

Skorlama Algoritmaları 1 (Scoring Algorithms 1)

Amine YEŞİLYURT¹, Şadi Evren ŞEKER²

1. Gebze Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü
2. İstanbul Şehir Üniversitesi, Yönetim Bilişim Sistemleri Bölümü

Özet

Bu makalede, kredi skorlama alanı ve işletmelerin finansal başarısızlığının önceden tahmin edilmesi alanında literatür taraması gerçekleştirilmiştir. Bu alanda yapılan çalışmaların sonuçları değerlendirilerek kullanılan skorlama algoritmalarının başarıları karşılaştırılmıştır. Bu makale, bu alanda yayınlanan uzun soluklu bir yazı serisinin ilkidir. YBSAnsiklopedi üzerinde bu alanda farklı algoritmalar ve problemler bundan sonraki sayılarda da tartışılacaktır.

Anahtar Kavramlar: tahmin modeli, kredi skorlama, skorlama algoritmaları

Abstract

In this article, literature review was conducted in the field of credit scoring and predicting the financial failure of businesses. Result of studies were evaluated. Accuracy of scoring algorithms were compared each other. This paper is an early introduction to a long term paper series. Follow up papers will be published in YBSAnsiklopedi in the next issues.

Keywords: predictive model, credit scoring, scoring algorithms

1. Giriş

Birçok firma işlerini optimize etmek için istatistiksel modellerden faydalanır. Buna verilebilecek en güzel örneklerden biri kredi skorlama modelleridir. Bu modeller müşterinin gelecekteki davranışının tahmin edilmesini sağladığından tahmin modelleri olarak adlandırılırlar. Kredi skorlama özel bir tahmin modelidir. Bankacılık sektöründeki büyüme ile beraber geliştirilen modellerde istatistiksel yöntemlerin dışında makine öğrenmesi yöntemleri de kullanılmıştır. İncelenen çalışmalarda kullanılan modeller karşılaştırılarak yüksek başarı gösteren algoritmalar belirlenmeye çalışıldı. Literatürde, geleneksel yöntemleri kullanan modeller ile makine öğrenmesi kullanan modelleri karşılaştıran çalışmalar mevcuttur. Bunun yanında aynı sınıfta yer alan farklı algoritmaları karşılaştıran çalışmalar da mevcuttur. Bu çalışmada literatürdeki çalışmalar incelenerek skorlama alanında kullanılan başlıca algoritmalar ve algoritma başarıları belirlenmiştir.

2. İlgili Çalışmalar

Demirbulut ve arkadaşlarının [1] gerçekleştirdiği istatistiksel ve makine öğrenmesi metotları ile kredi skorlama çalışmasında geliştirilen yazılımda, bankacılık alanındaki veri setleri aracılığıyla kredi skorlama modelleri oluşturmak ve bu modellerin uygunluğunu karşılaştırmak amaçlanmıştır. Bu amaçla sekiz farklı algoritma kullanılmışlardır. İstatistiksel algoritmalar için Lojistik Regresyon, Probit Regresyon, Poisson Regresyon ve genelleştirilmiş katkı modeli yöntemlerini tercih etmişlerdir. Makine öğrenimi algoritmalarından ise K En Yakın Komşu, C4.5 Karar Ağacı, DVM (Destek Vektör Makineleri) ve YSA (Yapay Sinir Ağları) kullanmışlardır. Bu çalışma sonucunda makine öğrenmesi algoritmalarının istatistiksel algoritmalara göre zaman maliyeti açısından daha verimli olduğu sonucuna ulaşılmıştır. Regresyon algoritmaları numerik verilerden oluşan veri setlerinde daha verimlidir. Regresyon algoritmaları kategorik verileri kukla verilere dönüştürerek veri kümesinin boyutunu enine artırmaktadır. Makine öğrenimi algoritmaları ise veri kümesinin boyutunda artmaya sebep olmaz, veri boyutu sabit tutulur.

