Social Sciences Research Journal (SSRJ)
Sosyal Bilimler Araştırma Dergisi

Aralık/December 2020, 9(4)
Başvuru/Received: 22/10/2020
Kabul/Accepted: 13/11/2020
Araştırma Makalesi/Research Article
http://dergipark.gov.tr/ssrj
http://socialsciencesresearchjournal.com

Türküz, E., Ç. Akay, E. (2020). Ekonometri ve Makine Öğrenmesi: Tercih Modelleri ve Sınıflandırma Algoritmaları Açısından Değerlendirmeler. *Social Sciences Research Journal*, 9 (4), 181-194.

Ekonometri ve Makine Öğrenmesi: Tercih Modelleri ve Sınıflandırma Algoritmaları Açısından Değerlendirmeler[1]

Elanur Türküz
Arş. Gör., İstanbul Kültür Üniversitesi
İktisadi ve İdari Bilimler Fakültesi, Girişimcilik Bölümü
Orcid: 0000-0002-5176-7792
e.turkuz@iku.edu.tr

Ebru Çağlayan Akay
Prof. Dr., Marmara Üniversitesi
İktisat Fakültesi, Ekonometri Bölümü
Orcid: 0000-0002-9998-5334
ecaglayan@marmara.edu.tr

Özet

Ekonometri ve makine öğrenmesi geniş kullanım alanlarına ve tekniklere sahiptir. Bu çalışmada ekonometride bağımlı değişkenin nitel özellik gösterdiği durumda kullanılan nitel tercih modelleri ile makine öğrenmesinde kullanılan sınıflandırma algoritmalarına yer verilmiş olup, bu doğrultuda ekonometri ile makine öğrenmesi arasında nasıl bir köprü kurulabileceğinin araştırılması amaçlanmıştır. Büyük verilerin ekonometride yarattığı sorunlar ve makine öğrenmesinin yapabileceği katkılar araştırılmış ve kestirim tabanlı sınıflandırma algoritmalarının çekimser kaldığı nedensellik araştırmalarındaki konumu incelenerek ekonometrinin sağlayabileceği katkılar ortaya konulmuştur.

Anahtar Kelimeler: Ekonometri, Makine Öğrenmesi, Nitel Tercih Modelleri, Sınıflandırma Algoritmaları, Yüksek Boyutluluk

Econometrics and Machine Learning: A Review of Choice Models and Classification Algorithms

Abstract

Both econometrics and machine learning operate in a broad area of study. Therefore, this paper limits the scope to where the dependent variable is categoric and investigates the relationship between discrete choice models and classification algorithms. In particular, we address the challenges of big data in econometrics and the contributions of machine learning. The article also gives an overview of why classification algorithms have abstained from causality and how the machine learning community could benefit from econometrics.

Key Words: Econometrics, Machine Learning, Qualitative Choice Models, Classification Algorithms, High Dimensionality

[1] Bu çalışma Elanur Türküz'ün, Prof.Dr. Ebru Çağlayan Akay'ın danışmanlığında hazırladığı doktora tezinden üretilmiştir.

Giriş

"2000 yılı civarlarında istatistik dünyasında bir ayrışma oldu." ***Bradley Efron***

Günümüz bilgi teknolojileri ve veri üretiminde yaşanan gelişmeler 21.yy'ın bilimsel araştırmalarında önemli değişikliklere neden olmaktadır (Fan vd., 2020). Bu dönemde özellikle sosyal bilimler alanında çalışan araştırmacıların belirsizliği modellemek için kullandığı istatistiksel yöntemlerin giderek daha karmaşık hale geldiği görülmektedir (Randolph ve Myers, 2013). Önceleri doğayı anlamak üzere araştırmacıların topladıkları veriler ile başlayan istatistiksel analizler, ölçüm tekniklerinin değişmesi, yeni veri tiplerinin elde edilmesi ve veri toplamanın kişilerden kurumlara geçişi ile günümüze değin varlığını koruyan bir araştırma sahası oluşturmuştur (Godin, 2009). Bugün ise bu oluşumun, bir yanda yeni veri üretim mekanizmalarından akan verilerin oluşturduğu büyük veri ile bu verilerin işlem görmesi için gerekli olan yeni teknikler diğer yanda ise istatistiksel modellemeye konu olan farklı yaklaşımların oluşturduğu kültürel ayrılıklar içerisinde harmanlandığı görülmektedir.

Son yirmi yılın literatürü, istatistiksel çıkarım tekniklerini temel alan ekonometri ile kestirimlerin amaçlandığı makine öğrenmesi arasındaki yaklaşımların daha sık vurgulandığına işaret etmektedir. Bu anlamda yapılan çalışmalara bakıldığında, 2001 yılında Leo Breiman'a ait *"Statistical Modelling: The Two Cultures"* makalesi, istatistiksel analizler arasındaki farklılıkların ortaya oldukça keskin bir şekilde konduğu ve Cox (2001)'a göre istatistiksel modellemede bir dönüm noktası olarak adlandırılan, ilham verici çalışmalar arasında yer almaktadır. Breiman (2001), bu makalede veriden öğrenmek üzere kullanılan istatistiksel tekniklerin iki farklı yaklaşımın izlerini taşıdığını öne sürmektedir. Bu yaklaşımlarından biri veri yaratma sürecinin çoğunlukta parametrik yöntemlerle stokastik bir model tarafından tanımlandığını varsayarken diğeri ise olayları veri yaratma süreci bilinmeyen ilişkilerden öğrenen algoritmik modeller olarak gösterilmiştir. Bu ayrışmanın amaç, yöntem ve düzenlenme şekilleri itibari ile birbirinden farklılıklar gösteren ekonometri ile makine öğrenmesi disiplinlerinde de etkisini gösterdiği söylenebilir.

Her iki disiplin günümüze kadar kendi alanlarında oldukça başarılı analizler icra ederken günümüz trendi olarak görülen büyük veri özellikle ekonometrinin analizlerinde temel aldığı yöntemlere farklı bir açıdan bakılmasına neden olmuştur. Verinin ölçeğinde meydana gelen artış, daha önceleri mümkün görülmeyen yapıların keşfedilmesine ve önceleri ölçülemediğinden araştırmalara konu olamamış birçok olayın yeni tip değişkenlerle analizine olanak tanımaktadır. Ancak aşina olmadığımız boyutlara sahip bu verilerin hacim, çeşitlilik ve hız gibi özellikleri, bu özelliklerle başa çıkabilecek yeni ekonometrik tekniklerin gelişimini veya farklı disiplinlerle bir araya gelerek kendisini daha güçlü bir araca dönüştürmesini gerektirmektedir. Zira *Econometrica* dergisinin ilk sayısında Ragnar Frisch'in de belirttiği gibi *"ekonometri, önlem alınmaz ise başıboş kalacak dev bir veri kütlesini düzene sokabilecek ilkeler topluluğu ihtiyacını karşılamak üzere hareket etmektedir"* (Green, 2018). Her ne kadar Frisch, "dev veri kütlesi" ile bugün sahip olduğumuz büyük verileri kastetmese de bir disiplin olarak ekonometri, bugün büyük verilere entegre edilmiş/edilmeye çalışılan teknikleri ile gerçeğe en yakın model tahminini gerçekleştirmeye bir adım daha yaklaştığı bir çağda bulunmaktadır. Bu nedenle ekonometrinin hangi durumlarda yetersiz kaldığı saptandıktan sonra bu sorunlarla mücadele edebilmek için farklı disiplinlerden hangi tekniklerin ödünç alınabileceğine odaklanmak, bilimsel gelişme açısından önemli bir adım olarak görülebilir.

Bu çalışmada, özellikle veri miktarında meydana gelen artış sonrası, iki yaklaşım arasındaki boşluk genişlemeden nasıl bir köprü kurulabileceği araştırılmakta ve sadece ekonometri alanında çalışanlara makine öğrenmesinin sağlayabileceği katkılar değil aynı zamanda makine öğrenmesi çalışanlarına ekonometrinin amaçları, teknikleri ve uygulanabilirliği noktasında fikirler vererek iki disiplin arasında bir sinerji oluşturulması amaçlanmaktadır. Her iki disiplin geniş kullanım alanlarına ve tekniklere sahip olduğundan bu çalışmada sadece ekonometride bağımlı değişkenin nitel özellik gösterdiği durumda kullanılan Nitel Tercih Modelleri (NTM) ile bu modellere kimi noktalarda benzerlik taşıyan sınıflandırma algoritmalarına yer verilmiş, iki teknik arasındaki temel farklılıklar saptanmış ve birbirlerine olan katkıları ortaya konulmuştur.

Büyük Veri ve Ekonometride Yarattığı Sorunlar

"Veride boğuluyor olsak da bilgiye olan açlığımız devam ediyor." ***Rutherford D. Roger***

Geleneksel veri işleme araçları ile analizi yapılamayan ve karmaşık bir yapıya sahip büyük miktardaki veri setlerini ifade etmek üzere *"Büyük Veri"* teriminden yararlanılmaktadır. Bu terim, verilerin boyutları düşünüldüğünde büyük hacimli ve/veya yüksek boyutlu verileri tanımlamak üzere kullanılmaktadır. Büyük veriler, satır – sütun yapılı bir veri seti için uzun ($n > p \; veya \; n \gg p$) ve/veya geniş ($p > n \; veya \; p \gg n$) özellikte olabilirler (Varian, 2014). Boyutları $p > n$ ve/veya $p \gg n$ ise bu tip veriler açıklayıcı değişken sayısının gözlem sayısından fazla olduğunu belirtmek üzere *"yüksek boyutlu veriler"* olarak adlandırılmaktadır. Gözlem sayısının açıklayıcı değişken sayısından daha hızlı arttığını ifade eden $n \gg p$ formundaki veriler ise *"büyük hacimli veriler"* olarak adlandırılmaktadır (Einav ve Levin, 2013). Her ne kadar verinin her iki formu da analizlerde kullanılacak bilgi

miktarında bir artışa neden olsa da büyük verinin bu özellikleri dikkate alınmadığında geleneksel ekonometrik analizleri bir kabusa dönüştürebilmektedir. Yüksek boyutluluk durumunda ortaya çıkan gürültü birikimi (Fan ve Fan, 2008), sahte korelasyon (Fan ve Lv, 2008) ve tesadüfi içsellik hesaplanma sorunlarına (Fan ve Liao, 2014) neden olurken büyük hacimli olma özelliği ise yüksek değişkenlik ve sapmalara (Fan vd., 2011) yol açmaktadır. Ekonometrik analizler büyük verinin yapısal özelliklerinden kaynaklanan bu sorunları dışında küçük örnek özelliklerine adapte olmuş tekniklerinin $n \gg p$ ve $p \gg n$ boyutlu veriler için yetersiz kalması sorunu ile de karşı karşıyadır.

