MEF UNIVERSITY

CHURN PREDICTION IN VODAFONE TURKEY

Capstone Project

Gokhan Genel

İSTANBUL, 2017

MEF UNIVERSITY

A CHURN PREDICTION MODEL FOR VODAFONE TURKEY

Capstone Project

Gokhan Genel

Advisor: Assoc. Prof. Semra Ağralı

İSTANBUL, 2017

EXECUTIVE SUMMARY

A CHURN PREDICTION MODEL FOR VODAFONE TURKEY

Gokhan Genel

Advisor: Assoc. Prof. Semra Ağralı

SEPTEMBER, 2017, 18 pages

This Capstone Project focuses on finding a churn pattern in Vodafone postpaid consumer subscribers. The churn value refers to disconnection from subscription via port-out / Mobile number portability (MNP). It is one of the most important items that demonstrate revenue-loss. The subscriber who churned with MNP switches to a rival GSM operator.

The cost of keeping an existing customer is generally cheaper than the cost of acquisition of a new customer. Focusing on customer retention is one of the most profitable strategy for growth. Statistical analysis and machine learning can help analyze churn activities and they can even alert companies when their existing customers are likely to churn.

By using machine-learning algorithms, this project aims to detect Vodafone postpaid consumer subscribers who are likely to churn. This project will help the company to decrease its revenue loss.

Key Words: Data Analytics, Machine Learning, Churn Prediction, Random Forest, Decision Tree

ÖZET

A CHURN PREDICTION MODEL FOR VODAFONE TURKEY

Gokhan Genel

Tez Danışmanı: Doç. Dr. Semra Ağralı

EYLUL, 2017, 18 sayfa

Bu Capstone projesi, faturalı mobil telefon hattına sahip Vodafone Türkiye bireysel abonelerinin numara taşıma vasıtasıyla diğer operatorlere geçiş sürecindeki paternleri belirlemeye ve henüz numara taşımasi yapmamış aktif müşterilerde bu olasılığı tahminlemeye odaklanmaktadır.

Rakip GSM operatorüne numara taşıma, bu şirketler için en önemli gelir kaybı kalemlerinden biridir. Numara taşıma ile hattı iptal edilen abone, geçiş yaptığı operatörde aktif olarak yaşam döngüsüne başlar. Sektöründe pazar liderliği hedefleyen şirketler için de churn oranının azaltılması çok önemli bir noktayı temsil etmektedir. Numara taşıma vasıtasıyla kaybedilen her abone rakip firmalarda yeni abone olarak sayılacağı için bir numara taşıma işlemi, abone sayısı bakımından incelendiğinde diğer operatörlerle abone sayısı farkı eksi yönde "iki abone" olarak değişmektedir.

Numara taşıma işlemi çok onemli bir gelir kaybi kalemi yaratmasının dışında müşteri memnuniyetinin izlenmesi açısından da önemlı kalemdır. Bır kullanıcı Net Promoter Score veya call center aramaları ıle memnuniyet derecesini belirtebilir. Buna karşın aldığı hizmetin yeteri kadar iyi olmadığını düşünen müşteriler mevcut hatlarını geri bildirim olmaksınız iptal ettirebilirler. Numara taşıma yapan abonelerin yarattığı paternlere yakın paterne sahip aboneler belirlenerek gerekli aksiyonlar alınabilir. Bu abonelerin belirlenerek numara taşımaktan kaçındırılması, dolaylı yoldan müşteri sadakatinin artmasına yardımcı olacaktır.

Anahtar Kelimeler: Data Analytics, Machine Learning, Churn Prediction, Random Forest, Decision Tree

TABLE OF CONTENTS

1. INTRODUCTION

The idea of this project stem from challenges the author came in terms with in while he was working in data warehouse department in Vodafone. One of the most important subjects of the company was the loss of subscribers due to lack of customer satisfaction. Usually this satisfaction can partially be measured by using NPS (Net Promoter Score) and call center calls of the active subscribers. However, it is quite difficult to get feedback from subscribers who are already disconnected. This project illustrates that it is possible to detect subscribers are likely to churn via MNP. Therefore, it facilitates the communication with these subscribers.