Ferdi Sönmez tarafından yapılan çalışmada [2] bir bankadan kredi talep eden müşterilerin başvurularının kabul edilmesi ya da reddedilmesi kararının verilmesinde, yapay sinir ağları (YSA) metodolojisini temel alan bir model önerilmiştir. Çalışmada çok katmanlı ağ (ÇKA) Levenberg-Marquard geriye yayılım algoritması (LM) ile kredi skoru hesaplayan bir yazılım modeli oluşturulmuştur. Ve model sonuçları karar ağacı modeli ile karşılaştırılmıştır. Çalışmada 10 kat çapraz doğrulama kullanılmıştır. Çalışma sonucunda, YSA ile oluşturulan modelden elde edilen sonuçlar karar ağacının kullanıldığı modele göre daha yüksek başarı oranına sahip olduğu görülmüştür. YSA modelinin ortalama başarısı %65,3 iken karar ağacı modelinin ortalama başarısı %61,5'dir.

Finansal başarısızlık tahmini ile ilgili 1966 yılında Beaver tarafından yapılan çalışmada, tek değişkenli model olan oran analizi kullanılmıştır. Tek değişkenli modeller finansal oranları teker teker ele alarak başarısızlık tahmini gerçekleştirir. Bu sebeple incelenen finansal orana göre farklı sonuçlar üretilir. Altman [5] tarafından gerçekleştirilen çalışmada kurumsal iflas tahmini probleminde geleneksel oran analizi yerine çoklu diskriminant analizi (Multiple Discriminant Analysis/MDA) kullanılmıştır. Çalışmada şirketlerin iflas tahmini %95 başarı oranına ulaşmıştır. Bu alanda yapılan bir diğer çalışmada [6] lojistik regresyon analizi kullanılarak ilk yıl için %96,12 başarı oranı ile şirketlerin iflas durumları tahmin edilmiştir. Lojistik regresyon yöntemi kullanarak yapılan bir diğer çalışma ise Salome Tabagari [7] tarafından gerçekleştirilmiştir. Bu çalışma bir bankaya kredi başvurusunda bulunan 500 müşteriye ait veri kümesi üzerinde yapılmıştır. Gerekli tüm hesaplamalar için, IBM SPSS istatistik yazılımının lojistik regresyon prosedürü kullanılmıştır. Çalışma sonucunda %82,8 başarı oranı elde edilmiştir.

Firmaların finansal başarısızlığının önceden tahmini için yapılan diğer bir çalışmada ise [10] veri madenciliği yöntemlerinin diskriminant analizine göre daha iyi sonuç verdiği gözlemlenmiştir.

Moraes [11] tarafından yapılan çalışmada müşterilerin kredi profillerini iyi ve kötü sınıflandırmak için YSA ve C4.5 karar ağacı algoritmaları kullanılmıştır. Karar ağacı modeli %94.08 YSA (MLP algoritması) ise %91.59 başarı göstermiştir.

Finansal başarısızlık alanında yapılan bir başka çalışmada [12] SVM algoritması ve Lojistik regresyon kullanılmıştır. SVM algoritmasının başarısı %75, Lojistik regresyon algoritmasının başarısı ise %71,8 olarak kaydedilmiştir.

Birçok çalışma [13] [14] karar ağaçlarının ya da yapay sinir ağlarının (YSA), diskriminant analizi ya da lojistik regresyona göre müşterinin kredi skorunun düşük olma ihtimalinin tahmininde daha doğru sonuçlar verdiğini göstermiştir. Kredi skorlama problemlerinde Neural Network, Lojistik Regresyona göre genel olarak daha iyi performans gösterir [4]. Fakat örnek boyutları küçük olduğunda, Lojistik Regresyon, tahmin gerektiren daha az

sayıdaki parametre nedeniyle daha iyi performans gösteren modeller üretebilir. Veri kümesinin büyümesi Neural Network için yararlı olurken Support Vector Machine'in performansı düşebilir [3].

