Machine Learning istifadə edərək Bank Marketinq Kampaniyasının Uğurunu Proqnozlaşdırmaq

Nisə Aslanzadə

- Business Understanding
- Data Understanding
- Exploratory Data Analysis
- Data Preparation
- Model Building and Evaluation
- Result

BUSINESS UNDERSTANDING

Bu project-də bir bankın marketinq kampaniyalarının nəticələri təsvir edilib. Keçirilən kampaniyalar birbaşa telefon zənglərinə əsaslanır və bank müştərisinə müddətli depozit yerləşdirməyi təklif edir. Əgər bütün bu cəhdlərdən sonra müştəri depozit qoymağa razılıq veribsə, qarşısına "yes" qeyd edilir, əks halda "no" qeyd olunur. Bu məlumatlara əsasən, bankın gəlirində azalma müşahidə edilir və bu vəziyyəti düzəltmək üçün hansı tədbirlər görəcəklərini bilmək istəyirlər.
Araşdırmadan sonra məlum olur ki, əsas səbəb müştərilərin əvvəlki kimi tez-tez banka depozit qoymamasıdır.

DATA UNDERSTANDING

Data-mızda 41188 sətr və 20 sütun mövcuddur. Amma verilənlər bazasında təkrar sətrlər də çoxdur və ilk öncə onları handle edirik. Daha sonra isə 39404 sətr və 20 sütun qalır. Sütunlarımızda həm categoric, həm də numeric dəyərlər vardır. Target (y) sütunu isə ikili dəyərə malikdir (‘yes’ və ‘no’).

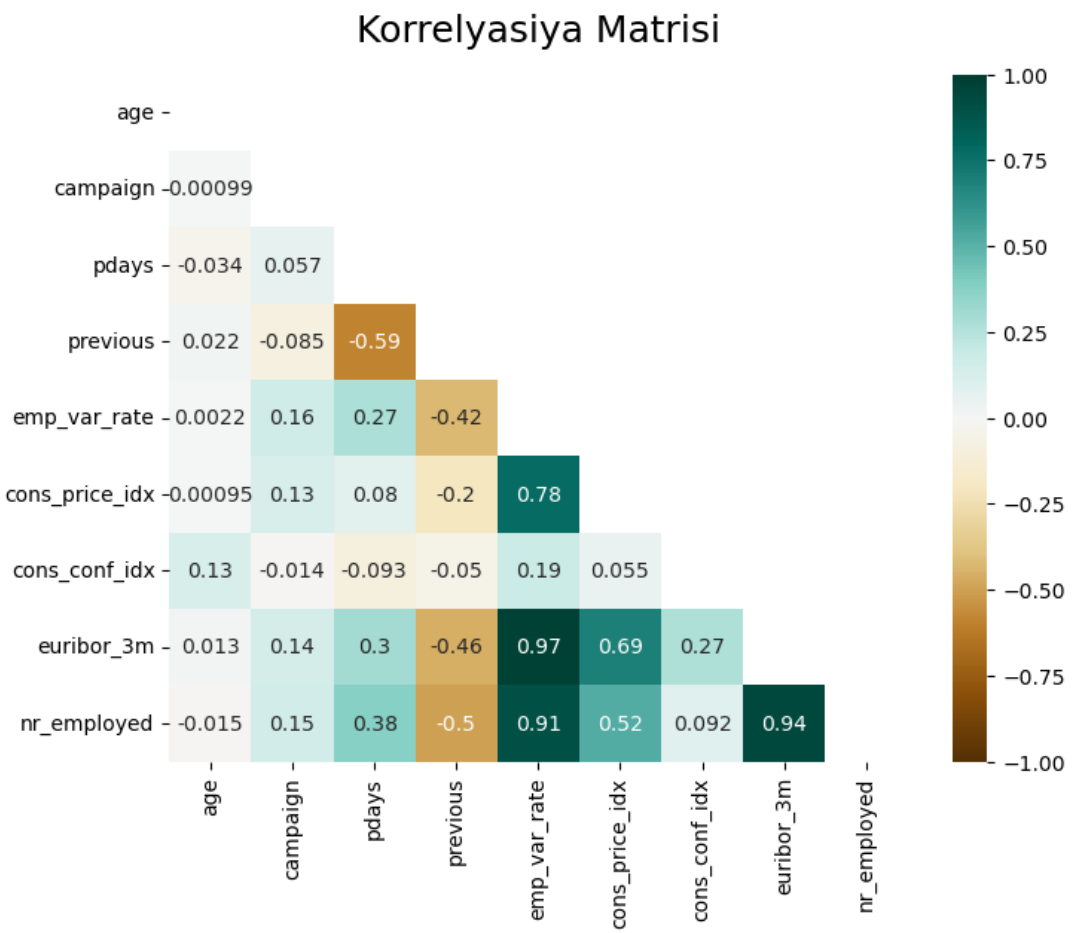
Sütunlarımız haqqında məlumatlar:

