

MEF UNIVERSITY

PREDICTIVE CACHE MANAGEMENT

Capstone Project

Olcaý Gürsel Baltaođlu

İSTANBUL, 2017

MEF UNIVERSITY

PREDICTIVE CACHE AMANGEMENT

Capstone Project

Olcay Gürsel Baltaoğlu

Advisor: Assistant Professor Vahid Akbari

İSTANBUL, 2017

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name

Olcaý Gürsel Baltaoğlu

Date

.../.../.....

Signature

EXECUTIVE SUMMARY

PREDICTIVE CACHE MANAGEMENT

Olcay Gürsel Baltaoğlu

Advisor: Asst. Prof Vahid Akbari

09, 2017, 30

Major dependency of a mobile application performance is the response time of backend services. Building a cache layer can be a solution in architectural way to provide better experience to user but it cannot affects when the cache is empty for the first usages. A costly and non-efficient method to deal with this problem is to prepare the cache for the all users in advance. Our purpose is to predict the users' time of use and do the cache preparation just for them.

In the project, we will analyze users' application usage statistic and demographic information to build machine-learning model to predict the users' time of use. As a result, the model will provide a list that contains forecasted user identifier.

Key Words: cache management, predict, mobile application, faster response time

ÖZET

ÖNGÖRÜSEL ÖNBELLEK YÖNETİMİ

Olca Gürsel Baltaoğlu

Tez Danışmanı: Assitan Profesör Vahid Akbari

09, 2017, 30

Mobil uygulama kullanıcılarının uygulama içerisinde daha hızlı bir deneyim yaşamaları, servis aldıkları sistemlerin hızlı dönüşleri ile doğru orantılıdır. Bu servislerin hızlı yanıt dönmesi için yapılması gereken teknolojik altyapı oldukça karmaşık ve yüksek bütçe ihtiyacı gerektirebilir. Bu sebeple mobil uygulamalar kendi katmanlarında ilk çağırılan servis sonucunu önbellekte tutarak belirli bir süre tekrar servis çağırısı yapmamayı tercih ederler (önbellekleme). Bu yöntem müşterinin ilk deneyiminde önbellek boş olduğu için hız konusunda bir fayda sağlamaz fakat kullanım tekrarında hızlı bir deneyim yaşatır.

Amacımız; müşterinin demografik ve uygulama kullanım bilgilerinden yola çıkarak ; müşterinin olası kullanım zamanını tahminlemek ve önbellekleme işlemini önceden yapabilmektir. Böylece müşterinin ilk deneyimini hızlandırmayı amaçlamaktadır.

Projemizde; kullanıcılarımızın mobil uygulama kullanım istatistiklerini ve kişisel bilgilerini makine öğrenme modelimizde kullanarak kullanıcımızın ilgili günde kullanım yapıp yapmayacağı öngörülme çalışılacaktır. Sonuç olarak, modelimiz ilgili gün kullanım gerçekleştirilmesi öngörülen kullanıcı listesi sağlayacaktır.

Anahtar Kelimeler: Ön bellek yönetimi, mobil uygulama, kullanıcı kullanım öngörülleri

TABLE OF CONTENTS

Academic Honesty Pledge	vi
EXECUTIVE SUMMARY	vii
ÖZET	viii
TABLE OF CONTENTS.....	ix
1. INTRODUCTION	1
2. ABOUT THE DATA	2
2.1. Data - Usage Information.....	2
2.2. Data – Users’ Demographic Information.....	3
3. PROJECT DEFINITION	6
3.1. Objective	6
3.2. Scope.....	6
4. METHODOLOGY	7
4.1. Evaluation of Data	7
4.1.1. Missing Values	7
4.1.2. Meaningless attributes	8
4.1.3. Imputation.....	8
4.1.4. Excluding some observation.....	9
4.2. Data Explore	9
4.3. New Features	11
4.3.1. Clustering for user behavioral information	12
4.3.2. Aggregation on User information.....	14
4.4. Model Implementation.....	14
4.4.1. Selection of Model.....	14
4.4.1.1. Naive Bayes Algorithm	15
4.4.1.2. Random Forest Algorithm.....	16
5. RESULTS	21
6. REFERENCES	21
7. APPENDICES	22

1. INTRODUCTION

A cache is a digital place to store something temporarily in a computing environment. In computing, active data is often cached to shorten data access times, reduce latency and improve input/output (I/O). Because almost all application workload is dependent upon I/O operations, caching is used to improve application performance.

Caching stores content or data retrieved by a portal user request. The cached data is stored in memory for a preconfigured amount of time. This means that the data is stored closer to portal users so that when they request it, it can be retrieved from the closest source without going back to the original data source.

Successive identical data requests first access the cache, rather than resubmitting the query to the data source. If the cache has not expired, the information stored in the cache is used instead. Caching improves response time and overall system performance by reducing the load on the information source.

A costly and inefficient method to deal with this problem is to prepare the cache for the all users in advance. The other option is to provide an intelligent system to suggest each user's possible visits. Our purpose is to predict the users' time of use and do the cache preparation just for them.

2. ABOUT THE DATA

We have two kinds of datasets. The first one is about the usages of mobile applications (e.g. login information, usage information inside the application, mobile system properties) and the second one is about the users' demographic information.

The exported data contains information about the usages of the mobile application for the 5th weeks.

Note: The identifier of user is gsm number (msisdn) and it has been provided as encrypted for entire data.

2.1. Data - Usage Information

This data table contains all usage information of mobile application includes inside the actions' info of user.

Table Name: MALT_CLIENT_LOG

Table row counts: 2.5 Billion

IP	Client Ip	217.31.248.73
OPERATION_TIME	Time of method call	18.04.2017 23:57
MSISDN	Gsm information	5459258876
CHANNEL	Channel information (mostly filled by android widget information)	Widget_2345
CLIENT_NAME	If client has a name on network	
USER_HEADERS	Client header parameters	
URL	the url information about where the request came from	https://m.vodafone.com.tr/maltgtwaycbu/api
USER_AGENT	Client information	iphone_VodafoneMCare/1 CFNetwork/808.3 Darwin/16.3.0
TRANSACTION_CODE	Internal transaction code	[r:15830921]
API_METHOD_NAME	Method name	getOptionList
STATUS	Response status	SUCCESS
RESULT_CODE	Response code	S0999000100
DEVICE_MODEL	Device model	
SESSION_ID	SSO session id	96776c8b-4229-4707-9398-52948bd5337d
BYTES	Bytes in traffic	
DURATION	Method execution time	2069
AUTH_TYPE	Authorization type	1
INSERT_TIME	Database insertion time	11.04.2017 00:00

