Statistik maşın öyrənməsindən istifadə edərək insan resurslarının məlumat analitikası

written by Hikmat Abdulazizov Hikmət Əbdüləzizov

Hissə 1 – Vəzifədə Yüksəlmə

Günümüzdə İR(insan resursları) analitikası tədqiqatçılar və sənaye üçün cazibədar mövzu halına gəlib. İR ilə əlaqəli daha çox məlumata sahib olduğumuz üçün statistik metodlardan istifadə edərək bunların analizinə ehtiyac daha çox əhəmiyyət kəsb edib. İR rəhbərliyi əlaqəli məlumatları analiz edərək, qiymətləndirərək, daha səmərəli bir sistemə çevrilir. Burada birinci hissədə bir işçinin daha yüksək vəzifəyə yüksəlmə ehtimalını qiymətləndirmək üçün açıq mənbəli İR məlumatlarını istifadə edirik. İstifadə etdiyimiz metodlar *statistika-ekonometrika* və *maşın öyrənməsi*ndən qaynaqlanır. Bunlar arasında sərt bir sərhəd olmadığı üçün buna *statistik maşın öyrənməsi* adını veririk. Bizim yanaşmamız xüsusiyyətlərin mənalı şərhindən daha çox dəqiq proqnozlaşdırma ilə əlaqəli olduğu üçün maşın öyrənməyə daha yaxındır. Statistik öyrənmə hissəsi şərti və cəzalandırılmış logistik reqressiyalardan da istifadə etməyimizdən irəli gəlir. Sərhəd xətti metodu isə maşın öyrənmə metodlarına giriş məlumatları kimi istifadə etməyə həm proqnozlaşdırma, həm də xüsusiyyət seçimi üçün LASSO operatoru olacaqdır. İstifadə olunan metodlar hər hansı bir Azərbaycan məlumatı üçün də mükəmməl şəkildə təkrarlana bilər. Azərbaycan məlumatları üçün başqa bir istifadə də nepotizmin aşkarlanmasıdır. Bir işçinin vəzifəsini yüksəltmə ehtimalı yüksəkdirsə (və ya yüksəldiləcəyi təxmin edilirsə), ancaq vəzifəsini yüksəltməmişsə və əvəzinə vəzifəsini yüksəltmə ehtimalı daha az olan birisi vəzifəsini almışsa, bu aşkar bir nepotizm əlaməti ola bilər. Bu metodla regional və

cinsi diskriminasiya da aşkarlana bilər. Yenidən vurğulamaq üçün qeyd edək ki, xüsusiyyətlərin mənası ətrafında spekulyasiya etmək əvəzinə, tamamilə proqnozlaşdırma gücünə diqqət yetirəcəyik. İkinci hissədə müəyyən bir işçi üçün işdən ayrılma ehtimalını qiymətləndirmək üçün fərqli məlumatlar dəsti ilə işləyəcəyik.

Data və Metodologiya

20-yə yaxın ilkin xüsusiyyətə sahib 50 000-dən çox müşahidəmiz var. Cavab dəyişən bir işçinin yüksəldilib-yüksəldilməməsini göstərir. Başlanğıcda uyğun olanlardan kvadratlar və ya kublar götürərək daha çox xüsusiyyət yaratdıq, lakin daha çox proqnozlaşdırıcı güc qazana bilmədik. Üstəlik, dəyişənlərin standartlaşdırılması və normallaşdırılması da daha yaxşı qiymətləndirmələr əldə edilmədən sınanmışdır. Bu səbəbdən xüsusiyyətləri normal vəziyyətə gətirməyimiz lazım olan KNN (*k-nearest neighbors* [*k-yə ən yaxın qonşu*]) qiymətləndiricisi xaricində dəyişənləri sadə səviyyə formalarında istifadə edirik. Verilənlərin təxminən 8% -i 1-ə bərabər vəzifədə yüksəlmə etiketinə malikdir. Ümumiyyətlə, məlumatları treyninq dəsti, doğrulama dəsti və test dəstinə bölməliyik. Bununla birlikdə sadəliyi təmin etmək üçün məlumatları treyninq və test dəstinə bölürük. Bölünmə nisbəti [90%; 10%] ilə [70%; 30%] arasında uzanan bir döngü üzərində seçildi və [71%; 29%] ilə olan modelin dəqiqlik baxımından digərlərindən daha yaxşı bölündüyü aşkar edildi. Səliqə üçün bölünmə nisbətini [70%; 30%] seçdik. Bunu etməkdə çox əhəmiyyətsiz bir itki var.

