EasyChair Preprint

№ 4458

Credit Usage Prediction in the Banking Sector Using Big Data Processing and Analysis Techniques: A Case Study

Ümit Tigrak, Nail Taşgetiren, Erdal Bozan, Güven Gül, Emir Demirci, Hakan Sarïbïyïk and Mehmet S. Aktaş

October 24, 2020

Büyük Veri İşleme ve Analizi Teknikleri Kullanılarak Bankacılık Sektöründe Kredi Kullanımı İhtiyacı Tahmini: Durum Çalışması

Credit Usage Prediction in the Banking Sector Using Big Data Processing and Analysis Techniques: A Case Study

Umit Tigrak[1], Nail Taşgetiren[2], Erdal Bozan[1], Guven Gul[1], Emir Demirci[1], Hakan Saribiyik[1], Mehmet S. Aktas[2]

[1]R&D Center, Fibabanka, Istanbul, Turkey
[2]Yildiz Technical University, Istanbul, Turkey
umit.tigrak@fibabanka.com.tr, nail.tasgetiren@std.yildiz.edu.tr, {erdal.bozan, guven.gul, emir.demirci, hakan.saribiyik}@fibabanka.com.tr, aktas@yildiz.edu.tr

***Özetçe*— Bu araştırma kapsamında, bankacılık alanında müşterilerin kredi ihtiyacının büyük veri işleme ve analizi platformları üzerinde makine öğrenmesi modelleri kullanılarak tahmin edilebilmesi için bir metodoloji önerilmektedir. Önerilen metodolojinin prototip uygulaması tasarlanmış, geliştirilmiş ve bankacılık müşterilerinin kredi davranış veri setleri üzerinde uygulanmıştır. Geliştirilen prototip uygulama üzerinde, yöntemin tahmin başarısı, ölçeklenebilirliği, modellerin oluşturulmasında eğitim için gereken çalışma süresi metrikleri için performans testleri yapılmıştır. Elde edilen sonuçlar, önerilen yöntemin bankacılık sektöründe kullanılabilirliğini ortaya koymaktadır.**

Anahtar Kelimeler — bankacılık sektörü, kredi kullanım tahmini, kredi satın alma eğiliminin modellenmesi, büyük veri analizi.

***Abstract*— Within the scope of this research, a methodology is proposed for estimating the loan needs of customers in the banking field using machine learning models on big data processing and analysis platforms. Prototype application of the proposed methodology has been designed, developed and applied on the credit usage behavior data of banking customers. Performance tests were carried out on the prototype application developed for the estimation success of the method, its scalability, and the working time metrics required for training in creating the models. The results obtained reveal the usability of the proposed method in the banking sector.**

Keywords — banking sector, credit usage prediction, modeling the credit purchase propensity, big data analysis.

I. GİRİŞ

Finans kuruluşları arasındaki rekabetle beraber, kurumların sağladığı ürün ve hizmetlerde çeşitlilik artarken, pazarlama ve teknoloji alanlarındaki gelişmelerle beraber kurumların günümüz koşullarına adapte olma zorunluluğunu ortaya çıkmaktadır. Değişen koşullarla beraber bireysel kredilere yönelik pazarlama faaliyetleri için yenilikçi stratejilere ihtiyaç bulunmaktadır.

Bireysel ihtiyaç kredileri, bireylerin ticari amaç gütmeyen her türlü kişisel ihtiyaçlarının, belirlenen bir ödeme planına göre taksitle geri ödeme imkânı sağlayan bir finansman türüdür. Bankaların bireysel ihtiyaç kredisi hizmeti vermeye başladığı zamanlar, belirli bir malın satın alımı ile sağlanan kredilerde teminat alımı ve taksit ödemeleri belirli kurallar çerçevesinde yapılırken, zamanla koşullar esnetilerek herhangi bir ihtiyacın karşılanması ile yaygınlaştırılmaktadır. Otomobil, dayanıklı tüketim malları ve ev eşyaları alımında tüketicilere finansman sağlamak amacıyla verilen tüketici kredileri yanında eğitim harcamaları, sağlık ve seyahat masraflarını ve hatta sigorta primlerini finanse eden tüketici kredileri de kullandırılmakta olduğunu görmekteyiz.

Günümüzde bankalar, milyonlarca müşteriye hizmet vermektedir. Bankaların müşterilere sunduğu finansal çözümler çok çeşitli olmakla beraber, bu ürünler için uygulanan pazarlama faaliyetleri, uzman görüşüne dayalı kurallarla ya da farklı sezgisel yöntemler kullanılarak uygulanmaktadır. Uzman görüşüne dayalı yaklaşımını istatiksel tahmin modelleri ve optimizasyon teknikleri ile geliştirmeyi hedeflemektedir. Ayrıca iletişim kanallarının çeşitlenmesi ile beraber müşteriye ulaşma seçenekleri artarken, iletişim maliyetiyle beraber hedef alınacak müşterilerin kim olacağı, iletişimin ne zaman kurulacağı gibi sorunlar ortaya çıkmaktadır.