IS_DATA_VALID	Valid input information or not	1
ALL_DATA_REPORTED	Complex responses or not	0
DATA_ID	Request id from backend system	5390676110
RESULT_DESCRIPTION	Response message	İşleminiz başarıyla gerçekleştirilmiştir.
REPORT_ADV_ID	Report advertisement id	3626cc84-ebee-461e-8b4a-714b7486ed19
PUSH_NOTIFICATION_ID	Push notification ID	eD3Qbkl2DDc:

With the scripts in below the data aggregated for usage information purposes;



capstone.sql

- Daily total login count of the user for first 4 weeks as training data. ~5M

msisdn	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
5421111111	34	13	0	0	0	13	11

- in which day do the user login for 5th week as targets. 1 means user logged, 0 means not logged. ~2.7M

msisdn	T1Monday	T2Tuesday	T3Wednesday	T4Thursday	T5Friday	T6Saturday	T7Sunday
5421111111	1	0	1	0	0	0	1

Explanation of the sample data: the user used our application mostly on Monday (sum for 4 weeks is 34) and in 5th week; user used the application on Monday, Wednesday and Sunday

2.2.Data – Users’ Demographic Information

This data table contains all users’ demographic information.

Table Name: subscriber_freq_used_m_hist

Table row counts : The information has been exported just for the user who has mobile usage information

Coloumd Name	Description	Sample data
subscriber_sk	Database level unique id of subscription (max 10 digits)	220773321
report_day	the day of report proceed	20170430
customer_id	unique id of customer	1-I2T0-2027000

customer_sk	Database level unique id of customer (max 10 digits)	35948510
account_id	billing account id of subscription	3,11E+28
activation_city	city information that activation has been made	KONYA
activation_region_name	region information that activation has been made	AKDENIZ
activation_shop_type	Shop type information that activation has been made	17
act_shop_code	activation shop code	S066712
Birthdate	birthdate of customer	24.05.1978
business_segment	bussiness segment	Regular
calculation_yearmonth	segmentation process calculation month	201703
credit_limit	credit limit information of customer	200
credit_risk	credit risk informaiton of customer	2
cred_risk_calc_date	credit processs calculation date	201703
customer_type	customer type (F : firm , S : individual)	F
disconnection_date	disconnection date	20170304
first_package_id	first_package id	RED0032
Gender	gender (E / K)	E
gsm_no	gsm no (encrypted)	5423873102
is_3g_subscriber	is_3g subscriber	1
mnp_flag	mnp flag (In / Out)	I
org_shop_code	org_shop code	S066712
package_id	package id	RED0043
previous_package_id	previous package id	RED0032
previous_status	previous status (Active, Suspended, Cancel, Duning ...)	A
previous_vip_code	previous vip code	100
risk_score	risk score (1..10)	2
segment_current_value	segment value of subscription(HV1, HV2, LV1...)	HV1
segment_given_value	segment given value (new processed)	HV1
segment_given_value_monthly	segment given_value monthly (new processed monthly)	LV1
sm_market_seg	market segment information	HV

sm_market_seg_last_calc_month	sm_market_seg_last_calc month	201703
sm_micro_seg	sm_micro seg	HV
sm_micro_seg_last_calc_month	sm_micro_seg_last_calc month	201703
sm_orgcur_seg	sm_orgcur_seg (yougth, mass, premium)	Youth.OrgSegCur
sm_orgcur_seg_last_calc_month	sm_orgcur_seg_last_calc month	201703
sm_orggiv_seg	unknown	NULL
sm_orggiv_seg_last_calc_month	unknown	NULL
start_date	the day of activation	
Status	current subscription status (Active, Suspended, Cancel, Duning ...)	A
subscription_type	subscription type	1
Tenure	tenure information	180
vip_code	vip code	111
source_system_sk	source_system sk	22
load_date	table load date	20170314
load_ett_date	load_ett date	20170314
billing_account_code	billing_account code	6048101285
billing_account_id	billing_account id	VAZAY924-1
has_m2m_opt_flag	has_m2m_opt flag (1, 0)	1
sm_glob_seg	global segmentation information	GlobSeg.11
sm_glob_seg_last_calc_month	sm_glob_seg_last_calc month	201703
customer_display_id	customer_display id	26742566
last_port_in_date	last_port in date	30.04.2017
last_port_out_date	last_port out date	NULL
billing_account_sk	billing_account source system id, databaselevel	6332505
contact_sk	contact source system id , database level	36645862
contact_id	contact id	NWCE--1 2844034
is_volte_subscriber	is_volte subscriber (1/ 0)	1
is_4g_subscriber	is_4g subscriber	1
is_active_fut_enterprise_user	is_active_fut_enterprise user	0
is_fut_enterprise_user	is_fut_enterprise user	1
package_sk	package sk	47004
previous_package_sk	previous_package sk	47003
cancel_date	cancel date of subscriber	20170302
Pmonth	Report taken date	201704

3. PROJECT DEFINITION

3.1.Objective

Forecasting each user's most probable visit days is the main objective for this study. Thus, the system will prepare the caches for the incoming user to provide a better user experience (UX) while minimizing the memory cost.

3.2.Scope

As outputs, the model will provide a user identifier list (GSM number) about the estimation of who will use the application today. Therefore, the relevant caches will be prepared daily bases usage forecast. The contribution of this kind of cache optimization will have a significant commercial contribution in terms of revenue generation.

4. METHODOLOGY

4.1. Evaluation of Data

Different types of data will require different types of cleaning methods. For this dataset, I have searched for (1) missing values, (2) removed meaningless attributes, (3) did relevant imputations, (4) excluded some of the observations from the dataset.

4.1.1. Missing Values

Identify the attributes that have “null” values more than 80% to exclude because the imputation will be meaningless. I have checked the meaning of “null” if there is.

```
> sort(sapply(Dataset_Main, function(x) { sum(is.na(x)) }), decreasing=TRUE)
```

credit_limit	credit_risk	cred_risk_calc_date
100000	100000	100000
risk_score	sm_market_seg	sm_market_seg_last_calc_month
100000	100000	100000
sm_micro_seg	sm_micro_seg_last_calc_month	sm_orgcur_seg_last_calc_month
100000	100000	100000
sm_orggiv_seg	sm_orggiv_seg_last_calc_month	last_port_out_date
100000	100000	97877
disconnection_date	billing_account_sk	cancel_date
97556	80068	63906
gender	last_port_in_date	mnp_flag
61795	44647	43993
previous_vip_code	contact_sk	contact_id
43849	40048	39074
business_segment	billing_account_code	billing_account_id
39047	38973	38973
segment_given_value_monthly	segment_given_value	previous_package_id
16059	15862	14190
previous_package_sk	previous_status	segment_current_value
14190	14079	9964
activation_city	activation_region_name	activation_shop_type
3825	2713	1860
sm_glob_seg	sm_glob_seg_last_calc_month	org_shop_code
1856	1856	1746
act_shop_code	customer_display_id	calculation_yearmonth
1028	715	458
birthdate	subscriber_sk	report_day
80	5	5
customer_id	customer_sk	account_id
5	5	5
customer_type	first_package_id	gsm_no
5	5	5
is_3g_subscriber	package_id	sm_orgcur_seg
5	5	5
start_date	status	subscription_type
5	5	5

