MEF UNIVERSITY

CHURN PREDICTION OF A DEAL E-COMMERCE WEBSITE CUSTOMERS

Capstone Project

Müge Çevik

İSTANBUL, 2017

Thank you for all support of my lecturers at MEF University Big Data Analysis Programme and many thanks to my father and mother who always supported me.

Müge Çevik

MEF UNIVERSITY

CHURN PREDICTION OF A DEAL E-COMMERCE WEBSITE CUSTOMERS

Capstone Project

Müge Çevik

Advisor: Asst. Prof. Dr. Hande Küçükaydın

İSTANBUL, 2017

EXECUTIVE SUMMARY

CHURN PREDICTION OF
A DEAL E-COMMERCE WEBSITE CUSTOMERS

Müge Çevik

Advisor: Asst. Prof. Dr. Hande Küçükaydın

SEPTEMBER, 2017, 36 pages

Today, there is a lot of deal e-commerce sites which are essentially marketplaces. They provide deals which are offered by merchandisers. Because of the nature of these sites there is no subscription model; customers continue because of price or interest or quality not because of subscription.

It is normal to have some customers who stop buying, which is defined by "churn". Data mining is now a new technique to define "churned" customers and to have prediction who will churn and what should be against.

In this project customers are clustered via unsupervised clustering technique for clusters as "newly purchased", "frequently purchased" and "mostly payed" and "churned" . Random Forest Classifier is used to prove that the "churned" customer clusters have homogeneous character and also it has been proved that the "churned" labelled customers have actually no deal order after the observed time period.

To recommend what should be done to regain the churned customers to the site the deal order history of these customers have been explored and the deal categories from which they have bought have been found.

Key Words: Data mining, e-commerce churn, unsupervised clustering, Random Forest Classification, visualisation of data

ÖZET

BİR FIRSAT E-TİCARET SİTESİNİN KAYBEDİLMİŞ MÜŞTERİ TAHMİNİ

Müge Çevik

Tez Danışmanı: Yrd. Dç. Dr. Hande Küçükaydın

EYLÜL, 2017, 36 Sayfa

Bugün, temelde bir pazar yeri olan birçok fırsat e-ticaret sitesi var. Bu siteler mağazaların ve dükkanların sundukları fırsatlar gösterirler. Bu sitelerin doğası gereği abonelik modeli yoktur, müşteriler fiyat veya ilgi veya kalite nedeniyle kalır, abonelik nedeniyle değil.

Bazı müşterilerin alışverişi bırakması normaldir ki bunlar "kaybedilmiş" olarak tanımlanır. Veri madenciliği teknikleri, "kaybedilmiş" müşterileri tanımlamak, hangi müşterilerin "kaybedileceğini" tahmin etmek ve buna karşı ne yapılması gerektiğini bulmak için yakın zamanlarda kullanılmaya başlanmıştır.

Bu projede, müşteriler denetimsiz kümeleme tekniği kullanılarak"yeni satın almış", "sık satın almış" ve "en çok para ödemiş" ve "kaybedilmiş müşteri" kümelerine bölünmüştür. Kaybedilmiş müşteri sınıflarının karakteristiğinin homojen olduğunu kanıtlamak için Rastgele Orman Sınıflandırıcısı kullanılmıştır, ayrıca "kaybedilmiş müşteri" etiketli müşterilerin gözlem yapılan zaman periyodu sonrasında hiçbir fırsat satın alımı gerçekleştirmediği de ispatlanmıştır.

Bu projede, "kaybedilmiş" müşterileri siteye geri kazanmak üzere ne yapılması gerektiğini önermek için, bu müşterilerin fırsat sipariş geçmişleri keşfedilmiş ve satın aldıkları fırsatların kategorileri bulunmuştur.

Anahtar Kelimeler: Veri madenciliği, e-ticarette kaybedilmiş müşteri, denetimsiz kümeleme, Rastgele Orman Sınıflandırıcısı, veri görselleştirme

TABLE OF CONTENTS

.

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

In all e-commerce websites "Churn" is defined as to have the "paying" customers lost. This "churn value" can be easily found at recurring payment businesses but it is not so easy to find the churned customers when there is no recurring payment. Furthermore to "predict" the customer who "will" churn one should have a "model" that gives a prediction so that it can be prevented.

To predict churn is important because "it is easier to sell to an existing customer than to attract and sell a new one. The loyal customers are less sensitive to price changes. Statistics show that it costs five to six times more to gain a new customer than to keep the existing ones" as stated in paper by Hudaib et al. (2015) [3].

The deal e-commerce site is an e-commerce website where customers pay for deals when there is a big discount offered. So there is no recurring payment. While some customers use the website frequently, others pay only one time in their customer lifetime. Furthermore there also exist subscribed users who click on the deals but have no purchase in their customer lifetime.

1.1. Objective of the Project

The objective of this project is twofold:

1. To find customer segments where customers are still actively buying, where customers are once buying and then "churned" and where users are only "watchers" for deals.
2. To define the model for the "churn" prediction and to validate the model.

According "Retentionscience" [1] customer churn is defined for different business models:

1. "In a pure subscription model, without postponement, customer churn is simply users who have unsubscribed."
2. "In ad-hoc purchase models, however, such as most e-commerce sites, customer churn is defined as customers who stop being paying customers."

3. "It may involve defining churners statically, like people who haven't purchased for a time period that is a few standard deviations away from the average purchase time, or someone whose time on the site without purchase far exceeds the average customer lifetime."

Thus, defining churned customers depends on business type as well as on definition type. One can create customers by unsupervised clustering techniques whereas others use classic statistics. In Section 1.2 related studies are further investigated.

1.2. Literature Review

In literature "churn" problem is mostly investigated through recurring paying customers where only classification methods are used. For example, in the famous KDD-Cup for data science projects, training and test data of a telecommunication company is given where churned values of customers are known [2]. The nature of subscribed customers for a telecommunication company is recurring payment, so it is certain to label people who have churned or not.

In this project there is no recurring payment. So, similar problems and the solutions for that problem have been further investigated.

To predict churning customers of a business, where recurring payments occur, mostly supervised classification algorithms are used. In a new study, a hybrid model is used: In the study by Hudaib et al. (2015) [3] customers are segmented through clustering algorithms at first, then classification method is used. They explain their method as : "First, the clustering algorithm is applied to filter the data from outliers and unrepresentative behaviours. The resulted filtered data of the clustering algorithm become the input to the classification algorithm. Then the classification algorithm uses these data in the learning process and builds the final churn prediction model."

Using clustering before classification algorithm has increased the accuracy according to their paper.

Another approach is to cluster customers by "Recency, Frequency and Monetary metrics.".

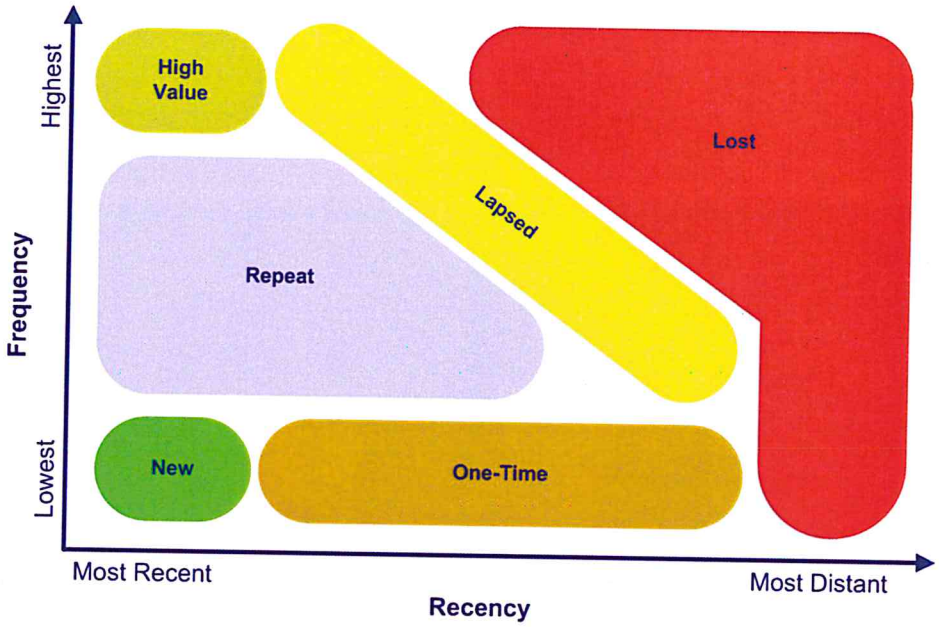

Figure 1. Recency- Frequency Chart Diagram for Customer Segments

As explained in presentation by Porzak (2008) [4] "Recency" is defined as count of days since last purchase of customer until the day of modelling. "Frequency" is defined as the count of orders in analysis period of modelling. "Monetary" is defined as total dollar value of all orders in analysis period of modelling.

