



# Customer Churn Analytics using Microsoft R Open

Malaysia R User Group Meet Up  
16<sup>th</sup> February 2017  
Poo Kuan Hoong

<https://github.com/kuanhoong/churn-r>

Disclaimer: The views and opinions expressed in this slides are those of the author and do not necessarily reflect the official policy or position of Nielsen Malaysia. Examples of analysis performed within this slides are only examples. They should not be utilized in real-world analytic products as they are based only on very limited and dated open source information. Assumptions made within the analysis are not reflective of the position of Nielsen Malaysia.

# Agenda

- Introduction
- Customer Churn Analytics
- Machine Learning Framework
- Microsoft R Open and Visual Studio
- Model Performance Comparison
- Demo

# Malaysia R User Group (MyRUG)

- The Malaysia R User Group (MyRUG) was formed on June 2016.
- It is a diverse group that come together to discuss anything related to the R programming language.
- The main aim of MyRUG is to provide members ranging from beginners to R professionals and experts to share and learn about R programming and gain competency as well as share new ideas or knowledge.

# Malaysia R User Group - (myRUG)

[Home](#)[Members](#)[Sponsors](#)[Photos](#)[Pages](#)[Discussions](#)[More](#)[Group tools](#)[My profile](#)[Change photo](#)

Kuala Lumpur,  
Malaysia

Founded Jun 5, 2016

[About us...](#)

## Welcome to Malaysia R User Group (myRUG)

[+ Schedule a new Meetup](#)[Upcoming](#)[Past](#)[Calendar](#)

There are no upcoming  
Meetups

You can schedule one!

[Schedule a Meetup](#)

## Recent Meetups

Oct 20, 2016 · 7:00 PM

[Rate this Meetup](#)

## What's new



<https://www.meetup.com/MY-RUserGroup/>

Page

Messages


Notifications

Insights

Publishing Tools

Settings

Help



R User Group Malaysia  
@rusergroupmalaysia

Home

About

Photos


Events



Likes

Liked


Following

More







Learn More




224 likes 0 this week  
Andy Low and 18 other friends




226 follows



See Pages Feed  
Posts from Pages you've liked as your Page




Invite friends to like this Page



309 post reach this week

The R User Group Malaysia is a diverse group that come together to discuss anything related to the R programming language.



224 Likes  
Andy Low and 18 other friends like this

<https://www.facebook.com/rusergroupmalaysia/>



# Malaysia R User Group

MyRUG





# Introduction

- **Customer churn** can be defined simply as the rate at which a company is losing its customers
- Imagine the business as a bucket with holes, the water flowing from the top is the growth rate, while the holes at the bottom is **churn**
- While a certain level of churn is unavoidable, it is important to keep it under control, as high churn rate can potentially kill your business









## Data Scientists Quick Apply

PositiveLinks Asia

Kuala Lumpur, Malaysia

 Posted 7 days ago  165 views



2 connections work here



Apply

Save

### Job description

We are looking for data science and analytics candidates with the following experience:


- Experts in Python, SQL and R.
- Experienced in working with large data sets with the aim of developing predictive models.
- Have carried statistical modeling, analytics modeling, customer segmentation and profiling, social network analysis and customer insights.
- Have contributed to the marketing campaign strategically and tactically through the use of various models (descriptive, predictive, optimisation).
- Knowledge of Hadoop and Spark would be beneficial.



## Data Scientist Quick Apply

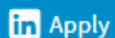
Axiata

Kuala Lumpur, Malaysia

 Posted 21 days ago  1161 views



3 connections work here



Apply

Save

### Job description

#### Responsibilities

- Designs experiments, test hypotheses, and build models.
- Build, maintain, and improve on multiple real-time decision systems.
- Leads discovery processes with stakeholders to identify the business requirements and the expected outcomes.
- Works with and alongside business analysts by suggesting other products of interest to the client.
- Models and frames business scenarios that are meaningful and which impact on critical business processes and/or decisions.
- Identifies what data is available and relevant, including internal and external data sources, leveraging new data collection processes such as smart meters and geo-location information or social media.
- Collaborates with subject matter experts to select the relevant sources of information.
- Makes strategic recommendations on data collection, integration and retention requirements incorporating business requirements and knowledge of best practices.

digi

Data Scientist, BIRS  
Digi Telecommunications Sdn Bhd

### Job Description

#### JOB SUMMARY

This individual will be the expert modeler in the data science and modeling team and have deep knowledge of machine learning, data mining and statistical analyses.

#### KEY RESPONSIBILITIES

- Hands on building models utilizing the various analytical techniques.
- Engage regularly with the campaign mgt team to come up with new and innovative campaigns.
- Develop predictive models (e.g. Churn Prediction model, Next Best offer model, Market Basket analysis) to leverage existing information assets for optimal Marketing activity.
- Develop descriptive models (e.g. behavioural segmentation, lifetime value model, social network analysis) to enhance customer insights.
- Develop optimization models (e.g. network optimization, campaign optimization) to enhance return on investment.
- Capitalize opportunities for revenue enhancement through targeted campaigns by development of:
  - Behavioral analytics, measurement and modelling
  - Customer and Audience segmentation, clustering and profiling
  - Geo/demographic attribution and segmentation

# Churn analytics

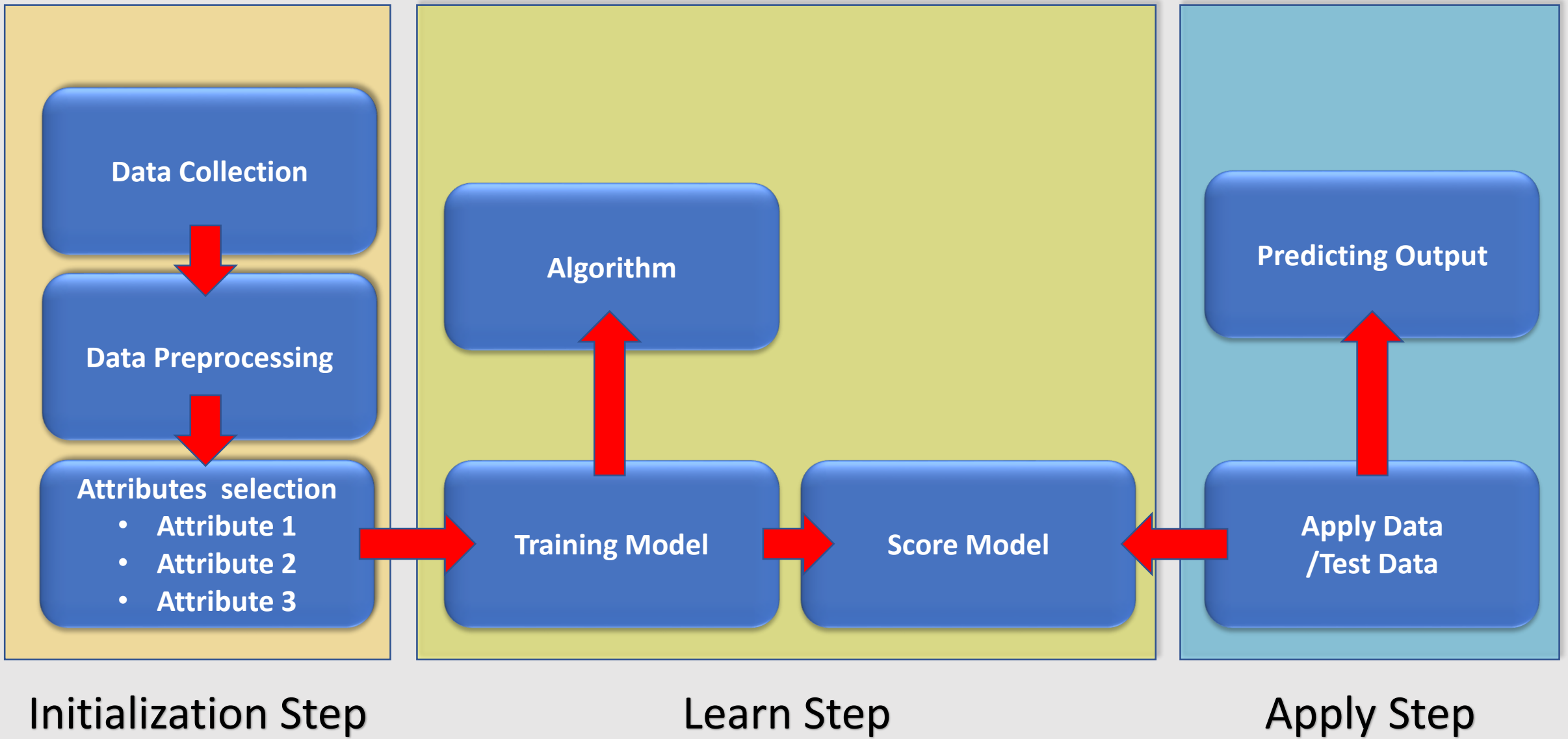
- Predicting who will switch mobile operator

