1. **Gereksinimin Tanımlanması**
   1. İş gereksiniminin tanımı

Acentelerimizin, şirketimize olan borçları üzerinden her dönem için risk gruplarını oluşturarak, riskli olan acentelerimizi belirlemek için çalışma yapılmıştır..

* 1. Problem sınıfı (supervised, unsupervised, semi-supervised, reinforcement)

Regresyon (Supervised Learning) ve kümeleme problemi.

* 1. Test senaryoları

Veri setini % 80 Train için ayrıldıktan sonra, geriye kalan veri setinde validasyon sonuçlarını görmek ve model doğrulama işlemi işlemi için k katlı çapraz doğrulama işlemi yapılacak.

Model doğrulanma işlemi gerçekleştikten sonra modelbaşarısı değerlendirilecek. (rmse veya mse v.b. )

Modelin eğitim hatası ile test hatası arasındaki durum ele alınacak (overfit, underfit v.b.)

Son olarak en 2023 yılına ait çeyrekler üzerinden 3 farklı canlı veride test işlemi gerçekleştirilecek

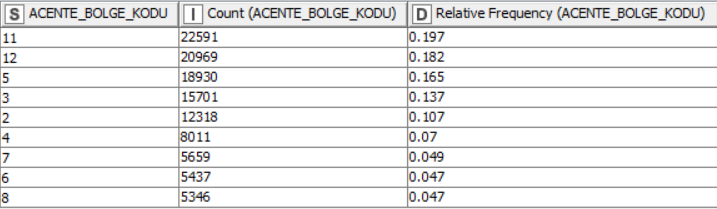
* 1. İş çıktılarının paylaşılacağı taraflar ve yetkilendirme

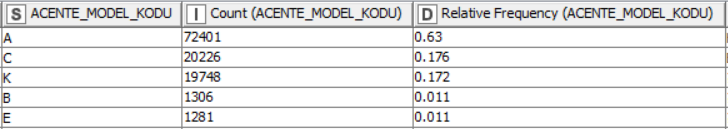
Yetkilendirme Sadece KNIME Web Portal üzerinde olacak olup, ilgili müdürlükteki arkadaşların kullanımına açılacaktır.

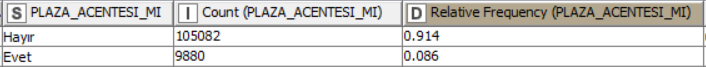
* 1. KNIME statistics node sonuçlarının paylaşılması

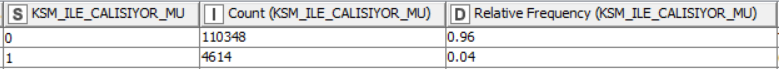
Kategorik Değişken : Şirket Tipi, Sadece AS, Plaza Acentesi Mi?, Ödeme Süresi, Acente Bölge Kodu, Acente Model Kodu değişkenleri , kategorik değişkenlerimizdir.







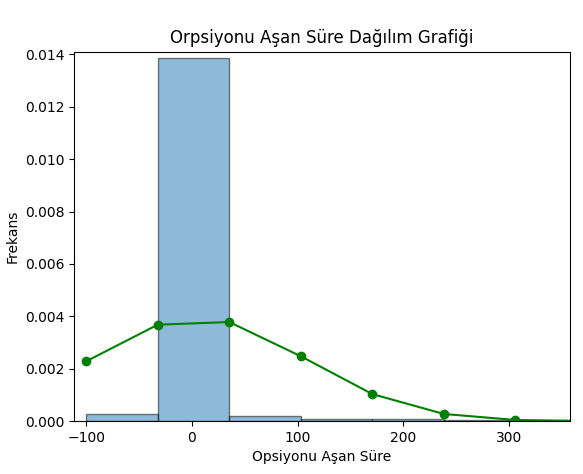
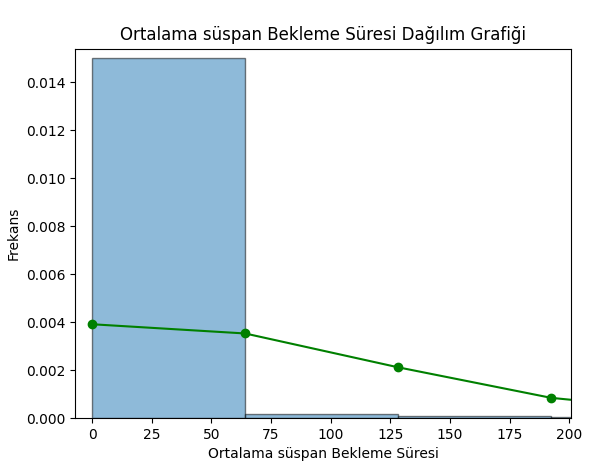






