Is this transaction fraudulent?

Jessy Sun, Nada Liang, Duncan Wingfield, Travis Li

Machine Learning Final Project

Agenda

Business Problem

Dataset

Data Exploration

Models

Classification Tree
Logistic Regression (Lasso)
Random Forest
Stacking (SVM, Log)

Business Interpretation & Application

Summary

Part 1

Business Scenario

Business Problem

Who are we?

E-Commerce company

What's our Goal?

To **identify fraudulent customers** at their first transaction.

What type of question is it?

Classification: 0 (not fraudulent) or 1 (fraudulent)

Business Interpretation

		Predicted Result: Not fraudulent (0)	Predicted Result: Fraudulent (1)
Actual	Not fraudulent (0)	**TN**	**Type error I (FP):** The transaction is not fraudulent, but we predicted it was.
	Fraudulent (1)	**Type error II (FN):** The transaction is fraudulent, but we predicted not.	**TP**

Part 2

Dataset & Feature Engineering

Dataset: Each user first transaction

How large is it?

151,112 records / transactions

11 predictors

Response:

“Class”

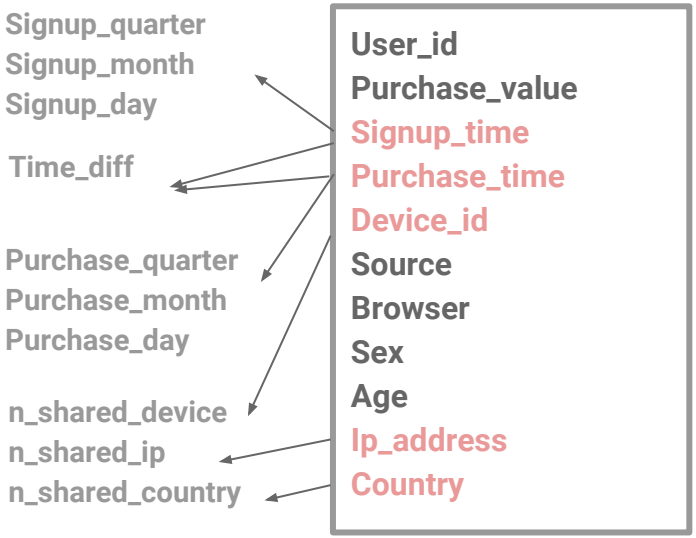
- **0 (not fraudulent)**
- **1 (fraudulent)**

User_id [numeric]	Unique User ID
Signup_time [date,time]	Time when user created account
Purchase_time [date,time]	Time when user bought an item
Purchase_value [numeric]	Cost of the item purchased
Device_id [string]	Unique physical device identification
Source [string]	How the user found and clicked on the site
Browser [string]	The internet browser the user used
Sex [string]	User sex (male/female)
Age [numeric]	User age
Ip_address [numeric]	User numeric IP address
Ip_address_to_country [string]	Shows what Country the IP address is in upon purchase

(including dummy variables)

Feature engineering: 11p → 15p → 48col

1. Delete user_ID
2. Divide time
3. Calculate Time difference
4. Transform ip, device id, country info
5. Transform string into categorical for tree based model
6. Transform categorical data to dummy variables for logistic regression
7. Scale data for logistic regression



Part 3

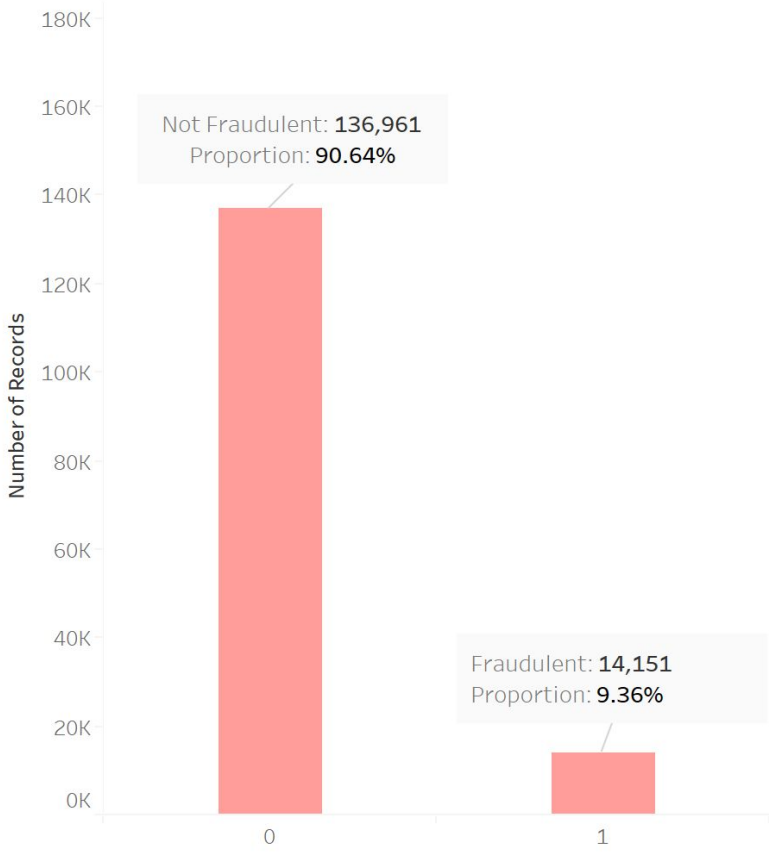
Data Exploration

Naive model

(If we predict that every transaction is not fraudulent)

