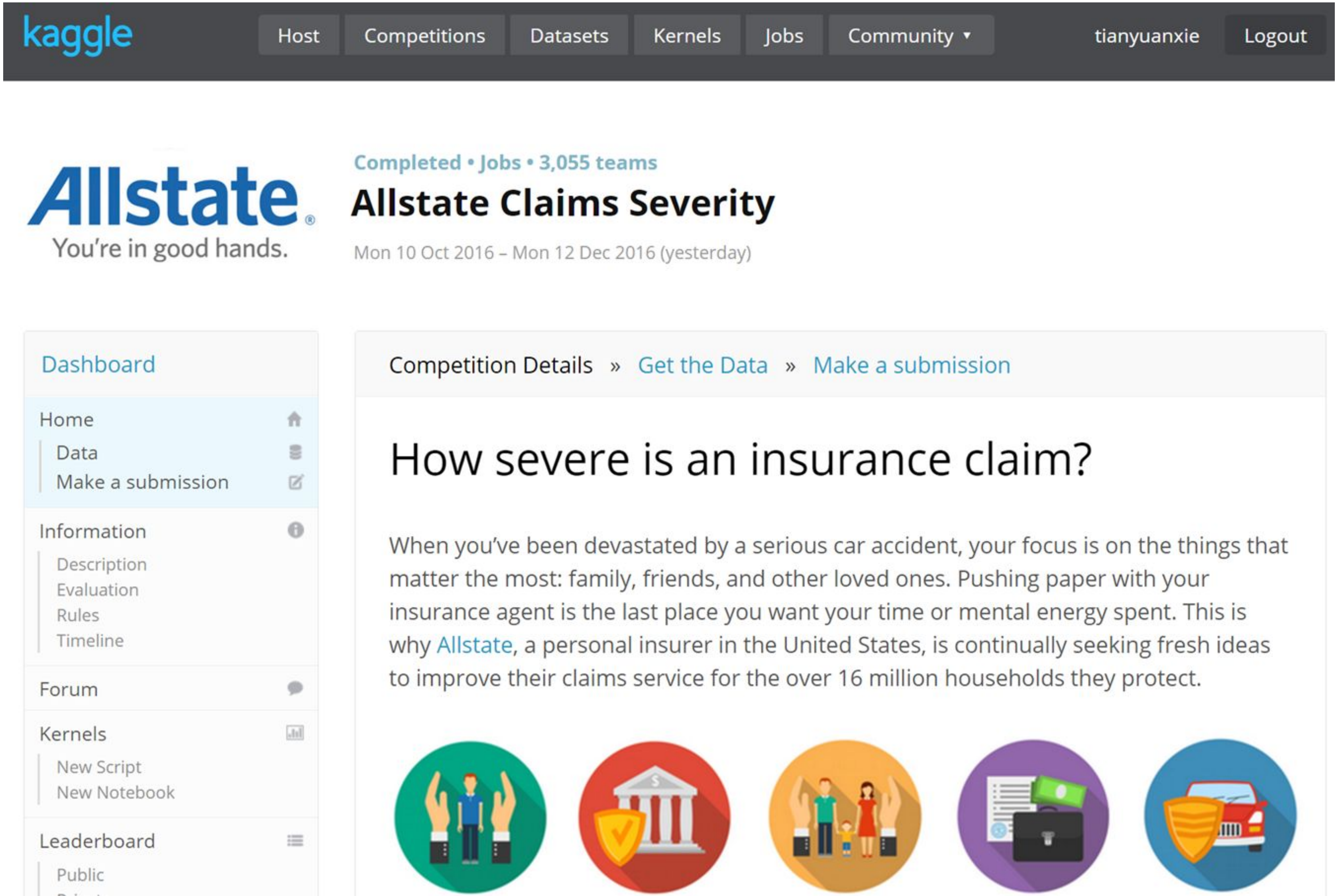
How severe is an insurance claim?

Team Member

Siyu Nan, Li Ding, Tianyuan Xie

Project Objective



Data Overlook

	id	cat1	cat2	cat3	cat4	cat5	cat6	cat7	cat8	cat9	cat10	cat11	cat12	cat13
0	1	A	B	A	B	A	A	A	A	B	A	B	A	A
1	2	A	B	A	A	A	A	A	A	B	B	A	A	A
2	5	A	B	A	A	B	A	A	A	B	B	B	B	B
3	10	B	B	A	B	A	A	A	A	B	A	A	A	A
4	11	A	B	A	B	A	A	A	A	B	B	A	B	A

	cat14	cat15	cat16	cat17	cat18	cat19	cat20	cat21	cat22	cat23	cat24	cat25
0	A	A	A	A	A	A	A	A	A	B	A	A
1	A	A	A	A	A	A	A	A	A	A	A	A
2	A	A	A	A	A	A	A	A	A	A	A	A
3	A	A	A	A	A	A	A	A	A	B	A	A
4	A	A	A	A	A	A	A	A	A	B	A	A