Besides the revenue-loss it is important to prevent churn for prestigious companies. Vodafone Turkey aims to be the number one GSM operator in Turkey. Though the company continues the gain new subscribers it should also sustain the current customer satisfaction and loyalty.

Though the churn value refers to disconnection from subscription via port-out / Mobile number portability (MNP) it is one of the most important items that demonstrate revenue-loss. The subscriber who churned with MNP switches to a rival GSM operator.

The cost of keeping an existing customer is generally cheaper than the cost of acquisition of a new customer. Focusing on customer retention is one of the most profitable strategy for growth. Statistical analysis and machine learning can help analyze churn activities and they can even alert companies when their existing customers are likely to churn.

By using machine-learning algorithms, this project aims to detect Vodafone postpaid consumer subscribers who are likely to churn. This project will help the company to decrease its revenue loss.

2. ABOUT THE DATA

2.1. About the Company

Vodafone Turkey was formed in 28 December 2005. On 24 May 2006 Telsim brand name was changed into Vodafone Turkey. Incorporated by the Vodafone Group, one of the biggest international mobile communication companies of the world in terms of revenues, Vodafone Turkey is the 2nd biggest mobile communication company of with ~22.5 million users as of 01 January 2017, an increase of 105,000 customers from the fourth quarter of 2011. As the second biggest direct international investment of Turkey, the total investment of Vodafone Turkey, including the acquisitions, since 2006 has 11 billion. Vodafone Turkey operates in 81 cities of Turkey with its approximately more than 3300 employees, more than 1200 stores, 23 thousand points of sale and a family of 53 thousand people.

2.2. Dataset

The Dataset extracted from Vodafone DWH Databases includes 15K still active postpaid subscribers by May'17 and 5K churned subscribers who left via MNP in Apr'17. The dataset does not include any identification information. It has 20000 observations and 10 attributes.

Name	Type	Description
SUBSCRIBER_SK	Number	Unique Subscriber surrogate key
TENURE_RANGE	Char	Categorical, the number of months since subscriber start date.
TENURE	Number	the number of months since subscriber start date
BLACKLIST_FLAG	Number	Subscriber exists in black list in last 12 months 1/0
RESTRICTION_FLAG	Number	Line restriction in last 12 months 1/0
PRE_TO_POST_FLAG	Number	Migration from prepaid line to postpaid line in last 12 months 1/0

LATE_PAYMENT_COUNT	Number	Nof late payment count in last 12 months
OVERUSAGE_COUNT	Number	Nof overusage (exceed the quota) count in last 12 months
NOF_COMPLAINT	Number	Nof complaint in last 12 months
CHURN_FLAG	Number	Churn 0/1
TARIFF_CHURN_RATIO	Number	Churn ratio in subscriber tariff = (Port Out Count in the month / AVG NOF Active Subscriber in the tariff)

SUBSCRIBER SK	TENURE RANGE	TENURE	BLACKLIST FLAG
Unique Key Class: Number	Length:20000 Class :character Mode :character	Min. : 0.00 1st Qu.: 20.00 Median : 71.00 Mean : 88.98 3rd Qu.:126.00 Max. :269.00	Min. :0.0000 1st Qu.:0.0000 Median :0.0000 Mean :0.2544 3rd Qu.:1.0000 Max. :1.0000
PRE TO POST FLAG	LATE PAYMENT COUNT	OVERUSAGE COUNT	RESTRICTION FLAG
Min. :0.00000 1st Qu.:0.00000 Median :0.00000 Mean :0.06605 3rd Qu.:0.00000 Max. :1.00000	Min. : 0.0 1st Qu.: 0.0 Median : 3.0 Mean : 4.2 3rd Qu.: 8.0 Max. :13.0	Min. : 0.000 1st Qu.: 0.000 Median : 2.000 Mean : 2.677 3rd Qu.: 4.000 Max. :13.000	Min. :0.00000 1st Qu.:0.00000 Median :0.00000 Mean :0.01685 3rd Qu.:0.00000 Max. :1.00000
TARIFF CHURN RATE	NOF COMPLAINT	CHURN FLAG	
Min. : 0.00 1st Qu.: 9.77 Median : 27.48 Mean : 43.98 3rd Qu.: 54.08 Max. :2222.22	Min. : 0.000 1st Qu.: 0.000 Median : 0.000 Mean : 0.187 3rd Qu.: 0.000 Max. :12.000	Min. :0.00 1st Qu.:0.00 Median :0.00 Mean :0.25 3rd Qu.:0.25 Max. :1.00	

All attributes were derived from the Vodafone DWH datamart tables except for subscriber surrogate key. SQL was used for data extraction process. The churn dataset data has not been manipulated and extracted randomly.