According to result of the attributes lists that have NA values;

- Some attributes have not store any value, so it is easy to decide the exclusion for these.
(we analyzed 100K observation and some attributes have NA for all)

4.1.2. Meaningless attributes

Some attributes will be excluded because they represent only a number which has meaning just for the database level ecosystem

- System or relational row Ids and unique numbers that are used for foreign key. These attributes will be excluded from our dataset

(subscriber_sk, customer_id, customer_sk, account_id, act_shop_code, org_shop_code, source_system_sk, billing_account_code, billing_account_id, customer_display_id, billing_account_sk, contact_sk, contact_id)

- Date information about when the data analysis has been made. These attributes will be excluded from our dataset.

(report_day, calculation_yearmonth, red_risk_calc_date, segment_given_value_monthly, sm_market_seg_last_calc_month, sm_micro_seg_last_calc_month, sm_orgcur_seg_last_calc_month, sm_orggiv_seg_last_calc_month, load_date, load_ett_date, sm_glob_seg_last_calc_month, last_port_in_date, last_port_out_date, cancel_date, pmonth)

- Some information is related with system processes itself; it is not related with user. These attributes were excluded from the dataset.

(activation_region_name: it is not related with region value of where the subscription activated. it has value to where the activation process ended by activation channel)

- Duplicated attributes which contain same values; while exporting and aggregating the data msisdn and gsm_no attributes have been used on joining the data table. One of them can be excluded (gsm_no will be excluded from dataset)

4.1.3. Imputation

By exploring and analysing the data some imputation should be done accordingly

- Mnp_flag (mobile number portability – Changing gsm operators with same gsm number) : the feature keeps value about the subscriber has portIn or portOut flow for mobile number portability. around 44K rows have missing values for the subscriber who didn't do the mnp process. So it can be assigned for new categorical value as None "N"
- Some attributes are used for previous information of current attributes. If the current information doesn't change its previous attributes' value is NULL, so it can be filled as "None"
(previous_vip_code, previous_package_id, previous_status)

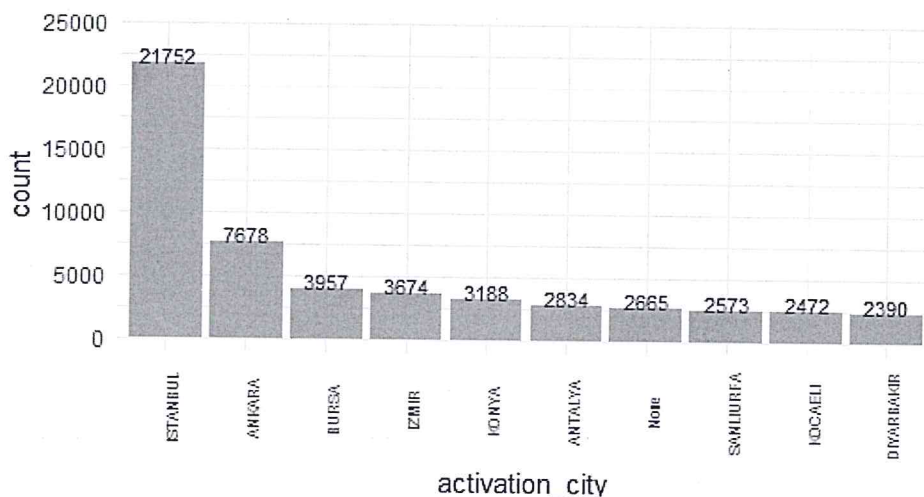
- Some attributes have analytic values, before completion of analysis these values are NULL, so it can be filled as "None"
(business_segment, segment_current_value, segment_given_value, sm_glob_seg)
- Activation_shop_type : most common activation shop type can be used for imputation (filled as "1" for 1855 observations)

4.1.4. Excluding some observation

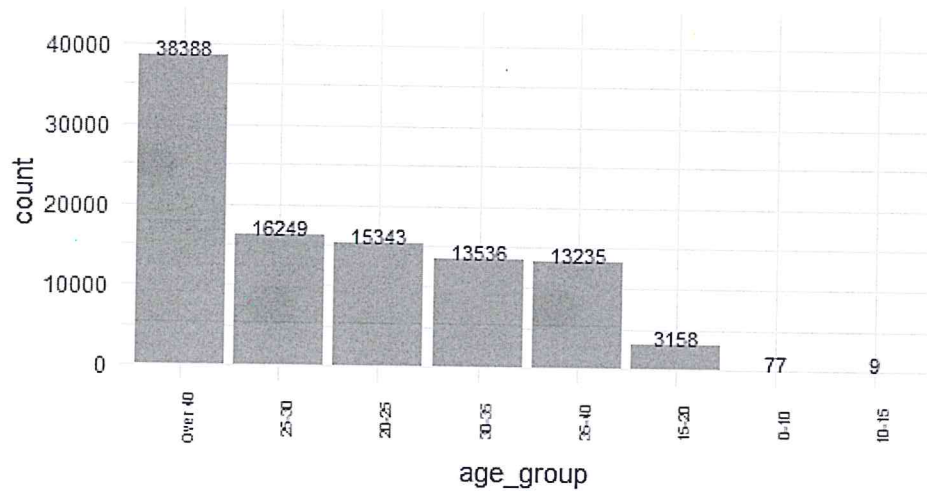
We used "msisdn" (at usage table) and "gsm_no" (at user information table) attributes to join the usage information and user demographic information tables. If "gsm_no" is empty it means we cannot find the relevant customer information, it can be related with the exported date of customer information table, these records can be deleted.

