Click Fraud Detection In Online Advertisements

PROBLEM??

- Dishonest publishers might generate clicks on advertisements on their websites using manual and/or automated techniques to increase their revenue.
- Dishonest advertisers might also generate false clicks on their competitor's advertisements to drain their advertising budgets.

Dataset

PartnerID	Bank Account	Address	Status
8iaxj		14vxbyt6sao00s	Fraud
8jljr			Ok

Table 1: Publisher sample in raw training data.

id	iplong	agent	Partner id	cid	cntr	timeat	category	referer
134178	364840	GT-I9	8iaxj	8fj2j	ru	0:00:00	ad	26okyx5
134176	375696	Samsun	8jljr	8geyk	in	0:11:05	es	15vynjr
134178	693232	SonyEri	8jljr	8gkkx	ke	1:21:00	es	
134178	288420	Nokia	8jljr	8gp95	vn	0:47:13	es	
134180	364840	GT-I	8iaxj	8fj2m	ru	5:30:00	ad	24w9x4
134183	78135	Nokia	8iaxj	8fj2j	ru	7:16:53	ad	4im48n

Table 2: Click sample in raw training data.

Feature Extraction

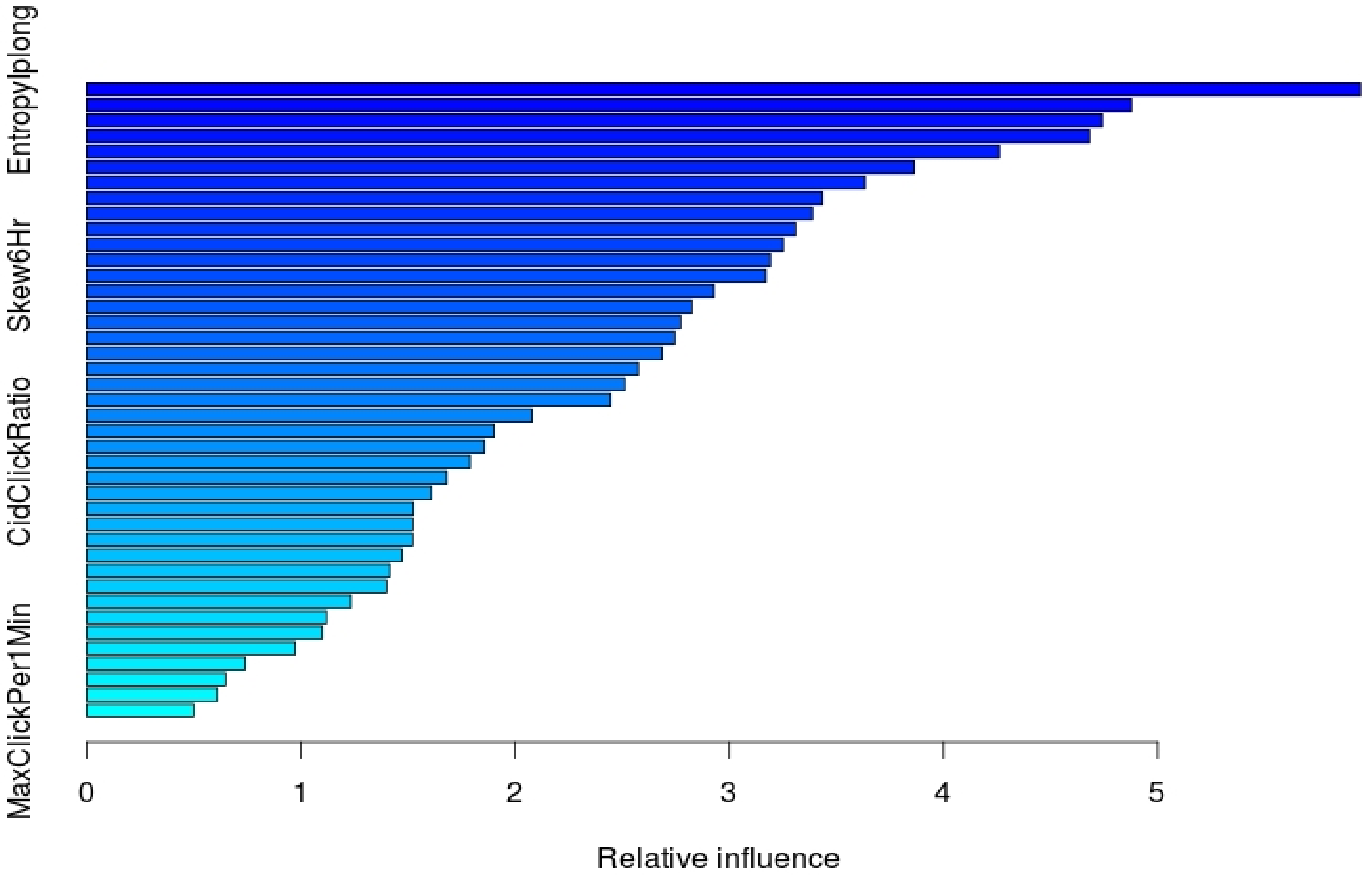
- Attribute: iplong
  - MaxSameIpCount
  - NoOfUniqueIp
  - IpClickRatio
  - Ipvariance
- Attribute: Time-At
  - No. of click per 1 minute
    - Average
    - Max
    - Variance
    - Skewness
  - No. of clicks per 1 hour
  - No. of clicks per 3hour