age: Müştərinin yaşı
job: Müştərinin işlədiyi sahə
marital: Müştərinin ailə vəziyyəti
education: Müştərinin təhsil səviyyəsi
default: Müştərinin ödənilməmiş krediti varmı?
housing: Müştərinin mənzil krediti varmı?
loan: Müştərinin şəxsi krediti varmı?
contact: Müştəri ilə əlaqə növü
month: Müştəri ilə son əlaqə saxlanılan ay
day_of_week: Müştəri ilə son əlaqə saxlanılan həftənin günü
campaign: Kampaniya zamanı bu müştəri ilə həyata keçirilən kontaktların sayı
pdays: Əvvəlki kampaniyadan sonra müştəri ilə sonuncu əlaqədən keçən günlərin sayı
previous: Kampaniyadan əvvəl bu müştəri ilə həyata keçirilən kontaktların sayı
poutcome: Əvvəlki marketinq kampaniyasının nəticəsi
emp_var_rate: Məşğulluq dəyişmə dərəcəsi - rüblük göstərici
cons_price_idx: İstehlak qiymətləri indeksi - aylıq göstərici
cons_conf_idx: İstehlakçı inamı indeksi - aylıq göstərici
euribor_3m: Avropa banklararası təklif dərəcəsinin 3 aylıq məzənnəsi - gündəlik göstərici
nr_employed: İşçilərin sayı - rüblük göstərici
y: Müştəri müddətli depozitə abunə olubmu?

Numeric sütunlarımızın statistik dəyərlərinə baxaq və analiz edək:

	count	mean	std	min	25%	50%	75%	max
age	39404.0	40.116105	10.460328	17.000	32.000	38.000	47.000	98.000
campaign	39404.0	2.618744	2.814780	1.000	1.000	2.000	3.000	56.000
pdays	39404.0	960.847097	190.869184	0.000	999.000	999.000	999.000	999.000
previous	39404.0	0.178738	0.503172	0.000	0.000	0.000	0.000	7.000
emp_var_rate	39404.0	0.064067	1.577041	-3.400	-1.800	1.100	1.400	1.400
cons_price_idx	39404.0	93.577538	0.583820	92.201	93.075	93.798	93.994	94.767
cons_conf_idx	39404.0	-40.499604	4.644327	-50.800	-42.700	-41.800	-36.400	-26.900
euribor_3m	39404.0	3.601243	1.742337	0.634	1.334	4.857	4.961	5.045
nr_employed	39404.0	5165.986481	72.763866	4963.600	5099.100	5191.000	5228.100	5228.100

1. **age:**
   - Mean (orta): Müştərilərin orta yaşı təxminən 40 yaşdır.
   - Min və Max: Müştərilərin yaşı 17 ilə 98 arasında dəyişir. Bu, bankın geniş yaş spektrində müştərilərə xidmət etdiyini göstərir.
   - Std (standart sapma): Yaş dəyişkənliyinin kifayət qədər yüksək olduğunu (10.46) göstərir, yəni müştərilərin yaşı müxtəlifdir.
2. **campaign:**
   - Mean: Hər bir müştəri üçün ortalama olaraq 2.6 kampaniya həyata keçirilib.
   - Max: Bir müştəriyə 56 kampaniya yönəldilib. Bu, bəzi müştərilərin çox sayda kampaniyada iştirak etdiyini göstərir.
   - Std: Dəyərlərin müxtəlifliyinin kifayət qədər yüksək olduğunu (2.81) göstərir.
3. **pdays:**
   - Mean: Sonuncu kampaniyadan sonra orta hesabla 960 gün keçib.
   - Min və Max: Bəzi müştərilərlə heç bir gün keçmədən əlaqə saxlanılıb, bəzilərilə isə 999 gün sonra.
   - Std: Çox yüksək standart sapma (190.87), bu dəyərin yüksək dərəcədə dəyişkən olduğunu göstərir.
4. **previous:**
   - Mean: Əvvəlki kampaniyalarda müştərilərlə ortalama olaraq 0.18 dəfə əlaqə saxlanılıb, yəni çox az müştəri ilə təkrar əlaqə saxlanılıb.
   - Max: Bəzi müştərilərlə 7 dəfə əlaqə saxlanılıb.
5. **emp_var_rate (işsizlik dəyişmə dərəcəsi):**
   - Mean: Orta hesabla 0.064. Bu, işsizlik dəyişmə dərəcəsinin müsbət tərəfdə olduğunu göstərir.
   - Min və Max: İşsizlik dərəcəsi -3.4 ilə 1.4 arasında dəyişir.
6. **cons_price_idx (istehlak qiymətləri indeksi):**
   - Mean: Orta hesabla 93.58.
   - Min və Max: İstehlak qiymətləri 92.20 ilə 94.77 arasında dəyişir.
   - Std: Standart sapmanın çox aşağı (0.58) olması dəyərlərin nisbətən sabit olduğunu göstərir.
7. **cons_conf_idx (istehlakçı inamı indeksi):**
   - Mean: Orta hesabla -40.50.
   - Min və Max: İstehlakçı inamı -50.8 ilə -26.9 arasında dəyişir. Dəyərlərin mənfi olması istehlakçıların ümumi inamının aşağı olduğunu göstərir.
8. **euribor_3m (3 aylıq Euribor faizi):**
   - Mean: Orta hesabla 3.60.
   - Min və Max: Euribor faizi 0.63 ilə 5.04 arasında dəyişir. Bu, banklar arası təklif dərəcələrinin kifayət qədər müxtəlif olduğunu göstərir.
9. **nr_employed (işçilərin sayı):**
   - Mean: Orta hesabla 5165.98.
   - Min və Max: İşçi sayısı 4963 ilə 5228 arasında dəyişir, bu isə rüblük olaraq işçi sayısının nisbi sabit olduğunu göstərir.



