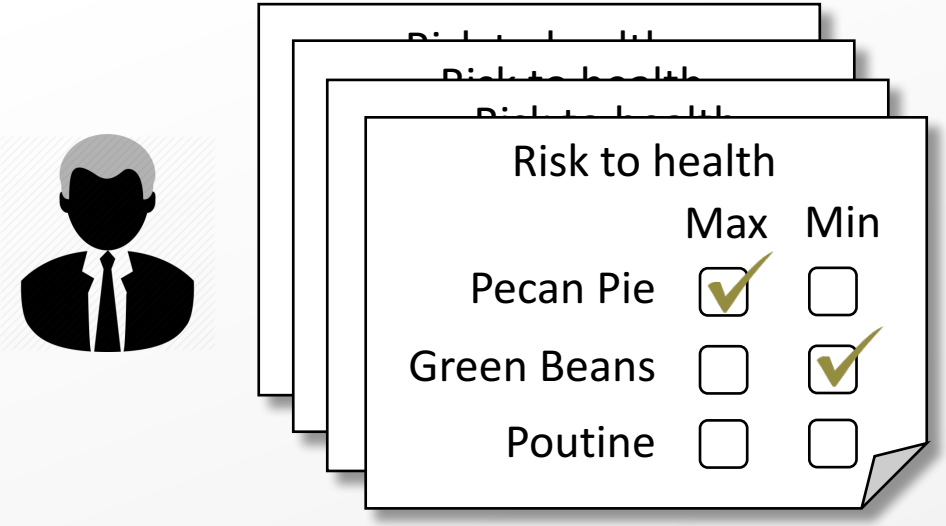
Risk to health

	Max	Min
Cheesecake	☑	☐
Poutine	☐	☐
Kale	☐	☑

Risk to health

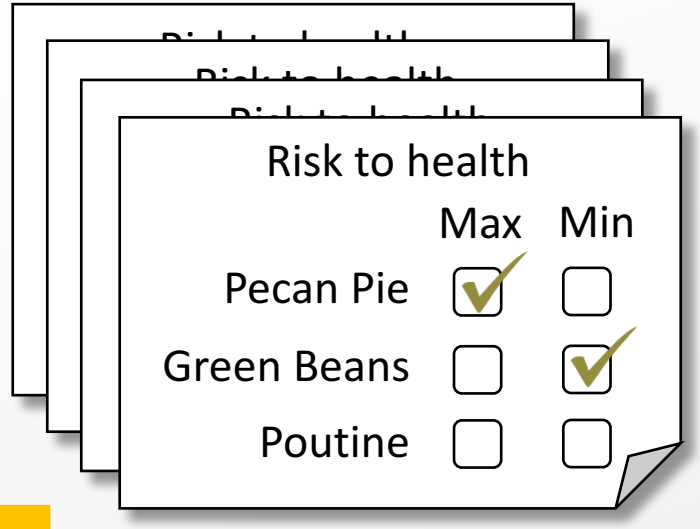
	Max	Min
Pecan Pie	☑	☐
Green Beans	☐	☑
Poutine	☐	☐

Methodology



	Absolute risk
Cheesecake	0.87
Green Beans	0.35
Poutine	0.54
Pecan Pie	0.92
Kale	0.23

Methodology



Choice to Score

Absolute risk

Cheesecake	0.87
Green Beans	0.35
Poutine	0.54
Pecan Pie	0.92
Kale	0.23

Choice Set Diversifier

Our Contributions

Applied to Risk Scoring in the Wild

Choice-to-Score

Choice set encoding

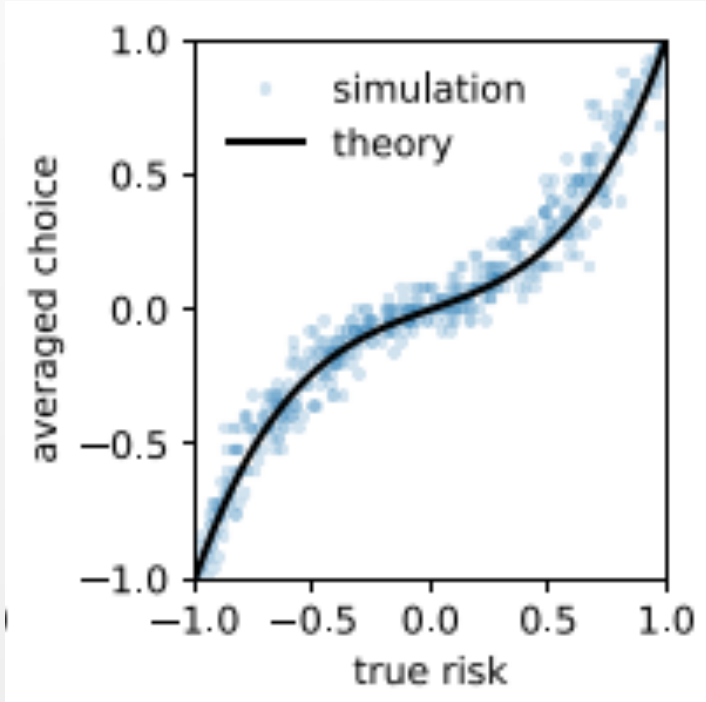
$$c(y_i|l) = \begin{cases} 1, & \text{if } y_i = \max_{j \in S_k} y^j \\ -1, & \text{if } y_i = \min_{j \in S_k} y^j \\ 0, & \text{otherwise} \end{cases}$$