In Figure 1 above the customer segments by "Buying Recency" vs "Buying Frequency" are seen. People who have bought most recent and one time are "new" customers. People who have bought one time and then do not buy for long time are labelled as "one-time customer". People who bought more than once and bought in recent time are labelled as "Repeat". People who buy very frequent and bought newly are labelled as "High Value". People who bought long time ago are labelled as "Lost".

Figure 1 shows only an overview; there is no exact border to define customer segments. Defining customer segments as "repeat", "high value", "lost" etc. depends on the business type and the loyalty of all customers. Therefore, it should be handled specific to business.

To understand customer behaviours customer life cycle for e-businesses should be taken into account: In Master Thesis of Jahromi (2009) [5] customer life cycle is represented as in Figure 2 below:

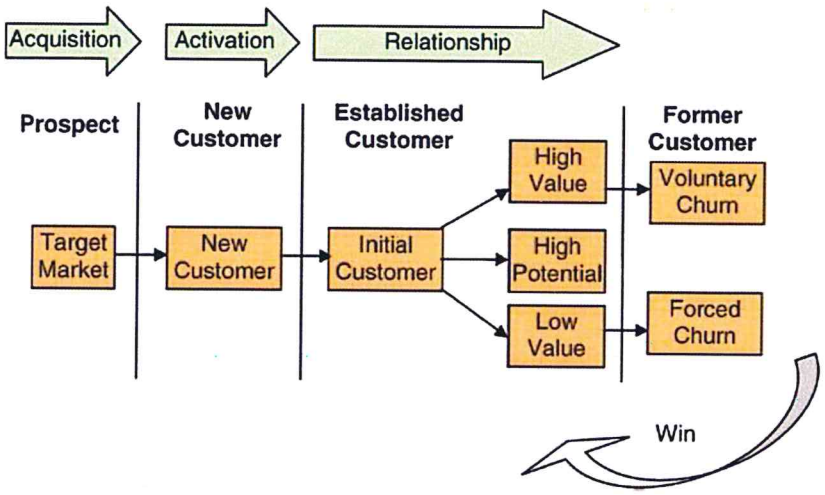

Figure 2 Customer life cycle in general

In e-businesses in general people are first attracted to visit the site. A percentage of these people are subscribed and become users. A percentage of these users buy something and become customers. They may be High Value, High Potential or Low Value customers. Website makes some offers to gain churned customers. Website can prefer to make offers former High Value customers or all churned customers.

The thesis of Jahromi (2009) [5] tries to find out churned customer in pre-paid mobile phone sector. Due the nature of this business there is no subscription. Customer data are clustered based on their usage behavioural features (Recency- Frequency-Monetary) and then classification techniques are used.

The author creates six clusters of customers based on the recency and frequency of call behaviour, such as "Call ratio", "Average Call Distance" etc. For each developed cluster "Average Max-Distance" is calculated which is defined as the maximum count of days between two calls.

In the second step different features such as "Minutes of Use" and "Frequency of Use" and their change between periods are calculated for every cluster. For every cluster classification algorithms such as decision tree and neural networks are used to predict churned customers. The accuracy is up to 90% in this study.

2. PROJECT DEFINITION

In this project, customer data of a group-deal e-commerce site are used. This website has been founded in 2010 and the data used in this project are selected from database with condition of recording between the dates 01.10.2015 - 17.03.2017.

Group-deal e-commerce sites offer one-time deals which are discount prices for a certain period for businesses like restaurant, dance course, theater, spa, concert, holiday tours etc. In its nature there is no subscription for the customers, they can buy whenever they like. However there is a lot of group deal e-commerce sites, so when a customer has bought one or more deals and stops buying, it may mean the customer has churned and buys now from another group-deal e-commerce site.

So how will we understand if the customer has churned or not ? What should be our "churn definition" ?

2.1. Churn Definition for Deal E-Commerce Web Site

The data are big enough to split the data into four parts as seen in Figure 3:

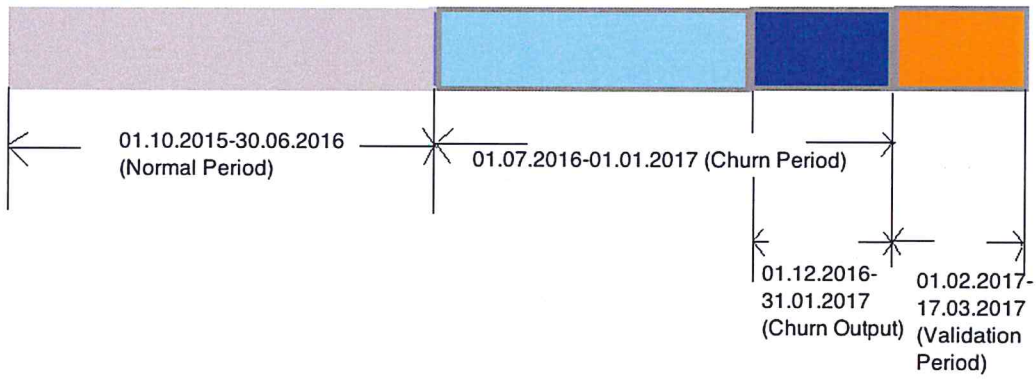

Figure 3 Segmentation of time period of the dataset

• Normal period : User has subscribed and has made order(s). It is possible that the user has subscribed and does not have any order. This period has been defined for the first nine months.

• Churn Period: Customers who have made order(s) in the normal period can have more orders than normal period or less orders than normal period. We consider the customers who have less orders than normal period. This period has been defined for the following seven months.

• Churn Output : Last two months of "Churn Period". Customers who has less orders in the Churn Period are expected not to have any order or even any action click in the website.

• Validation Period: Next two months after Churn Output. Customers who are defined as "churned" are expected not to have any order or even any action click in the website.

Customers who are registered in the first seven months (01.10.2015 - 30.04.2016) are selected for observation. Deal order counts and action clicks at the website of the deals in the normal period and in the churn period of these customers are considered. The main idea is to define the customers as "churned" who have less orders and deal action clicks in the churn period than in normal period.

2.2. Describing Data Set

The data set is totally about 1 GB. It is structured in tables which are:

A	B	C	D	E	F	G	H
id	id	id	id	id	id	id	id
fCityId	fMemberId	fMemberId	fDealOrderId	fMerchantId	fParentTagId	fDealId	fMemberId
Created	fDealId	fDealId	hasUsed	fCityId	Title	fTagId	fPublisherId
	actionKey	Quantity	isCanceled	realPrice			
	Created	totalPrice	created	dealPrice			
		Platform		created			
		Created					

Table 1 Data set tables

In Table 1 the database tables of the data are seen:

Column A: Member table. Member id, city of member, registration date

Column B: Member Deal Action table. Click type of member in deal page like "click to purchase", "cancel" etc.

Column C: Deal Order table. Who purchased which deal, how many coupons of the deal, what is the total price which is paid, what was the platform of buying (pc or mobile), the date of the purchase.

Column D: Coupon Code table. Member use coupons at the merchandisers. This table shows if coupon is used or cancelled and the deal id to which coupon belongs.

Column E: Deal table. Information about deals. The merchandiser of deal, the city where deal is offered, the real price and deal price of offer, creation date of deal.

Column F: Tag table. Parent tag id and title of tag.

Column G: Deal Tag table. Deal id and tag id of the deal

Column H: From which platform is the user redirected to the site (Google, Facebook etc.). If the user is registered, the member id is stored. If the user is not registered the member id is stored as zero value.

All data acquired are data of which "created" column is between the dates 01.10.2015 and 17.03.2017.

Since this deal e-commerce is 7 years old, the deal order and click tables in the data set have data from the members who have registered before 01.10.2015.

2.2. Problem Statement

Problem is to find out customers who have registered between 01.10.2015-30.04.2016 at deal e-commerce web site, have at least one order and churned after a while.

Churned customers are also clustered as "high value" and "low value" customers. "High value" customers are users who had many orders in normal period and then churned, "Low value" customers are users who had some orders in normal period and then churned. In this project, a prediction model is created to predict customers which will churn in churn period.

