DATA SCIENCE PROJECT

upx
Move up in life

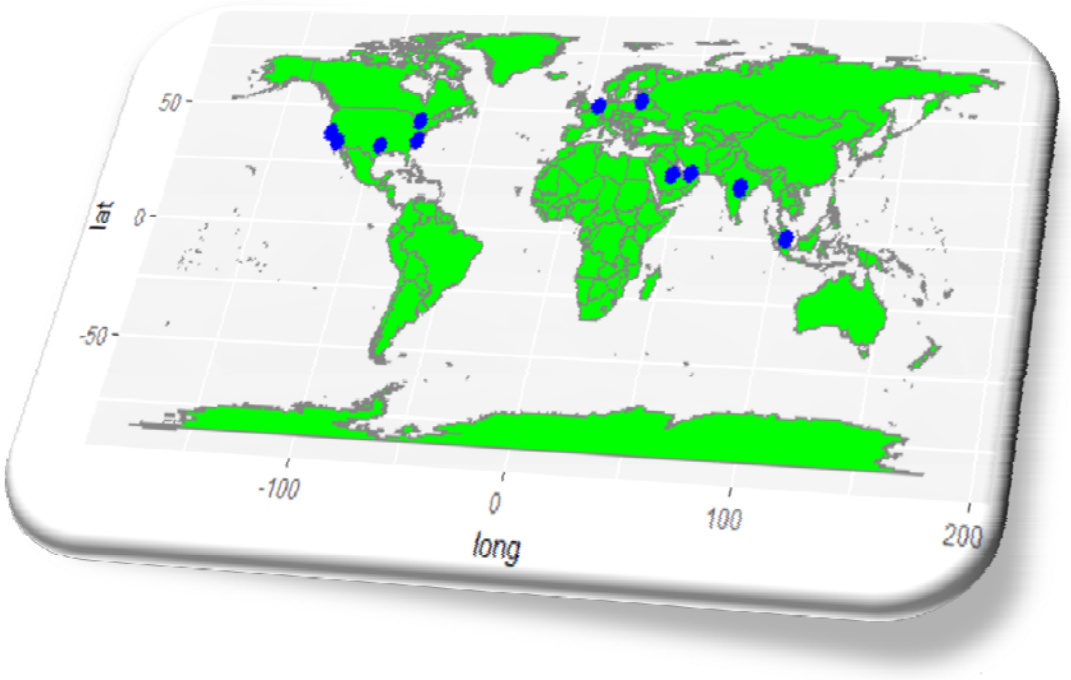
Data Diggers
Let's Uncomplicate Business

MACHINE LEARNING

Telecom Churn Analysis

Team Info

- *Ali Qaiser Syed*
- *Annada Prasad*
- *Chia Aik Lee*
- *Diego Jaramillo*
- *Kiran Devathi*
- *Manoj Kalamkar*
- *Praveen Parvataneni*
- *Subrata Roy*
- *Ravi Bodkai*
- *Vasudev Pendyala*



Domain Telecom

Topic Telecom Churn Analysis

Churn (loss of customers to competition) is a problem for telecom companies because it is expensive to acquire a new customer and companies want to retain their existing customers. Most telecom companies suffer from voluntary churn. Churn rate has strong impact on the life time value of the customer because it affects the length of service and the future revenue of the company. For example if a company has 25% churn rate then, the average customer lifetime is 4 years; similarly a company with a churn rate of 50%, has an average customer lifetime of 2 years.

In the targeted approach the company tries to identify in advance customers who are likely to churn. The company then targets those customers with special programs or incentives. This approach can bring in huge loss for a company, if churn predictions are inaccurate, because then firms are wasting incentive money on customers who would have stayed anyway.

Introduction

Project Objective

- To predict Customer Churn.
- Highlighting the main variables/factors influencing Customer Churn.
- Use various ML algorithms to build prediction models, evaluate the accuracy and performance of these models.
- Finding out the best model for our business case & providing executive Summary.