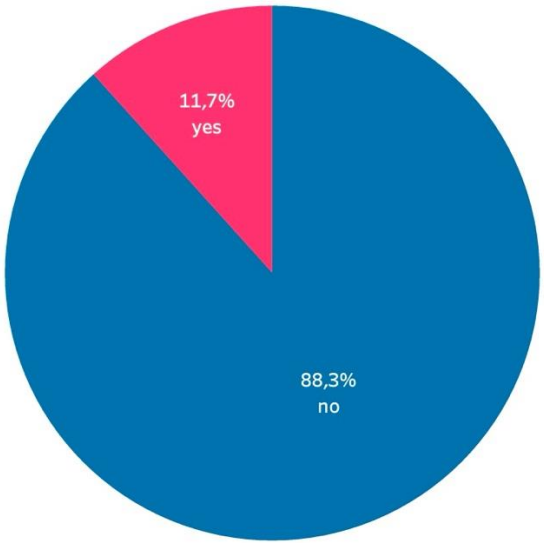
Matrisin əsas xüsusiyyətləri və müşahidə edə biləcəyimiz bəzi məqamlar:

1) Emp_var_rate və Euribor_3m: Bu iki dəyişən arasında ən güclü pozitiv korrelyasiya (0.97) var. Bu, məşğulluğun dəyişmə dərəcəsi və Avropa banklar arası təklif dərəcəsinin 3 aylıq məzənnəsi arasında güclü əlaqənin olduğunu göstərir. -muticol
2) Nr_employed və Cons_price_idx: İşçilərin sayı ilə istehlak qiymətləri indeksi arasında da yüksək pozitiv korrelyasiya (0.91) var. Bu da həmin iki dəyişənin bir-biri ilə əlaqəli olduğunu göstərir.
3) Previous və Pdays: Əvvəlki kampaniyadan müştəri ilə sonuncu əlaqə saxladıqdan sonra keçən günlərin sayı ilə kampaniyadan əvvəl bu müştəri üçün həyata keçirilən kontaktların sayı arasında orta dərəcədə mənfi korrelyasiya (-0.59) var.
4) Yüksək pozitiv korrelasiya: Cons_price_idx və Emp_var_rate arasında 0.78 dərəcəsində müsbət korrelyasiya var.
5) Zəif və ya əhəmiyyətsiz korrelasiyalar: Age, campaign, pdays və digər dəyişənlər arasında zəif və ya əhəmiyyətsiz korrelasiyalar müşahidə olunur.

