Fast or furious? - User Analysis of SF Express Inc.

Gege Wen, Yiyuan Zhang, Kezhen Zhao
Environmental Fluid Mechanics and Hydrology, Civil and Environmental Engineering, Stanford University

Motivation

- The motivation of this project is to predict the probability for a user to **file complaint for a certain delivery**. With the prediction, we can also determine the top triggers that make a user furious and improve the customer experience during a delivery.
- The S.F. express company is the second largest carrier in China. This project is proposed based on the company's current business needs.

Data

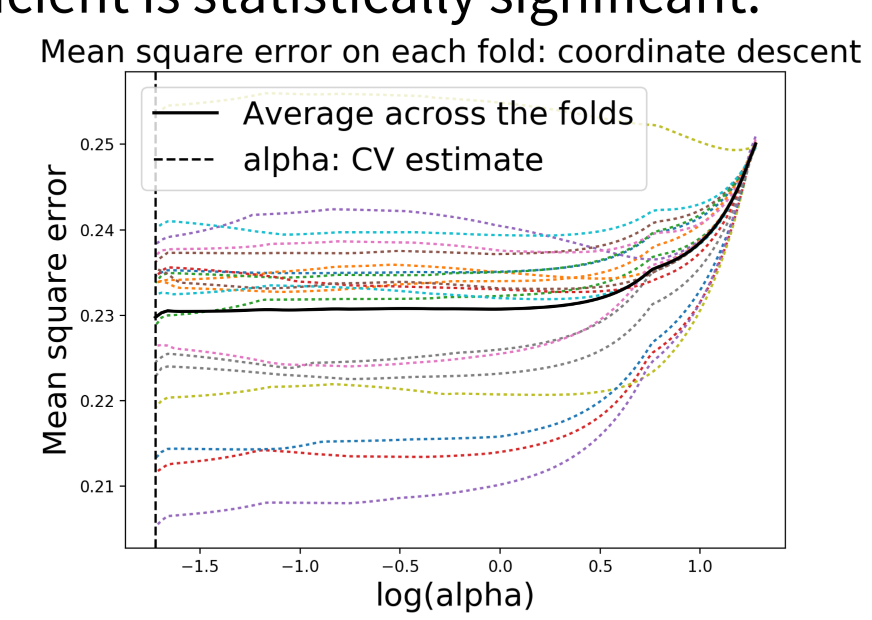
- The dataset comes from S.F's **real delivery history**, represented in CSV format with 40 columns as attributes and 22000 rows as input entries. After cleaning null and irrelevant attributes, 27 attributes are selected as potential features.
- **Data Pre-Processing**: the selected potential features exist in numerical, binary, and categorical formats. We normalized the features into the form of feature matrix as follows. Note that to avoid **Dummy Variable Trap**, a categorical variable with *n* categories is projected to *n-1* columns. The resulting feature matrix has 87 columns.

$$\begin{cases} x_1 \rightarrow numerical & \in \mathbb{R} \\ x_2 \rightarrow categorical & \in \{A, B, C\} \\ x_3 \rightarrow binary & \in 0, 1 \end{cases} \qquad \begin{bmatrix} x_1 \\ x_2 B \\ x_2 C \\ x_3 \end{bmatrix} \rightarrow \begin{bmatrix} 5 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

- **Class Unbalance Issue:** the dataset is highly unbalanced with 2000 complaint entries and 20000 non-complaint entries. To solve this issue, we created three balanced datasets each containing all 2000 complaint entries and 2000 randomly selected non-complaint entries. The following feature and model selection are performed on all three balanced datasets.

Feature Selection

- **Z test:** we assumed that a feature is considered relevant to the predication when we reject the hypothesis that a coefficient equals to zero. When $Z>2$, we have 95% confidence to reject null hypothesis and claim the coefficient is statistically significant.
- **Lasso**: we performed a cross-validation to estimate the expected generalization error for each λ. The value of λ was chosen based on CVMSE to be 0.02. An example on set 1 is presented on the right.



