MEF UNIVERSITY

CUSTOMER SEGMENTATION AND CUSTOMER CHURN PREDICTION FOR BABIL.COM

Capstone Project

Berk Çakar

İSTANBUL, 2021

MEF UNIVERSITY

CUSTOMER SEGMENTATION AND CUSTOMER CHURN PREDICTION FOR BABIL.COM

Capstone Project

Berk Çakar

Asst. Prof. Evren Güney

İSTANBUL, 2021

MEF UNIVERSITY

Name of the project: Customer Segmentation and Customer Churn Prediction for babil.com

Name/Last Name of the Student: Berk Çakar

Date of Thesis Defense: dd/mm/yyyy

I hereby state that the graduation project prepared by Berk Çakar has been completed under my supervision. I accept this work as a "Graduation Project".

dd/mm/yyyy

Asst.Prof. Evren Güney

I hereby state that I have examined this graduation project by Berk Çakar which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

dd/mm/yyyy

Director

of

Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Evren Güney	…….
2. ……..	……..

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.

Berk Çakar 25/01/2021
Signature

EXECUTIVE SUMMARY

CUSTOMER SEGMENTATION AND CUSTOMER CHURN PREDICTION FOR BABIL.COM

Berk Çakar

Advisor: Asst. Prof. Evren Güney

JANUARY 2021, 33 Pages

In the past decade, a lot of players have joined into e-commerce market and competition in the market has been increasing lately. The e-commerce companies want to use their resources more efficiently to stay ahead in the competition. Personal communication with customers, increasing customer loyalty, acquiring new customers and preventing customer churn are some of the ways to achieve this goal.

Babil.com is an e-retailer that sells books online and it is one of the companies which wants to stay ahead in the competition. It is founded in 2013 and now it is a 8 years old mature company. So, instead of spending much resources on acquiring new customers, trying to keep existing customers and increasing retention rate is a more ideal goal for the company. Also, personal communication with customers and reaching them with the right product in the right time is crucial.

In this project, a customer segmentation with two levels is implemented to help Babil.com. For the first level of segmentation, customers' value to company is identified by RFM segmentation and in the second level of segmentation customers' behaviors are identified by K-Means clustering. To prevent customer churn, a machine learning algorithm which predicts customers who will churn in the next 6 months. With this algorithm, it will be easy to take an action for the right customers in the right time.

Key Words: Customer Segmentation, Customer Churn Prediction, RFM Segmentation, K-Means Clustering, Classification

ÖZET

Babil.com için Müşteri Segmentasyonu ve Kayıp Müşteri Tahminlemesi

Berk Çakar

Tez Danışmanı: Dr. Öğr. Üyesi Evren Güney

OCAK, 2021, 33 Sayfa

Son yıllarda e-ticaret pazarına bir çok oyuncu katıldı, bununla birlikte pazardaki rekabet son zamanlarda artış göstermekte. E-ticaret şirketleri ellerindeki imkanları daha efektif bir şekilde kullanarak bu rekabette önde olmak istiyorlar. Müşteriler ile kişisel iletişim, müşteri sadakatini artırma, yeni müşteriler kazanma ve müşteri kaybını engelleme bu hedefe ulaşmanın yollarından bir kaçı.

Babil.com internet sitesi üzerinden kitap satışı yapan ve bu rekabette önde olmak isteyen oyunculardan biri. 2013 yılında kurulan şirket şuanda olgunluğa ulaşmış durumda. Bu yüzden yeni müşteri kazanımına kaynak ayırmaktansa varolan müşterilerin kaybedilmesini engellemek şirket için daha ideal bir yol olarak öne çıkıyor. Ayrıca müşteriler ile daha kişiselleştirilmiş iletişimler kurmak ve doğru zamanda doğru ürün ile müşteriye ulaşmak kritik önem arz ediyor.

Bu projede Babil.com'a yardımcı olmak amacıyla iki seviyeli bir müşteri segmentasyonu uygulandı. Müşteri segmentasyonun ilk seviyesinde müşterilerin şirket için ne kadar değerli oldukları GSP (Güncellik – Sıklık – Parasallık) methodu ile belirlenirken ikinci seviyesinde müşteri davranışlarını ayırt etmek adına K-Means kümeleme yöntemi kullanıldı. Müşteri kaybını engellemek için ise şirketin gelecek 6 ay içerisinde kaybedeceği müşterileri tahminleyen bir makine öğrenmesi modeli kuruldu. Bu algoritma ile birlikte doğru müşterilere doğru zamanda aksiyon alınabilmesi kolaylaştırıldı.

Anahtar Kelimeler: Müşteri Segmentasyonu, Kayıp Müşteri Tahminlemesi, Değer Segmentasyonu, K-Means Kümeleme, Sınıflandırma

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Overview

With increased access to data in every business, the way we do business has changed and it created a new era in marketing. In this new era of marketing, knowing your customers has become a necessity. Customers are expecting more personalized communications, campaign and marketing actions and they want you to understand them, know them. Smart marketers fulfill this request and try to understand and know their customers instead of just creating more clicks or acquiring more customers (Makhija, 2019). There are some analysis techniques and algorithms that help us to understand our customers.

One of these techniques is RFM Analysis. RFM stands for Recency, Frequency and Monetary, respectively. Recency is the time since last purchase of customer and it is considered as customers with less recency are more likely to engage with marketing actions. Frequency is number of visits in a specific time period. Frequent customers are more valuable to the companies and they are more likely to engage with marketing actions. And the last one is Monetary, customers' total value to the company. Customers with higher monetary value are more loyal to the company. With RFM analysis, we can evaluate these three metrics to calculate a customer's value to the company. RFM analysis helps us to show values of customers and it gets easier to create customer groups whose value are closer to each other. But their behavioral habits can be different.