Müşterinin kredi ürününe ihtiyaç duyup duymadığını istatistiksel olarak ölçmek, iletişim kanallarının kapasitesini hesaba katarak kredi eğilimi en yüksek müşterileri belirlemek, bankacılık sektörü açısından oldukça önemli bir gereksinimdir. Bankaların hizmet verdiği müşteri sayısının artmasıyla beraber kredi eğilimini ölçmek için kullanılan teknolojilerin de büyüyen müşteri verisiyle orantılı olarak ölçeklenebilir olması gerekmektedir. Bu nedenle kredi eğilimi tahmini için kullanılan çözümlerin, büyük veri işleme ve analizi tekniklerini kullanması önemli bir gereksinimdir.

Bu gereksinimleri karşılamak amacıyla, araştırmamız kapsamında, bireysel kredi ihtiyacı tahmininin büyük veri işleme teknolojileri kullanılarak nasıl gerçekleştirilebileceği üzerine çalışılmıştır. Bu amaçla bankacılık müşterilerinin kredi eğilimlerinin tahmin edilmesine olanak verecek bir veri işleme ve analizi iş süreci metodolojisi önerilmiştir. Önerilen metodoloji kapsamında, kredi kullanma eğilimini etkileyen öznitelikler belirlenmiştir. Her müşteri için özniteliklerden oluşan ve kredi eğilimi tahmininde kullanılacak davranış vektörleri oluşturulmuştur. Kredi eğilimi davranış vektörleri etiketleri, müşterinin kredi kullanıp kullanmaması bilgisine dair olarak belirlenmiştir. Müşteri kredi kullanma eğilimi davranış vektörlerinin bir araya gelmesi ile oluşan davranış matrisi üzerinde gözetimli makine öğrenmesi yöntemi uygulanarak, kredi kullanma eğilimi modelleri oluşturulmuştur. Bankacılık müşteri verilerinin zaman içinde hızla artan nitelikte bir veri olması nedeniyle, model oluşturmak için büyük veri işleme

platformları üzerinde makine öğrenmesi algoritmaları kullanılmıştır.

Önerilen metodolojinin kullanılabilirliğini ortaya koymak adına, bu mimarinin ilk prototip uygulaması geliştirilmiştir. Geliştirilen prototip uygulama üzerinde, doğruluk ve ölçeklenebilirlik açılarından, performans testleri yapılmıştır. Geliştirilen metodolojinin kredi kullanma eğilimi tahminindeki başarısı, iki farklı bankacılık müşteri veri seti üzerinde irdelenmiştir. Araştırmamız kapsamında, model oluşturmak için kullanılan algoritmanın büyük veri platformu ortamında ölçeklenebilirliği de irdelenmiştir. Elde edilen sonuçlar, önerilen metodolojinin büyük veri işleme platformu üzerinde başarılı ve ölçeklenebilir kredi kullanımı tahminleri yapabildiğini ve kullanılabilir nitelikte göstermektedir.

Bu bildirinin organizasyon yapısı şu şekildedir: 2. bölümde bu araştırma kapsamında üzerinde çalışılmakta olan araştırma soruları belirtilmektedir. 3. bölümde konuyla ilgili literatür taraması anlatılarak, 4. bölümde araştırma kapsamında önerilen metodoloji tanıtılmaktadır. 5. bölümde ise önerilen metodolojinin prototip uygulaması üzerinde yapılan değerlendirme ve elde edilen sonuçları paylaşılmaktadır. 6. bölüm, araştırma sonucunda elde edilen deneyimleri özetlemekte ve gelecekteki çalışmaları aktarmaktadır.

II. ARAŞTIRMA SORULARI

Bankalar, müşterilerine ekonomik faydaya dönüşecek bankacılık kredi ürünlerini sunma zorunluluğundadır. Kredi ürünü önerilerinin de kredi kullanma ihtiyacı ya da eğilimi içinde olan müşterilere yapılması gerekmektedir. Bankacılık alanında, kredi ürünü çeşitliliğinin ve müşteri sayısının fazla olması, doğru müşteriye kredi ürünü önerisinin yapılabileceği bir sistemi gerektirmektedir. Uzman görüşüne dayalı kredi önerisi yaklaşımları, makine öğrenmesine dayalı tahmin yöntemleriyle geliştirmek, bankacılık alanında öncelikli bir hedef halini almıştır. Bu kapsamda, müşterilerin kredi kullanma eğilimlerini tahmin edecek makine öğrenmesi yöntemlerime dayalı veri analizi iş süreçlerine ihtiyaç duyulmaktadır. Bunun yanı sıra bankaların hizmet verdiği müşteri sayısının artması ve bu müşterilerin kredi davranışı verilerinin de sürekli artması; müşterilerin kredi kullanma eğilimini ölçmek için kullanılan teknolojilerin, büyüyen müşteri verisiyle orantılı olarak ölçeklenebilir ve yüksek performansta tasarlanması ve geliştirilmesini gerekmektedir. Bu gereksinimlere yanıt verebilmek için, bu çalışmada belirlediğimiz araştırma sorularını aşağıda maddeler halinde belirtiyoruz.

(1) Kredi başvurusu yapan müşterilerin kredi kullanımını tahmin edecek bir veri analizi süreci aracının yazılımın mimarisi nasıl olabilir?