Accuracy: 90.64%

Distribution of Response Values



First Glimpse of Target by Age & Gender

Part 4

Model & Results

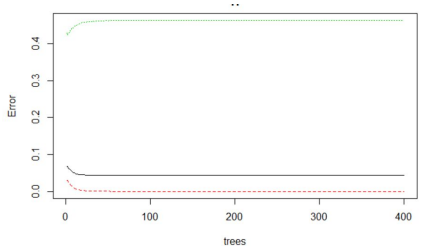
Metric Selection

1. Since the dataset is imbalanced, we are **not** going to use **ACCURACY** as the **sole metric**
2. **AUC and ROC** curve will be leveraged as a trade-off metric

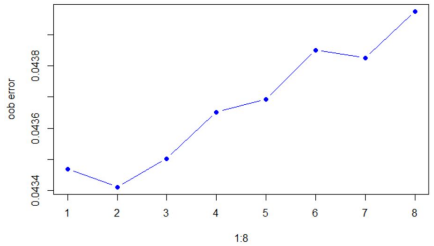
Model 1: Random Forest

Type I error	0.00%
Type II error	45.97%
Accuracy	95.68%
AUC	0.848

ntree: 97

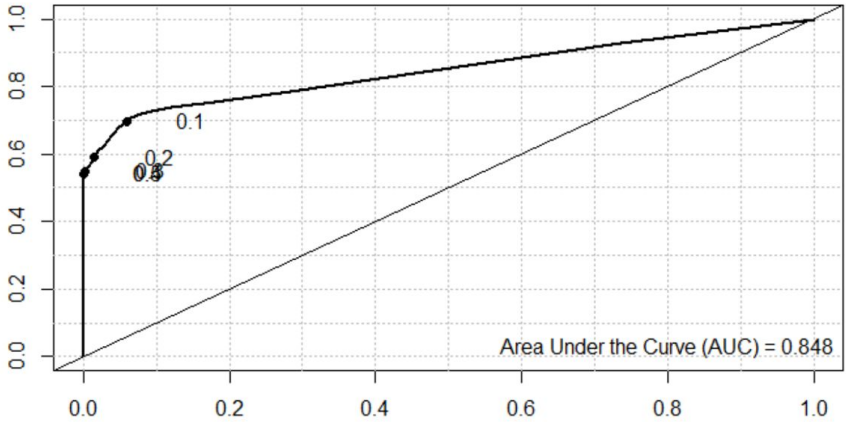


mtry: 2

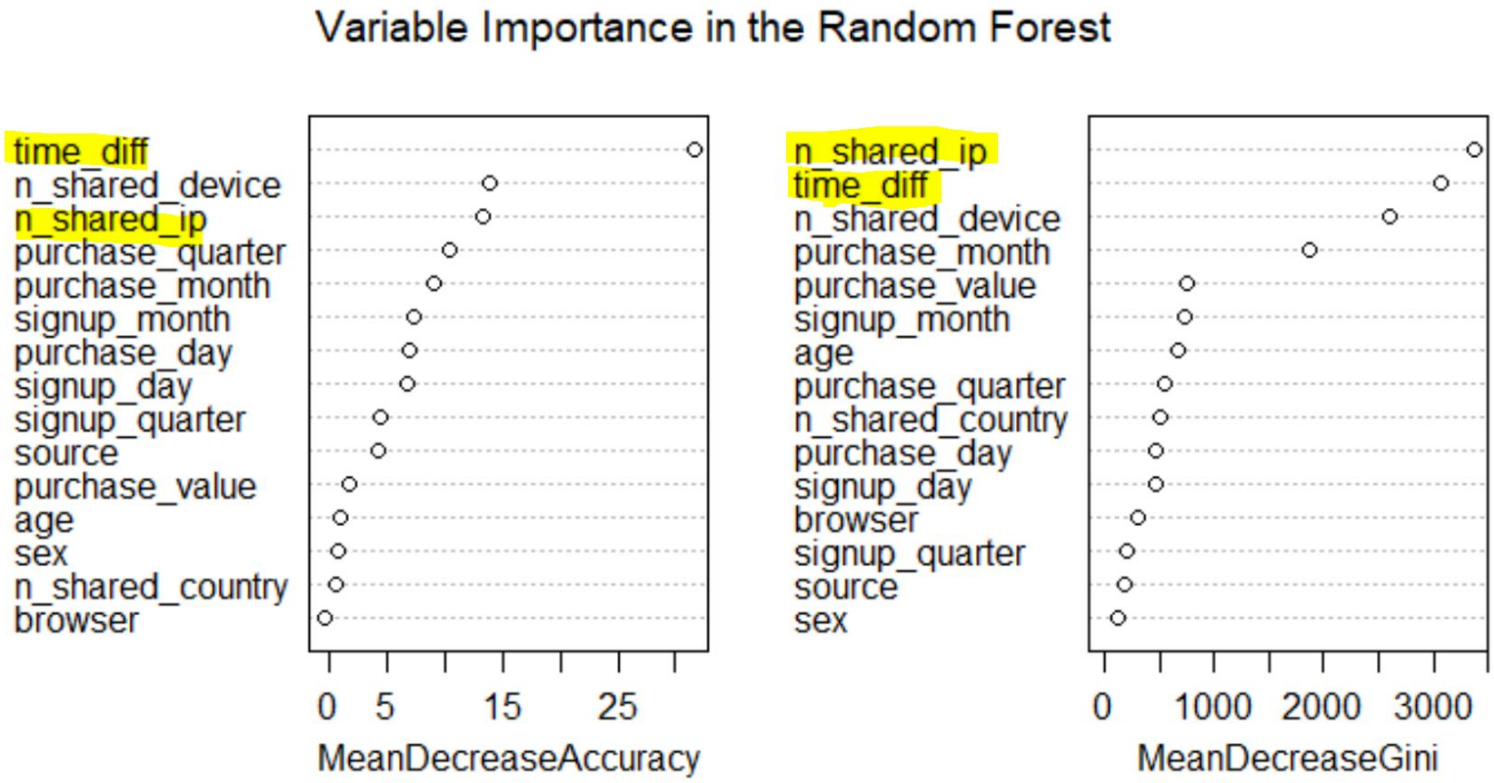


		Prediction	
		Not fraud	Fraud
Actual	Not fraud	27380	0
	Fraud	1307	1536

ROC Curve for Random Forest



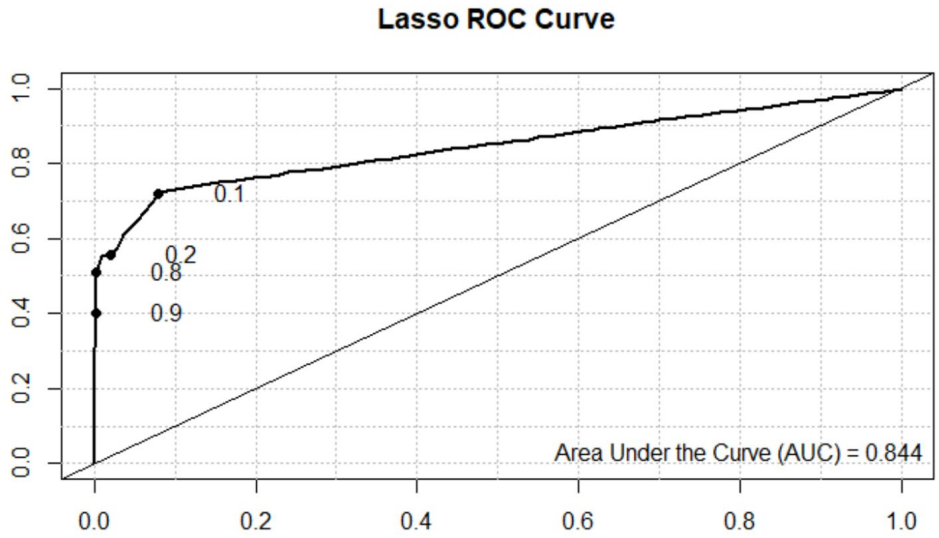
Model 1: Random Forest



Model 2: Logistic Regression (Lasso)

Type I error	0.46%
Type II error	45.48%
Accuracy	95.30%
AUC	0.844

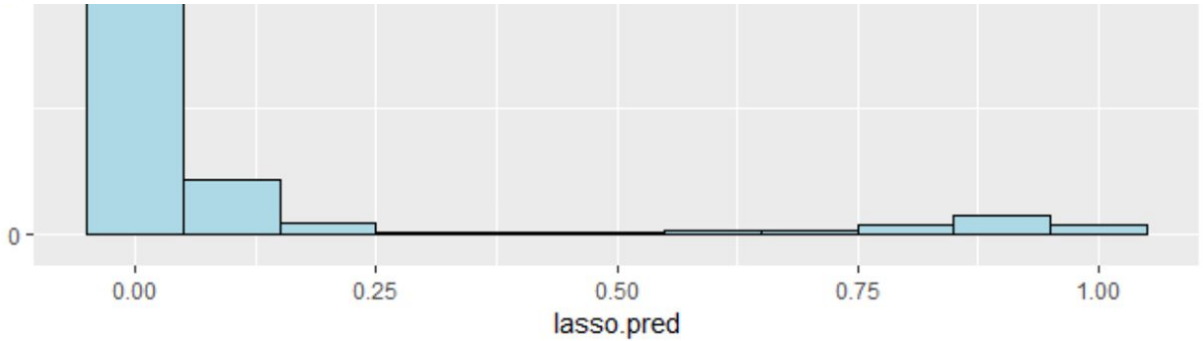
		Prediction	
		Not fraud	Fraud
Actual	Not fraud	27253	127
	Fraud	1293	1550