Top Features

Click Behaviour			High Risk Click Behaviour		
Rank	Feature	Relative inf.	Rank	Feature	Relative inf.
1	EntropyIplong	5.97	2	cntr_id_%	5.49
3	RefClickRatio	4.79	9	cntr_sg_%	3.09
4	IpClickRatio	4.69	10	cntr_othr_%	2.72
5	NoOfUniqueRef	4.56	12	cntr_us_%	1.44
6	MaxSameAgentCnt	4.15	13	cntr_th_%	1.39
7	IpVariance	3.77	14	cntr_uk_%	1.34
8	NoOfUniqueAgent	3.75	15	cntr_in_%	1.26
11	Skew5Min	1.621	19	cntr_ng_%	0.95
16	TotalClicks	1.09	21	cntr_tr_%	0.94
17	AvgClickPer1Min	1.08	32	cntr_ru_%	0.66

Table 3: Top-10 features by type

Relative Influence



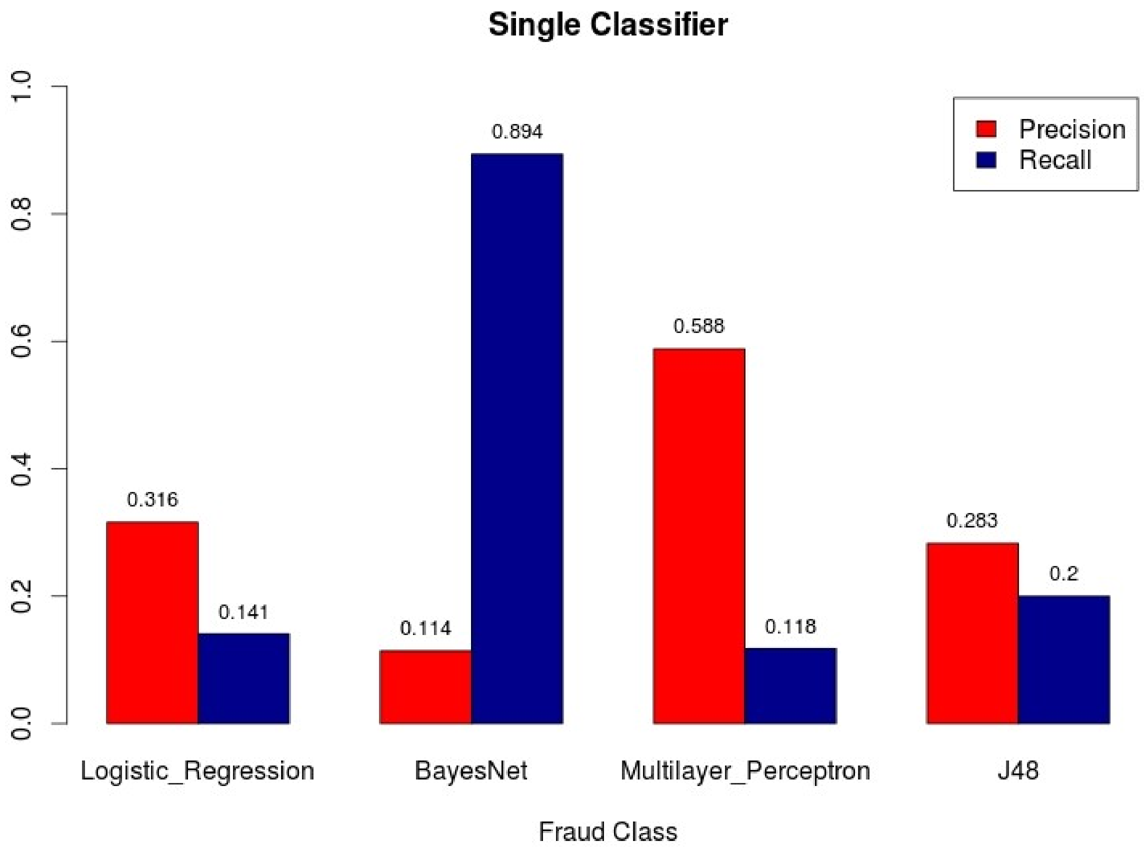
Feature Selection

- Chance of data overfitting.
- Methods
  - Principal Component Analysis (PCA)
  - Common Spatial Patterns (CSP)
  - rapper subset evaluation
- Not yet tested.

