



ABOUT ME

Srini is a lead engineer on the Instrumentation team within PayPal's Enterprise Data Services organization. In this role, he focuses on building the next generation tracking and analytics platform at PayPal. He is passionate about developing high performance applications and designing large-scale distributed software architecture systems in addition to continuous learning and mentoring fellow developers.

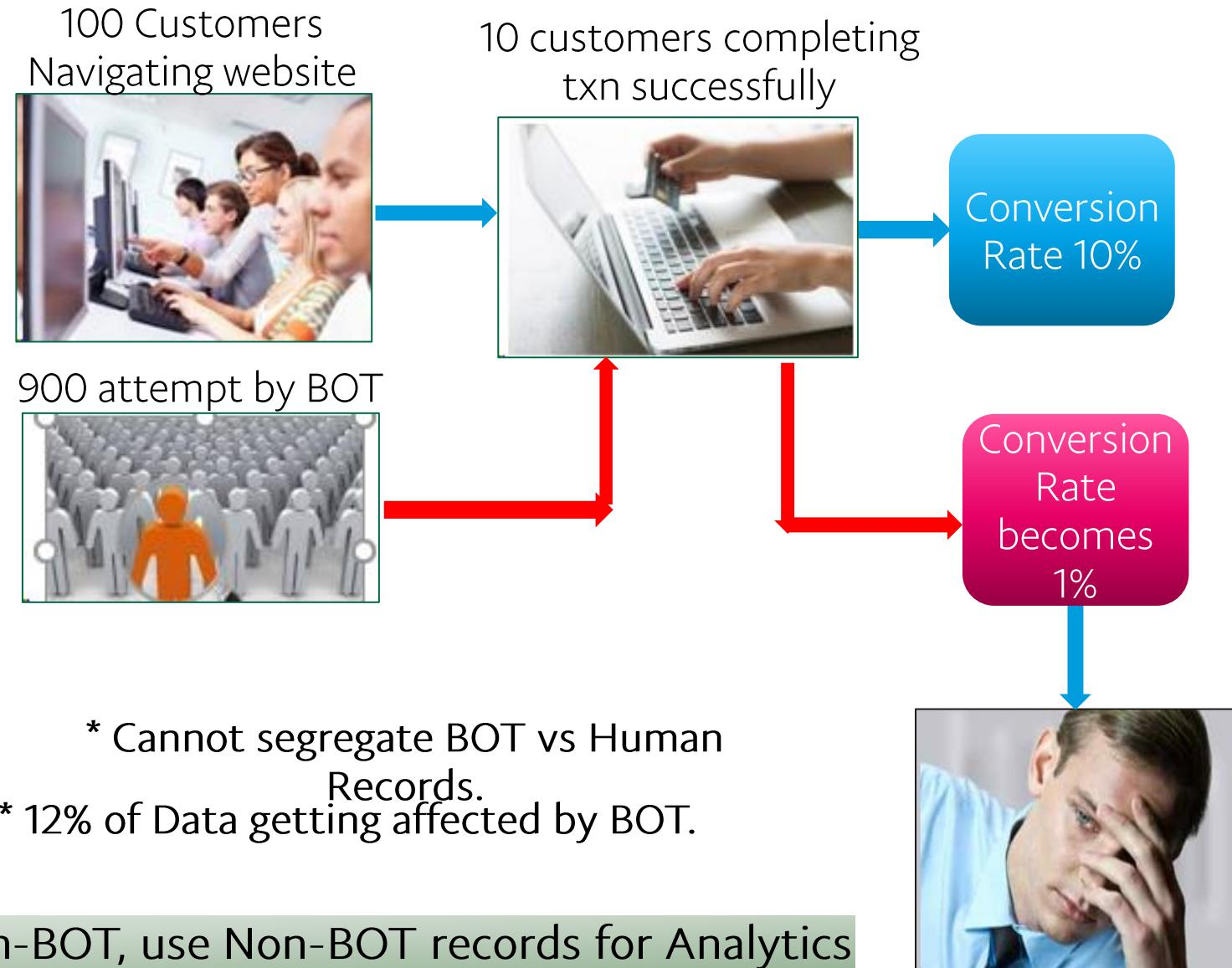
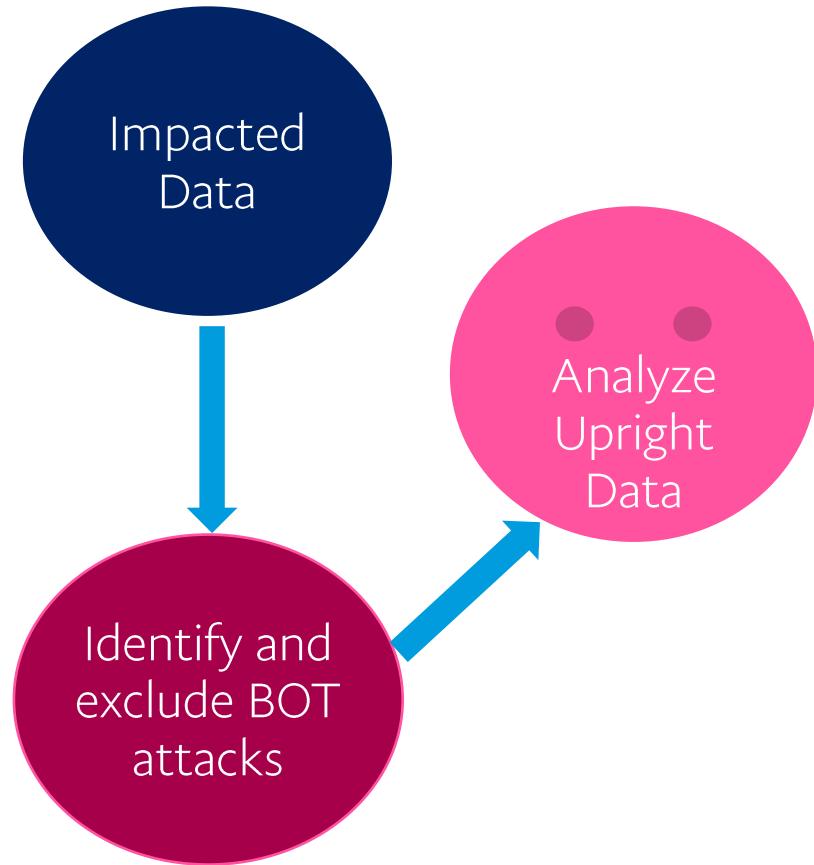
SRINIVASAN KUMAR
Lead Engineer