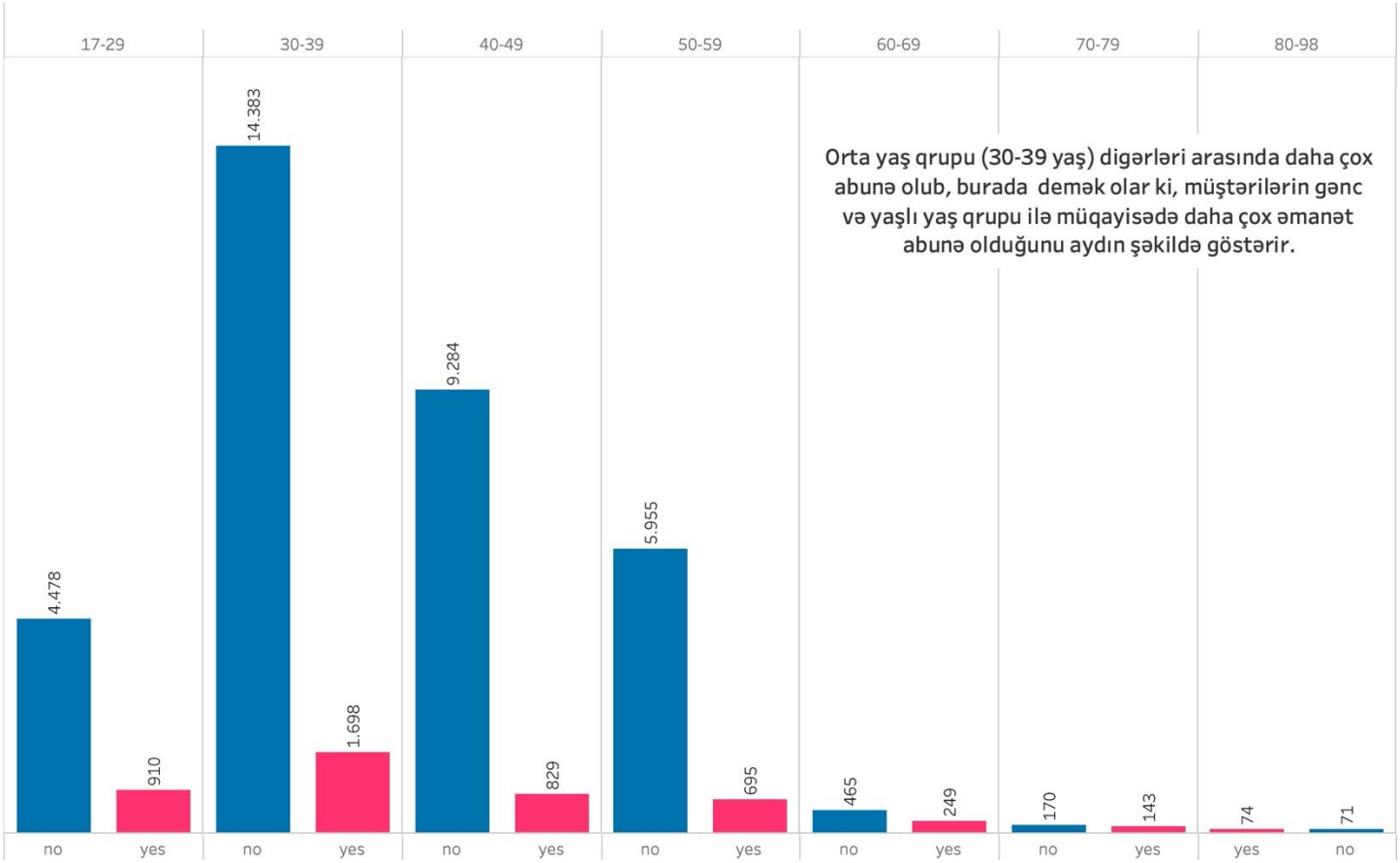
EXPLORATORY DATA ANALYSIS

Müştərinin müddətli depozitə abunə olma proporsiya

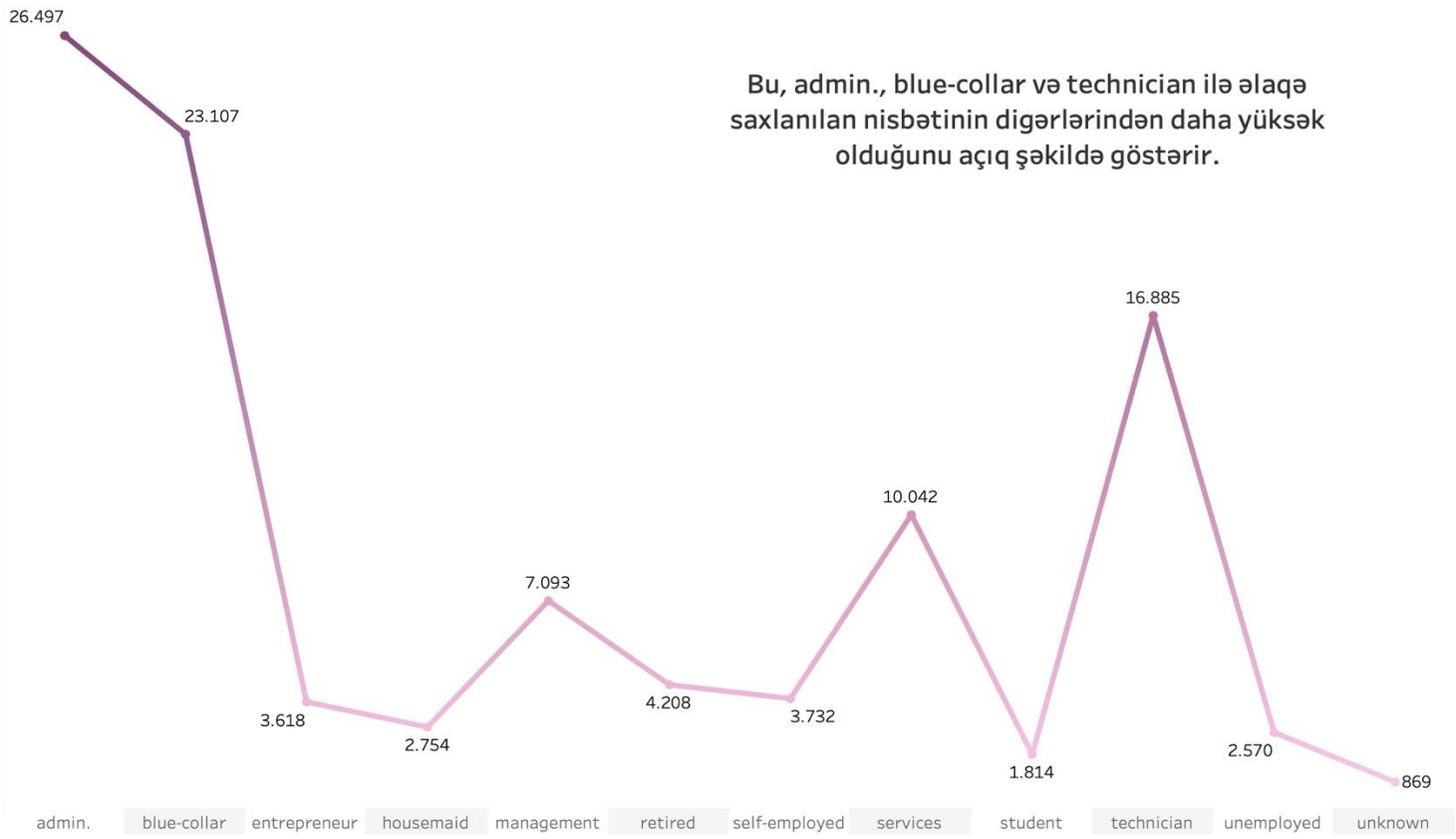
Bank Telefon Zəngləri ilə birbaşa marketinq kampaniyaları həyata keçirmişdir. Ümumilikdə 39404 nəfərlə əlaqə saxlanılıb, onlardan təxminən 12%-i yalnız abunə olub, 88%-i isə bankın təklif etdiyi məhsula abunə olmaq istəməyib.