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	customerI	gender	SeniorCiti	Partner	Dependen	tenure	PhoneServ	MultipleLi	InternetSe	OnlineSec	OnlineBac	DevicePro	TechSupp	Streaming	Streaming	Contract	Paperless	PaymentM	MonthlyCl	TotalChar	Churn
2	7590-VHV	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to	Yes	Electronic	29.85	29.85	No
3	5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed ch	56.95	1889.5	No
4	3668-QPY	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to	Yes	Mailed ch	53.85	108.15	Yes
5	7795-CFO	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank trans	42.3	1840.75	No
6	9237-HQI	Female	0	No	No	2	Yes	No	Fiber opti	No	No	No	No	No	No	Month-to	Yes	Electronic	70.7	151.65	Yes
7	9305-CDSI	Female	0	No	No	8	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	Month-to	Yes	Electronic	99.65	820.5	Yes
8	1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber opti	No	Yes	No	No	Yes	No	Month-to	Yes	Credit car	89.1	1949.4	No
9	6713-OKC	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to	No	Mailed ch	29.75	301.9	No
10	7892-POC	Female	0	Yes	No	28	Yes	Yes	Fiber opti	No	No	Yes	Yes	Yes	Yes	Month-to	Yes	Electronic	104.8	3046.05	Yes
11	6388-TAB	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank trans	56.15	3487.95	No
12	9763-GRSI	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to	Yes	Mailed ch	49.95	587.45	No
13	7469-LKBC	Male	0	No	No	16	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Two year	No	Credit car	18.95	326.8	No
14	8091-TTV	Male	0	Yes	No	58	Yes	Yes	Fiber opti	No	No	Yes	No	Yes	Yes	One year	No	Credit car	100.35	5681.1	No
15	0280-XJGE	Male	0	No	No	49	Yes	Yes	Fiber opti	No	Yes	Yes	No	Yes	Yes	Month-to	Yes	Bank trans	103.7	5036.3	Yes
16	5129-JLPI	Male	0	No	No	25	Yes	No	Fiber opti	Yes	No	Yes	Yes	Yes	Yes	Month-to	Yes	Electronic	105.5	2686.05	No
17	3655-SNQ	Female	0	Yes	Yes	69	Yes	Yes	Fiber opti	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit car	113.25	7895.15	No
18	8191-XWS	Female	0	No	No	52	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Mailed ch	20.65	1022.95	No
19	9959-WOI	Male	0	No	Yes	71	Yes	Yes	Fiber opti	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank trans	106.7	7382.25	No
20	4190-MFL	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to	No	Credit car	55.2	528.35	Yes

# Customer churn - who do customers change operators?

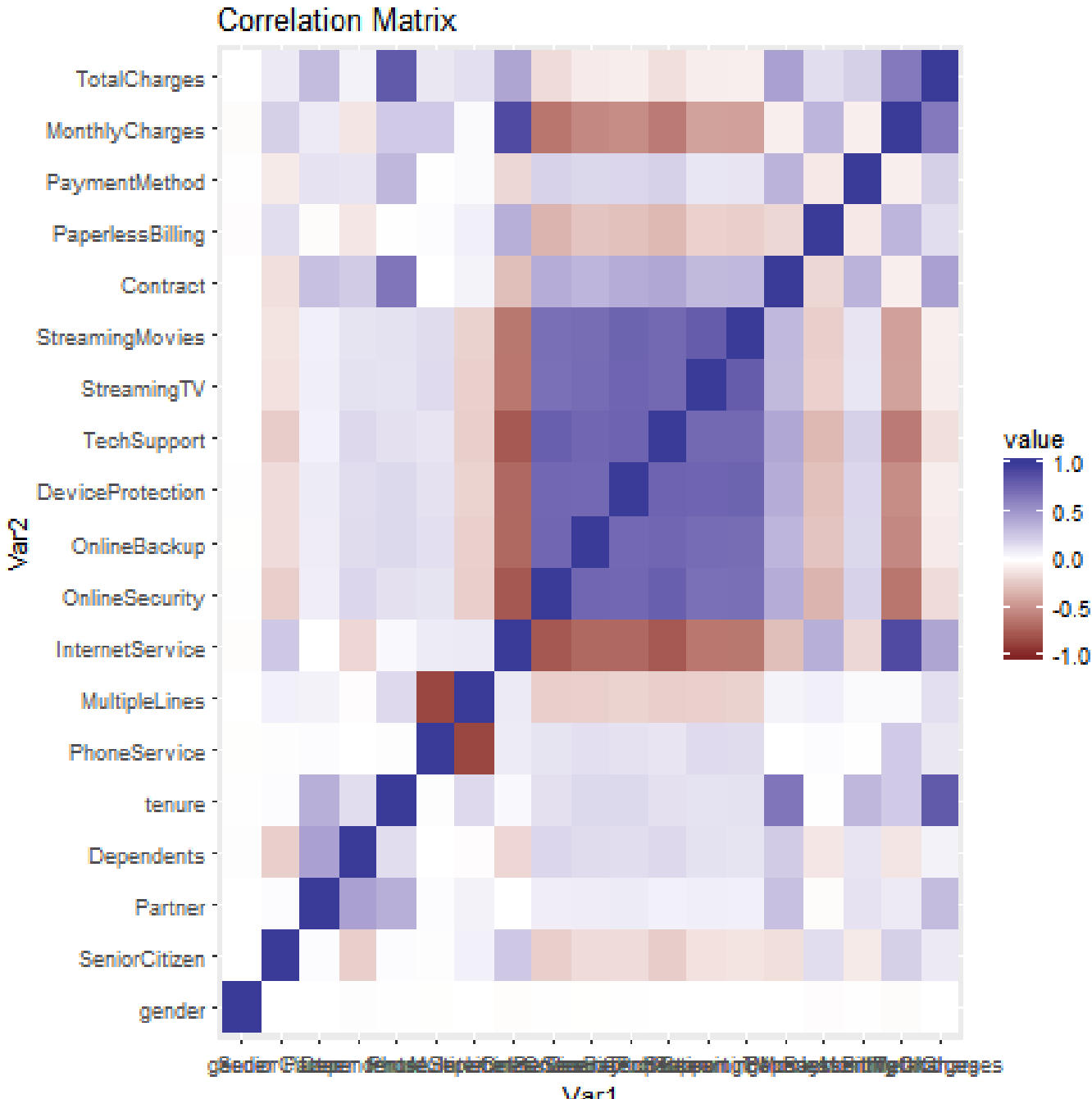
- The top 3 reasons why subscribers change providers:
  - They want a new handset
  - They believe they pay too much for calls/data
  - Providers do not offer additional loyalty benefits





# Machine Learning Framework

- **correlation matrix**, which is used to investigate the dependence between multiple variables at the same time.



