An intelligent system for churn prediction and customer retention: a case of telecommunications company

Parth Sarangi

Computer Science and Information Management
Asian Institute of Technology Thailand

Thesis Progress, September 2017

Table of Contents

Overview

- Telecom industry is highly competitive
- Government deregulation policies
- Affordable handsets and Technological advancements
- Disruptive plans and services by rival companies
- Database reveals trends of service usage
- Data mining to identify churn customers

Overview(contd.)

Operators	Customer Count in Aug 2016	Increase or Decrease in Period				Customer Count in Dec 2016
		Aug - Sep	Sep – Oct	Oct – Nov	Nov – Dec	
Airtel	257	2	2	1	2	265
Vodafone	200	0.5	1	0.8	1.8	204
Tata Indicom	58	- 1	- 1	- 1	- 1.6	52
Reliance Jio	0	15	19	16	20	72

- TRAI - Telecom regulatory authority of India
- Reports subscribers at end of every month
- Table shows Relance Jio acquiring 72 million at end of 4 months
- Launch of 4G services by Jio have jolted the revenues of Airtel, Vodafone, Tata Indicom.
- High customer churn noticed

Problem Statement

- Constant product marketing by competitors
- Various cost effective data schemes. Night time free or high speeds, or unlimited usage plans
- Proactive mindset of incumbent services provider to identify unfaithful customers
- High investment cost of acquiring new customer
- Not a fully integrated system developed for churn prediction

Objectives

Overall objective - develop an intelligent system for churn prediction and customer retention ICPCR.

Specific objectives :

- Design models and evaluate prediction performance for churn.
- Build the system of intelligent churn prediction and customer retention system.
- Evaluate the system for reliable performance.

Limitations and Scope

- Many models for churn prediction.
- Scope of this thesis is tentatively limited to build ICPCR with 3 models - Decision tree , Support Vector Machine , Artificial Neural Network

Table of Contents

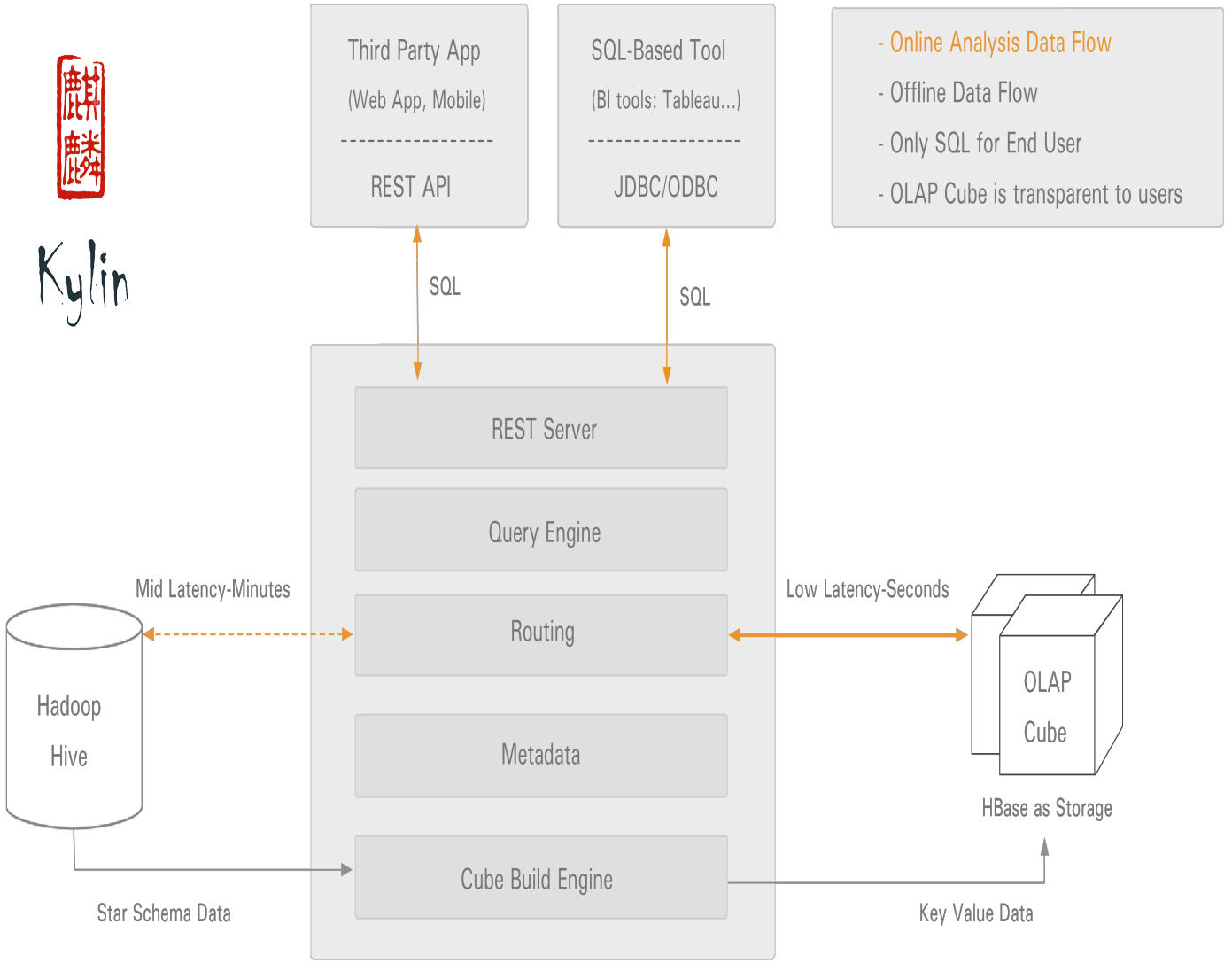
Customer Churn & Retention

- In the research paper it was found that Companies profit if they can retain customer
- Customers are most valuable asset.
- Long serving customers influence new customers to buy contracts
- Average Revenue Per User ARPU for telecom company is high if a customer stays
- If customers churn there is a loss of revenue
- Also acquiring new customer is expensive
- Forbes predicted a 10% swing in revenue if customers are retained

OLAP & Datawarehouse

- Data warehouse is a collection of Data marts
- Data marts are generally summarized tables of important data from Business units
- OLAP - Online analytical processing
- OLAP cube is the heart of an OLAP system
- There are two types - MOLAP & ROLAP. MOLAP is most common
- Apache Kylin is an open source OLAP solution

OLAP & Datawarehouse(contd.)



Data Mining

- John Naisbett (author of famous 'Megatrends') said "We are drowning in information but starved for knowledge"
- Data mining techniques can broadly be classified into two categories
  - Supervised learning
  - Un-Supervised learning

Data Mining(contd.)

Supervised Learning : The dependent and control variables are known. Classification and regression algorithms