⋮ **188318 * 132**

	cat109	cat110	cat111	cat112	cat113	cat114	cat115	cat116	cont1	cont2
0	BU	BC	C	AS	S	A	O	LB	0.726300	0.245921
1	BI	CQ	A	AV	BM	A	O	DP	0.330514	0.737068
2	AB	DK	A	C	AF	A	I	GK	0.261841	0.358319
3	BI	CS	C	N	AE	A	O	DJ	0.321594	0.555782
4	H	C	C	Y	BM	A	K	CK	0.273204	0.159990

	cont3	cont4	cont5	cont6	cont7	cont8	cont9
0	0.187583	0.789639	0.310061	0.718367	0.335060	0.30260	0.67135
1	0.592681	0.614134	0.885834	0.438917	0.436585	0.60087	0.35127
2	0.484196	0.236924	0.397069	0.289648	0.315545	0.27320	0.26076
3	0.527991	0.373816	0.422268	0.440945	0.391128	0.31796	0.32128
4	0.527991	0.473202	0.704268	0.178193	0.247408	0.24564	0.22089

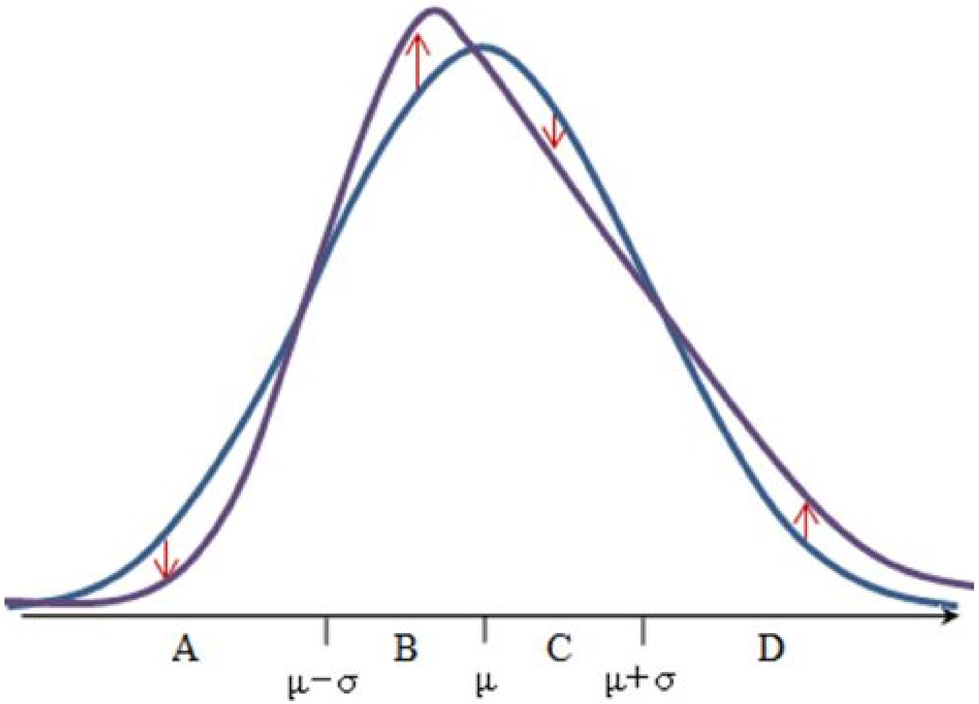
	cont10	cont11	cont12	cont13	cont14	loss
0	0.83510	0.569745	0.594646	0.822493	0.714843	2213.18
1	0.43919	0.338312	0.366307	0.611431	0.304496	1283.60
2	0.32446	0.381398	0.373424	0.195709	0.774425	3005.09
3	0.44467	0.327915	0.321570	0.605077	0.602642	939.85
4	0.21230	0.204687	0.202213	0.246011	0.432606	2763.85

Attributes?