2.2. Project Objectives

Project objectives are to find customers who churned and to cluster churned customer as "high value" and "low value". The objective to cluster them as "high value" and "low value" is to give the opportunity to website owner to create special campaigns for customers.

2.3 Project Scope

Main objective of the project is to create a model for the prediction of customers who are registered between 01.10.2015 - 30.04.2016, have at least one order and made no payment after 01.12.2016.

Project does not give any model for the customers who did not registered between these dates.

3. USING UNSUPERVISED CLUSTERING TO DEFINE CUSTOMER SEGMENTS

In the data set there are a lot of tables defined; we have a lot of features of customers, deals and orders, but they don't give any idea without processing. First step is to create independent and dependent variables of prediction model using appropriate data in the data set.

3.1. General Methodology of Data Mining

In general, these steps should be followed for Data Mining as also mentioned in Jahromi (2009) [5]:

- Data Collection : Data can be gathered from databases or from text files, can be retrieved as sound and image files.
- Data Selection: Selecting data which are appropriate for the project. This step may require joining multiple data sources.
- Data Pre-Processing : Eliminating or modifying examples from the selected data which are either noisy, inconsistent or having too much missing values. In this step missing values may be imputed with appropriate values.
- Transformation: Data are transformed and consolidated into forms appropriate for the modelling step.
- Modelling: Selection of the algorithm. If there are labelled data, which means if output is known, supervised algorithms are used. If the output is unknown, unsupervised algorithms are used. If the output is numerical, regression algorithms are used. If the output is categorical, classification algorithms are used.
- Interpretation / Evaluation : Visualising and interpreting the discovered knowledge for the user and evaluation of the discovered information with respect to validity, novelty, usefulness and simplicity.

3.2. Preparing Data for Data Mining Step

3.2.1. Data collection

Data set is retrieved in ".sql" file format from website owners and then transformed into ".csv" file format via "MySQLWorkbench" program.

3.2.2. Data selection

Data set is so big and has great variance which leads to think several outputs for many projects, but to define churn is simple: Churned customers will have output as one and non-churned customers will have output as zero.

As stated in Section 1.2 customers are mainly clustered in e-businesses to "Recency-Frequency-Monetary" clusters. To find recency of customers we need "Deal Order" and "Deal Action Click" tables. To find frequency of customers we need "Member", "Deal Order" and "Deal Action Click" tables. To find monetary of customers we need "Deal Order" tables.

3.2.3. Data pre-processing

There is no missing value in the tables, but big outliers have been removed for deal action clicks.

As we see in Figure 4 below there are very big outliers for deal action clicks in normal period. Normal period is 9 months which are 270 days. Even 10 click a day gives 2700 clicks. Clicks which are higher than 2000 may be realised through programs.

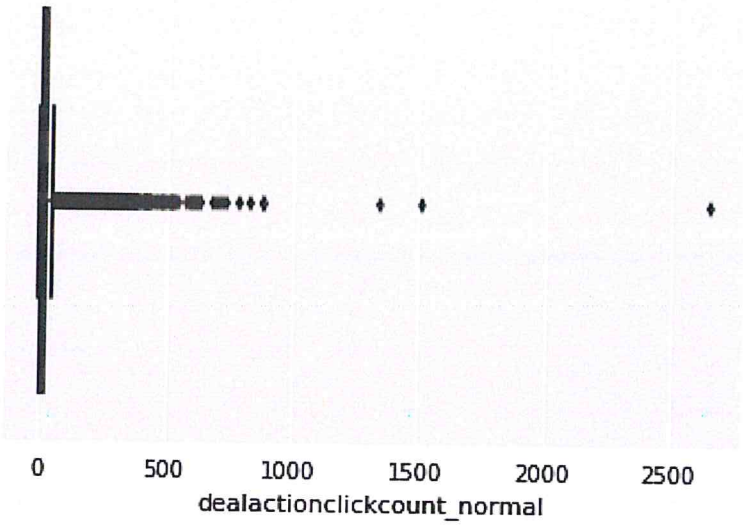

Figure 4 Box-plot for deal action click count in normal period (x:deal action click count in normal period, y: count of people)

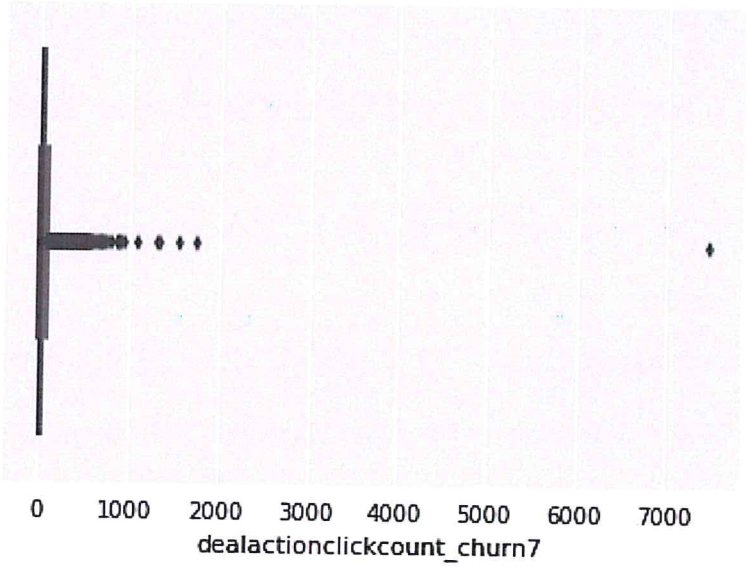

Figure 5 Box-plot for deal action click count in churn period (x:deal action click count in churn period, y: count of people)

3.2.4. Data transformation

For training and test data, members who are registered 01.10.2015 - 30.04.2016 are considered. "Member tenure" is defined as count of days from the registration date of member until 17.03.2017.

Day of counts of first order until the last day of training data is defined as "first deal order tenure". Day of counts of last order until the last day of training data is defined as "last deal order tenure". Day of counts of last deal action click until the last day of training data is defined as "last deal action click tenure".

Monthly average and monthly standard deviation of the order counts and deal action click counts are calculated. Tenure of last deal orders, first deal orders and last deal action clicks are also calculated.

3.3 Modelling

Since we do not know which customers have churned, we should make first clustering. For this purpose, "K-means" algorithm is selected. K-means algorithm is very fast and suitable for big data but the performance is dependent to the appropriate "K" value. It should not be very low and also it should not be very high. To find the optimum "K" value,"Elbow" method is used.

Elbow method is to run k-means clustering on the data set for a range of values of "K" (maybe 1 to 50) , and for each value of *k,* we calculate the sum of squared errors (SSE) as stated in [7]. The lowest K value which gives lowest SSE for clusters is selected as "K" value for the clustering algorithm.

3.3.1 Exploring the data

Before applying "K-means" algorithm some visualisations of data have been made to see what to expect.

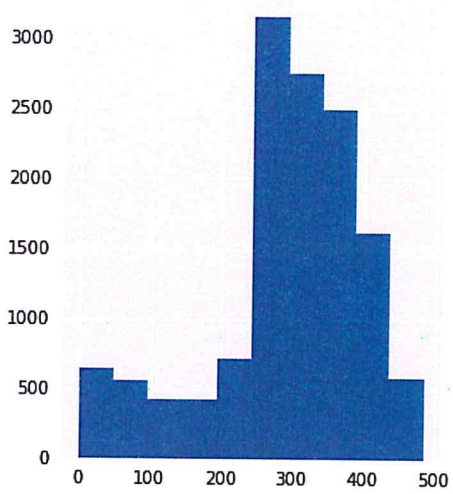

Figure 6 Histogram for last deal order tenure (x: last deal order tenure, y: count of people)

In Figure 6 we see that half of the customers have last deal order tenure more than 300 days. People who have low last deal order tenure are active customers.

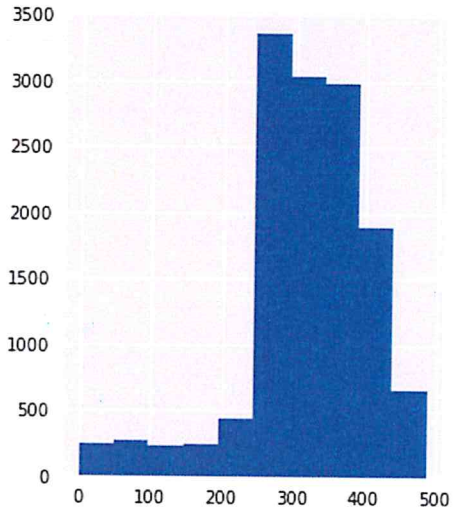

Figure 7 Histogram for first deal order tenure (x: first deal order tenure, y: count of people)

In Figure 7 we see again that more than half of the customers have last deal order tenure more than 300 days. These customers are most probably "one-time" customers.