# Microsoft R Open

- [Microsoft R Open](#), formerly known as Revolution R Open (RRO), is the enhanced distribution of R from Microsoft Corporation.
- It is a complete open source platform for statistical analysis and data science.

## Key enhancement

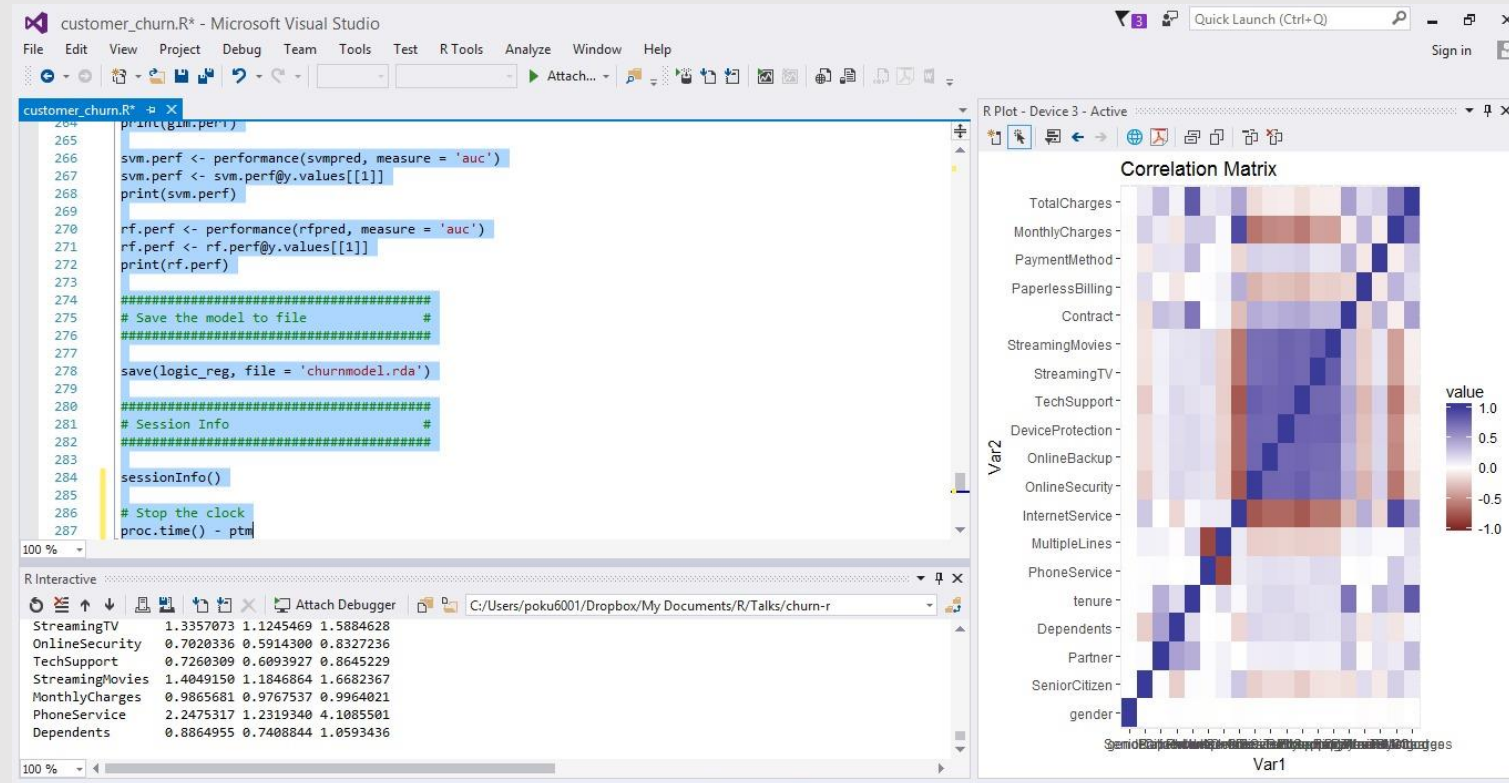
- Multi-threaded math libraries that brings multi-threaded computations to R.
- A high-performance default CRAN repository that provide a consistent and static set of packages to all Microsoft R Open users.
- The checkpoint package that make it easy to share R code and replicate results using specific R package versions.