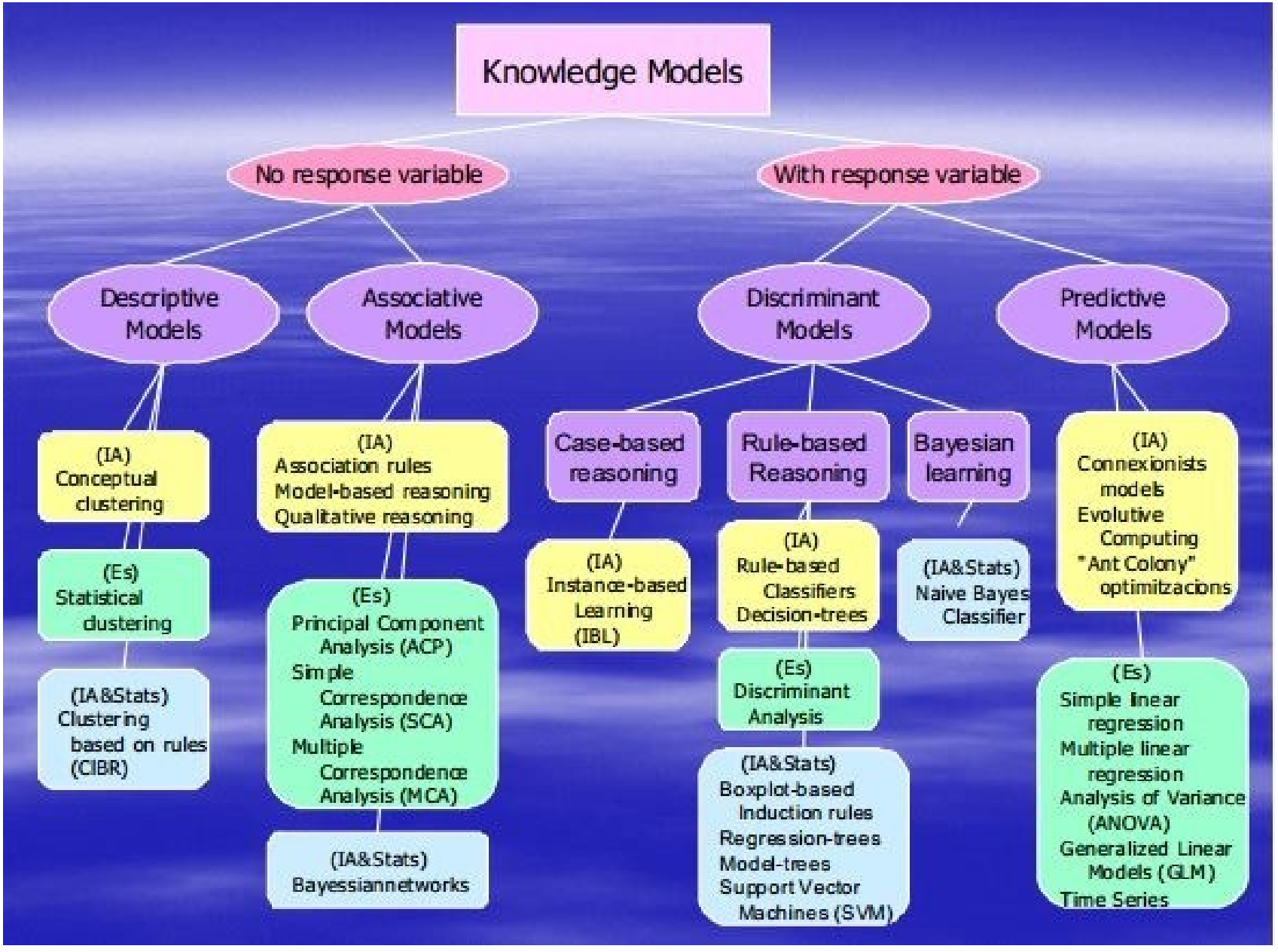
- Linear
- Multiple
- Nonlinear
- Logistic
- Decision tree
- Random forest

Data Mining(contd.)

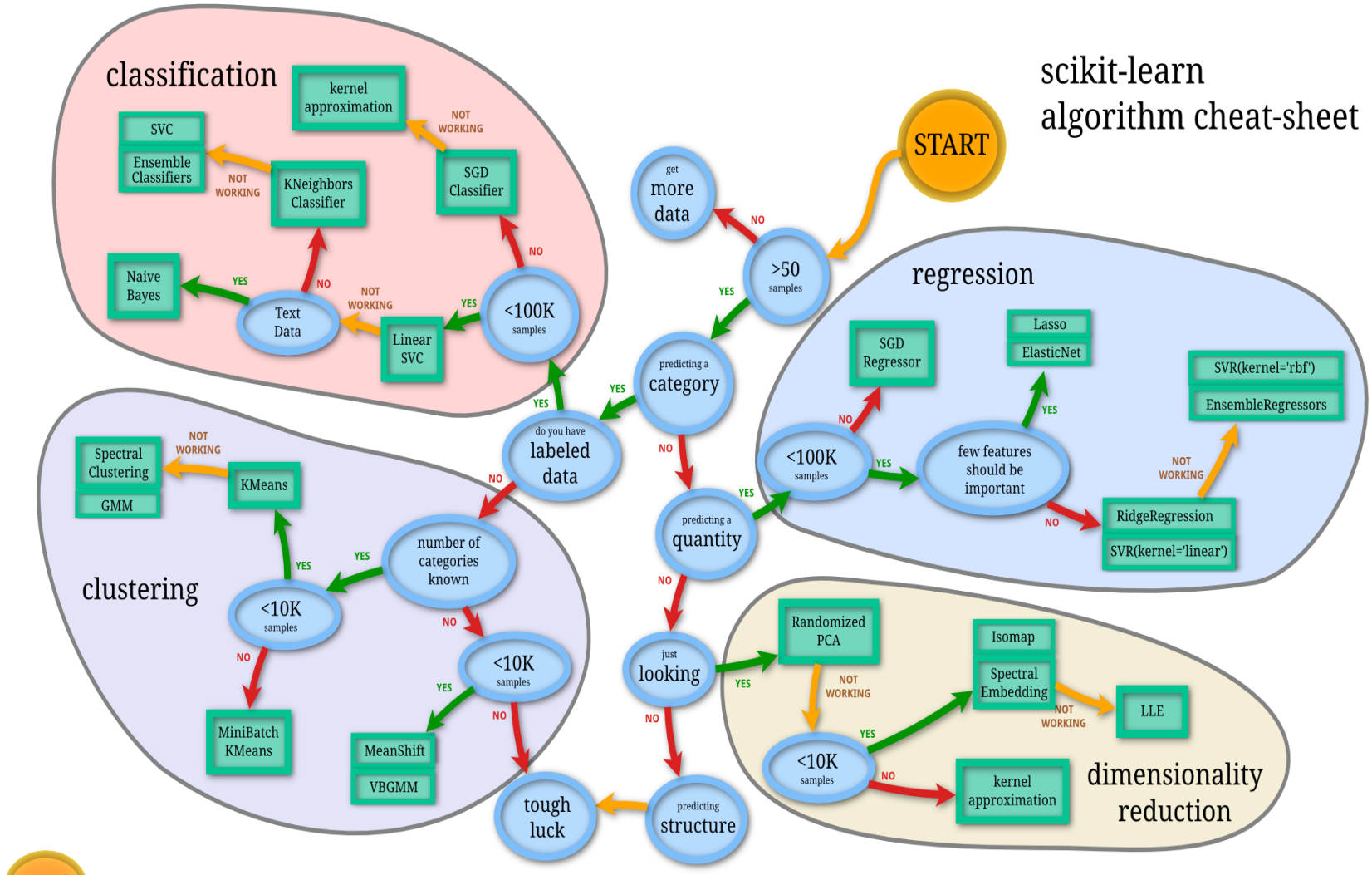
Un-Supervised Learning : The dependent and independent variables are unknown

- k means clustering
- Apriori clustering
- Hierarchical clustering
- Hidden Markov models
- Self Organizing Maps

Choosing the Right Data Mining Technique



Scikit model selection



Data Mining(contd.)