Geleneksel istatistiksel çıkarım tekniklerini temel alan ekonometri, çoğunlukta parametre tahmincilerinin asimtotik özelliklerini p sabit iken $n \to \infty$ için araştırmaktadır. Örnekleme teorisinden hareketle geliştirilen test istatistikleri, çoğunlukta örnek sayısının artan bir fonksiyonu olduğundan $n \gg p$ durumunda güven aralıkları daralmakta, parametrelerin neredeyse tümü anlamlı bulunmakta ve çok ufak etkiler dahi anlamlı bulunarak I. Tip hata olasılığına neden olabilmektedir. Böylece tahmini yapılan modelin hatalı bulguları tutarsız model seçimi ve hatalı bilimsel sonuçlara kapı aralayabilmektedir (Lin vd., 2013). Asimtotik tahmin teorisi p'nin sabit olduğu varsayımında bulunduğundan yüksek boyutlu özellik gösteren verilerin varlığında merkezi limit teoremi ile büyük sayılar kanunu geçerliliğini yitirmektedir. Bu durum ekonometrinin istatistiksel çıkarımlarda bulunabilmesi için odağını p'nin sabit tutulmayarak yüksek boyutlu asimptotik özelliklerin araştırıldığı yeni tekniklere çevirmesini gerektirmektedir (Yao vd., 2015). Yüksek boyutluluğun bir diğer sonucu ise kendini örnekten kaynaklanan çoklu doğrusal bağlılık (ÇDB) problemi ile göstermektedir. Çok değişkenli istatistiksel analizlerde, özellikle açıklayıcı değişkenlere ait kovaryans matrisinin tersinin hesaplanmasına dayanan teknikler, ÇDB durumunda kararsız sonuçlara veya parametrelerin hesaplanamamasına neden olmaktadır (Serdobolskii, 2000). Olası çözümler arasında bazı değişkenlerin analiz dışında bırakılması yer alsa da yaşanacak bilgi kaybı neden ile tercih edilmemektedir. Bunun yerine, analizlerde açıklayıcı değişkenler arasındaki yüksek korelasyonları dikkate alan makine öğrenmesi tekniklerinden yararlanıldığı görülmektedir. Ancak bu tekniklerin istatistiksel çıkarıma adapte edilmeden kullanımı, ekonometrinin amaçladığı bulgularla tezatlık oluşturacağından makine öğrenmesinin doğrudan kullanımı, amaç kestirim olmadığı müddetçe uygun bir tercih olmayacaktır.

Ekonometrinin büyük verilerin kullanımında karşılaştığı bu sorunlar, literatürde büyük veri analitiğinin ekonometriden çok kestirim amaçlı tahminlerde bulunan makine öğrenmesi ile anılmasına neden olsa da iki disiplin arasındaki yaklaşım farklılıkları sadece büyük veriden kaynaklanmamaktadır.

Ekonometri ve Nitel Tercih Modelleri

"Önlem alınmaz ise başıboş kalacak dev bir veri kütlesini düzene sokabilecek ilkeler topluluğu ihtiyacını karşılamayı amaçlıyoruz" ***Econometric Society***

Bir terim olarak literatürde kazandırılışı 19.yy'a isabet eden ekonometri istatistik, iktisat ve matematiğin güçlü bir birleşimidir ve genel olarak iktisadi olayların deneysel olmayan verilerle sistematik analizini içermektedir (Spanos, 1986). Çoğu ekonometrik uygulama bir değişkende meydana gelen değişimin *ceteris paribus* koşulu altında diğer değişkende neden olduğu etkinin araştırılması ile ilgilenmektedir (Athey, 2019). Ancak sosyal bilimler doğası itibariyle deneysel olmayan veriler ürettiğinden ekonometrinin arayışında olduğu nedensel ilişkiler deneysel verilerden elde edilebilen nedensel çıkarımlardan farklılık göstermektedir. Ekonometri genel olarak iktisat teorisinden aldığı destek ile veri kısıtları, sosyal bilimlerde deney gerçekleştirmenin zorluğu ve etik problemlerini aşarak iki değişken arasındaki korelasyonu *ceteris paribus* koşulu altında nedensel çıkarımlar için yeterli görmektedir (Wooldridge, 2010).

Ekonometrik modeller teori-tabanlı modeller olarak bilinmekte ve gerçek hayat ilişkilerinin çözülebilmesi için matematiksel kalıpta tanımladığı iktisadi teorinin istatistiksel modellerle araştırılmasını hedeflemektedir (Sevüktekin, 2000). İyi belirlenmiş bir model, karmaşık olayların analizini kolaylaştıracağından ekonometrik bir analizin en önemli aşaması, Green (2018)'e göre tahmin edilecek modelin belirlenmesidir. Ancak gerçek dünya ile ilgili bilgileri bir model ile kesin olarak tanımlamak, insan davranışlarının tesadüfiliği dikkate alındığında, neredeyse imkânsız olduğundan seçilen modelin gerçeğin en iyi temsili olması istenmektedir. Modelin fonksiyonel şekli ile analize katılacak değişkenler, bu ilişkinin temsilinde anahtar role sahiptir. Ekonometrik modelde bağımlı değişken, açıklayıcı değişkenlerin bir fonksiyonu olup değişken seçimi ile fonksiyonel şeklin belirlenmesi noktasında iktisat teorisinden, önceki yapılan çalışmalardan ve araştırmacının uzman bilgisinden yararlanılmaktadır. Tahmin edilen fonksiyonun, bağımlı değişkenin gerçek değerleri ile uyumlu olmama olasılığına karşılık, en iyi modelin temsili için modellerde tesadüfi hata terimine yer verilmektedir. Model tahmininde çoğunlukta temeli ölçülere dayanan parametrik yöntemler kullanılmaktadır. Parametrik yöntemlerde, dağılımın önceden bilindiğini varsayılmakta ve verilerin fonksiyonel şekli bilinen bir eğriye uydurulması amaçlanmaktadır. Ancak varsayımların geçerliliği konusunda şüphe duyulması, varsayımların sağlanmaması veya ilişkiyi daha iyi açıklayabileceği düşünülen farklı esnek fonksiyonel formların araştırılması istendiğinde nonparametrik yöntemler de kullanılabilmektedir (Çağlayan, 2012). Kullanılacak model, her ne kadar, teori

tarafından belirlense de model seçimine hangi aileden başlanacağı araştırma konusu ile ilişkilidir. Araştırma problemi, bağımlı değişkenin nitel değerler aldığı kategorik yapıda bir değişken ise seçim, NTM arasından yapılmaktadır.

Tahminde kullanılacak veri seti $V = \{(\boldsymbol{x}_i, y_i) | \, \boldsymbol{x}_i \in \mathbb{R}^p, y_i \in \{1, \dots, k\}\}_{i=1}^{N}$ şeklinde tanımlarsak, burada $\boldsymbol{x}$, açıklayıcı değişkenler vektörü olup $N \times p$ boyutlu bir matristir. k, kategoriler iken $N \times 1$ boyutlu nitel bağımlı değişken ise $y_i \in \{1, \dots, k\}$ ile gösterilmiştir. NTM'nde açıklayıcı değişkenlere koşullu bağımlı değişken kategorilerinin seçim olasılıklarının modellenmesi için $f(\boldsymbol{x}) = P(Y = k|X = x)$'in örnekten tahmini yapılmaktadır. Yapılacak seçime ait olasılığın maksimize edilmesi istediğinden benzerlik fonksiyonundan ve en çok benzerlik (EÇB) tahmincisinden yararlanılmaktadır. EÇB tahmincisi, tutarlı ve asimtotik etkin özellikleri dışında aynı zamanda asimtotik normal olduğundan doğrusal olmayan model parametrelerinin istatistiksel anlamlılıkları hipotezlerle test edilebilmektedir. Böylece açıklayıcı değişkenlerin tercih olasılıkları üzerindeki etkileri araştırılabilmekte ve $\boldsymbol{x}_i$'de meydana gelen değişimin $y_i \in \{1, \dots, k\}$ kategorilerinin seçimini nasıl etkilediğine dair çıkarımlarda bulunulabilmektedir (Green ve Hensher, 2010). Nitel bağımlı değişkenin iki değer aldığı durumda ($k = 2$) kullanılan modeller ***"ikili tercih modelleri"*** olarak adlandırılmaktadır. Burada karar birimlerinin (bireyler, firmalar, kurumlar, hanehalkları vs.) tercih yapmak üzere iki seçenek ile karşı karşıya oldukları varsayılmakta ve karar birimlerine ait özelliklerin belli bir tercihte bulunma olasılığı arasındaki ilişkinin ortaya konulması amaçlanmaktadır. $f(\boldsymbol{x})$, olasılıklı bir model olarak tanımlandığından NTM'nde amaç, aynı zamanda seçimin olasılığının belirlenmesidir. Karar sürecinin olasılıklı yapısı hakkında geliştirilen farklı varsayımlara sahip çeşitli ikili tercih modelleri bulunmaktadır. Bunlar *doğrusal olasılık modeli*, *logit* ve *probit* modelleridir (Cameron ve Trivedi, 2005). Öte yanda, karar vericinin ikiden fazla seçenekle ($k > 2$) karşılaşması durumda kullanılabilecek modeller ise *multinomial logit/probit, sıralı logit/probit* gibi modeller olup ***"çoklu tercih modelleri"*** başlığı altında incelenmektedir (McFadden, 1987).

