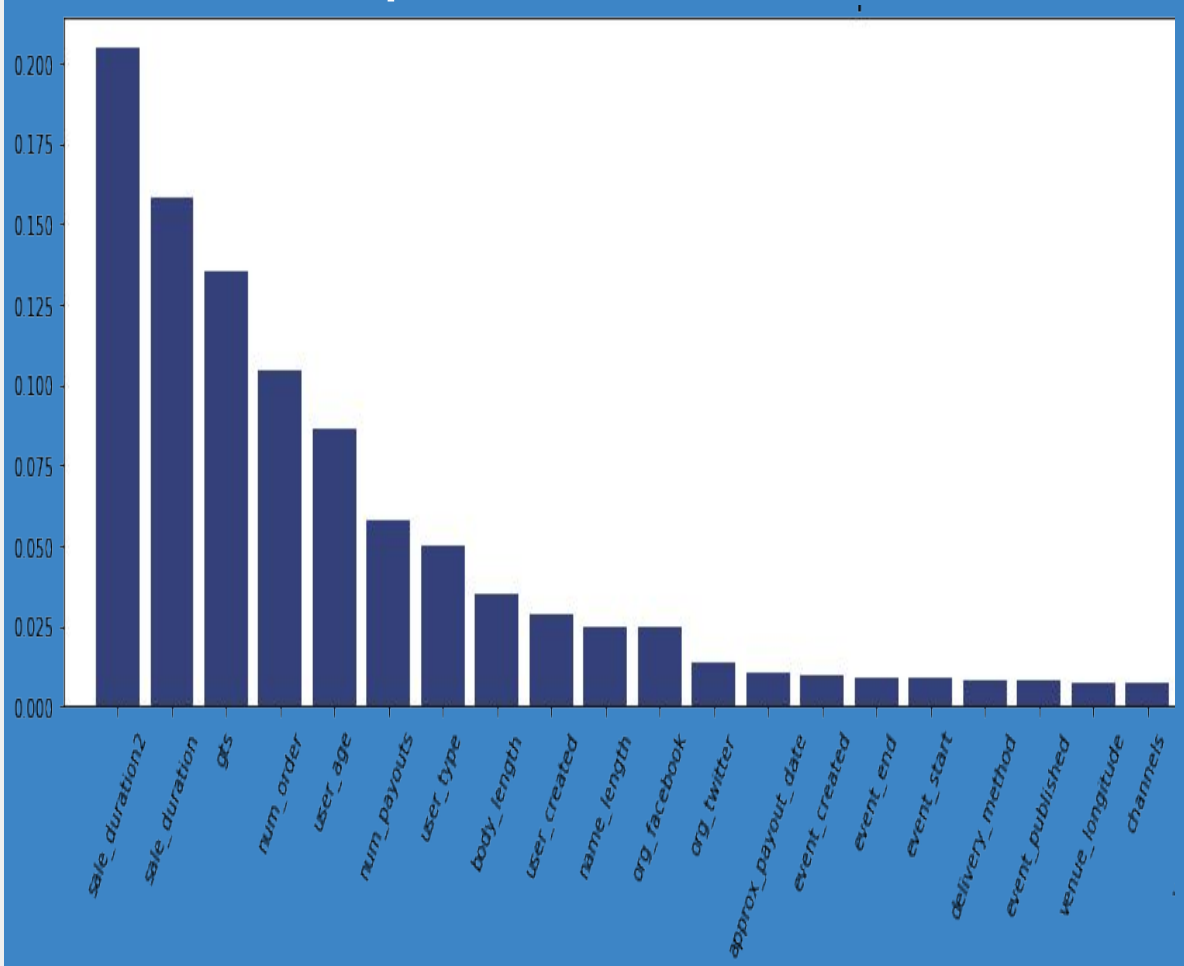
Event Data
Fraud Detection

Katie Slayden, Prantik Ghosh, Jenny Tang, Ziru Fan, Vicky Kwan

Avoiding Leakage

payout_type
sales_duration
sales_duration2

Feature Importances



Baseline

Model Type	Accuracy	Recall	Precision
Logistic Regression	92	18	78
Random Forest	95	59	79