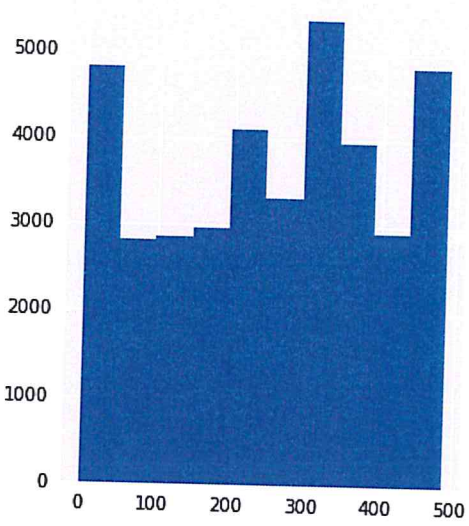

Figure 8 Histogram for last deal action click tenure (x: deal action click tenure, y: count of people)

Figure 8 shows histogram for last deal action click tenure. From Figures 6 and 8 we understand that there are many people who bought long time ago but are still actively visiting the website.

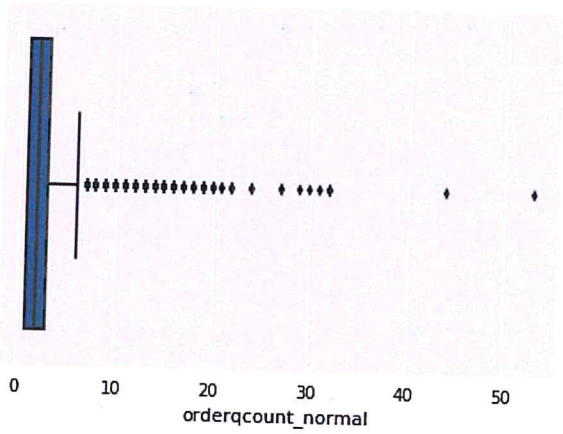

Figure 9 Box-plot for total order count for each customer in normal period (x: order count, y: count of people)

Figure 9 shows that there are outlier customers who make purchase in normal period but most of the customers have less than 5 orders in normal period.

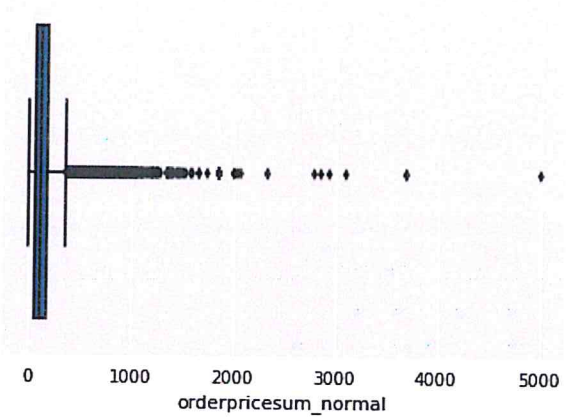

Figure 10 Box-plot for order price in normal period (x: payment of total orders, y: count of people)

Figure 10 shows that there are outlier customers who make purchase in normal period, but most of the customers pay less than 300 TL for deal coupons.

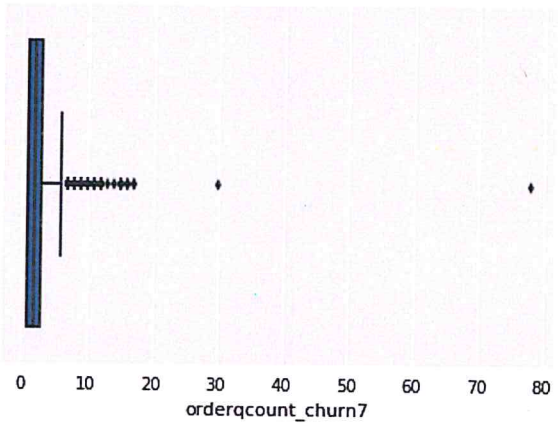

Figure 11 Box-plot for order count in churn (x: payment of total orders, y: count of people)

Figure 11 shows that there are outlier customers who make purchase in churn period, but comparing with corresponding figure of the normal period shows us there are less outliers.

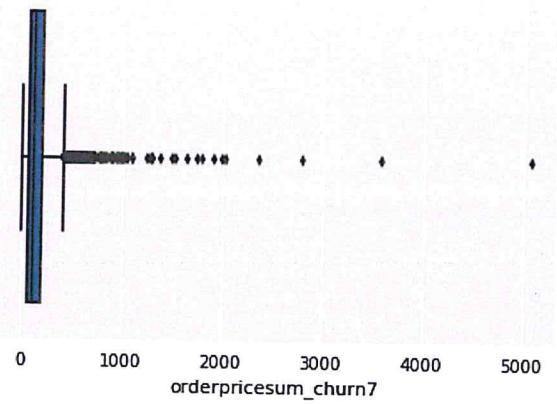

Figure 12 Box-plot for order price in churn period (x: payment of total orders, y: count of people)

Figure 12 shows that there are outlier customers who make purchase in churn period, but comparing with corresponding figure of the normal period shows us there are less outliers.

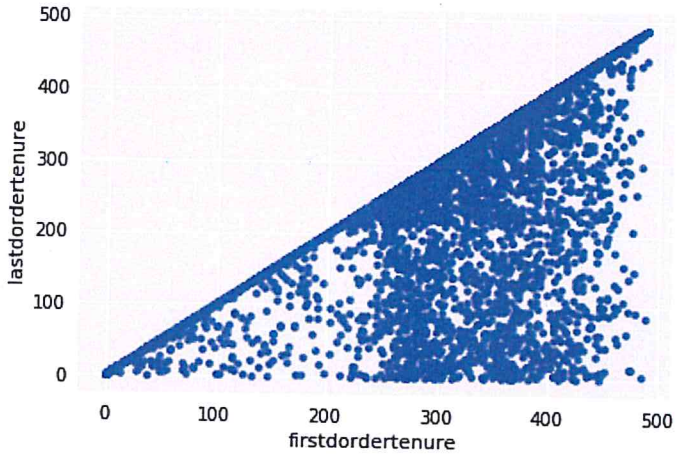

Figure 13 Scatterplot for first deal order tenure versus last deal order tenure for each customer

Figure 13 shows that people who are on the x=y line have only one order. We define here people who have high tenure as churned. How much "high tenure" should be is further investigated.

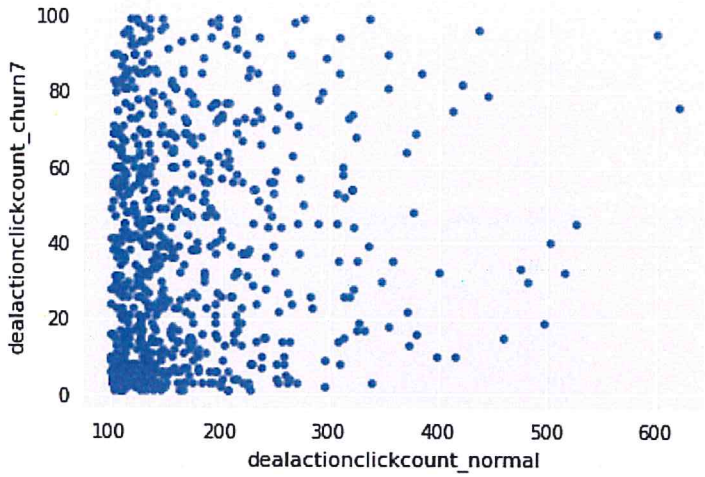

Figure 14 Scatterplot for deal action click count in normal period which are higher than 100" versus "deal action click count in churn period which are less than 100

Scatterplot for "deal action click count in normal period which are higher than 100" versus "deal action click count in churn period which are less than 100" can be seen in Figure 14. These people are most likely churned.

Scatterplot for "order count in normal period which are higher than 1" vs "order count in churn period which are less than 5" is seen in Figure 15. People who have high order counts in normal period but less or zero order count in churn period are most likely churned. Order count value "-1" in churn period means "no order" in this figure.