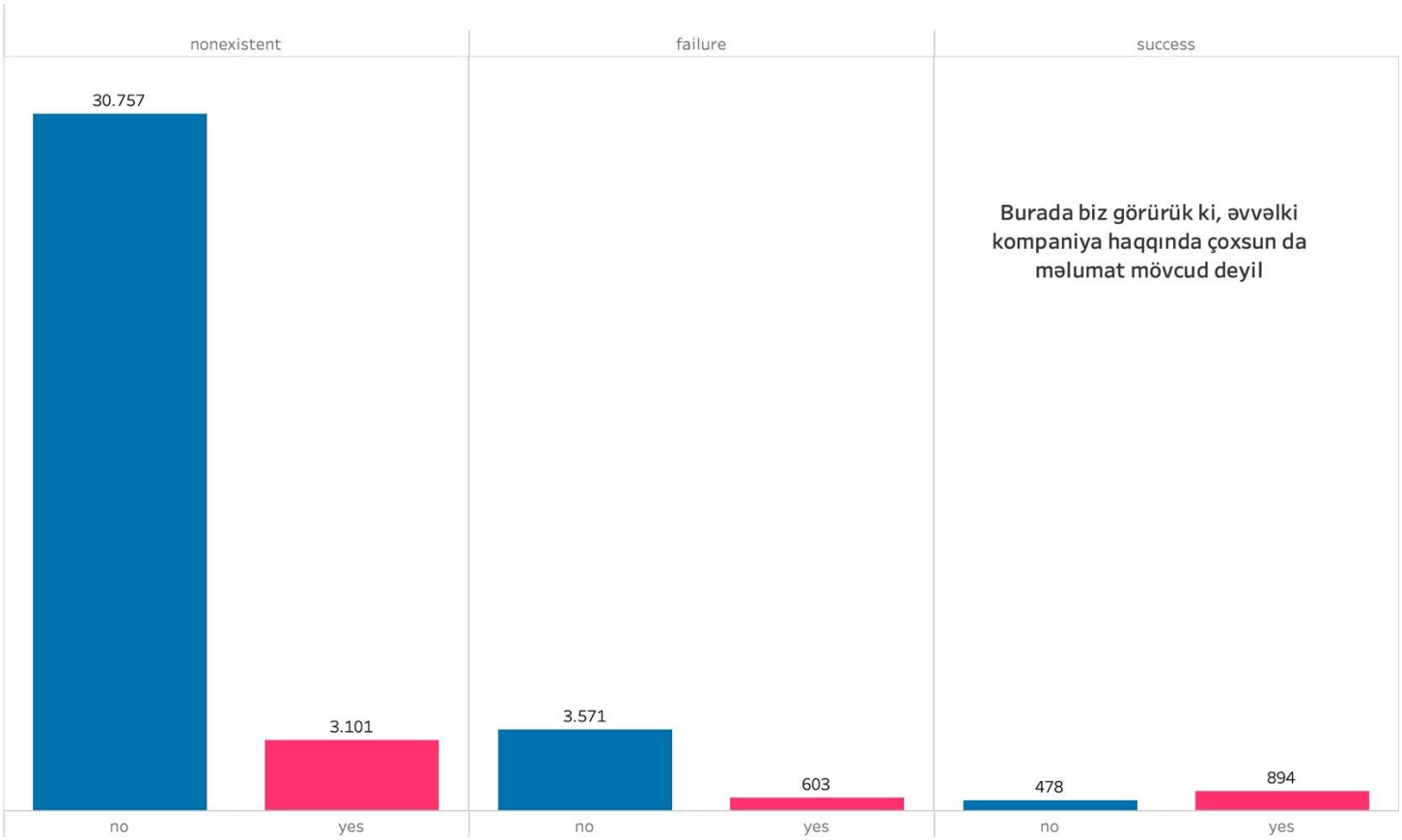
Yaşa görə abunə.