Intuitive Segmentation of Customer Analytics Data

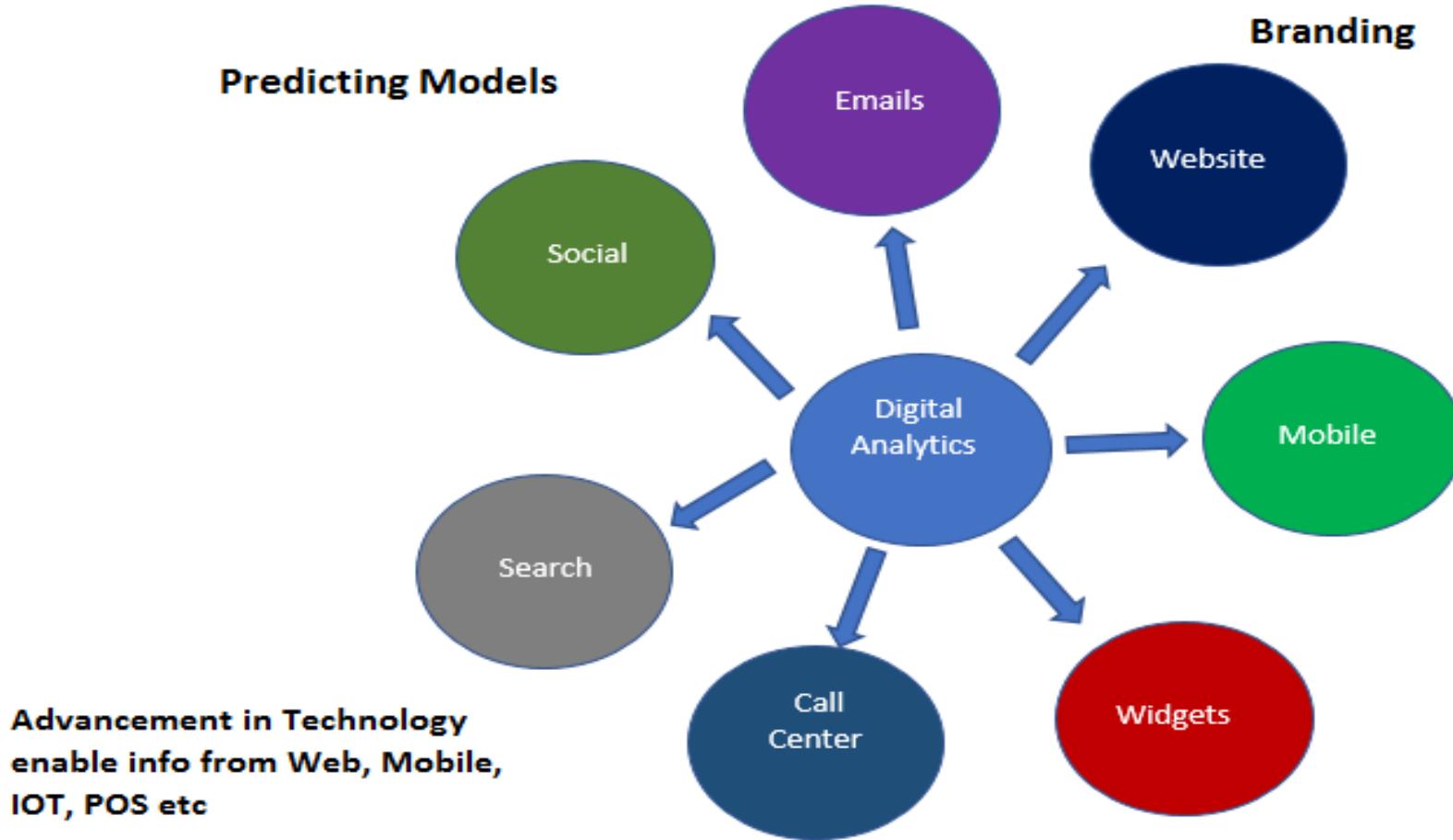
“The foundation of sound decisions is accurate data. So why introduce risk using impacted data?”