Əvvəlcə sadə bir şərti logistik reqressiya və 11 cərimələnmiş büzülmə logistik reqressiyasından başlayırıq. Alfa parametrini 0.1 aralıqlarla 0-dan 1-ə uyğunlaşdıraraq 11 model əldə edirik. Alfa sıfıra bərabər olduqda reqressiya Ridge, 1 olduqda isə LASSO adlanır. Bu modellər arasındakılar isə elastik-net adlandırılacaq (xüsusilə alfa 0.5-ə bərabər olduqda). Alfa parametri ilə bağlı əlavə müzakirə bu məqalədə daha sonra aparılacaq. Bu modelləri treyninq dəstində həyata keçiririk və müqayisə üçün AUROC[1] (bundan sonra sadəcə AUC)

əldə edirik. Sadə şərti logistik reqressiya üçün p-dəyərinə əsasən statistik cəhətdən əhəmiyyətli xüsusiyyətləri seçirik. Büzülmə modelləri üçün bütün xüsusiyyətləri daxil edirik və modelin hansını saxlayacağını, hansının sıfıra enəcəyini qərar verməsinə icazə veririk (beləliklə, əhəmiyyətsizliyini sınamış oluruq). Alfa parametrinin sıfıra ayarlandığı Ridge reqressiyasının bütün xüsusiyyətləri saxladığını vurğulamalıyıq.

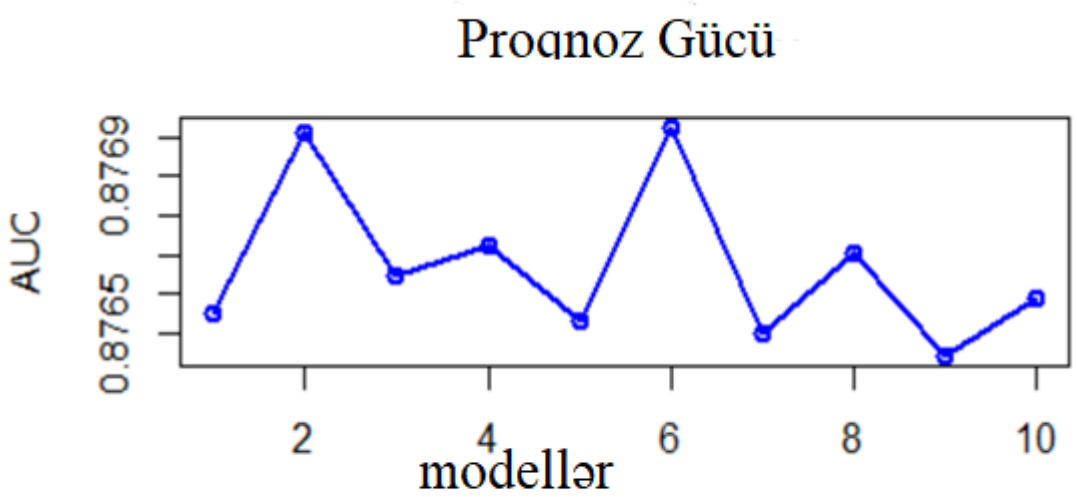
Statistik Öyrənmə üçün nəticələr

Əvvəlcə sadə ənənəvi logistik reqressiyanın nəticələrini təqdim edirik. Treyninq dəsti üzərində bu model üçün AUC 0.8735-dir. Bu bir şəkildə növbəti addımlarımız üçün bir meyardır. Bizdə 15 xüsusiyyət qalır və psevdo R^2 0.32 civarındadır, bu da bu tip modellər üçün yaxşı uyğunlaşma əlamətidir. Daha sonra, R proqramının glmnet paketini istifadə edərək büzülmə logistik reqressiyalarına uyğunlaşırıq. Daha əvvəl də deyildiyi kimi, 11 modellə nəticələnən 0.1 aralıqlarla 0-dan 1-ə qədər alfa parametrini seçirik.