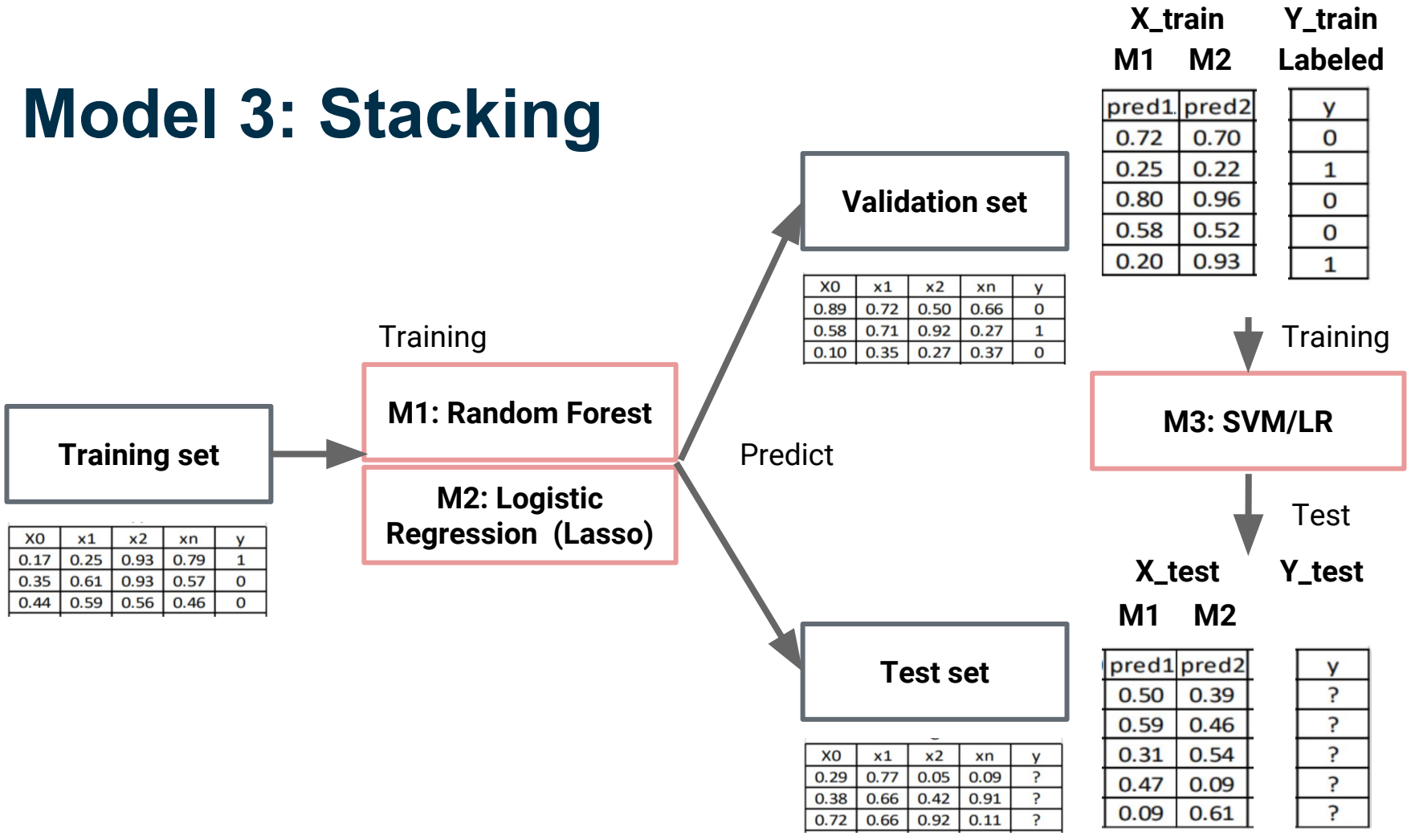
Model 2: Logistic Regression + New Loss

$$\mathrm{CEL}_{fl} = \begin{cases} -\alpha(1-y')^{\gamma} log y' & , \quad y = 1 \\ -(1-\alpha)y'^{\gamma}\log(1-y'), & \quad y = 0 \end{cases}$$

se.