The second technique is K-Means algorithm which is a good example of unsupervised learning and create desired number of clusters within data points according to their distances to each other. If we use customers' basket penetrations or revenue penetrations among the departments, we can create meaningful and distinguishable clusters. With this clustering, we can know what our customers are into and love more about our company.

If we combine these two techniques with each other, we know our customers' value to the company and what they love about our company. And this knowledge helps us to create suitable marketing actions for customers and it will increase success rate of marketing actions. Also, it may help us to use marketing budget efficiently.

Getting to know our customers is the very first step of combining big data analytics and marketing. The second step is taking targeted marketing actions which appeals our customers with existing knowledge about customers. Some purposes of this marketing actions are new customer acquisition, creating incremental revenue for every customer, preventing customer churn and etc.

Preventing customer churn costs way more than gaining new customers, so identifying and predicting customers who will leave our services in a short period of time are really important for every organization (Shaaban et al., 2012). If a company manages to predict those customers who will churn successfully it will be able to take propriety actions on time to save its customers and budget. Predicting churn is relatively easier than before, thanks to advanced supervised machine learning algorithms.

While K-Means algorithm and RFM analysis help us to know our customers better, and make them feel like our company cares about their behaviors. Churn prediction helps you to save your customers and money. However, big data applications that assisting marketing actions are not limited with these three. I will use these three methods in this paper and implement these techniques to an e-commerce company, Babil.com.

1.2. Literature Review

Customer segmentation is an essential project for any company to have a better knowledge about their customers. This knowledge can be used for increase profits by customizing marketing actions and offering appropriate products for each of the customers (Christy et al., 2018). There are many methods to create customer segments. RFM segmentation is one of these methods, and it helps to identify customer groups with their recency, frequency and monetary values. Customers with similar scores can be grouped to serve the goal of the company. While customers with high scores can be seen as the most loyal ones, customers with lower scores can be counted as lost. With this kind of segmentation, loyal customers are rewarded and they can become early adopters of new products and help promote the brand (Makhija, 2019) and customers with lower scores can be targeted to prevent churn.

Another method for customer segmentation is using unsupervised clustering algorithms such as K-Means algorithm. The goal of the algorithm is to group customers into separate non-overlapping subgroups with transactional customer data. One of the key applications of K means

clustering is segmentation of customers in order to gain a better understanding of them, which, in effect, could be used to maximize the company's revenue (Sagar, 2019).

The last approach to customer segmentation will discussed is Fuzzy C-Means. The algorithm is similar to the K-Means; the main difference is Fuzzy C-Means algorithm calculates relation of a data point to each cluster centers. In this algorithm, a data point can be belong to two different centroids with different weights (Dubey, 2013). As Dubey (2013) compared Fuzzy C-Means and K-Means algorithms in his paper, he stated that K-Means algorithm has higher purity score then Fuzzy C-Means. Therefore, K-Means algorithm may be suited better than Fuzzy C-Means in our business problem to create best customer groups.

Customer churn is another business problem that will be addressed in this paper. Customer churn can be explained as loss of a customer from customer base by leaving the services or products of the company. The companies are making huge efforts to prevent customer churn with data analysis. There two approaches for this problem, the first one is a predictive machine learning algorithm and the other one is uplift modeling.

For predictive approach creating right variables and finding the best-suited algorithm for the problem is crucial. In today, machine-learning algorithms are getting better every day, and there are many algorithms that one of them might be able to suit your data better. Some algorithms are well suited for a few domain types, although they may not hold true in all cases. It's mainly the underlying data that powers the performance of various algorithms (Subramanya, 2016). To evaluate all algorithms some metrics like accuracy, precision, recall scores. In addition to these scores, confusion matrixes are a good qualifier for an algorithm.

The uplift modeling is the second approach for customer churn problem. Uplift modeling is trying to estimate the outcome to a specific treatment (Devriendt et al., 2021). In uplift modeling, efficiency of treatment is the key point and the main goal of this approach is getting the best outcome with limited resources. Instead of targeting all customers of a prediction model, only customers who might be convinced with a campaign are targeted. Target categories of an uplift model can be seen in Figure 1.

Churn when targeted	Yes	Do-Not-Disturbs	Lost Causes
	No	Sure Things	Persuadables
		No	Yes
		Churn when not targeted	

Figure 1: Four Classes for Uplift Modeling

2. ABOUT DATA

2.1. Data Explanation & Preparation

Babil.com shared two datasets for the project in json format. The first dataset is called sales data which includes every sale has made on website. Dataset contains information about order id, customer id, order date, product id, product price and order total price. The second dataset is product data which includes every product in company's inventory. The product data contains information about product id, product name, publisher id and category tree.