Problem Statement:

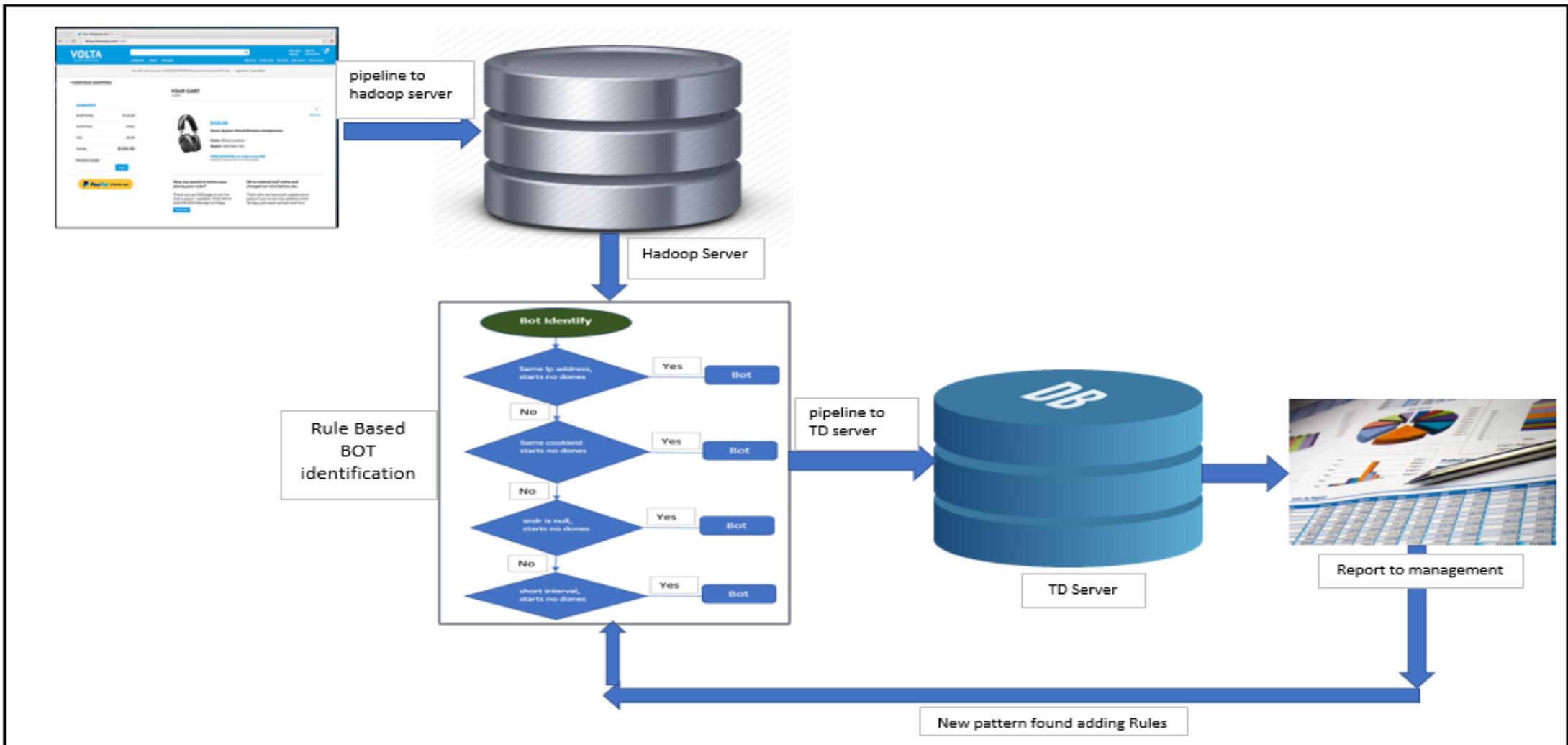


Aim: To segregate BOT vs Non-BOT, use Non-BOT records for Analytics

