Available online at www.alphanumericjournal.com

alphanumeric journal
The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 5, Issue 1, 2017

Received: September 27, 2016
Accepted: May 8, 2017
Published Online: June 30, 2017

AJ ID: 2017.05.01.OR.03
DOI: 10.17093/alphanumeric.323836

Predicting Second-Hand Car Sales Price Using Decision Trees and Genetic Algorithms

Mehmet Özçalıcı | Department of International Trade and Logistics, Kilis 7 Aralık University, Turkey, mozcalici@kilis.edu.tr

ABSTRACT

It is important to predict the sales price of second-hand car for both persons and institutions who are operating in second-hand market. The sales price of cars are affected by many factors which makes predicting difficult. Especially there is no readily available method to determine which factors are affecting the sales price most. The purpose of this study is to predict the sales price of second-hand cars with decision trees. Genetic Algorithm is used to select the most relevant features. For this purpose, 252645 advertisements are scanned fort his study. For each advertisement there are 139 features available. Different models are examined using genetic algorithms with selecting 5, 10, 15 and 20 features. The best predicting performance in the out-of-sample experiment is %65,67. Proposed model can be used as a decision support system for those operating in second-hand car market.

Keywords: Data Mining, Decision Trees, Genetic Algorithm, Second-Hand Cars, Predicting

Karar Ağaçları ve Genetik Algoritmalar ile İkinci El Otomobil Satış Fiyat Tahmini

ÖZET

İkinci el araçların satış fiyatlarının önceden tahmin edilmesi, ikinci el piyasada alım satımla ilgilenen kişi ve kurumlar için önem arz etmektedir. Ancak araçların fiyatlarının birçok faktörden etkilenmesi, satış fiyatının tahmin edilmesini zorlaştırmaktadır. Özellikle araçlara ilişkin hesaplanabilen birçok değişken arasında hangilerinin en iyi tahmin performansını sergileyeceğini belirlemeye yönelik hazır bir yöntem mevcut değildir. Bu çalışmanın amacı karar ağaçları ile ikinci el otomobil satış fiyatını tahmin etmektir. Karar ağaçları için değişken seçimi genetik algoritma ile gerçekleştirilmiştir. Bu amaçla Türkiye'de faaliyet gösteren e-ticaret sitelerinden birinde yer alan 252645 adet otomobile ait satış ilanı taranmıştır. Her bir otomobile ilişkin 139 adet değişken mevcuttur. Genetik Algoritmalar yardımıyla sırasıyla 5, 10, 15 ve 20 adet değişkenin seçildiği modeller incelenmiştir. %65.67 ye varan oranda doğru tahmin gerçekleştirilebilmiştir. Önerilen yöntemi, ikinci el piyasada işlem yapan taraflar karar destek sistemi olarak kullanabilirler.

Anahtar Kelimeler: Veri Madenciliği, Karar Ağaçları, Genetik Algoritma, İkinci El Otomobil, Tahmin

1. Giriş

İkinci el piyasadan araçlarını almayı düşünen kişi veya kurumlar için aracın gerçek değerini belirlemek oldukça önemlidir. Aynı şekilde araçlarını ikinci el piyasada satmak isteyen kişi veya kurumlar için satış fiyatını belirlemek de aynı derecede önemlidir. İkinci el piyasada alım satım gerçekleştirecek taraflar, araçlarını gerçek değerinde alıp satmak isteyeceklerdir. Ancak ikinci el araçların fiyatları birçok faktörden etkilenmektedir. Bu durum satış fiyatlarının tahmin edilmesini zorlaştırmaktadır. Bu nedenle araçların satış fiyatı genellikle bu piyasada uzman olan kişiler tarafından belirlenmektedir. Literatürde ikinci el araçların satış fiyatlarını yapay sinir ağları (Asilkan ve Irmak, 2009; Ecer, 2013; İşeri ve Karlık, 2009), hedonik model (Daştan, 2016; Erdem ve Şentürk, 2009; Hadinejad ve Shabgard, 2011), ANFIS (Wu vd, 2009) ve C4.5 karar ağacı (Pudaruth, 2014) teknikleri ile tahmin eden çalışmalar mevcuttur. Bu çalışmalarda başarılı tahmin sonuçları elde edilmektedir. Ancak iyi bir tahmin modelinde hangi değişkenlerin yer alması gerektiğine ilişkin hazır bir teknik mevcut değildir.

n adet değişkenin bulunduğu bir veri setinden m adet değişkenin seçilmesi durumunda olası sonuçların sayısı $\binom{n}{m}$ formülü ile hesaplanmaktadır ve çok büyük veri setleri için bu problemdeki bütün olasılıkların tek tek denenmesi imkansızdır (Oreski ve Oreski, 2014:2053).

Veri madenciliği ve örüntü keşfi (pattern recognition) için özellik seçimi prosedürü özellikle büyük hacimli veri setleri ile işlem yapılan analizler için önemli bir görevdir ve amacı birbirlerine benzemeyen özellik alt kümesini seçmektir (Wan vd 2016). Özellik seçimi (feature selection) veri madenciliğinde önemli ön işleme (preprocessing) yöntemlerinden birisidir (Tsai vd, 2013:240). Populasyon tabanlı bir arama stratejisine sahip Genetik algoritma yöntemi literatürde, değişken seçimi için yaygın bir şekilde kullanılmaktadır (Ghareb vd 2016; Wan vd, 2016; Oreski ve Oreski, 2014; Tsai vd 2013; Tong ve Mintram, 2010).

Öte yandan, karar ağacı yöntemi, şartlı olasılıklarla yaratılan ve çıktıları (sonuçları, maliyetleri, olayları vb) ağaç şeklinde gösterilebilen bir sınıflandırma yöntemdir (Geetha ve Nasira, 2014). Karar ağaçlarında kullanılan çıktı değişkeni kategorik olduğu durumda sınıflandırma ağacı (classification tree), çıktı değişkeni sürekli bir değişken olduğunda ise regresyon ağacı (regression tree) olarak adlandırılır. Karar ağaçları ile gerçekleştirilen tahmin analizi zengin bir şekilde görselleştirilmektedir. Bu durum karar ağaçlarının, uzman olmayan kişiler tarafından bile kolay değerlendirilmesini sağlayan çıktılar oluşturmasına olanak sağlamaktadır.