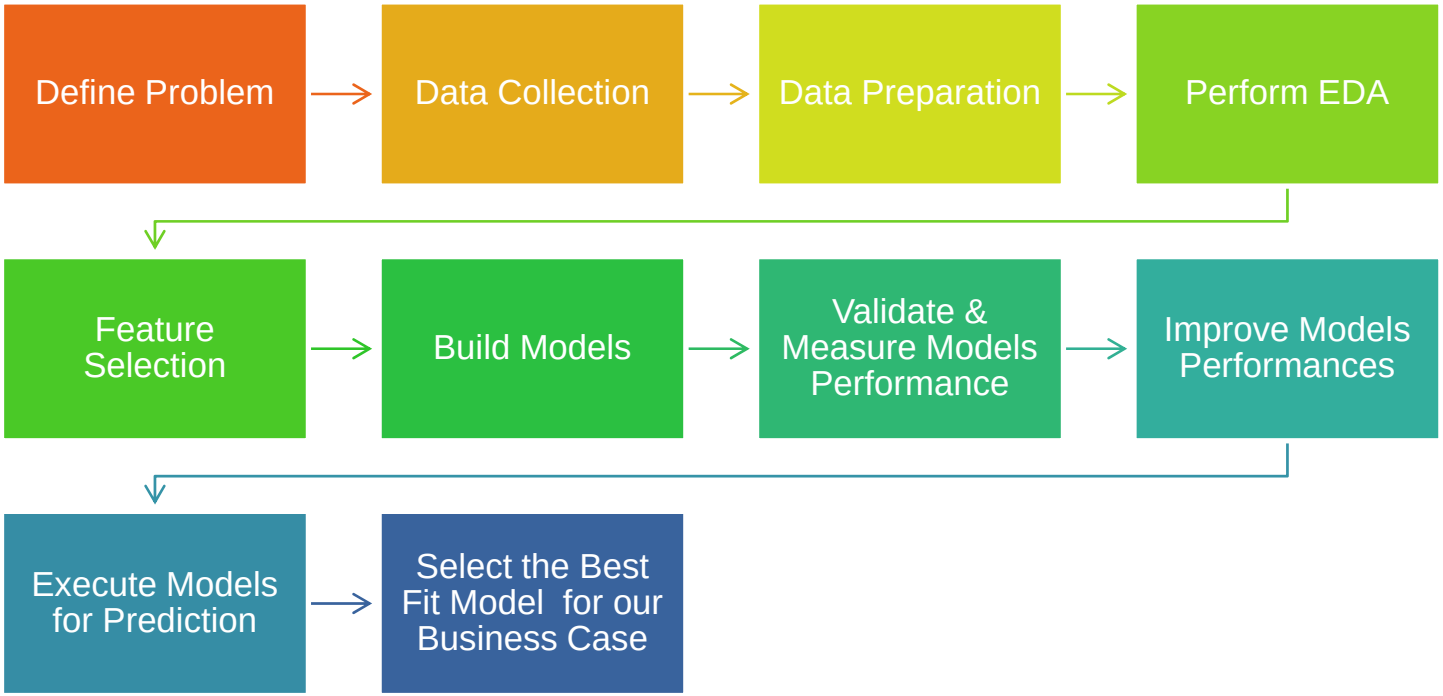
Dataset Description

- Source provided by Upx Academy for data science machine learning project evaluation
- Source dataset is in txt format with csv.
- Dataset contains 4617 rows and 21 columns
- There is no missing values for the provided input dataset.
- Churn_status is the variable which notifies whether a particular customer is churned or not. And we will be developing our models to predict this variable only.

Column Names	
State	factor
Account_Len	integer
Area	integer
Ph_No.	factor
Int_Plan	factor
Vmail_Plan	factor
messgs	integer
tot_day_mins	number
tot_day_calls	integer
tot_day_chrgs	number
tot_evening_mins	number

Column Names	
tot_evening_calls	integer
tot_evening_chrgs	number
tot_ngt_mins	number
tot_ngt_calls	integer
tot_ngt_chrgs	number
tot_int_mins	number
tot_int_calls	integer
tot_int_chrgs	number
cust_calls_made	Integer
churn_status	factor

Model Building Steps



Steps Description

Define Problem

In this project, we will be addressing Customer churn problem for a fictitious US Telecom company. UpX academy has provided us the usage pattern for 4617 customers for over a period of time along with the info whether particular customer is churned or not.
Utilizing the input data, we have to build a model which can predict the customer going to be churned well in advance.

