Google

Modeling labels for conversion value prediction

Ashwinkumar Badanidiyuru
Guru Guruganesh

Google Research
Google Research

ADKDD 2021

Conversion value prediction - Problem definition

- User does activity $A_1, A_2, \ldots, A_n$ with value $l_1, l_2, \ldots, l_n$ on advertiser website

A_i=purchase with l_i=Dollar spent

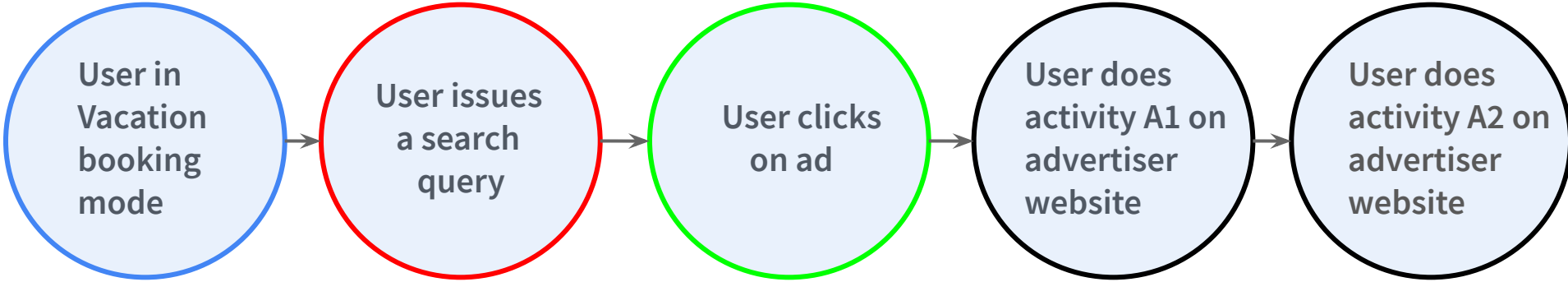
A_i=visit with l_i=time spent

A_i=email signup, l_i=long term value estimate

Conversion value prediction - Problem definition

- Goal: Optimize for total value of all events $l = \sum_i l_i$

Conversion value prediction - Problem definition



- Means: In performance based advertising
  supervised machine learning model to predict E[total value | click]
  bid ~ E[total value | click]
- Challenges
  - Different advertisers report label in different scale.
    - Someone might report time spent and someone else dollars spent
    - Someone might report in dollars and someone else in a different currency
  - Overfitting to outliers
  - E[total value | click] predicts mean of the distribution and neglects information in the complete distribution

Existing literature

- Proper scoring rules and loss functions which give unbiased estimators.
  Brier 1950, many other follow up papers
  Loss functions of the form Gradient(loss function) = f(prediction) * (prediction - label)

- Mean estimation with heavy tailed distributions
  A survey by Lugosi, Mendelssohn 2019
  Median of means technique

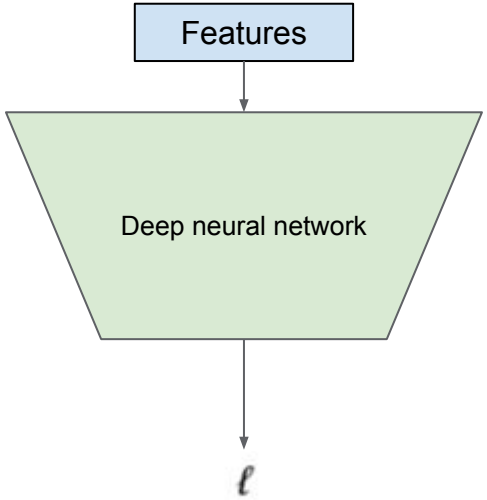
- Using information in the distribution of label
  Zero inflated poisson (ZIP) regression to handle zero inflation

- Allocate model capacity to optimize for final business metric
  Vasile, Lefortier, Chapelle. 2017. Cost-sensitive Learning for Utility Optimization in Online Advertising Auctions.
  Reweight the examples

Handling label scale

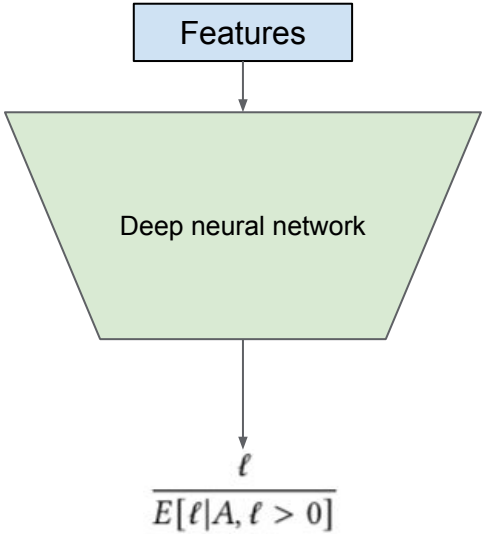
Handling label scale - Label normalization