NTM, gözlemlenebilir verilere sahip çok sayıda birey, kurum veya birimin davranışları ile ilgilenmektedir. Bu modellerde benimsenen fayda teorisi bireylerin veya kurumların birim bazındaki davranışlarının ne olduğu dışında, nasıl şekillendiği ile de ilgilenmektedir. Çoğunlukta spesifik durumlar yerine *"tercihlere ait nedenlerin bir özeti"* arayışındadır. Birim bazında incelenen, örneğin bir bireyin, karar alma sürecinde etkili birçok tercihi bulunmaktadır. İktisat, akılcı bireylerin, bu tercihler arasından faydalarını maksimize edecek yönde rasyonel kararlar alacağını varsaymaktadır. Bu teoriden hareketle NTM ile karar vericinin alternatifler arasından en olası seçeneğin tercihinde bulunması için bir karar kuralının elde edilmesinin amaçladığı söylenebilir (Ben-Akiva ve Lerman,1985). Karar vericide değişen durumların (açıklayıcı değişkenlerin aldığı değerlere koşullu bağımlı değişkenin atandığı kategorilerin) yarattığı etkilerin belirlenmesi için model parametreleri, hipotezlerle test edilmekte, tahminlerde örnekleme hatasından kaynaklı oluşabilecek sapmalara karşılık güven aralıkları hesaplanarak raporlanmakta ve modele ait katsayılar yorumlanmaktadır. Model parametrelerinin açıklayıcı değişkenler ile bağımlı değişkenin kategorileri arasındaki ilişkilerin iyi bir tahmini olması içinse verilere ait dağılımın bir eğriye uydurulmasında araç olacak tahmincinin tutarlı ve etkin olması istenmektedir.

Denetimli Makine Öğrenmesi ve Sınıflandırma Algoritmaları

"Alınan kararların arkasındaki nedenlerin açıklanamadığı veri odaklı tahminler dünyasına geçiş" ***Anonim***

Makine öğrenmesinin bir alt dalı olan denetimli öğrenme, bağımlı değişkenin aldığı değerlerin önceden bilindiği durumda kullanılan öğrenme tekniğidir. Burada öğrenilmiş soru-cevaplardan öğrenilmemiş soruların cevaplanabildiği bir tür genelleme kabiliyetinin öğrenilecek fonksiyona kazandırılması amaçlanmaktadır (Sugiyama, 2016). Makine öğrenmesinde fonksiyonel ilişkiler için çoğunlukta algoritmalardan destek alınmaktadır. Makinelere genelleme kabiliyeti kazandırmak için en güvenilir öğrenme tekniği makinenin, giriş çıkışları bilinen bir sistemin altında yatan mekanizmayı keşfettiği ve sistemin kendisini bildikleriyle test edebildiği bir ortamda gerçekleşmektedir. Bunun için başlangıçta uygulamaya konu olan $n + m = N$ büyüklüğündeki veri seti n gözlemli $\mathrm{E} = \{(x_i, y_i) | \, x_i \in \mathbb{R}^p, y_i \in \{1, \dots, k\}\}_{i=1}^{n}$ eğitim seti ile m gözlemli $\mathrm{T} = \{(x_i, y_i) | \, x_i \in \mathbb{R}^p, y_i \in \{1, \dots, k\}\}_{i=1}^{m}$ test seti olmak üzere ikiye bölünmektedir. Böylece eğitim setinden öğrenilen model sistemin bilinen bir parçası olan test setinde kendi performansını araştırabilmektedir (Hastie vd., 2017). Temel amacı doğru kestirimler yapmak olan makine öğrenmesinde veriden öğrenilen $\hat{f}(\boldsymbol{x})$, kestirim fonksiyonunun açıklayıcı ve bağımlı değişkenlere ait bileşik olasılık dağılımının en iyi temsili olması istenmektedir[2]. Öğrenme algoritmasının doğru kestirimler üretebilmesi içinse eğitilen algoritmanın, kayıp fonksiyonu ile hesaplanan hatalarının minimize edilmesi gerekmektedir (Vapnik, 1998). Bunun için örnekten hesaplanan $\hat{f}(\boldsymbol{x})$'e ait kestirim hatalarının ($f(\boldsymbol{x}) -$

[2] Burada kullanılan "öğrenme" terimi ile olasılık fonksiyonları arasından en doğru kestirimleri verecek fonksiyonun seçim süreci kastedilmektedir.

$\hat{f}(\boldsymbol{x}) \neq 0$) araştırılması gerekmektedir. Eğitim ve test setinden ayrı ayrı hesaplanan kestirim hataları sırasıyla eğitim ve test hatası olarak adlandırılmaktadır. Bu hataların araştırılabilmesi içinse eğitim ve test setinin $i.i.d$ dağıldığı varsayılmaktadır (Mello ve Ponti, 2018). Verilerin bağımsız ve özdeş dağılması ($i.i.d$) aynı zamanda eğitim ve test setine ait veri yaratma sürecinin tek bir örneğin olasılık dağılımı ile araştırılabilmesine olanak sağlamaktadır (Goodfellow vd., 2016).

Denetimli öğrenme, bağımlı değişkenin bilinen değerlerinin yapısına bağlı olarak regresyon ve sınıflandırma olmak üzere iki tip araç kullanmaktadır. ***Regresyon***, bağımlı değişkenin sürekli değerler aldığı durumda kullanılan veri odaklı kestirim modelidir. NTM ile benzerlik taşıyan ***sınıflandırma*** ise bağımlı değişkenin nitel özellik gösterdiği durumda kullanılan denetimli öğrenme türüdür. Sınıflandırma, NTM aksine herhangi bir iktisadi teoriyi dayanak görmediğinden ateorik bir yapıya sahiptir. Ancak kaos içerisinde bir düzen yaratabilmek adına sınıflandırma teorisi çerçevesinde hareket etmektedir. Sınıflandırma teorisi, birbirlerinden farklı özellikler taşıyarak heterojen yapıda yer alan birimlerin düzenlenerek homojen alt gruplar oluşturabileceği varsayımına dayanmaktadır (Baldwin, 1985). Bu teori, nesnelere ait benzerliklerin (veya farklılıkların) belirlenmiş bir ölçü kriteri ile düzenlenmesini ilke aldığından yapı içerisinde keşfedilmemiş örüntüleri ortaya çıkartabilmek için önemli bir araç olarak görülmektedir (Abrera, 1974).

Çoğunlukta benzerlik fonksiyonunun tahmin edilmesine odaklansa da NTM'nden farklı olarak sınıflandırma algoritmaları öncelikle eğitim setinden kategorileri birbirinden ayıran karar sınırlarını öğrenmektedir. Böylece her sınıfın kendisini temsilen bir bölgesi olabilecek ve yeni gözlemler bulundukları bölgenin kategorisine atanarak sınıflandırılabilecektir. Bu nedenle sınıflandırma için önce kategorileri oluşturan yapıları öğrenen bir sınıflandırma fonksiyonu sonra bu fonksiyon ile yeni örnekleri nitel kategorilerine atayan bir karar kuralının tahmin edilmesi gerekmektedir. Sınıflandırma fonksiyonu $f(\boldsymbol{x})$ ile gösterilmekte ve karar kuralının hesaplanmasında kullanılmaktadır (Webb, 2002). Uygulamada $f(\boldsymbol{x})$ bilinmediğinden eğitim setinden öğrenilmesi gerekmektedir. Bu fonksiyonun doğru sınıflandırma yapabilmesi için sınıflandırma hatalarının mümkün olduğunca küçük olması veya karar sınırlarını ihlal eden gözlem sayısının mümkün olduğunca düşük olacak şekilde eğitilmesi gerekmektedir. Küçük bir sınıflandırma hatasına sahip tahmincinin seçilebilmesi için ise sadece eğitim setini değil aynı zamanda test setini de minimum hata ile sınıflandırması gerekmektedir. Sadece eğitim hatası düşük olan bir modelin test setinde iyi performans göstereceği beklentisi fazla iyimser bir bakışdır. Bunun nedeni ise algoritmaların sabit olmayan parametre sayısı ile eğitim setinden öğrenebilme özelliğine sahip olmasıdır. Başka bir ifade ile algoritmalar karar sınırlarını doğru öğrenebilmek için eğitim setinden kategorileri tanımlamaya yardımcı olacak yeni tip değişkenler üreterek örnek uzayını genişletebilmektedir (Belloni vd., 2014). Algoritmaların bu özelliği, onların daha esnek fonksiyonel yapıları öğrenmesini sağlasa da tahmin edilecek parametre sayısında meydana gelen artış, algoritmanın eğitildiği verilere mükemmel uyum göstermesine neden olmaktadır. Bu, her ne kadar, olumlu bir sonuç gibi görünse de eğitim setini ezberlemiş bir algoritmanın, test setinde yer alan farklı kalıpları tanıyamamasına neden olmakta ve modelin genelleştirme amacıyla çelişmektedir. *Aşırı uyum* olarak adlandırılan bu durum eğitim seti ile $i.i.d.$ dağılan test setinin sanki eğitim setinden farklı bir dağılıma sahiplermiş gibi hatalı sınıflandırılmasına yol açmakta ve kendini düşük eğitim hatasına karşılık yüksek test hatası durumunda ele vermektedir. Test hatası, tahmincinin aynı zamanda eğitimde kullanılmayan verilere olan uyumunu temsil ettiğinden yeni gözlemlerin de benzer bir hata ile tahmin edileceği düşünülmektedir. Bu nedenle sınıflandırma algoritmasının performansı genelleme hakkında daha gerçekçi bir bilgi sunan test hatasına bakılarak değerlendirilmektedir. Aşırı uyum problemi ile başa çıkmak içinse öğrenme modelinin veri setinden öğrenebileceği parametre sayısının veya başka bir ifade ile model kapasitesinin kontrolü gerekmektedir. Yüksek parametre sayısına sahip tahminciler, veri içerisinde yer alan bütün ile uyumlu ve anlamlı bilgiler dışında veride istenmeyen ve bütünü temsil etmeyen tesadüfi yapıları da öğrenmek konusunda oldukça başarılı olduğundan modelin gürültüyü öğrenmek yerine *"tanıyacak kadar"* parametre sayısına sahip olması istenmektedir[3]. Modelin parametre sayısı model karmaşıklığı ile ilişkili olup uygun karmaşa seviyesi için varyans – sapma dengesi ile optimum kapasite kullanımı gerekmektedir. Varyans-sapma dengesi tüm kestirim modellerinin aranan bir özelliğidir ve denetimli öğrenmenin temel problemidir. Bu nedenle öğrenme algoritmaları mümkün olan en iyi modelin elde edilmesi için varyans ve sapma dengesini ayarlamayı amaçlamaktadır (Çağlayan Akay, 2020). Varyans-sapma dengesi için kullanılabilecek ayarlama teknikleri model parametrelerinin ayarlanması dışında öğrenme algoritmasının sınıflandırma başarısının arttırılması için de kullanılabilmektedir (Goodfellow vd., 2016).