All features of sales data have object data type and it is not possible to make mathematical calculations without data type conversions. Order total price and product price features converted into float to make it possible to do mathematical operations. Order date feature contains date-time information about order. Time part of this feature was not needed for further analysis and it is converted into date format without order time. Rest of the features do not require a data type conversion, so they remained as object. The final version of Sales data set is shown in Figure 2 below.

order_id	customer_id	order_date	order_total_price	product_id	product_price
142113555014	11327	2015-07-14	58.50	9786055358921	19.50
142113555014	11327	2015-07-14	58.50	9786055358921	19.50
142113555014	11327	2015-07-14	58.50	9786055358921	19.50
142113556814	11327	2015-07-14	106.34	9789750833519	11.25
142113556814	11327	2015-07-14	106.34	9786059809122	12.00

Figure 2: Organized View of Sales Data

The product data has four features and all of them in object data type. There is no need for data type conversion but category tree feature which includes information about product tree is intertwined. To get more information about product, category tree feature separated into 6 different features which are "Type", "Main Category", "Category", "Sub_Cat0", "Sub_Cat1" and "Sub_Cat2" respectively. The final version of Product data set is shown in Figure 3.

id	name	publisher_id	category_tree	Type	Main Category	Category	Sub_Cat0	Sub_Cat1	Sub_Cat2
9789757860716	İstanbul'da Konut Girişleri ve Kapılar Residenti	5434	KİTAP>SEYAHAT>Referans	KİTAP	SEYAHAT	Referans	None	None	None
9789757860723	Kapadokya Cappadocia	5434	KİTAP>FOTOĞRAFÇILIK>Genel	KİTAP	FOTOĞRAFÇILIK	Genel	None	None	None
9799757860272	Monkey's Right to Paint	5434	KİTAP>SANAT>Genel	KİTAP	SANAT	Genel	None	None	None
9789757860693	Post Peripheral Flux	5434	KİTAP>REFERANS>Genel	KİTAP	REFERANS	Genel	None	None	None
9789758431724	Yapı Fiziği ve Malzemesi	11812	KİTAP>TEKNOLOJİ & MÜHENDİSLİK>İnşaat>Genel	KİTAP	TEKNOLOJİ & MÜHENDİSLİK	İnşaat	Genel	None	None

Figure 3: Organized View of Product Data

In the final step of data preparation Sales and Product data are joined together by using common feature in two data sets which is product id (as id in product data). The final dataset which will be called as "main data" in the rest, is used in further analysis to have access product and order information easily.

2.2. Exploratory Data Analysis

Main data contains records between April 4, 2014 and February 2, 2017. In total main data has 1,084,167 rows, which come from 236,627 distinct orders and 124,614 customers. Yearly order and customer counts are demonstrated in Figure 4 and Figure 5. Babil.com had managed to increase their customer and order numbers every year since 2014.

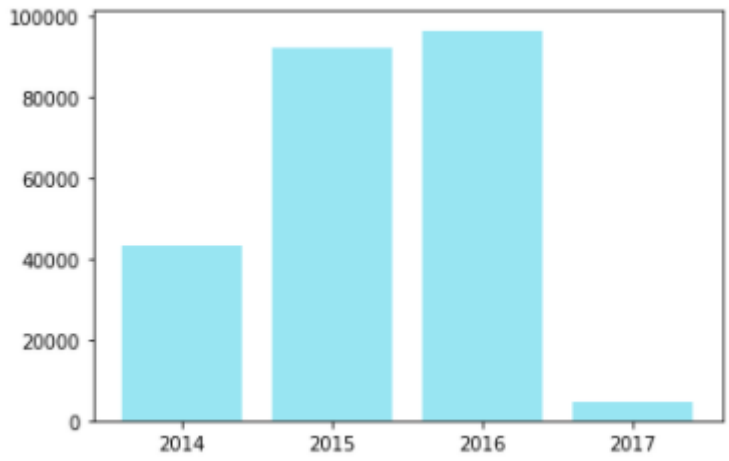

Figure 4: Yearly Order Counts

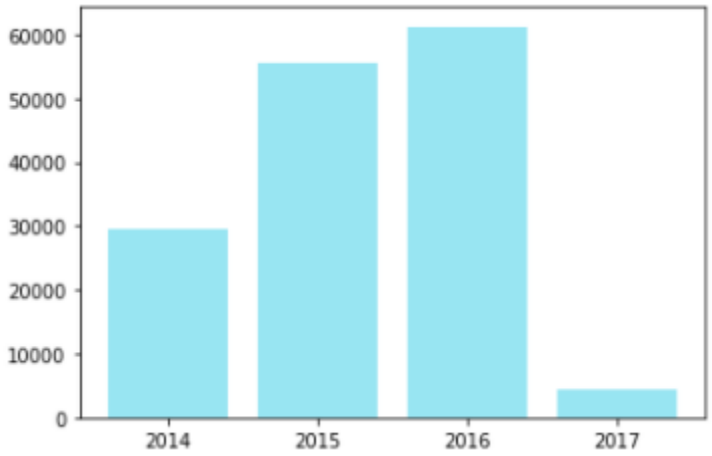

Figure 5: Yearly Customer Counts

To understand customers' shopping habits, analyzing sales based on product tree is crucial. There are 8 different types that has a sales record between 2014 and 2017. In Table 1 below, order counts and order penetrations of these 8 types are demonstrated. According to this table, "KİTAP", "GÜZEL ŞEYLER" and "YABANCI KİTAP" are the top three types among the sales data and these top three penetrated nearly all of the orders.