Train against label



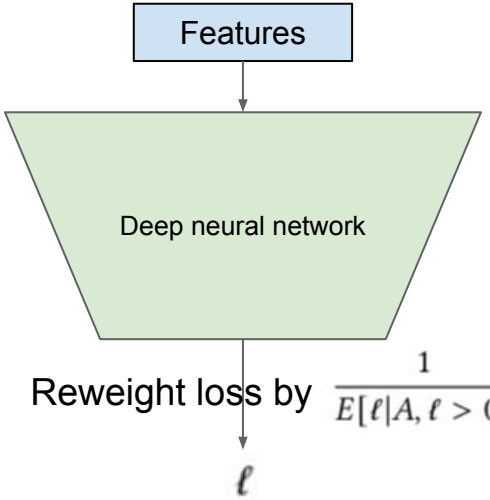
Gradient ~ $prediction - \ell$

Train against normalized label



Gradient ~ $prediction - \frac{\ell}{E[\ell|A, \ell > 0]}$

Reweight Examples



Reweight loss by $\frac{1}{E[\ell|A, \ell > 0]}$

Numerically unstable training

Gradient ~ $\frac{prediction - \ell}{E[\ell|A, \ell > 0]}$

Blows up when
prediction ~ constant
$E[\ell|A, \ell > 0]$ is very small

A - Advertiser

Handling label scale - Label normalization

Models	Relative negative log likelihood improvement	
	against label	against normalized label
S=Model trained against label	0.0%	0.0%
BN=Model trained against normalized label	+1.53%	-38.02%

Table 1: Negative log likelihood for label normalization

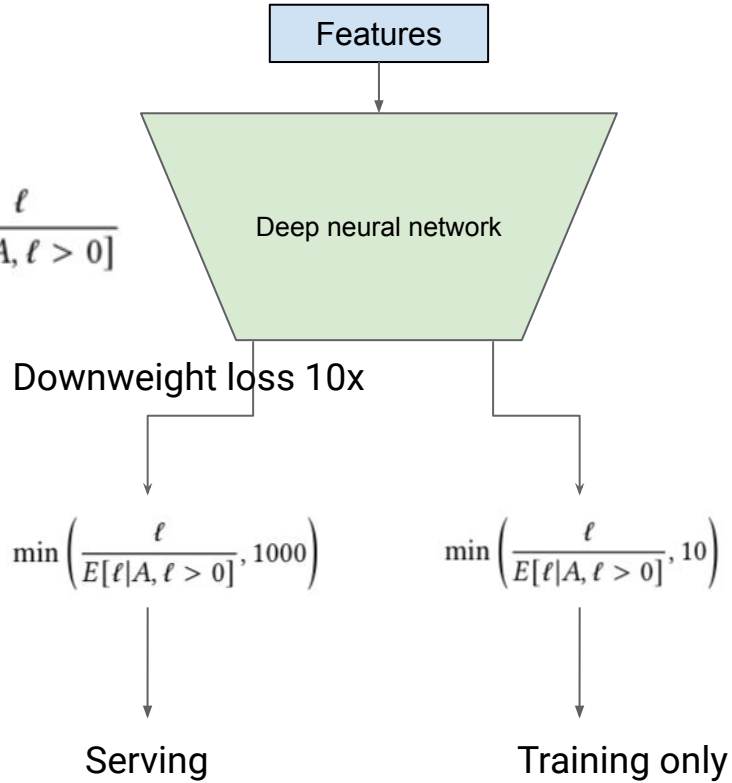
Increasing value of bucketized $E[\ell \mid A, \ell > 0]$	**S** un-normal-ized label	**BN** un-normal-ized label	**S** normal-ized label	**BN** normal-ized label
Bucket0	2.36	1.04	17.94	0.99
Bucket1	1.29	0.85	1.75	0.99
Bucket2	1.03	1.0	1.03	1.0
Bucket3	1.02	1.0	1.02	1.0
Bucket4	1.01	1.01	1.01	1.0
Bucket5	1.03	1.10	1.11	1.04
Bucket6	1.0	1.04	1.01	1.01