Sınıflandırmada kullanılabilecek teknikler, bağımlı değişkenin nitel kategorilerinin sayısına göre $k = 2$ için ***"ikili sınıflandırma"*** $k > 2$ içinse ***"çoklu sınıflandırma"*** olarak iki grupta incelenmektedir. Her bir kategoriyi temsilen, örnek uzayından rastgele bir nokta seçilirse k sayıdaki noktadan $k - 1$ sayıda doğru geçeceği varsayımı ile k kategorinin varlığında hesaplanması gerekecek karar sınırı veya sınıflandırma fonksiyonlarının sayısı da

[3] Burada gürültü verinin bütünü ile ilişkisiz bilgileri ifade etmek üzere kullanılmıştır.

kategori sayısının bir eksiği olacaktır. Böylece ikili sınıflandırma durumunda örnek uzayı tek bir karar sınırı ile bölünürken çoklu sınıflandırmada ise birden fazla karar sınırı ile bölündüğü görülecektir. Bu karar sınırları öğrenme algoritmasının seçimine göre doğrusal veya eğrisel özellik göstermektedir. Bunlar arasından hangi karar sınırına sahip tahmincinin daha iyi olduğu sorusuna verilebilecek evrensel bir cevap bulunmamaktadır. Yapılacak seçim verinin hacmi, boyutu, kalitesi, değişkenlerin yapısı, kategori sayısı ve araştırmanın konusu gibi birçok özellikten etkilendiğinden uygun öğrenme algoritmasının seçimi zor bir süreçtir. Bu nedenle makine öğrenmesinde tek bir sınıflandırma tahmincisi yerine birden fazla sınıflandırma tahmincisinin eğitilmesi, parametrelerinin optimize edilmesi ve test seti ile karşılaştırılarak nihai seçime ulaşılması gerekmektedir.

Verilerin bir kara kutu tarafından tanımlandığı algoritmik yaklaşımda araştırma problemini bileşenlerine ayırmak ve hangi sınıflandırma algoritmasının seçileceğini belirlemek için öncelikle algoritmaların varsayımları, güçlü ve zayıf yönlerinin bilinmesi ve her sınıflandırma algoritmasının tüm araştırma problemleri için başarılı olamayacağının hatırlanması gerekmektedir. Sınıflandırma algoritmaları veriden öğrenme yaklaşımlarına göre farklı kategorilere ayrılmaktadır. Genel olarak, tahmin edilecek parametre sayısının sabit olup olmamasına göre *parametrik* ve *nonparametrik sınıflandırma algoritmaları*, karar sınırlarının eğriselliğine göre *doğrusal* ve *doğrusal olmayan sınıflandırma algoritmaları*, karar sınırlarına ait olasılık fonksiyonun ve model parametrelerinin öğrenme ve ayarlanma şekline göre *üretken (generative)* ve *ayrıştırıcı sınıflandırma algoritmaları* ve son olarak karar sınırlarının doğrudan ve dolaylı tahmini durumunda sırası ile *geometrik yaklaşım* ve *olasılık yoğunluk fonksiyonunun tahminini* şeklinde gruplandırılmaktadır. Algoritmaların tamamının veriden esnek fonksiyonel ilişkilerle öğrendiğini söylemek doğru değildir. Kimi sınıflandırma tahmincileri daha geleneksel bir yapı sergilemekte ve sadece ilgili yöntem varsayımları sağlandığı takdirde istatistiksel açıdan güvenilir sınıflandırma sonuçları üretmektedir. Ancak günlük hayat problemlerinde veri seti çoğunlukla bu yöntemlerin gerektirdiği varsayımlara uymayabileceğinden geleneksel yöntemlerin aksine, daha az varsayım gerektiren, daha esnek ve dayanıklı tahminciler üreten sınıflandırma algoritmalarına başvurulmaktadır (Aşkın, 2019).

Nitel Tercih Modelleri ve Sınıflandırma Algoritmaları Arasındaki Temel Farklılıklar

"Bir organizma ne kadar çok eldeki koşullara uyarlanırsa, yeni değişikliklere o kadar az uyum sağlayacaktır."
Ronald Fisher

NTM ve sınıflandırma algoritmaları amaçları, odak noktaları ve yöntemleri itibari ile farklı disiplinlerin teknikleridir. Sınıflandırma algoritmaları genel olarak nedensel ilişkilere önem vermeyen ve kara kutuya benzetilen sistemin bir parçası olarak görülmektedir. Söz konusu kategorilerin kaotik yapısına bir düzen vermek olduğunda modellerin kara kutudan sınıflandırma algoritmaları ile tahmini yeterli görülmektedir. Yüksek hesaplanma gerektiren teknikleri ekonometri kadar istatistiksel olmayıp kestirim odaklıdır. Ekonometri ise kestirimden çok hipotezler ile araştırma konusuna açıklık getirmeye çalışmaktadır. Nedensel çıkarımlara istatistiksel teknikler ve iktisat teorisi aracılığı ile ulaşılmaktadır.

Sınıflandırmada veriye uygun şekilde karar sınırlarının tahmini amaçlandığından kestirim performansının yüksek olması için parametreler, veriye uygun şekilde optimize edilmektedir. Veriden öğrenilebilecek bilginin miktarı, modelde yer alan parametre sayısı ile ilişkili olduğundan algoritmaya esneklik kazandırmak için öğrenilecek parametre sayısının sabit tutulması doğru bir yaklaşım olarak görülmemektedir (Goodfellow vd., 2016). Bu noktada sınıflandırmanın, sabit parametre sayısı ile analizler gerçekleştiren parametrik ekonometriden farklılaştığı ve nonparametrik ekonometri ile benzerlik taşıdığı söylenebilir. Sınıflandırma algoritmalarının $f(\boldsymbol{x})$'e ait herhangi bir fonksiyonel form varsayımında bulunmaması ise nonparametrik ekonometri ile arasındaki diğer bir benzerliktir (Athey ve Imbens, 2019).

Sınıflandırma, sıralı olmayan kategorilerin analizinde kullanıldığından sıralı olmayan NTM ile benzerlik göstermektedir. Ancak örnek uzayını en doğru şekilde karar bölgelerine ayıracak karar sınırlarının öğrenilmesi amaçlandığından açıklayıcı değişkenlere koşullu bağımlı değişken kategorilerinin tercih edilme olasılıklarının hesaplandığı NTM'den farklı bir yapıya sahiptir. Koşullu olasılık dağılımının tutarlı tahmini yerine test hatasının minimizasyonu ön plandadır. Bu nedenle örnek uzayını minimum sınıflandırma hatası ile karar bölgelerine ayırmak üzere hareket eden algoritmanın $P(Y = k|X = x)$'in tahmini sonrasında elde ettiği sapmalar, sınıflandırma sonucunu etkilemediği müddetçe göz ardı edilmektedir. İki teknik arasındaki diğer bir farklılık sapma, varyans ve ortalama hata kare kavramlarına olan yaklaşımlarından ileri gelmektedir. Sınıflandırmada algoritmaların sapmasız kestirimlerde bulunması istense de bu durum yüksek varyans nedeni ile genele uyarlanamayan (veya test hatası yüksek) bir tahminciye neden olacağı için bir miktar sapmaya izin verilmektedir. Bu nedenle sınıflandırma algoritmaları aynı zamanda sapmalı tahmin yöntemleri olarak adlandırılmaktadır (Mitchell, 1980). NTM ise katsayılara ait sapmasız/tutarlı tahmincilerin elde edilmesi arayışındadır (Storm vd., 2020). Modele ait parametrelerin istatistiksel anlamlılıklarının araştırılması ve standart hatalar ile güven aralıklarının hesaplanarak örnekten kaynaklı belirsizliklerin kontrol edilebilmesi için tahminciler ile testlerin tutarlı olmaları, normal dağılmaları ve etkinlik özelliğine sahip olmaları istenmektedir. Ancak sınıflandırma, bir

tahmin problemi olmadığından bu beklentiler ne yazık ki makine öğrenmesi teknikleri tarafından henüz karşılanamamaktadır. Optimize edilen parametreler istatistiksel çıkarımlar için tutarlı tahminler veremese de kestirim tabanlı algoritmalar, sınıflandırma hatalarının alt-üst limitlerini hesaplayarak hatalı sınıflandırma oranını kontrol altında tutulabilmektedir (Mullainathan ve Spiess, 2017).

Genel olarak ekonometrik modeller, parametrelerini tüm veri setinden veya sınıflandırma ile benzerlik taşıması için sadece eğitim setinden tahmin etmektedir. Makine öğrenmesinde ise parametrelerin eğitim setinden tahmini yerine öğrenimi söz konusudur. Öğrenilen parametreler, test setini minimum hata ile sınıflandırmak için geçerlilik seti ile optimize edilirken iktisadi teorinin egemen olduğu ekonometride ise model geçerliliğine çoğunlukta yer verilmemektedir. Bunun nedeni ise Varian (2014)'a göre, sosyal bilimlere konu olan gözlem sayısının veri setini eğitim – test seti olarak bölmeye yeterli olmaması olarak gösterilmiştir. Bunun doğal bir sonucu ise NTM'inin örnek-içi kestirimlerinin aşırı uyumla sonuçlanmasıdır. Ancak NTM'nde amaç kestirim olmadığından bu durum dikkate alınmamaktadır (Varian, 2014). Bu nedenle model geçerliliği, model varsayımlarının araştırılmasına dayanmaktadır. Çoğunlukta anakütlenin bilinmeyen parametrelerinin örnekten doğru tahmini için tahmincinin asimtotik özelliklere sahip olması istenmektedir.