Table 1: Order Counts and Order Penetrations Based on Types

Type	Orders (#)	Orders (%)
KİTAP	217,168	91.8%
GÜZEL ŞEYLER	14,249	6.0%
YABANCI KİTAP	9,600	4.1%
E-KİTAP	5,743	2.4%
MÜZİK	1,084	0.5%
ELEKTRONİK	1,041	0.4%
FİLM	957	0.4%
HOBİ & OYUNCAK	450	0.2%
TOTAL	**236,627**	**100%**

To see what customers choose in "KİTAP" type, sales metrics of main categories that belongs to "KİTAP" type are examined with order counts, order penetrations, revenue and revenue penetrations. "KURGU", "ÇOCUK & GENÇLİK KURGU" and "DİN" are the most desirable main categories for customers according to order penetration and these three main categories are

penetrated to 73.8% of all orders and 36.3% of total revenue. All details about top 10 categories are demonstrated in Table 2 below.

Table 2: Sales Metrics of Main Categories in Type KİTAP

Main Category	Orders (#)	Orders (%)	Revenue (TL)	Revenue (%)
KURGU	116,245	49.1%	4,405,341	27.0%
ÇOCUK & GENÇLİK KURGU	30,215	12.8%	726,967	4.4%
DİN	28,212	11.9%	801,434	4.9%
BİYOGRAFİ & OTOBİYOGRAFİ	24,744	10.5%	479,704	2.9%
EDEBİ KOLEKSİYONLAR	23,122	9.8%	393,827	2.4%
TARİH	22,310	9.4%	616,236	3.8%
SİYASAL BİLİM	15,496	6.5%	374,067	2.3%
ŞİİR	14,508	6.1%	236,035	1.4%
KİŞİSEL GELİŞİM	14,211	6.0%	299,993	1.8%
FELSEFE	14,048	5.9%	287,995	1.8%
TOTAL	**236,627**	**100%**	**16,342,460**	**100%**

In the final step of exploratory data analysis, top 10 products according to shoppers are examined and demonstrated in Table 3 below. “Küçük Prens”, “Bülbülü Öldürmek” and “Kürk Mantolu Madonna” are the top 3 products according to customers who bought these books. These top 3 products are penetrated 11.2% of all customer base.

Table 3: Top 10 Products Based on Customer Counts

Product Name	Customers (#)	Customers (%)
Küçük Prens	5,482	4.4%
Bülbülü Öldürmek	4,576	3.7%
Kürk Mantolu Madonna	3,867	3.1%
Kafamda Bir Tuhaflık	3,553	2.9%
Sineklerin Tanrısı	3,211	2.6%
1984	3,137	2.5%
Zamanın Kısa Tarihi	2,984	2.4%
İçimizdeki Şeytan	2,884	2.3%
Hayvan Çiftliği	2,767	2.2%
Yabancı	2,627	2.1%
TOTAL	**124,614**	**100%**

3. PROJECT DEFINITION

3.1. Problem Statement

In the competitive retail market, efficiency is one of the key performance indicators for every company. They need to take effective marketing actions to gain market share and to attract new customers. The best way to do this is taking more personal marketing actions and preventing customer churn. These two topics are two main problems that babil.com is facing in their marketing operations. The company do not have enough information about their customers for more personal and effective marketing actions and there is no information or action to prevent customer churn.

3.2. Project Objective

To help babil.com with the problems they are facing, a customer segmentation that identifies shopping habits and behaviors of customers so the company will be able to take different marketing actions for different customers. A machine learning algorithm which calculates probability of churn for every customer will be implemented for customer base to identify lost customers and it will help company to take effective actions in the right time.

3.3. Project Scope

In this project RFM (Recency – Frequency – Monetary) segmentation and K-Means algorithm is used to create customer segmentation with two levels which contains information about habit and behavior of customers. With RFM segmentation it will be easy to identify customers with similar shopping habits. Using customers' basket and revenue penetrations with K-Means algorithm is helped to create customers clusters with similar main category interests. By these two level it will be possible to have a marketing action for loyal customers who have interest to history books.

To predict customers who will churn in the next period, regression algorithms, tree based algorithms and boosting algorithms are used with hyper parameter tuning. Almost hundred features are created for customer churn model and customer segments are also used as an input for the algorithm.

4. METHODOLOGY

4.1. Habit Segmentation (RFM Segmentation)

RFM segmentation is a method that makes it easier to take customized marketing actions (Makhija, 2019). The method takes recency, frequency and monetary value of a customer into account while creating different segments. According to the general market knowledge, recent, more frequent and customers with higher volume tend to respond marketing actions and promotions so, it provides warrant targeting (Yang, 2004).

While applying RFM segmentation, there are many ways to score your customers according to their recency, frequency and monetary values. One of the most common ways is scoring customers according to quartiles of variables. Serrano (2020) says that using quartiles is the most helpful and easiest way to create RFM segments. In this project, instead of quartiles five different score categories identified by 20 percent of each variable to achieve more detailed customer segments.

In this project, orders between January 1, 2015 and December 31, 2015 are used while creating customer segments. Recency of customers are calculated by substituting last day of customers' visit from January 1, 2016. Frequency of customers are calculated by counting distinct order ids for the selected time period and monetary values are calculated by summing order values. Detailed calculations can be seen in Figure 6.

```
customer_revenue = invoice_data.groupby("customer_id")[["order_total_price"]].agg("sum")
customer_invoice = invoice_data.groupby("customer_id")[["order_id"]].agg("nunique")
customer_lastdate = invoice_data.groupby("customer_id")[["order_date"]].agg("max")
rfm_data = pd.concat([customer_revenue, customer_invoice,customer_lastdate], axis=1)
rfm_data.columns = ['total_revenue', 'total_transactions','last_visit_date']

rfm_data.reset_index(inplace=True)

now = dt.date(2016,1,1)
rfm_data["recency"] = (rfm_data['last_visit_date'].dt.date).apply(lambda x: (now - x).days)
rfm_data
```

	customer_id	total_revenue	total_transactions	last_visit_date	avg_basket	recency
0	100007	16.20	1	2015-03-15	16.200000	292
1	100009	48.65	1	2015-03-15	48.650000	292
2	100010	471.13	6	2015-12-24	78.521667	8
3	100012	5.00	1	2015-03-15	5.000000	292
4	100017	14.40	1	2015-03-15	14.400000	292