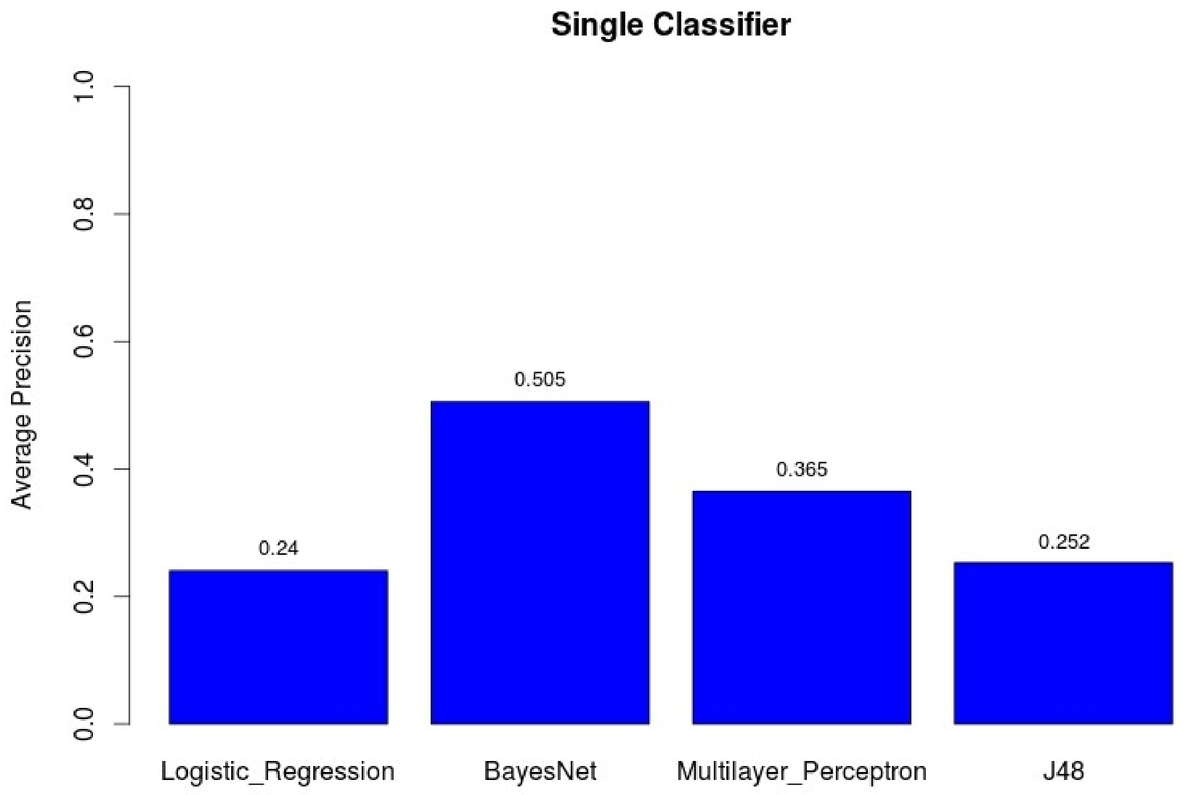
Single Algorithms

	Precision		Recall		F-Measure		ROC	
Algorithm	Ok	Fraud	Ok	Fraud	Ok	Fraud	Ok	Fraud
Logistic Regression	0.976	0.316	0.991	0.141	0.984	0.195	0.936	0.906
Bayesian Net	0.996	0.114	0.803	0.894	0.889	0.203	0.871	0.867
MLP	0.975	0.588	0.998	0.118	0.986	0.196	0.825	0.825
J48	0.977	0.283	0.986	0.200	0.981	0.234	0.589	0.589