Bu çalışmanın amacı ikinci el otomobil fiyatlarını sınıflara ayırmak ve araçların diğer özelliklerinden yola çıkmak suretiyle aracın ait olacağı fiyat sınıfını tahmin etmeye çalışan bir model geliştirmektir. Çalışmada analiz için görsel değerlendirmeye olanak sağlayan karar ağaçları kullanılmıştır. Bir araç için çok fazla sayıda değişkenin (özelliğin) depolanması ve kaydedilmesi mümkündür. Bu çalışmada her bir araç için 139 adet değişken hesaplanmıştır. Bu değişkenler arasında en doğru tahmini gerçekleştirecek değişkenler genetik algoritma ile seçilmiştir. Bu şekilde tasarlanan uzman sistemi, ikinci el araç piyasasında işlem yapan kişi ve kurumlar, alım satım işlemlerinde karar destek sistemi olarak kullanabilirler.

Çalışma beş bölümden oluşmaktadır. Bu giriş bölümünden sonra, ikinci bölümde literatür özeti yer almaktadır. Üçüncü bölümde çalışmada kullanılan analiz teknikleri hakkında temel bilgiler sunulmaktadır. Dördüncü bölümde veri seti ve analiz sonuçları yer almaktadır. Beşinci bölümde ise sonuç ve tartışma yer almaktadır.

2. Literatür Özeti

Karar ağaçları ile literatürde farklı işletmecilik problemlerinin çözümü için başarılı ile kullanılmaktadır. Bu çalışmalardan bazılarını şu şekilde özetlemek mümkündür. Chen (2016) çalışmasında hileli finansal tabloların belirlenmesinde veri madenciliği yöntemlerini karşılaştırmıştır. Çalışma sonuçlarına göre %87 oranında doğru bir şekilde hileli tablolar belirlenebilmiştir. Panigrahi ve Mantri (2015) çalışmalarında metin-tabanlı (text based) karar ağacı yöntemini (C4.5) hisse senedi fiyat tahmini için uygulamışlardır. Önerilen metin-tabanlı yöntemin normal karar ağacı yönteminden daha iyi sonuç ortaya çıkaramadığını raporlamaktadırlar.

Liu vd 2013 çalışmalarında lisans mezunu öğrencilerin sayısının artması üzerine işsizliğin sosyal bir problem haline geldiğini vurgulamaktadır ve çalışmalarında mezun öğrencilerin istihdam durumunu tahmin etmek için karar ağacı tabanlı bir yöntem önermektedirler. Çalışmalarında genetik algoritmayı özellik seçimi için kullanmışlardır. Ma (2013) çalışmasında karar ağacı (CART) yöntemini kullanmak suretiyle hisse senedi kapanış fiyatının bir sonraki güne ilişkin yönünü tahmin etmiştir. Çalışma sonunda %50'nin üzerinde doğru bir şekilde tahmin gerçekleştirilebildiği raporlanmaktadır.

Liu ve Jiang (2009) çalışmalarında finansal risk tahmini için hem finansal bilgilerin hem de finansal olmayan bilgilerin kullanılmasını önermektedirler. Ayrıca modelleme sürecine C4.5 karar ağacını eklemişlerdir. Xiao vd (2006) çalışmalarında müşterilerin kredi skorlarını modellemek için farklı veri madenciliği tekniklerini bir arada kullanmışlar ve sonuçları karşılaştırmışlardır. Çalışmaları sonunda Destek Vektör Makineleri yönteminin başarılı bir şekilde tahmin gerçekleştirdiğini buna rağmen, karar ağaçları yönteminin açıklayıcı gücünün daha yüksek olduğunu raporlamaktadırlar. Emel ve Taşkın (2005) çalışmalarında perakendeci bir işletmenin müşterilerine göre kişiselleştirilmiş satış hareketlerinin içeren veri tabanını kullanmak suretiyle satış analizi gerçekleştirmiştir. Analiz yöntemi olarak karar ağacı (CART) tekniğinden yararlanılmıştır. Zekic-Susac vd (2004) çalışmalarında Hırvatistan'da faaliyet gösteren bir bankadan edindikleri veri seti üzerini kullanmak suretiyle, küçük işletmelerin kredi derecelerini modellemişlerdir. Modellemek için yapay sinir ağları, lojistik regresyon ve karar ağacı (CART) olmak üzere üç farklı model kullanmışlar ve sonuçları karşılaştırmışlardır.

Literatürde yayınlanan çalışmalar incelendiğinde, karar ağaçlarının farklı amaçlar doğrultusunda kullanıldığı, buna karşın ikinci el otomobil satış fiyatı için nadir uygulandığı ortaya çıkmıştır. Bu çalışmada ise literatürden farklı olarak değişken seçim yöntemi ile entegre edilen bir yöntem yardımıyla ikinci el otomobil satış fiyat tahmin işlemi gerçekleştirilecektir.

3. Metodoloji

Çalışmanın bu bölümünde, karar ağaçları ve genetik algoritmalar hakkında temel tanımlayıcı bilgilere yer verilecektir.

3.1. Karar Ağaçları

Karar ağacı grafiğinde kök düğüm, dallar ve yapraklar bulunmaktadır. Yapraklar sınıflandırmanın oluştuğu yerlerdir, dallar ise her bir olayın sonucunu içermektedir. Sınıflandırma kurallarının oluşturulmasında, kök düğümden yaprak düğümlere doğru giden yollar göz önünde bulundurulur (Geetha ve Nasira, 2014).