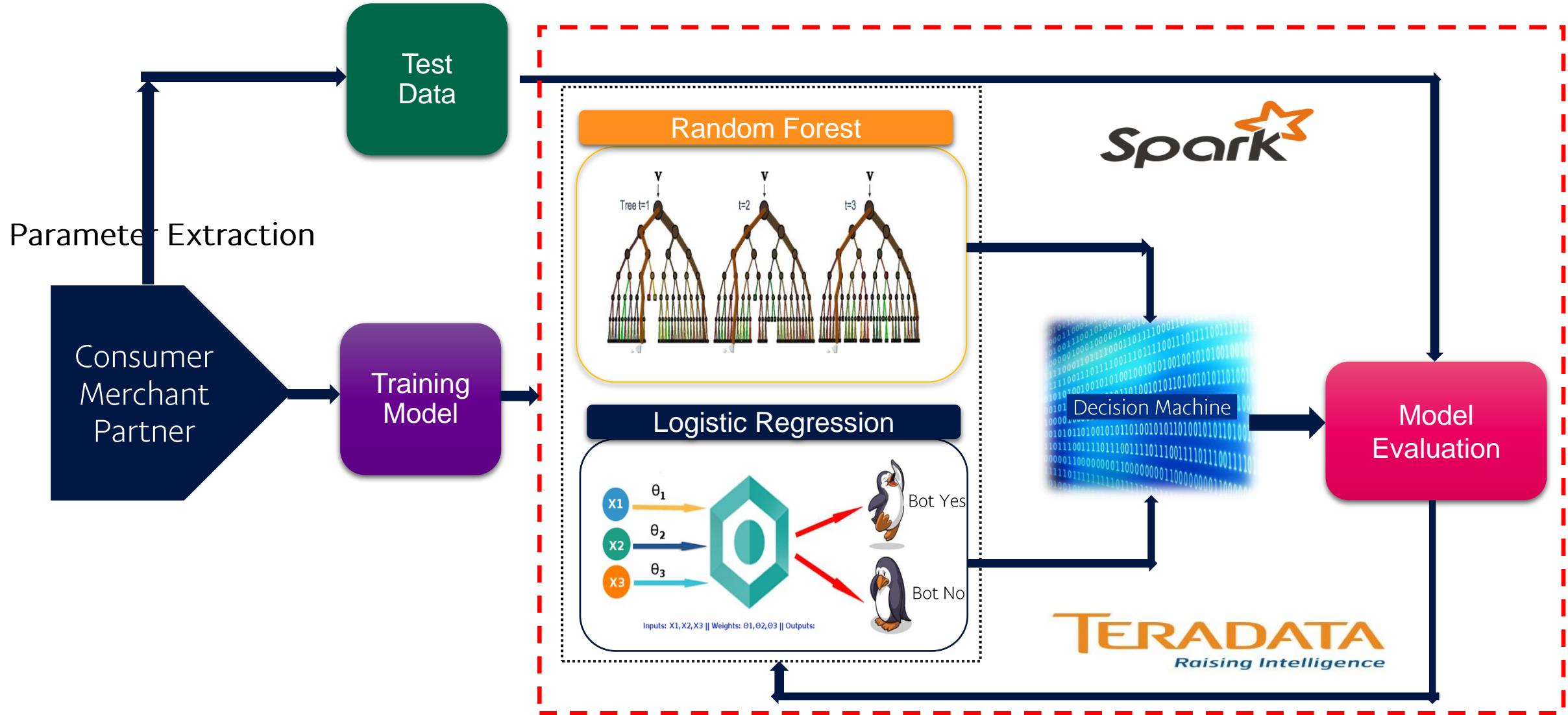
Behavioral & Digital Analytics:



Existing Rule Based Flow:

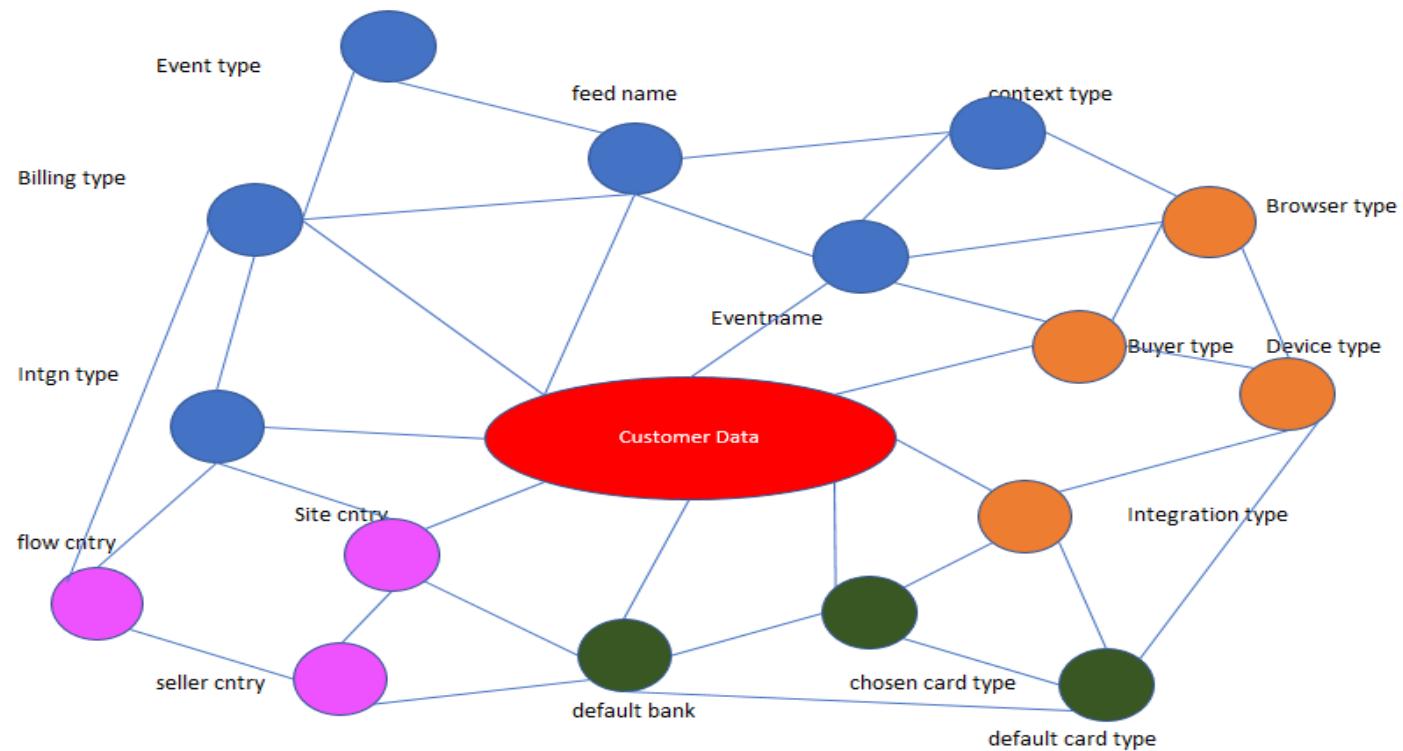


Supervised Learning:



Parameter Extraction:

Initial Parameter Count - 86



Final Parameter Count Used- 63

Hyperparameter Changes - Model:

- Number of decision trees
- Maxdepth
- Impurity – Entropy and Gini

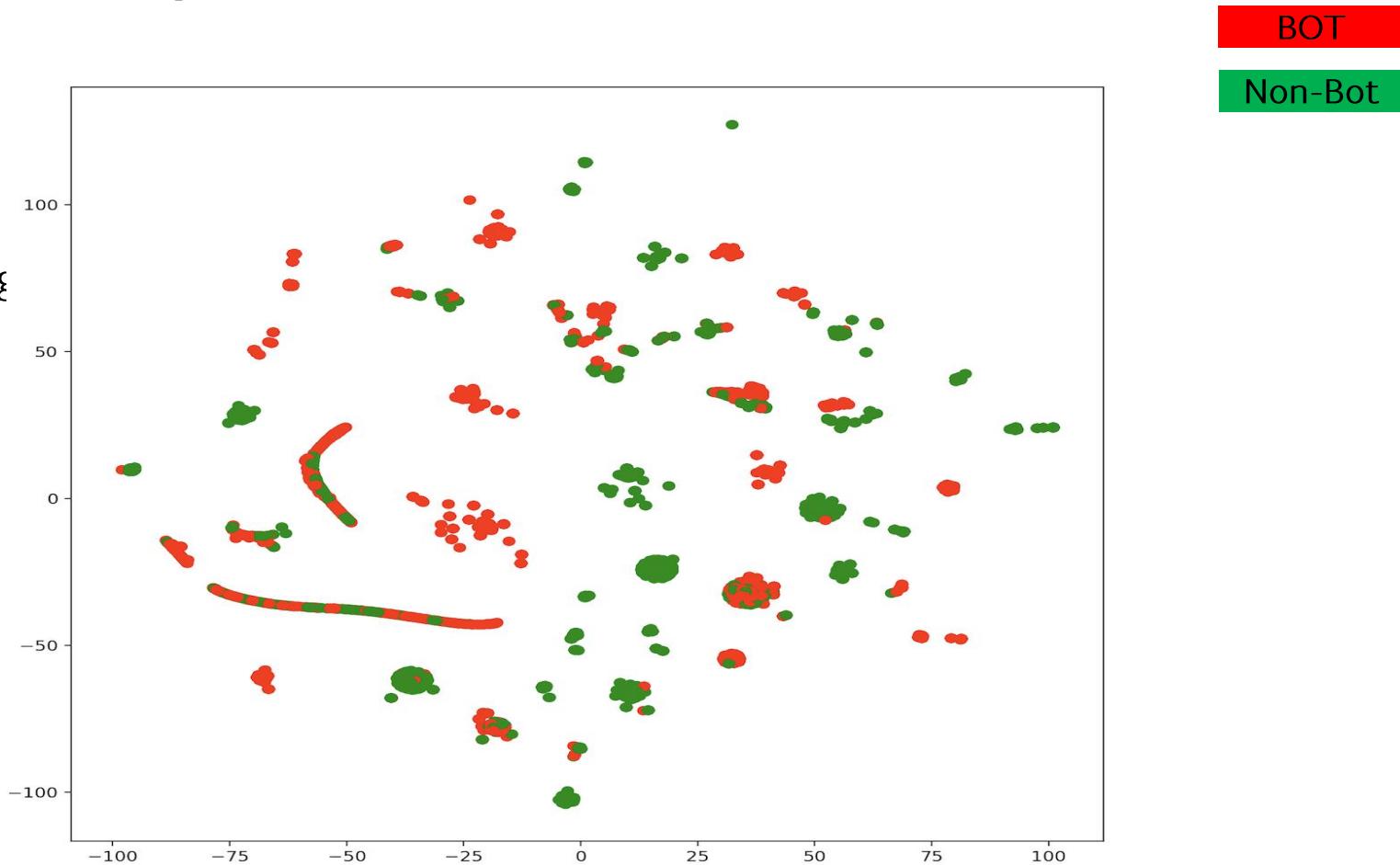
- Accuracy – Random Forest
- Accuracy – Logistic Regression

- Final Model decision.

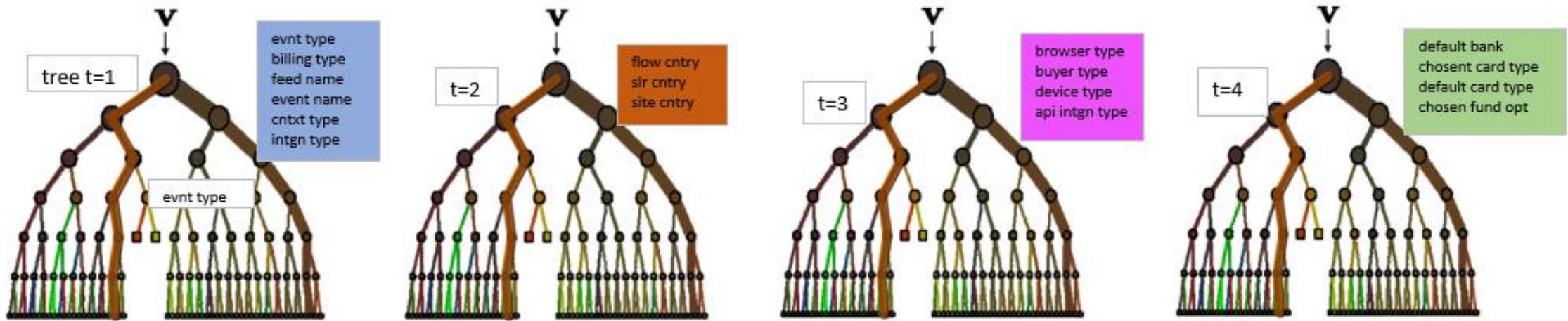


Visualization through T-SNE Algorithm:

Reducing the dimensions from 1153 to 10 using PCA (Principal Component Analysis).