4.2. Data Explore

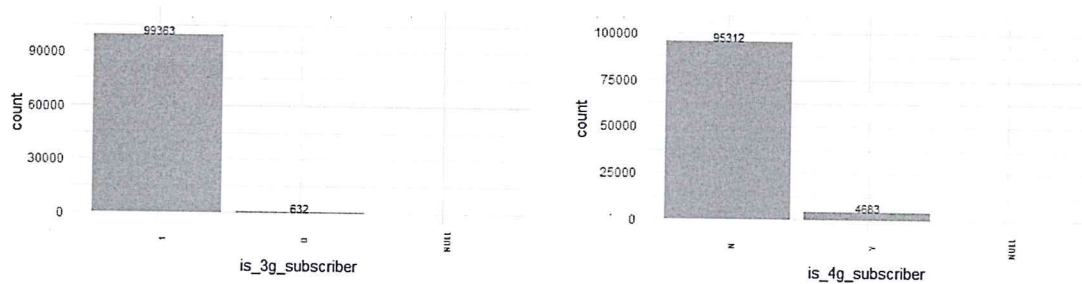
- Activation City : as you can in below graph for top 10 city , %30 activation of 100K subscription 'who use our mobile application' have been done on Istanbul, Ankara



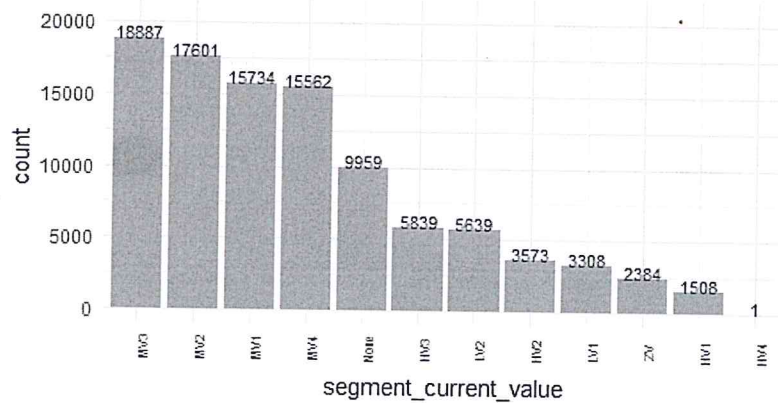
- Age distribution of mobile application users: as you can in below graph, age distribution of 100K subscription 'who use our mobile application' mostly over 40

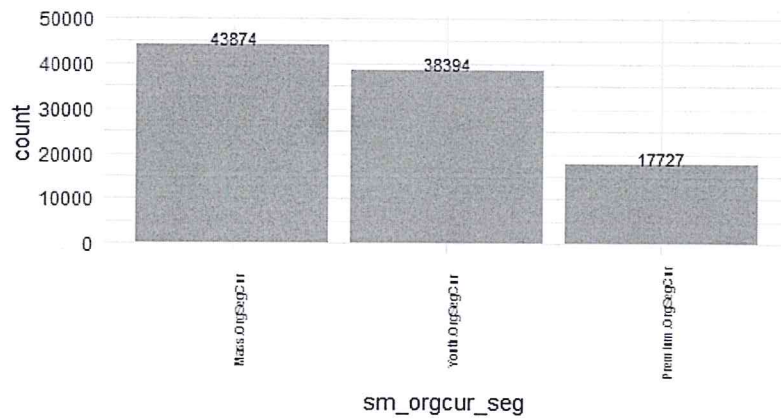


- Network distribution : 3g percentage is higher than 4g and almost all users use mobile data. Only 564 users has no usage 3g or 4g services, probably they use wireless to connect the application

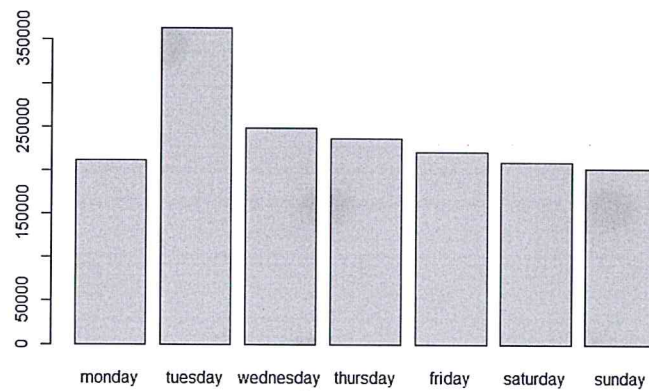


- Segmentations: Our mobile user mostly in MV (Middle Value) segmentation and also our Mass segment



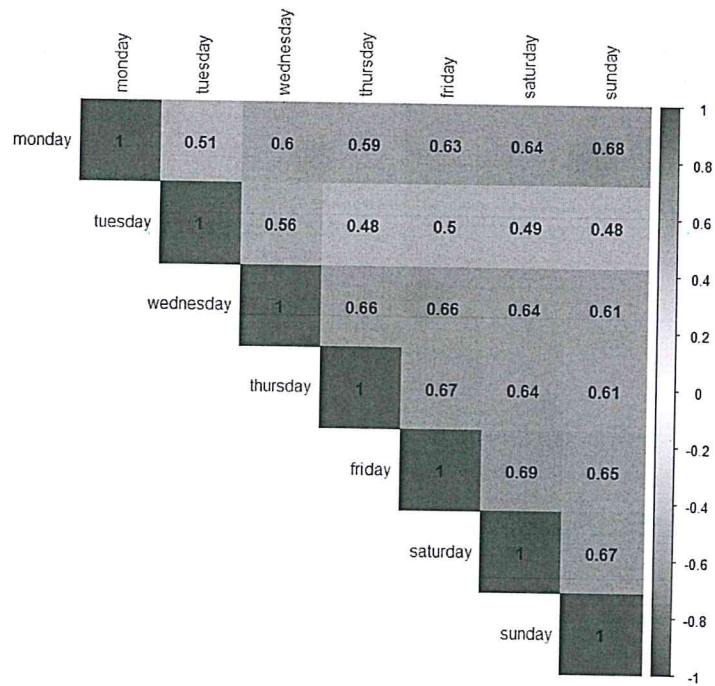


- Daily usage counts for 4 weeks



4.3.New Features

You can find below the correlation of daily usage (behavioral) information; as you can see we have correlation between usage of the days but it's not strong as expected(strong means if correlation is higher then 0,7). We will use this information to create clustering as usage pattern of user.