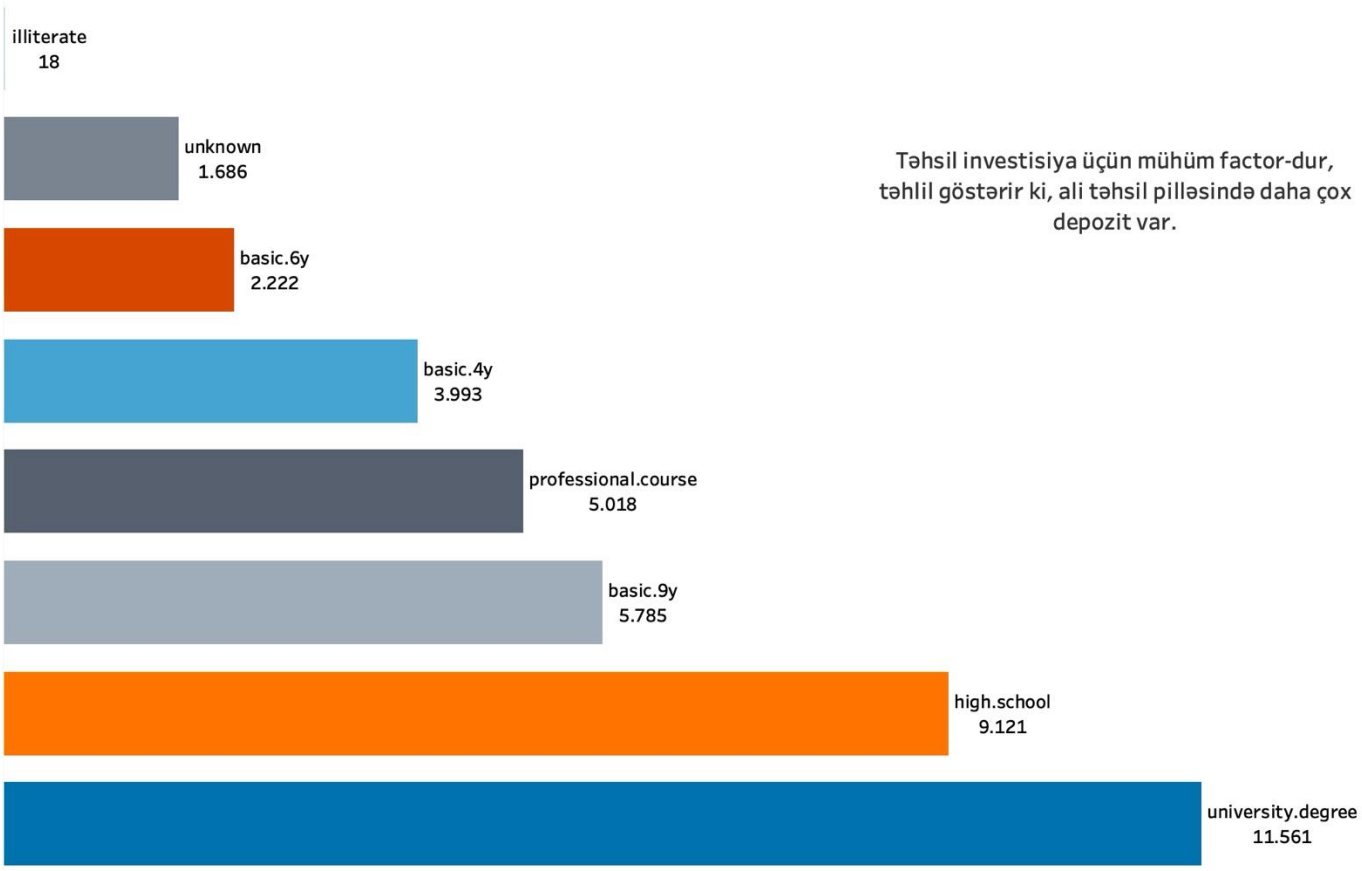
Kompaniya zamanı müştərilərin işinə görə edilən zəglər.



Əvvəlki kompaniyanın təsiri.