Overview of Our Approaches

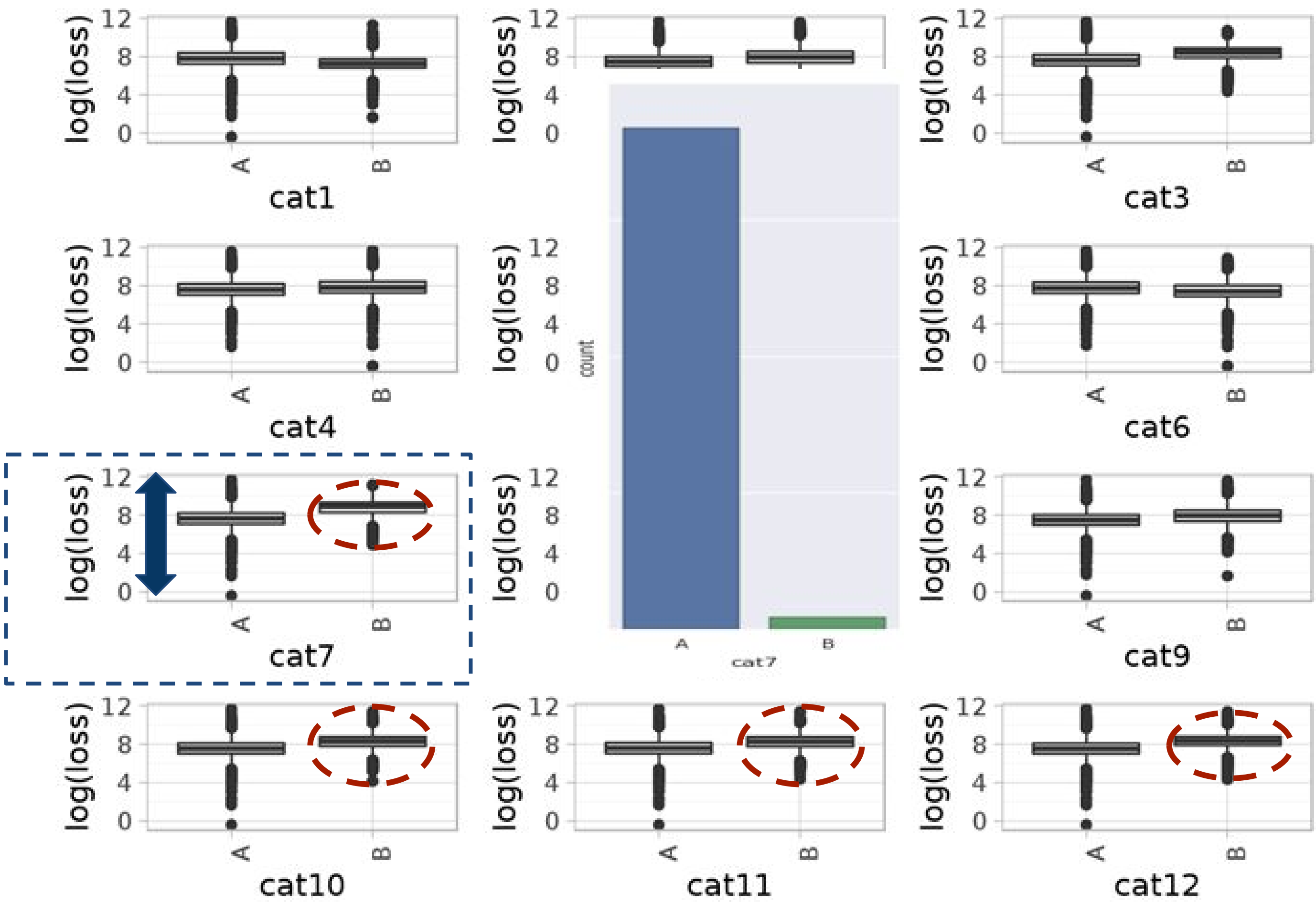
- Data Preprocessing
  - Categorical
  - Numerical
- Machine Learning Algorithm
  - Baseline: Linear Regression
  - Improvement
    - Linear Regression with l_1-loss objective function
    - Interaction between variables
    - XG-Boosting
    - Deep Learning
    - Neural Networks + XG-Boosting
- Result
- Conclusion