(2) Müşteriye kredi önerisinin yapılması için; müşterinin bankacılık işlemleri için gerçekleştirdiği davranışlarını en iyi tanımlayan; müşterinin ihtiyaç ve davranışlarının modellenmesine olanak veren; kredi kullanma eğilimi tahmininde kullanılabilecek, öznitelikler neler olabilir?

(3) Bu öznitelikler kullanılarak oluşturulacak modeller, kredi kullanma eğilimi tahmininde başarılı sonuçlar ortaya koyabilir mi?

(4) Kredi kullanım eğilimi tahmini için kullanılan gözetimli makine öğrenmesi yöntemlerinin, büyük veri işleme ve analizi platformları üzerinde çalıştırıldığında ölçeklenebilirliği ve performansı, büyüyen eğitim veri setleri düşünüldüğünde, nasıldır?

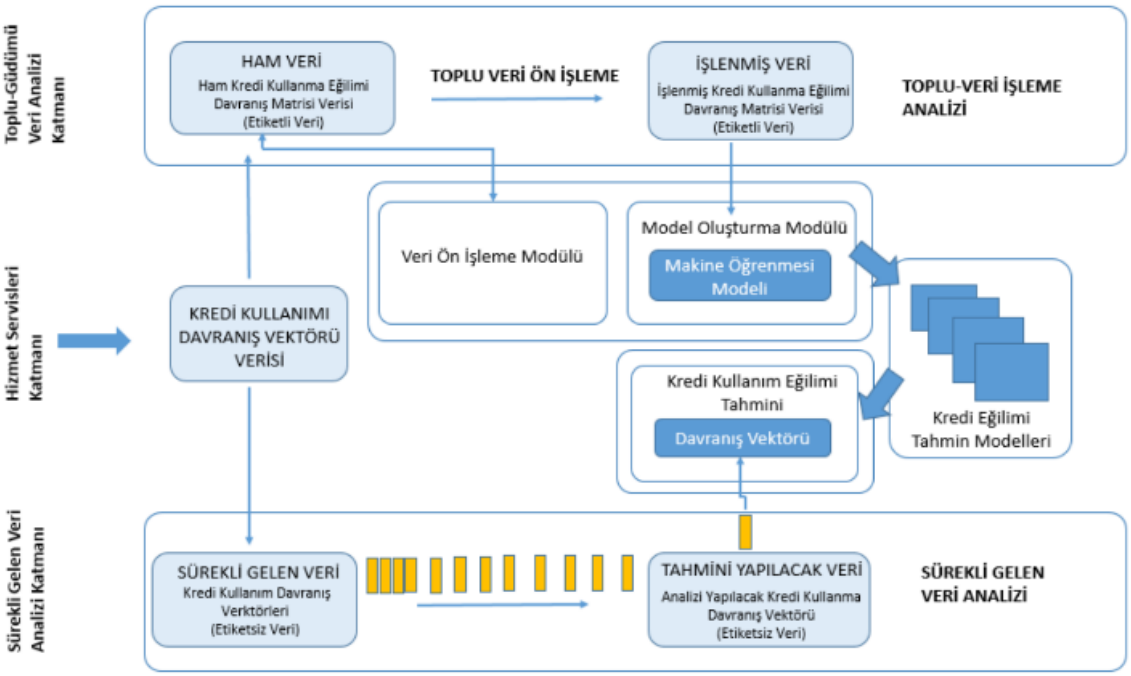
III. LİTERATÜR TARAMASI

Müşteri kredi kullanım eğilimi tahmininde farklı yöntemlerin uygulandığı görülmektedir. Bu alanda yapılan araştırmalar sırasıyla Lojistik regrasyon ve random forest [1], destek vektör makineleri (SVM) [2], karar ağaçları [3], ve j48, bayesNet, NaiveBayes [4], yöntemlerini kullanmıştır. Ancak bu araştırmada makine öğrenmesi yöntemi olarak XGBoost kullanılmıştır. XGBoost karar ağacı tabanlı, bütünleştirilmiş (ensemble) bir makine öğrenmesi algoritmasıdır. XGBoost parelel işleme tekniklerini uygulayan yapısı, bu algoritmanın veri paralelismi sağlandığında daha hızlı çalışabileceğini işaret etmektedir. Bu yüzden bu araştırma kapsamında makine öğrenmesi yöntemi olarak XGBoost yöntemi kullanılmıştır.

