Statistik maşın öyrənməsindən istifadə edərək insan resurslarının məlumat analitikası hissə 2 – kim gedəcək?

written by Hikmat Abdulazizov Hikmət Əbdüləzizov
Hissə – 2 – Kim gedəcək?

Birinci hissədə[i] bir müəssisədə müəyyən bir işçinin yüksəlmə ehtimalını və ya etiketini proqnozlaşdırdıq. Bu ikinci məqalədə işçinin işdən çıxma ehtimalı və ya etiketi ölçülür. Bu, bütün dünyadakı təşkilatlar üçün həlledici bir tədbirdir, çünki işçilərin işdən çıxmasının maddi və görünən qiymətləndirilməsi bir təşkilata çox vaxt və pul qazandıracaqdır.

Mövcud bir işçinin işdən çıxma ehtimalını bilməklə yanaşı, yeni bir işçinin qurumu tərk edib etməyəcəyinə də yaxşı baxa bilərik. Azərbaycan şirkətləri yaxın və ya orta gələcəkdə dataya əsaslanan strategiyalar tətbiq etməkdən uzaq olsalar da, bu sənəddə tətbiq olunan metodlar sağlam insan resursları (İR) məlumatlarına sahib olan hər bir təşkilat üçün son dərəcə faydalı olacaqdır.

Məsələn, Azərbaycandakı bir bankın bir kredit mütəxəssisi başqa bir banka keçərsə, çox güman ki, yığılmış kredit portfelini yeni bir banka aparacaq. Bu, böyük bir riskdir və konkret bir kredit işçisinin ayrılma ehtimalını bilmək banka ilkin tədbirlər görməyə imkan verir. Eyni nümunə digər vəziyyətlərdə və iş yerlərində də tətbiq oluna bilər.

Bu yazıda tətbiq olunan metodologiya digər ölkələrdə olduğu kimi, Azərbaycanda da şirkətlər üçün uyğun və dəyərlidir. Onun

İR-də tətbiqi iş şəraitinin inkişafına kömək edəcək. Bu metodologiyanın dəyəri Azərbaycanın əmək bazarı üçün daha da böyükdür, çünki iş şəraiti zəifdir və işçilərin sədaqəti aşağıdır.

Yanaşmamızı vurğulamaq üçün xüsusiyyətlərin mənası barədə fərziyyə etmirik. Bu, geniş işçi məlumatlarına malik olmayan Azərbaycan təşkilatları üçün də faydalıdır. Məlumatların çoxluğu proqnozları daha dəqiqləşdirsə də, yanaşmamız mövcud olan xüsusiyyətlərlə işdən çıxma ehtimalını qiymətləndirməkdir. Beləliklə, metodikamız istənilən təşkilat üçün universal olur. Birinci hissə ilə oxşar olaraq, cəzalandırılmış logistik reqressiyalar (Ridge, Lasso) kimi statistik öyrənmə vasitələrindən və ya KNN (k-nearest neighbors [k-yə ən yaxın qonşu]) və SVM (Dəstək Vektor Maşını) kimi maşın öyrənmə metodlarından istifadə edirik.

Data və Metodologiya

Təxminən 6 ilkin xüsusiyyətə malik 12 mindən çox müşahidəmiz var. Cavab dəyişən bir işçinin şirkətdən çıxıb çıxmadığını göstərir. Uyğun davamlı olanlardan kvadratlar və ya kublar alaraq başlanğıclardan daha çox xüsusiyyət yaradıldı, lakin proqnozlaşdırma gücündə bir qazanc olmadı.

Üstəlik, dəyişənlərin standartlaşdırılması və normallaşdırılması da sınaqdan keçirildi, lakin proqnozlaşdırma gücündə bir qazanc əldə edilmədi. Nəticə olaraq, dəyişkənlər ilkin səviyyələrində istifadə edilmişdir. Normallaşma, əks halda işləmədiyi üçün yalnız KNN qiymətləndiricisi üçün istifadə edilmişdir.

Məlumatların təxminən 17% -ində 1-ə bərabər bir çıxış etiketi var. Verilənlər bazasını iki hissəyə böldük: modellərimizi öyrətmək üçün 70%, onları yoxlamaq üçün 30%-ini istifadə etdik. Data hər iki hissənin eyni müvəffəqiyyət dərəcəsini bölüşəcəyi şəkildə avtomatik olaraq bölünür (burada çıxış etiketi, çıxan işçi).