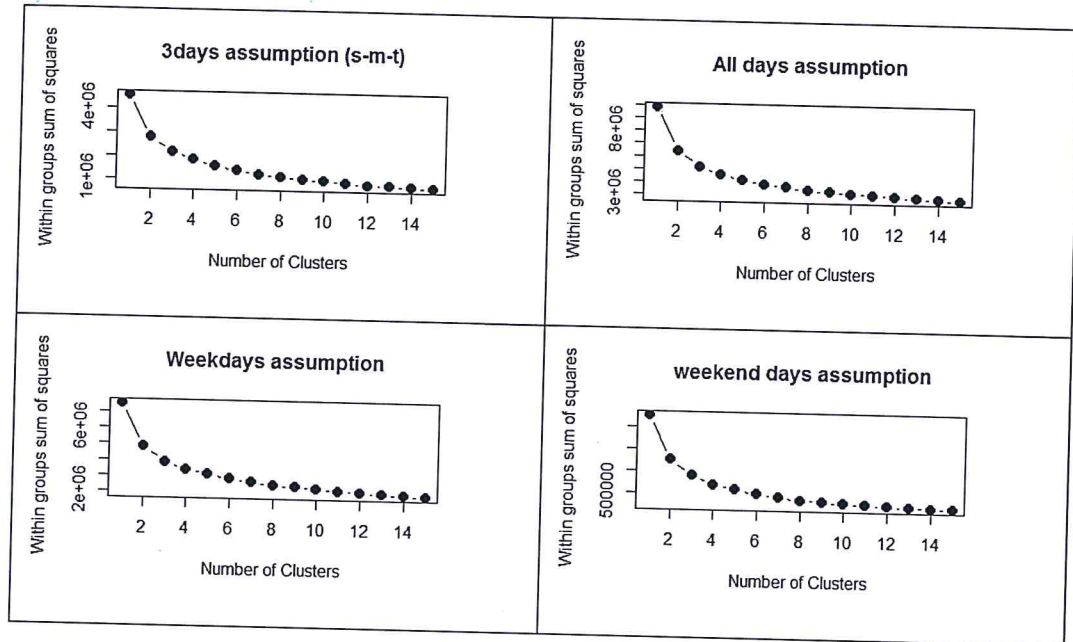
4.3.1. Clustering for user behavioral information

The aim is to segregate groups with similar traits and assign them into clusters by selecting some initial attributes

- Assumption-1: 3 days period usage information can be a profile for a user. For example, if we want to forecast the login on Monday, we can focus the day itself and 2 days before the day. Target is Monday and important attributes are Saturday, Sunday and Monday.
- Assumption-2: Working and weekend days related to the usage information can be a profile attribute for a user.
- Assumption-3: Daily usage information can be a profile for a user.

With these assumptions above we created new features and clusters for the behavioral inputs and assessing the optimal number of clusters with the Elbow method:

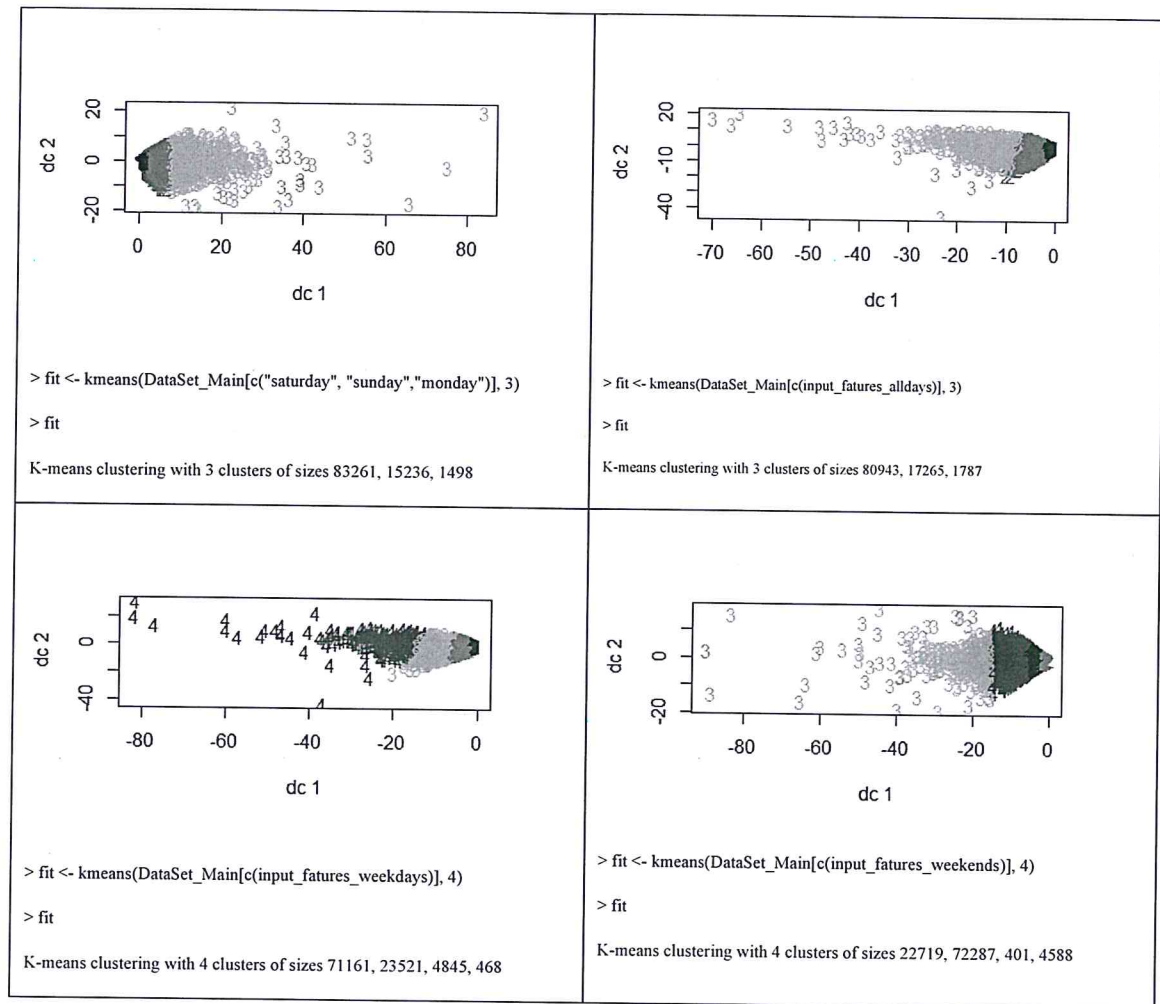
```
input_features_behavioural=c("monday","tuesday","wednesday","thursday","friday","saturday","sunday",
"cluster_3day_monday","cluster_3day_tuesday","cluster_3day_wednesday",
"cluster_3day_thursday","cluster_3day_friday","cluster_3day_saturday",
"cluster_3day_sunday",
"cluster_all_day",
"cluster_weekdays","cluster_weekends")
```



Note: One method to validate the number of clusters is the elbow method. The idea of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 15 in the examples above), and for each value of k calculate the sum of squared errors (SSE). Then, plot a line chart of the SSE for each value of k . If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best.

- optimal number of clustering is 3 for 3-day usages. (We examined for all 3 days combinations, results are same)
- optimal number of clustering is 3 for all day usages
- optimal number of clustering is 4 for weekdays usages
- optimal number of clustering is 4 for weekend usages

Lets see the clustering results in graphs:



4.3.2. Aggregation on User information

There are two date information about user, birthdate and start_date, both have been added to our data set after the aggregation as “age” and “customerSince” features.