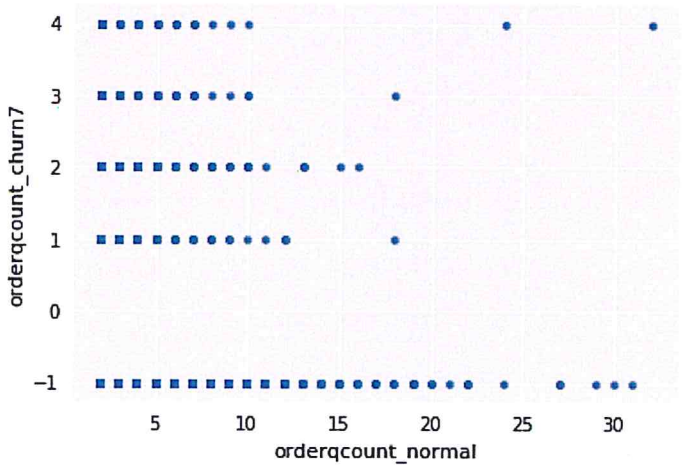

Figure 15 Scatterplot for "order quantity count in normal period which is higher than 1" vs "order quantity count in churn period which is less than 5"

3.3.2 Finding "K" value to create customer clusters

Training data should have independent variables to define clusters of customers. We define "churn" as it is in literature , people who buy less or zero in churn period than in normal period are churned. What the exact "less" should be depends on data. Hence, we should form these clusters and find the border to define "churned" customers.

"firstdordertenure"	First deal order tenure
"lastdordertenure"	Last deal order tenure
"lastdactionclicktenure"	Last deal action click tenure
"orderqcount_normal"	Order count in normal period
"orderqcount_churn7"	Order count in churn period
"orderqcount_churn5"	Order count in last 5 months of churn period
"orderqcount_churn3"	Order count in last 3 months of churn period
"orderqcount_churn2"	Order count in last 2 months of churn period
"orderpricesum_normal"	Payment for orders in normal period
"orderpricesum_churn7"	Payment for orders in churn period
"orderpricesum_churn5"	Payment for orders in last 5 months of churn period
"orderpricesum_churn3"	Order count in last 3 months of churn period
"orderpricesum_churn2"	Order count in last 2 months of churn period
"avgdealorderquantity_normal"	Average quantity of orders in normal period
"avgdealorderquantity_churn7"	Average quantity of orders in churn period
"avgdealorderquantity_churn5"	Average quantity of orders in last 5 months of churn period
"avgdealactionclick_normal"	Average quantity of deal action clicks in normal period
"avgdealactionclick_churn7"	Average quantity of deal action clicks in churn period
"avgdealactionclick_churn5"	Average quantity of deal action clicks in last 5 months of churn period

Table 2 Training data features

Applying PCA to the training data gives explained variance ratio for these features:

```
[ 5.16599212e-01  3.08936708e-01  9.25730161e-02  6.33902325e-02
  6.76620491e-03  5.01101133e-03  2.35092229e-03  1.98156183e-03
  1.32881192e-03  9.09446982e-04  1.26947503e-04  1.35319730e-05
  8.96774226e-06  1.39781642e-06  9.75989313e-07  5.33115941e-07
  3.39571639e-07  1.47847346e-07  3.07209543e-08]
```

The first 5 components of PCA output explains %98 of the training data, but to understand the features which form this PCA components we should calculate the correlation of training data features with PCA components.

Correlation of the original data with the first 5 components of PCA data gives:

PCA component 1	PCA component 2	PCA component 3	PCA component 4	PCA component 5
firstdordertenure 0.927882	lastdactionclicktenure -0.759079	firstdordertenure -0.611306	orderqcount_normal -0.620098	avgdealorderquantity_churn7 -0.569174
lastdordertenure 0.939720		lastdordertenure -0.707736	orderqcount_churn5 0.511061	avgdealactionclick_churn7 -0.573656
lastdactionclickte nure 0.657135		orderqcount_churn7 0.803217	orderqcount_churn3 0.508681	avgdealactionclick_churn5 -0.521741
orderqcount_normal 0.872147		orderqcount_churn5 0.834819	orderqcount_churn2 0.507680	
orderqcount_churn7 -0.550074		orderqcount_churn3 0.850149	orderpricesum_normal -0.812736	
orderqcount_churn5 -0.609956		orderqcount_churn2 0.843841	orderpricesum_churn7 0.597917	
orderqcount_churn3 -0.656033		orderpricesum_churn7 0.908647	orderpricesum_churn5 0.623295	
orderqcount_churn2 -0.670511		orderpricesum_churn5 0.923780	orderpricesum_churn3 0.615516	
orderpricesum_normal 0.682076		orderpricesum_churn3 0.905375	orderpricesum_churn2 0.605945	
orderpricesum_churn7 -0.594135		orderpricesum_churn2 0.876350	avgdealorderquantity_normal -0.594441	
orderpricesum_churn5 -0.656594		avgdealorderquantity_normal -0.536615	avgdealorderquantity_churn5 0.503499	

PCA component 1	PCA component 2	PCA component 3	PCA component 4	PCA component 5
orderpricesum_churn3 -0.696387		avgdealorderquantity_churn7 0.701346		
orderpricesum_churn2 -0.698592		avgdealorderquantity_churn5 0.751830		
avgdealorderquantity_normal 0.907275				
avgdealorderquantity_churn7 -0.501853				
avgdealorderquantity_churn5 -0.576935				

Table 3 Correlation values of training data features with PCA components.

Every feature except "Member Tenure" are correlated to PCA components more than 50% so they have high explanation of training data, thus taking all features but "Member Tenure" would be good to form clusters.

To find "K" value which should be a parameter for K-Means algorithm the "Elbow" method is used. "Elbow diagram" plotted for this training data can be seen in Figure 16.

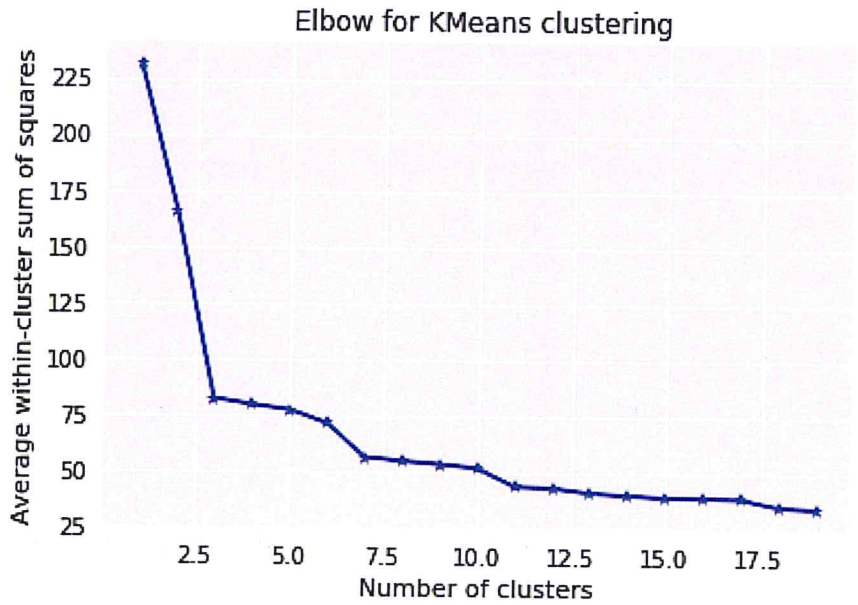

Figure 16 Elbow diagram for training data set

Figure 16 shows "Number of clusters" vs "Average within-clusters sum of squares". This means as "Average within-clusters sum of squares" gets lower clusters gets more homogeneous. 7 clusters seem highly homogeneous. Taking too high cluster number makes the evaluation more difficult. So, 7 seems very appropriate for this project.

3.3.3 Creating customer clusters with K-Means algorithm

Giving the data and cluster number to the K-Means algorithm results to have cluster centres and count of members in every cluster. So the count of member in each cluster are:

cluster number	member count in cluster
1	10323
2	6196
3	11252
4	31667
5	3927
6	433
7	30

Table 4 Member counts in K-means clusters

	1	2	3	4	5	6	7
firstdordertenure	0	357	-1	0	324	352	282
lastdordertenure	0	354	-1	0	261	279	31
lastdactionclicktenure	222	318	407	7	83	185	23
orderqcount_normal	0	2	0	0	2	6	4
orderqcount_churn7	0	0	-1	0	0	0	10
orderqcount_churn5	0	-1	0	0	0	0	9
orderqcount_churn3	0	0	0	0	0	0	7
orderqcount_churn2	-1	-1	-1	0	0	0	5
orderpricesum_normal	-1	120	-1	0	129	796	552
orderpricesum_churn7	0	0	0	2	44	46	1532
orderpricesum_churn5	0	0	-1	1	26	28	1344

	1	2	3	4	5	6	7
orderpricesum_churn3	0	0	0	1	15	14	1043
orderpricesum_churn2	0	0	0	0	9	9	864
avgdealorderquantity_normal	0	2	-1	0	1	4	1
avgdealorderquantity_churn7	0	-1	0	0	0	0	3
avgdealorderquantity_churn5	0	0	-1	0	0	0	3
avgdealactionclick_normal	8	10	6	0	17	19	24
avgdealactionclick_churn7	2	0	0	1	15	8	30
avgdealactionclick_churn5	0	0	-1	1	12	6	30
Member Counts	**10323**	**6196**	**11252**	**31667**	**3927**	**433**	**30**

Table 5 Feature values for cluster centres

As seen Table 5, Cluster 1 includes customers who have very high last deal action click tenure and they had no order in customer life time. "-1" means that the feature is undefined for these customers.