Precision-Recall



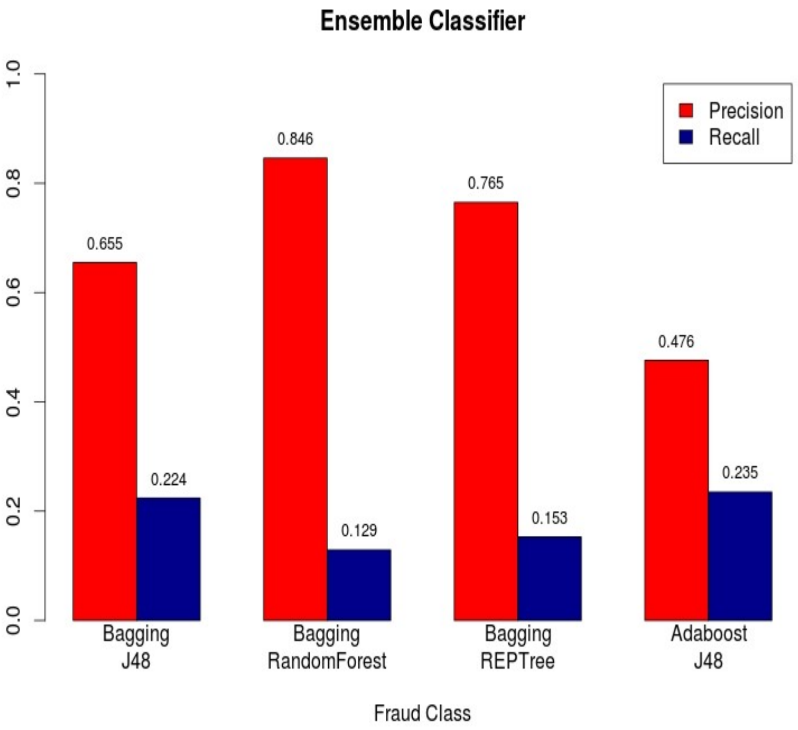
Average Precision

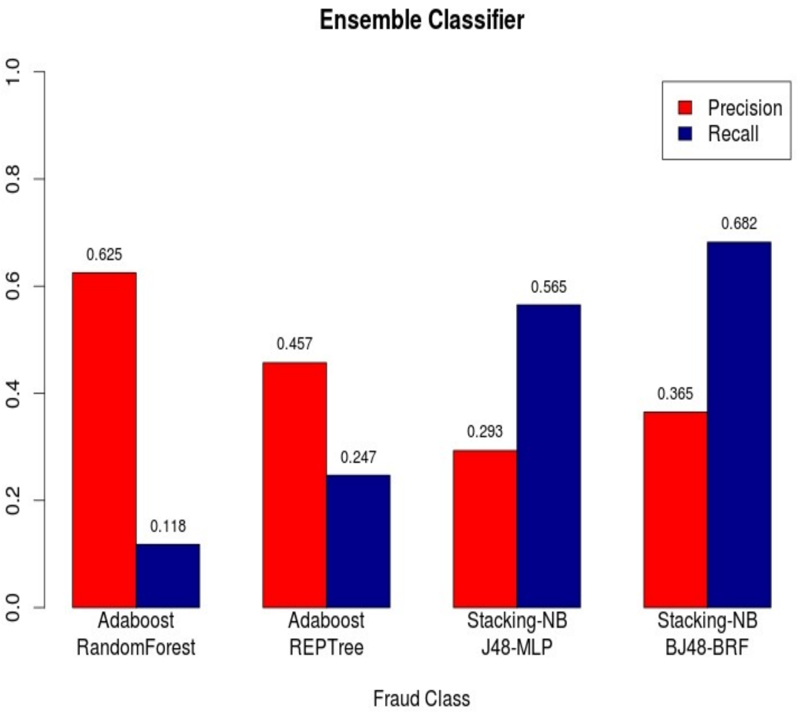


Ensemble Approach

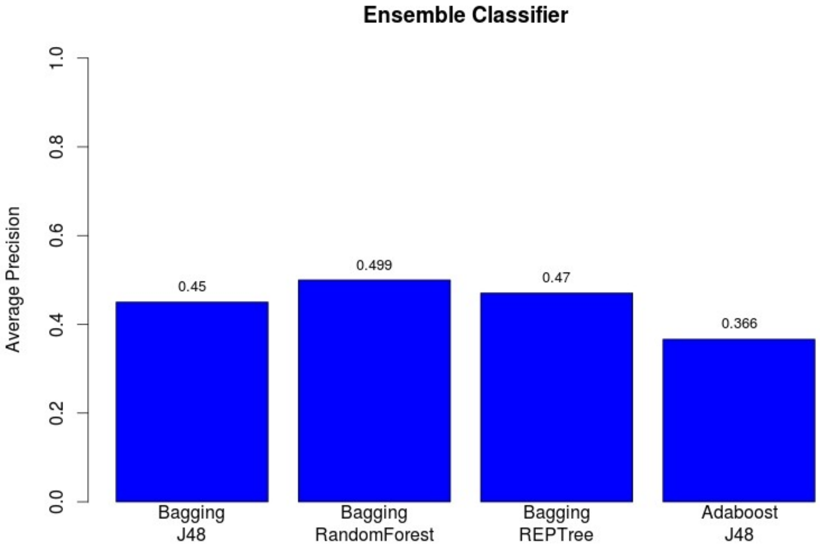
	Precision		Recall		F-Measure		ROC	
Algorithm	Ok	Fraud	Ok	Fraud	Ok	Fraud	Ok	Fraud
Bagging J48	0.978	0.655	0.997	0.224	0.987	0.333	0.928	0.928
Bagging RF	0.976	0.846	0.999	0.129	0.987	0.224	0.932	0.932
Bagging REPTRee	0.976	0.765	0.999	0.153	0.987	0.255	0.934	0.934
Adaboost J48	0.978	0.476	0.993	0.235	0.986	0.315	0.846	0.871
Adaboost RF	0.975	0.625	0.998	0.118	0.987	0.198	0.923	0.923
Adaboost REPTree	0.979	0.457	0.992	0.247	0.985	0.321	0.910	0.910
Stacking NB<J48<MLP	0.987	0.293	0.961	0.565	0.974	0.386	0.907	0.907
Stacking NB<BJ48<BRF	0.991	0.365	0.966	0.682	0.978	0.475	0.931	0.931

