İnternet Üzerinden Alışverişlerde Ürün İade Tahmini
Prediction of Return in Online Shopping

İsmail Bilgen, Ömer Sinan Saraç
Bilgisayar Mühendisliği Bölümü
İstanbul Teknik Üniversitesi
İstanbul, Türkiye
{ibilgen,sinan.sarac}@itu.edu.tr

***Özetçe* —İnternet üzerinden yapılan alışverişlerin yaygınlık kazanması şirketler açısından bazı olumsuzlukları beraberinde getirir. Bunlardan biri alınan ürünlerin çeşitli sebeplerle iade edilmek istenmesidir. Şirketler için amaç müşteri memnuniyetini zedelemeden satılan ürünlerin geri iade oranını azaltacak önleyici tedbirler almaktır. Bu çalışmanın amacı makine öğrenmesi yöntemlerini kullanarak sipariş edilen ürünün geri iade edilip edilmeyeceğini tahmin etmektir. Bunun için geçmişte gerçekleştirilen siparişlere ait veriler kullanılmaktadır. Bu veriler sipariş işlemine, müşteriye ve ürüne ait çeşitli bilgilerden oluşur. Bu çalışmanın en önemli katkılarından biri, veri kümesinde bulunmayan bir takım niteliklerin, uzman tecrübesi ve mevcut bilginin harmanlanmasıyla elde edilmesidir. Bir diğeri de, mevcut verilerden tahmin başarısını arttıracak yararlı yeni özniteliklerin keşfedilmesidir.**

Anahtar Kelimeler—internet üzerinden alışveriş, ürün geri iade tahmini.

***Abstract*—Mail order business gains popularity day by day. One major problem for the retailers is the high return rates. The incurred cost of returns forces companies to take measures to reduce the number of returns without affecting the customer satisfaction. The aim of this study is to predict whether a purchase results with a return or not based on historical purchase data using machine learning techniques. The data consist of various information about the purchase, customer and the items. Major contribution of this study is to show that using external information to generate features which would otherwise be impossible to extract from data directly improves prediction accuracy significantly.**

Keywords—online shopping, return of a shipment.

I. GİRİŞ

İnternet üzerinden alışveriş alışkanlığı gün gittikçe daha fazla yaygınlık kazanmaktadır. Bunun firmalar açısından birçok avantajı olduğu gibi dezavantajları da vardır. En önemli dezavantajlarından biri posta yoluyla alınan ürünlerin çeşitli sebeplerle iade edilmek istenmesidir. İnternet ve posta yoluyla ticaret yapan birçok şirket, müşteri memnuniyetini sağlamak açısından kolay ve çoğu zaman da ücretsiz yoldan ürünü geri iade etme imkanı sunar. Ancak bu, satılan ürünlerin yüksek oranlarda geri iade edilmesine sebep olur. Ürünün iadesinden doğan ücretler de şirketlerin üstüne kalır. Şirketler için amaç müşteri memnuniyetini zedelemeden satılan ürünlerin geri iade oranını azaltacak önleyici tedbirler almaktır. Bu çalışmanın amacı makine öğrenmesi yöntemlerini kullanarak sipariş edilen ürünün geri iade edilip edilmeyeceğini tahmin etmektir. Bunun için geçmişte gerçekleştirilen siparişlere ait veriler kullanılmaktadır. Bu veriler sipariş işlemine, müşteriye ve ürüne ait çeşitli bilgilerden oluşmaktadır. Bu çalışmanın en önemli katkılarından biri, veri kümesinde bulunmayan bir takım niteliklerin, uzman tecrübesi ve mevcut bilginin harmanlanmasıyla elde edilmesidir. Bir diğeri de, mevcut verilerden tahmin başarısını arttıracak yararlı yeni bilgilerin keşfedilmesidir.

Bu çalışmada kullanılan veri kümesi, "Data Mining Cup Competition 2014" (DMC-2014) [1] para ödüllü projesi çerçevesinde sunulmuştur. Projede, sınıfları bilinen eğitim kümesi verisi üzerinden eğitilen yapay öğrenme modeli ile eğitim kümesinden tamamen bağımsız ve sınıf bilgisi içermeyen sınama verilerinin sınıflandırılması istenmiştir.