NTM'nde bilinmeyen ancak var olduğu kabul edilen doğru modelin seçimi genelde birbirine benzer modeller arasından hipotez testleri ile yapılmaktadır. Sınıflandırmada ise örnek uzayını karar bölgelerine ayıran sınıflandırma fonksiyonunun test setindeki (örnek-dışı) kestirim başarısı ön plandadır. NTM'nde model seçimi için çoğunlukta uyum iyiliği ölçüleri kullanılırken sınıflandırmada çapraz geçerlilik tekniğinden yararlanılmaktadır. Algoritmalar belirsizlikleri kontrol altında tutmak için parametrelerini veri odaklı bir yaklaşım ile optimize etmekte ve tek bir model tahmini yerine veriden farklı fonksiyonel formların öğrenilmesi ile test hatasını minimize eden tahminciden yana seçimde bulunmaktadır. NTM'nde ise tahminciler bağımlı değişkendeki değişimin açıklanması ile ilgili ölçüler ürettiğinden amaç, özel olarak, en iyi örnek-dışı öngörüler verecek olan modelin seçimi olmadığı müddetçe çapraz geçerlilik tekniğine başvurulmamaktadır. Ekonometride çoğunlukta doğru modelin bilindiği varsayıldığından sınıflandırmaya kıyasla daha az sayıda model karşılaştırılmaktadır. Denetimli öğrenmede ise sınıflandırma için karar ağaçları, tesadüfi orman ve yapay sinir ağı gibi yüksek işlem gerektiren teknikler dışında lojistik regresyon gibi istatistiksel özelliklere sahip birden fazla tahminci, örnek uzayında örüntü aramak üzere kullanılmaktadır. Sınıflandırma algoritmaları, NTM'nin aksine tek bir model yerine farklı formlarda birden fazla model ile senkron çalışabilmektedir. Amaç sınıflandırma hatasını minimize etmek ise sınıflandırmada, birden fazla modelin ortalaması alınarak tek bir modelle, daha düşük test hatasına sahip olmak mümkündür (Rokach, 2010). Artırma ve torbalama gibi tekniklerin kullanıldığı bu öğrenme tekniğine *"topluluk (ensemble) öğrenmesi*" adı verilmektedir.

Yüksek sınıflandırma performansına odaklanan algoritmalar, test setindeki başarılarına eğitim setinden öğrendikleri doğrusal olmayan ilişkiler ile ulaşmaktadır. İnsan davranışlarının tesadüfiliği dikkate alındığında doğrusal olmayan yapıların öğrenimi, sınıflandırmanın doğruluğu açısından önemli bir özellik olacaktır. Sınıflandırma algoritmaları, esnek fonksiyonları öğrenebilme kapasitesine sahip olduğundan özellikle doğrusal olmayan ilişkilerin saptanmasında NTM'e kıyasla daha başarılı bulunmaktadır (Liu ve Xie, 2019). NTM'nde bu ilişkilerin tahmini için çoğunlukta araştırmacının uzman bilgisine başvurulmakta ve hesaplamalarda kolaylık sağlamak için doğrusal olmayan ilişkiler doğrusallaştırılarak modele eklenmektedir. Bu teknik araştırmacının önsel bilgisi dışında yer alabilecek doğrusal olmayan ilişkilerin tahmine olanak vermediğinden sınırlı bir modelleme yaklaşıma neden olmaktadır. Karar sınırlarının eğriselliği test hatası, aşırı uyum ve varyans – sapma dengesi ile doğrudan ilişkilidir. Dengenin korunması için optimum model karmaşıklığının belirlenmesi gerekmektedir. Minimum test hatasına sahip karar sınırlarını elde etmek için ise ayarlama parametreleri veriden öğrenilmektedir. Kullanılan ayarlama teknikleri, ekonometrinin ÇDB durumunda başvurduğu sapmalı tahmin yöntemleri ile benzerlik taşımaktadır. Ancak ekonometride ayarlama parametreleri, kullanıcı tarafından seçildiğinde bir öğrenme yaklaşımı oluşturmamakta ve sınıflandırmadaki kullanımından farklılaşmaktadır.

NTM, birimlerin tercih olasılıklarını tahmin edebilse de sınıflandırma için uygun değildir. Bunun yerine istatistiksel modelin yapısal parametrelerinin tahmini ve yorumlanmasına odaklanmaktadır. Mullainathan ve Spiess (2017)'e göre amaç sadece nitel kategorilere gözlemler atamak ise algoritmalar özellikle yüksek boyutlu veri setlerinin sınıflandırılmasında NTM'lerine göre daha başarılıdır. Burada parametrelerin tahmini ve testi temel amaç olmadığından yüksek boyutlu veriler arasındaki ÇDB, sınıflandırma başarılı olduğu müddetçe ilişkili değişkenlerden herhangi biri ile çalışılmasına olanak tanıdığından bu durum model kapasitesi anlamında destekleyici bir araç olabilmektedir. Temelde düşük test hatasına sahip model tahmini amaçlandığından ilişkili değişkenlerden hangisinin sınıflandırmada kullanılacağı önem arz etmemektedir. Ancak bu avantaj model bulgularının yorumlanabilir olması istendiğinde kolaylıkla bir dezavantaja dönüşebilmektedir. Öyle ki iktisadi teoriye katkı sağlamayan bir kestirim modelinin birimlerin tercih olasılıklarını açığa kavuşturmaktaki yetersizliği sınıflandırma algoritmalarını sahip olduğu avantajlara rağmen NTM karşısında güçsüz bırakmaktadır. Sınıflandırmanın ötesinde bir analiz amaçlanmış ise algoritmalar, verinin boyutları ne olursa olsun istatistiksel

çıkarımlar ile nedensel ilişkilerin tahmininde yetersiz kalmaktadır. Öte yandan yüksek boyutlu bir veri setinin açıklayıcı değişkenleri arasındaki korelasyonlar, NTM'ne ait parametrelerin güven aralıklarının genişlemesine neden olduğundan ekonometrik analizlerde tehlike yaratmaktadır.

NTM'i açıklayıcı değişkenlerin bilgisine koşullu kategorilerin ortalama seçim olasılıklarını araştırmakta ve çoğunlukta tercihte bulunmaya ilişkin özet bir tablo sunmaktadır. Bu durum karmaşadan uzak ve müşterek ilişkilerin tahminine imkân verse de bütünü görebilmek için (makine öğrenmesinin kolaylıkla saptayabildiği) birimlere has özellikler göz ardı edilebilmekte veya alt örnek gruplarına has olduğu için genele uymayan veriler *"aşırı değer"* olarak görülebilmektedir. Kimi araştırmalarda bu bilgilerin dikkate alınmaması önemli sayılabilecek bir bilgi kaybına neden olmaktadır (Fan vd., 2014). Bu nedenle NTM, doğru kestirimler için bilginin her zerreciğine ihtiyaç duyulan sınıflandırmada başarısız olmaktadır.

Sınıflandırma algoritmalarının veri odaklı model seçimi her ne kadar kestirim performansında önemli bir artışa neden olsa da tutarlı parametre tahmincilerinin hesaplanamamasına, parametrelerin test edilememesine ve nedensel ilişkilerin araştırılamamasına neden olmaktadır (Berk vd., 2010).

Sınıflandırma Algoritmalarının Nitel Tercih Analizi ile Buluşması

"Hiçbir insan bir makineden daha iyi değildir ve hiçbir makine de makinesi olan bir insandan daha iyi değildir." ***Paul Tudor Jones***

Ekonometrinin geleneksel NTM çalışmalarına bakıldığında, istatistiksel çıkarımlar ile nedensel ilişkilerin araştırılması için yetersiz kalan sınıflandırma algoritmalarına yer verilmediği veya tercih edilmedikleri görülmektedir. Ancak güncel bir literatür taraması yapıldığında bilgi teknolojilerinde meydana gelen gelişmelerin, veri üretimindeki artışların ve veriye daha kolay erişebilir olmanın sağladığı avantajların, nitel tercih analizi (NTA)'ne olan yaklaşımları da etkilediği görülecektir. Literatürdeki bu çeşitlenme sadece ekonometriden makine öğrenmesine değil aynı zamanda makinelerin de nedensellik araştırmalarına dahil olmaları için yapılan araştırmalardan oluşmaktadır. İki disiplin arasındaki bu alışveriş henüz çok yeni olsa da NTA'ne karşı daha esnek olabileceğimizi göstermektedir (Kleinberg vd., 2015). Bu entegrasyon, sınıflandırmada etkili olduğu saptanan yapısal olmayan verilerin NTM ile analizini içerdiği gibi aynı zamanda makine öğrenmesini ekonometrik bir analizin önsel bir aşaması olarak sunmaktadır (Zheng vd., 2017).

Yapılan çalışmalar çoğunlukta yapısal olmayan verilerin uygun tekniklerle dönüşümüne ve sınıflandırmada etkili bulunan değişkenlerin ekonometrik analizlerdeki kullanıma odaklanmaktadır. Her iki disipline katkı sağlayan çalışmalar arasında cep telefonu verileri ile müşterilerin kredi geri ödemelerini araştıran Björkegren ve Grissen (2017), hangi mahallelerde fiziki gelişmelerin olduğunu uydu ve harita verileri ile araştıran Naik vd. (2017), yüksek hacimli verilerin kullanımı ile bireylerin işsizlik riskini araştıran Gerunov (2016, 2020), kredi derecelendirme (Fu vd., 2016), müşteri sadakati (Deliana ve Rum, 2017), kredi risk yönetimi (Meng vd., 2019) ve firma bazında toplanan insan kaynakları verilerinden oluşturulmuş bir veri seti ile meslek seçimi ve çalışma durumuna ilişkin birimlerin tercihleri üzerine çalışan Ikudo vd. (2018) dışında birçok e-ticaret sitesinin tavsiye sistemleri oluşturmak üzere kullanıldığı doğal dil işleme teknikleri ile birimlerin tercih davranışlarını araştıran Penczynski (2019), hanehalkı anketleri ile bireylerin ulaşım tercihleri için karar ağaçlarına başvuran Brathwaite vd. (2017), 10-K forumları ile finansal piyasa oynaklığını araştıran Kogan vd. (2009), cep telefonu verileri ile kişilerin servet kategorilerinin araştırıldığı Bernheim vd. (2013) ile Türkiye'de henüz işsiz olan bireylerin istihdama geçiş veya işsiz kalmaya devam etme olasılıklarının tahmini için sınıflandırma algoritmalarını araç olarak kullanan Kütük ve Güloğlu (2019) örnek gösterilebilir.