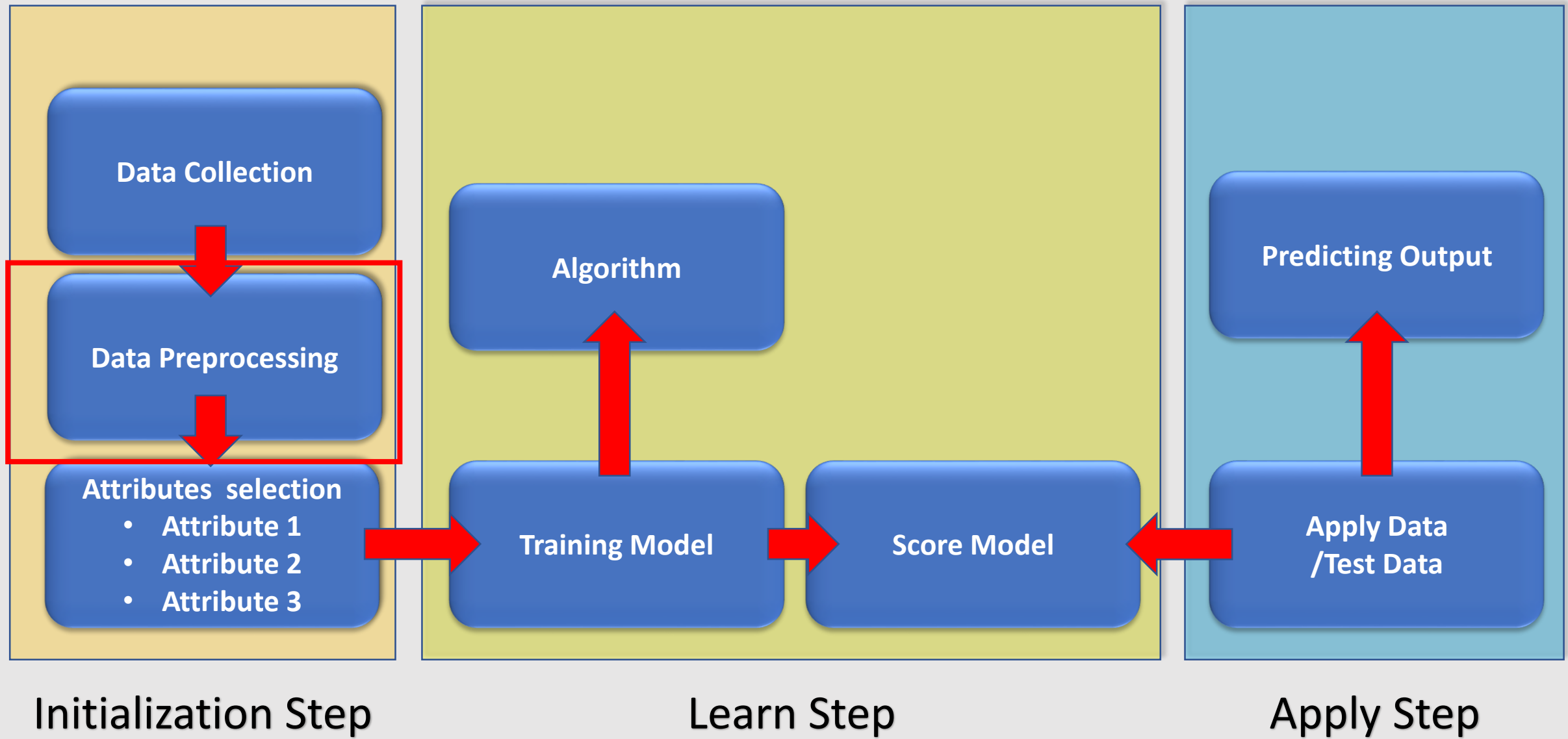
# R Tools for Visual Studio

- Turn Visual Studio into a powerful R development environment
- [Download R Tools for Visual Studio](#)



# R Tools for Visual Studio

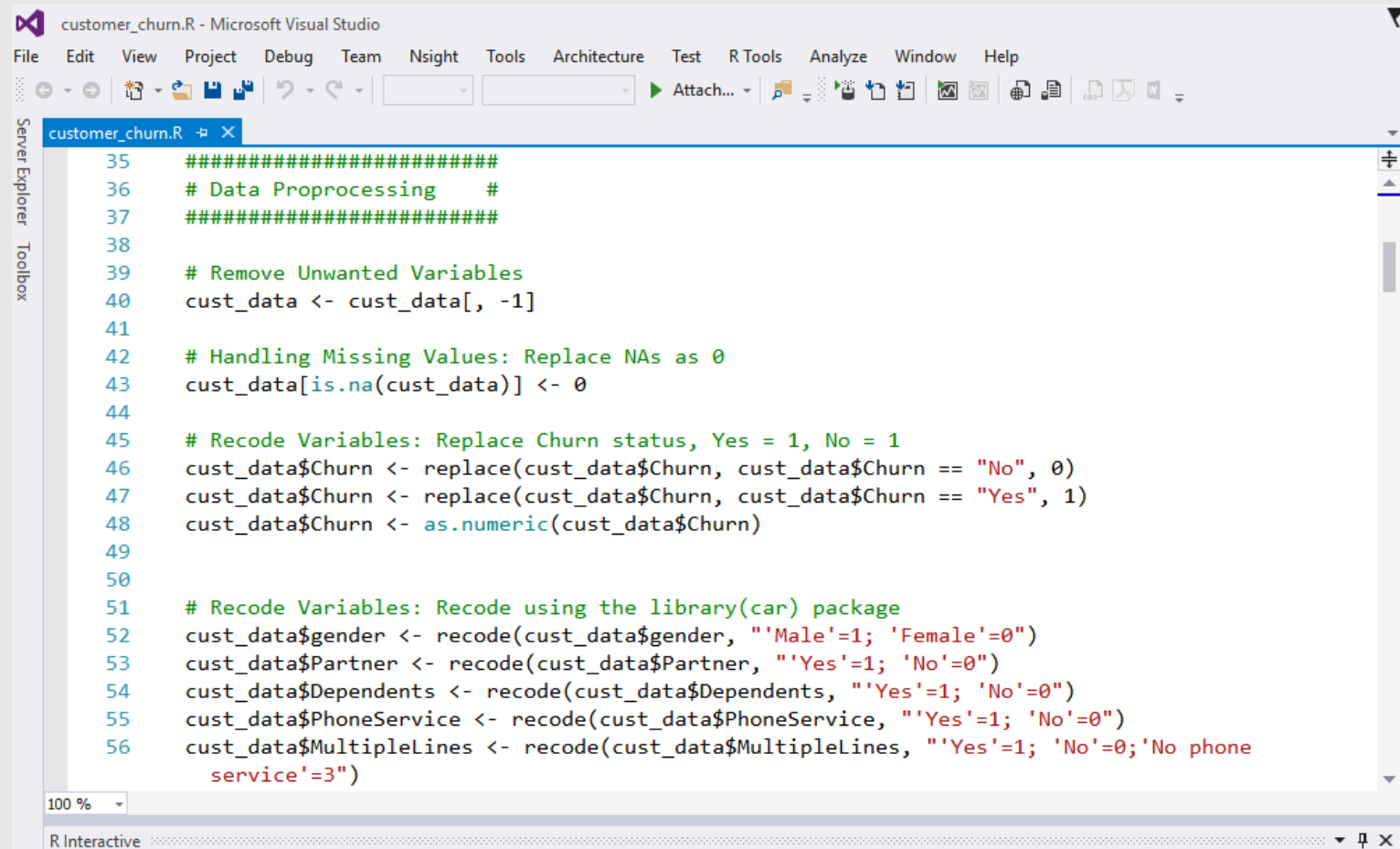
- Visual Studio IDE
- Intellisense
- Enhanced multi-threaded math libs, cluster scale computing, and a high performance CRAN repo with checkpoint capabilities.
- Learn more about R Tools from here:  
<https://microsoft.github.io/RTVS-docs/>