3. PROJECT DEFINITION

3.1. Problem Statement

Customer retention refers to the percentage of customer relationships that, once established, a business is able to maintain on a long-term basis. Customer retention is a simple concept, happy customers who feel important and are regularly communicated with in the right way will keep coming back. It is a major contributing factor in the net growth rate of businesses (Customer Retention, 2017).

There are two main headline in market size: keeping an existing subscriber and acquisition of new subscribers. The cost of keeping an existing subscriber is generally cheaper than the cost of acquisition of a new subscriber. According to a research about comparison between customer acquisition and customer retention done by Khalid Saleh, the articles listed below are obtained.

- %44 of companies have greater focus on customer acquisition while %18 that focus on retention.
- While more than 89% of companies see customer experience as a key factor in driving customer loyalty and retention, 76% of companies see Customer Lifetime Value as an important concept for their organization (Customer Lifetime Value is a prediction of the net profit attributed to the entire future relationship with a customer)
- Only 42% of companies are able to measure Customer Lifetime Value accurately.
- The probability of selling to an existing customer is 60 – 70%, while the probability of selling to a new prospect is 5-20%.
- Existing customers are 50% more likely to try new products and spend 31% more, when compared to new customers (Saleh K, 2017).

Customer churn rate is a metric that measures the percentage of customers who end their relationship with a company in a particular period. The more customer churn rate increases the more revenue and prestige loss will occur.

3.2. Project Objectives and Scope

This project aims to detect Vodafone postpaid consumer subscribers who are likely to churn. The subscribers that were investigated as a part of this project were active at the beginning of May'17. Some of these subscribers disconnected via MNP in May'17. By looking at the activities of all these subscribers during the last year the features that are effected in churn will be examined. In this study will be subscriber based and it may not lead to customer based results. For instance, a customer may have more than one line simultaneously and he may not be satisfied with one of them. This situation does not point to a dissatisfaction with the company but it may demonstrate a dissatisfaction with the subscription.

The Dataset extracted from Vodafone DWH Databases and it has not been manipulated and extracted randomly. These tables may not contain the whole activity of a subscriber and therefore, they do not contain the factors that are outside of the company that effect the churn decision of a subscriber. These factors that are outside of the company inherit situations such as emergence of a new campaign or discount in a particular tariff.

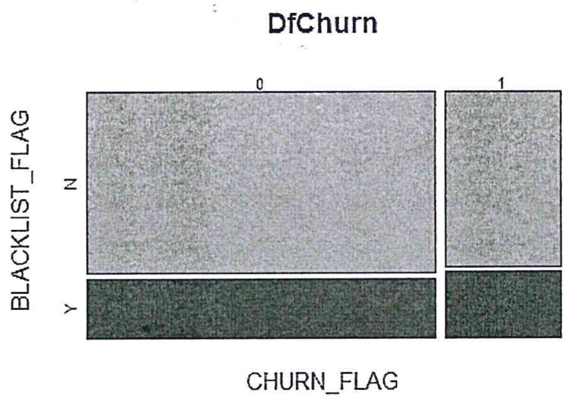
As an outcome of the project a model that predicts the churn tendency of a subscriber is developed. This model constructed and tuned with a dataset which is based on condition during May'17.