4.4.Model Implementation

4.4.1. Selection of Model

We have to build a repeatable model to see the accuracy of our model for each day, for example, our model can forecast the usage of Monday better than the others can or user information’s can describe the usage just for weekend usage. Therefore, we should choose the best accurate model for our data then we should tune and deep analysis for each targets.

Input feature lists :

input_features_behavioural=c("monday","tuesday","wednesday","thursday","friday","saturday","sunday",
"cluster_3day_monday","cluster_3day_tuesday","cluster_3day_wednesday",
"cluster_3day_thursday","cluster_3day_friday","cluster_3day_saturday",
"cluster_3day_sunday","cluster_all_day","cluster_weekdays","cluster_weekends")

input_features_demographic=c("activation_city","activation_shop_type","business_segment",
"customer_type","first_package_id","is_3g_subscriber","mnp_flag","package_id",
"previous_package_id","previous_status","previous_vip_code","segment_current_value",
"segment_given_value","sm_orgcur_seg","status","subscription_type",
"tenure","vip_code","has_m2m_opt_flag","sm_glob_seg","is_volte_subscriber",
"is_4g_subscriber","is_active_fut_enterprise_user","is_fut_enterprise_user",
"age_calculated","customerSince_calculated","age_group")

4.4.1.1.Naive Bayes Algorithm

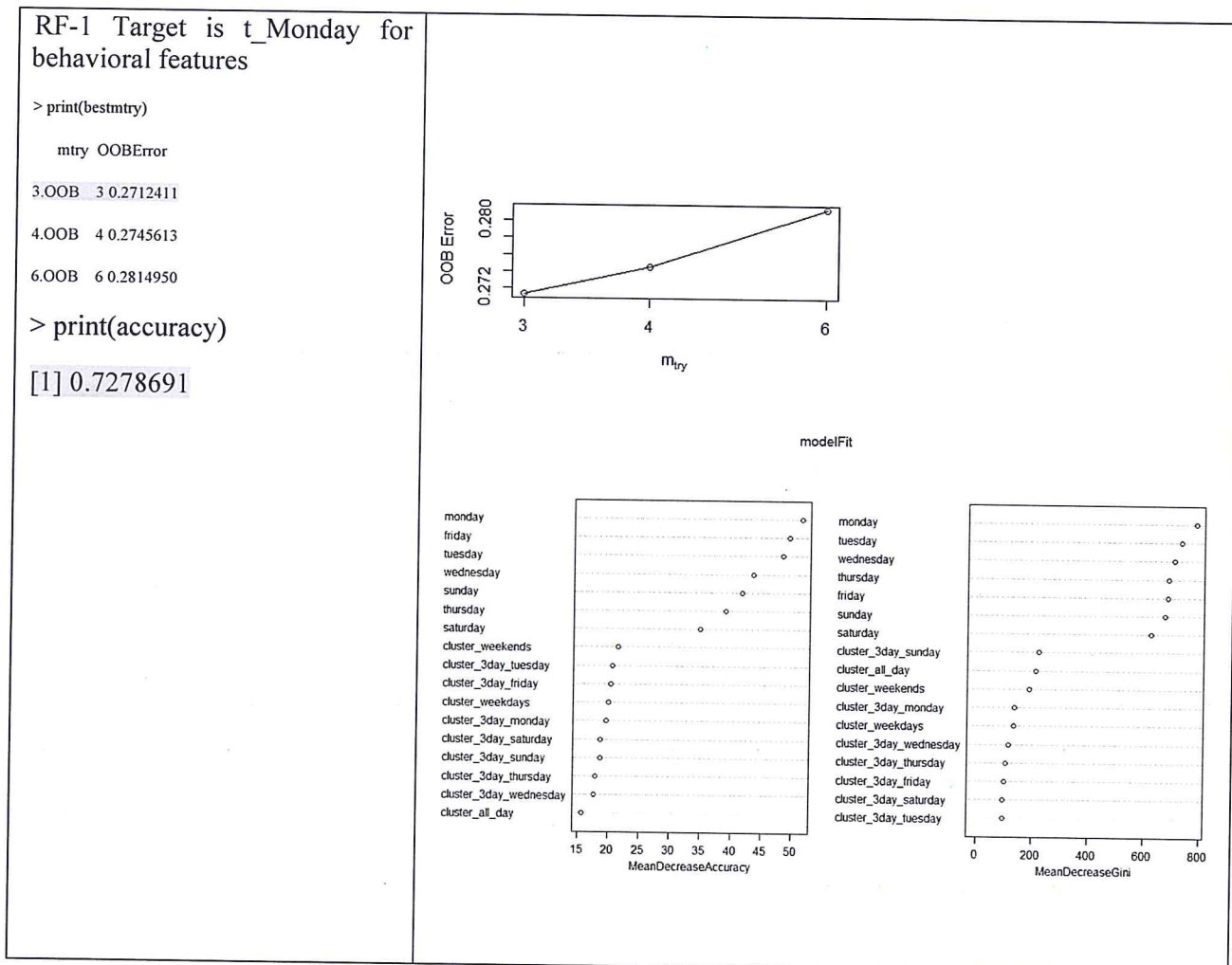
Here are the naive bayes model results have been listed in below for various targets

Target	Type	confusionMatrix			Accuracy
t_Monday	Behavioral	Prediction	Y	N	0.6947
		Y	14924	4673	
		N	2959	2443	
t_Tuesday	behavioral	Prediction	Y	N	0.9165
		Prediction	Y	N	
		Y	22879	185	
		N	1903	32	
t_Wednesday	behavioral	Prediction	Y	N	0.65
		Y	12591	6252	
		N	2497	2497	
t_Thursday	behavioral	Prediction	Y	N	0.695
		Y	14817	5345	
		N	2279	2558	
t_Friday	behavioral	Prediction	Y	N	0.692
		Y	14871	5293	
		N	2406	2429	
t_Saturday	behavioral	Prediction	Y	N	0.7011
		Y	15072	5114	
		N	2357	2456	
t_Sunday	behavioral	Prediction	Y	N	0.6822
		Y	14472	5647	
		N	2298	2582	

As you can see above we have accurate results ~0.6 and forecasting the usage of Tuesday is higher with 0.91.