11 cərimələnmiş büzülmə logistik reqressiyasından başlayaraq, 11 model- 0.1 aralıqlarla 0-dan 1-ə qədər alfa parametrini tənzimləyərək əldə edildi. Nə vaxt ki, alfa sıfıra bərabərdir, reqressiya *Ridge* adlanır. 1-ə bərabər olduqda *LASSO* adlanır. Aradakı hər hansı bir dəyərə, xüsusən, alfa 0.5-ə bərabər olduqda *elastik-net* deyilir. Əlavə tədqiqatlar daha çox hesablama gücü və vaxt olan təqdirdə xeyli daha çox alfa dəyərindən istifadə edilə bilər.

Bu modellər təlim dəstində hazırlanır və müqayisə üçün AUROC[ii] (bundan sonra sadəcə AUC) əldə edirlər. Büzülmə modelləri üçün ənənəvi olaraq, bütün xüsusiyyətləri daxil edirik və modelin hansının saxlanacağını və hansının sıfıra doğru büzüləcəyini qərar verməsinə icazə veririk (0 olduqda əhəmiyyətsiz olur). Alfa parametrinin sıfıra ayarlandığı Ridge reqressiyasının bütün xüsusiyyətləri qoruduğunu vurğulamalıyıq.

Statistik Öyrənmə üçün Nəticələr

Əvvəlcə, cəzalandırılan logistik reqressiyaların nəticələrini təqdim edəcəyik. Bu modellər üçün təlim dəsti üzərində AUC müqayisə ediləcəkdir. Parsimonluq xatirinə və vaxta qənaət etmək üçün ənənəvi cəzasız logistik reqressiyanın üstündən keçirik, çünki daha müasir metodları zatən istifadə etməli olacağıq.

Cədvəl 1. Cəzalandırılmış Logistik Reqressiya üçün Nəticələr

alfa	*AUC*
0	0.852727678189757
0.1	0.851843491786532
0.2	0.851098038353651
0.3	0.85130858403536
0.4	0.851295735120991
0.5	0.850227322390076

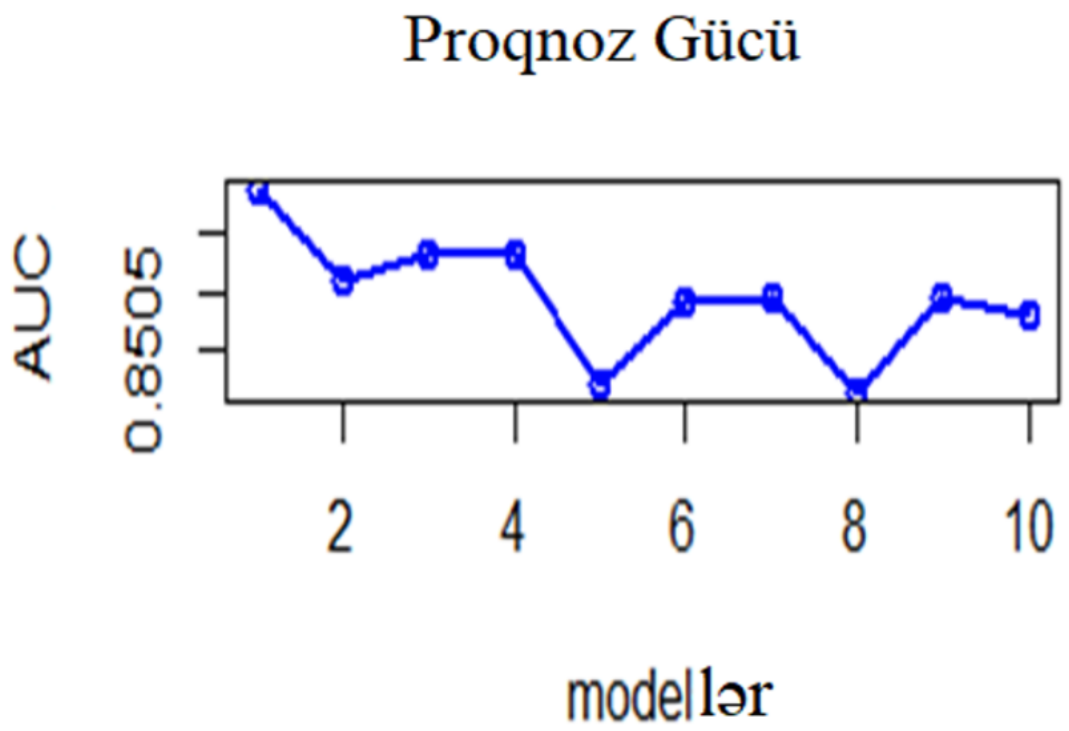
0.6	0.850910558042536
0.7	0.850941817777063
0.8	0.850163845304546
0.9	0.850946921836748
1	0.850797344962633