4. METHODOLOGY

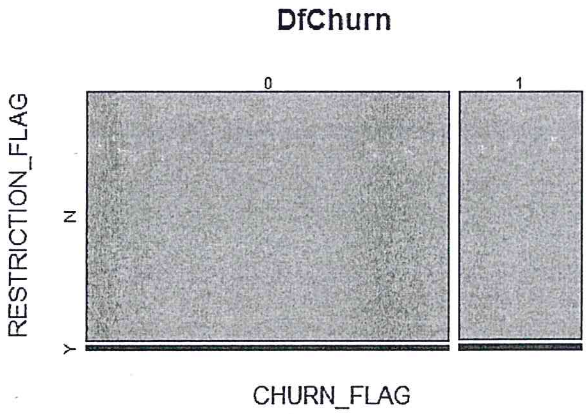
4.1. Exploratory Data Analysis

10 features were explored for feature selection. The selection criteria was about a strong positive or negative correlation between target (Churn_flag) and their own value. Exploratory data analysis and excel pivot tables were used for this process.

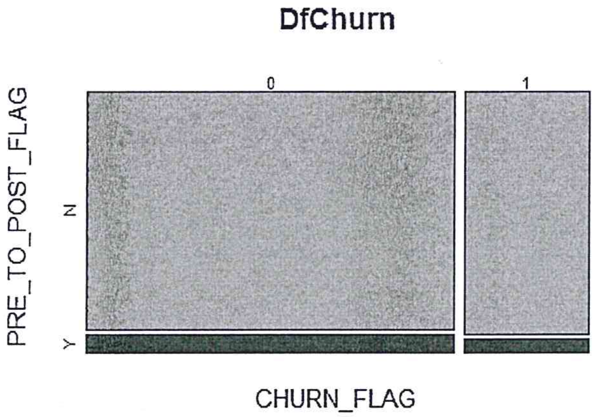
Black List Flag: There is a weak positive correlation