Figure 6: Calculation of RFM Values

After creating this data, percentiles of RFM variables are examined to understand and interpret distributions and habits of yearly customer base. There are 55,506 shoppers in 2015 and nearly 70% of them have only 1 transaction. This is a bad performance indicator for the company and rate of customers with single transaction is above the retail market according to my experiences in three different retail players. But this is understandable when Turkish culture and reading habits of Turkey are taken into account. When monetary values are examined, 64% of customers spent less than 100 TL in a year. For online market, 100 TL can be considered as a small basket. Detailed distributions of variables are shown in Table 4.

Table 4: Distributions of RFM Variables

Variable	P10	P20	P30	P40	P50	P60	P70	P80	P90	P95	MAX
Recency	17	37	70	96	129	154	196	268	320	345	360
Frequency	1	1	1	1	1	1	2	2	3	4	7
Monetary	22.4	38.6	47.1	55.0	68.3	92.7	111.8	167.8	268.4	356.2	715.8

When recency variable is examined, %18 of customers have an order in the last month. On the other hand, %33 of customer base does not have an order in the last six months and nearly %15 percent of the customers had not ordered an item from babil.com since 11 months. The company need to give attention those customers to prevent customer churn with appropriate marketing actions. Monthly recency distribution of customer base are shown in Figure 7.

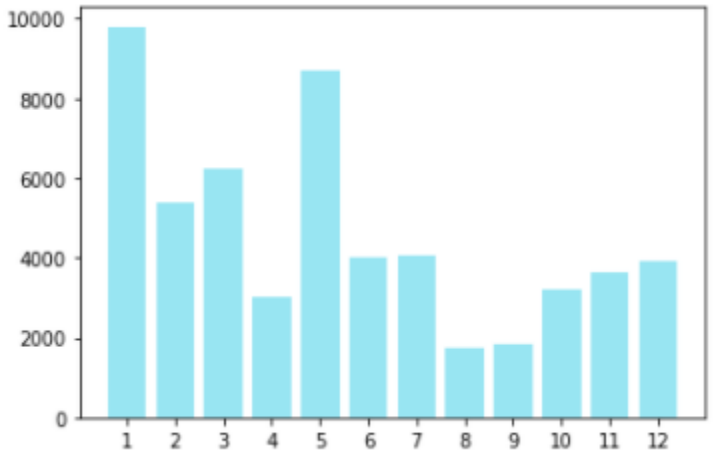

Figure 7: Monthly Recency Distribution of Customer Base

When examining of RFM variables are done, each variable divided into quintiles and customers ranked from 1 to 5 for each variable. The most recent, frequent and valuable customers got 5 points and the least recent, frequent and valuable customers got 1 point. After the ranking, each point from every variable are combined together and RFM score are created for all customers. Customers are divided into 7 segments according to their RFM scores. These segments are;

- Champions: Customers who get 5 points from all variables and they are the most valuable customers for babil.com
- Loyals: Customers who get 4 or more points from recency and frequency. Monetary values were not taken into account for this segment. These customers are more frequent and constant shoppers of babil.com.
- Spenders: Customers who get 4 or more points from monetary ranking, 2 or more points from recency ranking and 2 or less points from frequency ranking. These customers are not constant shoppers of babil.com but they have higher baskets.
- Traffic Makers: Customers who get 3 or more points from frequency ranking. 2 or more points from recency ranking. These customers are constant shoppers of babil.com but they have smaller baskets than usual.
- Low F – Low M: Customers who get 4 or more points from recency ranking, 1 from frequency ranking and 3 or less from monetary ranking. These customers are one the most recent customers of the company but they do not have valuable baskets. The company needs to take an action to make them more loyal to the company.

- Likely to Lose: Customers who get 3 or less points from recency ranking, 1 from frequency ranking and 3 or less from monetary ranking. These customers had ordered long time ago and their customers were not valuable. The company needs to engage with this customers to make them more frequent and valuable.
- Lost: Customers who get 1 point from recency ranking. Their frequency and monetary scores are ignored because they had ordered from company long time ago. The company needs to take an action for this customers and try to get them as Champions or Loyals. Detailed customer numbers and customer distribution for each segment are shown in Table 5.

Table 5: Customer Distributions for Habit Segments

Customer Segment	Customers (#)	Customers (%)
1) Champions	2,016	3.6%
2) Loyals	6,056	10.9%
3) Spenders	6,173	11.1%
4) Frequents	8,100	14.6%
5) Low F - Low M	9,324	16.8%
6) Likely to Lose	12,772	23.0%
7) Lost	11,065	19.9%
Total	**55,506**	**100%**

4.2. Behavioral Segmentation (K-Means Clustering)

For the second level of the customer segmentation, main category behaviors of customers are taken into account. So, the company will be able to target customers with items they are interested. Clustering is an efficient method for discovering subtle thus important patterns and relationships in a dataset (Ezenkwu, 2015).