2.1 Yapay sinir ağları ile doğrusal skorum modellerinin karşılaştırılması

Desai [9] tarafından yapılan araştırmada, YSA, doğrusal diskriminant analizi (LDA) ve Lojistik Regresyon algoritmaları kullanılmıştır. Çalışmada kullanılan veri kümesi 3 farklı firmadan alınmıştır. Veri kümesi, kredi başvurusunda bulunan müşterilerin bilgilerini içerir. YSA negatif tahminleri yüksek başarı oranı ile sınıflandırmıştır. Genel olarak YSA ve Lojistik Regresyon başarıları birbirine yakın sonuçlar vermiştir. LDA yönteminin başarısı ise diğer yöntemlere göre düşük çıkmıştır.

3. Kredi Skorum

Kredi başvuru sahibinin verilerinin istatistiksel yöntemler ile analiz ederek, müşterinin gelecekteki ödeme performansının iyi mi kötü mü olacağını bir model yardımıyla tahmin etmeyi amaçlar. Bu modelde kredi başvurusu esnasında müşteriden bilgi olarak alınan parametrelere ağırlık değerleri atanarak müşterilere puan verilmektedir. Böylece kredilerin temerrüde düşme olasılığı müşteri bazında tahmin edilmektedir. Modelde geçmişte kredi başvurusunda bulunmuş müşterilerin tüm mevcut bilgileri veri olarak kullanılabilir.

Kredi skorum tipleri:

- Başvuru Skoruması
- Davranış Skoruması

3.1. Skorum modelinde kullanılan parametrik yöntemler:

Lojistik regresyon, diskriminant analizi, doğrusal olasılık modeli gibi yöntemler kullanılarak geliştirilir. Güçlü olmakla birlikte sağlanması gereken varsayımlar vardır ve genellikle bu varsayımlar gerçekleşmez. Doğrusal olasılık modeli kolay ve hızlı bir yöntemdir ve uzun yıllar tercih edilmiştir. Diğer iki yöntemle göre daha yavaş olan lojistik regresyon ise ikili değişkenlerin modellenmesinde daha başarılıdır ve yaygın olarak kullanılan bir yöntemdir. Diskriminant analizi tekniğinin kullanılabilmesi için aşağıdaki varsayımların sağlanması gereklidir. Bu sebeple her durumda kullanılmamaktadır ve bu sebeple az tercih edilen bir yöntemdir.

- Bağımlı değişken veri kümesi, çok değişkenli normal dağılıma uygundur.
- Grupların kovaryans, varyans matrisleri homojendir.
- Bağımlı değişkenler arasında anlamlı bir korelasyon bulunmaz.
- Değişkenler arasında çoklu doğrusal bağ yoktur.
- Bağımsız değişken kümesi gereksiz değişken içermez.

3.2 Skorum modelinde kullanılan parametrik olmayan yöntemler:

Parametrik yöntemlerde varsayımları sağlama şartı olmasından ötürü çoğu kişi varsayım şartı gerektirmeyen parametrik olmayan yöntemleri tercih etmiştir. Yapay sinir ağları, karar ağaçları, k-n yakın komşu, doğrusal programlama bu yöntemler arasındadır.