CART algoritması eğitim setini her bir yaprak olmayan düğümlerde iki parçaya bölmektedir ve bölme işlemi eğitim setimi tamamen ayrışmaz (inpseparable) duruma geldiğinde durmaktadır (Ma, 2013). Karar ağaçları, ağacın oluşturulması (building the tree) ve budanması (cutting the tree) süreçlerinden oluşmaktadır. Ağacın oluşturulması sürecinde ilk olarak kök düğüme karar verilmektedir. Bu kök düğümüne eğitim setindeki bütün gözlemler atanmaktadır. Daha sonra yinelemeli süreç yardımıyla yeni düğümler yaratılmaktadır ve sonuçta tamamen gelişmiş bir ağaç ortaya çıkmaktadır. Düğümler yaratılırken, hesaplanan Gini indeksininençoklanması amaçlanmakta ve böylelikle optimal çözüme erişilmektedir (Ma, 2013:160).

Ancak veri madenciliği yöntemleri aşırı uyum (overfitting) tehlikesi ile karşı karşıya kaldığından ağacın budanması gerekmektedir. Karmaşıklık parametresi (complexity parameter) alt ağaçlar için hesaplanmaktadır ve yüksek karmaşıklığa sahip alt ağaçlar budanmaktadır (Ma, 2013:160).

Karar ağacı yöntemi, ağaç şeklinde bir grafik oluşturduğundan görsel bir değerlendirme imkânı sunmaktadır. Bu özellik karar ağacı yönteminin önemli bir avantajıdır. Ayrıca İş zekâsı (Business Intelligence) için ideal olması, geleneksel istatistiksel yöntemlerin ikamesi olarak kullanılabilmesi, niteliksel ve niceliksel veri setini destekliyor olması gibi avantajları da mevcuttur (Geetha ve Nasira, 2014). Karar ağaçlarının bazı dezavantajlarını ise şu şekilde sıralamak mümkündür (Dahan vd, 2014:8). Karar ağaçlarının sağlıklı işleyişi, çok fazla birbirleriyle ilişkili değişkenlerin girdi setinde kullanıldığı durumlarda tehlikeye girmektedir. Çok fazla sayıda değişken girdi olarak kullanıldığında ağacın boyutu incelemeyi zorlaştıracak düzeyde karmaşık hale gelmektedir. Birçok uygulamada karar ağaçlarında (ve kurallarda) analize sunulan değişkenlerden az bir kısmının kullanılmaktadır.

3.2. Genetik Algoritmalar

Genetik Algoritmalar popülasyon bazlı evrimsel algoritmalardandır ve ilk kez Holland tarafından 1975 yılında önerilmiştir. GA doğanın evrim mekanizmasını kullanmaktadır ve doğal seçim sürecini taklit etmek suretiyle optimal bir çözüm aramaktadır (Ghareb, 2016:32). Genetik algoritmalar özellikle sezgisel bir bilginin var olmadığı optimizasyon problemlerinin çözümü için başarılı sonuçlar üretmektedir (Wan vd, 2016:250).

Genetik algoritmaların işleyişini şu şekilde özetlemek mümkündür (Tsai vd, 2013:241). Genetik algoritmalarda kromozom olarak adlandırılan bireylerden oluşan bir ana kütle üzerinde işlem gerçekleştirilmektedir. Kromozomlar, optimizasyon problemi için bir çözümü temsil etmektedir. Bir sonraki nesli oluşturmak için, bu ana kütlede yer alan bireylere genetik işlemler uygulanmaktadır. İki adet temel genetik

işlem söz konusudur. Bunlar çaprazlama ve mutasyon işlemlerdir. Çaprazlama işlemi ile birlikte, eski ana kütleden seçilen iki adet birey üzerinde bilgi değişimi gerçekleştirilmektedir ve bir sonraki nesil için iki farklı birey oluşturulmaktadır. Buna karşın mutasyon işleminde bir bireydeki bilgi değişikliğe uğramaktadır. Bununla birlikte uygunluk fonksiyonu bir bireyin bir sonraki nesiller boyunca hayatta kalmasını sağlayacak bir değeri içermektedir.

Genetik algoritmaların bazı avantaj ve dezavantajları mevcuttur. Avantajları arasında şunlar sıralanabilir (Haupt ve Haupt, 2004:23). Sürekli ve kesikli değişkenlerle işlem gerçekleştirebilir, türev hesaplamalarına gerek yoktur, eşzamanlı arama gerçekleştirmektedir, çok fazla sayıda değişkenle işlem yapılabilir, donanımın elverdiği ölçüde paralel işlem gerçekleştirilebilir. Bununla birlikte, uygunluk fonksiyonun her bir problem için ayrı ayrı tasarlanması gerekmesi, bazı parametrelere karar verilmek zorunda olunması, çalışmanın ne zaman durdurulacağına ilişkin etkin bir yöntemin bulunmaması, uygunluk fonksiyonunun çok fazla sayıda çalıştırılması gerekmesi gibi bazı dezavantajları mevcuttur (Svanandam ve Deepa, 2008: 35).

4. Analiz

Çalışmada ilk olarak geniş nitelikli veri seti bir araya getirilmiştir. Söz konusu veri seti ham haldedir ve analizde kullanılmadan önce ön-işleme işlemine tabi tutulmalıdır. Bir sonraki adımda, çalışmada tanıtılan uygunluk fonksiyonuna sahip genetik algoritma ile optimizasyon süreci yer almaktadır. Optimizasyon süreci sonunda optimal değişkenler belirlenmektedir. Son olarak veri setinden rastgele 1000 adet araç seçilmektedir. Bu araçların 700 tanesi optimal olarak belirlenen değişkenlerle eğitime tabi tutulmaktadır. Eğitilen modeller, 300 adet test setinde denenmektedir ve bu performans değerlendirmeye tabi tutulmaktadır.