K-Means clustering is an unsupervised learning algorithm that sees data points in dimensional space and groups them into given number of clusters with many iterations. (Al-Masri, 2019). In the first iteration of the process, algorithm randomly choses center points in space for k clusters and calculates distances of each data point to cluster centers. After calculating distances, every data pointed is assigned to the nearest center. For the second iteration, cluster centers are

recalculated within the groups and the rest of the process continues with the same steps until number of max iterations is reached or cluster centers remains constant (Jeevan, 2018).

In this project, a customer-based dataset with basket penetrations of top 10 main categories are created to catch buried patterns and creates customer groups with similar category interest with K-Means algorithm. An example subset from K-Means dataset is shown in Figure 8.

customer_id	kurgu_penet	cocuk_kurgu_penet	din_penet	biyografi_penet	tarih_penet	cocuk_kurgudisi_penet	edebi_penet	arapca_penet	felsefe_penet	siir_pe
100007	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000
100009	1.000000	0.0	0.0	1.000000	0.000000	0.0	0.0	0.0	0.0	0.000
100010	0.333333	0.0	0.5	0.333333	0.166667	0.0	0.0	0.0	0.0	0.166
100012	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	1.000
100017	0.000000	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.000

Figure 8: Sample Data Used for K-Means Clustering

As mentioned above, K-Means divides data into desired groups and there are number of methods to specify optimal number of clusters. Elbow method is one of many ways to decide optimal number of clusters for given dataset. In this method, dataset fits into K-Means algorithm for a range of K values and scoring parameter which is the sum of squared distances from each point to the assigned center for each K represented in a line chart. When the chart forms an arm, the inflection point on the curve shows the best number of K (Benfort, 2020). For the created dataset, 5 clusters are the optimal number as shown in the Figure 9.

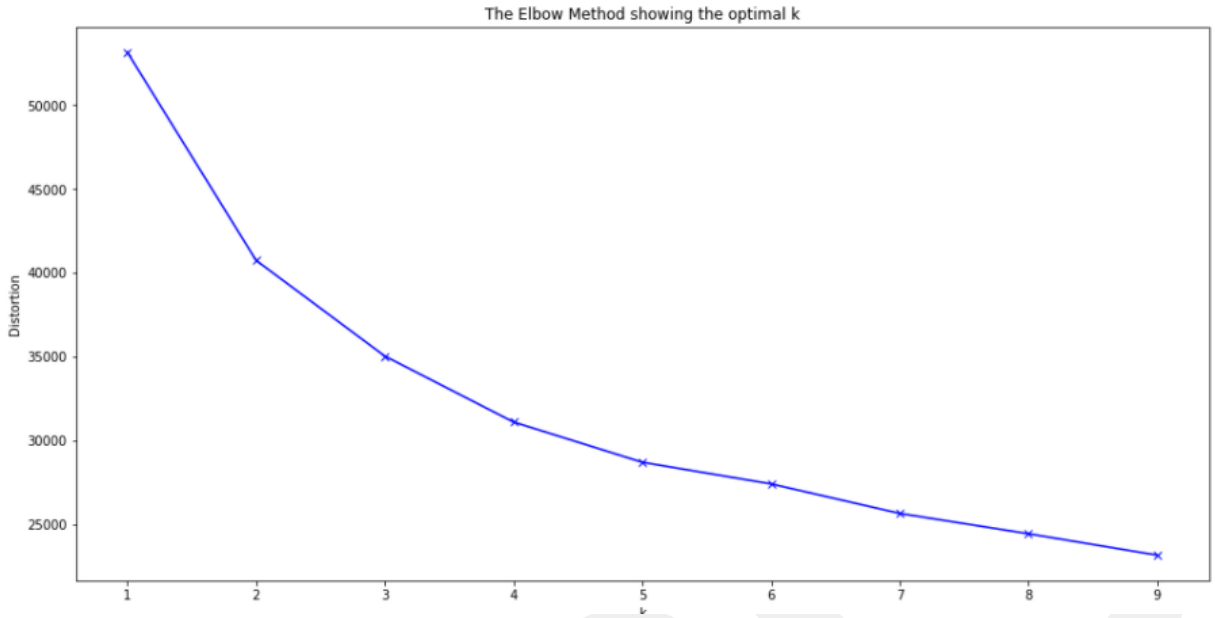

Figure 9: Elbow Graph

K-Means algorithm with 5 clusters and 1,000 iterations is applied to the created dataset. Clusters are named by examining cluster centers and segments are created manually. These five segments are;

- Fiction (Cluster-2)
- Fiction & Others (Cluster-3)
- Child Books (Cluster-1)
- Religious (Cluster-4)
- Others (Cluster-1)

Detailed cluster centers and customer distribution among behavioral segments are shown in Table 6 and Table 7 below.

Table 6: K-Means Cluster Centers

Cluster	Fiction	Child Fiction	Religion	Biography	History	Child (Fiction Excl.)	Philosophy	Poetry	Others
0	58%	95%	7%	10%	6%	21%	4%	7%	48%
1	3%	1%	7%	5%	7%	3%	5%	2%	98%
2	96%	4%	8%	10%	6%	1%	4%	8%	11%
3	91%	5%	12%	16%	14%	2%	10%	9%	93%
4	5%	2%	40%	20%	20%	9%	7%	7%	9%

Table 7: Customer Number and Distribution of Behavioral Segments

Behavioral Segment	Customers (#)	Customers (%)
1) Fiction	11,532	20.8%
2) Fiction & Others	12,865	23.2%
3) Child Books	6,039	10.9%
4) Religous	5,718	10.3%
5) Others	19,352	34.9%
Total	**55,506**	**100%**

4.3 Customer Churn Prediction

Customer churn is one of the biggest problems that many companies are facing. Babil.com founded in 2013 and it is a mature company right now. They have acquired many new customers since foundation and now they are trying to switch their effort from acquiring new customers to retaining existing ones. Because for E-commerce companies, the cost of acquiring new customers is a big effort in comparison to retaining existing customers (Subramanya, 2016).