Cədvəl 1. Cəzalandırılmış Logistik Reqressiya Üçün Nəticələr

alpha	*AUC*
0	0.8654773
0.1	0.8765511
0.2	0.8770106
0.3	0.8766488
0.4	0.8767241
0.5	0.8765327
0.6	0.8770255
0.7	0.8764979
0.8	0.8767012
0.9	0.8764419
1	0.8765889

Yuxarıdakı cədvəldə AUC və müvafiq alfa parametrlərinə görə ən yaxşı 3 modeli qeyd edirik. Üçü də elastik-net reqressiyası kimi etiketlənə bilər. Bunlardan görünməmiş test məlumatlarında ən yüksək proqnozlaşdırma gücünə sahib olanı seçmək üçün istifadə edəcəyik. Rəqəmlər bir-birinə yaxın görünsə də, aşağıdakı qrafik onun nə qədər dəyişkən ola biləcəyini göstərir. Nəzərə alın ki, yalnız alfa dəyəri 0 olan Ridge reqressiyası ənənəvi logistik reqressiyanı üstələyə bilir. Beləliklə, müasir daralma məntiqli reqressiyalardan istifadə etməli olduğumuzu bilirik. Üstəlik, büzülmə modellərinin xəstə halları (yüksək korrelyasiyalı dəyişənlər, mükəmməl ayrılma problemləri və s. kimi) sağaltdığı və daha aşağı proqnoz səhvinə səbəb olduğu məlumdur.



İndi bu modellər arasından seçim etmək çətin olan bir mərhələyə qədəm qoyuruq. Bildiyimiz kimi, test dəstimiz həqiqi sıfır və birlərə malikdir. Lakin modellər ehtimalları sıfır ilə bir arasında proqnozlaşdırır. Sinifləri 0 və ya 1 kimi konkret proqnozlaşdıran funksiyanı istifadə edə bilərik, amma proqram ehtimallar üçün kəsmə dəyərini 0.5 olaraq seçir, bu da etibarlı deyil və həddindən artıq sadələşdirmədir. Kəsmə dəyəri 0.5, 0.5-dən yuxarı dəyəri 1 seçmək demək, əks halda

isə sıfır olaraq proqnozlaşdırmaq deməkdir. Məntiqli səslənsə də, belə bir 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 sonrakılar ilə müqayisə etmək üçün daha universal meyarlara ehtiyacımız var. Yeri gəlmişkən, test dəsti üçün AUC alfa dəyəri 0.6 olan yeddinci modeldə ən yüksəkdir.

Üç anlayışı anlamalıyıq: dəqiqlik, həssaslıq və spesifiklik. Dəqiqlik model tərəfindən düzgün şəkildə çəkilmiş hissədir. Dəqiqlik 90% -dirsə, bu, sıfırların və birlərin 90% -nin düzgün proqnozlaşdırıldığı deməkdir. Həssaslıq düzgün proqnozlaşdırılan birlərin həqiqi müsbət nisbəti deməkdir. Spesifiklik isə sıfırlar üçün eyni şeydir.

Kəsmə seçimi üçün ümumilikdə üç meyarımız var (beləliklə, hər model üçün üç kəsik). Biri dəqiqliyi maksimum dərəcədə artıran kəsilmədir. İkincisi, həm həssaslığı, həm də spesifikliyi artıran kəsikdir, beləliklə, ROC əyrisi qrafikinin yuxarı sol küncü ilə əyrinin özü ilə arasındakı məsafəni minimuma endirir (göstəriləcək qrafiklə daha aydın olur). Başqası, özümüzün müəyyən yanaşmayla təyin etdiyimiz itki funksiyasını minimuma endirən kəsilmədir. Bu itki funksiyası yanlış mənfi və yanlış müsbət halları, yanlış neqativləri (birləri sıfır kimi proqnozlaşdıran) yanlış pozitivlərdən iki (və ya daha çox) dəfə baha başa gələcək şəkildə cəmləşdirir. Bizim vəziyyətimizdə isə bu o deməkdir ki, bizim üçün yüksələn işçini yüksəldilməmiş kimi etiketləmək yüksəldilməmiş işçini yüksəldilmiş kimi etiketləməkdən daha bahalıdır (və ya yaxşı işçini itirmək daha xərclidir). Bir model əldə etmək üçün seçilmiş üç modelin nəticələrini nəzərdən keçirək.