Table 2: Avg Prediction/Avg Label

Learning properties of the distribution via multi-task learning

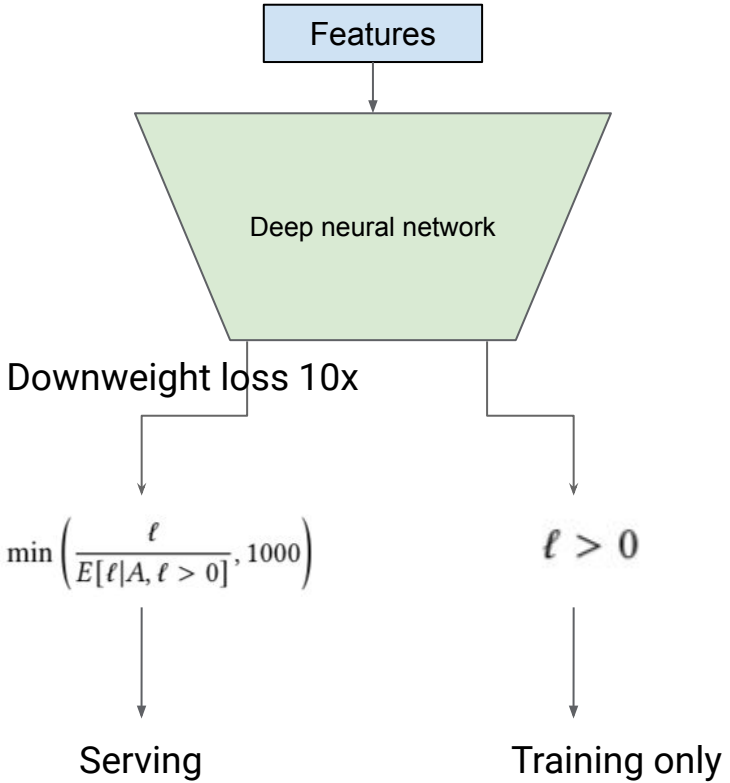
Handling outliers

- Most normalized labels are between 0 and 10.
- But a small fraction can be much larger due to outliers
- Can cause overfitting because Gradient ~ $prediction - \frac{\ell}{E[\ell|A, \ell > 0]}$



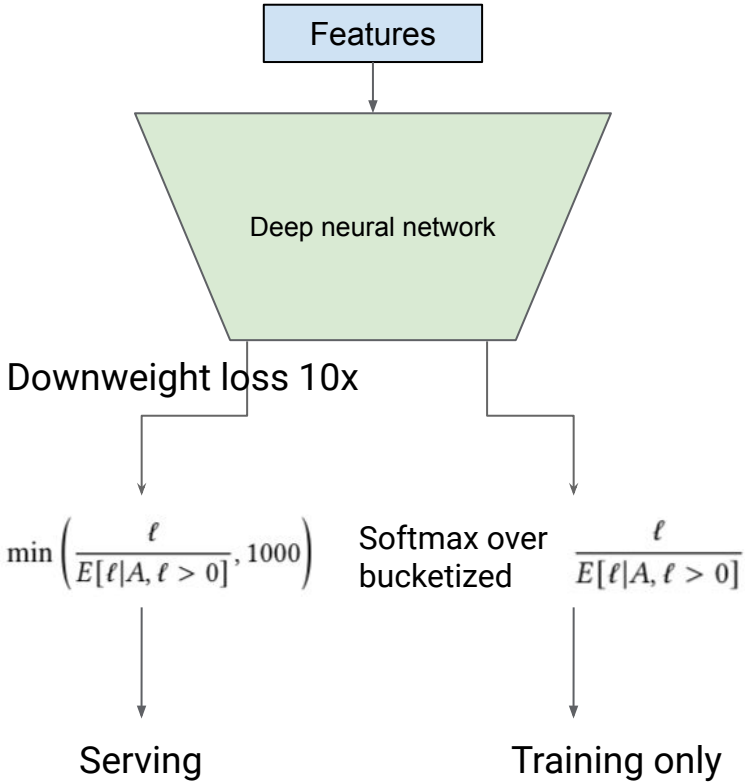
Handling zero inflation

- A very large fraction of clicks have value 0
- Zero inflated poission regression is one way to get better accuracy
- Our solution add a label>0 logistic regression head