II. BENZER ÇALIŞMALAR

Postayla ticaret işinde en önemli meselelerden biri ürün iadesidir. Benzer bir bir çalışmada müşterilerin davranışlarının tahmin edilmesi için öncelikli olarak S Biçimli Bağlanım (Logistic Regression), Yapay Sinir Ağları (Neural Networks), Naive Bayes ve En Yüksek Düzensizlik (Maximum Entropy) yöntemlerini denemişlerdir [2]. Ayrıca, Maximum Entropy yöntemini kullanarak, her bir nitelik değerinin sonsal sınıf dağılımlarına ağırlıklar kestirmiş ve Naive Bayes yöntemiyle birleştirmişlerdir. Bu yaklaşımla 2004 Data-Mining-Cup (Veri Madenciliği Yarışması)'da en iyi sonucu elde etmiştir.

Ürün geri iadesinin tahmin edilmesinde müşterilerin eski alışveriş kayıtlarına göre etiketlenmesi olumlu katkı sağlayabilir. Bir başka çalışmada IBM I-Miner aracında bulunan demografik kümeleme yöntemi kullanılarak müşterileri, yüksek kâr, yüksek değer ve düşük risk olacak şekilde kümelenmiştir [3].

Bir başka çalışmada geçmiş alışveriş verilerinden yararlı bilginin çıkarılması için Fuzzy Demetleme ve olgular arasındaki ilişkinin ortaya çıkarılması için de İlişkilendirme Kuralı (association rule) kullanılmıştır [4]. Satın alma zaman aralığı ve satın alma sayısına göre ve müşteri yaşı, cinsiyeti ve bölge kodu gibi çeşitli değişkenlere göre Fuzzy Kümeleme yöntemi kullanarak veri üç kümeye ayrılmış ve en fazla miktarda alışveriş yapan müşteriler tespit edilmiştir.

978-1-4799-4874-1/14/$31.00

III. VERİ KÜMESİ

Veri kümesi, posta yoluyla (online) alışveriş yapan bir ticari müessesede kayıtları tutulan geçmiş alışveriş hareketlerine ait bilgilerinden oluşmaktadır. Veri kümesinde 481092 alışveriş hareketi için müşteriyi ve yapılan alışveriş işlemini niteleyen değerler ile alışverişin geri iade ile sonuçlanıp sonuçlanmadığını bildiren sınıf etiketi bulunmaktadır. Veri kümesindeki değişkenler Tablo I'de gösterilmiştir. Veri kümesindeki nitelikler, manaları belirtilmek kaydıyla, daha anlaşılır olması için çoğunlukla orjinal isimleri ile kullanılmıştır.

Tablo I: Veri kümesine genel bakış.

Nitelikler	**Nitelik Sınıfı**	**Nitelik Değerleri**	**Eksik Veri sayısı**
orderItemID	sayı	1 2 3 4 5 6 7 8 9 10 ...	0
orderDate	tarih / 365	"2012-04-01","2012-04-02",..	0
deliveryDate	tarih / 327	"1990-12-31","2012-04-03",..	39419
itemID	sayı	186 71 71 22 151 598 15 ...	0
size	kategori / 122	"1","10","10+",..	0
color	kategori / 87	"almond","amethyst",..	143
manufacturerID	sayı	25 21 21 14 53 87 1 3 ...	0
price	sayı	69.9 70 70 39.9 29.9 ...	0
customerID	sayı	794 794 794 808 825 ...	0
salutation	kategori / 5	"Company","Family",..	0
dateOfBirth	tarih / 14308	"1655-04-19","1900-11-19",..	48889
state	kategori / 16	"Baden-Wuerttemberg",..	0
creationDate	tarih / 775	"2011-02-16","2011-02-17",..	0
returnShipment	sınıf etiketi / 2	"0","1"	0

Veri kümesinde eksik ve hatalı veriler bulunmaktadır. Örneğin renk niteliğinde eksik veriler bulunmaktadır. Bir kısım örneklerde ise doğum tarihi niteliği 19.11.1900 şeklindedir. Bu nitelik değerleri de hatalı olarak düşünülmüştür. Aynı şekilde teslim tarihi de eksik ve hatalı veriler içermektedir.

IV. YÖNTEM

Makine öğrenmesi, büyük veri kaynaklarından yararlı bilginin çıkarılmasında temel teşkil eder. Gelecek daima bir miktar geçmişe benzer varsayımı ile hareket eder. Geçmiş verilerden elde edilen tecrübe ile gelecekteki örnekler tahmin edilir.