Feature Selection (cond')

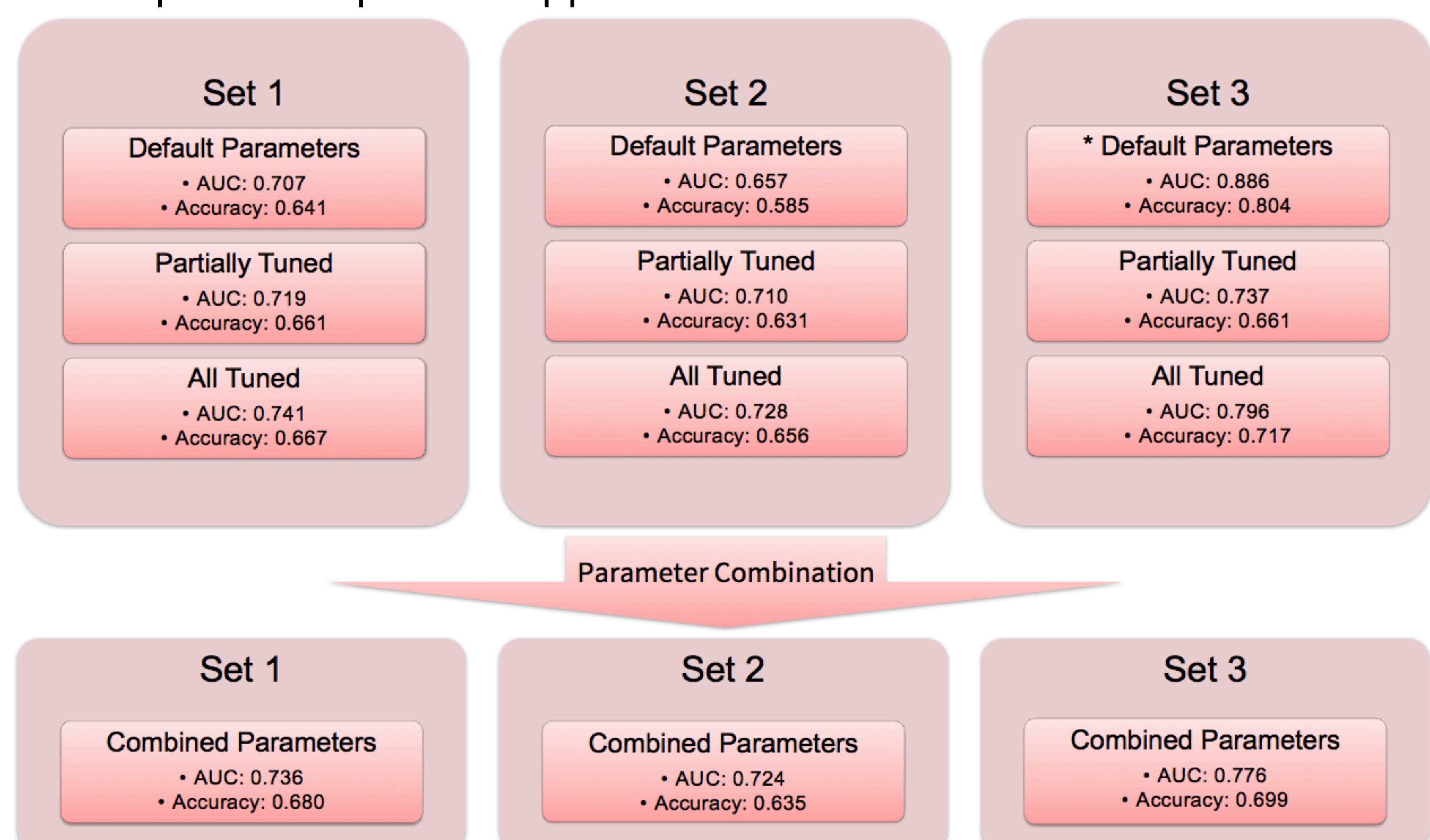
- **Recursive :** we eliminated least important features using the recursive method and compared the remaining feature in each dataset.
- **Stability :** using randomized-lasso as selection algorithm, this method was performed, we eliminated features with score lower than 0.85.

dest_hq_code_3	consign_last6mon_cnt	cons_type1_2	send_la_6mon_comp_waynm
all_fee_rmb	dest_type_2	cons_type1_5	industry_type_level1_8
freight_rmb	dest_hq_code_3	waybill_type_dest_3	receiv_la_6mon_comp_waynm
cons_value	dest_hq_code_8	custtype_3	consigin_emp_hire_mon

As a result of our feature selection, 40 features were eliminated and 16 above features were considered significant.