Data Preprocessing



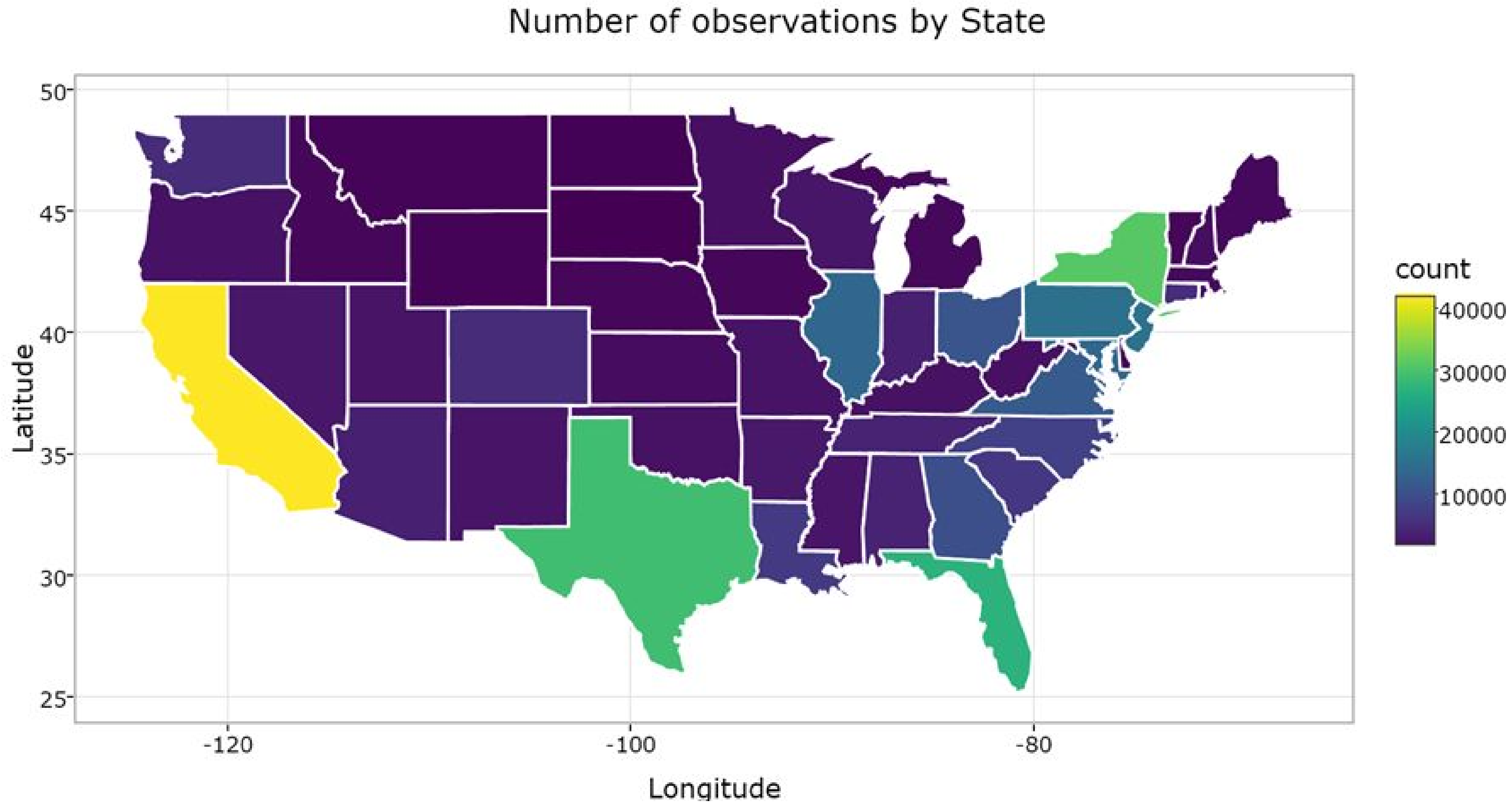
Categorical Feature Analysis

- *cat1 ~ cat72 have only two labels A and B. In most of the cases, B has very few entries*



Categorical Feature Analysis

- *cat73 ~ cat 116 have more than two labels*
- cat112 has 51 levels (50 + DC). California is well sampled, as well as some central states



Categorical Data Conversion: **One-Hot Encoding Technique**

Convert categorical to numerical data before doing linear regression, using Dummy variables.

Example: cat92 : [A, B, C, D, F, H, I]

- One way: A = 1, B = 2, …, F = 6,... I = 9
  - Confusing to algorithm
  - Meaningless, eg. location
- **One-Hot Encoding**