3.3 Kredi skorlama probleminde model karşılaştırması:

Kredi endüstrisinin büyümesi ile birlikte geliştirilen kredi skorlama modellerinin de sayısı artmıştır. Genel olarak kredi skorlama modelleri iki sınıfa ayrılır. Bunlar: klasik istatistiğe dayalı geleneksel madencilik modelleri ve yapay zeka tekniklerine dayalı modern madencilik modelleridir. Genetik programlama, sinir ağları, fuzzy kuralları gibi modern modeller gittikçe popülerleşen araştırma alanlarıdır. Yaygın kredi skorlama modelleri arasında lojistik regresyon, sinir ağları, kaba küme(rough sets), genetik programlama, SVM vardır. Lojistik regresyon en yaygın sınıflandırma modellerinden birisidir. Diğer istatistiksel regresyon modellerinin aksine, çeşitli dağıtım fonksiyonları ve kredi puanlama problemleri için daha uygundur. Lojistik regresyon lineer sınıflandırmada oldukça güçlüdür. Ancak doğrusal olmayan problemler için kötü performans sergiler. Rough Set, bilginin çelişkili olması ve belirsizliği ile baş eden matematiksel bir araçtır. Fuzzy kümeleri (fuzzy sets) ile karşılaştırılırsa, rough set verileri işlemek için herhangi bir varsayıma ihtiyaç duymaz. Tümevarımsal bir yöntemeye dayanır. Avantajı karar desteği sağlama yeteneğidir. Kredi skorlama problemlerine başarılı bir şekilde uygulanmış olsa da zayıf tarafı kötü tahminde bulunmasıdır. Yeni bir girdi için tahmin gerçekleştiremediği durumlar olabilmektedir. GP ise yaygın kullanılmasına karşın bazen çok zaman alır ve yeni bir müşteri için sonuç üretmemektedir. Yapay sinir ağlarının pek çok avantajı vardır fakat over-fitting(algoritmanın eğitim verisini öğrenmek yerine ezberlemesi) ve yerel-minimum(local minima) tuzağına düşmesi kolaydır. SVM genellikle optimal çözümler sunar. Ancak, hangi bilginin fazlalık olduğunu, hangi bilginin daha yararlı ve daha önemli bir rolünün olduğunu belirlemek çok zordur.

Defu Zhang ve arkadaşları [8] tarafından gerçekleştirilen bu çalışmada genetik programlama(GP) geri-yayımlı sinir ağları(BP) ve Destek Vektör Makinesi(SVM) kredi skorlama problemi üzerinde uygulanmıştır. Sonuçlar karşılaştırılarak modeller arasında karşılaştırma yapılmıştır. Çalışma sonucunda elde edilen sonuçlar:

1. Sinir Ağları: Karmaşık verilerde lineer olmayan ilişkileri ortaya çıkarmak için kullanılır. Kolay uygulanması ve genelleme yapmakta yetenekli olması nedeniyle geri beslemeli ağlarda yaygın olarak kullanılan bir mimaridir.
2. Genetik Programlama: GP, Darwin'in doğal seleksiyon ve evrim ilkesine dayanmaktadır. GP, sürekli daha iyi çözümleri kabul ederek en uygun olan çözümün hayatta kalması fikrini kullanır. Kalıtım, doğal seçim, mutasyon gibi genetik süreçlerden ilham alır. Genetik algoritmaları, geleneksel lineer olmayan optimizasyon algoritmalarından farklıdır. Tek bir çözümde iyileştirmeler yapmaktan ziyade daha iyi çözümlerin oluşturulduğu bir çözüm popülasyonunu yönetir.
3. Destek Vektör Makineleri (SVM)

Bu çalışmada [8] Almanya ve Avusturalya ülkelerine ait kredi verileri üzerinde yapılan çalışma sonuçları Tablo-1'de gösterilmiştir.

	Ortalama Başarı (Almanya)	Ortalama Başarı (Avusturalya)
BP	<u>79,77</u>	89,29
GP	79,53	<u>89,47</u>

SVM	78,94	88,50
-----	-------	-------

Tablo 1 Kredi verileri üzerinde algoritma başarıları (%)

Defu Zhang ve arkadaşları [8] tarafından yapılan bu çalışmada Tablo-1’de görüldüğü üzere GP ve BP modellerinin başarısı her iki veri kümesinde için SVM algoritmasına göre yüksek çıkmıştır.

4. Lojistik Regresyon için Bazı Modelleme Kuralları

Lojistik regresyon uygulanırken dikkat edilmesi gereken bazı ipuçları vardır. Bunlar;

- Veri kümesi en az 30 satırdan oluşmalıdır.
- Lojistik regresyon modeli her 30-50 satır için en fazla bir değişken içermelidir
- Model, o alanda profesyonel kişi tarafından belirlenmiş önceden seçilmiş değişkenler içermelidir. Önceden seçilmiş değişkenler, modelleme sürecinden önce model kararlarını etkileyeceği düşünülen değişkenlerdir.
- Lojistik regresyon modelleri minimal değişken kümesine sahip olmalıdır. Bu kümenin büyüklüğü hakkına sayısal bir kısıttan söz edilemez. Fakat modelin performansını çok az etkileyen değişkenler modele dahil edilmemelidir.
- Analiz sürecinde istenen parametreler kesin değildir. Örneğin; insan davranışlarını içeren modeller fiziksel olgu içeren modeller ile karşılaştırıldığında daha büyük p-değeri ve daha düşük başarı oranına sahip olabilir.