4.4.1.2.Random Forest Algorithm

The model results have been listed in below for various targets

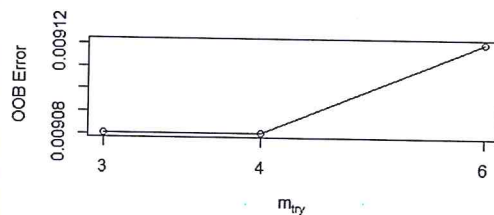


RF-2 Target is t_tuesday for behavioral features

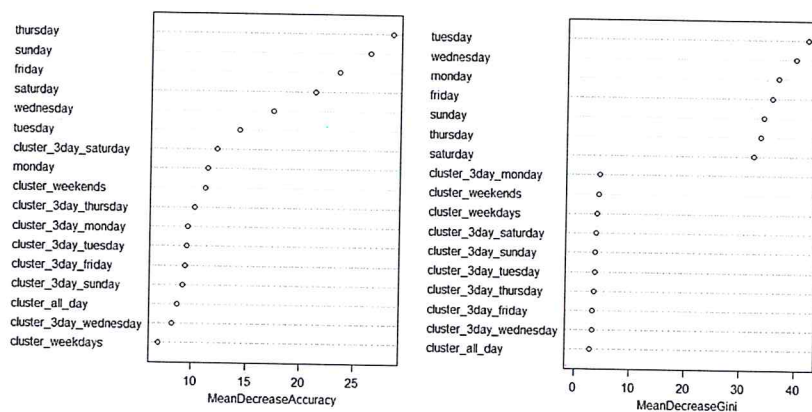
```
> print(bestmtry)

mtry  OOBError
3.OOB 3 0.009080484
4.OOB 4 0.009080484
6.OOB 6 0.009120486

> print(accuracy)
[1] 0.9913197
```



modelFit

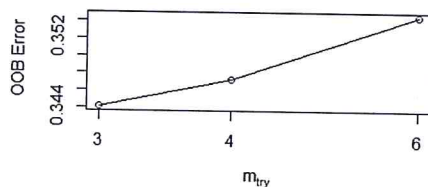


RF-3 Target is t_Wednesday for behavioral features

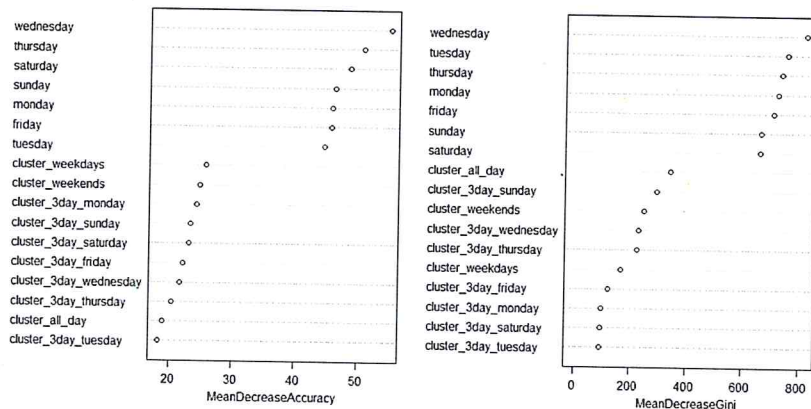
```
> print(bestmtry)

mtry  OOBError
3.OOB 3 0.3441117
4.OOB 4 0.3472185
6.OOB 6 0.3545256

> print(accuracy)
[1] 0.6549062
```



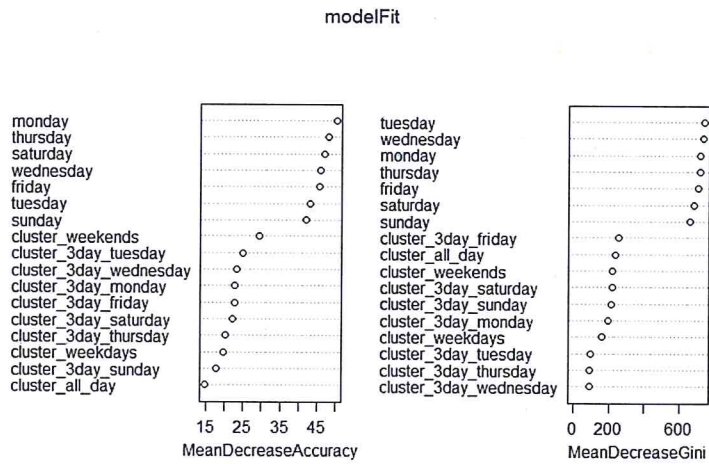
modelFit



RF-4 Target is t_thursday for behavioral features

```
> print(accuracy)
```

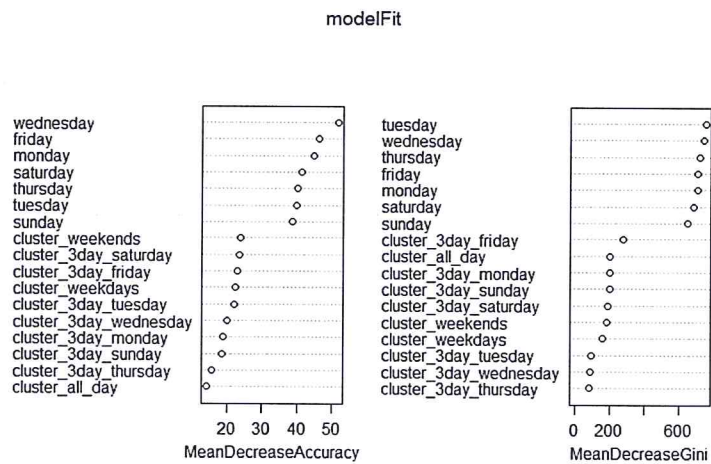
```
[1] 0.7075883
```



RF-5 Target is t_friday for behavioral features

```
> print(accuracy)
```

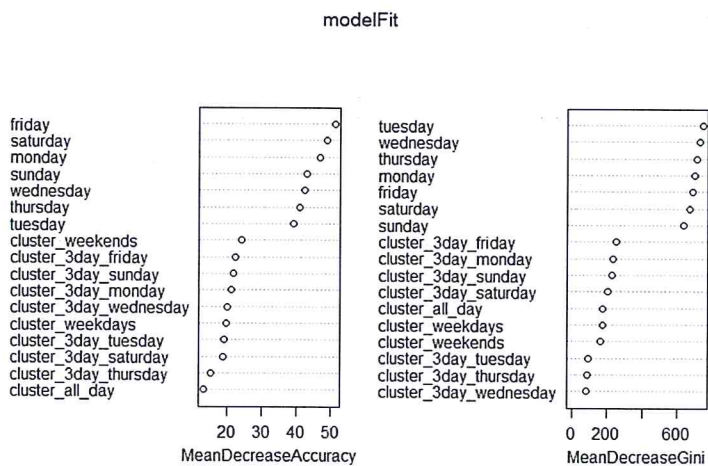
```
[1] 0.7071883
```



RF-6 Target is t_saturday for behavioral features

```
> print(accuracy)
```

```
[1] 0.7155486
```