Model 3: Stacking

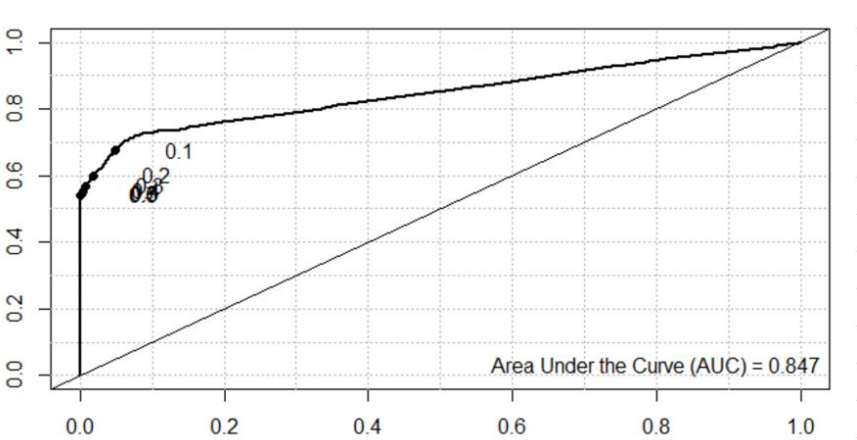


Model 3: Stacking

M3: logistic regression

Type I error	0.18%
Type II error	45.55%
Accuracy	95.55%
AUC	0.847

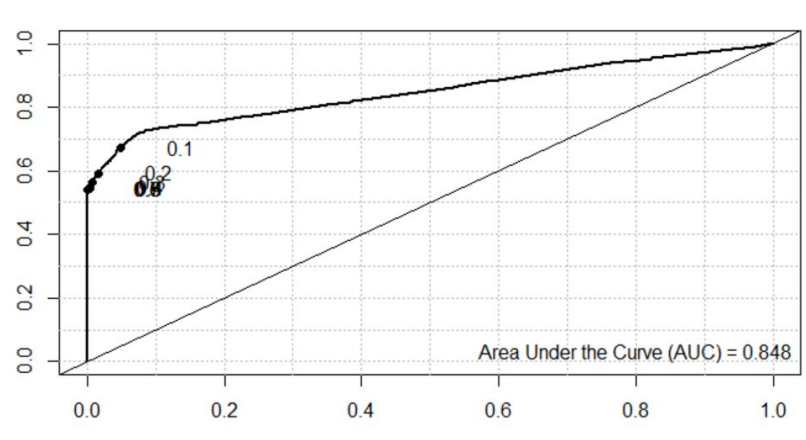
		Prediction	
		Not fraud	Fraud
Actual	Not fraud	27330	50
	Fraud	1295	1548



M3: SVM

Type I error	0.00%
Type II error	45.97%
Accuracy	95.68%
AUC	0.848

		Prediction	
		Not fraud	Fraud
Actual	Not fraud	27380	0
	Fraud	1307	1536



Summary of Models

Random Forest is the best model.

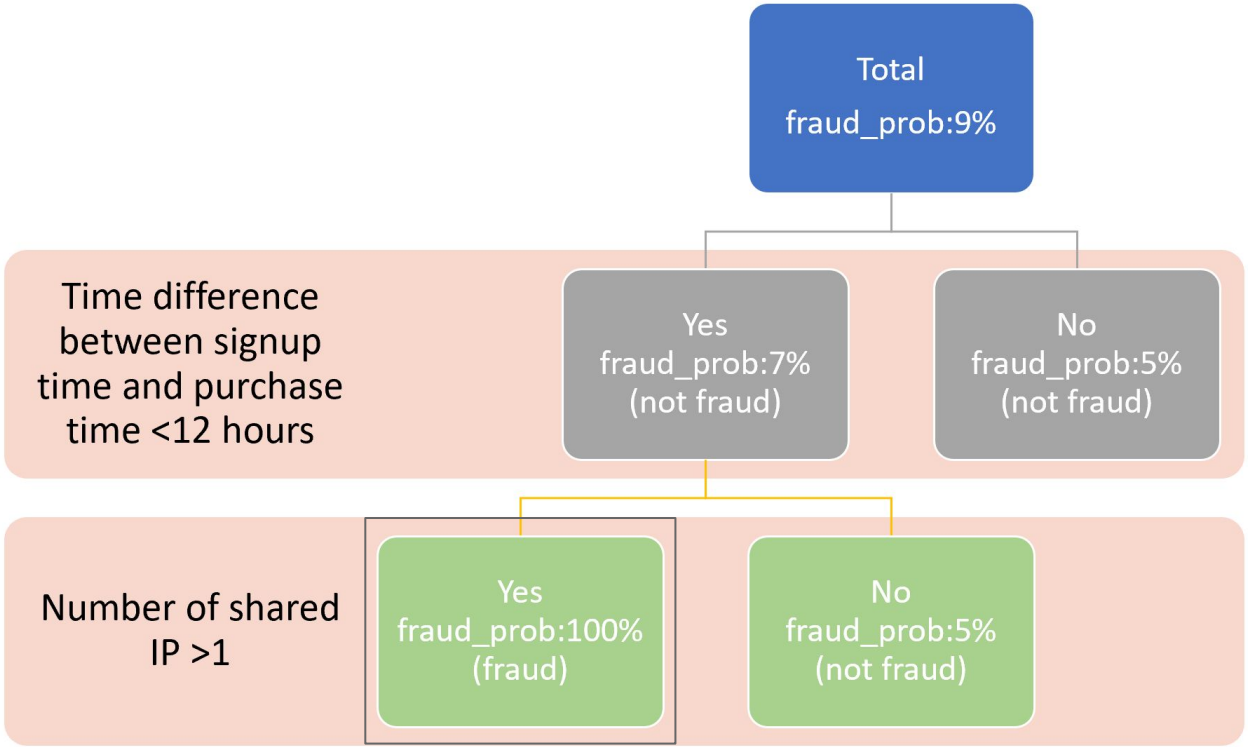
(overall performance is great, more simple than stacking)