Data Collection

The input data has been provided by UpX academy in txt format with comma separated values. Read in the data into R workbook by appropriate methods ensuring without loosing the data. Here for this loading the data into R we have used read.table command with sep = ‘,’ . As columns names were not part of input dataset and we have provided them separately, hence we have provided the col.names as well to read.table command.

Steps Description

Data Preparation

Perform general housekeeping activities such as checking for the completeness of the content, look at the dimensions & review the structure of input dataset, peek into the data, summarize the data, and get a snapshot of all the features. Check if there is any missing data for particular customers, luckily for our case there is no missing data.

As a final step, transform the data into a format ready to apply with the algorithms and modeling.

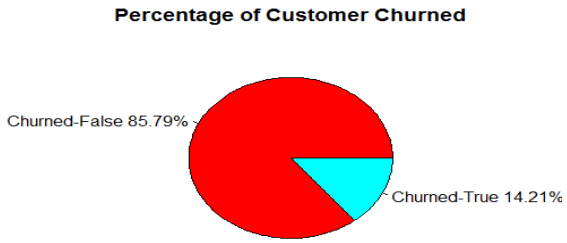
Perform EDA

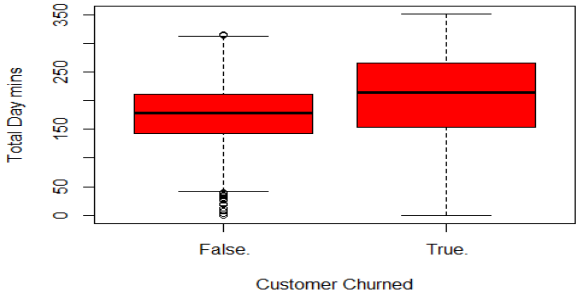
Perform EDA (exploratory data analysis) to understand the data and its applicability on the problem. Understand how the data and its features are interrelated & correlated, evaluate presence of outliers and its effects. Here we will use various box plots and bar plots to understand what are the features which majorly impact on our outcome variable ‘churn_status’. We will see in our next slides the various outcomes from our EDA.

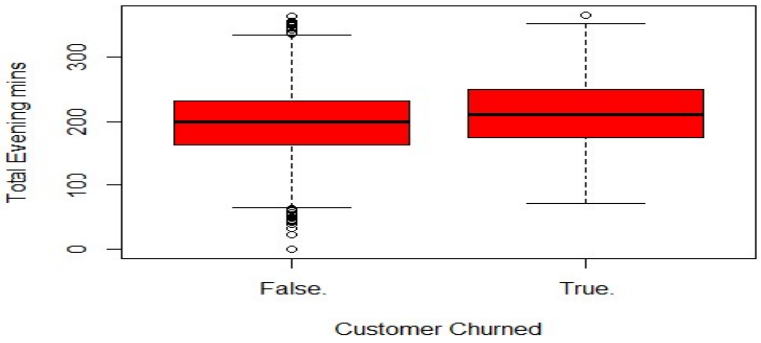
Steps Description

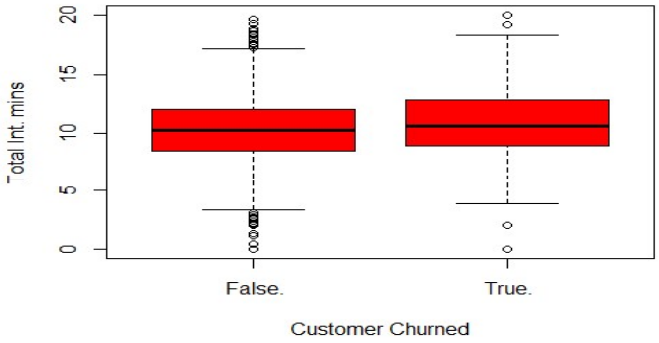
contd..