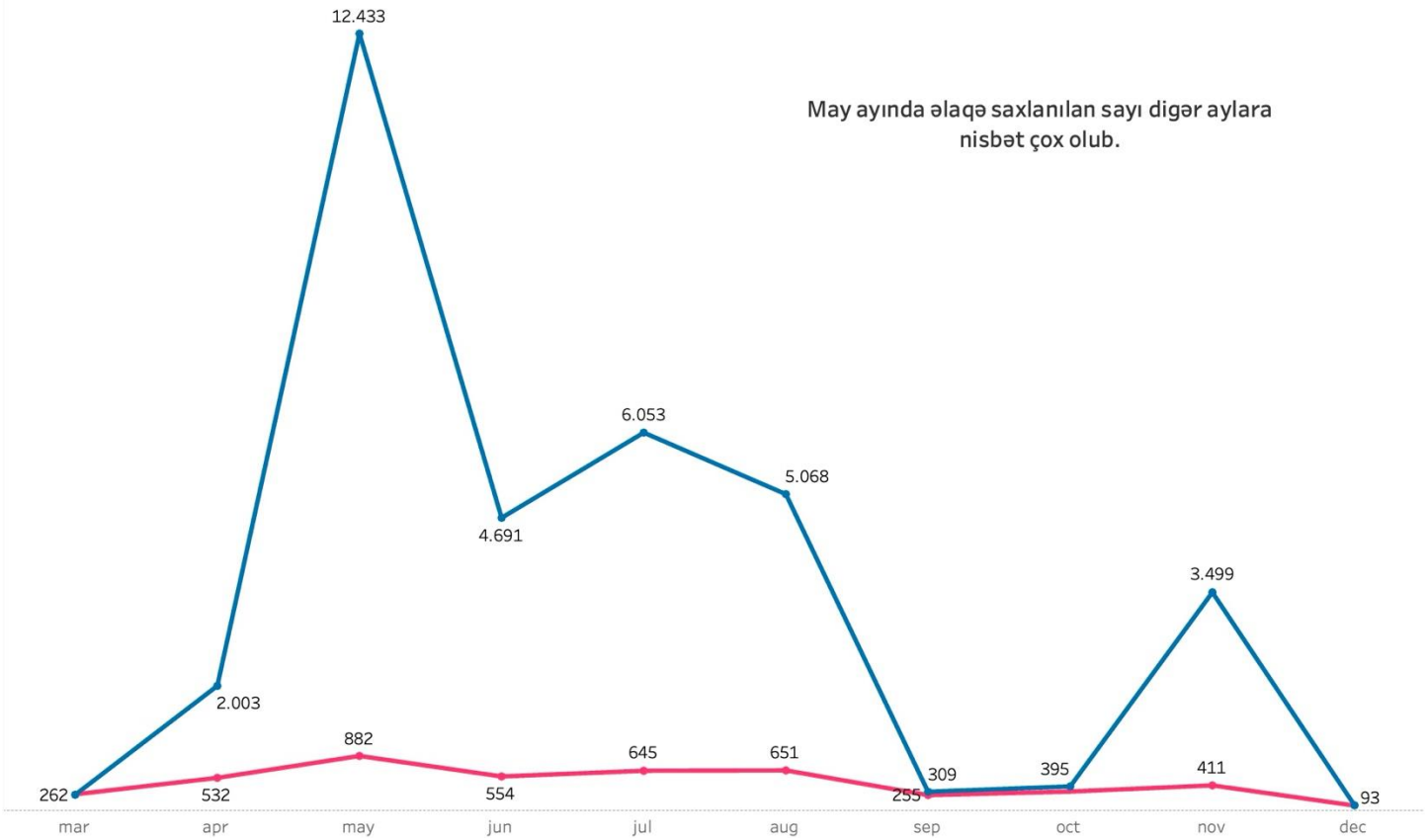


Kompaniya zamanı müştərilərin təhsilinə görə edilən zəglər.



Təhsil investisiya üçün mühüm factor-dur, təhlil göstərir ki, ali təhsil pilləsində daha çox depozit var.

Aylar üzrə kontakt sayı.



May ayında əlaqə saxlanılan sayı digər aylara nisbət çox olub.

DATA PREPARATION

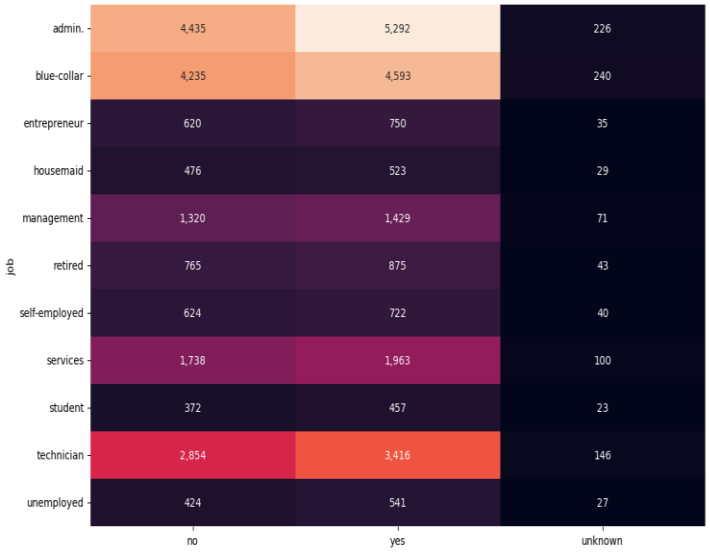
Chartlardan görmək olar ki, bir çox sütunlarda **unknown** dəyərlər mövcuddur. Çatışmayan məlumatları idarə etməyin bir çox yolu var. Yollardan biri sıradan imtina etməkdir, lakin bu, məlumatların azalmasına səbəb olacaq və bu səbəbdən də dəqiq proqnozlaşdırma modeli qurmaq məqsədimizə uyğun olmayacaq. Digər üsul isə "unknown" dəyərini digər dəyişənlərdən ağıllı şəkildə çıxarmaqdır. unknown dəyərləri olan dəyişənlər: **'job', 'marital', 'education', 'default', 'housing' və 'loan'**.

'job' və 'education' sütunları üçün fərziyyəm odur ki, 'job' insanın 'education'səviyyəsindən təsirlənir. Beləliklə, insanın təhsilinə əsaslanaraq 'job' nəticəsini təxmin edə bilərik.

Bu şəkildə, "unknown" dəyərlərin idarə edilməsində daha ağıllı və məlumatlı yanaşma tətbiq edərək, məlumat itkisini minimuma endirib, proqnozlaşdırma modelinin dəqiqliyini qorumaq mümkündür.