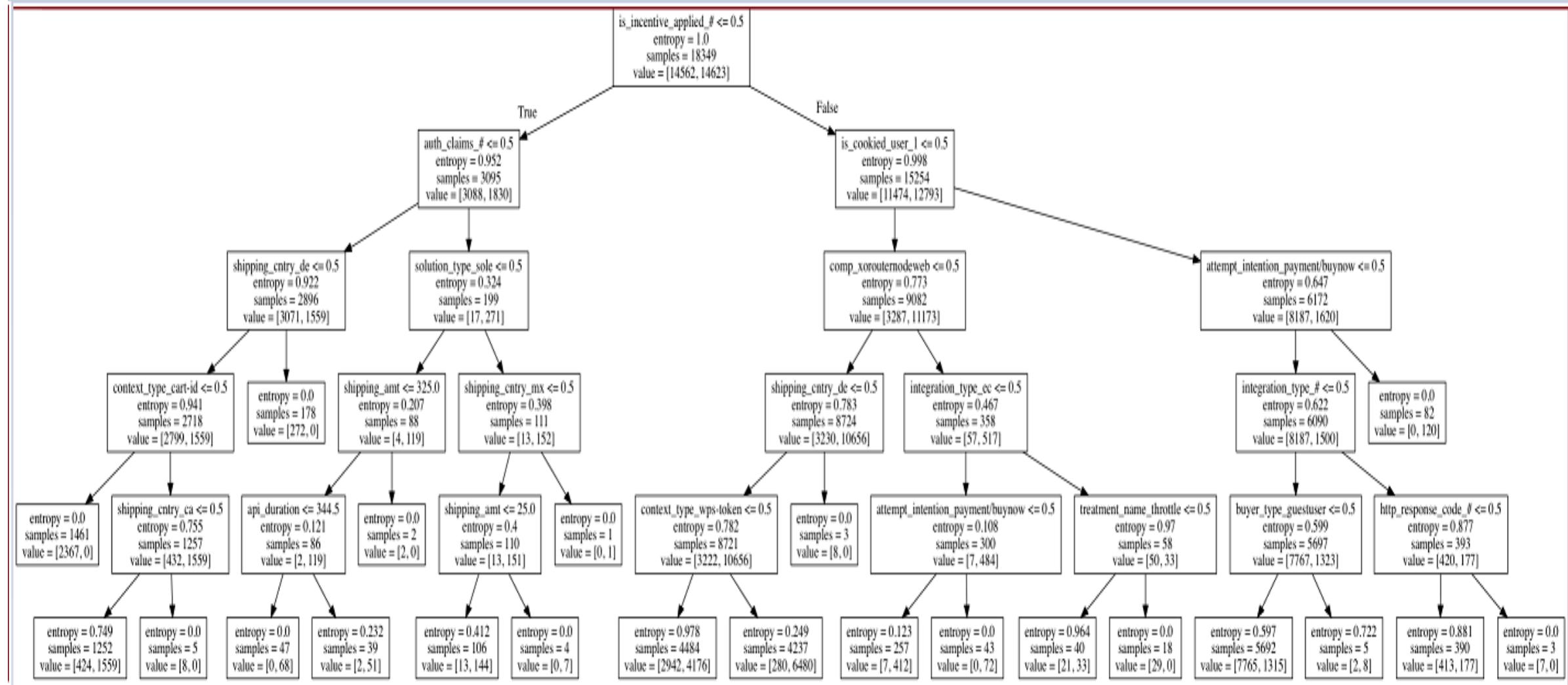


Random Forest feature split to identify BOT:



Final Result by algorithm using voting

Decision Tree after Encoding:



Model Results:

Total Sample set = 41,694 records (20,847 bots, 20,847 non-bots)

Training Records: 29185

Testing Records: 12509

Split: 70—30 (shuffle split)

Confusion Matrix : Bot Identification						
n=12.5K	Predicted: No			Predicted: Yes		
	Actual: No	TN	5943	FP	352	6295
	Actual: Yes	FN	346	TP	5868	6214
6289			6220			