Perform EDA





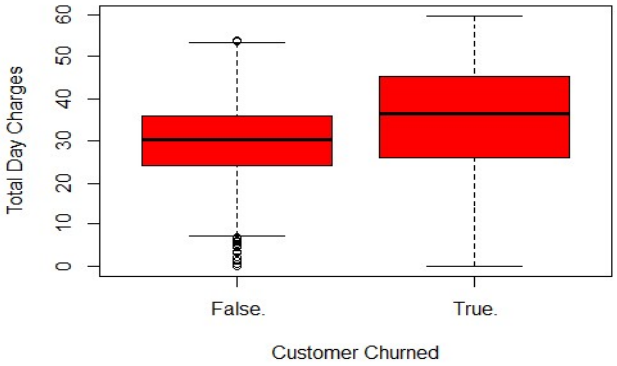


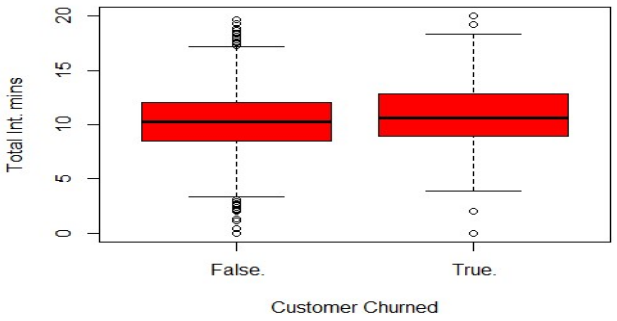


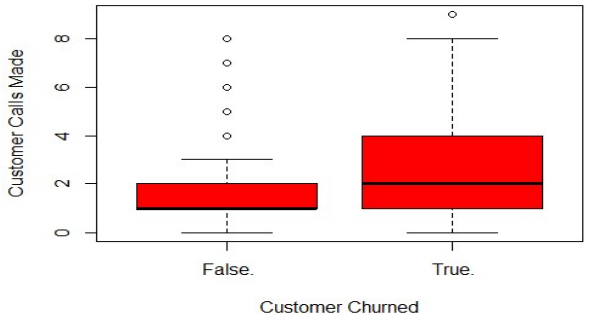
Steps Description

contd..

Perform EDA



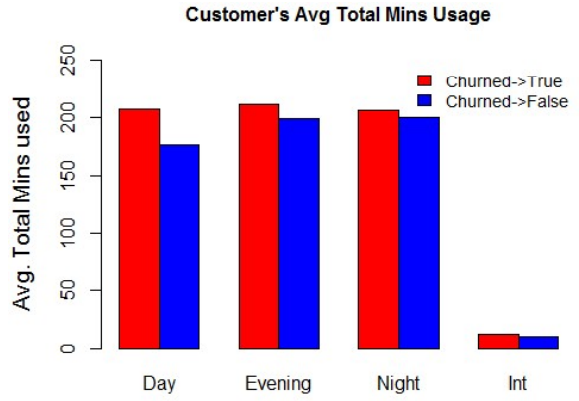


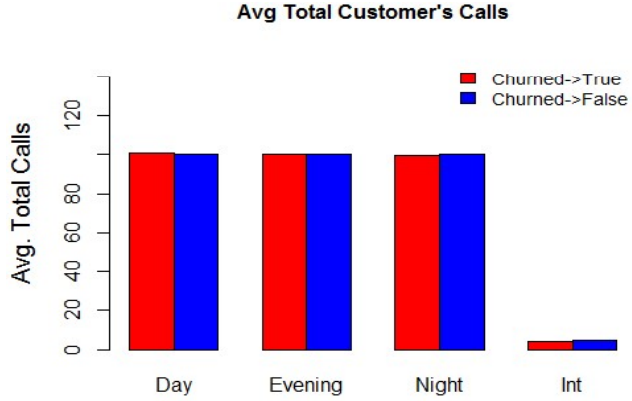


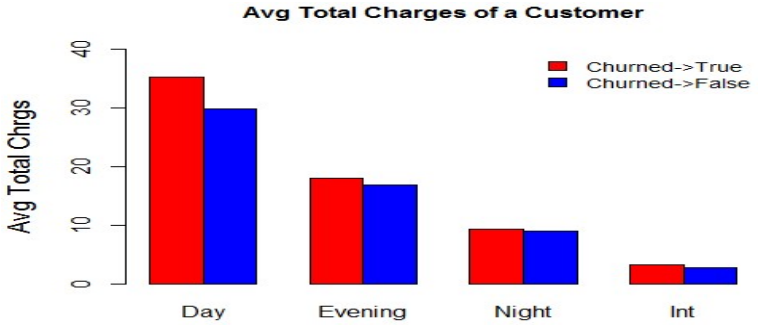
Steps Description

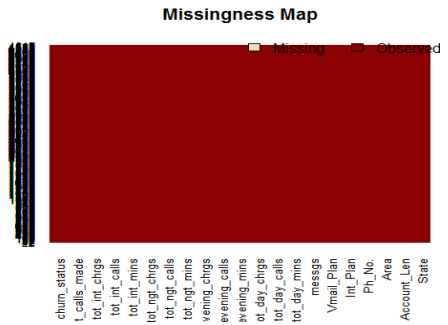
contd..