4.1. Veri Seti

Çalışmada ihtiyaç duyulan veri seti mevcut değildir. Bu nedenle Türkiye'nin önde gelen e-ticaret sitelerinden olan sahibinden.com sitesinde yer alan otomobil ilanları kullanılmıştır. 2016 yılının Temmuz ve Ağustos aylarında yer alan ilanlar taranmış ve 410470 adet araç ilanına ilişkin bilgiler bir araya getirilmiştir. Bununla beraber aynı araca ilişkin çeşitli sebeplerden dolayı iki kere okuma gerçekleştirilmiş olabilir. Buna ilişkin tarama gerçekleştirilmiş ve aynı ilanlar veri setinden çıkarılmıştır. Aynı şekilde uç değerlerin de veri setinden çıkarılması gerekmektedir. Bu amaçla veri setinde yer alan sayısal değişkenler büyükten küçüğe doğru sıralanmış ve en yüksek yüzde beşlik dilimde ve en düşük yüzde beşlik dilimde yer alan ilanlar uç değer olarak nitelendirilmiş ve veri setinden çıkarılmıştır. Tarama yapılan sayısal değişkenler şunlardır: fiyat, model yılı, aracın kilometresi, motor hacmi, motor gücü, tork değerleri, şehir içi yakıt tüketimi, şehir dışı yakıt tüketimi ortalama yakıt tüketimi ve taşıma kapasitesi. Bu işlem sonunda veri setinde 252645 adet özgün ve analize uygun ilan kalmıştır. Veri setinde yer alan araçların sayısal değerlerine ilişkin tanımlayıcı istatistikleri Tablo 1'de sunulmuştur.

	Min	Max	Ortalama	Std spm	Medyan
Fiyat (TL)	13750	104500	40702,19	16268,85	38000
Yıl	1996	2015	2008,90	4,65	2011
Km (km)	0	254000	109887,22	60974,40	107000
Motor Hacmi (cc)	1229	1998	1504,07	144,35	1560
Motor Gücü (hp)	70	175	104,17	20,02	100

	Min	Max	Ortalama	Std spm	Medyan
Şehir içi yakıt tüketimi (lt/100km)	4,50	15,80	7,33	2,03	6,80
Şehir dışı yakıt tüketimi (lt/100km)	3,50	8,90	4,70	0,88	4,50
Ortalama yakıt tüketimi (lt/100km)	3,90	10,80	5,65	1,28	5,30
Net ağırlık (kg)	815	1768	1224,51	145,27	1235
Yıllık MTV (TL x 2)	33	688	285,76	132,88	217

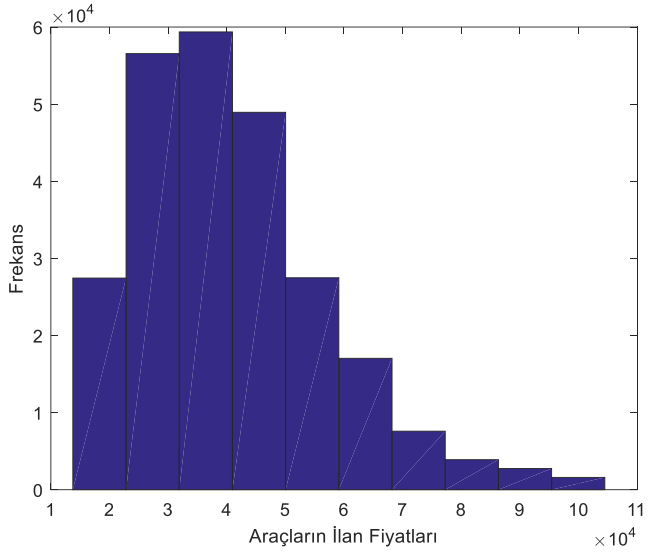
Tablo 1. Seçilmiş sayısal değişkenlerin tanımlayıcı istatistikleri

Çalışmada kullanılan değişkenler ve türleri Tablo 2'deki gibidir. Tablo 2'de fiyat sınıfı değişkeni çıktı değişkeni olarak kullanılmıştır ve bu nedenle de sıralamaya dâhil edilmemiştir.