To prevent customer churn, most of the companies are using machine-learning algorithms to predict which customers will be lost in the next period. Decision trees and logistic regression are two most common algorithms to predict customer churn with high accuracy performance and simplicity. On the other hand, decision trees are not good enough to understand linear relations between different variables and logistic regression cannot create interaction between variables (Caigny et al., 2018). In the last decade, boosting algorithms are getting more popular every day due to their high performance. In addition, they are highly customizable and that makes them desirable for any business problem (Natekin & Knoll, 2013). In this project, decision tree and logistic regression algorithms are selected for the base performance and a gradient boosting algorithm is selected to beat this base performance.

As discussed in customer segmentation, 55,506 customers had ordered at least one item from babil.com in 2015. When churn rate examined for 1 month, 3 months, 6 months and 12 months, churn rates are 93.4%, 84.7%, 80.3% and 74.3% respectively. According to Turkish Statistics Institution, an average Turkish person watches TV for 6 hours, surfing on the internet for 3 hours and reads a book for only a minute in a day. When this statistics and knowledge about Turkish people and culture is taken into account, analyzing customer churn for 6 months seems

acceptable. Detailed customer numbers who churned in the given period and churn rates can be seen in Table 8 below.

Table 8: Churn Rates for Different Periods

Churn Period	Customers (#)	Churn Rate (%)
1 Month	51,816	93.4%
3 Months	46,987	84.7%
6 Months	44,572	80.3%
12 Months	41,235	74.3%

For customer churn prediction, 122 features are created by using sales data, product data and customer segments. These 123 features include recency, tenure, transaction count and revenue for different time periods which are ever, last 12 months, last 9 months, last 3 months and last month. Also, transaction counts and revenue are calculated for 8 different product types for the same time periods. Segments that are created in the earlier steps of the project are used as an input for customer churn prediction as well. When customer segments are examined for customer churn, it is easily noticed that higher segments like Champions and Loyals have lower churn rate than other segments. On the other hand, there is not a significant difference behavioral segmentation but Fiction, Fiction & Others and Child Books segments have lower churn rate than customer base. Detailed churn rates of customer segments can be examined in Table 9 and Table 10 below.

Table 9: Churn Rates for Habit Segments

Customer Segment	Customers (#)	Churners (#)	Churn Rate (%)
1) Champions	2,016	691	34.3%
2) Loyals	6,056	3,446	56.9%
3) Spenders	6,173	5,292	85.7%
4) Frequents	8,100	6,320	78.0%
5) Low F - Low M	9,324	7,398	79.3%
6) Likely to Lose	12,772	11,498	90.0%
7) Lost	11,065	9,927	89.7%
Total	**55,506**	**44,572**	**80.3%**

Table 10: Churn Rates for Behavioral Segments

Behavioral Segment	Customers (#)	Churners (#)	Churn Rate (%)
1) Fiction	11,532	9,079	78.7%
2) Fiction & Others	12,865	9,580	74.5%
3) Child Books	6,039	4,821	79.8%
4) Religous	5,718	4,607	80.6%
5) Others	19,352	16,485	85.2%
Total	**55,506**	**44,572**	**80.3%**

Customer recency and historical transaction number of customers are taken into examination as well. For recency, more recent customer groups have lower churn rates. While customers with 1 month recency have 63% churn rate, customers with 12 months recency have 90.4% churn rate. For historical transaction numbers, customers with only 1 transaction have a churn rate of 87.9% while customers with 7 and more transactions have 39.0% of churn rate. Detailed churn rates of recency groups and historical transaction groups can be examined in Table 11 and Table 12 below.

Table 11: Churn Rates of Recency Groups

Month Recency	Customers (#)	Churners (#)	Churn Rate (%)
1	9,796	6,171	63.0%
2	5,367	3,942	73.4%
3	6,228	4,880	78.4%
4	3,018	2,365	78.4%
5	8,690	7,427	85.5%
6	3,989	3,493	87.6%
7	4,044	3,567	88.2%
8	1,752	1,485	84.8%
9	1,858	1,571	84.6%
10	3,226	2,896	89.8%
11	3,641	3,252	89.3%
12	3,897	3,523	90.4%
Total	**55,506**	**44,572**	**80.3%**

Table 12: Churn Rates of Historical Transaction Groups

Transaction Group	Customers (#)	Churners (#)	Churn Rate (%)
1) 1 Transaction	33,793	29,708	87.9%
2) 2-3 Transactions	15,543	11,579	74.5%
3) 4-6 Transactions	4,549	2,653	58.3%
4) 7+ Transactions	1,621	632	39.0%
Total	**55,506**	**44,572**	**80.3%**

Customer churn prediction is a classification problem and it can be solved with a supervised learning algorithm like decision trees, support vector machines, logistic regression, etc. Decision tree, logistic regression, random forest and gradient boosting classifier are used to predict customer churn. Decision tree and logistic regression used as a based algorithm and I tried to achieve better results with random forest and gradient boosting. Accuracy, precision, recall and F-1 scores for each algorithm can be seen in Table 13 below.