Softwares

- Weka
- Knime
- Rapidminer

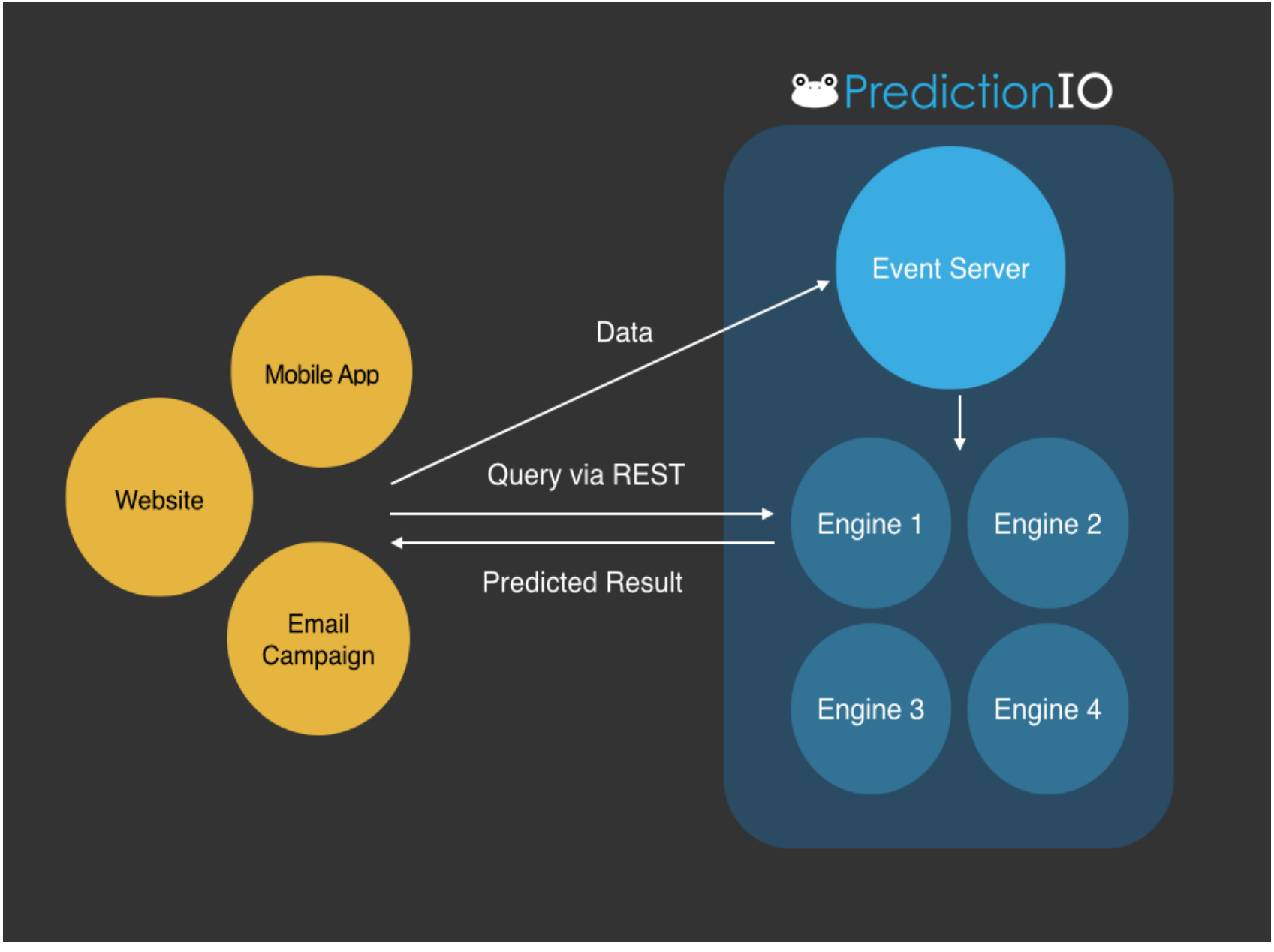
Libraries

- Tensorflow
- mlpack
- H20
- Mlib
- Scikit

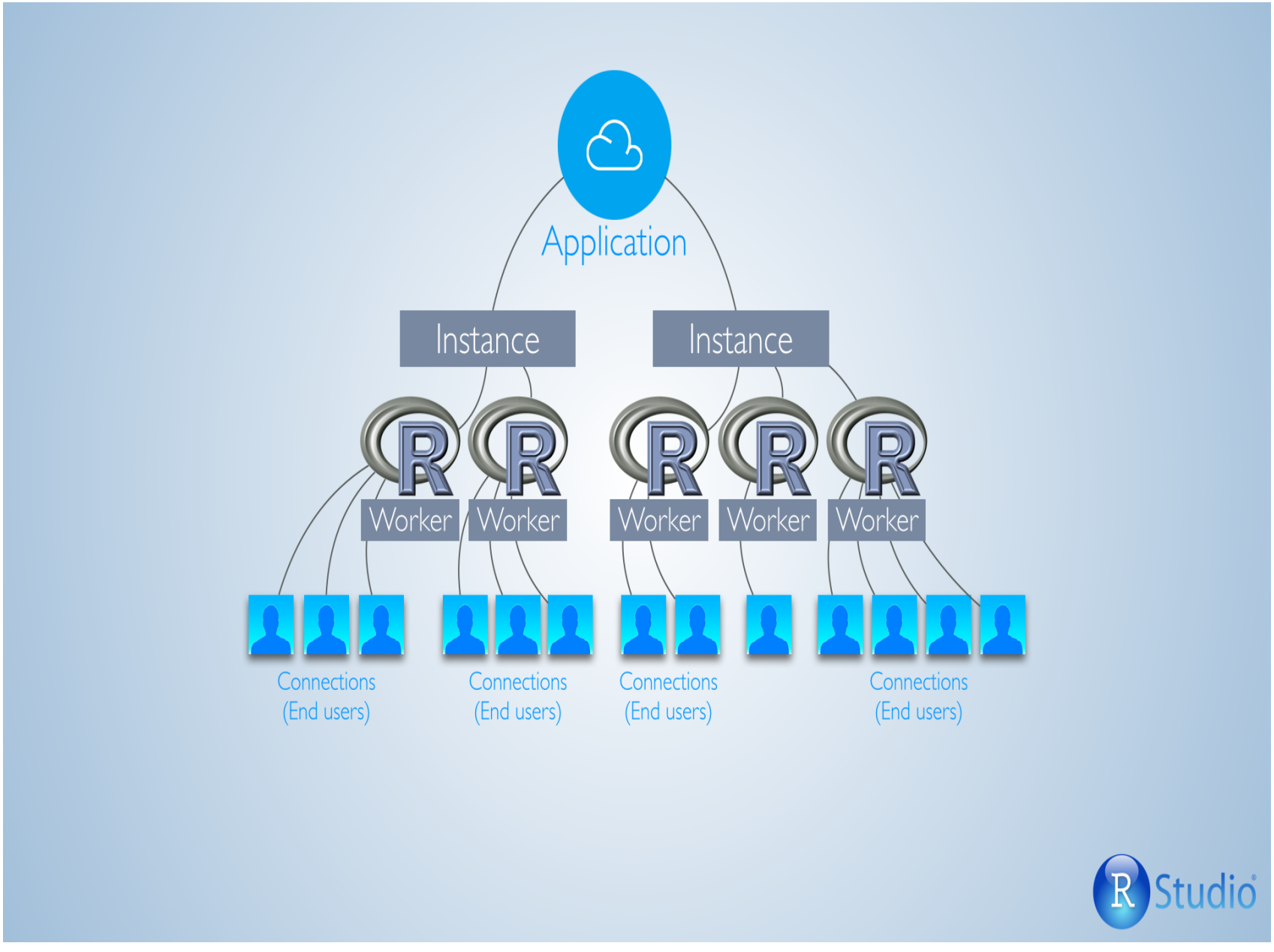
Servers

- DeepDetect
- Apache PredictionIO
- Shiny

Apache PredictionIO



Shiny Architecture



Model Evaluation Metrics

- Holdout technique
- k-fold Cross validation technique
- Sentivity and Specificity

Review of Selected Research Papers

Present 3 papers which are relevant to Churn prediction

- Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry
- A Hybrid Churn Prediction Model in Mobile Telecommunication Industry
- A comparison of machine learning techniques for customer churn prediction

Review of Selected Research Papers(contd.)

Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry

- Technique used by researchers was a Fuzzy inference system
- Combination of Neural network with fuzzy logic
- They modeled Membership functions for the attributes
- Data used was call detail record of 5000 subscribers
- It has 21 attributes and only 9 were selected
- Precision around 80%; Recall of 92.7% and Accuracy 95.8%

Review of Selected Research Papers(contd.)

A Hybrid Churn Prediction Model in Mobile Telecommunication Industry

- Presents a combination of Voted Perceptron and Logistic regression
- Compared performance with Logistic regression and Voted perceptron as individual prediction models
- WEKA software was used to model the predictors
- Call detail records of 2000 customers was sourced from Asian telecom company
- 23 attributes were used for modeling
- Results : Hybrid models performed better than individual models

Review of Selected Research Papers(contd.)