Bu çalışmada geçmiş sipariş bilgilerinden oluşan veri kümesindeki eksik ve hatalı veriler çeşitli yöntemler kullanılarak giderilmiştir. Mevcut nitelik değerlerinin yanında sınıflandırma performansını arttıracak yeni nitelikler çıkarılmıştır. Elde edilen veri kümesi, Rastgele Orman (RO) [5] ve Destek Vektör Makinaları (DVM) [6] makine öğrenmesi yöntemleri kullanılarak eğitildi. Yöntemler R ortamında gerçeklendi [7].

A. Veri Ön İşleme

Veri üzerinde makine öğrenmesi yöntemlerinden daha iyi sonuç alabilmek için öncelikle eksik nitelik değerlerinin giderilmesi gerekmektedir. Eksik verileri gidermek için çeşitli yaklaşımlar vardır. İlk akla gelen yöntem eksik veri içeren örnekleri veri kümesinden çıkarmaktır. Yeterli veri olması durumunda tercih edilebilir. Bir diğer yöntem, eksik verileri o niteliğe ait ortalama değerler ile doldurmaktır. Veri kümesinde ham halde hangi nitelikten ne kadar eksik değer olduğu Tablo I'de belirtilmişti. Bu bölümde bu eksik verilerin giderilmesi için hangi yöntemlerin kullanıldığı anlatılacaktır.

Teslim tarihi (deliveryDate) niteliğindeki eksik veriler için önerilen ve kullanılan çözüm siparişlerin ortalama iletim süresini hesapladıktan sonra, *ortalama (sipariş tarihi – teslim tarihi)*, bu süreyi sipariş tarihine ekleyerek teslime tarihini hesaplamaktır. Sipariş tarihinde eksik veri bulunmaması bu işlemi kolaylaştırmaktadır. Ancak bir diğer önemli husus teslim tarihindeki hatalı verilerdir. Teslim tarihi niteliğindeki 4660 adet sipariş "1990-12-31" değerine sahiptir. Halbuki sipariş verileri 2012 ve 2013 yıllarına aittir. Bu durumda ortalama süre hesaplanırken bu tarihler hariç tutulmalıdır ve diğer eksik veriler gibi bu veriler de ortalama süre üzerinden düzeltilmelidir.

Renk (color) niteliğindeki eksik verileri gidermek için en sık geçen renk kullanılabilir. Sipariş verilerinde en sık görülen renk niteliği olarak siyah geçmektedir (86252 adet). Renk değerine sipariş edilen ürün bazlı olarak da bakılabilir. Ancak bu veri kümesinde eksik olan bir renk değeri aynı ürüne ait bütün sipariş hareketlerinde eksik olduğu görülmektedir. Bundan dolayı eksik renk değerleri için en sık görülen nitelik "siyah" değeri atandı

Doğum tarihi (dateOfBirth) en fazla sayıda eksik ve hatalı değer içeren niteliktir. Bu değerleri düzeltmek mümkün olmadığı için eksik ve hatalı değer içeren sipariş verileri veri kümesinden çıkarılmıştır.

B. Yararlı Bilginin Keşfi

Ürün, sipariş hareketi ve müşteri hakkında veri tabanlarında tutulan, sipariş tarihi, müşteri IDsi gibi nitelikler genellikle makine öğrenmesinde doğrudan kullanılmaya uydun değildir. Daha çok, bu niteliklerin önemini ortaya çıkaracak yeni bir takım değişkenlerin elde edilmesi gerekmektedir. Örneğin ürün numarası (itemID), kategorik bir değişken olarak ele alındığında, çok fazla sayıda olduğu için yapay öğrenme yöntemlerinin bu nitelikler üzerlerinde bir bağıntı kurmaları zordur. Ayrıca kategorik değişkenlerin sayısının artması karmaşıklığı çok fazla arttırdığı için birçok yöntem bu değişkenin sayısını sınırlandırır. Ürün numarası, sıralı (nominal) olarak da bir anlam ifade etmez. O halde ürün numaraları yerine, ürünün karakteristiğini yansıtacak niteliklerin bulunması gerekmektedir. Bu veri kümesinden aşağıdaki nitelikler üretilmiştir.