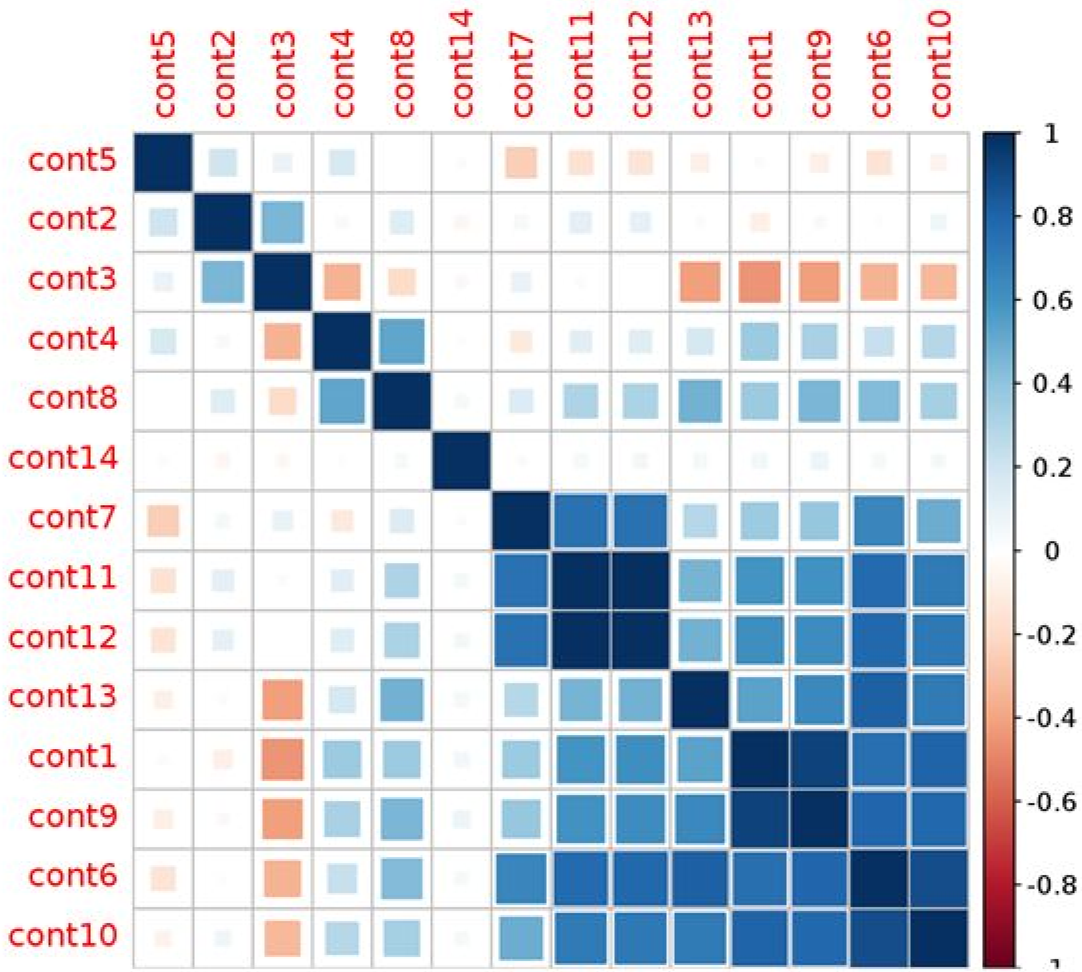
id	cat92
3	A
34	B

id	A	B	C	D	F	H	I
3	**1**	0	0	0	0	0	0
34	0	**1**	0	0	0	0	0

Number of features: 130 → 1176

cats are done

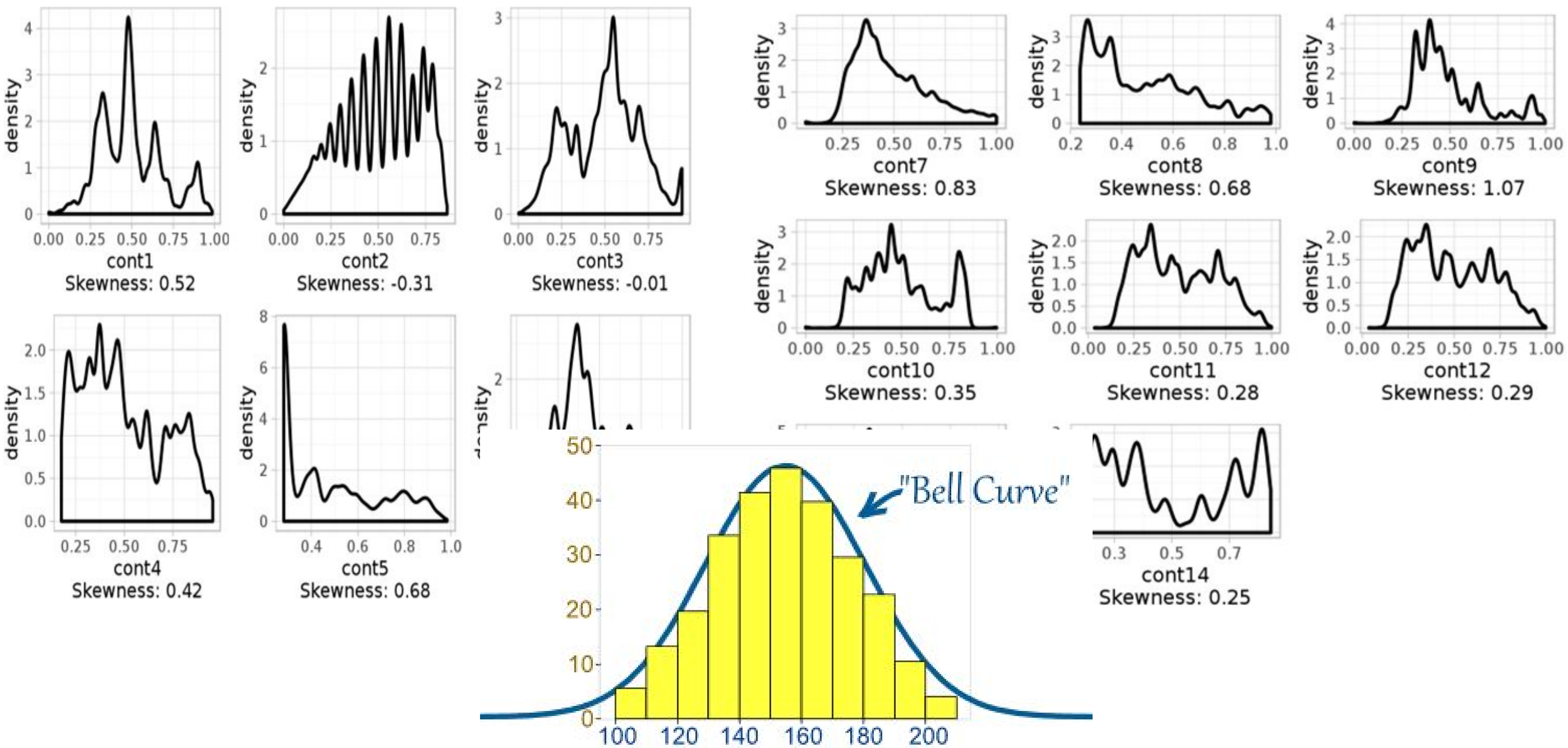
Continuous Feature Correlation



Highly correlated

cont11 and cont12 = 0.99
cont1 and cont9 = 0.93
cont6 and cont10 = 0.88
cont6 and cont13 = 0.82
cont1 and cont10 = 0.81
cont6 and cont9 = 0.80
cont9 and cont10 = 0.79
cont6 and cont12 = 0.79
cont6 and cont11 = 0.77
cont1 and cont6 = 0.76
cont7 and cont11 = 0.75
cont7 and cont12 = 0.74
cont10 and cont12 = 0.71
cont10 and cont13 = 0.71
cont10 and cont11 = 0.70
cont6 and cont7 = 0.66
cont9 and cont13 = 0.64
cont9 and cont12 = 0.63
cont1 and cont12 = 0.61
cont9 and cont11 = 0.61
cont1 and cont11 = 0.60
cont1 and cont13 = 0.53
cont4 and cont8 = 0.53