Perform EDA









Steps Description

Perform EDA

Survival Analysis (Cox Hazard Method)

There are several Survival analysis methods which are applied to examine the time it takes for particular event to occur. For our business problem, we can make a case like, the time it takes for a existing customer to churn. As we have one of the input variable as Account_Length (in number of days) which denotes period of stay for a particular customer before he/she churned, our input data fits right for applying Survival Analysis.

Create a Survival Object & Cox Model:-

```
train_survival$survival <- Surv(train_survival$Account_Len, train_survival$churn_status == 2)

results <- coxph(survival ~ Int_Plan + Vmail_Plan + messgs + tot_day_calls + tot_day_chrgs  +
tot_day_mins +
                + tot_int_calls + tot_int_chrgs + tot_ngt_chrgs
                + tot_int_mins + cust_calls_made, data = train_survival)
```

Steps Description

Perform EDA

Survival Analysis (Cox Hazard Method)

Now lets see what does the summary stats from Cox Model convey us :-

	coef	exp(coef)	se(coef)	z	p
Int_Plan yes	1.29e+00	**3.64e+00**	1.12e-01	**11.53**	<2e-16
Vmail_Plan yes	-1.74e+00	1.75e-01	5.54e-01	-3.14	0.0017
messgs	3.38e-02	1.03e+00	1.68e-02	2.02	0.0438
tot_day_calls	1.14e-03	1.00e+00	2.50e-03	0.46	0.6480
tot_day_chrgs	8.29e+00	3.97e+03	1.73e+01	0.48	0.6314
tot_day_mins	-1.40e+00	2.47e-01	2.94e+00	-0.48	0.6336
tot_int_calls	-6.52e-02	9.37e-01	2.26e-02	-2.89	0.0039
tot_int_chrgs	-1.46e+01	4.47e-07	1.79e+01	-0.82	0.4131
tot_ngt_chrgs	6.75e-02	1.07e+00	2.17e-02	3.11	0.0019
tot_int_mins	4.03e+00	5.64e+01	4.82e+00	0.84	0.4032
cust_calls_made	3.07e-01	**1.36e+00**	3.16e-02	**9.72**	<2e-16

Note:- From the summary we can see that the most statistically significant coef are Int_Plan_yes & cust_calls_made. Have also checked the Proportional Hazards for these.(check R code).

Steps Description

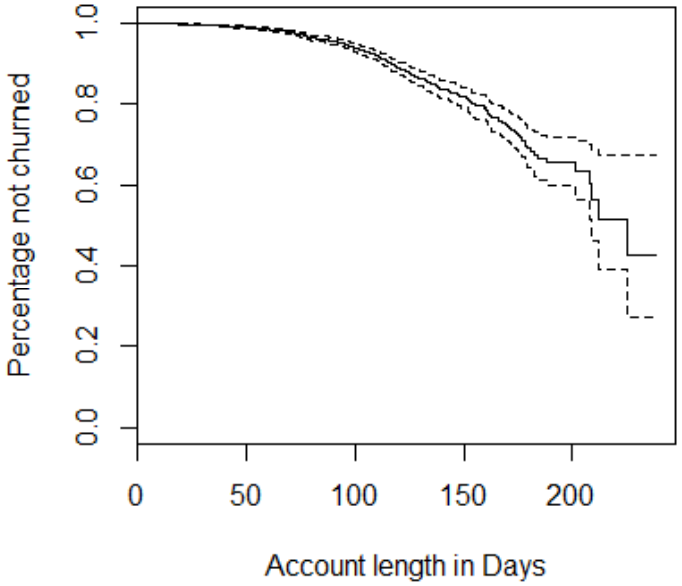
contd..

Perform EDA

Survival Analysis (Cox Hazard Method)

Lets plot now the survival plot for our Customers.