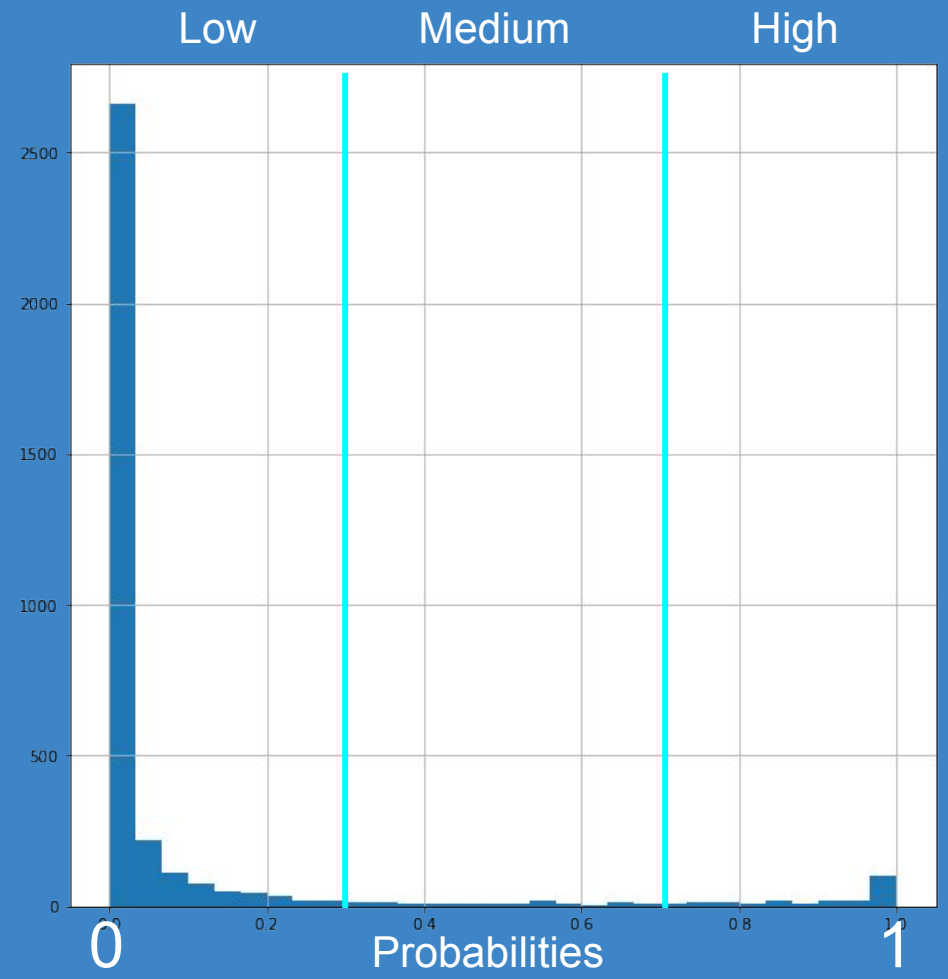
Final Model

Model Type	Accuracy	Recall	Precision
Random Forest	97	75	86

Final Features = ['gts', 'num_order', 'user_age', 'num_payouts', 'user_type', 'body_length', 'user_created', 'name_length', 'org_facebook']

Probability Thresholds

Low: 25 +below
Medium: 25 to 75
High: 75 +above



Countries

- Considered including countries
- Excluded from final model
- Data too skewed

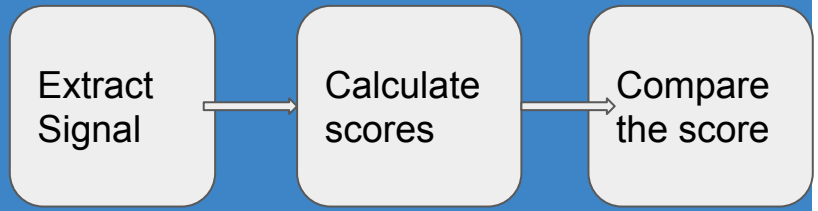
country	NotFraud	Fraud	Fraud_percent
US	8615	619	6.703487
GB	1789	294	14.114258
CA	1209	53	4.199683
AU	727	15	2.021563
	276	40	12.658228
NZ	59	14	19.178082
IE	48	0	0.000000
FR	37	3	7.500000
NL	30	6	16.666667
DE	28	15	34.883721
ES	26	1	3.703704
HU	18	0	0.000000
BE	17	1	5.555556
AT	8	0	0.000000
SG	7	0	0.000000
TH	6	0	0.000000
PR	6	4	40.000000
RO	5	3	37.500000
IT	4	0	0.000000
AR	4	4	50.000000
ZA	3	0	0.000000
IN	3	1	25.000000
SE	3	0	0.000000
GR	2	0	0.000000
MX	2	0	0.000000
ID	2	10	83.333333
HR	2	1	33.333333
UY	2	0	0.000000
RS	2	1	33.333333
FI	2	2	50.000000
...	...	...	...
IS	1	0	0.000000
IM	1	0	0.000000
AE	1	0	0.000000
VE	1	0	0.000000
VI	1	0	0.000000

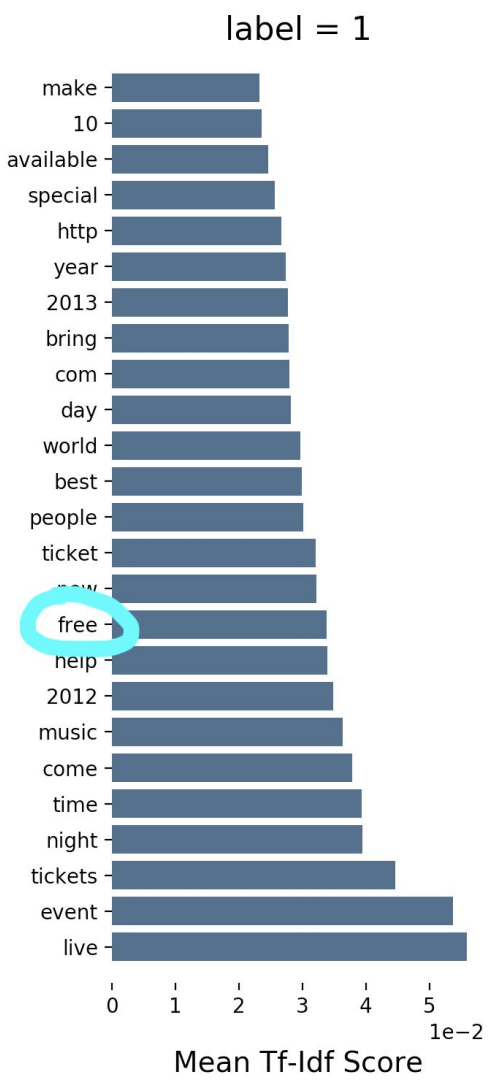
Text Signals

Label = 0: normal events

Label = 1: fraudulent events

Comparing Scoring System:
Fraud: 0 v.s. Not Fraud: 0.0323291





Thank you.