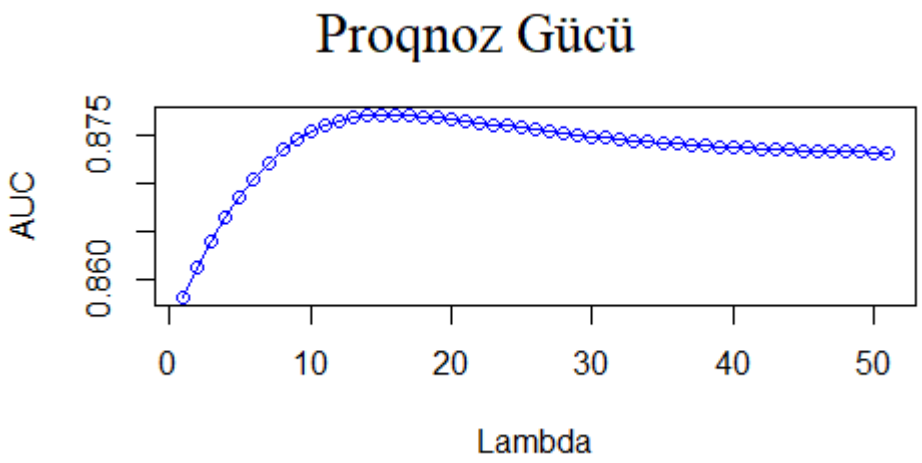
Model 3 alfa = 0.2					
Dəqiqlik Kəsimi	0.4095	İtki Minimumu Kəsimi	0.301	Minimum Məsafə Kəsimi	0.086
Dəqiqlik	0.9264	Dəqiqlik	0.9218	Dəqiqlik	0.74234
Həssaslıq	0.1984	Həssaslıq	0.2764	Həssaslıq	0.8299
Spesifiklik	0.9958	Spesifiklik	0.9833	Spesifiklik	0.7351

Model 5 alfa = 0.4					
Dəqiqlik Kəsimi	0.423	İtki Minimumu Kəsimi	0.303	Minimum Məsafə Kəsimi	0.088
Dəqiqlik	0.9268	Dəqiqlik	0.9229	Dəqiqlik	0.7464
Həssaslıq	0.1921	Həssaslıq	0.2724	Həssaslıq	0.8220
Spesifiklik	0.9968	Spesifiklik	0.9849	Spesifiklik	0.7391

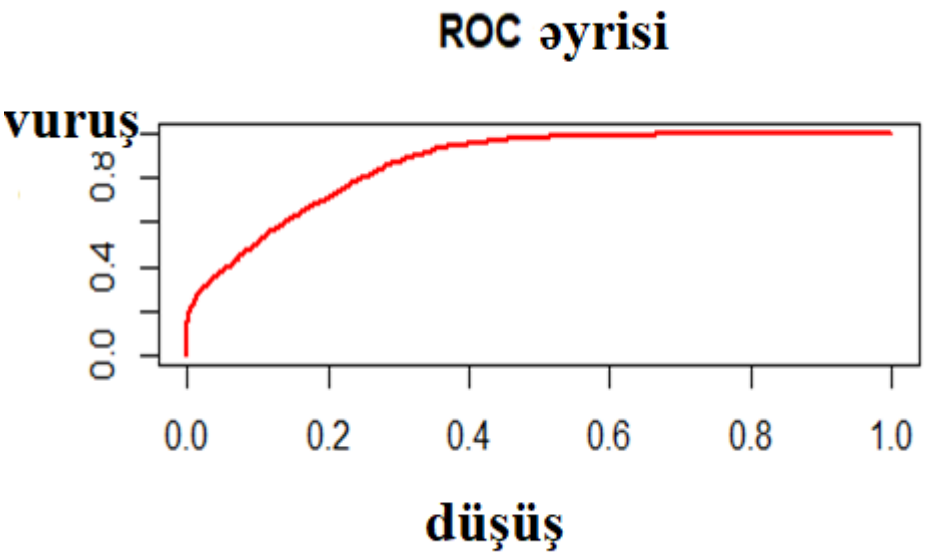
Model 7 alfa = 0.6					
Dəqiqlik Kəsimi	0.41	İtki Minimumu Kəsimi	0.306	Minimum Məsafə Kəsimi	0.081
Dəqiqlik	0.9271	Dəqiqlik	0.9234	Dəqiqlik	0.7327
Həssaslıq	0.2015	Həssaslıq	0.2756	Həssaslıq	0.8472
Spesifiklik	0.9962	Spesifiklik	0.9852	Spesifiklik	0.7217