Table 13: Metrics for Machine Learning Algorithms

Algorithm	Accuracy	Precision	Recall	F-1
Decision Tree	0.72	0.73	0.72	0.72
Logistic Regression	0.82	0.79	0.82	0.77
Random Forest	0.80	0.77	0.80	0.77
Gradient Boosting	0.82	0.79	0.82	0.78

Decision tree and random forest have poor results compared to logistic regression and gradient boosting. When logistic regression and gradient boosting are compared to each other, they have similar results. The only difference is F-1 score of gradient boosting algorithm is 0.01 point higher than logistic regression. To improve gradient boosting algorithm, hyper parameter tuning with grid search is applied. Comparative results and confusion matrix for Gradient Boosting v1 can be seen in Table 14 and Table 15. Gradient Boosting with Hyper Parameter Tuning is named as Gradient Boosting v1 in the table. Results of Gradient Boosting with Hyper Parameter Tuning did not improve as expected and it stayed in the same level.

Table 14: Metrics for Second Round Algorithms

Algorithm	Accuracy	Precision	Recall	F-1
Logistic Regression	0.82	0.79	0.82	0.77
Gradient Boosting	0.82	0.79	0.82	0.78
Gradient Boosting v1	0.82	0.79	0.82	0.78

Table 15: Confusion Matrix of Gradient Boosting v1

	Test	
Prediction	1	0
1	452	297
0	1,735	8,618

5. CONCLUSION

Competition has been increasing radically in the past decade for e-commerce market and companies try to reach their customers in a personalized manner to convince them in the right time. In this competitive environment, Babil.com wants to have a better knowledge about its customers to avoid staying behind the competition. It is usually better to analyze customers in homogenous groups instead of analyzing all customer base (Makhija, 2019). To help Babil.com, a customer segmentation has been implemented in two levels. In the first level of the segmentation, customer shopping habits like recency, frequency and monetary are used to create segments. With this kind of knowledge, Babil.com is able to know which customers are more valuable for their company. In the second level of the segmentation, customers' shopping histories are analyzed to identify different behaviors. K-Means Clustering method is implemented to identify different behaviors and we were able to create 5 unique behavioral groups. With this two level customer segmentation, Babil.com will be able to communicate with their customers in a more personal way.

Babil.com was founded in 2013, in the early periods of the company they focused on acquiring new customers as expected. But now, it is a mature company and it is getting much harder and more expensive to acquire new customers. Instead of acquiring new customers, Babil.com wants to focus on keeping existing customers active on their website. Also, actions to keep existing customers are three times cheaper than acquiring new customers (Subramanya, 2016). To achieve this goal, 5 classification models were developed and Gradient Boosting Algorithms with Hyper Parameter Tuning showed the best results.

To sum up, Babil.com has a great knowledge about their customers and able to communicate them personally with customer segmentation. Also, Babil.com is able to manage its resources more efficiently by trying to keep their existing customers. With the right actions, Babil.com can get ahead of the competition easily.

REFERENCES

[1] Al-Masri, A. (2019, May 15). How Does k-Means Clustering in Machine Learning Work? Available at: <https://towardsdatascience.com/how-does-k-means-clustering-in-machine-learning-work-fdaaaf5acfa0> [Accessed 22 January 2021]

[2] A. Joy Christy, A. Umamakeswari, L. Priyatharsini, A. Neyaa (2018). RFM ranking – An effective approach to customer segmentation

[3] Bengfort, Benjamin. "Elbow Method." Elbow Method - Yellowbrick v1.2 Documentation, Available at <www.scikit-yb.org/en/latest/api/cluster/elbow.html. > [Accessed 28 June 2020]

[4] Caigny, A. D., Coussement, K., De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760-772. doi:10.1016/j.ejor.2018.02.009

[5] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance Kalu, Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services, 2015

[6] Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. Information Sciences, 548, 497-515. doi:10.1016/j.ins.2019.12.075

[7] Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr, A Proposed Churn Prediction Model, International Journal of Engineering Research and Applications, 2012

[8] Jeevan, M. (2018, October 22). Possibly the simplest way to explain K-Means algorithm. Available at: <https://bigdata-madesimple.com/possibly-the-simplest-way-to-explain-k-means-algorithm/> [Accessed 22 January 2021]

[9] Karthik B. Subramanya, Enhanced feature mining and classifier models to predict customer churn for an e-retailer, Graduate Theses and Dissertations, 2016

[10] Makhija, P., 2019. RFM Analysis For Customer Segmentation. [online] CleverTap. Available at: <https://clevertap.com/blog/rfm-analysis/> [Accessed 20 May 2020]

[11] Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. doi:10.3389/fnbot.2013.00021

[12] Sagar, A., 2019. Customer Segmentation Using K Means Clustering. [online] Medium. Available at: <https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3> [Accessed 26 May 2020]

[13] Serrano, S. (2021, January 20). RFM Analysis Guide: 6 Key Segments for RFM Marketing. Available at: <https://www.barilliance.com/rfm-analysis/> [Accessed 22 January 2021]

[14] S. Ghosh, S. Dubey (2013). Comparative Analysis of K-Means and Fuzzy CMeans Algorithms, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4,

[15] Yang, Amoy X. "How to Develop New Approaches to RFM Segmentation." Journal of Targeting, Measurement and Analysis for Marketing, Palgrave Macmillan UK, 1 Sept. 2004.