# Machine Learning Framework

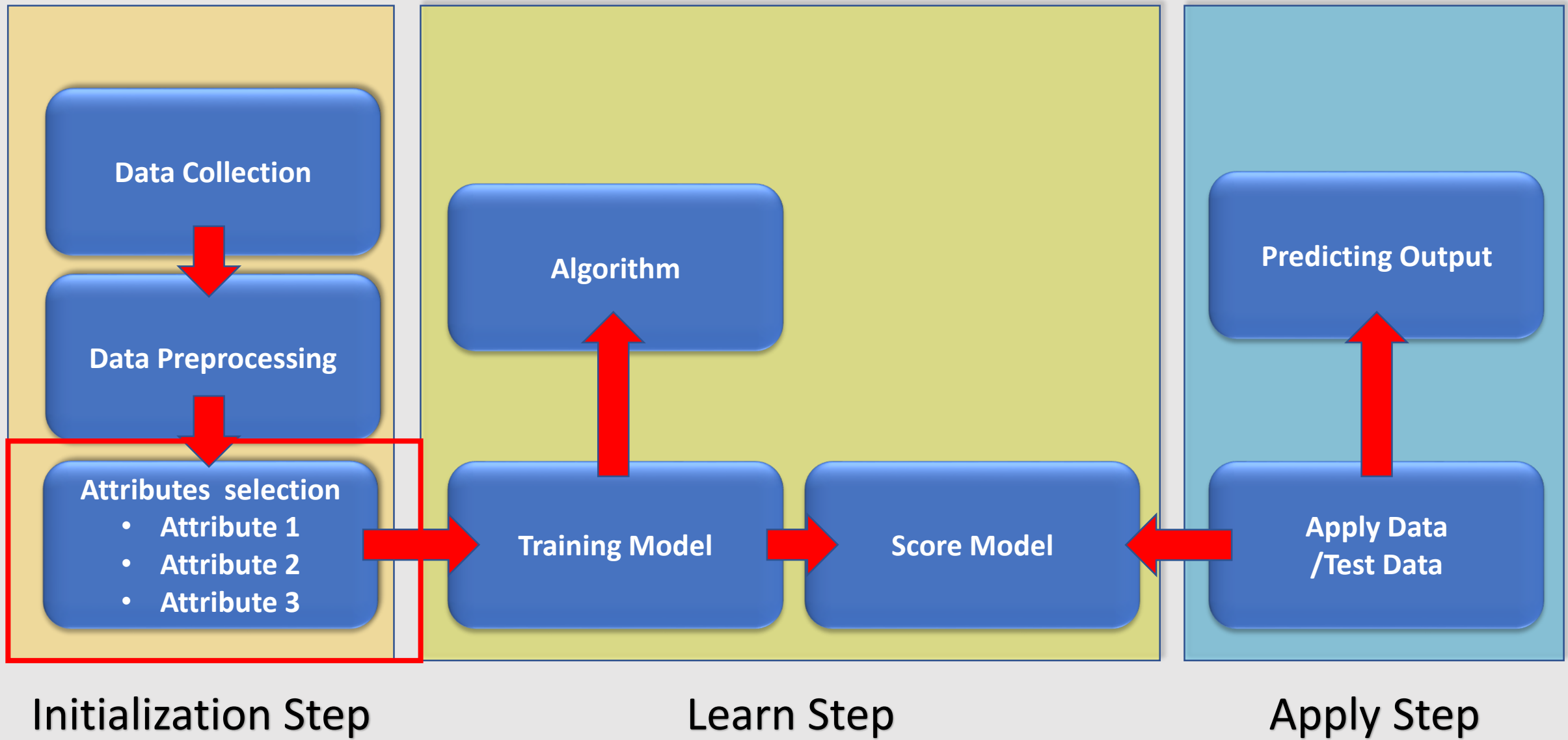
# Data Preprocessing

- Assign missing values as zero
- Detect outliers
- Remove unwanted variables
- Recode variables



The screenshot shows the Microsoft Visual Studio interface with a file named 'customer\_churn.R' open. The code is written in R and focuses on data preprocessing. It includes comments for each step: removing unwanted variables, handling missing values, and recoding variables. The 'car' package is used for recoding. The code is as follows:

```
35 #####
36 # Data Preprocessing #
37 #####
38
39 # Remove Unwanted Variables
40 cust_data <- cust_data[, -1]
41
42 # Handling Missing Values: Replace NAs as 0
43 cust_data[is.na(cust_data)] <- 0
44
45 # Recode Variables: Replace Churn status, Yes = 1, No = 0
46 cust_data$Churn <- replace(cust_data$Churn, cust_data$Churn == "No", 0)
47 cust_data$Churn <- replace(cust_data$Churn, cust_data$Churn == "Yes", 1)
48 cust_data$Churn <- as.numeric(cust_data$Churn)
49
50
51 # Recode Variables: Recode using the library(car) package
52 cust_data$gender <- recode(cust_data$gender, "'Male'=1; 'Female'=0")
53 cust_data$Partner <- recode(cust_data$Partner, "'Yes'=1; 'No'=0")
54 cust_data$Dependents <- recode(cust_data$Dependents, "'Yes'=1; 'No'=0")
55 cust_data$PhoneService <- recode(cust_data$PhoneService, "'Yes'=1; 'No'=0")
56 cust_data$MultipleLines <- recode(cust_data$MultipleLines, "'Yes'=1; 'No'=0; 'No phone
    service'=3")
```



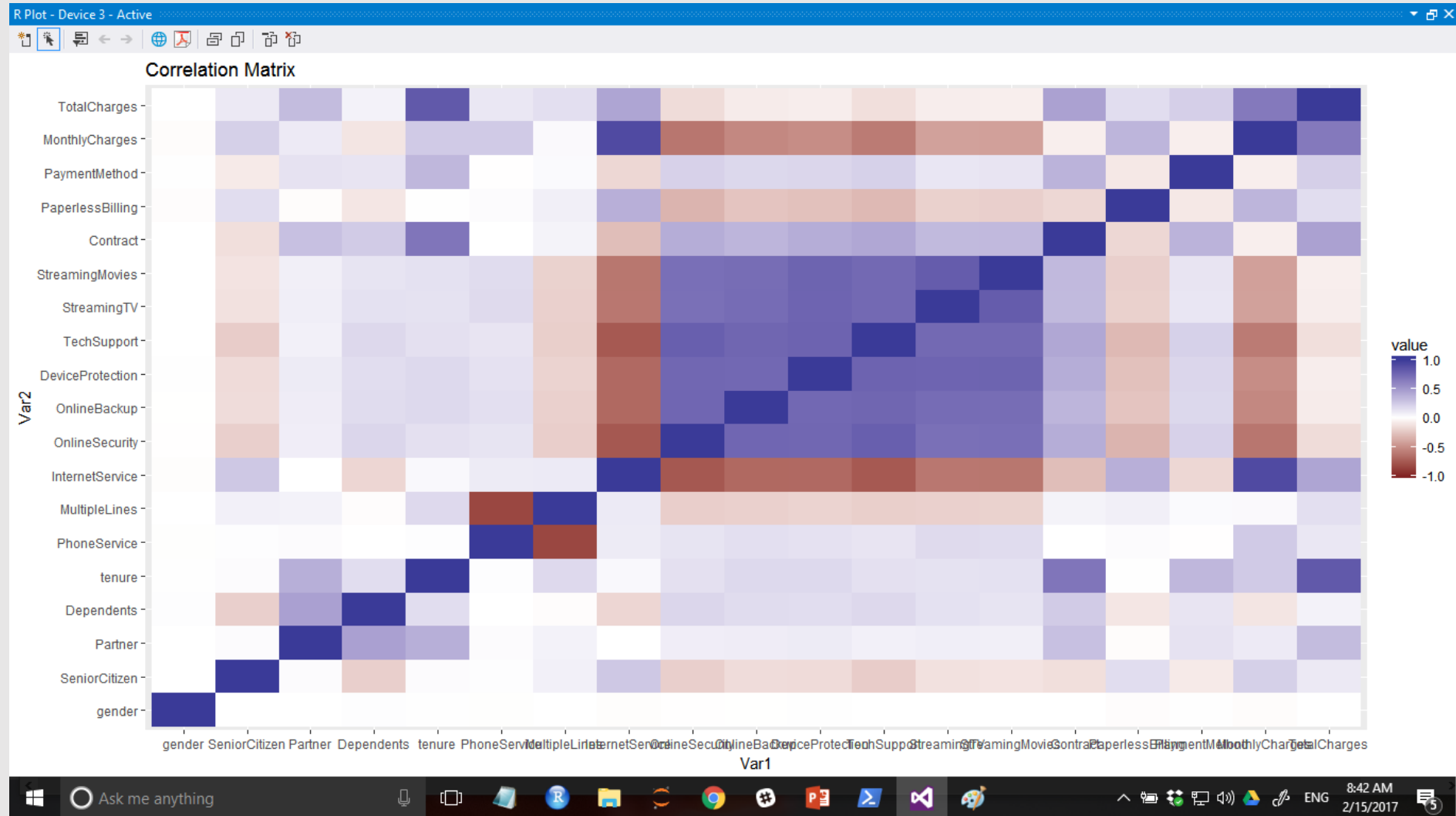
# Machine Learning Framework

# Features selection

- The process of selecting a subset of relevant features (variables, predictors) for use in model construction.
- Feature selection techniques are used for three reasons:
  - simplification of models to make them easier to interpret by researchers/users,
  - shorter training times,
  - enhanced generalization by reducing overfitting