A comparison of machine learning techniques for customer churn prediction

- researchers present a well meted out comparison between the normal model functions and their corresponding boosted models
- performance criteria was based on the F-score
- series of simulations based on the Monte Carlo method
- 5 DM techniques - Back-Propagation algorithm , Support Vector Machines, Decision Trees, Naive Bayes and Logistic Regression.
- churn dataset hosted at UCI Machine learning repository
- 100-fold cross validation technique was used to reduce bias
- Ratio of training to testing set is about 2 : 3

Review of Selected Research Papers(contd.)

- boosting algorithm Adaboost Adaboost.M1
- R programming was used for modeling the simulation experiment
  1. tested classifiers run with data and performance of F-score measured
  2. boosting algorithm was applied and performance F-score measured
- Results
  - Best performance : 2 layer BPN with 15 hidden nodes and Decision tree classifier
  - SVM scored lower followed by Naive Bayes and Logit Regression at last.
  - After boosting SVM got best performance accuracy of 97% and F-score 84%

Table of Contents

Research Methodology

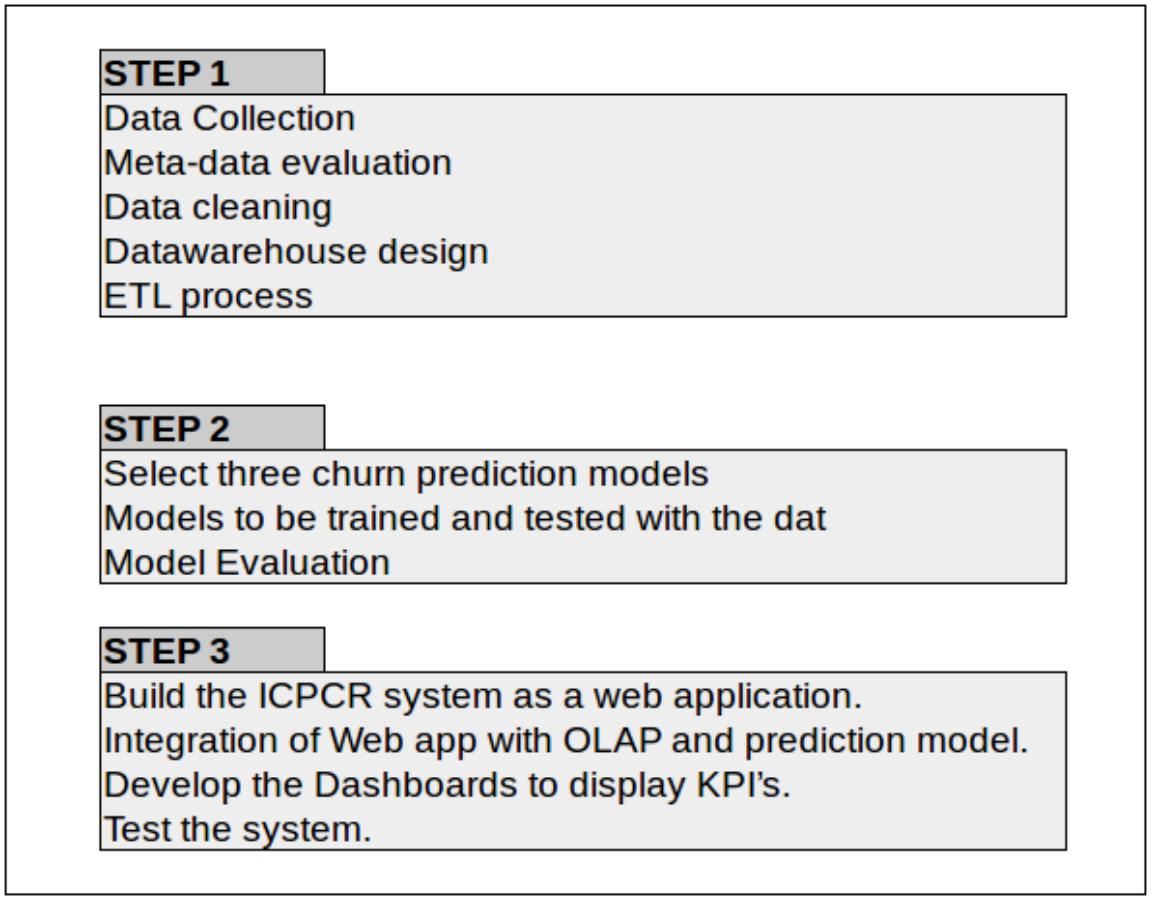


Figure: Research Methodology

Data Preprocessing and Datawarehouse Development

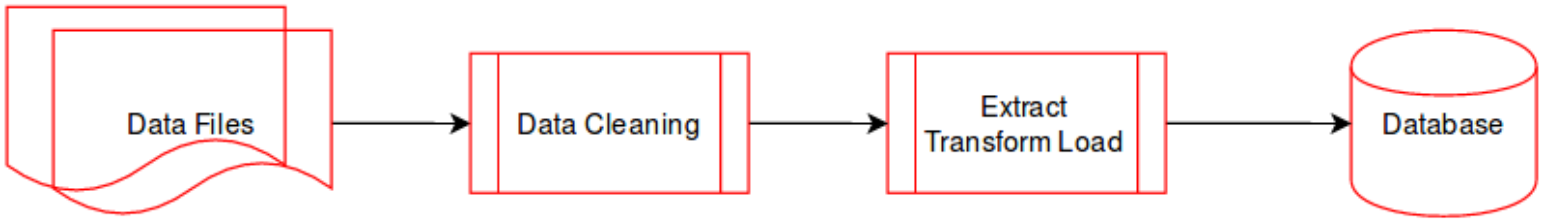


Figure: Data preprocessing

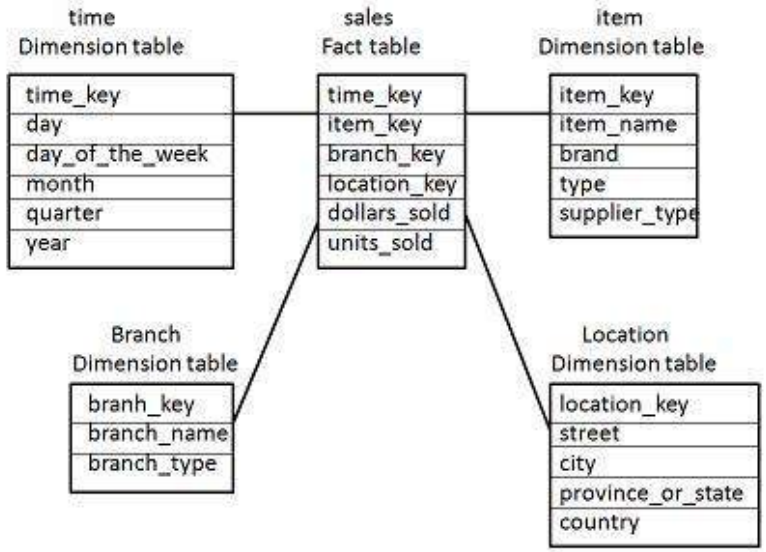


Figure: OLAP Star Schema

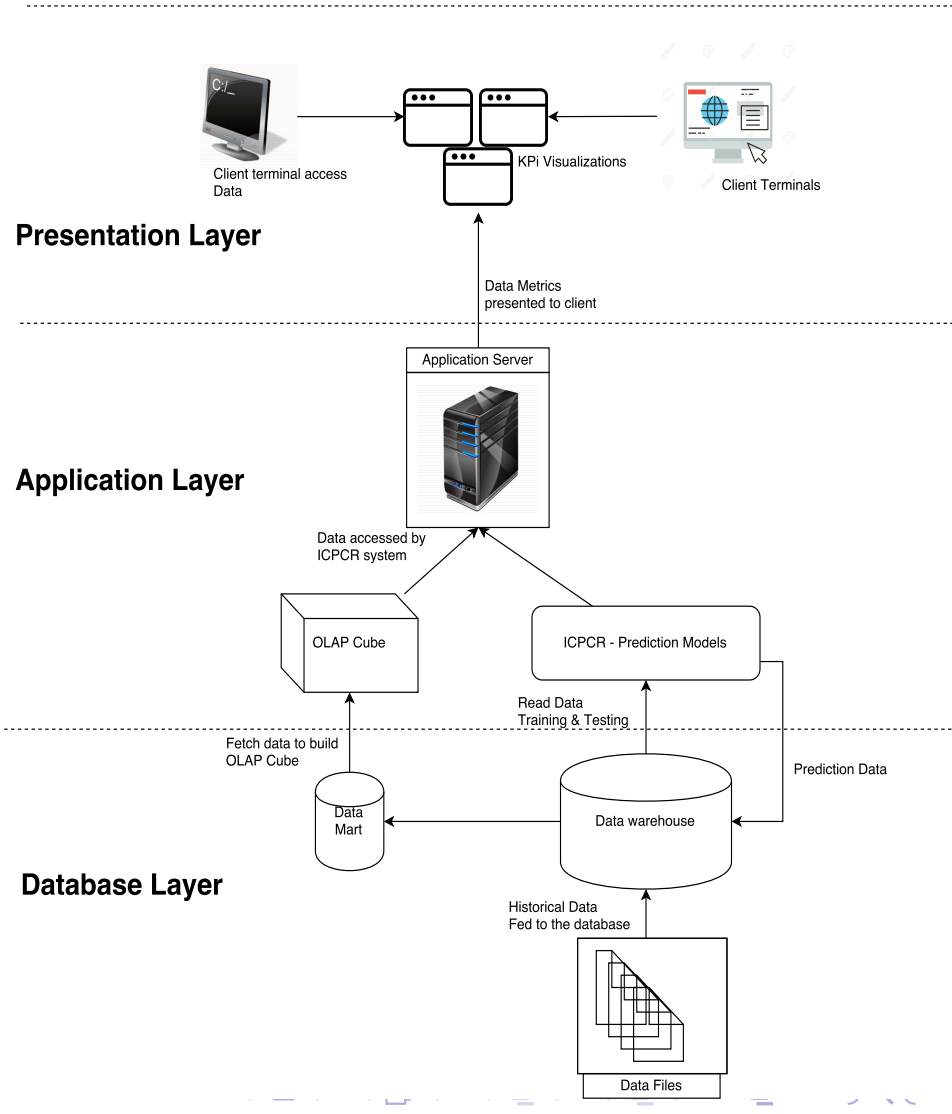
Development and Evaluation of the Prediction Models

- Model design
  - Proposal is to model 3 techniques based on Decision Tree, SVM, ANN
  - To use Machine learning libraries of either MLib, Scikit or R
  - Propose to implement a boosting algorithm Adaboost
- Model evaluation
  - K-fold cross validation technique
  - Confusion matrix with scoring of Sensitivity, Specificity, Precision and Recall, F-score

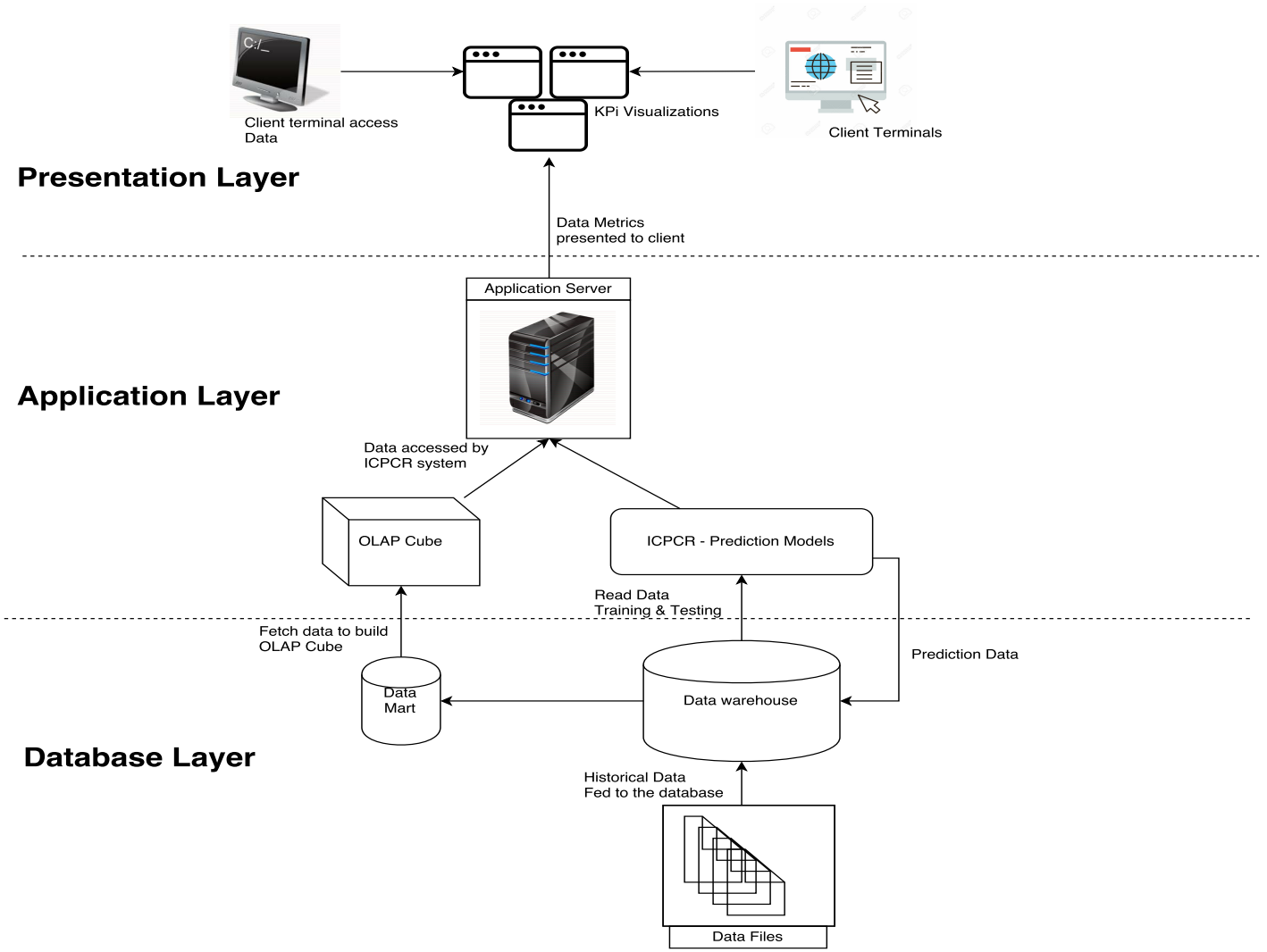
System Development & Evaluation

4 steps to be implemented.