Tercih teorisine ait nedensel çıkarımların amaçlandığı araştırmalarda akla ilk gelen teknik NTM'dir. Geçmiş verilerin gelecek veriler ile tutarlı olması durumunda bu modeller sadece nitel kategorilerin tercih olasılıklarını değil aynı zamanda bu olasılıkların kestirimi konusunda da oldukça başarılıdır. Ancak ne yazık ki, uygulamada yaşanan kimi olumsuzluklar NTM'nin her iki amacı eş zamanlı olarak yerine getirmesine engel olmaktadır. Bu noktada Zheng vd. (2017)'ne göre iki alternatif bulunmaktadır. Seçeneklerin birinde, araştırmacının NTM ile elde edebileceği bilgilerden vazgeçerek kestirim odaklı bir yaklaşım ile sadece nitel kategorilere gözlemler atayabildiği bir sınıflandırma algoritmasının tahmini yer almaktadır. Burada NTA'ne ait her araştırma probleminin istatistiksel çıkarımlara dayanmayabileceği ve kimi zaman doğru cevaba iyi bir kestirim modeli ile ulaşılabileceği hatırlatılarak araştırmacılar, sınıflandırma algoritmalarının kullanımına yönlendirmektedir. Disiplinlerarası çalışmaya teşvik eden ikinci seçenekte ise algoritmaların keşfettiği önemli bilgilere (değişkenlere) NTM'nde yer verilmesi yaklaşımı yer almaktadır. Burada ekonometrik analizlerden vazgeçmenin zorunlu olmadığı ortaya konulmaktadır. NTA'ne farklı bir perspektif kazandıran ikinci yaklaşım, son zamanlarda birçok araştırmacının analizine konu olmuş yaklaşım şeklidir. Böylece farklı tekniklerin kullanımı ile NTM'nin istatistiksel ve/veya nedensel çıkarımlar dışındaki sorulara da cevap verebileceği söylenebilir.

Kimi zaman bir araştırmadan olayların nedenleri yerine neler olduğunun ortaya konulması beklenmektedir. Tıpkı Amazon ve Netflix gibi tavsiye sistemlerinde uzmanlaşmış şirketlerin müşterilerinin neden bir ürünü aldığı veya filmi neden izlediği yerine bir sonraki tavsiyeye kulak asıp asmayacağını en doğru şekilde belirleyerek müşteri sadakatini kazanmayı amaçlaması gibi. Kestirim tabanlı sınıflandırma yöntemleri bu tip araştırmalar için oldukça yararlı tekniklerdir. Öte yandan hedef kitlenin ilgili tercihte bulunma nedenlerinin alınacak kararlarda etkili olduğu düşünülüyorsa nedensel ilişkilerin tahmini önem kazanacağından belirsizliği modellemede teorik tabanlı ekonometrik modeller daha doğru bir seçim olacaktır. Ancak veriler $p > n$ veya $p \gg n$ formunda ise NTM'inin yüksek boyutlu veri analizine ait bazı teknikleri ödünç alması gerekecektir (Belloni vd., 2018). Bunlar arasında alt-seçim yöntemleri, boyut indirgeme teknikleri ve daraltıcı yöntemler bulunmaktadır. Bu teknikler aynı zamanda yorumlanabilir sınıflandırma algoritmaları de için etkili araçlardır (Vellido vd., 2012).

Sınıflandırmada, algoritmaların esneklikleriyle sağladığı katkılar NTM'nin basit ve yorumlanabilir model arayışı ile çelişki oluşturmaktadır. Sınıflandırmada araştırmacı, sınıflandırma başarısında meydana gelen çok ufak bir artışın, hastanın kanser teşhisinde önemli olduğunu veya borsanın artış/azalışına göre yatırım yapacak bir yatırımcı için sınıflandırma hatasında meydana gelecek ufak bir azalışın kişinin yatırım tercihinde etkili olacağını düşünüyorsa artan model kapasitesinin doğurduğu maliyete katlanmak isteyecek ve yorumlanabilir model arayışı yüksek sınıflandırma başarısına sahip modeller tarafından gölgelenecektir. Ancak kimi durumlarda karar vericilerin doğru bir sınıflandırma ile beraber bulguların ve sistemin işleyiş şeması hakkında da bilgi sahibi olmayı istediği görülmektedir (Athey, 2017). Bu tip durumlarda algoritmalar önce yorumlanabilir sonra minimum sınıflandırma hatasına sahip olacak şekilde eğitilmektedir. Doğaları gereği karmaşık modellemelerde bulunan algoritmalar için yorumlanabilirlik ile sınıflandırma doğruluğu arasında korunması gereken bir denge bulunsa da bu durum bütün tahminciler için geçerli değildir. Örneğin; karar ağacının düğümlerinde yer alan açıklayıcı değişkenler sınıfları ayrıştıran özellikler hakkında bilgi verebildiği gibi aynı zamanda birimlerin tercihine ilişkin yorumlanması kolay bir görsel de sunabilmektedir. Bu tip bir analizin bir sonraki aşamasında kolaylıkla ekonometrik bir model tahmini yapılabilir. Mullainathan ve Spiess (2017)'e göre bu çalışmalar iktisat ve/veya tercih teorisine katkı sağlayabileceği gibi teorilerin test edilmesi için de bir ölçüt olarak da kullanılabilir.

Sınıflandırma Algoritmalarında Korelasyon ve Nedensellik

"İnsanlar verilerden daha akıllıdır. Veriler, nedenleri ve sonuçları anlamaz." ***Judea Pearl***

Makine öğrenmesinde kestirim gücü olan tüm fonksiyonlar birer model olarak adlandırılabilir. Modelden beklenen performans, bulgularının statik bir veri raporundan farklı ve daha esnek bir yapıya sahip olmasıdır (Burger, 2018). Makine öğrenmesi bu esnek yapıları veride yer alan korelasyonları dikkate alarak öğrenmektedir. Korelasyon, iki tesadüfi değişken arasındaki istatistiksel ilişkiyi incelemektedir. Güçlü bir korelasyon, veri değerlerinden biri değiştiğinde diğerinin de değişme olasılığının yüksek olacağı anlamına gelmektedir (Mayer-Schönberger ve Cukier, 2014). Tersi durum ise zayıf bir korelasyona işaret etmektedir. Korelasyonun önem kazanma süreci, bir veri setinin kendisine has değişkenliği olan varyans ölçüsü sonrasında gelişmiş olup iki değişkenin birlikte gösterdiği hareketin araştırılması ile gündeme gelmiştir (Fisher, 1934). Temeli Francis Galton ve Karl Pearson'a dayanan korelasyon analizinin ortaya çıkış sürecinde her iki araştırmacının da korelasyon öncesinde nedenselliği açıklamak üzere hareket ettikleri görülmektedir. Galton (1888), birbirleri ile korelasyonlu iki değişkenin *"ortak bir neden"* paylaştığını ifade ederken Pearson (1896) ise gözlemlenen korelasyonların *"birbirinden bağımsız şekilde katkı sağlayan neden"* tarafından oluştuğuna dikkat çekmiştir (Aldrich, 1995). Galton ve Pearson'a ait çalışmalar, araştırmacıların *"korelasyon ve nedensellik arasındaki ilişki"* tartışmasını sürdürmek ve/veya açığa kavuşturmak yerine matematiksel modellerle açıklanması çok daha kolay olan korelasyon analizine yoğunlaştığını göstermektedir. Korelasyon ile nedensellik arasındaki farklılık çoğunlukta görmezden gelindiğinden ayırt edilmesi zor kavramlara dönüşmüştür.

Kavramlar arasındaki ayrışma büyük verinin yer edinmeye başladığı çalışmalarda kendini daha da çok göstermektedir (Yalçıntaş, 2018). Kimi araştırmacılar 21.yy'ın cevheri olarak adlandırılan büyük verinin *"Neden?"*, *"... olsaydı ne olurdu?"* veya *"... yapılsaydı etkileri ne olurdu?"* gibi soruların sorulmasına ihtiyaç bıraktırmayacağını ve *"Ne"* sorusuna cevap arayan korelasyonun büyük veri kullanımında yeterli olacağını savunmaktadır. Aralarında özellikle bu gibi ifadeleri ile kendinden söz ettiren Chris Anderson, Wired Dergisine 2008 yılında verdiği röportajda büyük verinin sadece nedenselliğin değil aynı zamanda teorinin de bir sonu olduğunu iddia etmiştir. Anderson'un bu iddiası nedensellik ile korelasyonun farklı araştırma sorularına cevap arayan ve bu nedenle birbirlerinin ikamesi olmayan teknikler oldukları gerekçesi ile sıklıkla eleştirilmiştir (Cowls ve Schroeder, 2015).