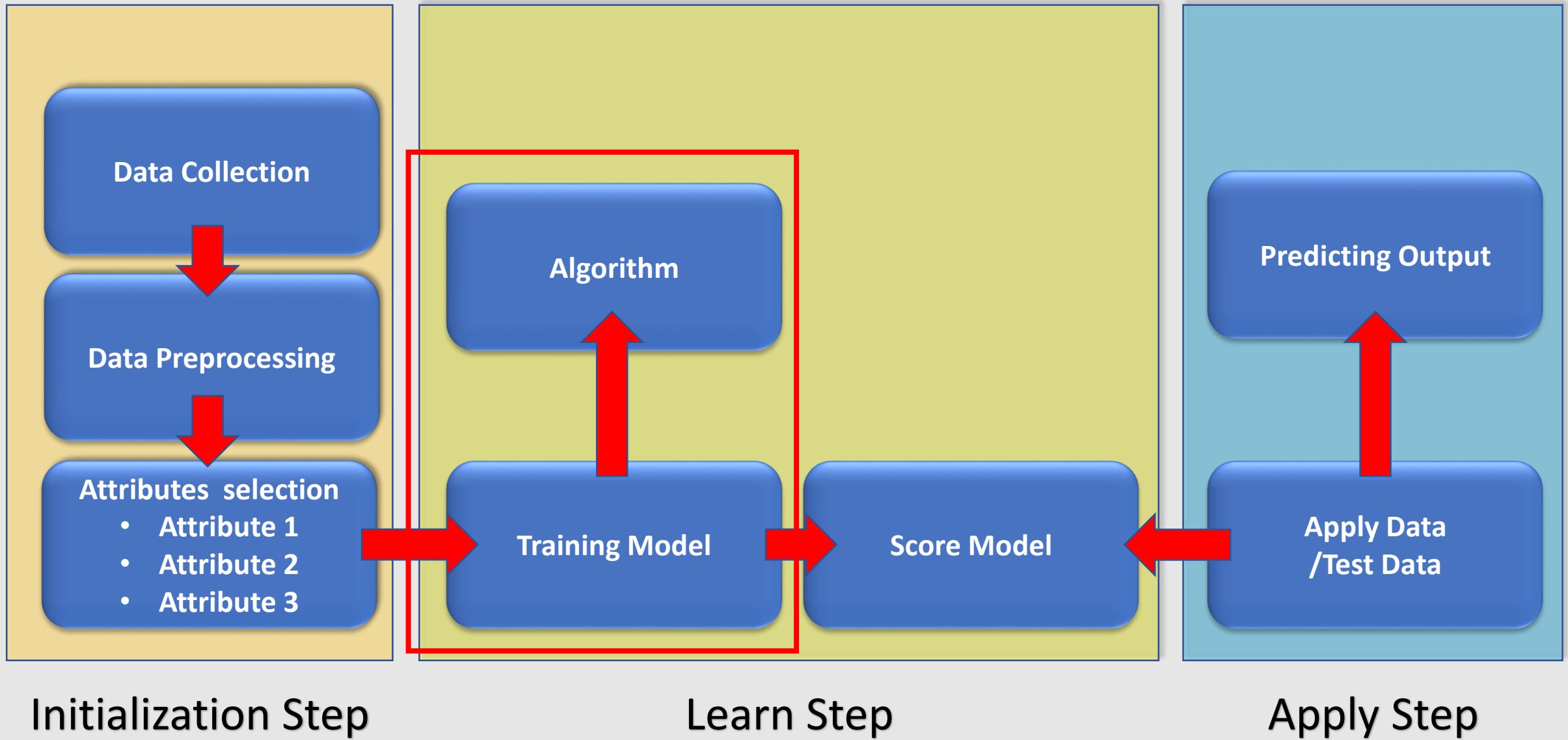


# Correlation Matrix



# Models Performance Comparison

- Logistic Regression
  - is a regression model where the dependent variable (DV) is categorical.
- Support Vector Machine
  - SVM is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis.
- RandomForest
  - is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