- Presentation Layer
  - GUI for KPI's
  - Plots of predictions
- Application Layer
  - Application processing
  - Predictive model
  - OLAP cube
- Database Layer
  - Data warehouse tables in star schema
  - Data from prediction
- System Testing
  - Unit testing and Latency tests



The Intelligent Churn Prediction Architecture



Timeline

Name	Begin date	End date
Proposal	3/17/17	4/20/17
Data collection and Meta data evaluation	4/10/17	5/5/17
Data cleaning	4/24/17	5/12/17
Data warehouse design	5/15/17	6/23/17
Model selection	5/19/17	6/29/17
Model Training and Evaluation	6/15/17	7/26/17
Model Evaluation	7/3/17	8/11/17
Build ICPCR system	8/1/17	9/11/17
Integration with OLAP and Model	9/1/17	9/28/17
Dashboard development	10/2/17	11/10/17
System testing	10/20/17	11/30/17
Final defence	12/1/17	12/14/17

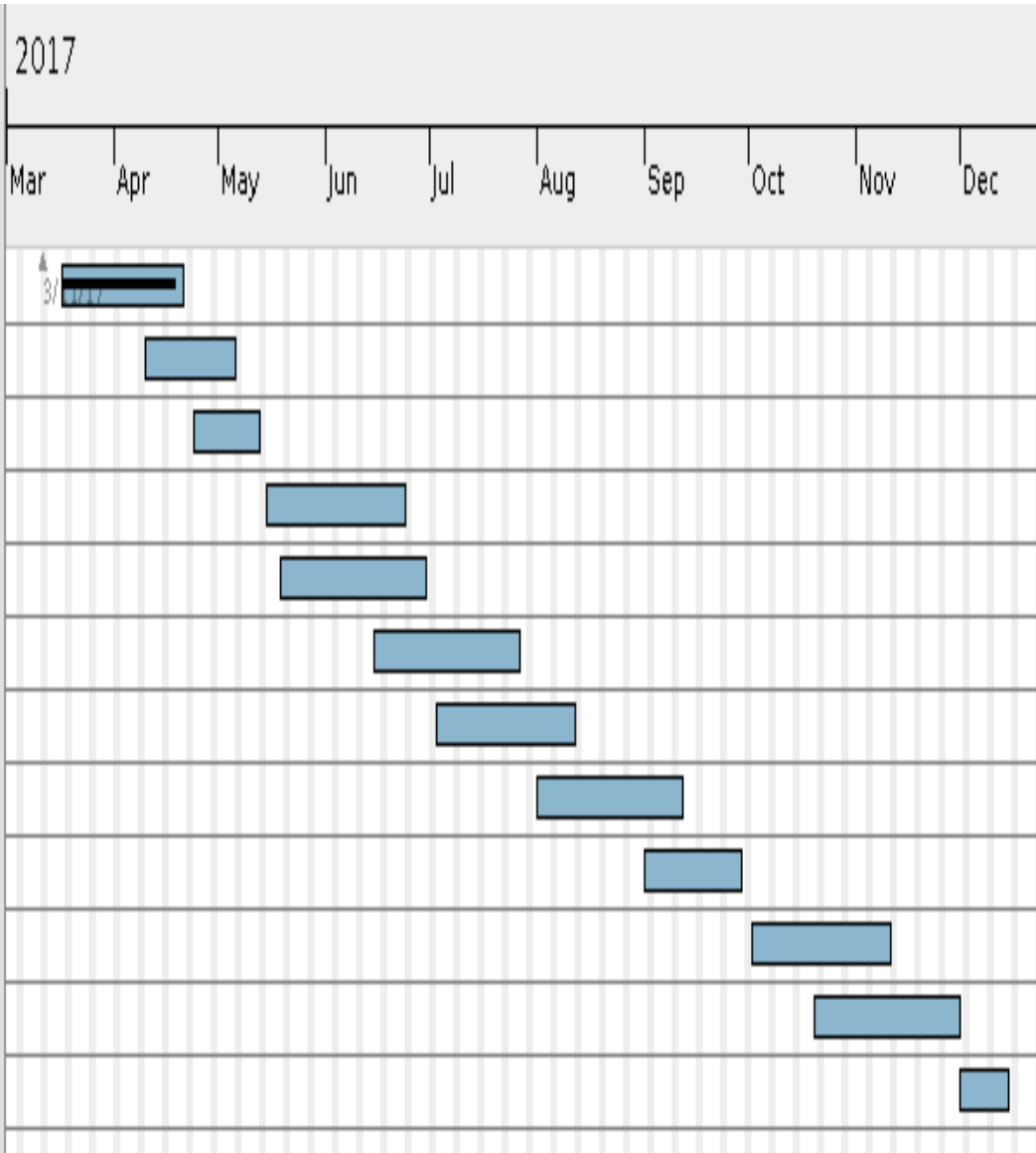


Table of Contents

Data Collection

- Churn data taken from the SGI Machine learning repository
- Dataset has 21 Dimensions
- Data has 5000 records of data

Meta-Data

Table: Variable descriptions

SNo	Name of Variable	Description	Type
1	State	state‘s of USA	discrete
2	Account Length	months of active usage	continuous
3	Area code	area code for phone	continuous
4	Phone number	phone number	discrete
5	voice mail plan	Subscribed to voice mail	discrete
6	number vmail messages	number of voice-mail messages	continuous
7	international plan	Subscribed to international plan	discrete
8	total intl minutes	total number of international calls	continuous
9	total intl calls	total charge of international calls	continuous
10	total intl charge	total charge of international calls	continuous

Meta-Data

Table: Variable descriptions

SNo	Name of Variable	Description	Type
11	total day minutes	total minutes of day calls	continuous
12	total day calls	total number of day calls	continuous
13	total day charge	total charge of day calls	continuous
14	total eve minutes	total minutes of evening calls	continuous
15	total eve calls	total number of evening call	continuous
16	total eve charge	total charge of evening calls	continuous
17	total night minutes	total minutes of night call	continuous
18	total night calls	total number of night calls	continuous
19	total night charge	total charge of night calls	continuous
20	number customer service calls	number of calls to customer service	continuous
21	churn value	if customer churned or not	discrete

Data evaluation

Churn data statistics

	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes
	WV : 158	Min. : 1.0	Min. :408.0	327-1058: 1	no :4527	no :3677	Min. : 0.000	Min. : 0.0
	MN : 125	1st Qu.: 73.0	1st Qu.:408.0	327-1319: 1	yes: 473	yes:1323	1st Qu.: 0.000	1st Qu.:143.7
	AL : 124	Median :100.0	Median :415.0	327-2040: 1			Median : 0.000	Median :180.1
	ID : 119	Mean :100.3	Mean :436.9	327-2475: 1			Mean : 7.755	Mean :180.3
	VA : 118	3rd Qu.:127.0	3rd Qu.:415.0	327-3053: 1			3rd Qu.:17.000	3rd Qu.:216.2
	OH : 116	Max. :243.0	Max. :510.0	327-3587: 1			Max. :52.000	Max. :351.5
	(Other):4240			(Other) :4994				

total day calls	total day charge	total eve minutes	total eve calls	total eve charge	total night minutes	total night calls	total night charg
Min. : 0	Min. : 0.00	Min. : 0.0	Min. : 0.0	Min. : 0.00	Min. : 0.0	Min. : 0.00	Min. : 0.000
1st Qu.: 87	1st Qu.:24.43	1st Qu.:166.4	1st Qu.: 87.0	1st Qu.:14.14	1st Qu.:166.9	1st Qu.: 87.00	1st Qu.: 7.510
Median :100	Median :30.62	Median :201.0	Median :100.0	Median :17.09	Median :200.4	Median :100.00	Median : 9.020
Mean :100	Mean :30.65	Mean :200.6	Mean :100.2	Mean :17.05	Mean :200.4	Mean : 99.92	Mean : 9.018
3rd Qu.:113	3rd Qu.:36.75	3rd Qu.:234.1	3rd Qu.:114.0	3rd Qu.:19.90	3rd Qu.:234.7	3rd Qu.:113.00	3rd Qu.:10.560
Max. :165	Max. :59.76	Max. :363.7	Max. :170.0	Max. :30.91	Max. :395.0	Max. :175.00	Max. :17.770

Data evaluation contd.

```
total intl minutes total intl calls total intl charge number customer service calls  churn value
Min.   : 0.00      Min.   : 0.000   Min.   :0.000     Min.   :0.00                    False.:4293
1st Qu.: 8.50      1st Qu.: 3.000   1st Qu.:2.300     1st Qu.:1.00                    True. : 707
Median :10.30      Median : 4.000   Median :2.780     Median :1.00
Mean   :10.26      Mean   : 4.435   Mean   :2.771     Mean   :1.57
3rd Qu.:12.00      3rd Qu.: 6.000   3rd Qu.:3.240     3rd Qu.:2.00
Max.   :20.00      Max.   :20.000   Max.   :5.400     Max.   :9.00
```

Churn prediction models

- Support Vector Machine trained Linear and Radial models
- Decision Tree trained and predicted with classification type dt
- Neural networks : Currently still in processing phase. Error with input variable types