Sürekli Değişken:

|  |
| --- |
| ORTALAMA\_SUSPAN\_BEKLEME\_SURESI (Number (double)) |
| OPSIYONU\_ASAN\_SURE (Number (double)) |
| SUSPAN\_CEK\_ADEDI (Number (double)) |
| SON\_YIL\_KARSILIKSIZ\_CEK (Number (double)) |
| ODEME\_EK\_SURESI (Number (double)) |
| SUSPAN\_CEK\_TUTAR (Number (double)) |
| TOPLAM\_TEMINAT (Number (double)) |
| TOPLAM\_ODEME\_SURESI (Number (double)) |
| TASFIYE (Number (double)) |
| NET\_RISK (Number (double)) |
| DIGER\_SIRKET\_ADEDI (Number (double)) |
| FAAL\_ALT\_ACENTE\_ADET (Number (integer)) |
| EKRAN\_KAPANMA\_ADEDI (Number (integer)) |
| Acente\_Yasi (Number (integer)) |
| Target (Number (double)) |
| Target(-1) (Number (double)) |
| Target(-2) (Number (double)) |
| Target(-3) (Number (double)) |
| Target(-4) (Number (double)) |
| Target(-5) (Number (double)) |
| Target(-6) (Number (double)) |
| Target(-7) (Number (double)) |
| Target(-8) (Number (double)) |



1. **Data Preprocessing – Keşifsel Veri Analizi**
   1. Hedef değişken ve train-test-validation kümeleri için zaman aralığının belirlenmesi,

Veri setinde, %80 train, %20 Validasyon verisi olacak şekilde random olarak ayarlandı. K- Katlı cross validasyon yöntemi kullanıldı.

Test için ise 2021 yılı üç dönem 03, 06, 09 dönem için ayrı ayrı model testleri yapılmıştır.

* 1. Hedef değişken unbalanced ise çözüm için uygulanan metot

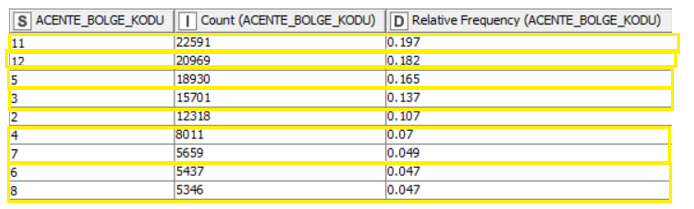
Unbalanced bir durum söz konusu olmamaktadır.

* 1. Data etiketleme ve temizleme için kullanılan yardımcı makine öğrenmesi metotları

Data temizleme için Lasso Regression Yöntemi kullanıldı.

* 1. Veri tiplerinin düzenlenmesi

Kategorik veri tipleri one hot yöndemi ile her bir kategori değişken olacak şekilde düzenlendi.



* 1. Veri tekilleştirme ve detay seviye tespiti

Her bir acentenin son 8 dönemlik (2 yıllık) verileri kullanılarak veri acente kodu seviyesinde tekilleştirilmiştir.