So, from the figure we can see that for the first 50-70 days we have no customer churn, that means we are able to retain all of our customers in the first 50-70 days. For next 50-70 days the percentage of not churned reduces almost from 100% to 80% approx. And decreases considerably in such a way that, after 200 days of tenure the percentage of not churned comes down almost to 40 %. That means by the end of completing 200 days tenure, 60% of our customers are churned who joined us on day 1.



Steps Description

Perform EDA

Main Observations from our EDA

- Usage pattern for Churned Customers which mainly includes minutes (Day/Evening/Night/Int) and their charges incurred is higher than the Not Churned Customers. So, Churned Customers seems to be relatively high value customers & more business relevant.
- Both Box and Bar plots are complementing each other with their findings.
- Box plot for Customer Calls made gives us enough proofs that customers going to be Churned Call Customer Care many times before getting churned.
- Lucky enough to have no missing values in our input dataset.
- From Survival Analysis it is confirmed that Customer Calls Made & International Plan are most influential features. Total Day Mins Used is the also one suggested by our traditional EDA methods.

Steps Description

Feature Selection

After EDA, its now possible for us to identify the major features/variables which influence significantly Customer Churn, we will now use these features for our model preparation. Further we can Drop the features that do/may not influence our outcome variable.

Build Models

Before starting with Building Model, we will segregate our input data in 3 datasets with proportions 60(train):20(test):20(pred for final prediction).

As our business case is for classification model, we will be building models for all possible ML algorithms which can be applied for Classification problem. Our approach will be to proceed with ML algorithm one by one and at the same time will be validating/improving the each ML algorithm outcomes.

Steps Description

contd..

Build Models

Logistic Regression

```
logit.model = glm(churn_status ~ Account_Len + tot_day_chrgs +
                  tot_evening_chrgs + tot_ngt_chrgs +
                  tot_int_chrgs + cust_calls_made, family = "binomial", data = train)
```

Validate & Measure

Now lets Test the above build model on test dataset, as the model is a logistic regression we assume for any observation if the probability of customer to be churned is > 0.5 we will assume that customer is churned and similarly if < 0.5 , we will assume its not churned. Lets validate and measure our assumption .

Logit	False.	True.
False.	741	115
True.	13	13

Accuracy	Specificity	Sensitivity
85.48	98.27	10.15

Steps Description

contd..

Improve Model Performance

Final Prediction

Variating the Threshold Value

In previous step we used our threshold value as 0.5 to decide whether customer is churned or not depending upon whether the value is above 0.5 or not. Lets variate the threshold value a bit and check if the performance of the model is increasing or not. Upon testing for various values (0.4 & 0.3), it was found that the performance of the overall model increases when threshold value is selected as 0.3 . We request you please refer to the attached R code for more details on this.

Final Prediction with threshold value as 0.3 on pred dataset, and calculate the model performance for one last time for logit model.

Logit	False.	True.
False.	756	86
True.	67	41

Accuracy	Specificity	Sensitivity
84.21	91.85	32.28

Steps Description

contd..

Build Models

K-nearest neighbour

knn_model <- knn(train = knn_train, test = knn_test,cl = train$churn_status, k=53)
k= 53 has been calculated as the sqrt of number of observations 2785

Validate & Measure

Validating the above model against our test data set, lets see what our confusion metrix and other performance parameters convey :-

Knn.Pre	False.	True.
False.	752	111
True.	2	17

Accuracy	Specificity	Sensitivity
87.55	99.73	13.28

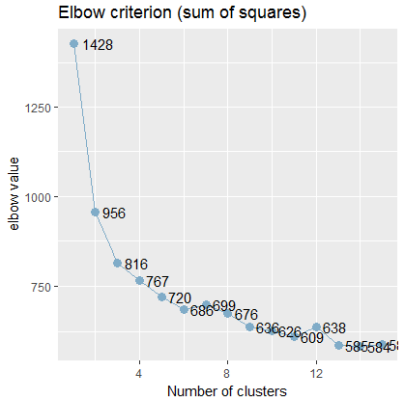
Steps Description

contd..

Improve Model Performance

-> Using Elbow Method to identify Optimum value of k.

-> Normalize the data to supress the scale irregularities as knn works on distance between the nearest neighbors.