Decision Tree 1

Decision Tree 1 Confusion Matrix

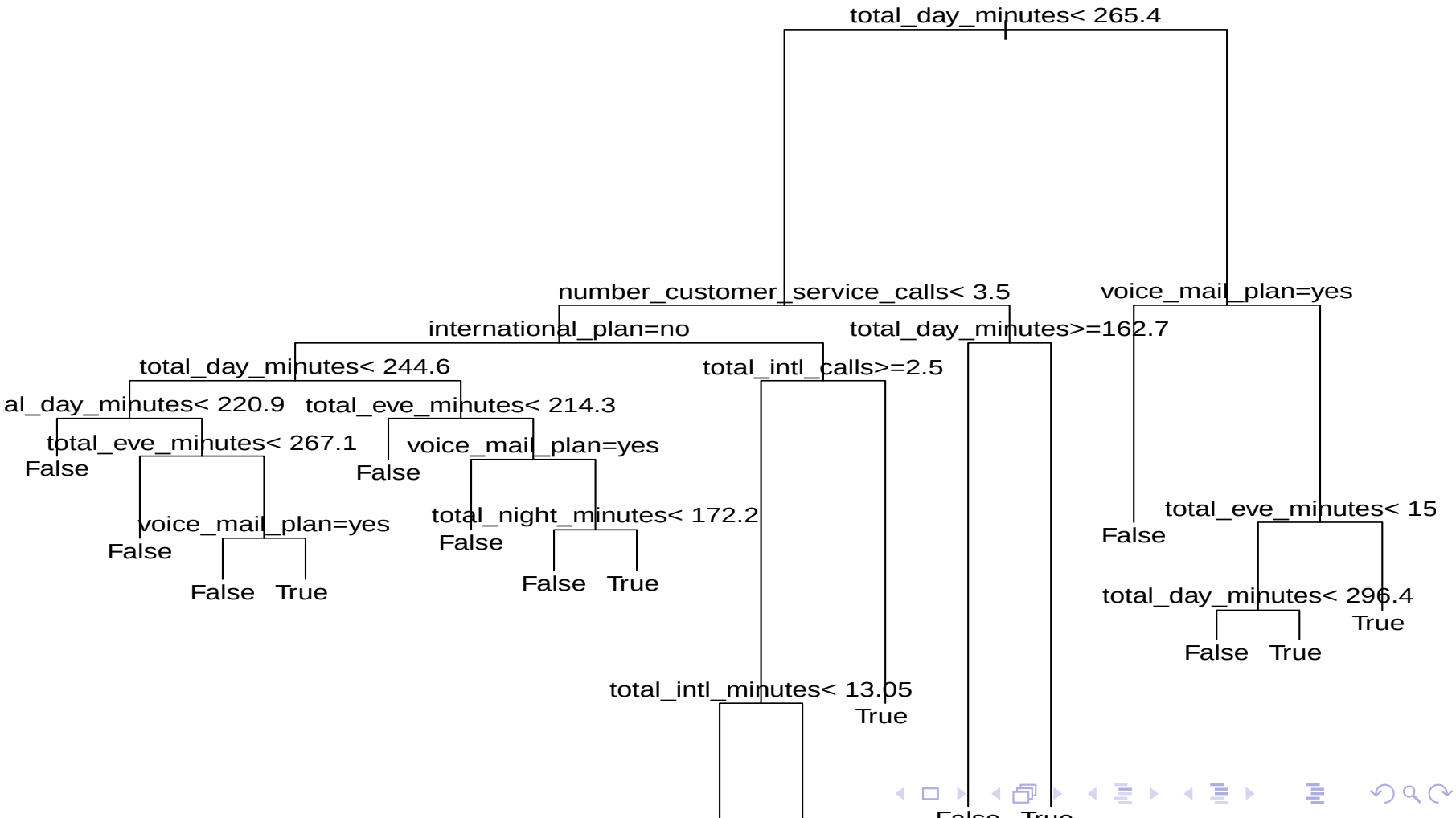
Prediction	False	True
False	1266	109
True	19	106

Decision Tree 1 Statistics

Table: DT-1 Stats

Accuracy	:	0.9146667
95% CI	:	(0.8993698, 0.9283156)
No Information Rate	:	0.8566667
P-Value [Acc >NIR]	:	0.00000000000511805609
Sensitivity	:	0.9852140
Specificity	:	0.4930233

Decision Tree 2



Decision Tree 2 Confusion Matrix

Prediction	False	True
False	1266	68
True	19	147

Decision Tree 2 Statistics

Table: DT-1 Stats

Accuracy	:	0.942
95% CI	:	(0.9289464, 0.9532869)
No Information Rate	:	0.8566667
P-Value [Acc >NIR]	:	0.00000000000511805609
Sensitivity	:	0.9852140
Specificity	:	0.4930233

SVM Linear Kernel Stats

- Training sameple : 3500
- Testing sample : 1500
- 18 predictors
- 2 classes: 'False', 'True'
- Pre-processing: centered (69), scaled (69)
- Resampling: Cross-Validated (10 fold, repeated 3 times)
- Resampling results
  - Accuracy 0.8595245
  - Kappa 0.003825618

SVM Linear Kernel Confusion Matrix

Prediction	False	True
False	1285	215
True	0	0

SVM Linear Kernel Statistics

Table: SVM Linear Stats

Accuracy	:	0.8567
95% CI	:	(0.8379, 0.874)
No Information Rate	:	0.8567
P-Value [Acc >NIR]	:	0.5182
Sensitivity	:	1
Specificity	:	0

SVM Radial Kernel Stats

- Training sameple : 3500
- Testing sample : 1500
- 18 predictors
- 2 classes: 'False', 'True'
- Pre-processing: centered (69), scaled (69)
- Resampling: Cross-Validated (10 fold, repeated 3 times)
- Resampling results
  - Accuracy 0.8594297
  - Kappa 0