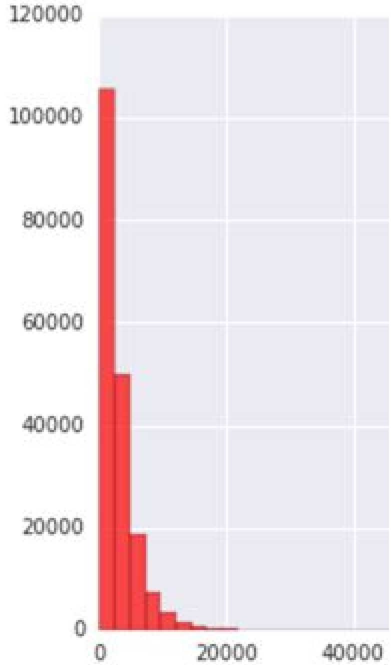
Continuous Feature Analysis -- Skewness

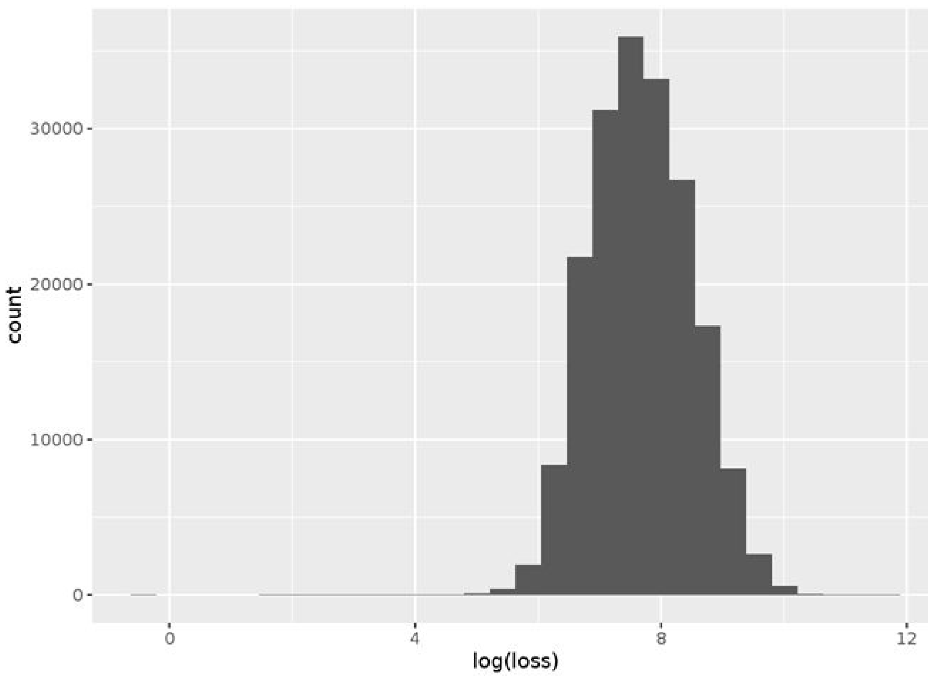


Continuous Feature Skewness Correction

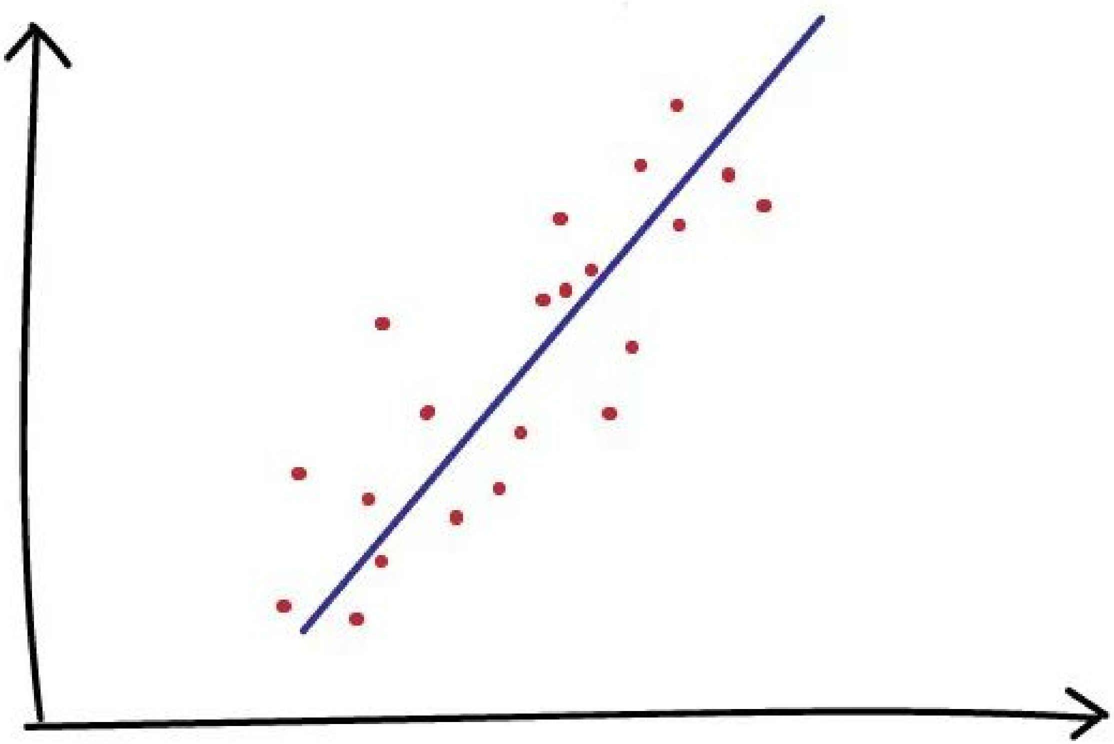
loss → log(loss + shift)

	skewness
cont1	0.516424
cont2	-0.310941
cont3	-0.010002
cont4	0.416096
cont5	0.681622
cont6	0.461214
cont7	0.826053
cont8	0.676634
cont9	1.072429
cont10	0.355001
cont11	0.280821
cont12	0.291992
cont13	0.380742
cont14	0.248674
loss	3.794958