Yuxarıdakı cədvəllərdə nəzər yetirməli bir neçə məqam var. 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 40% qazanırıq. Üçüncü sütunlarda dəqiqlik və spesifikliyə görə təxminən 20% itirərək həssaslıqda təxminən dörd dəfə qazandığımızı müşahidə edirik. Bizim üçün yüksəldiləcək işçiləri tapmaq (və bu işçiləri itirməmək) vacibdir və beləliklə ən yüksək həssaslıqla son modeli seçirik. Lakin son model üzərinə başqalarını seçməyimiz vəziyyətdən və ya rəhbərlik qərarından asılı ola bilər. Məsələn, kredit itkisi ehtiyatlarını qiymətləndirərkən minimum məsafəli kəsiyi seçsək, daha çox ehtiyat ayırmaq bahasına pis müştərilər üçün ən yüksək proqnoz gücünü qazanarıq. Burada isə bizim vəziyyətimizdə daha çox insanı yüksəltmək bahasına (bununla da daha aşağı kəsik) yüksələcəklər üçün proqnozlaşdırıcı güc qazanırıq. Hər bir halda bu modellər ehtiyac və yanaşmamızdan asılı olaraq bizə seçim azadlığı verir. Qeyd edək ki, kəsmə dəyərimiz sadəlövh 0.5-dən fərqlənir. İndiyə qədərki ən yaxşı modelimizə (model 7) daha yaxından baxaq.