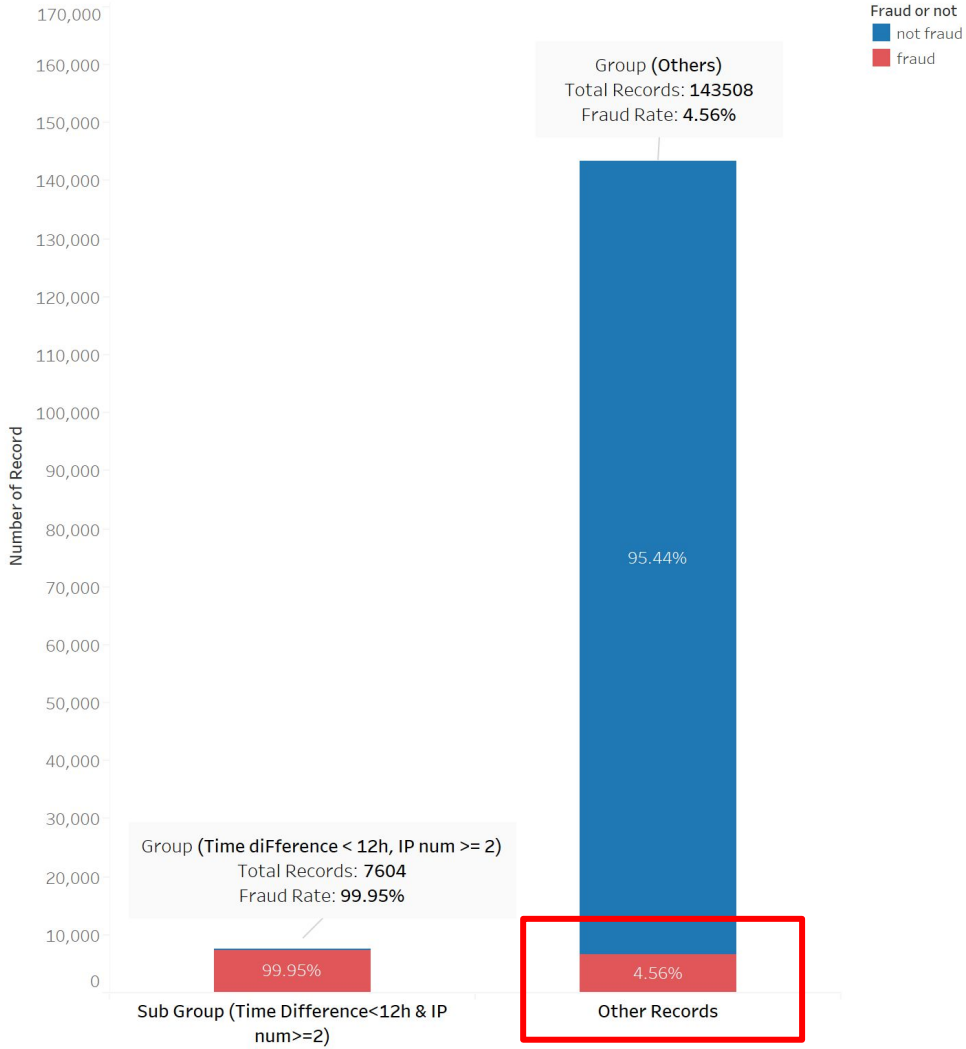
	Classification Tree	Logistic Regression	Random Forest	Stacking-SVM	Stacking-Logistic Regression
Type I error	0.00%	0.46%	0.00%	0.00%	0.18%
Type II error	45.94%	45.48%	45.97%	45.97%	45.55%
Accuracy	95.68%	95.30%	95.68%	95.68%	95.55%
AUC	0.7704	0.844	0.848	0.848	0.847

Part 5

Business Interpretation & Application

Patterns of Fraudulent Transactions





Transactions that were difficult to identify:

6551/151112

4.3% of total records → acc 95.66%

6551/14151

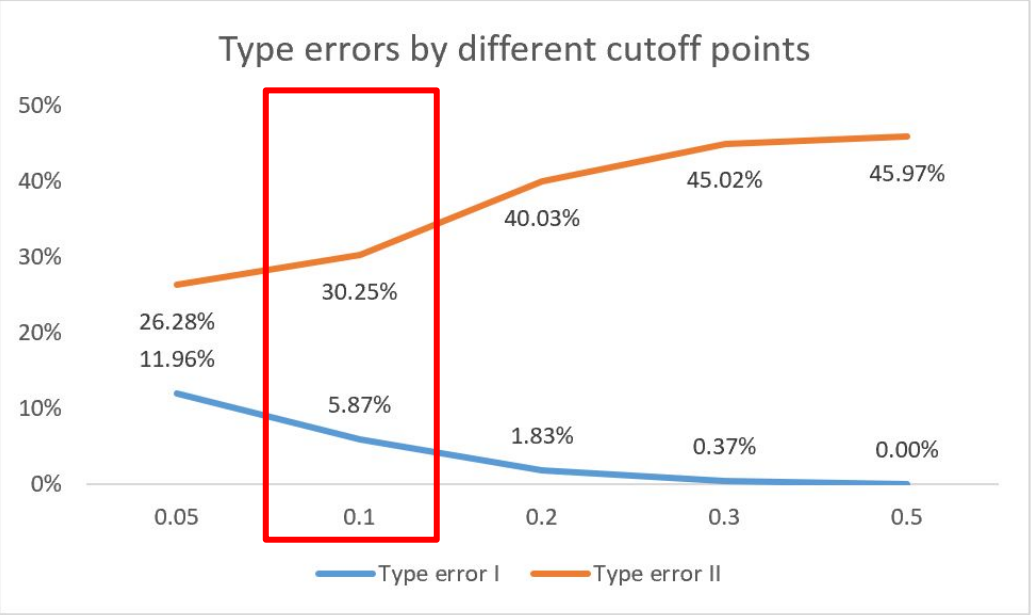
46.3% of fraud records → type error II

Lower Cutoff

To Get a Lower Type II Error

Based on Random Forest:

Cutoff Points	0.05	0.1	0.2	0.3	0.5
Accuracy	86.69%	91.83%	94.58%	95.43%	95.68%
Type error I	11.96%	5.87%	1.83%	0.37%	0.00%
Type error II	26.28%	30.25%	40.03%	45.02%	45.97%



Action Plan

Based on the fraud probability from our model, set **ALERT 1** (0.1) and **ALERT 2** (0.5)

Cutoff Points	0.1	0.5
Accuracy	91.83%	95.68%
Type error I	5.87%	0.00%
Type error II	30.25%	45.97%

If p < ALERT 1:

Normal, no alert

If ALERT 1 < p < ALERT 2:

The purchase is suspicious, we will ask the customer for additional authorization. For example, send email or SMS to the customer, let him/her authorize the purchase.

If p > ALERT 2 :

Then the purchase is highly suspicious. Not only ask the customer for additional authorization via email or SMS, but also put the purchase on hold and send the purchase information to relevant departments for further investigation.

Part 6

Summary

Summary and Recommendations

- In this scenario, **type II error** is more important.
- **Random forest** is our best model
- We lowered the **cutoff points** and deigned two alert levels based on 0.1 and 0.5 cutoff points.
- There is always a fraction of fraud that we can not distinguish (46%). Therefore, **more data** should be collected.

- For those who have had more than one IP address and purchased the products before signing up for 12 hours:

99.95% are **fraudulent**.

- Important variables:
  - **Time difference between purchase time and sign-up time**
  - **Number of shared devices**
  - **Number of shared IPs**
  - **Purchase value**
  - **Purchase time (month/quarter)**

Backup Slides

Stacking

Means making predictions of a number of models in a hold-out set and then using a different (Meta) model to train on these predictions.

Methodology

- Wolpert in 1992 introduced stacking. It involves:

1. **Splitting** the train set into two disjoint sets.
2. **Train** several base learners on the first part.
3. **Make predictions** with the base learners on the second (validation) part.
4. Using the **predictions** from (3) **as the inputs** to train a higher level learner.

Still confused about Stacking?

A

X0	x1	x2	xn	y
0.17	0.25	0.93	0.79	1
0.35	0.61	0.93	0.57	0
0.44	0.59	0.56	0.46	0
0.37	0.43	0.74	0.28	1
0.96	0.07	0.57	0.01	1

B

X0	x1	x2	xn	y
0.89	0.72	0.50	0.66	0
0.58	0.71	0.92	0.27	1
0.10	0.35	0.27	0.37	0
0.47	0.68	0.30	0.98	0
0.39	0.53	0.59	0.18	1

C

X0	x1	x2	xn	y
0.29	0.77	0.05	0.09	?
0.38	0.66	0.42	0.91	?
0.72	0.66	0.92	0.11	?
0.70	0.37	0.91	0.17	?
0.59	0.98	0.93	0.65	?

Train algorithm **0** on A and make predictions for B and C and save to **B1**, **C1**
Train algorithm **1** on A and make predictions for B and C and save to **B1**, **C1**
Train algorithm **2** on A and make predictions for B and C and save to **B1**, **C1**

B1

pred0	pred1	pred2	y
0.24	0.72	0.70	0
0.95	0.25	0.22	1
0.64	0.80	0.96	0
0.89	0.58	0.52	0
0.11	0.20	0.93	1

C1

pred0	pred1	pred2	y	Preds3
0.50	0.50	0.39	?	0.45
0.62	0.59	0.46	?	0.23
0.22	0.31	0.54	?	0.99
0.90	0.47	0.09	?	0.34
0.20	0.09	0.61	?	0.05