Linear Algorithm



Baseline: Linear Regression

$$\min_{x \in R^{1176*1}} \frac{1}{2}||y - \hat{y}||^2$$

$$where \quad \hat{y} = Ax + b$$

$$Closed \quad form: x = (A^T A)^{-1} A^T (y - b)$$

- y: Real cost of insurance (188,318 x 1)
- A: Matrix we have (188,318 x 1176)
- x: Attribute vector (1176 x 1)

Mean Absolute Error: 1278

Other Attempts

LASSO:

$$\min_{x \in R^{1176*1}} \frac{1}{2}||y - \hat{y}||^2 + \lambda||x||_1$$

$$where \quad \hat{y} = Ax + b$$

Mean Absolute Error: 1262.5

Ridge Regression:

$$\min_{x \in R^{1176*1}} \frac{1}{2}||y - \hat{y}||^2 + \lambda||x||_2$$

$$where \quad \hat{y} = Ax + b$$

Mean Absolute Error: 1267

Elastic Net Regression:

$$\min_{x \in R^{1176*1}} \frac{1}{2}||y - \hat{y}||^2 + \lambda_1||x||_1 + \lambda_2||x||_2$$

$$where \quad \hat{y} = Ax + b$$

Mean Absolute Error: 1260

- y: Real cost of insurance (188,318 x 1)
- A: Matrix we have (188,318 x 1176)
- x: Attribute vector (1176 x 1)

Improvement:

Linear Regression with l_1-loss objective function

Objective Function:

$$\min_{x \in R^{1176*1}} ||y - \hat{y}||_1$$

where $\hat{y} = Ax + b$

Closed form by SGD: $$x = \begin{cases} x - \gamma A_i & when \quad y - A_i x - b < 0 \\ x + \gamma A_i & when \quad y - A_i x - b > 0 \\ [-A_i, \mathrm{A}_i] & when \quad y - A_i x - b = 0 \end{cases}$$

- y: Real cost of insurance (188,318 x 1)
- A: Matrix we have (188,318 x 1176)
- x: Attribute vector (1176 x 1)

Mean Absolute Error: 1239

Tree-based Method

Tree-based Methods

Single Tree (CART) MAE: **1741** (by Santhosh Sharma)

Tree Ensemble:

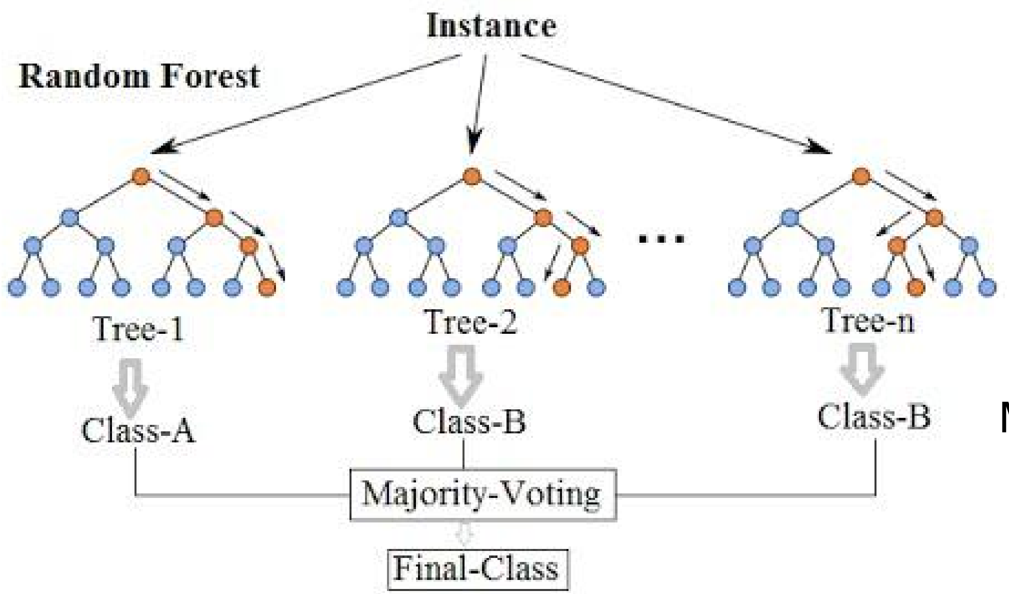
$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

f_k represents each tree

$\mathcal{F}$ represents the set of all possible trees

$$\text{obj}(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Benchmark: **Random Forest with CART**



Mean Absolute Error: **1228**

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}$$

$$\text{obj}(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Dominating: **XG-Boosting**