Qrafik 2. Hər Lambda Parametri Üçün Model AUCları

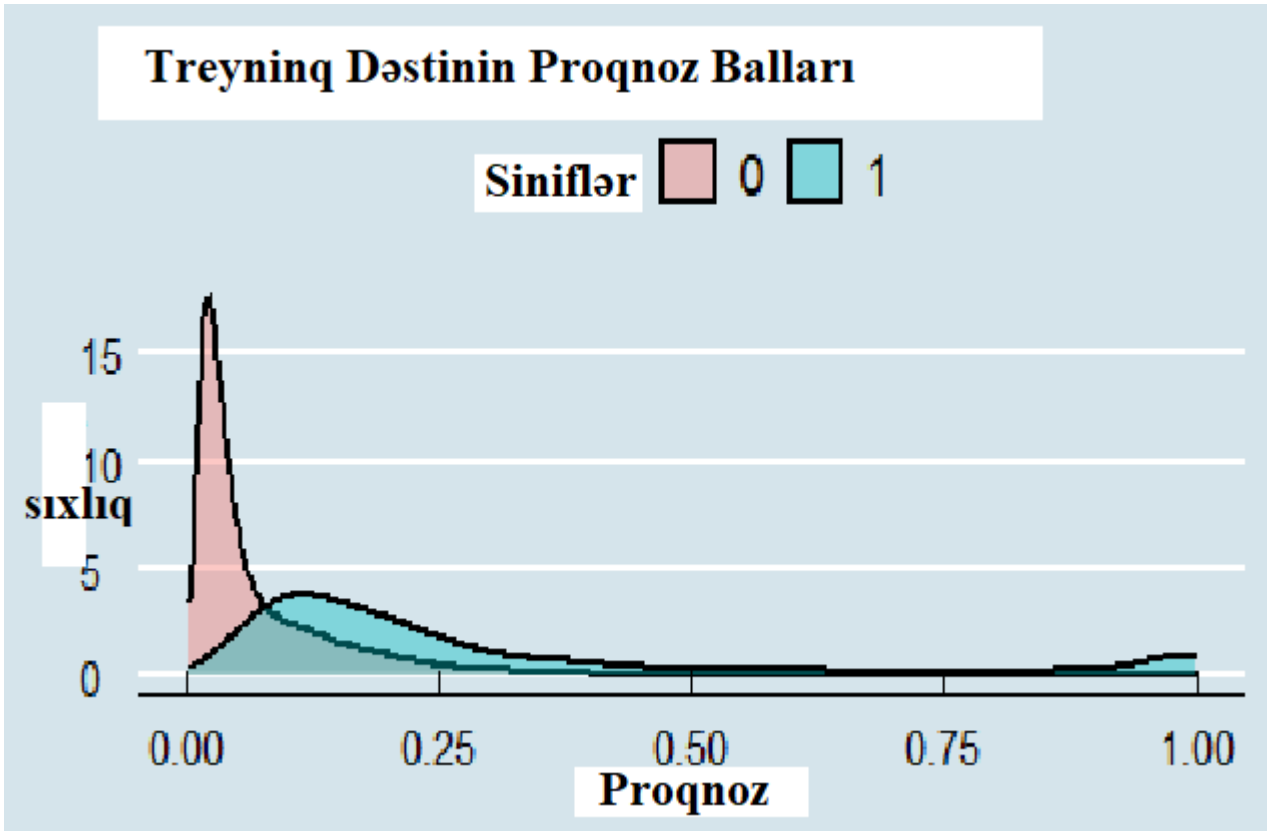


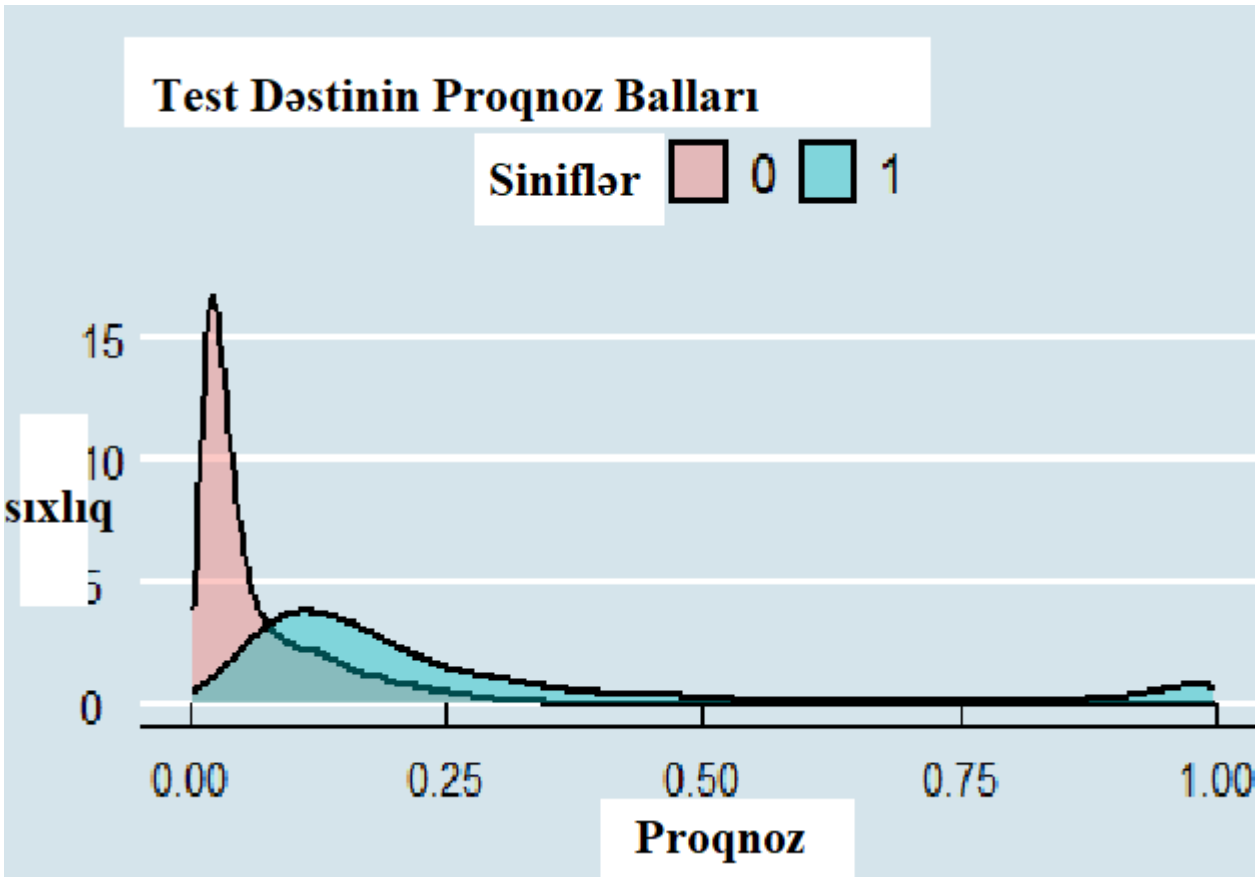
Qrafik 3. Modelin ROC əyrisi



Yuxarıdakı qrafikdə vuruş dərəcəsi həssaslıq, düşmə isə 1 – spesifiklik olan yanlış müsbət nisbətdir. Beləliklə, yuxarı sol küncdən əyriyə qədər olan minimal məsafə həssaslığı və spesifikliyi artırır, çünki ən yuxarı sol künc ideal nöqtədir.