SVM Radial Kernel Confusion Matrix

Prediction	False	True
False	1266	68
True	19	147

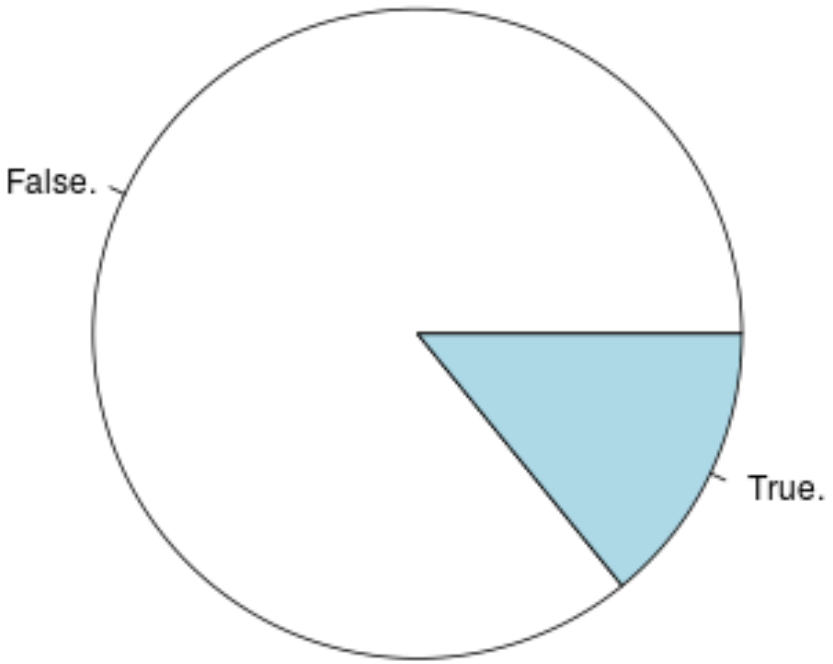
SVM Radial Kernel Statistics

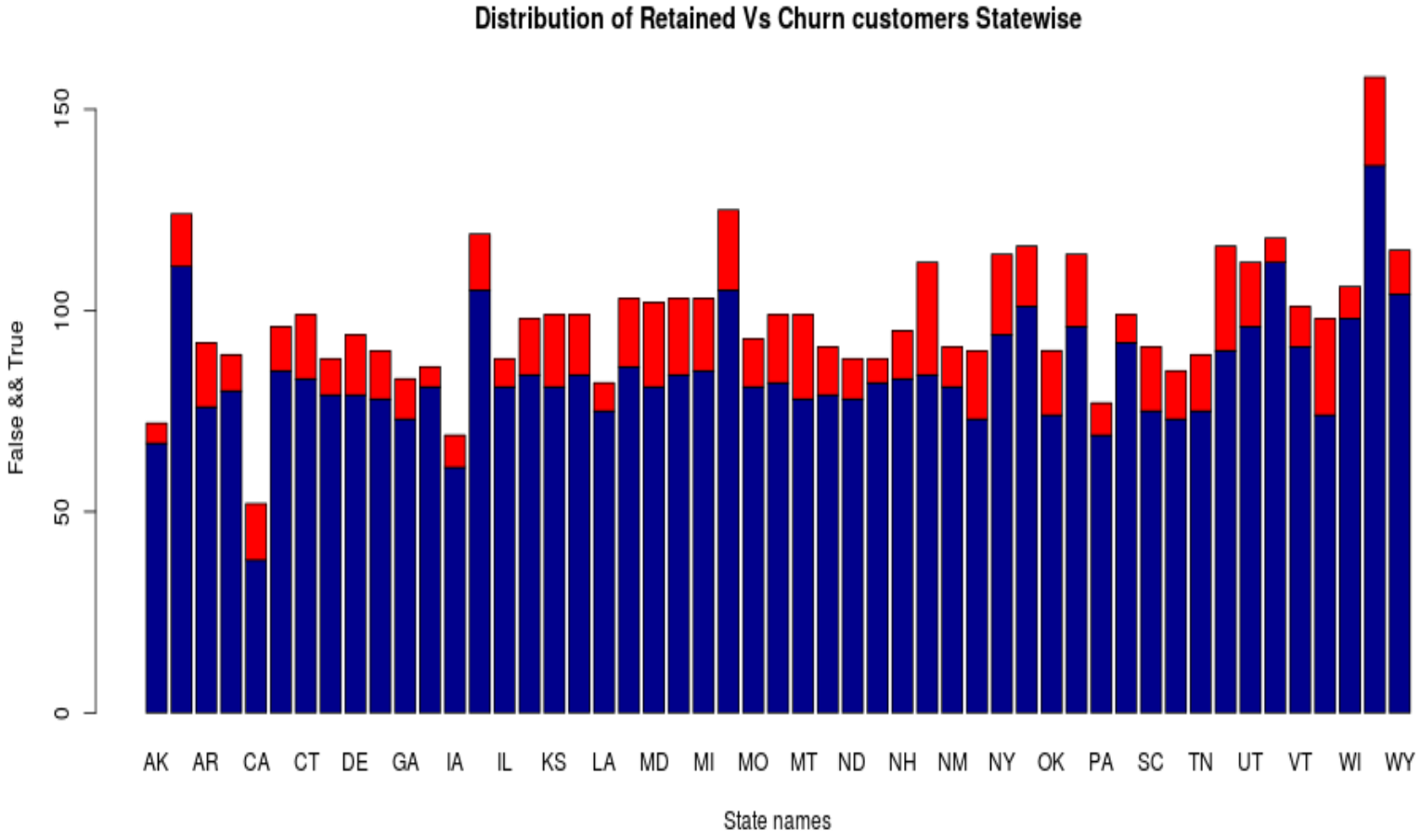
Table: SVM Radial Stats

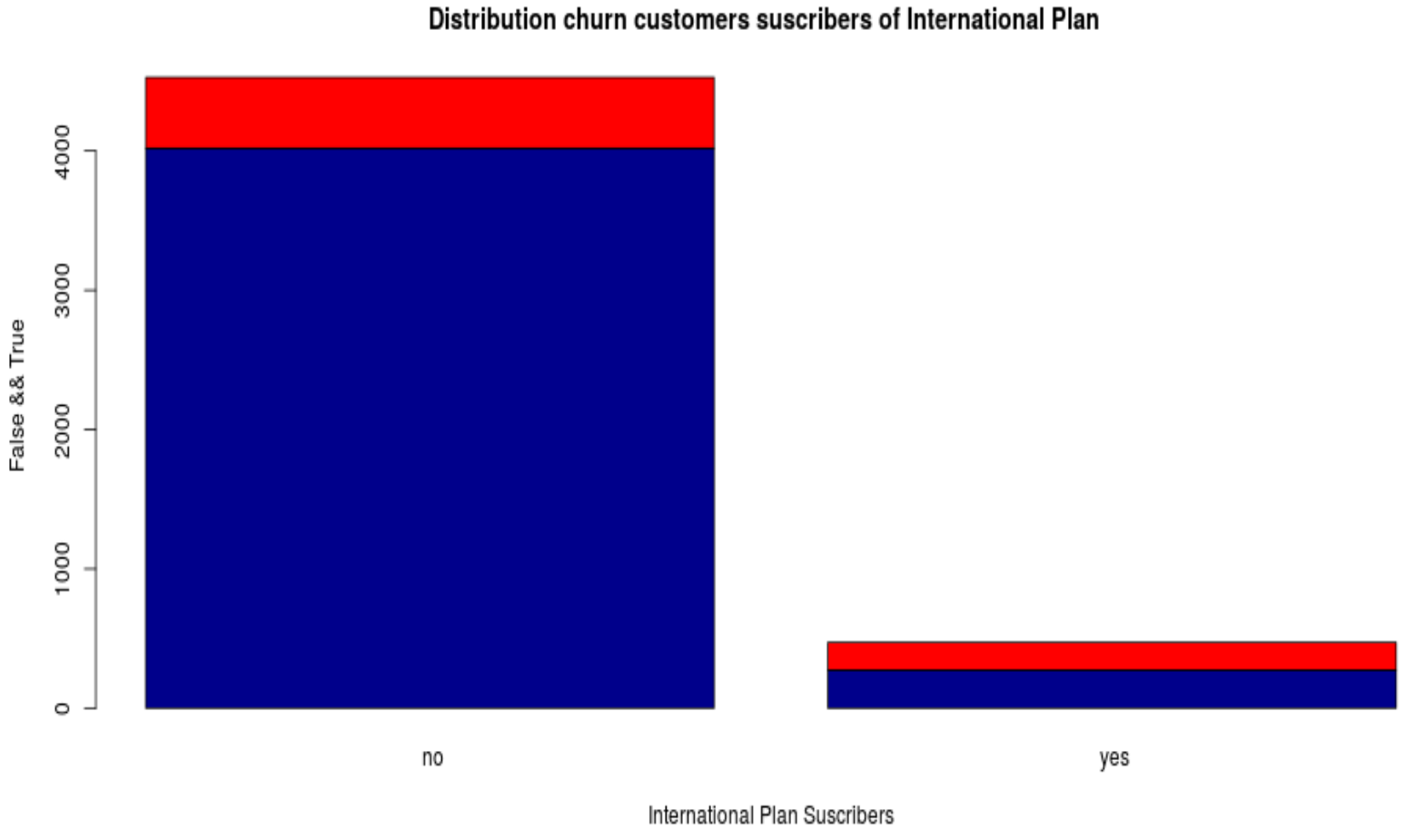
Accuracy	:	0.8566667
95% CI	:	(0.8379028, 0.8740215)
No Information Rate	:	0.8566667
P-Value [Acc >NIR]	:	0.5181819
Sensitivity	:	1
Specificity	:	0

ICPCR

It is a web application. Developed thus far is ui.r and server.r files .




Distribution of Retained Vs Churn customers Statewise
False && True
0
50
100
150
AK AR CA CT DE GA IA IL KS LA MD MI MO MT ND NH NM NY OK PA SC TN UT VT WI WY
State names


Distribution churn customers suscribers of International Plan
False && True
0
1000
2000
3000
4000
no
yes
International Plan Suscribers