Yuxarıdakı cədvəldə AUC və müvafiq alfa parametrlərinə əsasən üç ən yaxşı modeli qırmızı rənglə qeyd edirik. Bunlardan ikisi elastik-net reqressiyası kimi, biri də Ridge kimi etiketlənə bilər (Ridge yəni alfa = 0). Bunları görünməmiş test məlumatları üzərində ən yüksək proqnozlaşdırma gücünə malik modeli seçmək üçün istifadə edəcəyik.

Ən yaxşı model alfa parametri 0 olan Ridge modeli kimi görünür. Bu model bütün xüsusiyyətləri modeldə saxlayır. Bu məqalənin ilk hissəsində 20 xüsusiyyətdən təxminən 70% -i saxlanıldı və Ridge modeli ən yaxşı model hesab edilmədi. Fərq məlumatın mövcudluğu ilə bağlı ola bilər. Hər bir model xüsusiyyətlərin kombinasiyasını optimallaşdırmağa çalışır, lakin məhdud sayda xüsusiyyətə (6) sahib olan modellər heç bir məlumatı kənara buraxa bilmir.

Rəqəmlər bir-birinə yaxın görünsə də, aşağıdakı qrafik proqnozlaşdırma gücünün nə qədər dəyişkən ola biləcəyini göstərir. Büzülmə modellərinin yüksək dərəcədə korrelyasiya olunmuş dəyişənlər və kvazi-ayırma problemləri kimi qeyri-sağlam halları aşdığı və daha aşağı proqnoz səhvinə səbəb olduğu məlumdur. Qeyd edək ki, Ridge eyni modeldə çox əlaqəli dəyişənləri saxlaya bilər. Cüzi fərqlə demək olar ki, eyni dəyişənlərə sahib olsaq da, model heç bir məlumatı tərk etməyəcəkdir.

Qrafik 1. Təlim Dəsti Modellərinin AUC-ları



Tədqiqatın test dəsti aşağıdakı dəyərləri əhatə edir: 0 və 1. Bununla birlikdə modellər sıfır ilə bir arasında ehtimalları proqnozlaşdırırlar. Sinifləri proqnozlaşdıran bir funksiyadan istifadə edilə bilər, ancaq bu, proqramı ehtimallar üçün kəsmə dəyərini 0.5 kimi seçməyə məcbur edəcəkdir ki, bu da həddindən artıq sadələşdirmədir. Kəsmə dəyəri 0.5, 0.5-dən yuxarı dəyəri 1 seçmək demək, əks halda isə 0 olaraq proqnozlaşdırmaq deməkdir. Əvvəlcədən normal görünsə də, bu cür kəsilmənin əsaslandırılması üçün hər bir məlumatı yoxlamalıyıq. Burada kəsmə dəyərlərini hansı meyarlara görə əldə etdiyimizi təqdim etməli və sonra modelləri müqayisə etməliyik. Test dəsti üçün də AUC istifadə edə bilərik, lakin modelləri SVM kimi maşın öyrənmə həmkarları ilə müqayisə etmək üçün daha universal meyarlara ehtiyacımız var.

Üç anlayışın müəyyənləşdirilməsi lazımdır: dəqiqlik, həssaslıq və spesifiklik. Dəqiqlik test datada model tərəfindən düzgün şəkildə proqnozlaşdırılmış hissədir. Dəqiqlik 90%-dirsə, sıfırların və birlərin 90%-i düzgün proqnozlaşdırılıb deməkdir. Həssaslıq bir modelin həqiqi pozitivlərinin (HP) neçə faizinin düzgün proqnozlaşdırıldığını ölçür. Əksinə,

spesifiklik isə bir modelin həqiqi neqativlərinin (HN) neçə faizinin düzgün proqnozlaşdırıldığını ölçür.