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 seçmişik.





Yuxarıdakı qrafiklər sinif başına proqnozların sıxlığını göstərir. Verilən məlumatlarla olduqca üst-üstə düşdüyünü müşahidə etdiyimiz üçün diqqət tələb olunur. Üstəlik,

qrafiklər minimal məsafə kəsmə dəyərimiz 0.08-in olduqca məqbul olduğunu göstərir.

Maşın Öyrənməsi Nəticələri

Əlavə təsnifatlar üçün bir neçə əsas maşın öyrənmə vasitələrindən istifadə edəcəyik. 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. Burada çalışdığımızdan daha çox metod var. Məsələn, XGBOOST, Adaptive Lasso və Firth tənzimlənməsini də sınaqdan keçirdik, lakin əvvəllər görülən işlərlə əlaqəli bir yaxşılaşma əldə etmədik. Bəzi metodlar digərlərindən daha çox vaxt və hesablama gücü tələb edir. Bu metodları bütün xüsusiyyətlərdə işə salmaq və nəticələrin əldə edilməsi saatlar çəkə bilər. Beləliklə, xüsusiyyət seçimi üçün əvvəllər işləyən modellərdən LASSO operatorunu (alfa 1-ə bərabərdir) istifadə edirik. Bu yanaşma bizə çox vaxt qazandırdı. Aşağıda bəzi maşın öyrənmə üsullarının nəticələri verilmişdir.

Yuxarıda göstərilən nəticələr dəqiqliyi maksimallaşdıran kəsmə ilə əldə edilən ən yaxşı elastik- net modelimizə bənzəyir (əslində daha zəifdir).

Bu model DVM nəticələrindən bir az daha yaxşıdır, lakin indiyə qədərki ən yaxşı modelimizdən daha aşağı həssaslıq verir.

İxtiyari Meşə	
Dəqiqlik	0.935
Həssaslıq	0.17087
Spesifiklik	0.9952

Yuxarıda göstərilənlər də oxşar nəticələr verir. Burada

həssaslıq DVM-dən daha böyük, lakin KƏYQ-dən daha aşağıdır.

Yuxarıdakı nəticələr indiyə qədər ən maraqlı nəticələrdir. Bu metod hamıdan ən yüksək həssaslığı, minimum məsafəli kəsilmə ilə olan ən yaxşı modelimizdən daha yaxşıdır. Lakin təəccüblü bir şəkildə dəqiqlik və spesifiklik qəbuledilməz dərəcədə aşağıdır. Bu model sinif 1-i çox yüksək dəqiqliklə tuta bilir, lakin 0-ı yaxşı tuta bilmir (əslində 0 sinfində çox zəif nəticə göstərir). Bu, ən yaxşı modeli seçmək üçün birdən çox meyara ehtiyacımızın olduğunu sübut edir.

Nəticə və əlavə Müzakirə

Ən yaxşı modelimizin nizamlanmış logistik model və ya büzülmə modellərindən biri olduğunu gördük. Bu, bir qədər təmiz statistik öyrənmə ilə təmiz maşın öyrənmə metodları arasındadır. Bu yazının əsas nəticələrindən biri modellərimizi daha etibarlı şəkildə qiymətləndirmək və əsaslandırmaq üçün fərqli meyarlarla müqayisə etməyimizə ehtiyacın olduğudur. Digər maraqlı bir mövzu alfa parametri seçimidir. Burada səliqə və kompaktlıq məqsədi ilə 0 ilə 1 arasındakı 11 alfa dəyərini seçdik, əslində 0 ilə 1 arasında sonsuz sayda ədəd ola biləcəyini bilirik. Daha sonrakı araşdırmalarda alfa parametri seçimini böyütmək və misal üçün 100 (və ya hesablama gücündən asılı olaraq və mənasını itirmədən daha çox) dəyər verərək daha çox modellə nəticələndirmək olar. Baxılması lazım olan maraqlı bir məsələ də bir xüsusiyyətin seçilməyə başlandığı (və ya əksinə) zaman (və ya alfa dəyəri) olacaqdır.

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