Varsayımların az olması lojistik regresyonun tercih edilmesini etkilemektedir. Lojistik regresyonun sonuçlarının yorumlanması kolay olması bir avantajdır.

5. Sonuç

Yapılan literatür taraması sonuçlarına göre çoğu modelde yapay sinir ağları algoritması en yüksek başarı oranını vermektedir. Bu çalışmada incelenen makalelerden elde edilen sonuçlara göre algoritma başarı sonuçlarının karşılaştırılması Tablo-2’de özetlenmiştir.

YSA > Karar Ağacı
C4.5 > MLP
Lojistik Regresyon > Çoklu Diskriminant Analizi
SVM > Lojistik Regresyon
YSA > Lojistik Regresyon
YSA \cong Lojistik Regresyon > LDA
YSA > Genetik Programlama > SVM

Tablo 2 Algoritma Başarılarının Karşılaştırılması

Kaynakça

[1] Yunus Emre Demirbulut, Mehmet S. Aktaş, Oya Kalıpsız, Selçuk Bayracı, İstatistiksel ve Makine Öğrenimi Yöntemleriyle Kredi Skorum, http://ceur-ws.org/Vol-1980/UYMS17_paper_83.pdf

[2] Ferdi SÖNMEZ, Kredi Skorumun Belirlenmesinde Yapay Sinir Ağları ve Karar Ağaçlarının Kullanımı: Bir Model Önerisi, ABMYO Dergisi Sayı 40 - (2015) - (1-22)

[3] S.F. Crone, S. Finlay, Instance sampling in credit scoring: An empirical study of sample size and balancing, *International Journal of Forecasting* 28 (2012) 224– 238, 2012.

[4] J.N. Crook, D.B. Edelman, L.C. Thomas, Recent developments in consumer credit risk assessment, *European Journal of Operational Research*, 183(3), 1447– 1465, 2007.

[5] E. I. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”, *The journal of finance*, vol. 23, no. 4, pp. 589-609, 1968.

[6] Ohlso, J. A., “Financial Ratios and the Probabilistic Prediction of Bankruptcy”, *Journal of Accounting Research*, 18(1), 109-131.

[7] Tabagari, Salome, “Credit scoring by logistic regression, University of Tartu, 2015.

[8] Defu Zhang, Hongyi Huang, Qingshan Chen, “A Comparison Study of Credit Scoring Models”, *Third International Conference on Natural Computation (ICNC 2007)*

[9] Desai, Vijay S., Jonathan N. Crook, and George A. Overstreet. "A comparison of neural networks and linear scoring models in the credit union environment." *European Journal of Operational Research* 95.1 (1996): 24-37.

[10] Emre YAKUT, Bekir ELMAS, İşletmelerin finansal başarısızlığının veri madenciliği ve diskriminant analizi modelleri ile tahmin edilmesi, *Afyon Kocatepe Üniversitesi, İİBF Dergisi (C. XV, S. I, 2013)*

[11] Sousa, Marcos de Moraes, and Reginaldo Santana Figueiredo. "Credit analysis using data mining: application in the case of a credit union." *JISTEM-Journal of Information Systems and Technology Management* 11.2 (2014): 379-396.

[12] Önder, Ceren. "Bankruptcy prediction with support vector machines." (2010).

[13] Landajo M., Andr s, J., Lorca, P.(2007). Robust neural modeling for the cross-sectional analysis of accounting information, *European Journal of Operational Research*, 177 (2), s.1232-1252.

[14] Sheen, J.N.(2005). Fuzzy financial profitability analyses of demand side management alternatives from participant perspective, *Information Sciences*, 169 (3), s. 329-364.