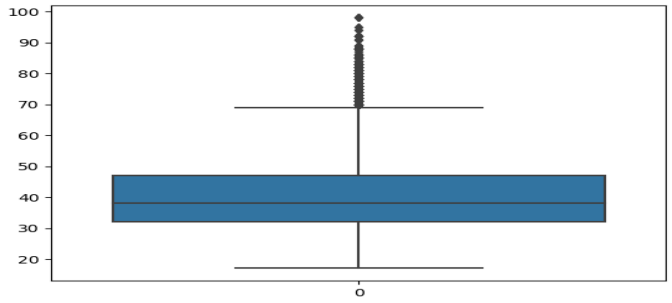
Bu məlumatlar vasitəsilə, 'education' və 'job' sütunları arasındakı əlaqələri daha aydın şəkildə görmək mümkündür. Bu əlaqələrə əsaslanaraq, 'education' və 'job' sütunlarındakı unknown dəyərləri təxmin etmək üçün bu cədvəldən istifadə edirəm.

'housing' və 'loan' sütunlarındakı unknown dəyərləri isə məntiqli şəkildə doldurmaq üçün 'job' sütununa əsaslanıram. Fərziyyəm odur ki, 'housing'hər bir 'job' kateqoriyası ilə nisbətdə olmalıdır. Beləliklə, hər bir iş kateqoriyasına uyğun olaraq unknown dəyərləri müəyyən edə bilərəm.

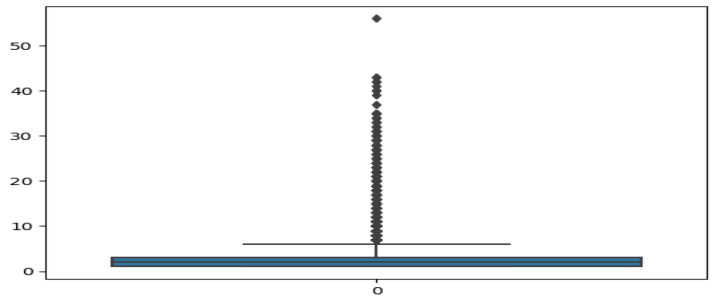


Outlier-lara baxdıqda isə, 'age' və 'campaign' sütunlarında outlier-ların çox olduğunu görürəm. Onları handle edirəm.

Age



Campaign



Növbəti olaraq, **'pdays'** sütunundakı bütün itkin dəyərlər "999" kimi kodlanmışdır. Bu sütun, əslində, çox sayda missing dəyər saxlayır. Daha ətraflı araşdırma göstərdi ki, bu missing dəyərlər əvvəllər heç vaxt əlaqə saxlanılmamış müştərilərə aiddir. Bu vəziyyəti idarə etmək üçün, 'pdays' sütununu müştərinin heç vaxt əlaqə qurulmadığı, 5 və ya daha az gün əvvəl əlaqə qurulmuş, 6-15 gün əvvəl əlaqə qurulmuş və s. kimi intervallara əsaslanan numeric sütunlarla əvəz etdim.

Data-mızda çox sayda kateqorik sütun olduğuna görə, **get_dummies()** funksiyasından istifadə edərək bu sütunlar əsasında yeni dəyişənlər yaratdım.

Daha sonra, **oversampling** üsulu ilə target sütunumuzdakı balanssızlıq problemini əlavə nümunələr yaradaraq aradan qaldırdım.

MODEL BUILDING AND EVALUATION

6 fərqli machine learning modellərindən istifadə edilib. Hər bir model fərqli qiymətləndirmə metrikləri üzrə test edilmiş və nəticələr müqayisə edilərək ən yaxşı model seçilmişdir.

Əvvəlcə dataset feature (X) və target (y) olaraq iki hissəyə bölündü. y target sütunu olaraq 'y_yes' seçildi, yəni müşahidənin müəyyən bir nəticəyə uyğun olub-olmaması. Train-Test Bölünməsi Dataset 70%-30% nisbətində train və test məlumatlarına bölündü, stratify parametri istifadə edilərək target dəyişəninin balansını qorumaq təmin olundu. Scaling edilməsi üçün MinMaxScaler istifadə edildi. Bu, hər bir xüsusiyyətin dəyərini 0 və 1 arasında sıxışdırmaq məqsədi daşıyır. 6 fərqli machine learning modelindən istifadə edildi: Logistic Regression, Random Forest, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Naive Bayes. Hər bir model 5-fold cross-validation ilə test edildi və 4 fərqli metrik üzrə qiymətləndirildi: F1 Score, Accuracy, Precision, Recall.

Cross-validation nəticələri:

- Logistic Regression: CV F1 Score: 0.71, CV Accuracy: 0.74
- **Random Forest:** CV F1 Score: 0.95, CV Accuracy: 0.95
- Support Vector Machine: CV F1 Score: 0.72, CV Accuracy: 0.76
- K-Nearest Neighbors: CV F1 Score: 0.84, CV Accuracy: 0.82
- Decision Tree: CV F1 Score: 0.93, CV Accuracy: 0.92
- Naive Bayes: CV F1 Score: 0.62, CV Accuracy: 0.69

RESULT

Bu project çərçivəsində **Random Forest** modeli verilən dataset üzrə ən yaxşı performansı göstərdi. Gələcək işlər üçün modelin hiperparametrlərini daha da optimallaşdırmaq və daha çox data toplamaq məqsədəuyğun ola bilər.