		Değişken Adı			Değişken Adı			Değişken adı
Temel Bilgiler		Fiyat sınıfı [k]	İç Donanım	48	Anahtarsiz Ç. [i]	Multimedya	94	kasetcalar [i]
	1	Marka [k]		49	Alti ileri vites [i]		95	Cd calar [i]
	2	Yıl [s]		50	Yedi ileri vites [i]		96	mp3 calar [i]
	3	Yakıt Türü [k]		51	Hidrolik Direksiyon [i]		97	Navigasyon [i]
	4	Vites Türü [k]		52	Fonks. Direksiyon [i]		98	Telefon [i]
	5	KM [s]		53	Ayarlanabilir Direks. [i]		99	USB [i]
	6	Kasa [k]		54	Deri Direksiyon [i]		100	AUX [i]
	7	Motor Hacmi [s]		55	Ahsap Direksiyon [i]		101	Ipod bağlantısı [i]
	8	Motor Gücü [s]		56	Isıtmalı Direksiyon [i]		102	6+hoparlör [i]
	9	Çekiş [k]		57	E. Koltuk [i]		103	Cd değiştirici [i]
	10	Renk [k]		58	Hafızalı Koltuk [i]		104	Arka eglnce pak. [i]
	11	Garanti [i]		59	Katlanır Koltuk [i]		105	Dvd deg [i]
Güvenlik	12	ABC [i]		60	Ön ısıtmalı koltuk [i]	Boyalı ve Değişen Parça	106	On tampon [k]
	13	ABS [i]		61	Arka ısıtmalı koltuk [i]		107	On kaput [k]
	14	ASR [i]		62	Soğutmalı koltuk [i]		108	Tepe [k]
	15	ESP [i]		63	Hız sabitleyici [i]		109	On sag camurluk [k]
	16	Airmatic [i]		64	Sogutmalı torpido [i]		110	On sag kapi [k]
	17	EDL [i]		65	Yol bilgisayarı [i]		111	Arka sag kapi [k]
	18	EBD [i]		66	Krom kaplama [i]		112	Arka sag çamrlk [k]
	19	TCS [i]		67	Ahsap kaplama [i]		113	On sol camurluk [k]
	20	BAS [i]		68	Head-up display [i]		114	On sol kapi [k]
	21	Distronic [i]		69	Start / stop [i]		115	Arka sol kapi [k]
	22	Yokuş Klkş Dest [i]		70	Geri gorus kamerası [i]		116	Arka sol çamrlk [k]
	23	Gece Görüş [i]		71	On gorus kamerası [i]		117	Arka kaput [k]
	24	Şerit Ayrlma İkaz [i]		72	3. sıra koltuk [i]		118	Arka tampon [k]
	25	Şerit Değiş. Yard. [i]	Dış Donanım	73	Hardtop [i]	Teknik Özellikler	119	Kapı Sayısı [s]
	26	Sürücü Hava Y. [i]		74	Led Far [i]		120	Maximum Torka [s]
	27	Yolcu Hava Y. [i]		75	Xenon Far [i]		121	Maximum Torkb [s]
	28	Yan Hava Y. [i]		76	Bixenon Far [i]		122	Vites Sayısı [s]
	29	Diz Hava Y. [i]		77	Sis Far [i]		123	Hizlanma [s]
	30	Perde Hava Y. [i]		78	Adaptif Far [i]		124	Azami Surat [s]
	31	Tavan Hava Y. [i]		79	Gece Far Sensörü [i]		125	Silindir Sayısı [s]
	32	Lastik Arıza Gös. [i]		80	Far Yikama [i]		126	Ş. ici Yakit Tuk [s]
	33	Yorgunluk Tespit [i]		81	Elektrikli Ayna [i]		127	Ş. Disi Yakit Tuk [s]
	34	Isofix [i]		82	Katlanir Ayna [i]		128	Ort. Yakit Tuk [s]
	35	Alarm [i]		83	İsitmali Ayna [i]		129	Depo Hacmi [s]
	36	Merkezi Kilit [i]		84	Hafizali Ayna [i]		130	Oktan [s]
	37	Immobilizer [i]		85	Arka Park Sensor [i]		131	Koltuk Sayısı [s]
İç Donanım	38	Deri Koltuk [i]		86	On Park Sensor [i]		132	Uzunluk [s]
	39	Kumas Koltuk [i]		87	Alasimli Jant [i]		133	Genislik [s]
	40	Deri/Kumas Kltk. [i]		88	Sunroof [i]		134	Yukseklik [s]
	41	E. Ön Cam [i]		89	Pan. Cam Tavan [i]		135	Net Agirlik [s]
	42	E. Arka Cam [i]		90	Yagmur Sensoru [i]		136	Tasima Kap [s]
	43	Analog Klima [i]		91	Arka cam buz çözücü [i]		137	Bagaj Kap [s]
	44	Dijital Klima [i]		92	Panoramik on cam [i]		138	Yillik MTV [s]
	45	O. Kar. Dikiz Ayn [i]		93	Romork Çeki Demiri [i]		139	Arac Segment [k]
	46	Ön Kol Dayama [i]	[s]: Sayısal, [k]: Kategorik, [i]: İkili					
	47	Arka Kol Dayama [i]						

Tablo 2. Çalışmada kullanılan değişkenler

Çıktı değişkeni ise fiyatların kategorik hale getirilmesinden oluşmaktadır. Sayısal olan fiyat değişkenine ilişkin histogram grafiği Şekil 1'deki gibidir. Şekilden de görülebildiği gibi veri setinde yer alan araçların fiyatları için sağa çarpık (pozitif asimetri) bir dağılım söz konusudur.



Şekil 1. Fiyata İlişkin Histogram Grafiği

Her grupta mümkün olduğu kadar eşit sayıda ilan olması istenmektedir. Göz kararı ile oluşturulan gruplar ve bu gruplardaki ilan sayıları Tablo 3'deki gibidir. Benzer bir uygulama (Pudaruth, 2014) çalışmasında yer almaktadır.

Çalışmadaki Endeksi	Fiyat Aralığı	Frekans	Yüzde (%)
1	10 000 – 30 000	72048	28,52
2	30 000 – 40 000	66408	26,29
3	40 000 – 50 000	52696	20,86
4	50 000 – 150 0000	61493	24,34

Tablo 3. Kesikli Hale Getirilen Fiyat Değişkeni

4.2. Parametreler ve Uygunluk Fonksiyonun Tasarımı

Çalışmada analizleri gerçekleştirmek için MATLAB yazılımı kullanılmıştır. Çalışmada kullanılan genetik algoritmaya ilişkin parametreler şu şekildedir: Ana kütledeki birey sayısı 200, elit sayısı 10, çaprazlama oranı 0.8, mutasyon oranı 0.2, nesil sayısı 2000. Ayrıca, en iyi uygunluk değerinde son elli nesilde bir iyileşme görülmediği takdirde genetik algoritma çalışmasını durdurmaktadır.

Çalışmada kullanılan karar ağacına ilişkin parametreler ise şu şekildedir. Her bir hatanın maliyeti eşit (1) kabul edilmiştir. Her bir gözlemin ağırlığı eşit kabul edilmiştir. Ayrıştırma (split) için Gini indeksi kullanılmıştır. Budama işlemi (prune) ve yaprakların birleştirilmesi (merge leaves) özellikleri açıktır. En fazla 20 adet bölünmeye (maximum split) olanak tanınmaktadır.

Bu çalışma için özel bir uygunluk fonksiyonu tasarlanmıştır. Sadece bir adet veri seti üzerinde işlem gerçekleştirmek rastgele etkilerden dolayı hatalı sonuçlar üretebilecektir. Bu nedenle uygunluk fonksiyonunda 20 adet çapraz doğrulama gerçekleştirilmiştir. Uygunluk fonksiyonu ilk olarak veri setinden rastgele 700 adet veri eğitim amacı ile seçilmekte ve 300 adet gözlem ise test amacı ile kullanılmaktadır.

Uygunluk fonksiyonu kromozomda kodlanan değişkenleri ve 700 adet eğitim setini kullanmak suretiyle karar ağacının eğitimini gerçekleştirmekte ve bu modeli 300 adet test setinde uygulamaktadır. Bu 700 adet eğitim setinin seçilme ve 300 adet test seti üzerinde performansı kaydetme işlemi 20 kere gerçekleştirilmektedir. Uygunluk fonksiyonun çıktısı bu 20 adet test performansının aritmetik ortalamasıdır.