- **Siparişten Teslime Geçen Gün:** Sipariş tarihi ile teslim tarihi arasında geçen zaman ürünün geri iadesinde önem teşkil edebilir. Bu sebeple veri kümesindeki her bir sipariş hareketi için sipariş tarihi ile teslim tarihi arasındaki fark gün bazında bulunmuştur. Bu değişken ayrıca belirli bir eşik değerlerle karşılaştırılıp, örneğin teslim süresi 7 günden fazla veya az olanlar şeklinde, ikili değişkenler de elde edilebilir.
- **Önemli Günlere Yakınlık:** Sipariş edilen ürün bilinen özel günlerden biri için istenmiş ancak geç teslimattan dolayı geri gönderilmiş olabilir. Bu durumda ürünün sipariş tarihi özel günlere yakın mı diye bakılabilir. Örneğin sevgililer günü ya da yıl başından önceki 2 hafta ya da 1 aylık periyotta sipariş edilen ürünler tespit edilir. Ayrıca bir başka önemli bilgi sipariş tarihi bu özel günlerden önce iken teslimat tarihi sonra olanlara da bakıldığında çıkar. Bu çalışmada özel günler olarak sevgililer günü ve yeni yıl tarihleri, yakınlık ölçütü olarak da bir aylık dönem seçilmiştir.

- **Müşterinin Yaşı:** Doğrudan doğum tarihi ile alakalı bir bilgidir. Bu bilgi sayısal olarak anlam ifade eder. Ayrıca müşterinin yaşından kategorik değişkenler de elde edilebilir. Örneğin müşteri genç, orta yaşlı ve yaşlı olarak sınıflandırılabilir. Örneğin 35 yaş altı genç (0), 35-55 yaş arası orta yaşlı (1), daha yukarısı da yaşlı (2) sayılabilir.
- **Müşterinin Üyelik Zamanı:** Müşterilerin ne kadar eski üye olduklarını gösterir. Hesap oluşturma tarihinden en son siparişi verdiği tarihe kadar gün bazında oluşan fark, bu niteliğe değerini vermiştir.
- **Ürünün Geri Gönderilme Riski:** Geçmişe dair siparişler incelendiğinde, siparişin geri gönderilme ile sonuçlanma riski ürün, üretici ve müşteri tabanlı olarak hesaplanabilir. Risk hesaplanırken basit olasılık hesabının ötesinde farklı stratejiler denenmiştir. Bu şekilde yapılmasının sebebi, sipariş verisinde bir ürüne ait olan satışların az sayıda olması durumunun risk hesaplamada doğuracağı olumsuz etkinin önüne geçmektir. Tablo II'de her bir tekil ürün için veri kümesi üzerinde ne kadarının geri gönderilme ile sonuçlandığı hesaplanmıştır.

Tablo II: Ürün bazlı olarak satışların geri gönderilme (r=1) ve gönderilmeme (r=0) adetleri.

itemID	# of r=0	# of r=1
1	295	381
2	214	143
3	320	147
4	105	191
…	…	…

Denenen ilk yaklaşımda bütün ürünler için bir ilk olasılık değeri ön görülmüştür. Ürünün geri gönderilme riski bu ilk olasılık değeri ile ürünün geri gönderilme oranı birleştirilerek elde edilmiştir. Formül 1'de r_1 geri gönderilme ve r_0 gönderilmeme sayısıdır.

$$risk_1 = \frac{5 + r_1}{10 + r_0 + r_1} \tag{1}$$

Diğer bir yaklaşımda ise Denklem 2'de verildiği şekilde hesaplanmıştır. Bu yaklaşım ile az sayıda satılan üründeki geri gönderilme oranı risk değerini daha az etkilerken, çok sayıda satılan üründeki oran daha fazla etkilemektedir. Önceki yaklaşımda geri gönderilme oranı 1/2 (2 üründen 1'i iade) ile 100/200 olanlar aynı risk değerini üretirken bu yaklaşımda 100/200 oranı daha kesin olduğundan risk değerini 0.5 başlangıç değerinden daha fazla uzaklaştırır, yani daha yüksek bir risk değeri üretir.

$$risk_2 = 1 - \frac{r_0^c + 50}{(r_0 + r_1)^c + 100} \tag{2}$$

c katsayısısı, verilerin değişeceği aralığı değiştirmektedir.
Bu şekilde ürünlerin risk değerleri hesaplandığında Tablo III'deki sonuçlar elde edilmiştir. Ürün 5 için gerçekleşen toplam 750 adet satışın 375 tanesi geri gönderilme ile sonuçlanmıştır. Burada $Risk_2$ hesaplama yönteminin önemi öne çıkmaktadır. Ayrıca ürün 3066'a bakıldığında sipariş edilen 5 ürünün tamamı geri gönderilmiştir. Normal olasılık hesabına göre oran risk %100 çıkmaktadır. Halbuki bir ürünün riskli olduğunu söylemek için 5 yeterli bir sayı değildir. Sonuç olarak $Risk_2$ yönteminin daha sağlıklı değerler ürettiği görülmüştür.