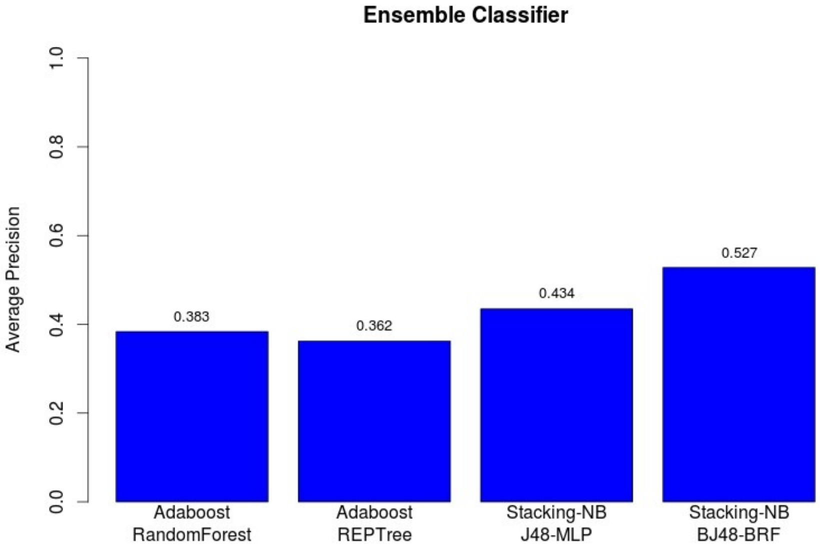
Precision-Recall





Average Precision

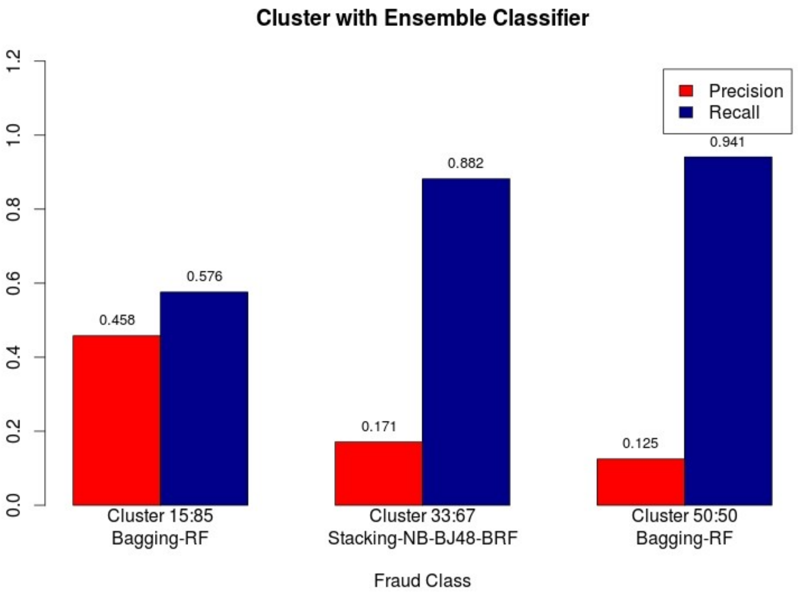


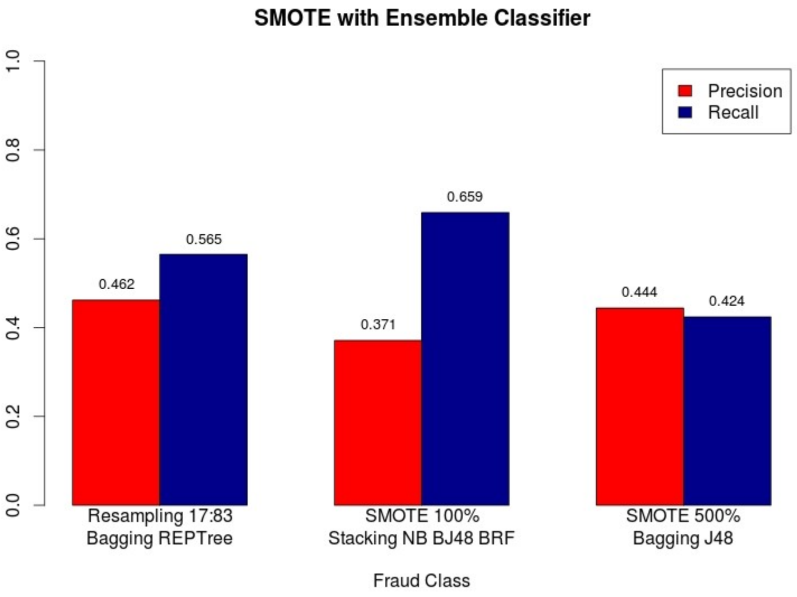


Sampling With Ensemble Learning

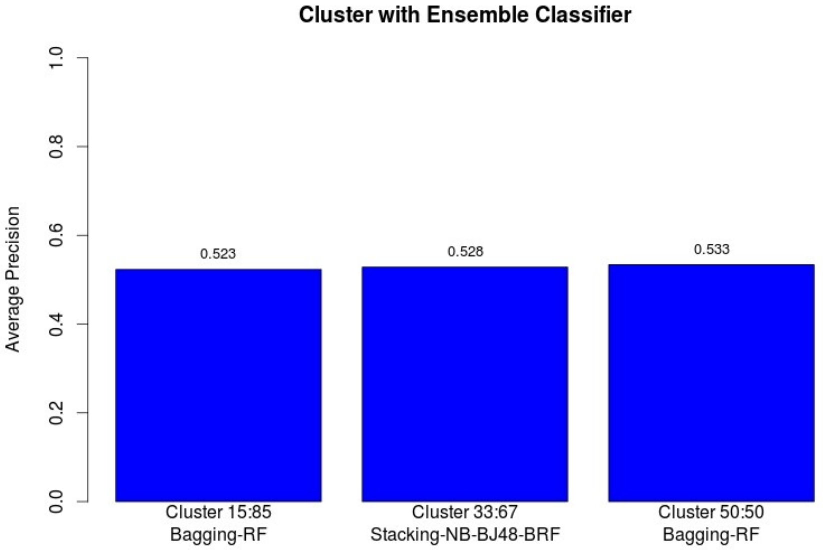
	Precision		Recall		F-Measure		ROC	
Algorithm	Ok	Fraud	Ok	Fraud	Ok	Fraud	Ok	Fraud
Cluster 15:85	0.988	0.458	0.981	0.576	0.984	**0.510**	0.928	0.928
Cluster 33:67	0.996	0.171	0.878	0.882	0.933	0.287	0.923	0.923
Cluster 50:50	0.998	0.125	0.812	0.941	0.895	0.220	0.916	0.916
Resampling	0.988	0.462	0.981	0.565	0.984	**0.508**	0.925	0.925
SMOTE 100%	0.990	0.371	0.968	0.659	0.979	0.475	0.924	0.924
SMOTE 500%	0.984	0.444	0.985	0.424	0.984	0.434	0.908	0.908

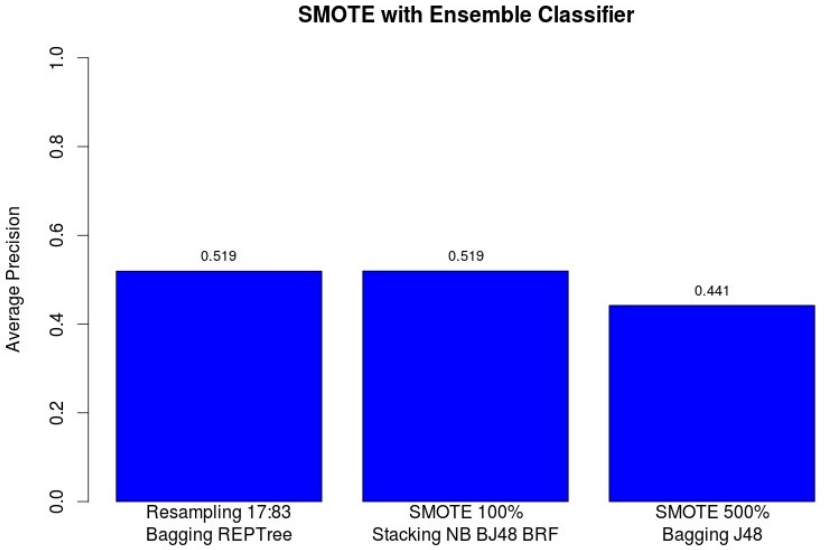
Precision-Recall





Average Precision





Whats Next

- Cost Based Algorithms
  - Minority class is more important than majority class.
  - Missclassifying a Fraud instance as OK instance has 10 times higher cost than misclassifying a OK instance as a fraud instance.
- Combining Multiple Algorithms
  - Simple averaging over the predicted confidence values for all models.
  - Majority voting.
  - Averaging on majority voting.

Whats Next

Category	Publisher Count	Fraud click (fraud %)	Night fraud click	Morning fraud click	Afternoon fraud click	Evening fraud click
adult	10	**47226 (37)**	**15435 (12)**	6439 (5)	**11299 (9)**	**14053 (11)**
Mobile content	**23**	41941 (33)	13589 (11)	**9284 (7)**	9623 (8)	9445 (7)
community	12	16411 (13)	7218 (6)	3301 (3)	2612 (2)	3280 (3)
lifestyle	14	14433 (11)	2649 (2)	3265 (3)	3573 (3)	4946 (4)
Search, portal	4	3180 (3)	682 (1)	572 (0)	689 (1)	1568 (1)
Premium portal	6	2926 (2)	351 (0)	608 (0)	732 (1)	904 (1)
Info.	3	893 (1)	49 (0)	284 (0)	428 (0)	132(0)
Total	72	127010	**39973 (31)**	23753 (19)	28956(23)	34328 (27)