Train algorithm **3** on B1 and make predictions for C1

Clustering

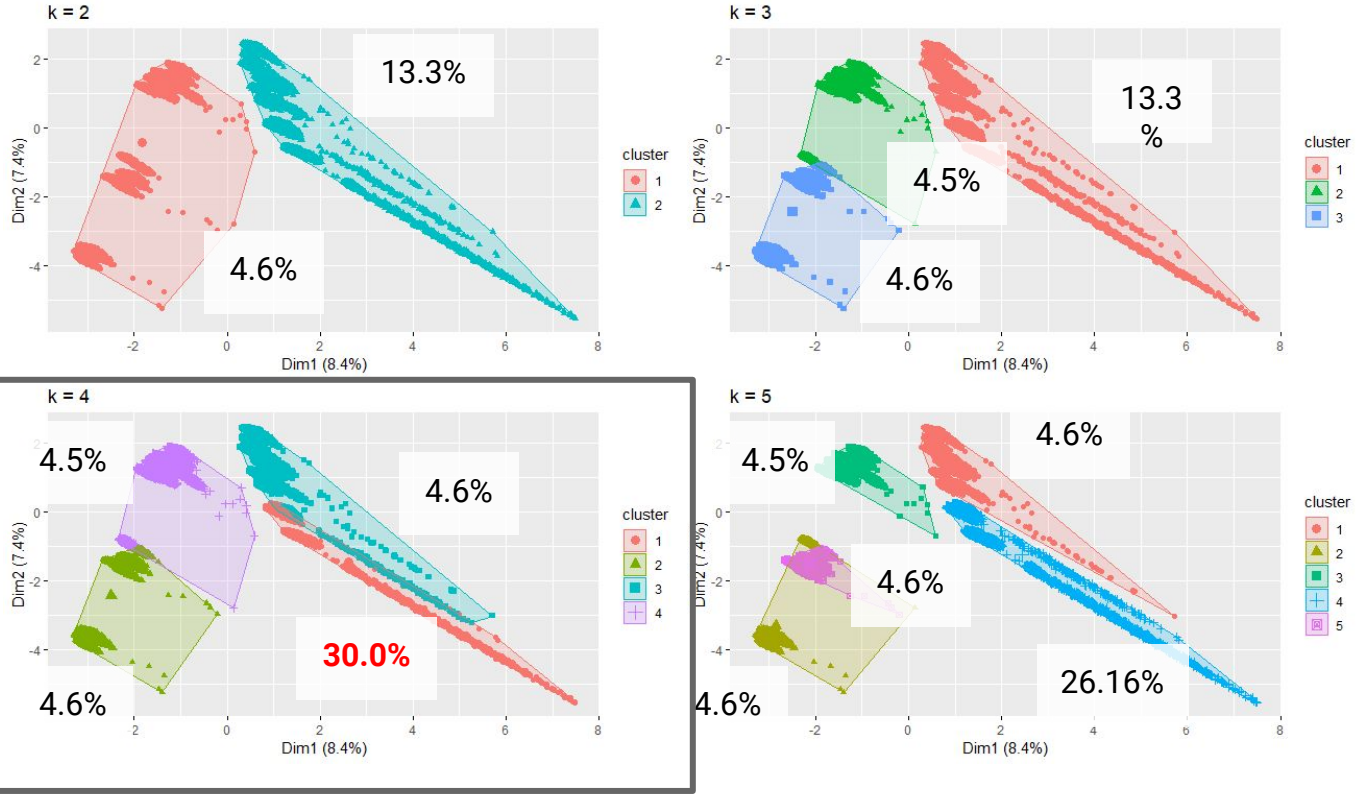
Method

k means (2,3,4,5)

Fraud rate:

Average: 9.36%

(k=4) Cluster1:30%



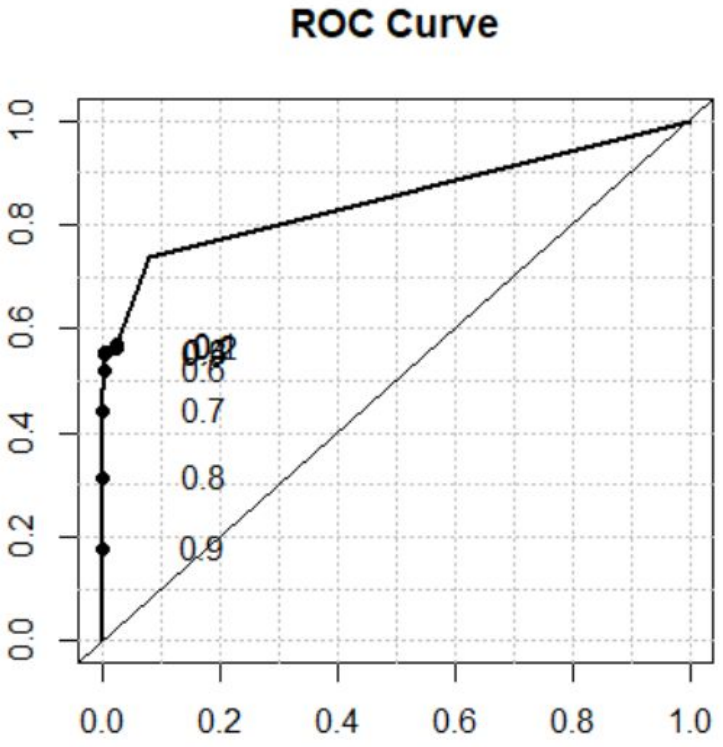
Lasso Regression

"acc for lasso is: 95.44 %"

pred_lasso	testy 0	1
0	27295	1270
1	107	1551

"TypeI: 0.39 % TypeII: 45.02 %"

"roc score is: 0.847"



RandomForest

Summary

Results

```
"acc for random forest is" "95.57%"

       Predicted
Actual     0     1
     0 41023     0
     1  2010  2301

"Type I Error:" "0.00%"
"Type II Error:" "46.62%"

"roc score is 0.849"
```