Tablo III: Ürünlerin geri gönderilme risklerinin farklı yaklaşımlarla hesaplanması.

itemID	r_0	r_1	$Risk_1$	$Risk_2$
1	295	381	0.5626822	0.7105232
2	214	143	0.4032698	0.5353684
3	320	147	0.3186583	0.4334420
4	105	191	0.6405229	0.7831655
5	**375**	**375**	**0.5000000**	**0.6457371**
6	320	328	0.5060790	0.6520519
...	...	...	...	...
3066	**0**	**5**	**0.6666667**	**0.5502802**
3067	1	0	0.4545455	0.4950495
3068	1	0	0.4545455	0.4950495
3069	64	7	0.1481481	0.1951383
3070	1	0	0.4545455	0.4950495
3071	1	0	0.4545455	0.4950495

- **Müşteri Karakteri:** Müşteri numarası (customerID) gibi nitelikleri öğrenmede kullanmak, sınıflandırıcıyı doğrudan müşteriler üzerinden bir nevi önyargı yaparak karar vermeye iter. Bu yüzden ID değerlerini kullanmak yerine, müşteri karakterini belirleyecek nitelikler türetilmesi gerekmektedir. Buna göre müşterinin geçmişte aldığı ürünlerin fiyat ortalaması müşteriyi tanımlayan bir nitelik olarak eklenebilir. Bunun gibi müşterinin yaşı, ne kadar eski üye olduğu, yaptığı alışveriş sayısı, hitap vasfı, en çok aldığı ürün rengi, aldığı ürünlerin fiyatı vb. de müşteriyi tanımlayan nitelikler olarak düşünülebilir.

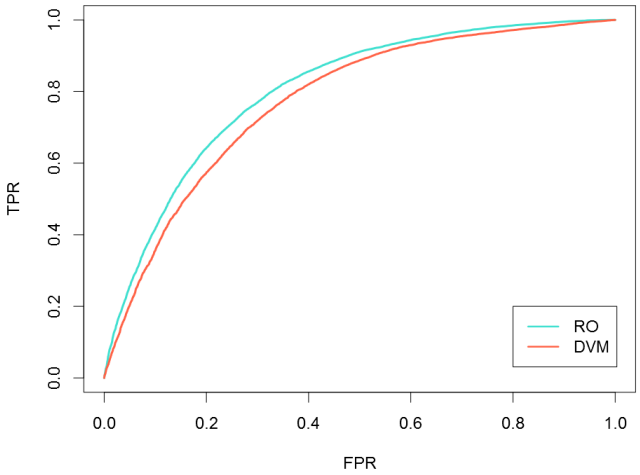
V. SONUÇLAR

Deneylerde değerlendirme ölçütü olarak ROC ve Kesinlik/Anma eğrileri kullanılmıştır. ROC eğrisi doğru pozitif (TPR) ve yanlış pozitif oranlarının (FPR) farklı eşik değerlere göre kıyasını verir. ROC eğrisi altında kalan alan (AUROC) bu eğriyi özetleyen tek bir değer verir. Doğruluk (PREC) ve anma (REC) grafiği de kesinliğin arttığı durumlarda anmanın durumunu verir. Bu grafiğin altında kalan (AUPRC) değeri de aynı şekilde bu grafiği tek bir değere indirger ve modeller arasındaki kıyası kolaylaştırır.

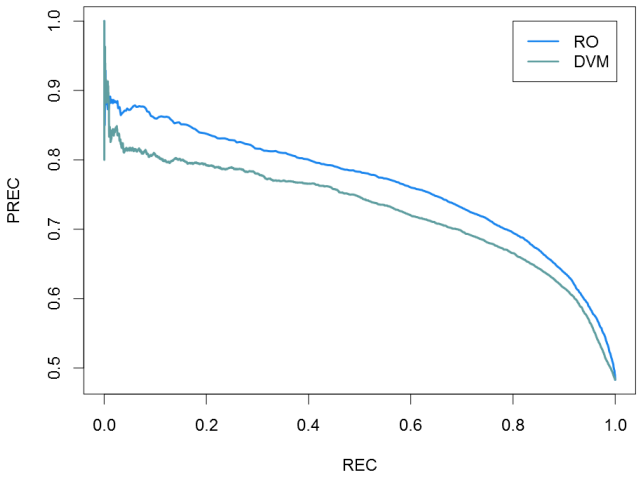
RO ve DVM yapay öğrenme yöntemleri veri kümesindeki bütün niteliklerle eğitildiğinde Tablo IV'deki sonuçlar alınmıştır. Bu sonuçlara göre RO, bu veri kümesinde DVM'den daha iyi sonuç vermiştir. Bu sonuçlar Şekil 1 ve Şekil 2'de rahatlıkla görülmektedir.