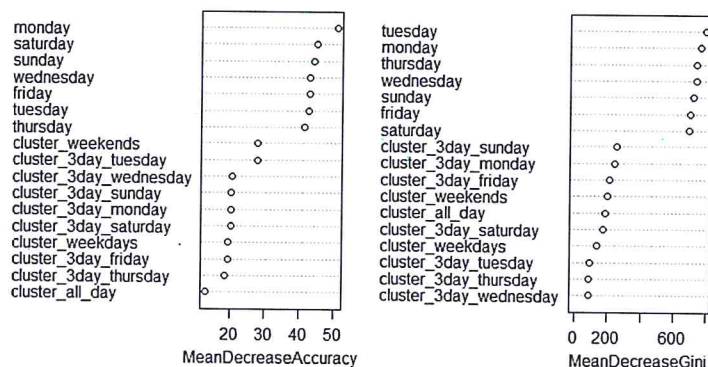


RF-7 Target is t_Sunday for behavioral features

```
> print(accuracy)
```

```
[1] 0.6913877
```

modelFit



RF-8 Target is t_Monday factor features of demographic information

```
> print(bestmtry)
```

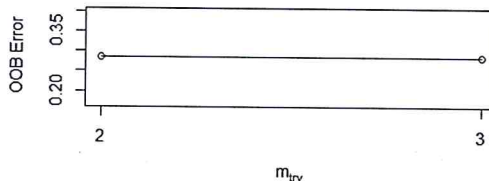
```
mtry OOBError
```

```
2.OOB 2 0.2834818
```

```
3.OOB 3 0.2834818
```

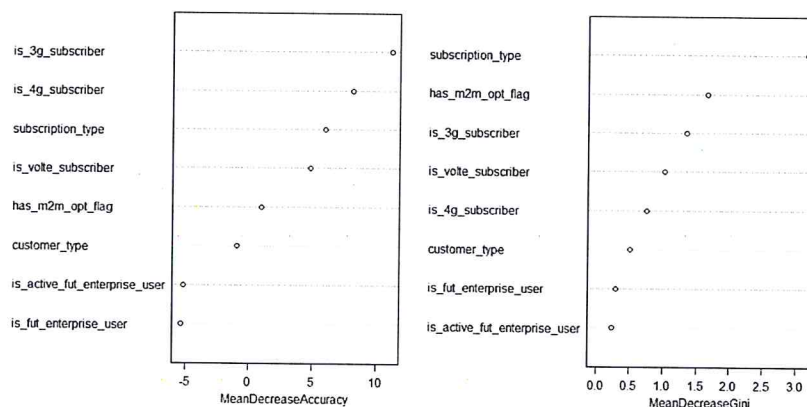
```
> print(accuracy)
```

```
[1] 0.7153486
```



"

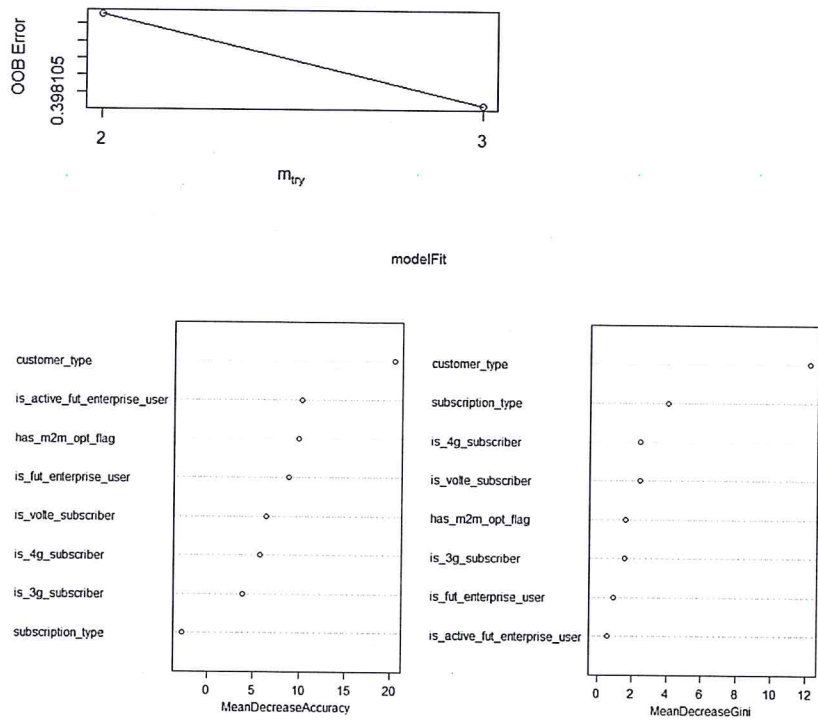
modelFit



RF-9 Target is t_wednesday
factor features of **demographic**
information

```
> print(bestmtry)
      mtry OOBError
2.OOB    2 0.3981279
3.OOB    3 0.3981012

> print(accuracy)
[1] 0.6035841
```



Comparing the two model results it seems more accurate results for random forest model

Target	type	Naïve Bayes Accuracy	Random Forest Accuracy
t_Monday	behaviroal	0.6947	0.72
t_Tuesday	behaviroal	0.9165	0.99
t_Wednesday	behaviroal	0.65	0.65
t_Thursday	behaviroal	0.695	0.70
t_Friday	behaviroal	0.692	0.70
t_Saturday	behaviroal	0.7011	0.71
t_Sunday	behaviroal	0.6822	0.69

5. RESULTS

- We built more accurate results with “Random Forest” algorithm.
- As one can see in the results of most important variables on the result of RF8 and RF-9 are “is_3g_subscriber” and “is_4g_subscriber” which are mostly “TRUE” for our dataset. It means the model that built on demographic information is not a reliable one.
- We have high accuracy (~0,7) on the model that built on behavioral information. It means building a predictive cache management model for mobile application is logical.
- Accuracy for forecasting “Tuesday” usage is too high because we have highly usage on Tuesday on our dataset. Thus, there might be bias related to this factor.
- Different models should be built for each day.
- According the importance variable result of our RF algorithm, our clusters has less importance than the daily usages features.

6. REFERENCES

- Clustering (<https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>)
- Elbow Method for optimal number of clustering (<https://blogs.ocks.org/rpgove/0060ff3b656618e9136b>)
- Example of K-Means Clustering with R (<https://rpubs.com/FelipeRego/K-Means-Clustering>)
- Naive Bayes algorithm (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930525/>)
- Predictive Caching (<http://dl.acm.org/citation.cfm?id=864864>)
- A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python) (<https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/>)
- Random Forest (https://www.tutorialspoint.com/r/r_random_forest.htm)

- CorPlot Visualization Methods (<https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html>)

7. APPENDICES

- CapstoneProjectV7.R : R code of the project
- Capstone.sql : The sql script that has been used for export the data
- export_enhv2.csv : the dataset