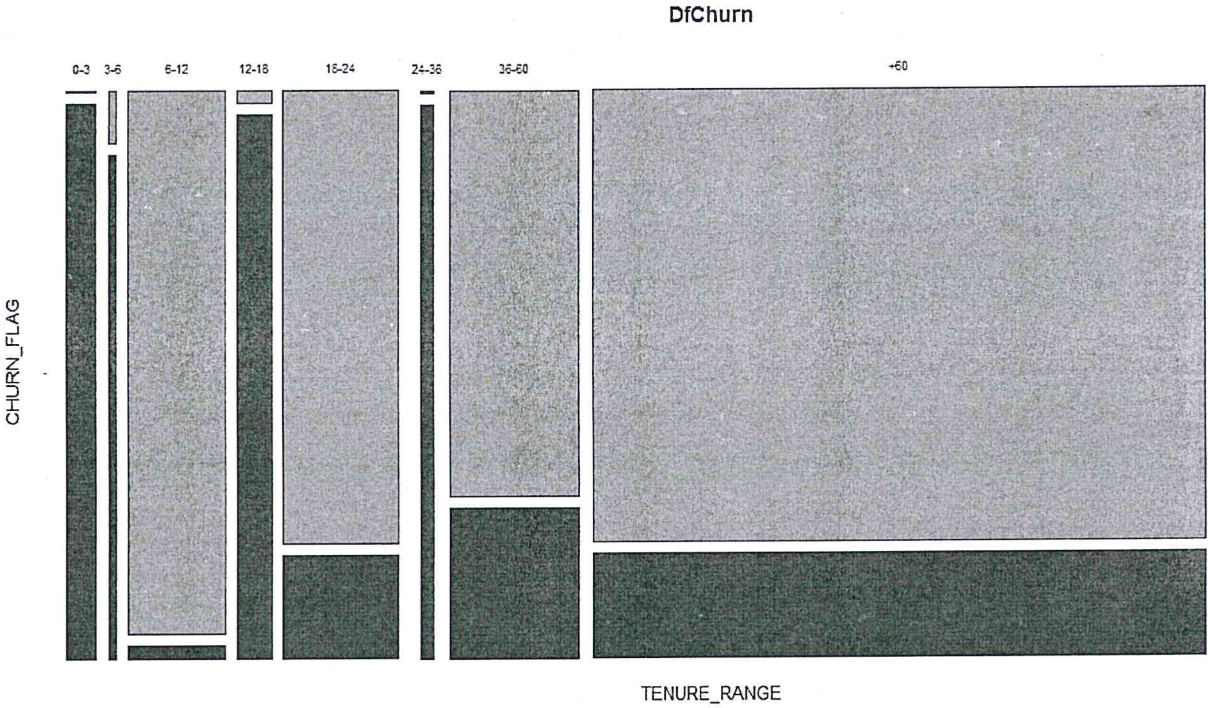
Restriction Flag: Weak correlation



Pre to post Flag: It has about weak negative correlation with Churn Flag

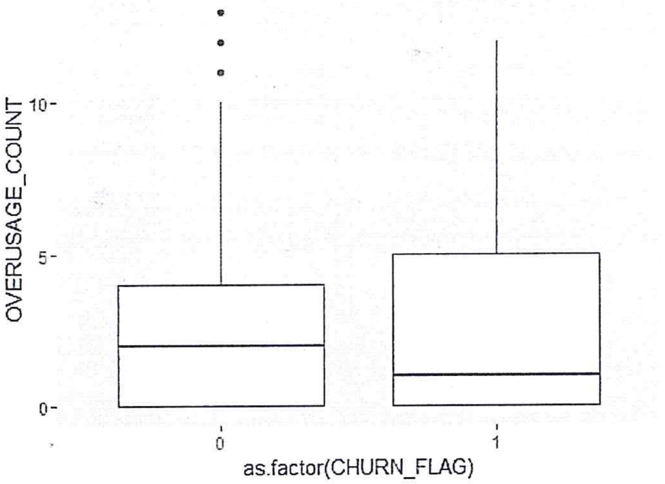


Tenure: There is a strong relation between tenure binds and churn. Subscribers who are in 0-3, 12 - 18 and 24-36 months tend to churn. It seems subscriber who started the subscription with 12 / 24 months handset contract tend to churn after their contract ends. The new subscribers whose tenure in 0-3 months might be disconnected because of the network and call quality problems in the districts they live in.

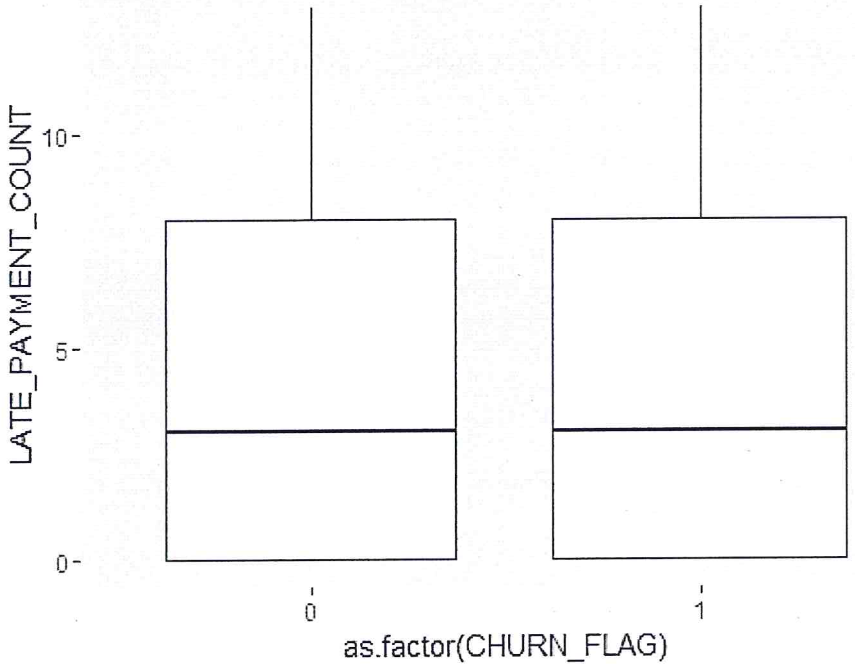


The items above were categorical and examined with mosaic plot.

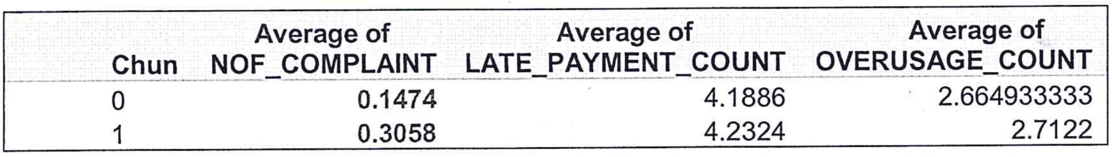
Overusage Count:



Late Payment Count: Late payment count does not effect churn flag explicitly.