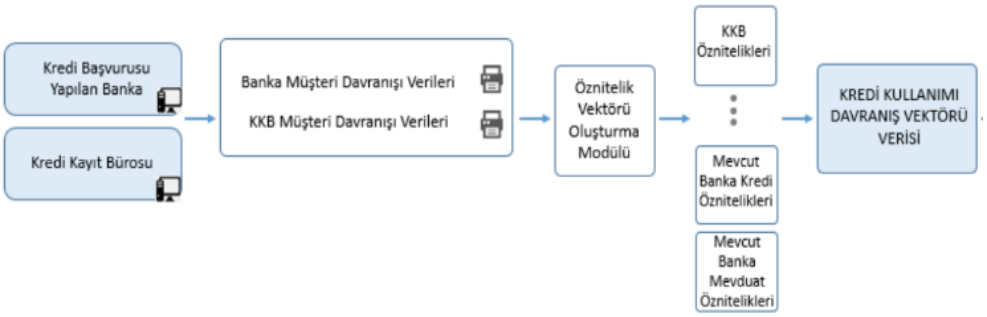
Veri ve veriyi anlatan üstverinin yönetiminin nasıl sağlandığı üzerinde farklı araştırmaların literatürde yer aldığı görülmektedir [5-7]. Dağıtık veri saklama platformları için bilgi sistemleri alanında ve dağıtık ortamda verinin aranması alanında da araştırmalar yapıldığı görülmektedir [8-13]. Bu araştırmalardan farklı olarak, bu çalışma kapsamında, veriden müşteri kayıp analizi tahmini amaçlı anlam çıkartılması üzerinde odaklanılmıştır.

IV. ÖNERİLEN METODOLOJİ

Bankacılık müşterileri, kredi kullanım davranışları ile ilgili tüm bilgileri bankaların farklı veri tabanları ve kredi kayıt bürosu kayıtları üzerinde bulunmaktadır. Bu araştırma kapsamında, bir bankaya ait bankacılık müşterilerinin kredi kullanım davranış verileri periyodik olarak (aylık bazda) bu iki veri kaynağı üzerinden toplanmıştır. Kredi kayıt bürosu kayıtları üzerinden tüm bankalara ait finansal verilerini içeren kayıtlar ve bankanın kendi kayıtlarından ise müşterilerin banka üzerindeki kredi kullanım davranışları, kredi kullanım eğilimi davranış matris veri setini oluşturmak için tek bir veri tabanı üzerinde toplanmıştır. Burada toplanan veri; zaman ekseni üzerinde, zaman ilerlerken, sürekli toplanması gereken veri olması ve çeşitlilik içermesi nedeniyle büyük veri niteliği taşıyan bir veri olarak nitelendirilmektedir. Bu araştırma kapsamında, bankacılık müşteri verileri üzerinde, kredi kullanımı eğilimi tahmini yapılabilmesi için lamda yazılım mimarisi [14] kullanımını öneriyoruz. Önerilen mimari, Şekil 1'de resmedilmektedir. Bu mimari yapıdaki veri analizi iş sürecinin girdisi olan verileri (kredi kullanımı davranış vektörü verisi) üretecek olan öznitelik çıkarma süreci Şekil-2'de resmedilmektedir.



Şekil 1: Bankacılık Müşterileri Kredi Kullanımı Tahmini İş Süreci Yazılım Mimarisi



Şekil 2: Kredi Kullanımı Tahmini İçin Özniteliklerin Oluşturulması Süreci

Şeki-1 ve Şekil-2'de görüleceği üzere, müşterilere ait kredi kullanım davranışı verileri, 4 temel veri analizi süreci aşamasını içermektedir: Kredi Davranışını Tanımlayan Öznitelik Vektörlerinin Oluşturulması (Şekil 2), Ham Kredi Kullanma Davranış Verişinin İşlenmesi (Şekil 1), Kredi Kullanma Eğilimi Tahmin Modelinin Oluşturulması (Şekil 1), Kredi Kullanma Eğilimi Tahmini (Şekil 1). Bu aşamaların sonucunda, her müşteri için kredi kullanma eğilimi skoru oluşturulmaktadır. Önerdiğimiz mimari yapıyla, bu bildirinin "Araştırma Soruları" bölümündeki 1. Araştırma sorusunu yanıtlıyoruz. Önerdiğimiz mimarinin ne kadar efektif olduğunu ise bu bildirinin "Prototip Uygulama ve Önerilen Metodolojinin Değerlendirilmesi" bölümünde irdeliyoruz. Aşağıda bu temel veri analizi süreçlerinin detayları verilmektedir.

1) Öznitelik Vektörlerinin Oluşturulması: Bankacılık müşterilerinin kredi kullanma davranışlarını tanımlayabilmek için öznitelikler belirlenmiştir. Bu verileri, aşağıda temel kategoriler kapsamında açıklıyoruz.

Kredi Kayıt Bürosundan elde edilen diğer banka kredi kullanımı öznitelikleri (KKB Öznitelikleri): Müşterilerin diğer banka kredi ürün sahipliği; Kredi yaşı; Kullanılan kredinin zaman içindeki değişimi; Ödeme performansı; Kredi kayıt bürosu risk skoru; Kredi kayıt bürosu borçluluk endeksi. Diğer banka kredileri ile ilgili öznitelik değerleri Kredi kayıt Bürosu' nda elde edilen veri seti üzerinden hesaplanmaktadır. Kredi başvurusu yapılan banka kredi kullanımı öznitelikleri (Mevcut Banka Kredi Öznitelikleri): Müşterilerin bankadaki kredi ürün sahipliği, çeşitliliği; Kredi yaşı; Kredi kullanım detayı ve zaman içindeki değişimi; Ödeme araçları ve performansı. Mevcut banka kredileri ile ilgili öznitelikler, müşterinin banka üzerinde bıraktığı, kredi kullanımı davranışı verileri üzerinden hesaplanmaktadır. Kredi başvurusu yapılan banka mevduat hesabı öznitelikleri (Mevcut Banka Mevduat Öznitelikleri): Müşterinin banka' da sahip olduğu varlıkların bilgisi; Varlıkların yönetildiği ürünler; Bu ürünlerin zaman içerisindeki değişimini ve kullanım alışkanlığını gösteren değişkenler. Mevduat hesabı ile ilgili öznitelikler bankanın müşteri mevduat