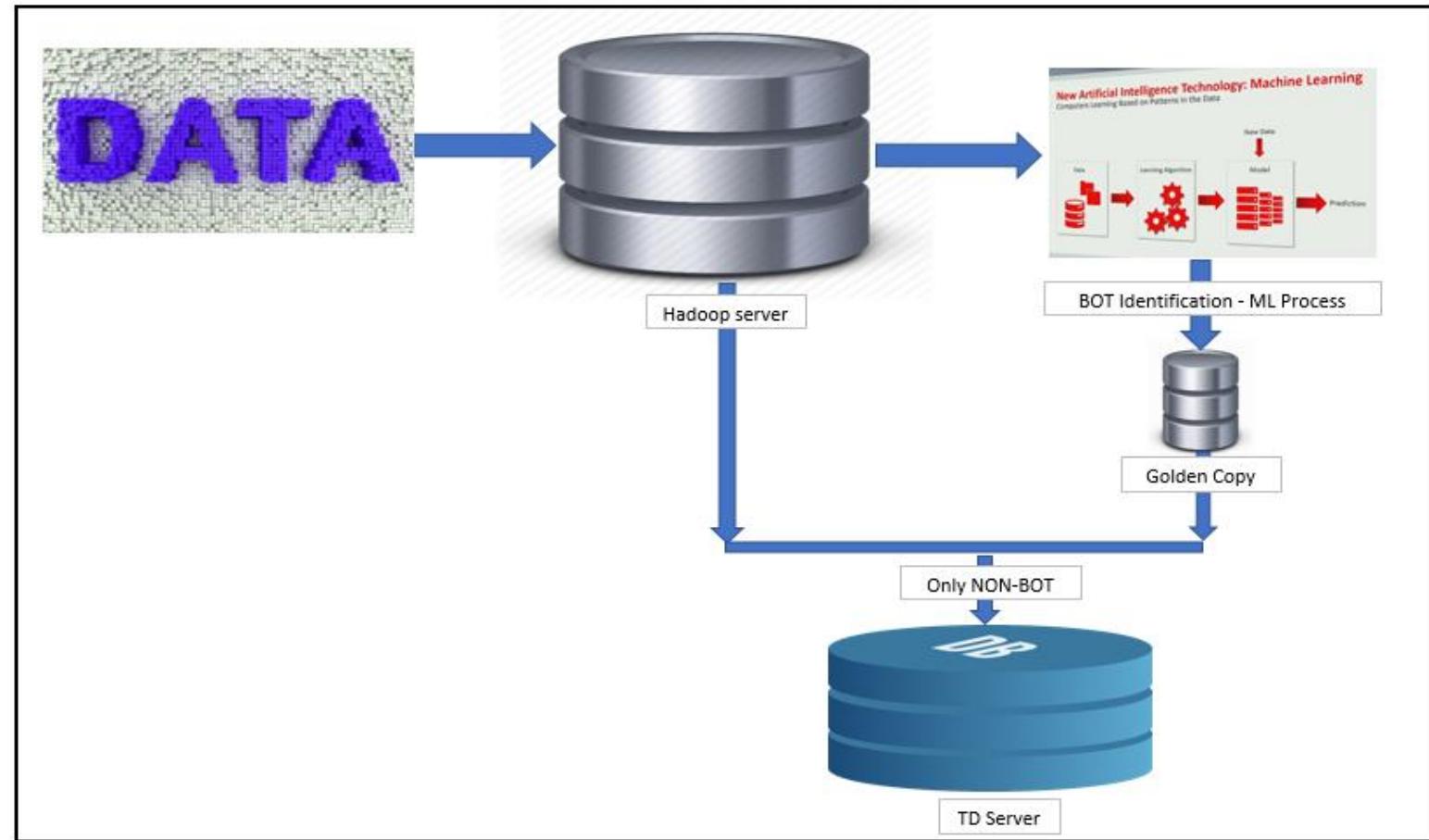
Accuracy (Tp+Tn)/total	Misclassification Rate (Fp+Fn)/total (Error Rate)	True Positive	Sensitivity or Recall (act:No, pred:Yes) (FP/actualno)	Specificity (act:No, Pred:No)(TN/actno)	Precision (Tp/predicted yes)	Prevalence (act yes/total)
0.944200176	0.055799824	0.944319279	0.055917395	0.944082605	0.94340836	0.496762331

Model Results – contd.. Feature Importance:

Sample Feature Importance
('context_type', 0.17920433754961568)
('feed_name', 0.11089302341100139)
('is_cookied_user', 0.10705843076595158)
('design', 0.084730435732830581)
('api_integration_type', 0.067297053027727274)
('http_response_code', 0.063585966042801367)

Opportunities:

- Minimize execution time.
- Reduce Storage.



Learning:

“Machines can identify Machines Better!”

Team: Srini & Vishal





Thank You !

Questions?

GREAT INDIAN **DEVELOPER** SUMMIT



2019™

Conference : April 23-26, Bangalore



Register early and get the best discounts!



www.developersummit.com



@greatindiandev



bit.ly/gidslinkedin



facebook.com/gids19



bit.ly/saltmarchyoutube



flickr.com/photos/saltmarch/