4.3. Optimal Değişkenlerin Belirlenmesi

Kaç adet değişken seçilmesi gerektiğini belirleyecek hazır bir yöntem söz konusu değildir. Bu nedenle çalışmada 5, 10, 15 ve 20 adet değişken seçilmiştir. Her bir model için ayrı ayrı sistem çalıştırılmış ve seçilen optimal değişkenler Tablo 4'de sunulmuştur. Tablo 4'deki verilerden, istisnasız her dört model için de yıl ve motor gücü değişkenlerinin optimal sette yer aldığı anlaşılmaktadır. Tabloda ayrıca 10, 15 ve 20 adet değişkenin seçildiği durumda arka kaput, ön sağ çamurluk, arka tampon gibi aracın boyalı, orijinal ve değişen parçalarının olup olmadığı bilgisini içeren değişkenlerin yer aldığı görülmektedir.

Değişken Sayısı	**Optimal Değişkenler**
5	Motor Yılı, Motor Gücü, Çekiş, Yokuş Kalkış Desteği, Genişlik
10	Motor Yılı, Vites Türü, Motor Gücü, Çekiş, EDL, Hareketli Koltuk, Soğutmalı Torpido, USB, Arka Kaput, Depo Hacmi
15	Marka, Model Yılı, Motor Gücü, ABS, ESP, Yan Hava Yastığı, Immobilizer, Arka Kol Dayama, Ön Isıtmalı Koltuk, Alaşımlı Jant, USB, Ön Sağ Çamurluk, Ön Sağ Kapı, Maksimum Tork(b), Koltuk Sayısı
20	Model Yılı, Motor Gücü, Çekiş, Yokuş Kalkış Desteği, Elektrikli Ön Cam, Anahtarsız Çalıştırma, Altı İleri Vites, Hareketli Koltuk, Soğutmalı Koltuk, Ahşap Kaplama, Led Far, Far Yıkama, Ön Sensor, Panoramik Cam Tavan, MP3, AUX, Ön Sağ Kapı, Arka Tampon, Genişlik, Taşıma Kapasitesi

Tablo 4. Seçilen Optimal Değişkenler

Seçilen değişkenlerin birbirleri ile düşük korelasyon katsayısına sahip olması beklenmektedir. Seçilen değişkenler arasındaki ikili Spearman korelasyon katsayıları hesaplanmıştır. Paramterik olmayan nitelikteki Spearman korelasyon katsayısının seçilmesinin nedeni bazı değişkenlerin kategorik olmasıdır. Her bir ikili karşılaştırmaya yer vermek çalışmanın hacmini arttıracaktır. Bu nedenle her model için sadece en yüksek korelasyon katsayısı ve bu katsayının hesaplandığı iki değişken Tablo 5'de raporlanmıştır. Tablo 5'de görüldüğü gibi her model için en yüksek korelasyon katsayısı 0.52 nin altındadır. Bu durum seçilen optimal değişkenlerin birbirleri ile neredeyse hiç ilişkili olmadığını göstermektedir. Bu bulgu, değişken seçiminin iyi bir şekilde yapıldığını göstermektedir.

Model	En yüksek K. Katsayısı	p	En yüksek kor. gerçekleştiği değişkenler
5	0.3927	~0	Motor Yılı – Yokuş Kalkış Desteği
10	0.5181	~0	Motor Gücü – Depo Hacmi
15	0.4572	~0	ABS - Immobilizer
20	0.5035	~0	Model Yılı – MP3 çalar

Tablo 5. Her modeldeki en yüksek Spearman Korelasyon Katsayısı

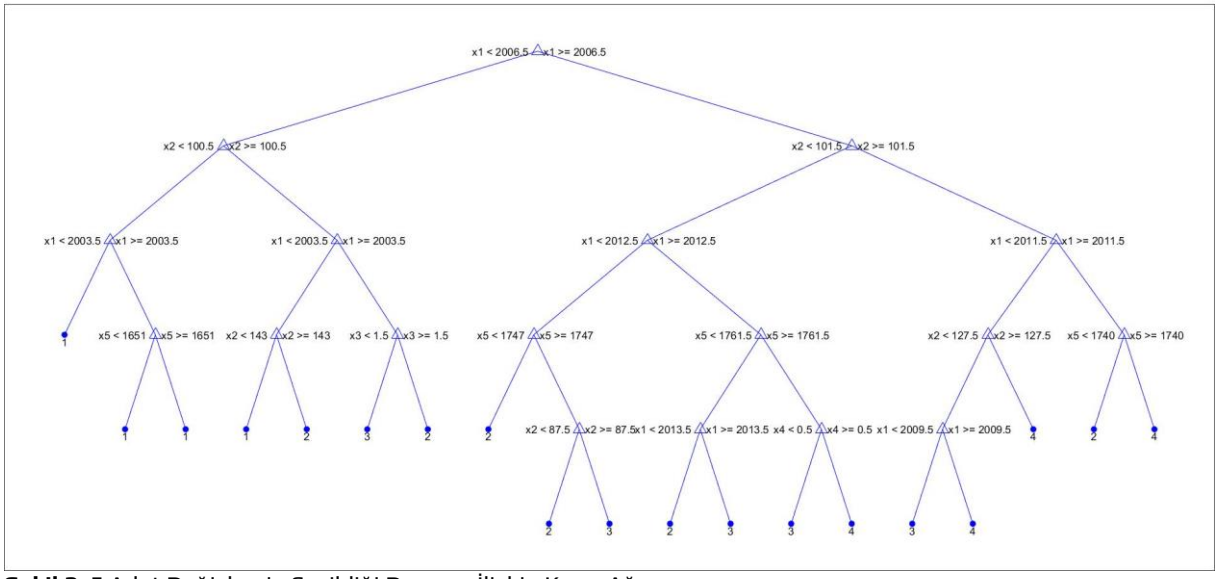
Her bir modelin eğitim sürecinde ve test sürecindeki performansları Tablo 6'da yer almaktadır. Bu tabloda pozitif sınıf her modelde birinci sınıfken, negatif sınıf diğer sınıflardan (2, 3 ve 4. sınıflardan) oluşmaktadır. Tabloda göze çarpan sonuçlardan ilki eğitim sürecinde ölçülen performansların örneklem dışı test sürecine göre daha yüksek olmasıdır. Tabloda yer alan sonuçlardan yola çıkmak suretiyle 10 adet değişkenin seçildiği modelde en yüksek örneklem dışı performansın elde edildiğini söylemek mümkündür (%65.67). Daha fazla sayıda değişken seçildiğinde modelin karmaşıklığı artmakta ve buna bağlı olarak da örneklem dışı performans düşmektedir.