Tablo IV: RO ve DVM yöntemleri ile elde edilen sonuçlar.

	AUROC	AUPRC
RO	0.803	0.761
DVM	0.771	0.727



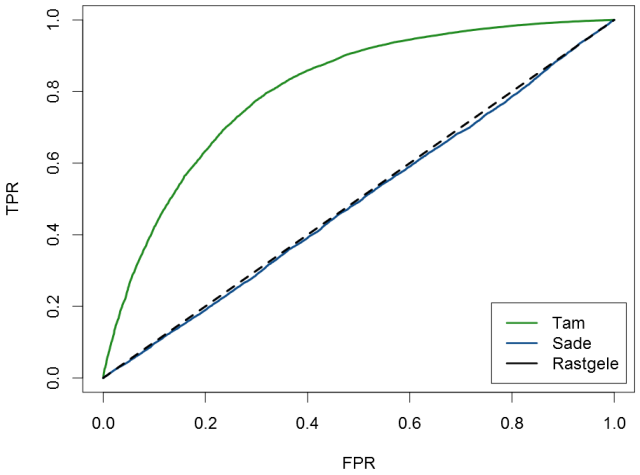
Şekil 1: RO ve DVM yöntemleri ile elde edilen TPR /FPR eğrisi sonuçları.



Şekil 2: RO ve DVM yöntemleri ile elde edilen PREC / REC eğrisi sonuçları.

Elde edilen yeni niteliklerin kullanımından önce ve sonra yapılan deneylerde Şekil 3'deki TPR/FPR eğrisi elde edilmiştir. Bu deneyde RO yöntemi kullanılmıştır. Buna göre yalnızca boyut, renk, fiyat, konum gibi mevcut niteliklerin kullanılmasına göre yeni niteliklerin eklenmesi model performansını çok ileriye taşımıştır.

RO yöntemi, sınıflandırma işlemine göre niteliklerin önemlerini hesaplar. Buna göre müşteri bazlı risk, ürün bazlı risk, teslim zamanı, fiyat, üretici bazlı risk, üyelik uzunluğu ve yaş sınıflandırmada en büyük öneme sahip nitelikler olarak öne çıkmıştır (Tablo V). Fiyat (price) dışındaki nitelikler sonradan elde edilen niteliklerdir.



Şekil 3: Yeni niteliklerin eklenmesinden önceki ve sonraki sonuçlar.

Tablo V: Niteliklerin RO yöntemine göre sınıflandırmadaki önemi.

	Ortalama Doğruluk Düşüşü	Ortalama Gini Düşüşü
riskByCustomer_2	1.060665e-01	10231.53046
riskByItem_2	3.327317e-02	4587.32103
deliveryTime	1.848463e-02	2516.37199
price	1.154662e-02	2513.97337
riskByManufacturer_2	8.626738e-03	2414.04649
membershipPeriod	7.000994e-03	3324.78814
age	2.367154e-03	2870.69529

KAYNAKÇA

[1] "Data-mining-cup (DMC) 2014 task", http://www.data-mining-cup.de/en/service/download-center/.

[2] A. Mauser, I. Bezrukov, T. Deselaers, D. Keysers, "Predicting customer behavior using naive bayes and maximum entropy", Winning the Data-Mining-Cup 2004, 2004.

[3] S. Rajagopal, Customer data clustering using data mining technique. *CoRR*, abs/1112.2663, 2011.

[4] J. Watada, K. Yamashiro, "A data mining approach to consumer behavior", In *Proceedings of the First International Conference on Innovative Computing, Information and Control - Volume 2*, ICICIC '06, pages 652–655, Washington, DC, USA, 2006. IEEE Computer Society.

[5] Leo Breiman., "Random forests", Mach. Learn., 45(1):5–32, October 2001.

[6] Cortes, C., Vapnik, V., "Support-vector network", Machine Learning, 20, 1–25. 1995.

[7] "Free software environment for statistical computing and graphics", http://www.r-project.org/.

[8] G. Adomavicius, A. Tuzhilin, "Using data mining methods to build customer profiles", Computer, 34(2):74–82, February 2001.