Veriden öğrendiği bilgiler ile sınıflandırmada bulunan makine öğrenmesi teknikleri, açıkça programlanmadıkları müddetçe öğrenme eylemini eğitim setindeki korelasyonları dikkate alarak gerçekleştirmektedir. Bu nedenle makinenin sınıflandırmada gösterdiği başarı bu korelasyonları ne kadar iyi öğrendiği ile ilişkilidir. Calude ve Longo (2017)'ya göre algoritmalar, sayısal olarak birbirlerine benzeyen değişkenler arasındaki tesadüfi ilişkileri

tespit etmek üzere görevlendirilmektedir. Algoritmaların özellikle yüksek boyutluluk durumunda eğitim setinden öğrendiği ve gerçekte anlamlı bir ilişkiye işaret etmeyen korelasyonlar *"sahte korelasyon"* olarak adlandırılmaktadır. Büyük veri analizinde sıkça karşılaşılan sahte korelasyon, sınıflandırma tahmincisinin test setindeki performansını olumsuz etkilediğinden uygun tekniklerin kullanımı ile sorunun giderilmeye çalışılması önerilmektedir (Fan, 2014). Büyük verilerin neden olduğu korelasyonla ilişkili diğer bir sorun ise *"tesadüfi içsellik"* problemidir. Tesadüfi içsellik, gerçekte korelasyonlu olmayan açıklayıcı değişkenler ile sınıflandırma hatalarının varlığında ortaya çıkmaktadır (Fan vd., 2014). Bu durum model seçiminde tutarsızlığa ve istatistiksel sapmalara neden olduğundan hatalı bilimsel sonuçlarla karşılaşmamak için uygun tekniklerin kullanılması gerekmektedir (Fan ve Liao, 2014). Nedenselliğin korelasyonla araştırılmasının doğru bulunmaması bir yana büyük verilerin beraberinde getirdiği kimi sorunlar da araştırma bulguları üzerinde önemli yaptırımlarda bulunabilmektedir. Ancak Pandey vd. (2015)'a göre büyük verinin bu özelliklerini dikkate alan tekniklerin kullanılması durumunda büyük verilerin daha öncesinde keşfedilmemiş bulguları ortaya çıkartması ve küçük veri setleri ile çalışmalar yürüten disiplinlere önemli katkılar sağlaması mümkündür.

Artık günümüzde nedenselliğin istatistiksel analizlerde araştırılmasına olanak tanıyan birçok teknik bulunmaktadır. Cowls ve Schroeder (2015)'e göre bu teknikler sadece deneysel verilerle değil aynı zamanda sosyal bilimlere ait gözlemlenebilir verilerle de çalışılmasına imkân tanımaktadır. Ekonometride istatistiksel çıkarımlar kadar nedensel çıkarımlar da önemli bir yere sahip olup ceteris paribus koşulu altında gözlemlenebilir verilerle araştırılan bu ilişkilerin tahmini kestirim modellerinin tahmininden farklılık göstermektedir. Değişkenlerin birlikte gösterdiği ilişkilere dayanan sınıflandırma algoritmalarının nedensel çıkarımlarda kullanılabilmesi içinse eğitim setinden öğrendikleri korelasyonları ekonometri gibi önceden bilinen bir teoriye dayandırması gerekmektedir. Ancak iktisadi bir teoriye dayanmayan algoritmaların saptadığı korelasyonlar olayların nedenine ilişkin bir bilgi vermese de nedenselliği nerede arayabileceğimiz konusunda bir ipucu sunabilmektedir.

Özellikle son on yılda sınıflandırma algoritmalarının nedensellik arayışına dahil olması için makine öğrenmesi tekniklerinde yapılan güncellemeler ekonometri ile makine öğrenmesi arasındaki güçlü birlikteliğe işaret etmektedir (Athey, 2019). Sınıflandırma alanında yapılan bu çalışmalar genel olarak iki başlık altında incelenmektedir. Bunlardan ilki, geleneksel NTM ile yüksek boyutlu verilerden nedensel çıkarımların amaçlandığı çalışmalardır. Bu tekniklerde çoğunlukta yüksek boyutluluğun ekonometrik modellerde yarattığı sorunların giderilmesi için çift seçim süreci kullanılmaktadır (Belloni ve Chernozhukov, 2013; Belloni vd., 2013). İki aşamalı en küçük kareler tekniği ile benzerlik gösteren bu çalışmalarda otomatik model seçiminde bulunan ve Tibshirani (1996) tarafından ortaya atılan LASSO tipi modellerden yararlanılmaktadır. NTM arasından logit ve multinomial logit modellerinin konu olduğu bu çalışmalara Belloni vd. (2014), Chernozhukov vd. (2015) ve Belloni vd. (2018) örnek gösterilebilir. Bir diğer yaklaşımda ise nedenselliğin sınıflandırma algoritmaları ile araştırılmasına dayanan çalışmalar bulunmaktadır. Bunlar arasında asimtotik varyans ve tutarlı tahmincilerin hesaplanabilmesi için tesadüfi orman tahmincisinden geliştirilmiş *"nedensel orman"* tahmincisi (Wager ve Athey, 2017), araç değişkenin nitel özellik göstermesi durumunda nedensel orman tahmincisinden geliştirilmiş *"genelleştirilmiş tesadüfi orman tahmincisi"* (Athey vd., 2018), nedensel çıkarımların derin öğrenme ile araştırıldığı Ramachandra (2018)'in çalışması ve ileri beslemeli yapay sinir ağının derin öğrenme çerçevesinde nedensel çıkarımlar için geliştirildiği Farrel vd. (2019)'nin çalışması örnek gösterilebilir.

Görüldüğü üzere daha önceleri sadece başarılı kestirimler üretmesi için eğitilen algoritmalar bugün ekonometrinin amaçladığı nedensel çıkarımlar için adapte edilmeye çalışılmaktadır (Athey ve Imbens, 2017). İki disiplin arasındaki yakın ilişkileri birçok çalışmasında vurgulayan ve birleşmenin kaçınılmaz olduğunu düşünen Susan Athey, çalışmalarında makine öğrenmesi algoritmalarının kestirimdeki başarılarını istatistiksel ve nedensel çıkarımlarda da gösterebilmesi için makine öğrenmesi ile ekonometrinin birlikte hareket etmesi gerektiğini vurgulamaktadır. Bu alanda yapılan yeni çalışmalarsa her iki disiplinin de ortak bir paydada buluşabileceğini ve gelecekte yeni bir ufuk yaratabileceklerinin göstergesi sayılabilir.

Sonuç

Birimlerin tercihlerinde meydana gelen değişmeleri istatistiksel bir model yardımı ile inceleyen ekonometride istatistiksel çıkarımlar ile nedensellik araştırmaları önemli bir role sahiptir. Amaç kestirim ise NTA'nde alternatif olarak kullanılabilecek yöntemler arasında makine öğrenmesi sınıflandırma algoritmaları yer almaktadır. Bu algoritmalar veri odaklı yaklaşımları ile en az hata ile kestirimde bulunacak tahminci arayışında olduğundan bu tekniklerin istatistiksel çıkarımlara adapte edilmeden kullanımı ekonometrinin amaçları ile tezatlık oluşturacağından her iki disipline ait sınırların iyi bilinmesi gerekmektedir. Sınıflandırma algoritmalarını istatistiksel ve nedensel çıkarımlarda kullanmak üzere geliştirilen yeni teknikler iki disiplini bir arada kullanmanın mümkün olduğunu göstermektedir. Algoritmalar, birimlerin neyi tercih ettikleri dışında neden tercih ettiklerine yönelik sorulara da cevap aramaya başlamış bulunmaktadır. Büyük veri çalışmalarına bakıldığında iki disiplin arasındaki ilişkinin tek yönlü olmadığı açıkça görülmektedir. Bu yeni ve heyecan verici alan iki disiplinin çift yönlü çalışıp karşı karşıya durmak yerine yan yana olabileceğinin bir göstergesi sayılabilir.

Kaynaklar

Abrera, J. B. (1974). Traditional Classification: Characteristics, Uses and Problems. In Painter, A. F. (Ed.), *Classification: Theory and Practice* (pp. 21-36). Philadelphia: Drexel University Press.

Aldirch, J. (1995). Correlations Genuine and Spurious in Pearson and Yule. *Statistical Science,* 10(4), 364-376.

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. Retrieved Mayıs 23, 2020, from https://www.wired.com/2008/06/pb-theory/

Aşkın, E. Ö. (2019). Karar Ağaçları. In Alp, S., & Öz, E. (Eds.) *Makine Öğrenmesinde Sınıflandırma Yöntemleri ve R Uygulamaları* (pp. 1-35). Ankara: Nobel Akademik Yayıncılık.

Athey, S. (2017). Beyond Prediction: Using Big Data for Policy Problems. *Science,* 355 (6324), 483-485. doi:10.1126/science.aal4321

Athey, S. (2019). The Impact of Machine Learning on Economics. In Agrawal, A., Gans, J., & Goldfarb, A. (Eds.). *The Economics of Artificial Intelligence: A Review* (pp.507-547). Chicago: University of Chicago Press.

Athey, S., & Imbens, G. W. (2017). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives,* 31(2), 3-32. doi: 10.1257/jep.31.2.3.

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods Economists Should Know About. *Annual Review of Economics,* 11, 685-725. doi:10.1146/annurev-economics-080217-053433

Athey, S., Tibshirani, R., & Wager, S. (2018). Generalized Random Forests. arXiv.org>stat>arXiv: 1610.01271v4. https://arxiv.org/abs/1610.01271.

Baldwin, J. T. (1985). Classification Theory: 1985. In Baldwin, J. T. (Ed.) *Classification Theory*. Berlin: Springer.

Ben-Akiva, M. E., & Lerman, S. R. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand (1. Basım). Londra: MIT Press.

Belloni, A., & Chernozhukov, V. (2013). Least Squares after Model Selection in High Dimensional Sparse Models. *Bernoulli,* 19(2), 521-547. doi:10.3150/11-BEJ410.

Belloni, A., Chernozhukov, V., & Wei, Y. (2013). Honest Confidence Regions for Logistic Regression with a Large Number of Controls. arXix.org>stat>arXiv: 1304.3969v1. https://arxiv.org/abs/1304.3969v1.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2), 29-50. doi:10.1257/jep.28.2.29

Belloni, A., Chernozhukov, V., Fernandez-Val, I., & Hansen, C. (2018). Program Evaluation and Causal Inference with High Dimensional Data. arXiv.org>math>arXiv:1311.2645. https://arxiv.org/abs/1311.2645.

Berk, R., Brown, L., & Zhao, L. (2010). Statistical Inference After Model Selection. *Journal of Quantitative Criminology*, 26(2), 217-236. doi: 10.1007/s10940-009-9077-7

Bernheim, D., Björkegren, D., Naecker, J., & Rangel, A. (2013). Non-Choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions (NBER Working Paper 19269). Cambridge, MA: National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w19269

Björkegren, D., & Grissen, D. (2017). Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment. arXiv.org>cs>arXiv:1712.05840v1. https://128.84.21.199/abs/1712.05840v1

Brathwaite, T., Vij, A., & Walker, J. L. (2017). Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice. arXiv.org>stat>arXiv:1711.04826. https://arxiv.org/abs/1711.04826

Brieman, L. (2001). Statistical Modelling: The Two Cultures. *Statistical Science*, 16(3), 199-215.