Üç modelin hər biri üçün kəsmə dəyərlərini seçmək üçün üç meyardan istifadə olunur. Birincisi, dəqiqliyi maksimum dərəcədə artıran kəsilmədir. İkincisi, həm həssaslığı, həm də spesifikliyi artırır, ROC əyrisi qrafikinin yuxarı sol küncü ilə əyrinin özü arasındakı məsafəni minimuma endirir. Son meyar, özümüzün təyin etdiyi xərc funksiyasını minimuma endirən kəsimdir. Xərc funksiyası yalan neqativ və yalan pozitivləri elə cəmləşdirir ki, yalan neqativlər yalan pozitivlərdən iki qat daha xərcli olsun. Bu isə bizim konteksdə işdən ayrılan işçini təşkilatda qalacaq kimi etiketləməyin, işləyən birisinə işdən çıxma etiketini yazmaqdan daha xərcli olduğu deməkdir. Seçilmiş üç modelin nəticələrini axırda bir modelə gəlib çıxmaq üçün nəzərdən keçirək.

Cədvəl 2. Modellər və Onların Test Datasında Performansları

Model 1 alfa = 0.0					
Dəqiqlik Kəsimi = 0.295		Xərc Minimum Kəsimi = 0.18		Minimum Məsafə Kəsimi = 0.182	
Dəqiqlik	84%	Dəqiqlik	80%	Dəqiqlik	81%
Həssaslıq	22%	Həssaslıq	91%	Həssaslıq	88%
Spesifiklik	97%	Spesifiklik	78%	Spesifiklik	80%

Model 2 alfa = 0.1					
Dəqiqlik Kəsimi = 0.396		Xərc Minimum Kəsimi = 0.169		Minimum Məsafə Kəsimi = 0.185	
Dəqiqlik	84%	Dəqiqlik	77%	Dəqiqlik	80%
Həssaslıq	22%	Həssaslıq	91%	Həssaslıq	83%
Spesifiklik	96%	Spesifiklik	74%	Spesifiklik	79%

Model 4 alfa = 0.3					
Dəqiqlik Kəsimi = 0.849		Xərc Minimum Kəsimi = 0.157		Minimum Məsafə Kəsimi = 0.181	
Dəqiqlik	83%	Dəqiqlik	75%	Dəqiqlik	79%
Həssaslıq	0%	Həssaslıq	91%	Həssaslıq	82%
Spesifiklik	100%	Spesifiklik	72%	Spesifiklik	79%

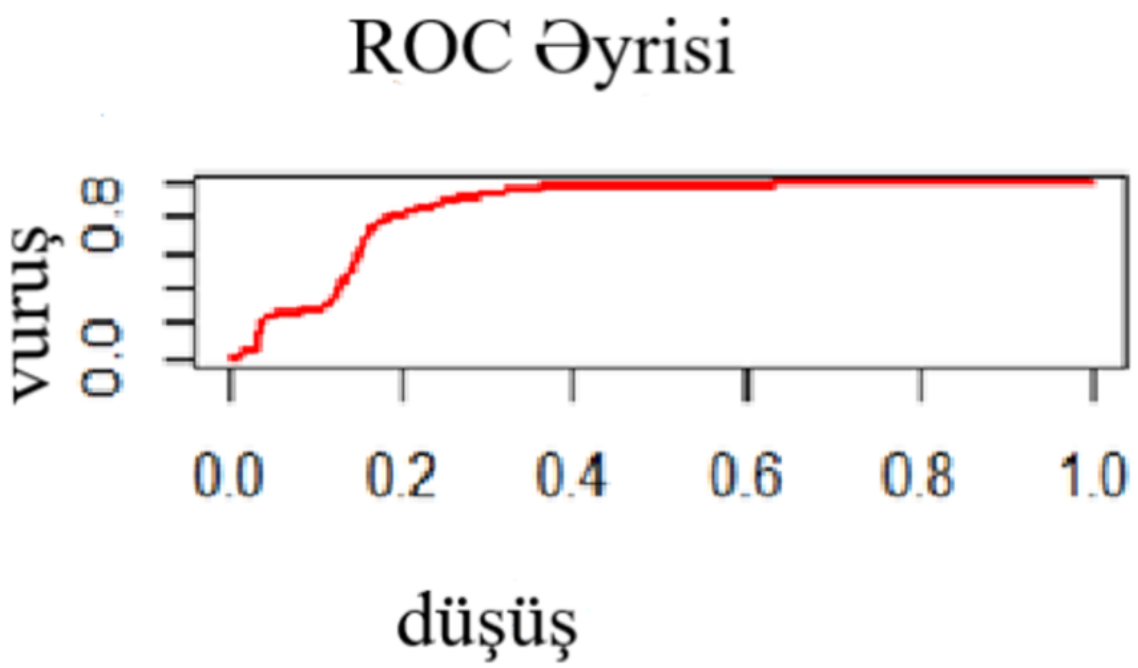
Bir neçə müşahidədən bəhs etmək lazımdır. Birincisi, dəqiqlik kəsiklərinin gözlənildiyi kimi bizə ən yüksək dəqiqliyi