hesabı verileri üzerinden hesaplanmaktadır. Kredi başvurusu yapılan banka diğer ürün öznitelikleri (Mevcut Banka Diğer Ürün Öznitelikleri): Müşterinin kredi başvurusu yapılan bankada bulunan kredi ve mevduat ürünleri dışında kalan ürünlerin sahipliği; Ürünlerin kullanımı ve bu ürünlerin zaman içerisindeki değişimi; Ürünlerin kullanım alışkanlığını gösteren değişkenler. Bu öznitelikler, müşterinin kredi başvurusu yapılan bankadan elde edilen diğer yatırım ürünleri ile ilgili verileri üzerinden hesaplanmaktadır. Müşteri demografik bilgileri ile ilgili öznitelikler: Müşterinin kredi başvurusu yaptığı bankadaki yaşı; Diğer bankalardaki kredililik yaşı; Mesleği, Yaşı ve Eğitim seviyesi. Müşteri demografik bilgileri ile ilgili öznitelikler, hem mevcut banka veri tabanı hem de KKB veri tabanı verileri üzerinden belirlenmektedir. Müşteri para transferi bilgileri ile ilgili öznitelikler: Müşterinin banka içi ve dışı, düzenli veya düzensiz para transferleri; Para transferlerinin zamana bağlı değişimi ile ilgili öznitelikler. Para transferi ile ilgili öznitelikler, kredi başvurusu yapılan bankanın veri tabanı üzerinden elde edilen veriler üzerinden hesaplanmaktadır. Bu kapsamda önerisini yaptığımız öznitelik seti ile bu bildirinin "Araştırma Soruları" bölümündeki 2. Araştırma sorusunu yanıtlıyoruz. Bu öznitelik seti kullanarak elde edilen modellerin tahmin başarısını ise yine bu bildirinin "Prototip Uygulama ve Önerilen Metodolojinin Değerlendirilmesi" bölümünde irdeliyoruz.

2) Ham Kredi Kullanma Davranış Verişinin İşlenmesi: Bu aşama eldeki ham verinin modellemeye uygun bir hale dönüşmesini sağlayacak, normalizasyon, aykırı değer tespiti, veri gruplaması, eksik veri yönetimi, zaman bazlı gruplama, kategorik değişken yönetimi adımlarını içermektedir.

3) Kredi Kullanma Eğilimi Tahmin Modeli Oluşturulması: Model tarafından işlenebilecek olan verinin eğitim, test ve validasyon gibi farklı örneklemelere bölünmesi, değişken, model ve model parametrelerinin seçimi bu aşamada yapılmıştır. Farklı metrikler için en yüksek performansa sahip olan öğrenme algoritması ve değişken seti bu aşamada belirlenmiştir. Yapılan çalışma sonucunda XGBoost gözetimli öğrenme algoritması'nın [15] en başarılı tahmin üreten modelleri oluşturduğu gözlemlenmiş ve kapsamlı deneysel çalışmamızda, bu algoritma kullanılmıştır. Bu algoritma kullanılarak, bu aşamada, kredi kullanma eğilimi tahmininde kullanılacak tahmin modelleri, eğitim verisinden oluşturulmaktadır.

4) Kredi Kullanma Eğilimi Tahmini: Bu aşamada, sisteme sürekli olarak sorgu şeklinde gelen müşteri kredi davranışı, davranış vektörlerinin tahmini gerçekleştirilmektedir. Bir önceki aşamada oluşturulan tahmin modeli kullanılarak, müşterinin kredi kullanma eğilimi tahmin yapılmaktadır. Müşteri kredi kullanma eğiliminin ikili karar şeklinde (kullanma eğilimi var / kullanma eğilimi yok) tahminlenmesinin yanı sıra; bu aşamada, tahmin için hesaplana doğruluk metriği verisi önceki aylara göre kalibre edilerek müşterinin kredi eğilimi skoru da hesaplanmaktadır.

V. Önerilen Metodolojinin Prototip Uygulamasının Değerlendirilmesi

Ölçeklenebilirlik testi için büyük veri işleme plaftormu olan Apache Spark veri işleme motorunun son versiyonu olan spark-3.0.0 versiyonu kullanılmıştır. Apache Spark platformu üzerinde ise XGBoost gözetimli öğrenme algoritması kullanılarak büyük ve küçük boyutlu verilerde eğitim işlemi yapılmıştır. Dağıtık veri mimarisi kullanılarak her bir node üzerinde Linux Ubuntu işletim sistemi çalıştırılmıştır. Testler sırasında tek-node ve çok-node hesaplama kümeleri oluşturulurken sanallaştırma teknolojileri kullanılmıştır. Sanallaştırma altyapısı Vagrant açık kaynaklı çerçeve programı kullanılarak oluşturulmuştur. Vagrant, tek bir iş akışında sanal makine ortamları oluşturmak ve yönetmek için geliştirilmiş bir araçtır [16]. 3 farklı sanal makine kurulmuştur. Makinenin özellikleri: 2GB RAM bellek, 2 Çekirdekli CPU ve hızı 2.9Ghz şeklindedir.