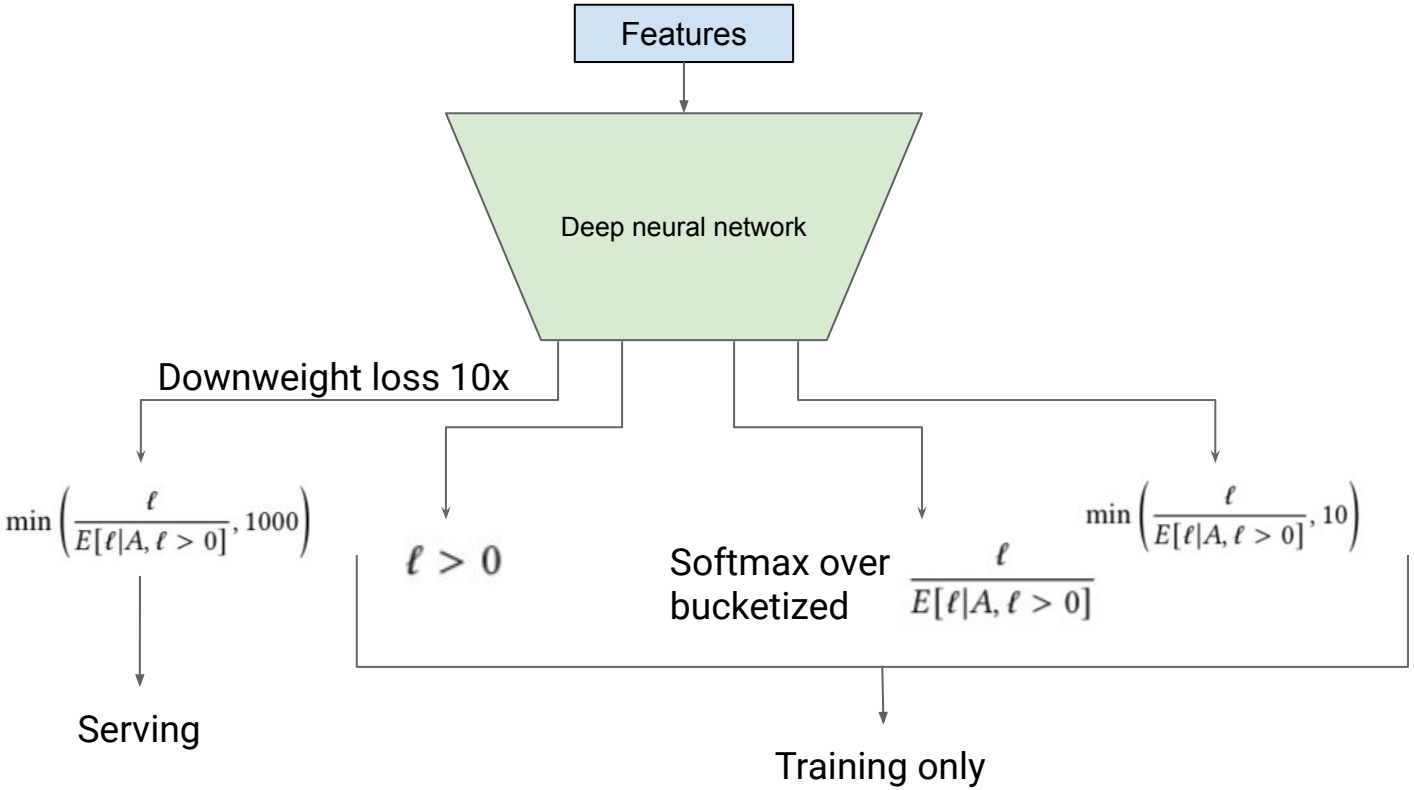


Learning distribution of labels and not just mean

- Regression predicts mean of the distribution conditioned on the covariates.
- Information present in the entire distribution itself
- Our solution add a softmax head over bucketized normalized label



All ideas together



Experiments

Experiments

- Baseline Model (BN) - none of the extra heads
- Full Model - (F) - All the ideas with all the extra heads

Experiments

- Baseline Model (BN) - none of the extra heads
- Full Model - (F) - All the ideas with all the extra heads

Ablation

- Median of Means - (MM)
- Full model without $\ell > 0$ head - (FP)
- Full model without Softmax Head - (FS)
- Full model without Winsorized Label Head (serving head weighted uniformly) - (FW1)
- Full model without Winsorized Label Head (serving head weighted down by 10x) - (FW2)
- Zero Inflated FullModel - (ZI)

Experiments

- Baseline Model (BN) - none of the extra heads
- Full Model - (F) - All the ideas with all the extra heads

Ablation

- Median of Means - (MM)
- Full model without $\ell > 0$ head - (FP)
- Full model without Softmax Head - (FS)
- Full model without Winsorized Label Head (serving head weighted uniformly) - (FW1)
- Full model without Winsorized Label Head (serving head weighted down by 10x) - (FW2)
- Zero Inflated FullModel - (ZI)

Models	Relative Negative log likelihood	
	All traffic	Advertisers with >2% winsorized +ve labels
BN	0.0%	0.0%
F	**-0.87%**	**-1.23%**
MM	+0.04%	0.12%
FP	-0.67%	-1.19%
FS	-0.79%	-1.07%
FW1	-0.50%	-0.81%
FW2	-0.50%	-0.33%
ZI	-0.47%	-0.75%

Table 3: Relative improvement in negative log likelihood

Q&A