verdiyini müşahidə edirik. Digər tərəfdən, dəqiqliyi və spesifikliyi çox az itirərək xərcləri minimuma endirən kəsikdən istifadə ediriksə, həssaslıqdan dörd dəfədən çox qazanırıq.

Üçüncü model, təəccüblüdür ki, 0 həssaslığı və 100% spesifikliyi ilə 85% dəqiqlik kəsiyini qaytarır. Biz bunun tam əksini istəyirdik. Yəni bu model və bu kəsilmə ilə heç kimin işdən ayrılacağını düzgün proqnozlaşdıra bilmərik. Sağdakı sütunlarda xərcləri minimuma endirən kəsikləri olanlarla ortadakılardan fərqli bir şey müşahidə etmirik.

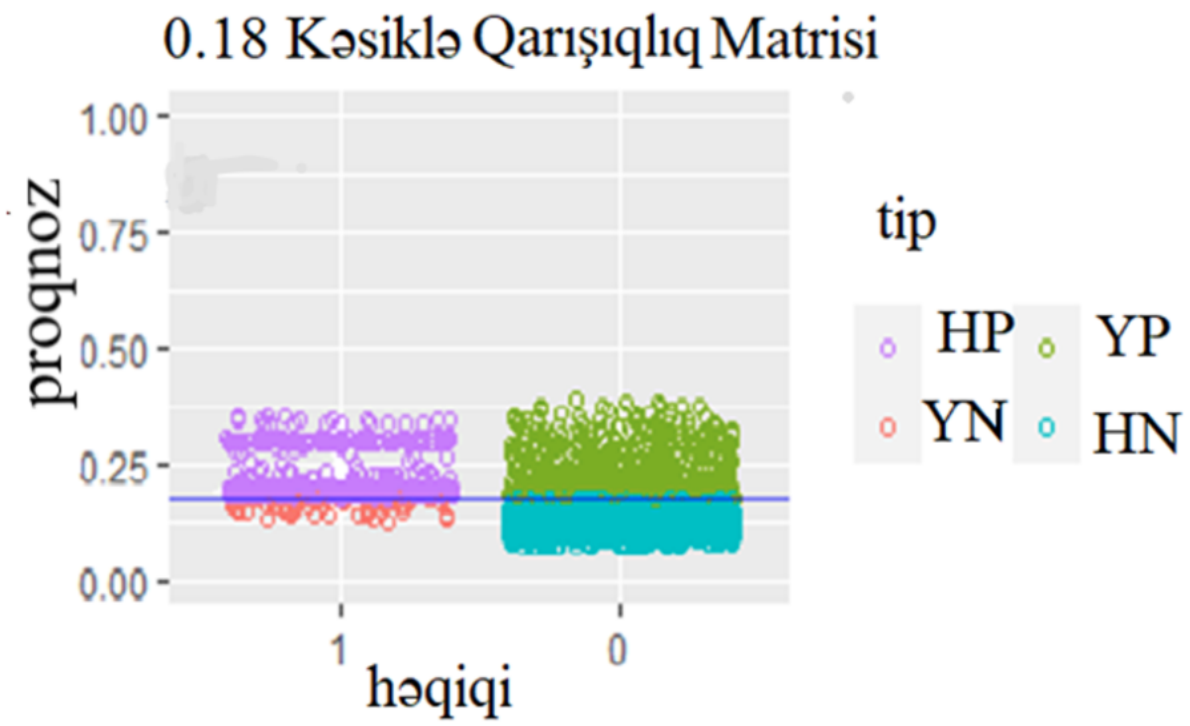
Əslində, xərci minimuma endirən və məsafəni minimuma endirən kəsmə dəyərləri bir-birinə çox yaxındır. Özümüz qurduğumuz xərclər funksiyamızın olduqca faydalı olduğuna əminlik yaranır. Ən yüksək həssaslığa və spesifikliyə malik olan bir model üçün ən az mümkün dəqiqlik itkisi ilə gedərdik. Bu, ilk modelimiz olacaq. Bu model bütün xüsusiyyətləri qoruyan 0 alfa parametrli və xərci minimuma endirən kəsmə dəyəri 0.18 olan Ridge olur. Qeyd edək ki, kəsmə dəyərimiz 0.5-in yarısından azdır, hansı ki, proqram avtomatik olaraq seçərdi. Ən yaxşı modelimizə (model 1, Ridge) daha yaxından baxaq.