NOF Complaint: Boxplot doesn’t work well with NOF_COMPLAINT attribute, a pivot table created in excel. As seen below, there is a strong relation between NOF_COMPLAINT and Churn.

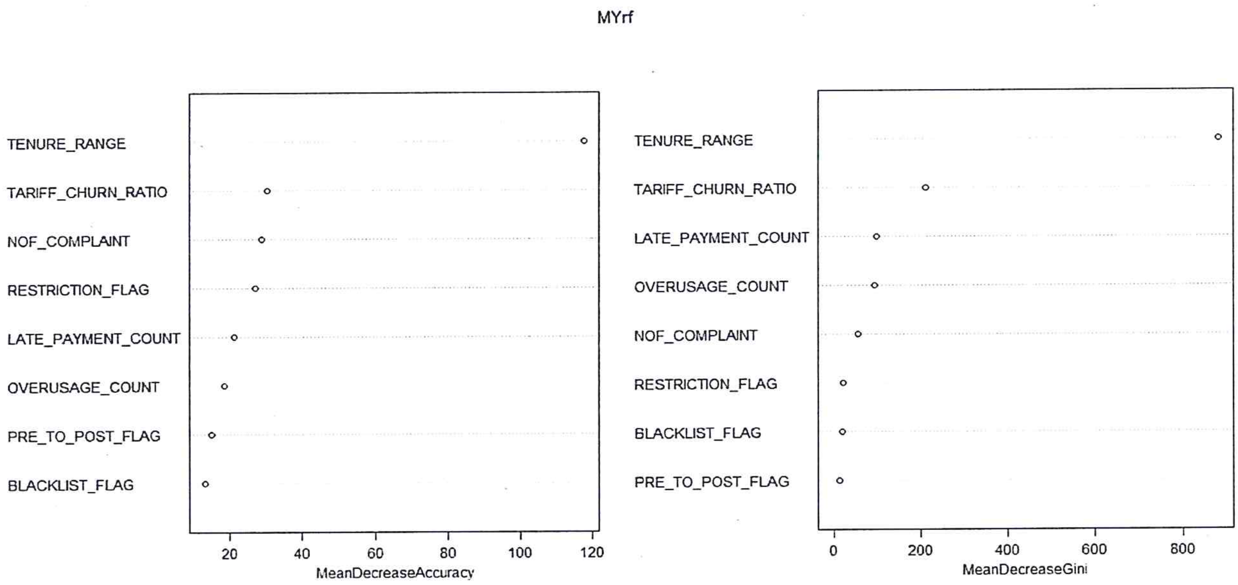


Chun	Average of NOF_COMPLAINT	Average of LATE_PAYMENT_COUNT	Average of OVERUSAGE_COUNT
0	0.1474	4.1886	2.664933333
1	0.3058	4.2324	2.7122

4.2. Feature Selection

To see how much effect which attribute has, a varImpPlot function was used. It is a Dotchart of variable importance as measured by a Random Forest.

- Using all features in the random forest model:



```
OOB estimate of  error rate: 16.88%
Confusion matrix:
      0    1 class.error
0 11233   25 0.002220643
1  2507 1235 0.669962587
```

- The most ineffective feature is excluded (Blacklist Flag):

```
OOB estimate of  error rate: 16.93%
```

```
Confusion matrix:
      0    1 class.error
0 11219   39 0.003464203
1  2501 1241 0.668359166
```

- 2 of the most ineffective features are excluded:

```
OOB estimate of  error rate: 16.89%
Confusion matrix:
      0    1 class.error
0 11189   69 0.006128975
1  2464 1278 0.658471406
```

- 3 of the most ineffective features are excluded:

```
OOB estimate of  error rate: 16.97%
Confusion matrix:
      0    1 class.error
0 11185   73 0.006484278
1  2473 1269 0.660876537
```

- 4 of the most ineffective features are excluded:

```
OOB estimate of  error rate: 16.98%
Confusion matrix:
      0    1 class.error
0 11202   56 0.004974241
1  2491 1251 0.665686799
```