Teknoloji ve Altyapı: Kredi eğilim modelleri Jupyter lab tabanlı bir ortamda geliştirildi. Ortam 64gb RAM ve 72 işlemci çekirdeğine sahip olup, LINUX Redheat işletim sisteminde çalıştırılmıştır.

Veri Seti: Bu araştırma kapsamında, kredi kullanma eğilimi tahmini başarısının araştırılması için iki farklı veri seti kullanılmıştır. Kredi kullanma eğilimi tahmininde kullanılan makine öğrenmesi yöntemi ve büyük veri işleme platformunun ölçeklenebilirlik performansını ortaya koymak için kullanılan bu veri setlerinin farklı büyüklüklerde olan versiyonları kullanılmıştır. Araştırmamız kapsamında kullanılan veri setlerini aşağıda açıklıyoruz: Veri Seti-1 ve Veri Seti-2: Öznitelik seti aynı olan bu iki veri seti, bağımlı değişken açısından farklılaşmaktadır. İlk sette müşterilerin, bağımlı değişkeni Kredi Kayıt Bürosu üzerinden elde edilen tüm bankalardan kullandığı kredilerden oluşurken, ikinci sette sadece kredi başvurusu yapılan bankadan yaptığı kullandırımlar ele alınmaktadır. Veri Seti-3: Farklı sayıda müşteri verilerinin olduğu verilerden oluşmaktadır. Veri seti toplamda 23 farklı öznitelik içermektedir ve özniteliklerden bir tanesi hedef değişken (etiket verisi) şeklindedir.

İlk iki veri seti için; Davranış matrisi oluşturulurken, 6 aylık bir gözlem periyodu düşünülerek, 6 aylık müşteri davranışı verisi (eğitim verisi) dikkate alınmıştır. Eğitim verisi üzerinde müşterilerin aylık bazda kredi kullanıp kullanmadığı bilgisi (bağımlı değişken/etiket verisi) bulunmaktadır. Tahmini yapılacak olan müşteri davranış verileri (test verisi) için, gözlem periyodu tamamlanması sonrasında gelen 1 aylık müşteri davranışı verileri kullanılmıştır. Test verisi içinde de müşterinin kredi kullanım bilgisi (etiket verisi) bulunmaktadır.

Gerçekleştirilen testlerde; 2020 yılı Şubat ve Mart aylarındaki bankacılık müşterisi kredi kullanma eğilimlerinin tahmini yapılmıştır. Kullanılan veriler, maskelenmiş, gerçek banka müşterisi kredi kullanım davranış verileridir.

Veri Seti-1 için; 2020 Ocak ayı tahminlerinin yapılabilmesi için kullanılan tüm veride 541,072 kayıt bulunmaktadır. Bu kayıtların 135,268 adedi Kredi Kayıt Bürosu kayıtlarına göre kredi kullanan müşterilerden oluşmaktadır. 2020 Şubat ayı tahminlerinin yapılabilmesi için kullanılan tüm veride 539,905 kayıt bulunmaktadır. Bu kayıtların 134,976 adedi Kredi Kayıt Bürosu bazında kredi kullanan müşterilerden oluşmaktadır. Veri Seti-2 için; 2020 Ocak ayı tahminlerinin yapılabilmesi için kullanılan tüm veride 133,200 kayıt bulunmaktadır. Bu kayıtların 33,300 adedi kredi başvurusu yapılan bankanın kayıtlarına göre kredi kullanan müşterilerden oluşmaktadır. 2020 Şubat ayı tahminlerinin yapılabilmesi için kullanılan tüm veride 140,372 kayıt bulunmaktadır. Bu kayıtların 35,093 adedi kredi başvurusu yapılan bankanın kayıtlarına göre kredi

kullanan müşterilerden oluşmaktadır. Veri Seti-3 için; farklı büyüklükte (0.342 MB, 0.742 MB, 1.48 MB, 10.4 MB, 104 MB, 509 MB) eğitim veri setleri kullanılarak, büyük veri platformu üzerinde makine öğrenmesi algoritmalarının ölçeklenebilirlik performansları irdelenmiştir.