		Eğitim Süreci			Test Süreci		
		Pozitif	Negatif	%	Pozitif	Negatif	%
5	Pozitif	134	20	69,43	53	12	63,67
	Negatif	58	488		25	210	
10	Pozitif	139	23	70,57	56	13	65,67
	Negatif	53	485		22	209	
15	Pozitif	157	26	70,00	60	17	61,00
	Negatif	35	482		18	205	
20	Doğru	149	29	70,86	54	18	62,00
	Yanlış	43	479		24	204	

Tablo 6. Eğitim ve Örneklem-Dışı Test Sürecinde Modellerin Performansları

4.4. Karar Ağacı ve Kurallar

Karar ağaçları ve kurallar için 5 adet değişkenin seçildiği model detaylı bir şekilde incelenmiştir. Değişken sayısı arttıkça ağaçların boyutları büyüyecek ve kural sayısı artacaktır. Bu durum incelemeyi karmaşık hale getirmektir. Bu nedenle beş adet değişkenin seçildiği modele ilişkin karar ağacı ve kurallara yer verilmiştir. Modele ilişkin karar ağacı Şekil 2'de, kurallar ise Tablo 7'de yer almaktadır.



Şekil 2. 5 Adet Değişkenin Seçildiği Duruma İlişkin Karar Ağacı

Düğüm	Kural
1	if *yıl* <2006.5 then node 2 elseif *yıl* >=2006.5 then node 3 else 2
2	if *motor gücü* <100.5 then node 4 elseif *motor gücü* >=100.5 then node 5 else 1
3	if *motor gücü* <101.5 then node 6 elseif *motor gücü* >=101.5 then node 7 else 4
4	if *yıl* <2003.5 then node 8 elseif *yıl* >=2003.5 then node 9 else 1
5	if *yıl* <2003.5 then node 10 elseif *yıl* >=2003.5 then node 11 else 1
6	if *yıl* <2012.5 then node 12 elseif *yıl* >=2012.5 then node 13 else 2
7	if *yıl* <2011.5 then node 14 elseif *yıl* >=2011.5 then node 15 else 4
8	class = 1 (10 000 < Fiyat <= 30 000)
9	if *genişlik* <1651 then node 16 elseif *genişlik* >=1651 then node 17 else 1
10	if *motor gücü* <143 then node 18 elseif *motor gücü* >=143 then node 19 else 1
11	if *çekiş* <1.5 then node 20 elseif *çekiş* >=1.5 then node 21 else 2
12	if *genişlik* <1747 then node 22 elseif *genişlik* >=1747 then node 23 else 2
13	if *genişlik* <1761.5 then node 24 elseif *genişlik* >=1761.5 then node 25 else 3
14	if *motor gücü* <127.5 then node 26 elseif *motor gücü* >=127.5 then node 27 else 4
15	if *genişlik* <1740 then node 28 elseif *genişlik* >=1740 then node 29 else 4
16	class = 1 (10 000 < Fiyat <= 30 000)
17	class = 1 (10 000 < Fiyat <= 30 000)
18	class = 1 (10 000 < Fiyat <= 30 000)
19	class = 2 (30 000 < Fiyat <= 40 000)
20	class = 3 (40 000 < Fiyat <=50 000)

Düğüm	Kural
21	class = 2 (30 000 < Fiyat <= 40 000)
22	class = 2 (30 000 < Fiyat <= 40 000)
23	if *motor gücü* <87.5 then node 30 elseif *motor gücü* >=87.5 then node 31 else 2
24	if *yıl* <2013.5 then node 32 elseif *yıl* >=2013.5 then node 33 else 3
25	if *yokuş kalkış desteği* <0.5 then node 34 elseif *yokuş kalkış dest* >=0.5 then node 35 else 4
26	if *yıl* <2009.5 then node 36 elseif *yıl* >=2009.5 then node 37 else 2
27	class = 4 (50 000 < Fiyat <= 150 000)
28	class = 2 (30 000 < Fiyat <= 40 000)
29	class = 4 (50 000 < Fiyat <= 150 000)
30	class = 2 (30 000 < Fiyat <= 40 000)
31	class = 3 (40 000 < Fiyat <=50 000)
32	class = 2 (30 000 < Fiyat <= 40 000)
33	class = 3 (40 000 < Fiyat <=50 000)
34	class = 3 (40 000 < Fiyat <=50 000)
35	class = 4 (50 000 < Fiyat <= 150 000)
36	class = 3 (40 000 < Fiyat <=50 000)
37	class = 4 (50 000 < Fiyat <= 150 000)

Tablo 7. Çalışmada oluşturulan kurallar

5. Sonuç

Çalışmada, ikinci el otomobil piyasasından elde edilen veri setini kullanmak suretiyle, ikinci el otomobillerin satış fiyatı tahmin edilmeye çalışılmıştır. İlanlarda araçlara ilişkin çok fazla sayıda değişken kaydedilmektedir. Bu değişkenlerden hangilerinin satış fiyatını başarılı bir şekilde tahmin edeceği genetik algoritma ile belirlenmiştir.