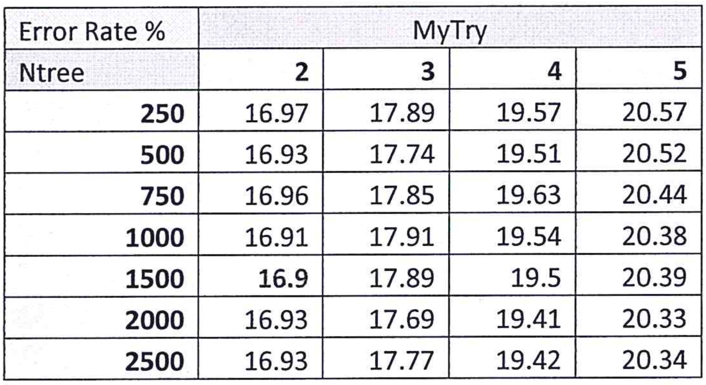
- 5 of the most ineffective features are excluded:

```
OOB estimate of  error rate: 17.18%
Confusion matrix:
      0    1 class.error
0 11159   99 0.008793747
1  2478 1264 0.662212720
```

Attempt 3 (2 of the most ineffective features are excluded) error rate is almost same as Attempt 1 error rate, Furthermore, Attempt-3 has the best churn prediction result. Therefore, Attempt-3's features are used in the model.

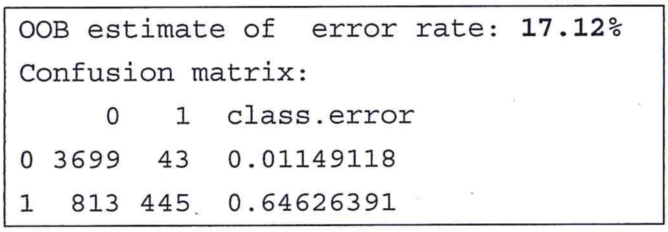
4.3. Model Development and Optimization

As the best practice, Random Forest and Decision Tree algorithms work well with the models such as churn and fraud analyses. First, a random forest model was built in R. After creating a train dataset, random forest run on it with ntree and mtry parameters values as seen below. The model which is run with Mtry = 2 and Ntree = 1500 got lowest error rate result. Mtry is Number of variables randomly sampled as candidates at each split and Ntree refers to number of trees to grow in the model.



Error Rate %	MyTry			
Ntree	2	3	4	5
250	16.97	17.89	19.57	20.57
500	16.93	17.74	19.51	20.52
750	16.96	17.85	19.63	20.44
1000	16.91	17.91	19.54	20.38
1500	**16.9**	17.89	19.5	20.39
2000	16.93	17.69	19.41	20.33
2500	16.93	17.77	19.42	20.34

The model also run with test dataset with same features and random forest optimization.



```
OOB estimate of  error rate: 17.12%
Confusion matrix:
     0   1  class.error
0 3699  43  0.01149118
1  813 445  0.64626391
```

Tenure is the most dominant feature in churn and the model error rate gets higher when it runs with the test data. It may cause overfitting that can lead to poor model performance. Overfitting is when a model is able to fit almost perfectly the training data but is performing relatively poorly on new data. The model is run on new features set whose tenure feature is excluded and got the lowest error rate in previous attempts.

```
OOB estimate of  error rate: 23.93%
Confusion matrix:
      0    1 class.error
0 11046 212  0.01883105
1  3377 365  0.90245858
```

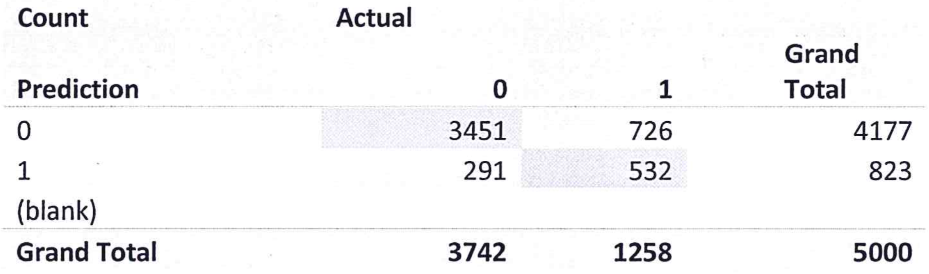
It also run with the set which includes all features without the tenure.

```
OOB estimate of  error rate: 23.79%
Confusion matrix:
      0    1 class.error
0 11158 100 0.008882572
1  3469 273 0.927044361
```

As clearly seen above, error rate of the model increased significantly when tenure feature is excluded. For this reason, the exclusion of tenure feature in the model is not a good idea as it increases the error rate from 16.89% to 23.93%.