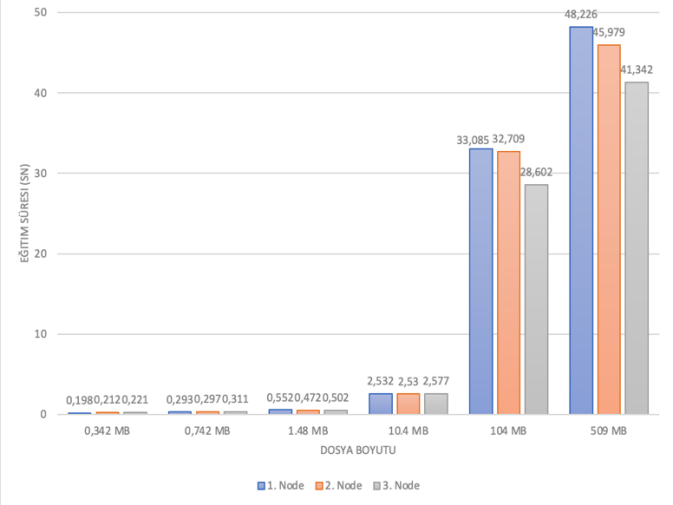
Gerçekleştirilen testlerde, önerilen veri işleme sürecinin tahmin başarı performansı incelenmiştir. Yapılan testlerde kullanılan farklı bağımlı değişkenler; 1) Kredi Kayıt Bürosu verileri üzerinden elde edilmiş bağımlı değişken ve 2) Kredi başvurusu yapılan bankanın verilerinden elde edilmiş bağımlı değişkene dayalı öğrenme modelleri oluşturulmuştur. Bu öğrenme modellerinin tahmin başarısı iki farklı veri setine ait test verileri üzerinde irdelenmiştir. Her iki öğrenme modeli kullanılarak, hem KKB test verileri, hem de Mevcut Banka test verilerinde tahminleme çalışması yapılmıştır. Sonuçlar; Bankadan elde edilen veriler üzerinde oluşturulan modellerin, kredi kullanımı tahmininde daha başarılı tahminler yapabildiğini göstermektedir. Farklı aylardan seçilmiş test verileri üzerinde oluşturulan her iki makine öğrenmesi modeli de benzer sonuçlar üretmiştir. Elde ettiğimiz bu bulgularla, bu bildirinin "Araştırma Soruları" bölümündeki 3. Araştırma sorusunu yanıtlamış oluyoruz.

TABLO I. DOĞRULUK TESTİ SONUÇLARI

Doğruluk (Accuracy) Metriği Sonuçları	KKB bilgileri eğitilen model – KKB Test Verisi	KKB bilgileri eğitilen model – Banka Test Verisi	Banka bilgileri ile eğitilen model – KKB Test Verisi	Banka bilgileri ile eğitilen model – Banka Test Verisi
Şubat Ayı Test Verisi İçin Tahmin Başarısı	% 75	% 74	% 86	% 87
Mart Ayı Test Verisi İçin Tahmin Başarısı	% 76	% 76	% 88	% 89

Ölçeklenebilirlik testinin gerçekleştirilmesi için, Apache Spark büyük veri işleme platformu üzerinde XGBoost gözetimli öğrenme algoritması koşturulmuştur. Şekil 3' te görüldüğü üzere, küçük ve büyük ölçekli eğitim verileri kullanılarak 1, 2 ve 3 node Spark kümeleri üzerinde, makine öğrenmesi algoritmasının eğitim süresi performansı ölçülmüştür.

Tablo 2 verilen sonuçları özetlemektedir. Küçük veri ölçeğinde (~1MB), Spark platformu üzerinde XGBoost gözetimli öğrenme algoritması tek node üzerinde daha başarılı sonuçlar vermektedir. Tek node'dan çok node' lu spark küme yapısına giderken, performans düşüklüğü gözlemlenmektedir. Bunun ana nedeni, çok-node Spark kurulumlarındaki node' lar arasındaki iletişim maliyetidir. Büyük ölçekli veriler için ise, tek-node' dan çok-node Spark kümesi kurulumuna giderken, performansın iyileştiği gözlemlenmektedir. Büyük ölçekli eğitim verileri (~500MB) üzerinde, Spark platformunun sağlamış olduğu veri paralelizmi ve iteratif eşle-indirge programlama modeline dayalı altyapısı performans artışı sağlamaktadır.



Şekil 3: Ölçeklenebilirlik Testi Sonuçları

TABLO II. BÜYÜK VERİ İŞLEME PLATFORMU ÜZERİNDE XGBOOST ÖLÇEKLENEBİLİRLİK PERFORMANSI

	Training XGBoost Ölçelebilirlik Performansı 1 node -> 2 node	Training XGBoost Ölçelebilirkik Performansı 1 node 3-> node
Küçük Ölçekte Eğitim Verisi	-%1,36	-%6,14
Büyük Ölçekte Eğitim Verisi	+%4,65	+%14,27

VI. SONUÇLAR VE GELECEKTEKİ ÇALIŞMALAR

Bu araştırma kapsamında, bankacılık sektöründe karşılaşılan kredi kullanımı eğilimi tahmini problemini çözmek amacıyla bir yazılım mimarisi önerisi yapılmaktadır. Bankacılık müşterileri kredi davranış verilerinin sürekli artan nitelikte bir veri olması ve müşteri sayılarının milyonlar mertebesinde olması nedenleriyle, önerilen yazılım mimarisi; büyük veri işleme ve analizi teknolojilerini kullanarak, bankacılık kredi davranışları verileri üzerinde, bir veri analizi iş sürecinin gerçekleştirilebilmesine olanak sağlamaktadır.