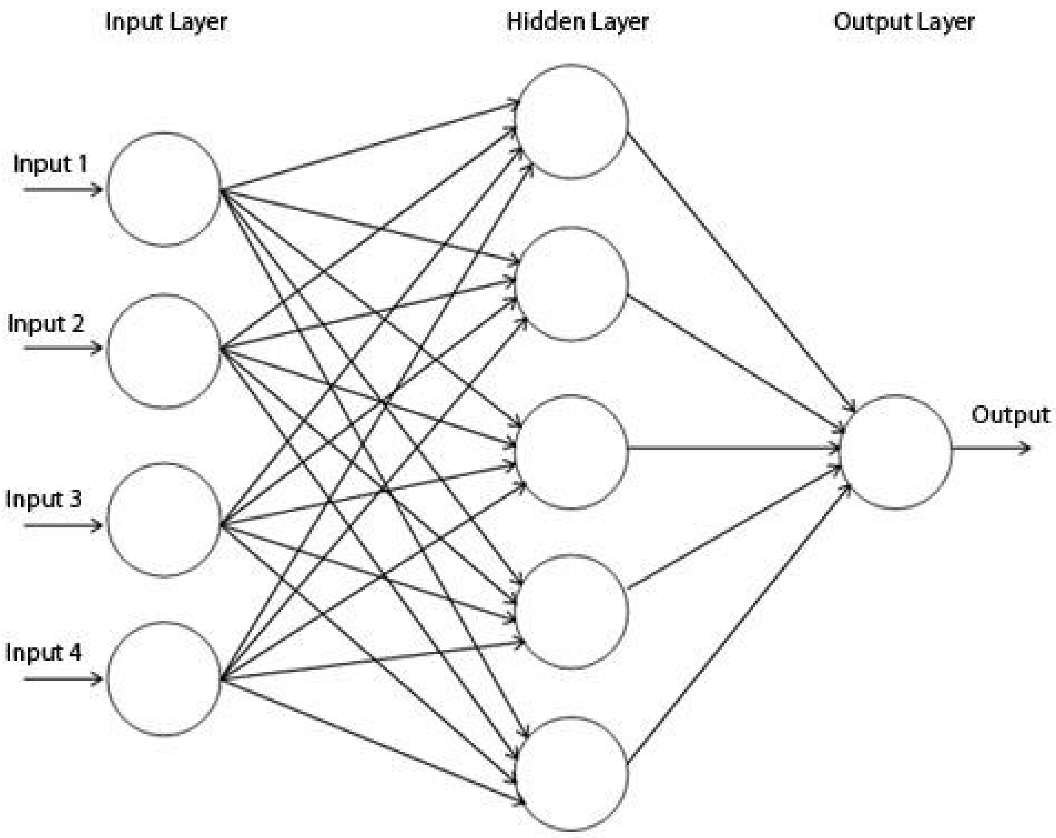
$$\hat{y}_i^{(0)} = 0$$
$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)$$
$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)$$
$$\dots$$
$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

$$\text{obj}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i)$$
$$= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant$$

Mean Absolute Error: **1169**

Fine-tune + Cross-validation:
Score: **1106 (Our final result)**

Deep Learning: **Multi-layer Perceptron**



Different Structures:

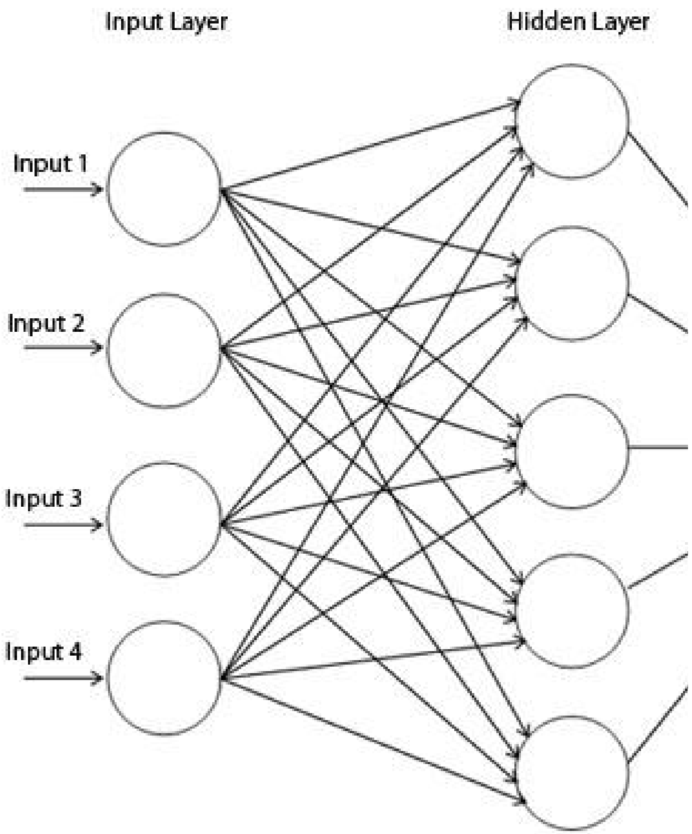
1024 - 2048 - 1
Relu Relu linear

1024 - 4096 - 1
Relu Relu linear

1024 - 4096 - 128 - 1
Relu Relu linear

Best achieve: **1168**

New Approach: **Neural Networks + XG-Boosting**



Used structure:

1024 - 2048 - 1
Relu Relu linear

To have 2048 outputs as the inputs for XG-Boosting

Best achieve: **1143**

Results

Algorithms	Resulting MAE
Linear Regression	1278 (Baseline)
LASSO	1262
Ridge Regression	1267
Elastic Net Regression	1260
Linear Regression (l_1 loss)	1239
CART	1741
Random Forest	1228 (Benchmark)
XG-Boosting (default)	1169
Multi-layer Perceptron	1168
NN+XG-Boosting	1143
Fine-tuned XG-Boosting	**1106** Ranking 533/3055 (18%)

Conclusion

1. It's important to choose a proper objective function. (Replace l_2 loss by l_1 loss in this case)
2. Tree-based methods performed better in this case, the data is not appropriate for the linear approaches.
3. Neural Network did give improvement on XG-Boosting, but not very much. Maybe deep learning structure is not very suitable for this case.

Future Work

- Multicollinearity
- Continuous value attribute processing

Thanks for your attention!

Questions?