* 1. Parametrelerin dönemselliğinin belirlenmesi

Dönemsellik yıllık bazda olduğu için her yıl belirlenecek olan enflasyon üzerinden bir kat sayı ile model değişkenleri (sürekli değişkenler), güncellenecek olan model tekrar train edilecektir.

* 1. Constant Column Filtering
  2. Variance Filtering

% 90 ve üzeri sabit değer olan kolonlar veriden çıkartılmıştır.

* 1. Kayıp değerler için kullanılan metotlar

Kayıp değerler bulunmamaktadır.

* 1. Dağılımların incelenmesi

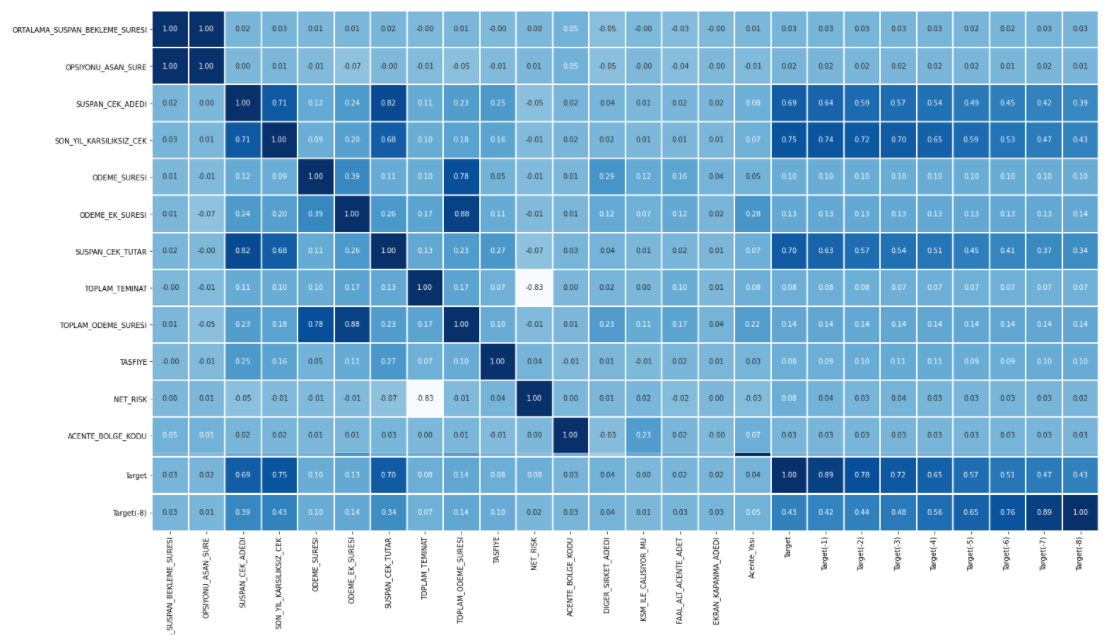
Veri dağılımlarında elde sürekli değişkenler 2.b ‘de gösterilmiştir.

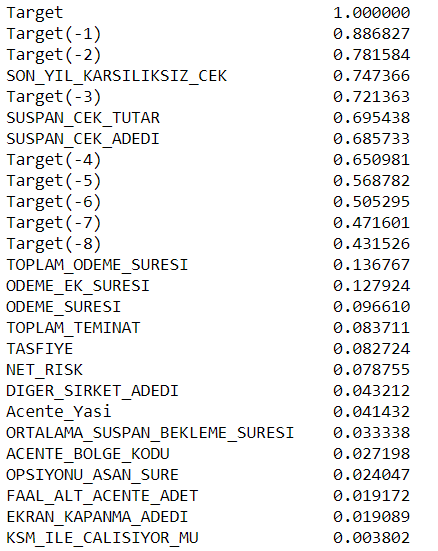
* 1. Outlier ların incelenmesi

Bir üst maddede gerçekleşen grafiklere ek olarak outlier belirlenen değişkenler özelinde incelenip en yakın değere indirgenecek şekilde veri setinde düzeltilmiştir.