Final Prediction

Final Prediction after applying above improvement techniques on pred dataset, lets calculate the Final model performance for knn algorithm

Knn.Pre	False.	True.
False.	805	84
True.	18	43

Accuracy	Specificity	Sensitivity
89.26	92.04	33.85

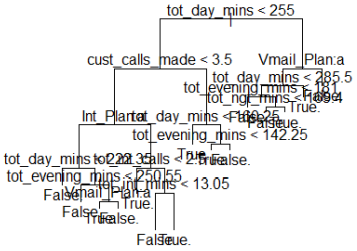
Steps Description

contd..

Build Models

Decision Trees

tree_model <- tree(churn_status ~.-State-Area-Ph_No., data = train)
Classification tree: tree
Variables actually used in tree construction:
"tot_day_mins" "cust_calls_made" "Int_Plan" "tot_evening_mins" "Vmail_Plan“
"tot_int_calls" "tot_int_mins" "tot_ngt_mins" , Number of terminal nodes: 15



Validate & Measure

Validating the above model against our test data set, lets see what our confusion matrix and other performance parameters convey :-

Tree.Pre	False.	True.
False.	739	38
True.	15	90

Accuracy	Specificity	Sensitivity
93.99	98.01	70.31

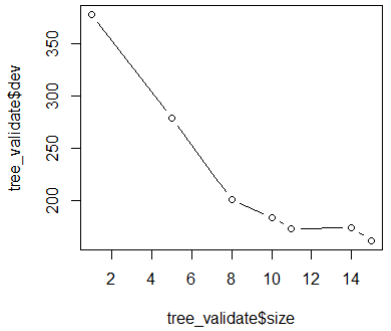
Steps Description

contd..

Improve Model Performance

-> Pruning, so that the model doesn't overfit the dataset.

```
tree_validate <- cv.tree(object = tree_model,FUN = prune.misclass )
plot(x=tree_validate$size,y=tree_validate$dev, type="b")
```



From the plot one can see that tree_validate$dev diff is same from for tree levels 11-14, so can we assume the best tree level size to be 12 (rather than original number 15) at the cost of some bias.

Final Prediction

- Final Prediction after applying above improvement techniques on pred dataset, lets calculate the

  Final model performance for decision trees algorithm

Tree	False.	True.
False.	801	44
True.	22	83

Accuracy	Specificity	Sensitivity
93.05	97.32	65.35

Steps Description

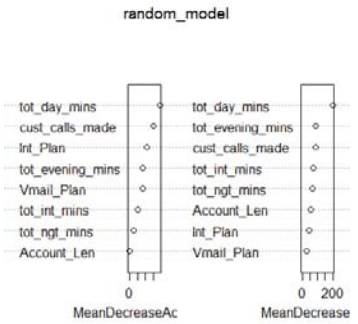
contd..

Build Models

Random Forest

random_model <- randomForest(as.factor(churn_status)~ Account_Len + Int_Plan + Vmail_Plan + tot_day_mins + tot_evening_mins + tot_ngt_mins
+ tot_int_mins + cust_calls_made,data = train,ntree = 2000,
importance = TRUE)

Please note the variable importance plot also compliments our earlier findings from EDA.



Validate & Measure

Validating the above model against our test data set, lets see what our confusion matrix and other performance parameters convey :-

Ran.Pre	False.	True.
False.	746	44
True.	8	84

Accuracy	Specificity	Sensitivity
94.10	98.93	64.62

Steps Description

Improve Model Performance

-> Performance Boosting by Cross Validation with folds k=10.

```
train_control <- trainControl(method="cv", number=10)

random_model_train <- train(as.factor(churn_status)~ Account_Len + Int_Plan + Vmail_Plan
                        + tot_day_mins + tot_evening_mins + tot_ngt_mins
                        + tot_int_mins + cust_calls_made, data=train, method="rf",
                          metric="Accuracy", trControl=train_control)
```

Final Prediction

Final Prediction after applying above improvement techniques on pred dataset, lets calculate the
Final model performance for Random Forest algorithm

Ran	False.	True.
False.	808	47
True.	15	80

Accuracy	Specificity	Sensitivity
93.05	98.17	62.99

Steps Description

Build Models

Support Vector Machine

svm_model<-svm(churn_status ,data=train,kernel='radial',gamma=1,cost=100)

SVM-Type: C-classification SVM-Kernel: radial
cost: 100
gamma: 1 Number of Support Vectors: 2785

Validate & Measure

Validating the above model against our test data set, lets see what our confusion matrix and other performance parameters convey :-

SVM.Pre	False.	True.
False.	754	128
True.	0	0

Accuracy	Specificity	Sensitivity
85.49	100	0

Steps Description

contd..