Qrafik 2. Modelin ROC əyrisi



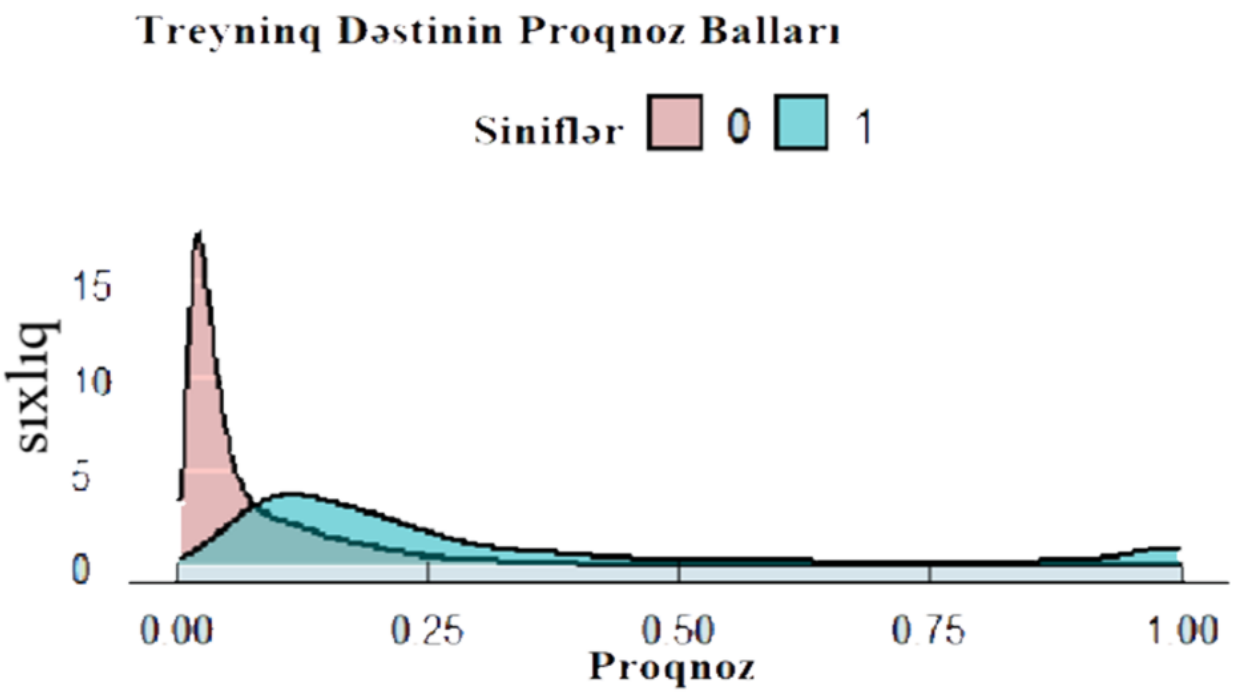
Yuxarıdakı qrafikdə vuruş dərəcəsi həssaslıq və düşüş yanlış pozitiv nisbətdir ki, bu da bir minus spesifiklikdir. Beləliklə, yuxarı sol küncdən döngəyə qədər olan minimal məsafə həssaslığı və spesifikliyi maksimallaşdırır, çünki ən yuxarı sol künc ideal nöqtədir.