Çalışmada ortaya çıkan sonuçlardan biri aracın model yılı ve motor gücü değişkenlerinin her bir modelde optimal olarak seçilmesidir. Bu durumda model yılı ve motor gücü değişkenlerinin fiyatı tahmin etmekte diğer değişkenlerden daha başarılı olduğunu söylemek mümkün olacaktır.

Çalışmada genetik algoritmanın seçtiği değişkenlerin birbirleriyle ilişkili olup olmadığı parametrik olmayan Spearman korelasyon katsayısı yardımıyla incelenmiş ve değişkenlerin hiçbirisinin birbiriyle düşük düzeyde korelasyona sahip olduğu belirlenmiştir. Bu durum değişken seçiminin doğru bir şekilde yapıldığını göstermektedir.

Çalışmada her ne kadar başarılı sonuçlar elde edilmiş olsa da bazı kısıtlar mevcuttur. Çalışmada araçlardan oluşan bir veri setine sınıflandırma algoritmalarını uygulamak için fiyat değişkeninde ayrıklaştırma (discretization) yoluna gidilmiştir. Sınıflar, çizilen histogram grafiği kullanılmak suretiyle göz kararı belirlenmiştir. Diğer çalışmalarda fiyat değişkeni kategorik hale getirilirken farklı teknikler uygulanabilir. Gelecek çalışmalarda daha geniş bir veri seti ve daha güçlü donanımlar yardımıyla farklı sonuçlar elde edilebilir.

Kaynakça

Asilkan I. ve Özcan, S. (2009) "İkinci El Otomobillerin Gelecekteki Fiyatlarının Yapay Sinir Ağları ile Tahmin Edilmesi" Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(2):375-391.

Chen, S. (2016) "Detection of Fraudulent Financial Statements Using the Hybrid Data Mining Approach" SpringerPlus, 5-89.

Dahan, H., Cohen, S., Rokach, L. ve Maimon, O. (2014) Predictive Data Mining with Decision Trees, New York, Springer.

Daştan, H. (2016) "Türkiye'de İkinci El Otomobil Fiyatlarını Etkileyen Faktörlerin Hedonik Fiyat Modeli ile Belirlenmesi" Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 18(1):303-327.

Ecer, F (2013) "Türkiye'de 2. El Otomobil Fiyalarının Tahmini ve Fiyat Belirleyicilerinin Tespiti" Anadolu Üniversitesi Sosyal Bilimler Dergisi, 13(4):101-112.

Emel, G. G. ve Taşkın, Ç. (2005) "Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması" Eskişehir Osman Gazi Üniversitesi Sosyal Bilimler Dergisi, 6(2):221-239.

Erdem, C. ve Şentürk, İ. (2009) "A Hedonic Analysis of Used Car Prices in Turkey" International Journal of Economic Perspectives, 3(2):141-149.

Geetha, A. ve Nasira, G.M. (2014) "Data Mining for Meteorological Applications: Decision Trees for Modeling Rainfall Prediction" IEEE International Conference on Computational Intelligence and Computing Research.

Ghareb, A.S., Bakar, A.A. ve Hamdan, A.R. (2016) "Hybrid Feature Selection Based on Enhanced Genetic Algorithm for Text Categorization" Expert Systems with Applications, 49:31-47.

Hadinejad, M. ve Shabgard, B. (2011) "Hedonic Price for Car in Iran", Sosyal Bilimler Dergisi, 2:118-127.

Haupt, R. L. ve Haupt, S. E. (2004) Practical Genetic Algorithms, New Jersey, John Wiley & Sons.

İşeri, A. ve Karlık, B. (2009) "An Artificial Neural Networks Approach on Automobile Pricing", Expert Systems with Applications, 36:2155-2160.

Liu, C. ve Jiang, Q. (2009) "Mixed Financial Forecasting Index System Construct and Financial Forecasting Study on the C4.5 Decision Tree" International Conference on Management and Service Science.

Liu, Y., Hu, L., Yan, F. ve Zhang, B. (2013) "Information Gain with Weight Based Decision Tree for the Employment Forecasting of Undergraduates" IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing.

Ma, Y. (2013) "The Research of Stock Predictive Model Based on the Combination of CART and DBSCAN" Ninth International Conference on Computational Intelligence and Security.

Oreski, S. ve Oreski, G. (2014) "Genetic Algorithm-Based Heuristic for Feature Selection in Credit Risk Assessment" Expert Systems with Applications, 41:2052-2064.

Panigrahi, S.S. ve Mantri, J.K. (2015) "A Text Based Decision Tree Model for Stock Market Forecasting" International Conference on Green Computing and Internet of Things.

Pudaruth, S. (2014) "Predicting the Price of Used Cars Using Machine Learning Techniques" International Journal of Information & Computation Technology, 4(7):753-764.

Sivanandam, S.N. ve Deepa, S.N. (2008) Introduction to Genetic Algorithms, New York, Springer.

Tong, D.L. ve Mintram, R. (2010) "Genetic Algorithm-Neural Network (GANN): A Study of Neural Network Activation Functions and Depth of Genetic Algorithm Search Applied to Feature Selection" International Journal of Machine Learning & Cybernetics, 1:75-87.

Tsai, C.F., Eberle, W. ve Chu, C.Y. (2013) "Genetic Algorithms in Feature and Instance Selection" Knowledge-Based Systems, 39:240-247.

Wan, Y., Wang, M., Ye, Z. ve Lai, X. (2016) "A Feature Selection Method Based on Modified Binary Coded Ant Colony Optimization Algorithm" Applied Soft Computing, 49:248-258.

Wu, J.D., Hsu, C.C. ve Chen, H.C. (2009) "An Expert System of Price Forecasting for Used Cars Using Adaptive Neuro-Fuzzy Inference" Expert Systems with Applications, 36:7809-7817.

Xiao, W., Zhao, Q. ve Fei, Q. (2006) "A Comperative Study of Data Mining Methods in Consumer Loans Credit Scoring Management" Journal of Systems Science and Systems Engineering, 15(4):419-435.

Zekic-Susac, M., Sarlija, N. ve Nensic, M. (2004) "Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network and Decision Tree Models" International Conference on Information Technology Interfaces.