* 1. Covariation / Correlation

Aşağıdaki şekil incelendiğinde değişkenlerin birbirleri ile alakalı korelasyon matrisleri gösterilmektedir. Bu matris incelendiğinde her bir değişkeni diğer değişkenler ve Target (Hedef değişken) üzerindeki ilişkisi gözlenlenebilmektedir. (İlişki 1 ile -1 arasında olmaktadır. 1 ‘e yaklaştığında pozitif yönlü ve kuvvetli bir ilişki durumu gözlemlenirken, -1 ‘e yaklaşıldığında negative yönlü ama yine kuvvetli bir ilişki durumu gözlenmlenmektedir.)





* 1. Normalizing and Scaling

Sürekli değişkenlere Normalizasyon çalışması (Z Score Normalization) yapıldı.

* 1. Dimension reduction

Değişken sayıları üzerinde birbiri üzerinde korelasyon olan değişkenler veri setinden çıkartıldı.

* 1. Encoding Categorical variables

Kategorik değişkenlere (Şirket Tipi, Sadece AS, Plaza Acentesi Mi?, Ödeme Süresi, Acente Bölge Kodu, Acente Model Kodu değişkenleri) encoding işlemi yapıldı. RareLabelEncoding.

1. **Uygulanan Çözümün Detayları**
   1. Çözüm seçimi ve nedenleri

Problem üzerinde regresyon çalışmaları yapılarak çözüm elde edildi.

Birden Fazla Makine Öğrenmesi Modelleri (Linear Regression, Random Forest Regression, Gboost Regression v.b.) denendi ve denemeler sonucunda, Adj R^2 değeri ve mae (Mean Absolude Error) değerileri üzerinden model seçimi yapıldı.

* 1. Çözüm Makine öğrenmesi algoritması kullanımını gerektiriyorsa;
     1. Yapılan ilk modele ilişkin bilgiler
        1. Algoritma seçimi

Gredient Boosting Regression Algoritması seçildi.

* + - 1. Test-train-validation setlerini ayırmak için kullanılan yöntem.

Train verisi ilk başta (EDA yapılmadan randomly olarak veri setinden ayrıldı.) Validation seti içinde 10 Katlı çapraz doğrulama kullanıldı.

* + - 1. Hyperparameters
      2. Cross Validation uygulaması

K- Katlı Çapraz Doğrulama (cv = 10) yöntemi.

* + - 1. Performans metriklerinin açıklanması

Mas (Mean absolute error) hataların ortalama değeri

* + - 1. Verilen geribildirimler ve bunlara ilişkin yapılan çalışmalar

-

* + - 1. Aynı algoritma ile devam mı edilecek / başka bir algoritma mı çalışılacak

-

* + - 1. Hata analizi (yeni eklendi)
  1. Çözüm Makine öğrenmesi algoritması kullanımını gerektirmiyorsa;
     1. Yöntem
     2. Parametre seçimi ve kullanılan parametrelerin açıklanması
     3. Performans metriklerinin açıklanması

1. **Sonuç**
   1. Modelin kaydedilmesi ve canlıya alınmasına ilişkin bilgiler

Gboost Regression modeli kaydedildi. Ek olarak normalizer sonucu elde edilen parametrelerde kaydedildi.

* 1. Modelin performansını otomatik olarak ölçümleyecek akışlara ilişkin bilgiler
  2. Uçtan uca modelin baştan eğitilebilmesi için yorumlar ile desteklenmiş uçtan uca workflow teslimi

Bu çalışma KNIME Web Portal üzerinden ilgili müdürlükle paylaşılıp iki opsiyon sunulmuştur:

Çalışma belirli (3 Aylık) dönemlerde ya kullanıcı bağımlı,

Kullanıcı bağımlı (Web Portalden kullanıcı istediği zaman)

Çalıştırılacak olan bir sistem düşünülmelidir.