Cluster 1 is defined as "non paying non active Customers"

Customers in Cluster 2 had orders in normal period, but did not have orders in churned time and they have high last deal action click tenure

Cluster 2 is defined as "churned customers"

Customers in Cluster 3 have high last deal action click tenure, no order

Cluster3 is defined as "non paying, very old customer"

Customers in Cluster 4 have low deal action click tenure, low order payment in normal period and low order payment in churn period

Cluster 4 is defined as "low value, new, active customer"

Customers in Cluster 5 have normal order payments in normal period and low last deal action click tenure.

Cluster 5 is defined as "active, normal value customer"

Customers in Cluster 6 have high order payments in normal period , low order price in churned period and high last deal action click tenure

Cluster 6 is defined as "churned high value customers"

Customers in Cluster 7 have high value order in every period and low deal action click tenure.

Cluster 7 is defined as "active high value customers"

3.3.4 Labelling customers as "Churned" or "Non-Churned"

According to the features of training data we can place the clusters in the Recency - Frequency- Monetary Chart as seen in Figure 17.

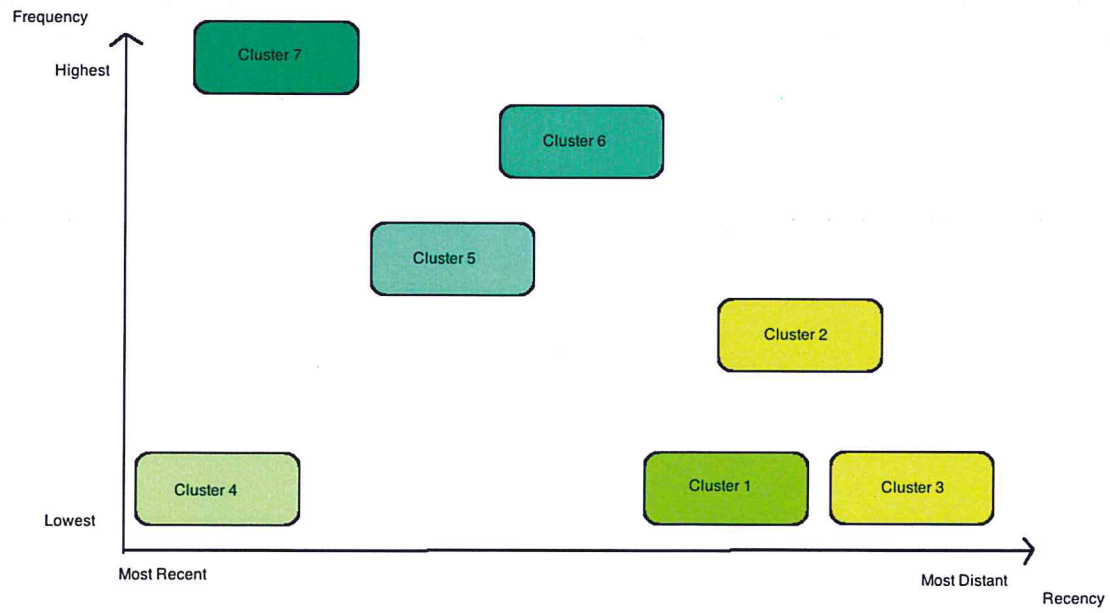

Figure 17 Recency -Frequency chart for dataset clusters

The colours of the clusters represent the monetary of the customers. As we see in the Figure 17 cluster 6 is high value and churned, cluster 2 is low value and churned.

We are not interested in clusters 1 and 3, since they have no order. We are interested in customers who have at least 1 order in normal period and had less or zero order in churn period.

The scatterplot for "x=order quantity count in normal period" vs "y=order quantity count in the last 2 months of churn period" for all customers describes clusters in colours where every colour represent a different cluster as seen in Figure 18.

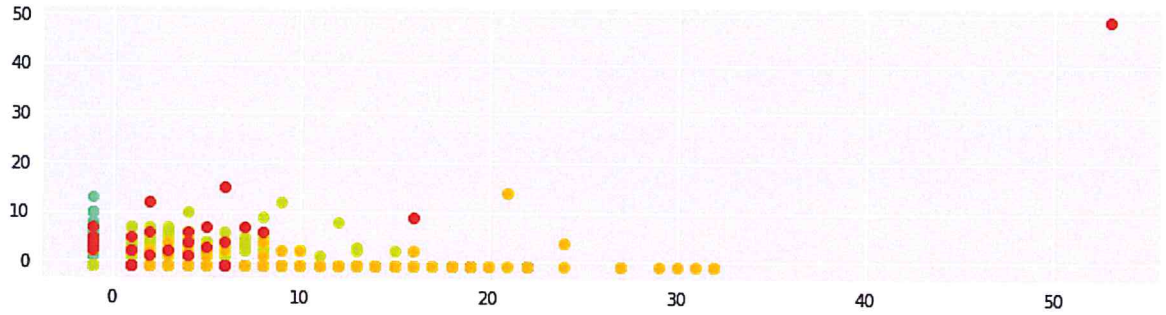

Figure 18 Order quantity count in normal period vs in churn output period in clusters (colours represent clusters)

We can conclude from the Figure 18 that the "orange" coloured points represent "churned" customers.

So we add a dependent variable column to the training data as "churn_label" where we label customers who are in Clusters 2 and 6 with label "1" and the others with label "0".

Figure 19 below we see a 3D scatter plot diagram where "x = Last deal action tenure" , "y = order quantity count in normal period" and "z= order quantity count in churn period" for customers who are labelled as "Churned".

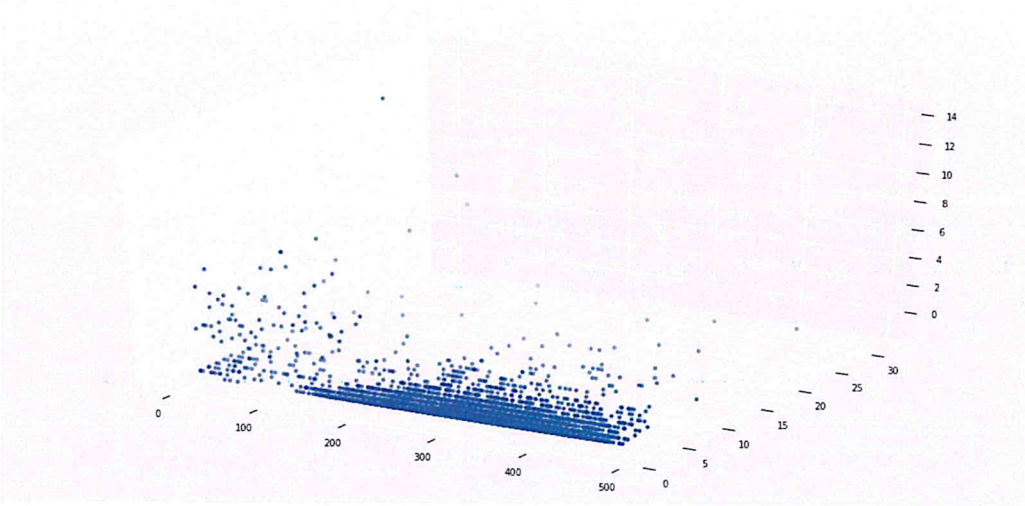

Figure 19 Order quantity count in normal vs in churned period vs last deal action click tenure for churned customers

In Figure 20 below we see scatterplot diagram "x=last deal action click tenure" vs "y=last deal order tenure" for all customers