Compute Mean Choice

$$\bar{c}_i(y_i) = \frac{1}{q} \sum_{l=1}^{q} c(y_i|l)$$

Expected Choice

$$\langle c(y_i) \rangle = \left(\int_{-\infty}^{y_i} f(y')\, \mathrm{d}y' \right)^{s-1} - \left(\int_{y_i}^{\infty} f(y')\, \mathrm{d}y' \right)^{s-1}$$



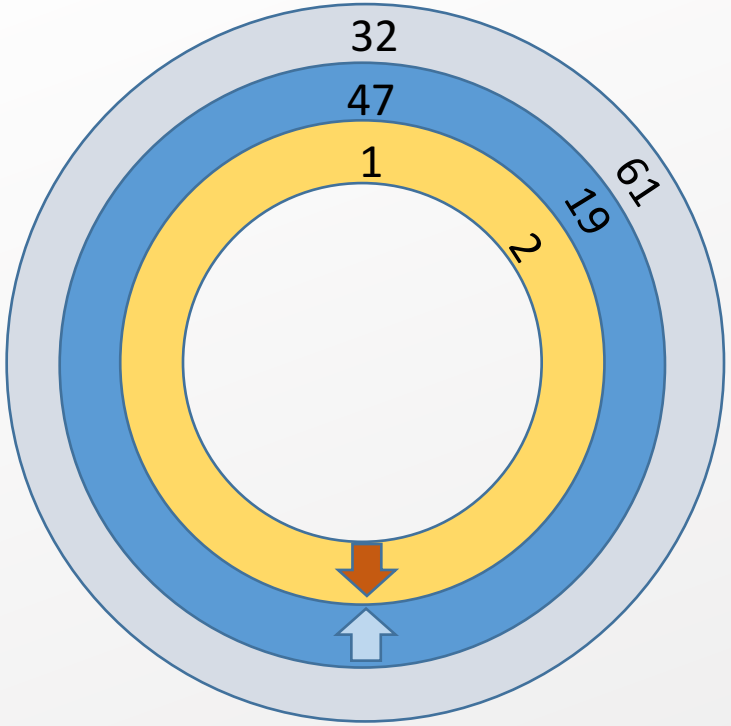
Choice Set Diversifier

Desired Property

Minimize the number of repeated profile key pairs for all generated questions

Algorithmic Guarantee

Select a Choice set size of 4. Compute the number of unique questionnaires that can be generated from the constructed *group representation*



$$\begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \end{pmatrix}$$

Application to Inherent Risk Scoring

Background: Understanding Know Your Customer (KYC) Risk is essential for the financial services industry.

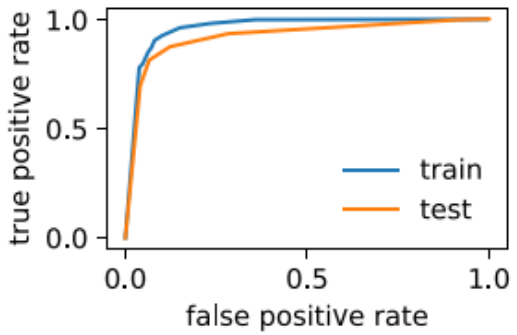
Problem: Modeling KYC risk is difficult because of lack of data and poor data quality.

Our Solution

Label synthetic examples needed to build a model that mimics human expert evaluation.

Out of Sample Performance

Population Group	Profiles	SMA Escalations	Escalation Rate
IRM Selected Alerted Profiles	1,500	28	1.87
Remaining Scenario Alerted Profiles	2,500	3	0.12



Metric	Train	Test
AUC	97%	93%
Classification Error	8%	11%

Takeaways

Choice to Score relation converts relative information to absolute information about risk

Choice Set Diversifier makes our training set data-efficient

Results Good performance on real-world data by a model trained with choice-based labels

https://arxiv.org/abs/1811.10791

Thank you!

Please visit our poster:

Accurate, Data-Efficient Learning from Noisy, Choice-Based Labels for Inherent Risk Scoring

EY