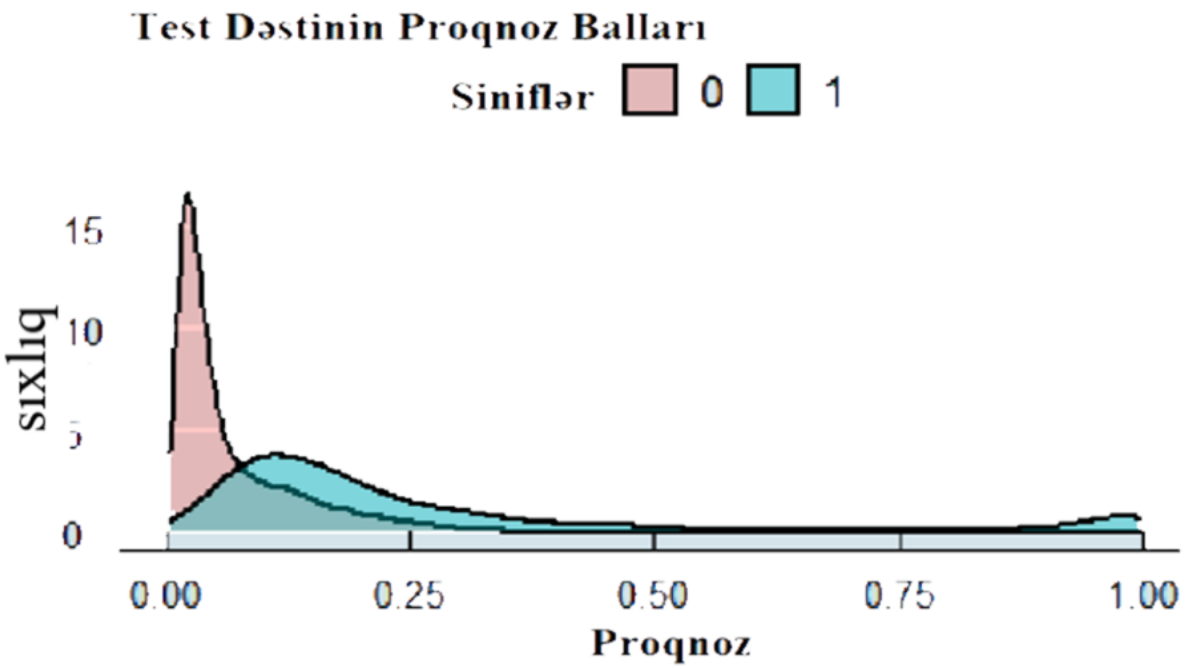
Qrafik 3. Qarışıqlıq Matrisi



Yuxarıda hər bir kateqoriya üçün uyğun nöqtələrin qarışıqlıq matrisi göstərilir (Həqiqi Pozitiv, Həqiqi Neqativ, Yanlış Pozitiv, Yanlış Neqativ). HP-lərin sayı olduqca yüksəkdir, çünki ən yüksək həssaslıqla (vuruş dərəcəsi) kəsilən hissəni seçdik.

Qrafik 4. Proqnozlaşdırılmış Təlim və Test Ballarının Örtüşməsi





Yuxarıdakı qrafiklər sinif başına proqnozların sıxlığını göstərir. Verilən siniflərə uyğun olduqca üst-üstə düşdüyünü müşahidə etdiyimiz üçün diqqət tələb olunur. Üstəlik, qrafiklər xərci minimuma endirən kəsmə dəyəri 0.18-in kifayət qədər məqbul olduğunu göstərir.

Maşın Öyrənməsi nəticələri

Əlavə təsnifatlar üçün bir neçə əsas maşın öyrənmə vasitəsi

istifadə olunur. Əvvəlkilərdən fərqli olaraq, bu metodlar proqnozlaşdırılan nəticələr kimi siniflər (0 və 1) verir. Beləliklə, heç bir kəsilmə problemi və ya başqa bir şəkildə qərar azadlığı yoxdur.