Improve Model Performance

-> Using Cross Validation technique to optimise the best values for gamma and cost.

```
svm.tune <- tune(svm,churn_status ~ Account_Len + Int_Plan + Vmail_Plan
                 + tot_day_mins + tot_evening_mins + tot_ngt_mins
                 + tot_int_mins + cust_calls_made,data=train,kernel='radial',ranges = list(cost =
c(0.1,1,10,100,1000), gamma = c(0.5,1,2,3,4)))
```

Final Prediction

Final Prediction after applying above improvement techniques on pred dataset, lets calculate the
Final model performance for SVM algorithm

Ran	False.	True.
False.	813	59
True.	10	68

Accuracy	Specificity	Sensitivity
92.74	98.78	55.54

Steps Description

Select the Best Fit Model

In the past few slides for our Classification Business Problem, we have created many models, measured their performances, applied various improvement techniques and at the end of each step we created best model for each corresponding ML algorithms. As we have the performance results with us for the models, lets select the best one out of them which suits best for our business case.

	Logit	kNN	Decision T	Random F	SVM
Accuracy	84.21	89.26	93.05	93.05	92.74
Specificity	91.85	92.04	97.32	98.17	98.78
Sensitivity	32.28	33.85	65.35	62.99	55.54

Decision Trees (CART) model / Random Forest suits best for our business scenario.

Summary

❖ Customers Churned were high business Value customers, their usage pattern was high as compared to not churned.

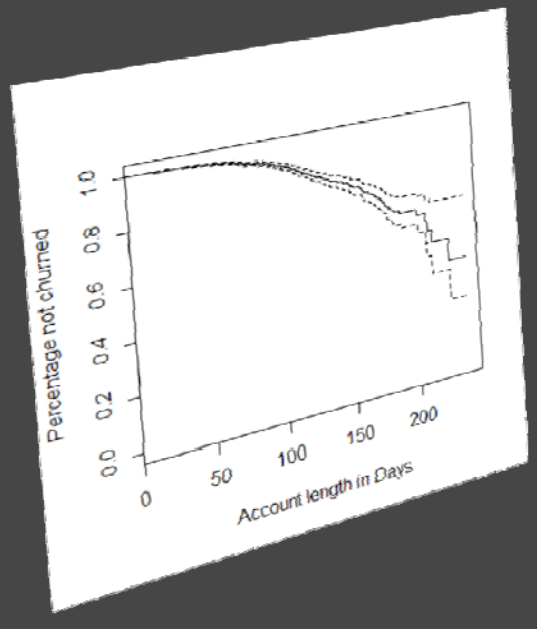
❖ Analysis from Survival Model suggests that we are able to retain all our customers in our first 50-70 days, then in next 50-70 days we are able to retain 80% and further this % reduces to 40% by the end of 200 days.

❖ Customers who are having International Plan OR those who call Customer Care for their service related queries churn more. These two factors are the most influential ones for a customer to be churned. So, one possible measure can be like company can always keep a check for all the customers who call Customer Care more frequently and can pool them in “Possible to be Churned Customers in future”, so that company can take their issues on priority or address them rapidly. Same measure/check can be implemented for International Plan Customers as well.

Summary

❖ Out of various models developed for Prediction of Customers going to be churned, Decision Trees Model (CART) and Random Forest suits best for our business case. Both the models are almost similar w.r.t each other. The accuracy for both the models is almost same i.e. 93.05 %. The only difference between their performance is, RF is able to correctly predict 62.99% of customers to be churned as true but also with the false out ratio of 1.83%. That means apart from predicting 62.99% correctly it is also predicting incorrectly for some 1.83% customers as churned(which is not true).

Where as Decision Tree(CART) model is able to predict correctly 65.35% customers to be churned as true but also predicts incorrectly for some 2.68% customers as churned(which also is not true). So it depends upon Business which model it would like to use if it would like to go for higher predictive power at cost of some Fall Out ratio/error deviation Decision Tree is best. ELSE if company wants to stress more on minimizing error rate than RF can be the best bet.



In God we trust, all others must bring Data…..

Data Diggers
Let's Uncomplicate Business

THANK YOU