Model Selection

- Logistic regression with the selected features and balanced data set achieved average AUC score of 0.64, which is selected as our **baseline model**. Logistic regression with Lasso reach an average AUC of 0.65.
- To achieve better performance, we attempted Random Forest (RF), Gradient Boosting Decision Tree (GBDT) and XGBoost. We performed a **Parameter Selection Process** on all of the three methods and a example of this process applied on RF is shown as fellows.



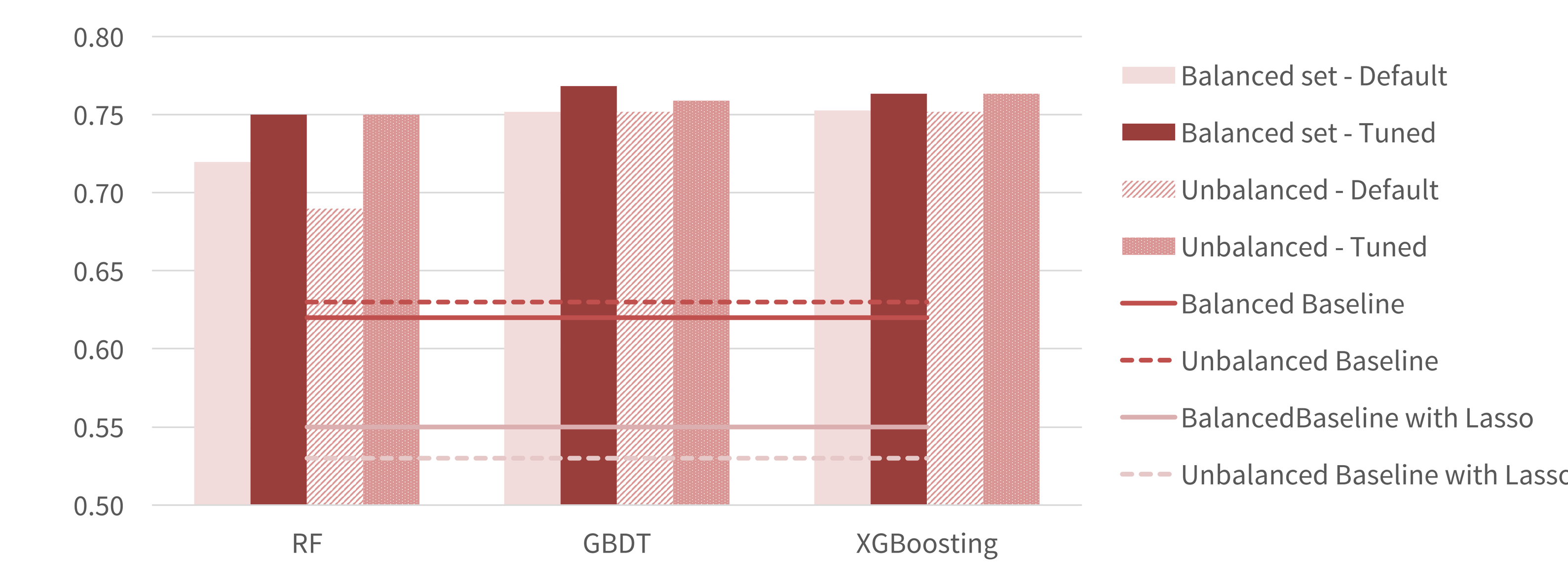
- Within each process, a **Parameter Optimization** technique is applied where we defined parameters to be either **Fixed** or **Inter-correlated**.
- For **Fixed** parameters (e.g. max feature in RF), we performed AUC optimization adjusting this parameter until reaching maximum AUC.
- **Inter-correlate** parameters (e.g. depth, min sample split, and min sample leaf RF) are adjusted in a recursive manner where a portion of this parameters group is adjusted while the rest are fixed. We repeat this adjustment until maximum AUC is reached.