Burger, S.V. (2018). *Introduction to Machine Learning with R: Rigorous Mathematical Analysis.* Beijing: O'Reilly.

Calude, S. C., & Longo, G. (2017). The Deluge of Spurious Correlations in Big Data. *Foundations of Science,* 22(3), 595-612. doi:10.1007/s10699-016-9489-4

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications* (1st ed.). New York: Cambridge University Press.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Valid Post-Selection and Post-Generalization Inference: An Elementary, General Approach. arXiv.org>math>arXiv:1501.03430. https://arxiv.org/abs/1501.03430.

Cowls, J., & Schroeder, R. (2015). Causation, Correlation, and Big Data in Social Science Research. *Policy & Internet,* 7(4), 447-472.

Cox, D. R. (2001). [Statistical Modeling: The Two Cultures]: Comment. *Statistical Science*, 16(3), 216-218.

Çağlayan, E. (2012). *Nonparametrik Regresyon Modelleri* (1st ed.). İstanbul: Derin Yayınları.

Çağlayan Akay, E. (2020). *Ekonometride Büyük Veri ve Makine Öğrenmesi: Temel Kavramlar* (1st ed.). İstanbul: Der Yayınları.

Deliana, Y., & Rum, I. A. (2017). Understanding Customer Loyalty Using Neural Network. *Polish Journal of Management Studies,* 16(2), 51-61. doi: 10.17512/pjms.2017.16.2.05

Einav, L., & Levin, J. D. (2013). The Data Evolution and Economic Analysis (NBER Working Paper 19035). Cambridge, MA: National Bureau of Economic Research. Retrieved from https://www.nber.org/papers/w19035.

Fan, J. (2014). Features of Big Data and Sparsest Solution in High Confidence Set. In Lin, X. (Ed.) *Past, Present and Future of Statistical Science* (pp.507-521). Boca Raton: CRC Press.

Fan, J., & Fan, Y. (2008). High-Dimensional Classification Using Features Annealed Independence Rules. *The Annals of Statistics,* 36(6), 2605-2637. doi:10.1214/07-AOS504

Fan, J., & Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society,* 70(5), 849-911. doi: 10.1111/j.1467-9868.2008.00674.x

Fan, J., & Liao, Y. (2014). Endogeneity in High Dimensions. *The Annals of Statistics*, 42(3), 872-917. doi:10.1214/13-AOS1202

Fan, J., Guo, S., & Hao, N. (2011). Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression. *Journal of the Royal Statistical Society,* 74(1), 37-65. doi:10.1111/j.1467-9868.2011.01005.x

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data Analysis. *National Science Review*, 1, 293-314. doi: 10.1093/nsr/nwt032

Fan, J., Runze, L., Zhang, C.H., & Zou, H. (2020). *Statistical Foundations of Data Science* (1st ed.). Florida: CRC Press.

Farrel, M. H., Liang, T., & Misra, S. (2019). Deep Neural Networks for Estimation and Inference. arXiv.org>econ>arXiv: 1809.09953. https://arxiv.org/abs/1809.09953.

Fisher, R. A. (1934). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd Ltd.

Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016). Credit Card Fraud Detection Using Convolutional Neural Networks. In Hirose, A., Ozawa, S., Doya, K. Ikeda, K., Lee, M., & Lui, D. (Eds.) *Neural Information Processing*. (pp. 483-490). Cham: Springer.

Galton, F. (1888). Co-relations and Their Measurement. *Proceedings of the Royal Society of London*, 45(1888-1889), 135-145.

Gerunov, A. (2016). Modelling Choice Under Radical Uncertainty: Machine Learning Approaches (MPRA Paper No. 69199). Retrieved from https://mpra.ub.uni-muenchen.de/69199/.

Gerunov, A. (2020). Binary Classification Problems in Economics and 136 Different Ways to Solve Them. *Center for Economic Theories and Policies*, Retrieved from http://www.bep.bg/p/papers.html.

Godin, B. (2009). The Culture of Numbers: The Origins and Development of Statistics on Science (INRS Working Paper No. 40). Retrieved from http://www.csiic.ca/PDF/Godin_40.pdf.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* (1st ed.). London: MIT Press.

Green, W. H. (2018). *Econometric Analysis.* Harlow: Pearson Education.

Green, W. H., & Hensher, D. A. (2010). *Modelling Ordered Choices* (1st ed.). New York: Cambridge University Press.

Hastie, T., Friedman, J., & Tibshirani, R. (2017). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.). New York: Springer.

Ikuodo, A., Lane, J., Staudt, J., & Weinburg, B. (2018). Occupational Classifications: A Machine Learning Approach (NBER Working Paper No. 24591). Cambridge, MA: National Bureau of Economic Research, Retrieved from https://www.nber.org/papers/w24951.pdf.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review: Papers & Proceedings*, 105(5), 491-495. doi:10.1257/aer.p20151023

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting Risk from Financial Reports with Regression. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, ACM, 272-280. Retrieved from http://public.kenan-flagler.unc.edu/faculty/sagij/N09-1031%5B1%5D.pdf.

Kütük, Y., & Güloğlu, B. (2019). Prediction of Transition Probabilities From Unemployment to Employment for Turkey via Machine Learning and Econometrics: A Comparative Study. *İktisat Araştırmalar Dergisi*, 3(1), 58-75. doi:10.24.954/JORE.2019.29

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research Commentary-Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research,* 24(4), 906-917. doi:10.1287/isre.2013.0480

Liu Y., & Xie, T. (2019). Machine Learning versus Econometrics: Prediction of Box Office. *Applied Economics Letters,* 26(2), 124-130. doi:10.1080/13504851.2018.1441499

Mayer-Schönberger, V., & Cukier, K. (2014). *Big Data: A Revolution that will Transform How we Live, Work, and Think* (1st ed.). London: John Murray.

McFadden, D. (1987). Regression-Based Specification Tests for the Multinomial Logit Model. *Journal of Econometrics*, 34(1-2), 63-83. doi:10.1016/0304-4076(87)90067-4

Mello, F. R., & Ponti, M. A. (2018). *Machine Learning* (1st ed.). Switzerland: Springer Intentional Publishing

Meng, C. Z., Liu, B. S., & Zhou, L. (2019). The Practive Study of Consumer Credit Risk Based on Random Forest. Advances in Intelligent Systems Research, 168, 101-106. doi:10.2991/masta-19.2019.17

Mitchell, T. M. (1980). The Need for Biases in Generalizations (Tech Report CBM-TR-117). New Jersey, Rutgers University: Rutgers CS Tech Report. Retrieved from http://www-cgi.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf.

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives,* 31(2), 87-106. doi:10.1257/jep.31.2.87

Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2017). Computer Vision Uncovers Predictors of Pyhsical Urban Change. *Proceedings of the National Academy of Sciences*, 114(29), 7571-7576.

Pandey, R., Dhoundiyal, M., & Kumar, A. (2015). Correlation Analysis of Big Data to Support Machine Learning. In 2015 Fifth International Conference on Communication Systems and Network Technologies, 996-999. doi:10.1109/CSNT.2015.32

Pearson, K. (1896). VII. Mathematical Contributions of the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of The Royal Society,* 187, 253-318. doi: 10.1098/rsta1896.0007.

Penczynski, S. P. (2019). Using Machine Learning for Communication Classification. *Experimental Economics,* 22, 1002-1029. Doi:10.1007/s10683-018-09600-z

Ramachandra, V. (2018). Deep Learning for Causal Inference. arXiv.org>econ>arXiv:1803.00149. https://arxiv.org/abs/1803.00149.

Randolph, K. A., & Myers, L. L. (2013). *Basic Statistics in Multivariate Analysis* (1st ed.). New York: Oxford University Press.

Rokach, L. (2010). *Pattern Classification Using Ensemble Methods*. Massachusetts: World Scientific Publishing Co. Pte. Ltd.

Serdobolskii, V. (2000). *Multivariate Statistical Analysis.* Boston: Kluwer Academic Publishers.

Sevüktekin, M. (2000). *Ekonometrik Model Kurma Teknikleri.* Bursa: Vipaş

Spanos, A. (1986). *Statistical Foundations of Econometric Modelling*. Cambridge: Cambridge University Press.

Storm, H., Baylis, K., & Heckelei, T. (2020). Machine Learning in Agricultural and Applied Economics, *European Review of Agricultural Economics,* 47(3), 849-892. doi:10.1093/erae/jbz033

Sugiyama, M. (2016). *Introduction to Statistical Machine Learning*. Amsterdam: MK Morgan Kaufmann.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1), 267-288.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: John Wiley and Sons.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), 3-28. doi:10.1257/jep.28.2.3.

Vellido, A., Martin-Guerrero, J. D., & Lisboa, P. J. G. (2012). Making Machine Learning Models Interpretable. In Proceedgins European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2012, Retrivied from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.431.5382

Wager, S., & Athey, S. (2017). Estimation and Inference of Heterogenous Treatment Effects Using Random Forests. arXiv.org>stat>arXiv:1510.04342. https://arxiv.org/abs/1510.04342.

Webb, A. (2002). *Statistical Pattern Recognition.* West Sussex: John Wiley ve Sons Ltd.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.

Yalçıntaş, A. (2018). n≥30 vs. n=all: Büyük Veri, Veri Obezitesi ve Kaybolan Nedensellikler. *Yıldız Social Science Review,* 4(2), 152-166.

Yao, J., Zheng, S., & Bai, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis* (1st ed.). New York: Cambridge University Press.

Zheng, E., Tan, Y., Goes, P., Chellappa, R., Wu, D J., Shaw, M., Sheng, O., & Gupta, A. (2017). When Econometrics Meets Machine Learning. *Data and Information Management*, 1(2), 75-83. doi: 10.1515/dim-2017-0012.