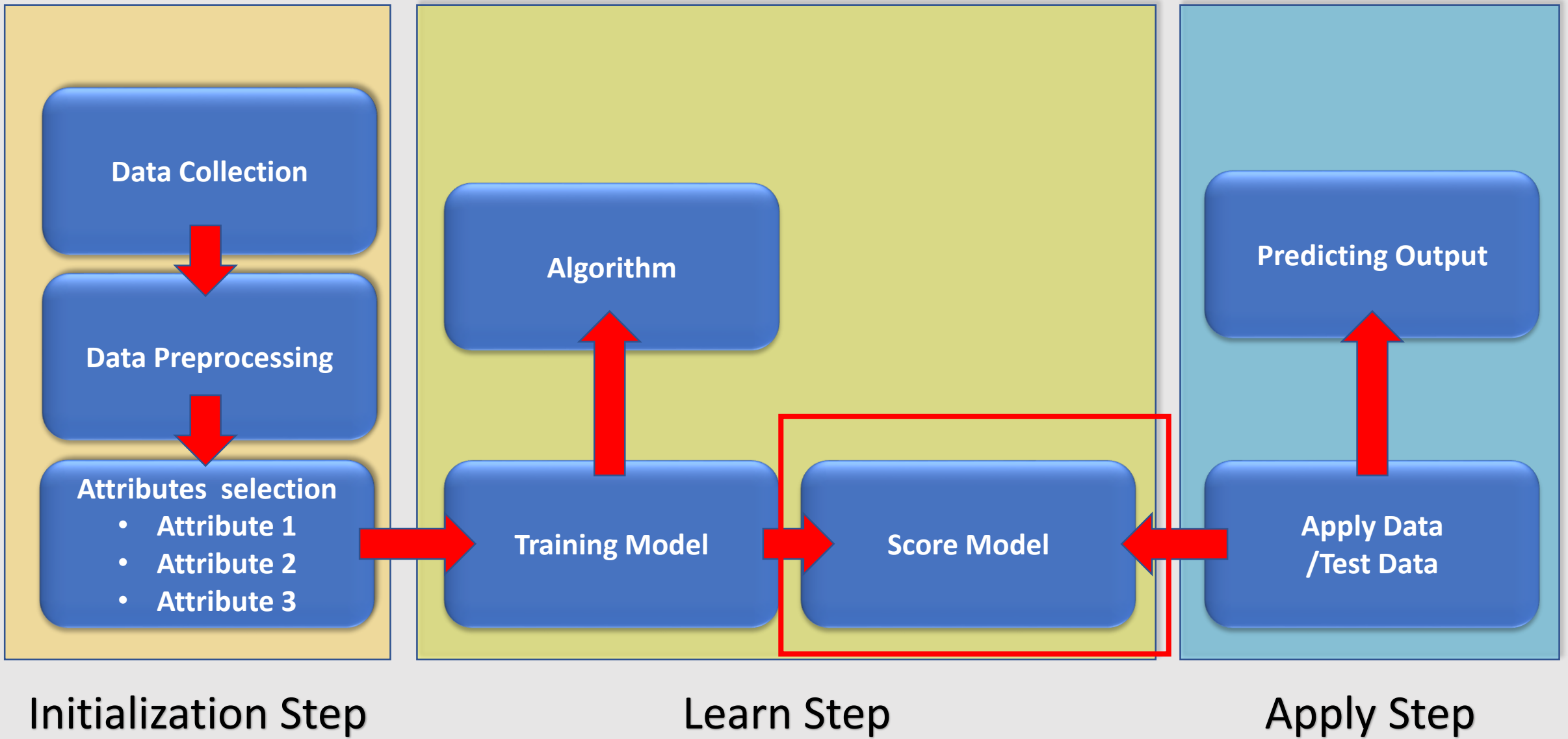


# Machine Learning Framework

# Training Model and Algorithm

- Split the data set into 80:20 using library(caret)
- Apply the algorithms: GLM, SVM and RF

```
customer_churn.R* X
87
88 #####
89 # Model Building #
90 #####
91
92 # For training and testing purpose,
93 # split the data to 80-20
94
95 library(caret)
96 set.seed(1234)
97 intrain <- createDataPartition(y = cust_data$Churn, p = 0.8, list = FALSE, times = 1)
98 training <- cust_data[intrain,]
99 testing <- cust_data[- intrain,]
100
101 #####
102 # Model 1: Logistic Regression Model #
103 #####
104
105 # Select the features to be used based on forward selection procedure
106 # Akaike information criterion (AIC = 2k - 2 log L) as the choice of
107 # metric. Lower AIC indicates better model
108
109 fullMod = glm(Churn ~ ., data = training, family = binomial)
110 summary(fullMod)
111 intMod <- glm(Churn ~ 1, data = training, family = binomial)
112 summary(intMod)
113 fwdSelection = stepAIC(intMod, scope = list(lower = formula(intMod), upper = formula(fullMod)),
```



# Machine Learning Framework

# Score Model

- **Confusion Matrix:** a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.
  - **true positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
  - **true negatives (TN):** We predicted no, and they don't have the disease.
  - **false positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
  - **false negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")



# Confusion Matrix: Generalized Linear Model (glm)

n=1407	Predicted: NO	Predicted: YES	
Actual: NO	TN = 929 (0.660)	FP = 105 (0.075)	1034
Actual: YES	FN = 168 (0.119)	TP = 205 (0.146)	373
	1097	310	

# Confusion Matrix: Support Vector Machine (SVM)

n=1407	Predicted: NO	Predicted: YES	
Actual: NO	TN= 878 (0.624)	FP= 156 (0.111)	1034
Actual: YES	FN= 180 (0.128)	TP= 193 (0.137)	373
	1058	349	

# Confusion Matrix: RandomForest

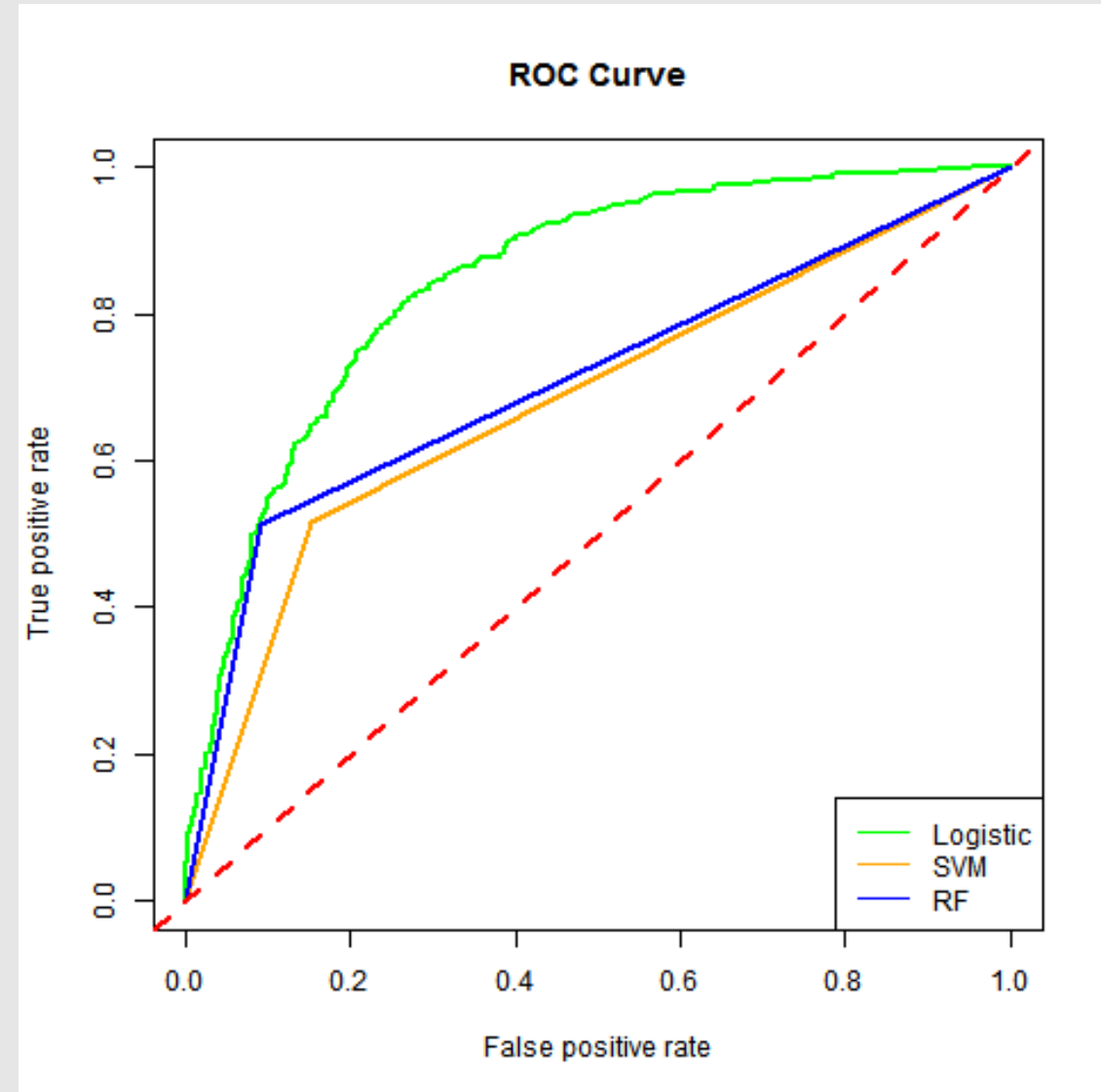
n=1407	Predicted: NO	Predicted: YES	
Actual: NO	TN= 940 (0.668)	FP= 94 (0.067)	1034
Actual: YES	FN= 181 (0.129)	TP= 192 (0.136)	373
	1121	286	

# Receiver Operating Characteristic (ROC) curve

- ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

# Models comparison

- ROC illustrates the performance of a binary classifier system as its discrimination threshold is varied.