Results

- Optimized Parameters

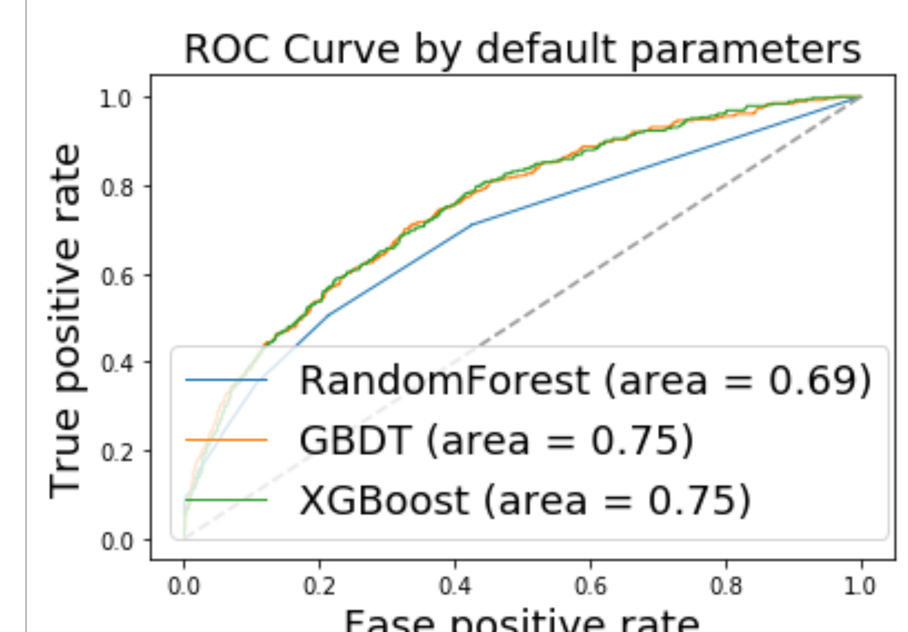
RF		
n_estimators= 110	max_depth=17	min_samples_split=50
min_samples_leaf=1	max_features=10	
GBDT		
n_estimators= 95	max_depth=11	min_samples_split=52
min_samples_leaf=6	max_features=6	
XGBoosting		
n_estimators=700	learning_rate = 0.07	gamma=0.2
subsample=0.8	colsample_bytree=0.8	reg_lambda=1e-5
reg_alpha=1e-5	max_depth=3	min_child_weight=1

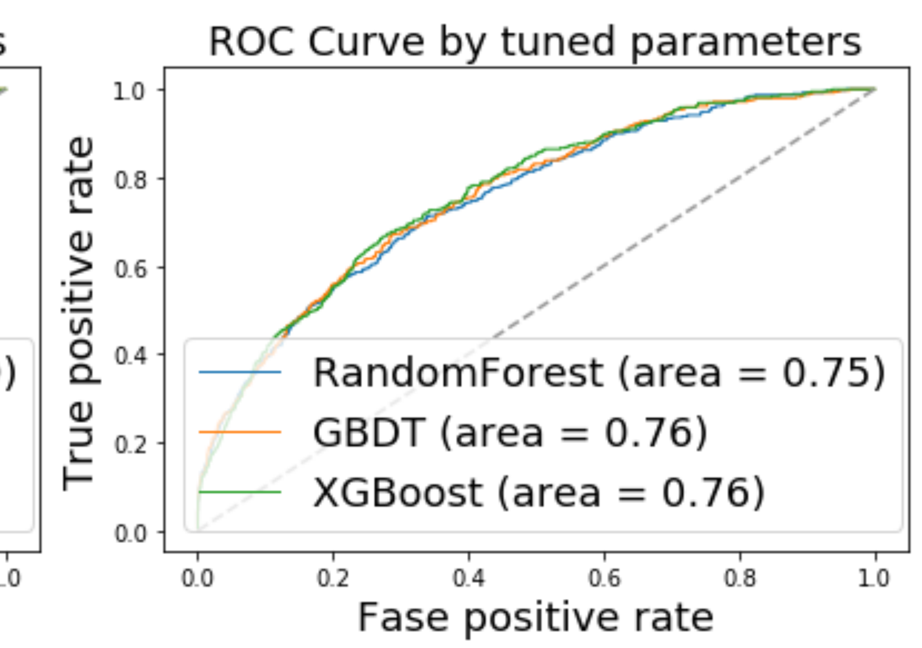
- Resulting AUC score



Discussion

- **Linear vs. Tree**: in this real life problem, most features do not have a linear property. Other linear models including stepback were also attempted. However, the improvement from feature selection in the linear models are not significant and a tree structured model performs better at capturing the non-linearity.
- **ROC Curves**: when using unbalanced dataset, GBDT and XGBoost has better performance of true positive rate, therefore better chance to distinguish complaint cases.





Future Work

- For future work, we are interested in improving the class bias problem. Since real-world complaint data only occupies a little portion in the whole dataset. Solving this problem can reduce the false negative classifications and meet real companys' needs.