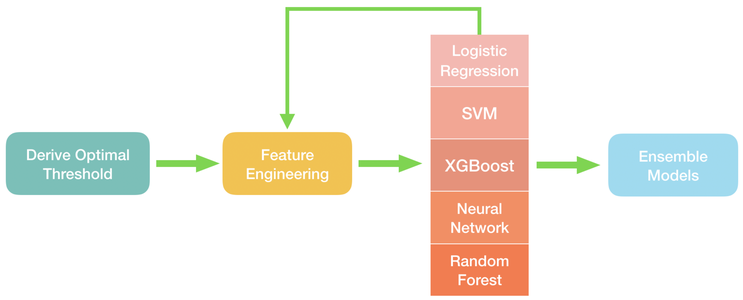
DATA MINING CUP 2019 — ISU team 2

Procedure



Optimal Threshold

The optimal decision rule for unbalanced loss function will be different from the classic Optimal $0-1$ Loss Classifier. Specifically, Let y and $\hat{y}$ be the truth and prediction in the 2-class classification problem and $\boldsymbol{x}$ be the vector of features. Denote $L(y,\hat{y})$ as the loss function in this contest where $L(0,1)=25$, $L(1,0)=5$, $L(0,0)=0$, $L(1,1)=-5$. we predict $\hat{y}=1$ when:

$$\mathbb{E}[L(y,1)|\boldsymbol{x}] < \mathbb{E}[L(y,0)|\boldsymbol{x}],$$

Solve the inequality, we have:

$$p(1|\boldsymbol{x}) > \frac{L(0,1)-L(0,0)}{L(1,0)-L(0,0)-L(1,1)+L(0,1)} = \frac{5}{7}$$

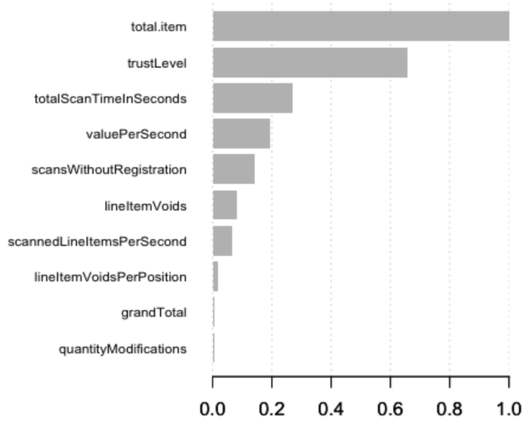
Therefore, **fraud is detected when** $\hat{p}(1|\boldsymbol{x}) > 5/7$.

Feature Engineering

Add important feature:

$$\text{TotalItem} = \text{scannedLineItemsPerSecond} \times \text{totalScanTimeInSeconds}\,.$$

Feature Relative Importance Plots from XGBoost:



Add another important feature by further investigation:

$$\begin{aligned}&\text{Inter}(\text{totalScanTimeInSeconds}, \text{valuePerSecond})\\ =&\text{totalScanTimeInSeconds} \times (1 + \text{valuePerSecond} + \text{valuePerSecond}^{\eta})\end{aligned}$$

Ensemble Logistic Models

Since the logistic regression is way better than other models, ensemble will be done within the logistic regression model. Let $f(\mathbf{x}) = p(1|\mathbf{x})$,

$$\text{logit}(f(\boldsymbol{x}_i)) = \beta_0 + \boldsymbol{x}_i^\top \beta,\ \ i = 1,\ldots,n.$$

We select three logistic models that have:

- good k-fold and LOOCV performance;
- robustness in repeated k-fold cross validation.

Feature set 1:

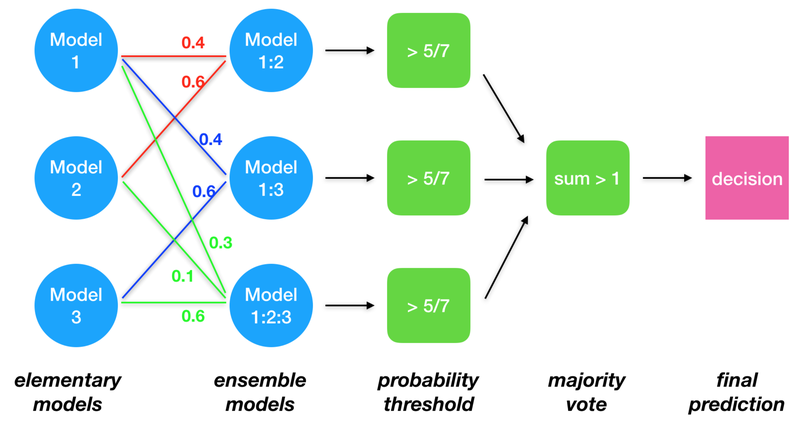
$$\begin{aligned}\text{fraud} \sim\ &\text{trustLevel} + \text{TotalItem} + \text{lineItemVoids}\\ &+ \text{scansWithoutRegistration} + \text{totalScanTimeInSeconds}\end{aligned}$$

Feature set 2:

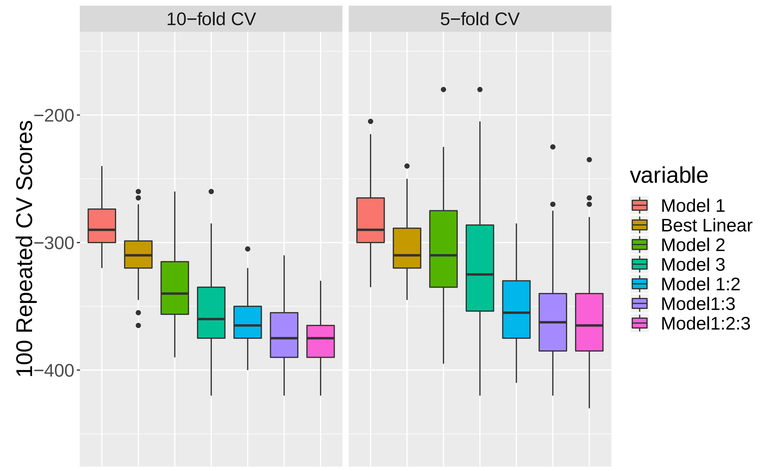
$$\begin{aligned}\text{fraud} \sim\ &\text{Feature set 1} + \text{totalScanTimeInSeconds} \times \text{valuePerSecond}\\ &+ \text{totalScanTimeInSeconds} \times \text{valuePerSecond}^2\end{aligned}$$

Feature set 3:

$$\begin{aligned}\text{fraud} \sim\ &\text{Feature set 1} + \text{totalScanTimeInSeconds} \times \text{valuePerSecond}\\ &+ \text{totalScanTimeInSeconds} \times \text{valuePerSecond}^{3.5}\end{aligned}$$



Logistic Ensemble Results



Compare Models

	RF	SVM	NN	XGBoost	Logistic	Logistic ensemble	Oracle
5-fold CV	-27.5	-203.8	-171.9	-272.2	-304.7	-359.8	-520.0
10-fold CV	-40.0	-175.2	-172.3	-296.0	-306.5	-376.8	-520.0

Table 1. The mean of 100 repeated CV scores

DATA MINING CUP 2019
International Student Competition

University	Iowa State University
Countrv	USA
Team Leader	Qihao Zhang
Team Member	Xingche Guo, Gang Han, Haoyan Hu, Qinglong Tian, Shaodong Wang, Yueying Wang, Haihan Yu, Lijin Zhang, Zerui Zhang, Wenting Zhao, Yifan Zhu