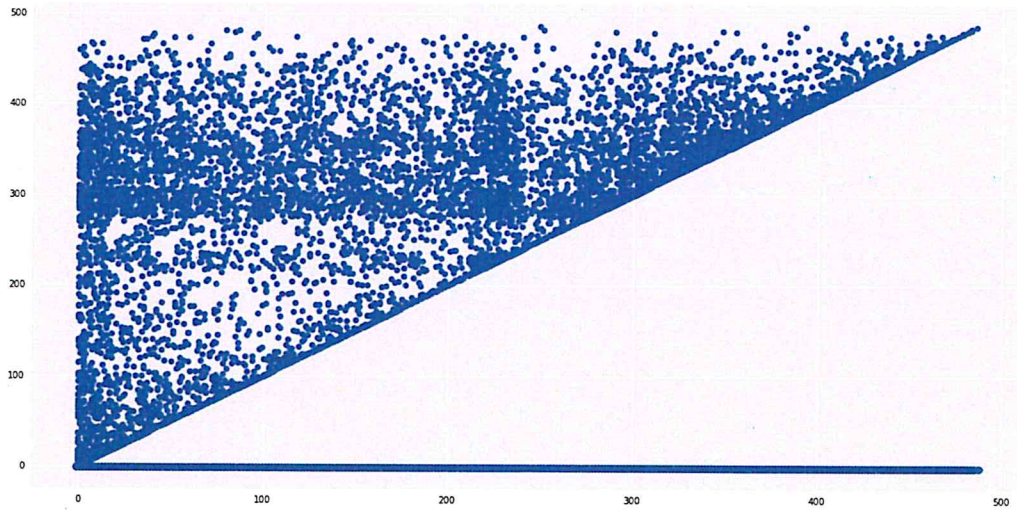

Figure 20 Last deal action click tenure vs last deal order tenure for all customers

If we concentrate only to "churned" customers and plot their "x=last deal action click tenure" vs "y=last deal order tenure" we have the following diagram in Figure 21 below:

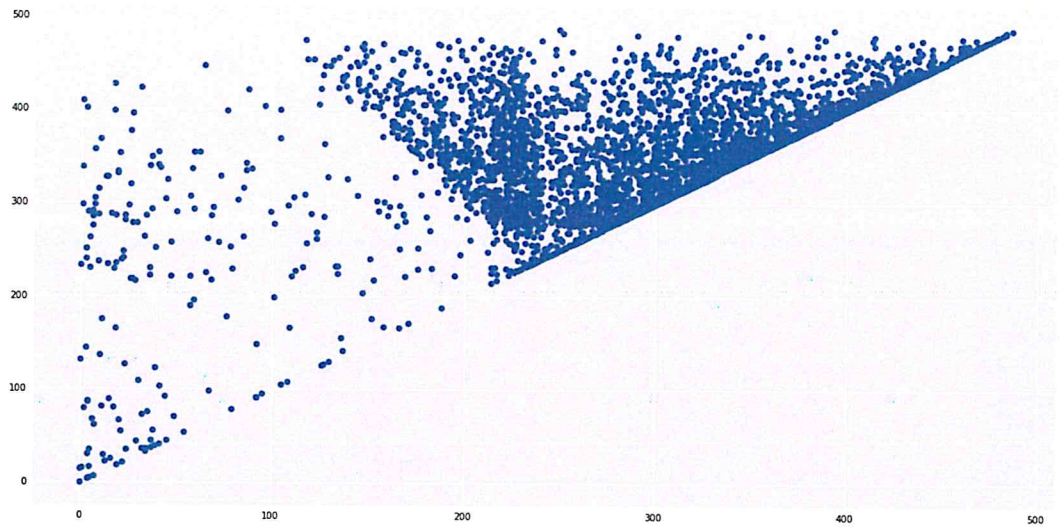

Figure 21 Last deal action click tenure vs last deal order tenure for churned customers

Comparing Figure 20 with Figure 21 shows us that last deal action click tenure and last deal order tenure are mostly very high for "churned" labelled customers.

4. USING SUPERVISED CLASSIFICATION METHODS TO PREDICT CHURNED CUSTOMERS

In Section 3 we have already labelled data with "Churned and "NonChurned", so at this point we have a supervised classification problem. We can create our train and test data and make a cross validation through dataset. We continue our modelling step with "Feature Scaling" , then we will create a "classification" model.

4.1. Feature Scaling

We have 19 independent features and 1 dependent feature. Independent features should be in the same scale so that classification algorithms can result better. For feature scaling Standard Scaler from Scikit library in Python is used.

4.2. Comparing Classification Algorithms

After merging tables with the index of members who are registered between 01.10.2015-30.04.2016 63828 member ids have been received.

Totally 6629 members of Clusters 2 and 6 are labelled as "Churned" and the other 57199 members are labelled as "Non-Churned". To balance "Churned" and "Non-Churned" labelled data, 6629 members of "Non-Churned" data are sampled and all data are shuffled.

There is a lot of classification algorithms so several of them are used and the accuracy results have been compared.

For cross validation K-Fold cross-validation with 5 folds has been used.

	LinearDiscriminantAnalysis(solver='lsqr', shrinkage='auto')	LinearDiscriminantAnalysis(solver='lsqr', shrinkage=None)	LinearDiscriminantAnalysis(solver="svd", store_covariance=True)	QuadraticDiscriminantAnalysis(store_covariances=True)
Mean accuracy score of the folds	0.971	0.972	0.972	0.975
Standard deviation of accuracy score of the folds	0.004	0.004	0.004	0.004

Table 6 Accuracy results for LDA and QDA algorithms with different parameters

	LogisticRegression (l1, C = 100)	Logistic Regression (l1, C = 1)	LogisticRegression (l1, C = 0.01)	LogisticRegression (l2, C = 100)	LogisticRegression (l2, C = 1)	LogisticRegression (l2, C = 0.01)
Mean accuracy score of the folds	0.983	0.983	0.978	0.983	0.983	0.978
Standard deviation of accuracy score of the folds	0.003	0.003	0.003	0.003	0.003	0.003

Table 7 Accuracy results for LogisticRegression with different parameters

	SGDClassifier(loss="hinge", penalty="l2")	KNeighborsClassifier(uniform, neigbour = 5)	KNeighborsClassifier(uniform, neigbour = 50)	KNeighborsClassifier(distance, neigbour = 5)	KNeighborsClassifier(distance, neigbour = 50)
Prediction score mean of the folds	0.901	0.996	0.996	0.992	0.993
Prediction score standard deviation of the folds	0.093	0.001	0.001	0.002	0.001

Table 8 Accuracy results for SGD and KNeighboursClassifier

	SVC((kernel='linear')	SVC((kernel='rbf')	SVC((kernel='poly')	LinearSVC()
Prediction score mean of the folds	0.994	0.994	0.994	0.994
Prediction score standard deviation of the folds	0.002	0.002	0.002	0.002

Table 9 Accuracy results for SVC and LinearSVC with different parameters

	DecisionTreeClassifier(min_samples_split=5)	RandomForestClassifier(min_samples_split=5)	ExtraTreesClassifier(min_samples_split=5)	MLPClassifier(solver='lbfgs')
Prediction score mean of the folds	0.997	0.996	0.994	0.892
Prediction score standard deviation of the folds	0.001	0.001	0.002	0.199

Table 10 Accuracy results for DecisionTree, RandomForest, ExtraTrees and MLPClassifiers

Because we have created homogeneous clusters before, all classifiers give successful results. Comparing all tables shows us DecisionTree, RandomForest and KNeighbours Classifiers are more successful than SVC, SGDClassifier, MLPClassifier, LogisticRegression, ExtraTrees , LinearDiscriminantAnalysis and QuadraticDiscriminantAnalysis classifiers. Dataset has too many values with "-1" because of imputation of missing values, hence tree algorithms may handle this situation better than the other algorithms.

4.3. Validation of Churned Customers

We have labelled 6629 members as "Churned" customers which means that they have made orders of deals in normal period and they are predicted not to make any order in the last 2 months of churn period.

This period is 01.12.2016 to 31.01.2017. In this period totally 15725 orders have been made. If we merge these orders with the churned member ids we receive zero row table, which means that they have made no order in this period.

A second validation is to test if they have ordered after the churned period which is between 01.02.2017 to 17.03.2017. In this period 20538 orders have been made. If we merge these orders with the churned member ids we receive zero row table, which means they have made no order in this period too.

We have validated our model for "Churned" labelled customers.

5. EVALUATION OF THE RESULTS

We have found out that 6629 customers have made purchase once and stopped buying with a reason that we do not know. We have also created a classification model with appropriate features and churn label using the information of clusters which we created with K-Means algorithm. The results of Decisiontree Classifier and Random Forest Classifier are very close to each other but for analysis Random Forest Classifier is selected since Random Forest Classifier avoids overfit by its nature.