Önerisi yapılan yazılım mimarisinin prototip uygulaması geliştirilmiştir. Bu kapsamda, kredi kullanımı davranışını en iyi tanımlayan öznitelikler belirlenmiş ve kategorize edilmiştir. Bu

araştırma kapsamında önerilen kredi kullanımı davranışı öznitelikleri Kredi Kayıt Bürosu verilerinden elde edilen öznitelikler ve Banka verilerinden elde edilen öznitelikler olarak iki ana başlık altında kategorize edilebilir. Belirlenen öznitelik vektörlerinin değerleri her müşteri için hesaplanmış ve tüm müşteri verilerini temsil eden kredi davranışı matrisi verisi oluşturulmuştur. Geliştirilen prototip uygulama kapsamında, veri ön işleme aşamaları uygulanarak, verinin makine öğrenmesine hazır hale gelmesi sağlanmıştır. Kredi davranış matrisi üzerinde çalıştırılan gözetimli makine öğrenmesi yöntemiyle modeller oluşturulmuş ve kredi kullanımı eğiliminin tahmininde başarısı irdelenmiştir. Elde edilen sonuçlar, Banka verilerinden çıkartılan özniteliklerin, kredi kullanım eğilimini tahmin etmede daha başarılı sonuçlar ürettiğini göstermektedir. Elde ettiğimiz sonuçlar, bankacılık sektöründe kredi kullanım eğilimi tahminlerinde, artan müşteri sayılarından dolayı oluşan büyük ölçekli verileri üzerinde çok-node Spark kümeleri üzerinde makine öğrenmesi algoritmalarının kullanımının performans ve ölçeklenebilirlik metrikleri açılarından daha başarılı sonuçlar elde edilebileceğini göstermektedir.

Bu çalışmanın gelecek adımlarında kredi eğilimi tahmin modeli için öznitelik seti zenginleştirilecektir. Bu kapsamda müşteri memnuniyeti ve şikayetleri ile ilgili veriler, banka servis kanalları ile ilgili kullanım sıklığı, çeşitliliği, dijital yatkınlığı ve zaman içindeki değişimi gibi öznitelikler eklenerek modellerin performansı iyileştirilecektir. Kredi kullanma eğilimi için tercih edilen sınıflandırma modeline alternatif veya destek olarak müşterilerin kolektif kredi kullanma davranışları incelenecek ve öneri sistemine dahil edilecektir.

TEŞEKKÜR

Bu araştırma kapsamında veri seti ve hesaplama ortamı sağlayarak bu çalışmanın yapılmasına olanak sağlayan Fibabanka' ya teşekkür ediyoruz.

KAYNAKLAR

[1] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, 2018, pp. 1-4, doi: 10.1109/ICACCAF.2018.8776802.

[2] Yoon, J. S., & Kwon, Y. S. (2010). A practical approach to bankruptcy prediction for small businesses: Substituting the unavailable financial data for credit card sales information. Expert systems with Applications, 37(5), 3624-3629.

[3] Sayed, H., Abdel-Fattah, M. A., & Kholief, S. (2018). Predicting potential banking customer churn using apache spark ML and MLlib packages: a comparative study. IJACSA) International Journal of Advanced Computer Science and Applications, 9, 674-677.

[4] Hamid, A. J., & Ahmed, T. M. (2016). Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal (MLAIJ) Vol, 3(1).

[5] Baeth, M.J., Aktas, M.S., (2019). Detecting misinformation in social networks using provenance data, Concurrency and Computation-Practice & Experience, Vol: 31, Issue:3.

[6] Riveni, M. et al. (2019). Application of provenance in social computing: A case study, Concurrency and Computation-Practice & Experience, Vol: 31, Issue:3.

[7] Tas, Y., Baeth M.J., Aktas, M.S., (2016) An Approach to Standalone Provenance Systems for Big Provenance Data, The International Conference on Semantics, Knowledge and Grids on Big Data (SKG-16).

[8] Aktas, M.S., (2018) Hybrid cloud computing monitoring software architecture, Concurrency and Computation: Practice and Experience, Vol: 30, Issue:21.

[9] Aktas, M.S. et al. (2004). A web based conversational case-based recommender system for ontology aided metadata discovery, The 5th IEEE/ACM International Workshop on Grid Computing, pp:69-75.

[10] Aktas, M.S. et al, (2007). Fault tolerant high-performance Information Services for dynamic collections of Grid and Web services, Future Generation Computer Systems, Vol:23, Issue: 3.

[11] Pierce, M.E. et al. (2008). The QuakeSim project: Web services for managing geophysical data and applications, Pure and Applied Geophysics, Vol:165, Issue: 3-4, pp. 635-651.

[12] Aydin, G. et al. (2005). SERVOGrid complexity computational environments (CCE) integrated performance analysis, The 6th IEEE/ACM International Workshop on Grid Computing.

[13] Jin, Xin & Han, Jiawei. (2017). K-Means Clustering. 10.1007/978-1-4899-7687-1_431.

[14] M. Villamizar et al., "Infrastructure Cost Comparison of Running Web Applications in the Cloud Using AWS Lambda and Monolithic and Microservice Architectures," 2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), Cartagena, 2016, pp. 179-182, doi: 10.1109/CCGrid.2016.37.

[15] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[16] Vagrant Web Site available at: https://www.vagrantup.com Access Date: 30/08/2020.