Bəzi metodlar digərlərindən daha çox vaxt və hesablama gücü tələb edir. Bu metodları bütün xüsusiyyətlərdə işlətmək nəticə əldə etmək üçün saatlar çəkə bilər. Beləliklə, xüsusiyyət seçimi üçün əvvəllər işləyən modellərdən LASSO operatorunu (alfa 1-ə bərabərdir) istifadə edirik. Aşağıda bəzi maşın öyrənmə üsullarının nəticələri verilmişdir.

Cədvəl 3. Maşın Öyrənməsi Nəticələri

Dəstək Vektor Maşını	
Dəqiqlik	97%
Həssaslıq	92%
Spesifiklik	98%

K-Ən Yaxın Qonşu	
Dəqiqlik	98%
Həssaslıq	91%
Spesifiklik	98%

İxtiyari Meşə	
Dəqiqlik	93%
Həssaslıq	87%
Spesifiklik	97%

Sadəlövh Bayes	
Dəqiqlik	89%
Həssaslıq	93%
Spesifiklik	95%

Qərar Ağacı	
Dəqiqlik	96%
Həssaslıq	89%
Spesifiklik	94%

Yuxarıdakı nəticələr üçün bir neçə məqam var. Hər şeydən əvvəl KNN qiymətləndiricisi ilə dəqiqliyin ən yüksək olduğunu, SVM-nin ikinci ən yaxşı olduğunu müşahidə edirik. Sözügedən modellərin dəqiqliyi əvvəlki büzüşən logistik modellərindən daha yüksəkdir. Bu, əvvəllər test edilmiş İR məlumatlarından çox fərqli bir nəticədir.

Həssaslıqlara baxsaq, SVM yuxarıda təqdim olunan maşın öyrənmə metodlarını və əvvəlki büzülmə modellərini üstələyir. Dəqiqlikdən yalnız 1% itirərək, işçilərin çıxış proqnozu üçün SVM qiymətləndiricisini seçərik. Qeyd etmək vacibdir ki, hansını və nə üçün istifadə edəcəyimizi anlamaq üçün bu modelləri həmişə sınamalıyıq.

Nəticə və əlavə Müzakirə

Bu məlumatlar üçün ən yaxşı metodumuzun SVM (Dəstək Vektor Maşını) olduğunu gördük. Son məqalədə İR məlumatlarını da oxşar bir məqsəd üçün istifadə etdik və fərqli bir modellə sona çatdıq. Qeyd edək ki, hər iki halda seçim etmək üçün 20 model istifadə edilmişdir.

Ən azı bu sənəddəki kimi və ya mümkünsə, eyni modellərdən istifadə edilməsi tövsiyə olunur. Bundan əlavə, hər bir məlumat dəsti üçün ayrı bir yanaşma lazım ola bilər, lakin məqsəd ən yaxşı proqnozu ən aşağı səhvlə qaytarmaqdır.

Bu ikinci məqalənin bir uğuru da odur ki, yalnız statistik öyrənmə və ya maşın öyrənmə istifadə etməyin kifayət olmadığını bildik. Qarışdırmağı tövsiyə edirik və daha dəqiq nəticə əldə etmək üçün hər iki tip modelləri də sınayın. Yalnızca maşın öyrənmə metodlarının istifadəsi nəzəri cəhətdən qidalandırıcı və təkmil olan logistik reqressiya kimi başqa əhəmiyyətli statistik güclü vasitələrə məhəl qoymamaq deməkdir. Eyni şəkildə, yalnız statistik öyrənmə metodlarından istifadə etmək, SVM kimi müasir, güclü proqnozlaşdırıcı vasitələrə məhəl qoymamaq deməkdir. Qeyd etmək vacibdir ki, LASSO hər ikisi üçün sərhəddir və seçmə xüsusiyyətə malikdir.

[i] https://bakuresearchinstitute.org/az/hr-data-analytics-using-statistical-machine-learning/

[ii] AUROC – The area under the receiver operating characteristic – Alıcının işləmə xüsusiyyətinin altındakı sahə

Hissə-1

https://bakuresearchinstitute.org/en/hr-data-analytics-using-statistical-machine-learning/