This model is dependent on the behaviour of the customers in a time period. Because there are high and low seasons through a year by the nature of e-commerce, the features should be gathered through a long time more than 6 months.

This deal e-commerce web site is growing everyday by member count, also deal count and deal variety. So, the model should be trained again for data for a different time period. If the model developed in this project will be used for time period before 1.10.2015 or after 17.03.2017, the accuracy would be lower because the monthly average and standard deviation of the customers in these time periods would be different. However, the rate of average orders between normal and churn periods in this model could used as an input for another model.

5.1. Analysis Of Random Forest Classifier Result

Data used for clustering are splited into %60 for training and %40 for testing. Random Forest Classifier gave the following confusion matrix as result for test data. Finding "Churn" label is defined as positive.

	Predicted - non churned	Predicted - churned
True - non-churned	2647 (True Negative - TN)	10 (False Positive - FP)
True - churned	8 (False Negative - FN)	2639 (True - Positive - TP)

Table 11 Confusion matrix of test result by Random Forest Classifier

Accuracy score: Overall, how often is the classifier correct?

0.99660 = (TP + TN/ all data)) :

Random Forest Classifier finds real churn and real non-churn 99% correct

Precision score : When it predicts true, how often is it correct?

0.99660 = (TP / (TP + FP)

Recall score: When it's actually true, how often does it predict true?

0.99660 = (TP / (TP + FN)

F1 score: 0.99660 : Weighted harmonic mean of the precision and recall

In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

We see on confusion matrix False Positive is a little bit higher than False Negative. This means the model tend to label customers "churned" rather than "non-churned". This is understandable since some clusters may overlap. But on overall we see that we have selected descriptive features for churning and thus making homogenous clusters for labelling.

5.2. Recommendations To Regain Churned Customers

The orders of the churned customers are analysed by deal tag count, order count, and for total purchased value in normal period and churn period.

There are 1804 customers who have made orders which include more than 2 different deal tags. The histogram for tag count of deal orders is as in Figure 22:

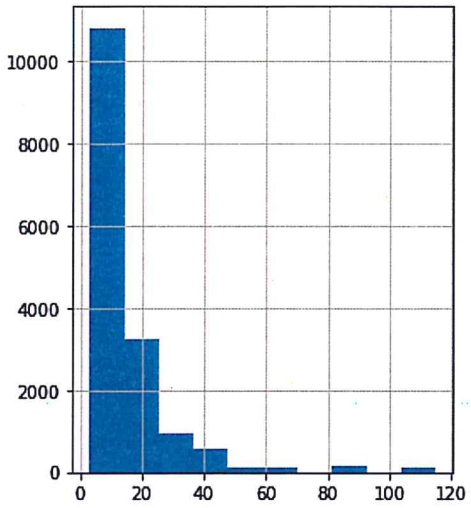

Figure 22 Tag count histogram of orders of customers who have orders more than 2 different deal tags

If we look at the histogram of the order payments of these customers in the normal period and churn period we see these figures below (Figure 23 and Figure 24):

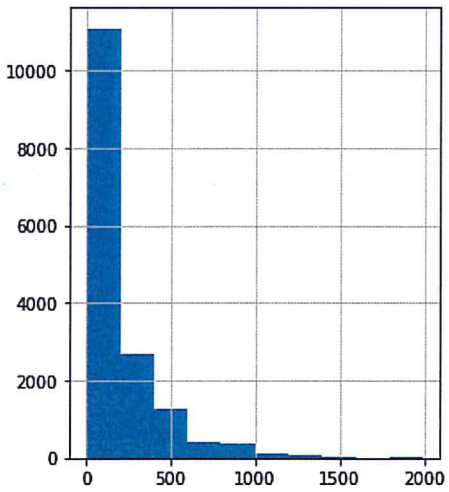

Figure 23 Price of order payments of customers who have orders more than 2 different deal tags in normal period

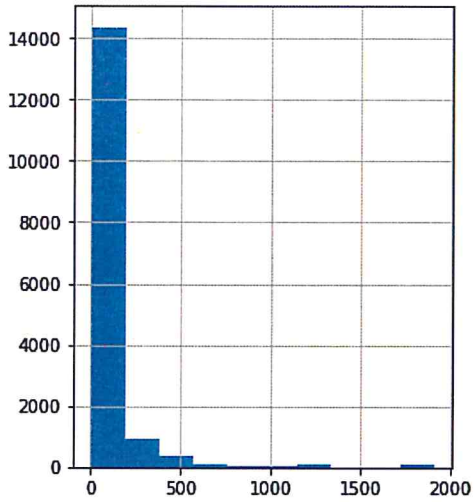

Figure 24 Order price of customers who have orders more than 2 different deal tags in churn period

These histograms show us that these 1804 customers who have made orders which include more than 2 different deal tags were valuable for the website.

The website owner may create a campaign special for these customers for which deal tags they are interested and evaluate the results.

5.3. Social Ethical Aspects

This project is a large exploration of ordering habits of deal customers. In this manner we can see how many people in Turkey can spend to deals for specific tags. We can see distribution of order quantities of deals for specific time period.

These information can be used to calculate for establishing a new deal e-commerce site.

For ethical reasons the name of this deal e-commerce website is not mentioned in this project, the tag values for deals which are mostly ordered are also not mentioned.

5.4. Value Delivered

According Hudaib,2015 [3] "Statistics show that it costs five to six times more to gain a new customer than to keep the existing ones".

In this project 1804 customers are found out which are likely to be regained and order new deals. This is possible because they are interested more than 2 sectors and they have had also orders even in churn period.

Because existing customers can be seen as 5 time valuable than new customers, 1804 existing customers correspond 9000 customers who would register new.

If a e-commerce site gives an advertisement on Facebook the conversion rate to order is %1 according to Smartinsights [6]. That means to get 9000 customers buying at the site the advertisement should reach 900 000 people. Not to mention that target of advertisement campaign should be specified for e-commerce interested people.

The value delivered in this project is to make a campaign which would make 1804 people (churned customers who have ordered more than 2 tags for deals) or even 6629 people (all churned customers) order again a deal from website would be cheaper than to give advertisement to show to 900 000 people on Facebook.

Making campaigns to regain existing customers would increase customer satisfaction. Analysing deal tags of the deals ordered by churned customers would also give an overview which tags and which merchandisers are less demanded. This problem has not been investigated in this project since it is out of scope.

If we used recent data with the model developed in this project we could predict the customers who would churn, and so we could make campaigns to prevent this and this would also increase customer satisfaction.

REFERENCES

[1] Dhandhania, V. (2017, August 10). Retention Metrics Explained: What Is Customer Churn? [RS Labs]. Retrieved August 13, 2017, from https://www.retentionscience.com/retention-metrics-explained-what-is-customer-churn-rs-labs/

[2] Computing Machinery, A. F. (2009, Fall). KDD Cup 2009 : Customer relationship prediction. Retrieved August 13, 2017, from http://www.kdd.org/kdd-cup/view/kdd-cup-2009/Intro

[3] Hudaib, A., Dannoun,, R., Harfoushi, O., Obiedat, R., & Faris, H. (2015). Hybrid Data Mining Models for Predicting Customer Churn. Hybrid Data Mining Models for Predicting Customer Churn,1-6. doi:10.4236/ijcns.2015.85012

[4] Porzak,, Jim. "Using R for Customer Segmentation." Proceedings of UseR! 2008, Dortmund, Germany. September 2013. Accessed August 13, 2017. https://ds4ci.files.wordpress.com/2013/09/user08_jimp_custseg_revnov08.pdf.

[5] Jahromi, A. T. (2009). Predicting Customer Churn in Telecommunications Service Providers. Luleå University of Technology. Retrieved August 13, 2017, from https://www.researchgate.net/profile/Ali_Tamaddoni/publication/267196933_Predicting_Customer_Churn_in_Telecommunications_Service_Providers/links/5555860c08ae6943a871c662/Predicting-Customer-Churn-in-Telecommunications-Service-Providers.pdf

[6] Chaffey, D. (2017, March 02). Ecommerce conversion rates 2017. Retrieved August 13, 2017, from http://www.smartinsights.com/ecommerce/ecommerce-analytics/ecommerce-conversion-rates/

[7] Gove, R. (2015, December 3). Using the elbow method to determine the optimal number of clusters for k-means clustering. Retrieved September 06, 2017, from https://bl.ocks.org/rpgove/0060ff3b656618e9136b