# Microsoft R Open vs R

```
[17] parallel_3.3.1      SparseM_1.74      RGtk2_2.20.31     stringr_1.1.0
[25] grid_3.3.1          nnet_7.3-12       survival_2.40-1    gdata_2.1.1
[33] scales_0.4.1         codetools_0.2-14  ModelMetrics_1.1.0 MASS_7.3-45
[41] labeling_0.3         quantreg_5.29     KernSmooth_2.23-15 stringi_1.1.2
>
> # Stop the clock
> proc.time() - ptm
  user  system elapsed
750.81    4.59   758.04
> |
```

R

```
loaded via a namespace (and not attached):
 [1] Rcpp_0.12.9      nloptr_1.0.4      plyr_1.8.4        class_7.3-16
 [9] partykit_1.1-1   lme4_1.1-12       tibble_1.2.1      nlme_3.1-142
[17] parallel_3.3.2   SparseM_1.74      RGtk2_2.20.31     stringr_1.1.0
[25] stats4_3.3.2     grid_3.3.2        nnet_7.3-12       survival_2.40-1
[33] magrittr_1.5     scales_0.4.1      codetools_0.2-15  ModelMetrics_1.1.0
[41] colorspace_1.3-2 labeling_0.3       quantreg_5.29     KernSmooth_2.23-15
>
> # Stop the clock
> proc.time() - ptm
  user  system elapsed
759.04    5.31   750.22
> |
```

Microsoft R Open

100 %

R Interactive

Attach Debugger C:/Users/poku6001/Dropbox/My Documents/

```
[19] RGtk2_2.20.31      stringr_1.1.0     caTools_1.17.1    gtools_3.5.0
[25] rtvs_1.0.0.0       stats4_3.3.2      grid_3.3.2        nnet_7.3-12
[31] minqa_1.2.4        Formula_1.2-1     reshape2_1.4.2    magrittr_1.5
[37] ModelMetrics_1.1.0 MASS_7.3-45       splines_3.3.2     assertthat_0.1
[43] labeling_0.3       quantreg_5.29     KernSmooth_2.23-15 stringi_1.1.2
  user  system elapsed
769.02    3.16   770.47
> |
```

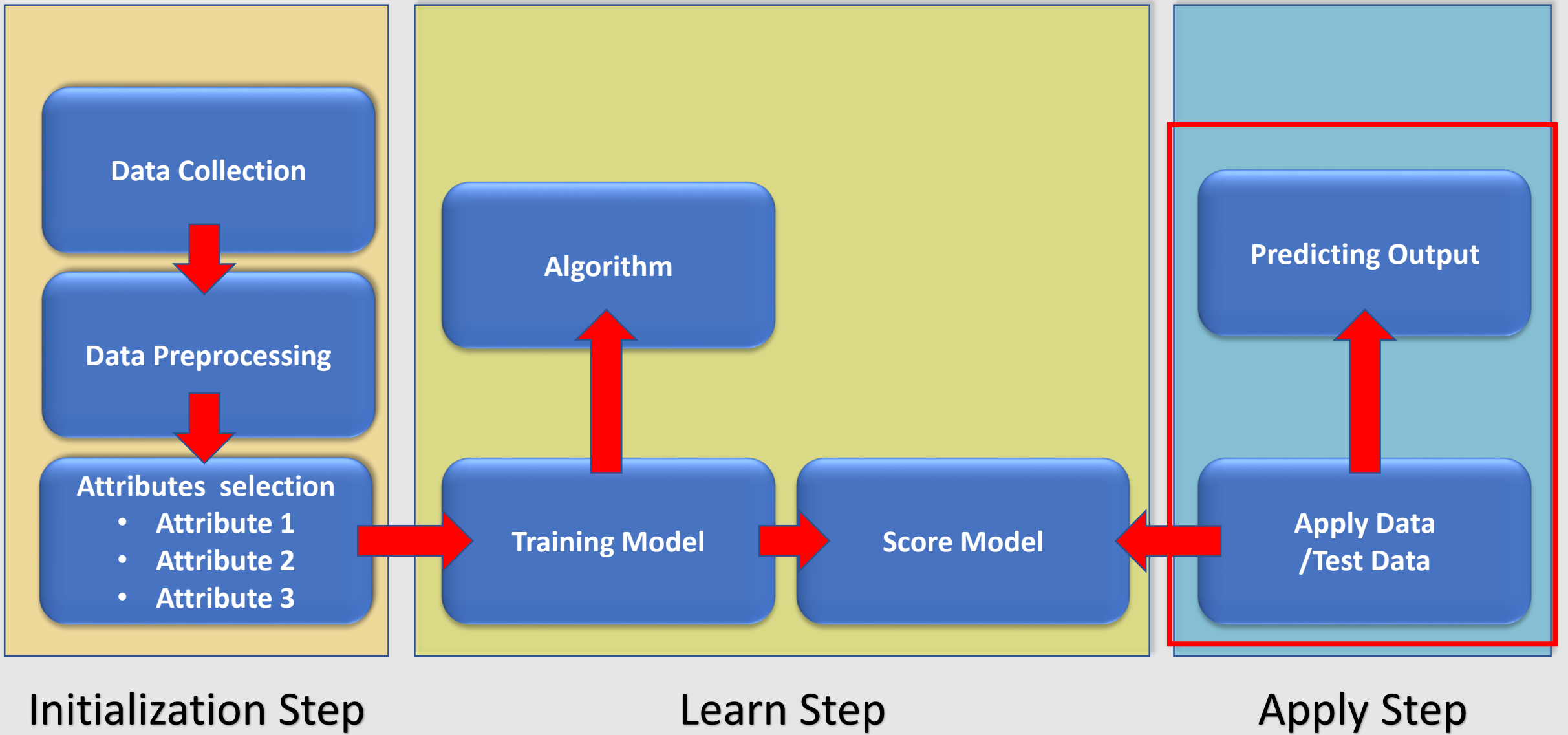
VS 2015

Console C:/Users/poku6001/Dropbox/My Documents/R/talks/chum-r/

```
[37] MASS_7.3-45       splines_3.3.2     assertthat_0.1    pkr
[41] colorspace_1.3-2  labeling_0.3       quantreg_5.29     Kern
[45] stringi_1.1.2     lazyeval_0.2.0    munsell_0.4.3
>
> # Stop the clock
> proc.time() - ptm
  user  system elapsed
783.30    3.77   783.80
> |
```

RStudio





# Machine Learning Framework

# Predict test data

- Based on the training model, select the best model to be used for test data prediction

```
load_model.R  X customer_churn.R*
44 #####
45 # load model #
46 #####
47 load('churnmodel.rda')
48
49 #logic_reg <- glm(Churn ~ Contract
50 #               + InternetService
51 #               + tenure
52 #               + PaperlessBilling
53 #               + TotalCharges
54 #               + MultipleLines
55 #               + PaymentMethod
56 #               + SeniorCitizen
57 #               + StreamingTV
58 #               + OnlineSecurity
59 #               + TechSupport
60 #               + StreamingMovies
61 #               + MonthlyCharges
62 #               + PhoneService
63 #               + Dependents
64
65 #for glm, requires to select the same variables used in training
66 cust_data1 <- cust_data1[,c
67   ("Contract", "InternetService", "tenure", "PaperlessBilling", "TotalCharges", "MultipleLines", "PaymentMethod", "SeniorCitizen", "StreamingTV", "OnlineSecurity", "TechSupport", "StreamingMovies", "MonthlyCharges", "PhoneService", "Dependents", "Churn")]
67
```

100 %

# Questions?



**kuanhoong@gmail.com**

DEMO