Therefore, the test model error rate is 17.12% and train model error rate is 16.89%. There is no significant difference between them and this does not yield to an explicit overfitting.

After random forest run with test dataset, the submit dataset was created which has prediction and actual churn value. It is loaded in an excel file and a pivot table was created for the results.



Count	**Actual**		
Prediction	**0**	**1**	**Grand Total**
0	3451	726	4177
1	291	532	823
(blank)			
Grand Total	**3742**	**1258**	**5000**

5. RESULTS

5.1. Overview of the results produced

This study started with an interest on the investigation of features that can effect churn of Vodafone postpaid consumer subscribers. Within this scope eight subjects were determined and via using PL-SQL these subjects were transformed into derived features. Subsequently, these features were examined with exploratory data analysis and their possible effects on the churn model are studied. These derived features were further elected through the monitoring of the model's performance and six features were designated for churn analysis. As a supervised machine learning algorithm random decision forest is used in the project. By chancing the parameters that are peculiar to random forest, the model performance is tuned.

75% of the whole dataset is separated randomly for train dataset. The model which was executed and tuned on the train dataset has an error rate of 16.9% estimation. This is the best error rate that was obtained during the train process. Subsequently, the same tuned model was run on the test dataset which is 25% of the whole dataset and got 17.12% error rate estimation. When this model was executed on the test dataset which has 5000 observation, approximately churn decision of 4000 subscriber out of 5000 were predicted correctly. Within these group of 4000 subscribers, the model was more successful in detecting the subscribers who did not churn.

When the features that effected the churn decision were examined, it was observed that the most effective feature was tenure. It is estimated that, subscribers whose contract ends are more likely to churn. The second most effective feature on churn decision is tariff churn ratio which is calculated by "MNP count in the month" divided by "average number of active subscriber in the tariff". This ratio calculates the speed of churn in the subscriber's tariff. The tariffs with a high churn ratio may refer to a campaign generated by rival companies that attract the subscribers using the tariff. The analysis also demonstrate that subscribers with a high complaint number are more likely to churn.

5.2. Analysis/Evaluation of expected and obtained results

When the essential features investigated in the beginning of the study the conditions that lead to churn were roughly estimated. When the analysis was executed it was observed that this was estimation was correct. However, the high impact of the tenure as the dominant feature was unexpected outcome.

The author did not have any predictions on the success rate on the model. The success rate was materialized during the model development and tuning.

6. VALUE DELIVERED

- As mentioned 3.1 problem statement the cost of keeping an existing subscriber is generally cheaper than the cost of acquisition of a new subscriber. This analysis will decrease revenue loss due to the churn of existing subscribers.

- The number of subscribers is very important for a leading company in the sector such as Vodafone Turkey. It will contribute to improve the company prestige.

- The analysis can implicitly increase the loyalty of subscribers by convincing the subscribers who are likely to churn to not leave the company.

As Vodafone already embodies the data and development and analysis tools such as like R, Microsoft excel, Oracle PL-SQL used in this project, the project has a high applicability with a low budget.

7. APPENDIX A

This project is composed of two sections: extract and derived features development via PL-SQL and model development and run in R. The R model and PL-SQL codes must be run or scheduled monthly. The data and order of importance of the features might change on a monthly bases. Furthermore, error rate of the model must be monitored after each model run and if it is necessary, the features that lost their importance must be excluded from the model.

8. REFERENCES AND CITING

Customer Retention. (2017). Retrieved from https://www.inc.com/encyclopedia/customer-retention.html.

Saleh K. (2017). Customer Acquisition Vs.Retention Costs – Statistics and Trends. Retrieved from https://www.invespcro.com/blog/customer-acquisition-retention/