Patika & EnerjiSA Veri Bilimi ve Analitiği Bootcamp Bitirme Sunumu Grup-2

1. **Sayfa - tanıtım**

Bootcamp bitirme sunumumuza hoşgeldiniz. Sizlere yaklaşık 2 haftadır üzerinde çalıştığımız, Nitelikli Kaçak Tahminleme projemizden bahsedeceğiz. Ekip arkadaşlarım …

2. **Sayfa - kullanılan tek.**
   Kullanılan Teknolojiler
3. **Sayfa - neler anlatacağız**

İlk önce problemin ne olduğundan kısaca bahsedecek olursak; **slayt oku.**

4. **Sayfa Projenin amacı ve konusu**

Bu proje ile Enerji sektörünün öncüsü EnerjiSA tarafından verilen veri setinden nitelikli kaçak tahminleme analizi yapıldı.

- Kayıp Kaçak: Kullanım yerine ilişkin olarak perakende satış sözleşmesi veya ikili anlaşma olmaksızın dağıtım sistemine müdahale ederek elektrik enerjisi tüketmesi.
- **Kayıp Kaçak Yöntemleri Nelerdir? :** Dağıtım sistemine, sayaçlara, ölçü sistemine ya da yapı bina giriş noktasından sayaca kadar olan tesisata müdahale edilerek tüketimin doğru tespit edilmesini engellemek. Müdahale sonucu sayacın eksik veya hatalı ölçüm yapması veya hiç ölçmemesine sebep olmak
- Biz Ne yaptık?: Verilen veri setindeki parametreler incelenerek, gerekli feature engineering aşamaları tamamlanarak veri setinden nitelikli kayıp kaçak tahminleme işlemi yapılmıştır.

5. **Sayfa - Verinin Anlaşılması**

Verimiz toplamda 15000 satır ve 50 sütundan oluşmaktadır. Verilerin önemli bir kısmı olan %72'sini tüketim ve demand sütunlarının oluşturduğu gözlenmiştir. Veri setinde kolonlarda yüzdesel olarak ne kadar eksik veri bulunduğu ve sayısal değer içeren kolonlar istatistiksel olarak incelenmiş, korelasyonu bulunmuştur. Korelasyon değerinin en fazla Risk Skoru kolonlarında %29 oranla çıktığı gözlenmiştir. Eksik veriler %21.17 ile en fazla DEMAND_M12 sütununda bulunmaktadır. Yapılan bu işlemler görselleştirilerek anlaşılması daha kolay hale getirilmiştir.

6. **Sayfa - Kaçak Oranı**

NK_FLAG değerlerinin piechart gösterimi ile toplamda %6.8 oranla 1020 adet kaçak kullanıcı olduğu görüldü. 15.000 adet verimizin olduğu göz önüne alındığında 1020 adet kaçak olması verinin imbalance bir veri seti olduğunu göstermektedir.

7. **Sayfa -tesisat_tipi plot**

Veri setinde 50 adet satırın tüketim harici hiçbir değerinin olmadığı gözlendi. 50 adet verinin 19 tanesinde yani %38 oranında kaçak kullanıma ait olduğu bulundu. 50 verilik bu pattern gözönüne alınarak tesisat tipi bazında grafiğe dönüştürüldüğünde bunu elde ettik. Bu 50 adet eksik verilerin büyük bir çoğunluğunu ticarethanelerin oluşturduğu görüldü.

8. **Sayfa -Veri Hazırlama (Feature Engineering):**

Eksik verilerin doldurulmasında 0, yüksek bir sayı, 'EKSIK' değeri ve özellikle demand ile tüketim sütunlarının doldurulmasında bir önceki ve bir sonraki değerlerin ortalamasının alınması yöntemiyle doldurma işlemleri yapılmıştır.

Yeni kolonlar oluşturduk bunlar:

SAYAC_YAS = güncel tarihten SAYAC_BASLANGIC_TARIHI çıkarılarak yeni bir sütun oluşturuldu.

ABONELIK_SURESI = Kullanıcının en son tüketimine ait ayın tarihi güncel tarihten çıkarılarak kaç aylık abone olduğu bilgisine ulaşıldı.
TUKETIM_std = Her bir tüketicinin tüketim verilerinin standart sapmasından oluşturuldu.
DEMAND_std = Her bir tüketicinin demand verilerinin standart sapmasından oluşturuldu.

Mevcut/Oluşturulan kolonların istatiksel çıkarımlar doğrultusunda gruplanması:

SAYAC_YAS_group = Oluşturduğumuz SAYAC_YAS kolonundaki verilerin dağılımı dikkate alınıp, 0-3 yaş, 4 yaş ve 4+ yaş şeklinde 3 gruba ayrıldı. Gruplamada edindiğimiz bussiness bilgisi ile TEDAŞ'ın elektronik sayaç şartnamesine göre bir elektrik sayacı maksimum 10 yılda 1 değiştirilmelidir. Buna göre sayaçları yeni olan abonelerin kaçak kullanıma daha yatkın olduğunu düşündük. Grafikten de anlaşılacağı üzere sayaç yaşı düşük olanlarda kaçak kullanım oranı daha yüksektir.

Sayaç Marka = Sayaç markalarına göre kaçak oranlarının gösterimidir. Sayaç marka ile model bilgisi verimizde var ve iki veri incelendiğinde sayaç modelinin kaçak bulmada daha faydalı bir pattern oluşturduğu görülmüştür.

SAYAC_MALZEME_ID = 76 tane sayaç modeline karşılık 76 tane SAYAC_MALZEME_ID olduğunu gördük. İki sütunun da aynı durumu karşıladığını gördük. Bu sebeple SAYAC_MALZEME_ID sütunu yerine SAYAC_MODEL sütununu kullanmaya karar verdik.

SAYAC_MODEL_group = Sayaç modellerindeki verilerin KAÇAK ORANI göz önüne alınarak gruplama işlemi yapıldı. Sayaç modellerinden LUN-10B modelinde diğerlerinden çok daha fazla kaçak tespit edildi. Toplam 2261 adet LUN-10B modelinden 462 tanesi (%20si) kaçak kullanıcıdır.

Model Hazırlama (Design Modelling Data)

- Modeli eski verilerle eğitip, en güncel verilerde test etmek için, train-test split işlemi yapılmadan önce veriler sayaç başlangıç tarihine göre sıralandı.
- Daha doğru sonuçlar alabilmek için bize verilen 15000 satırlık veride, en güncel 3000 veri validasyon, geri kalan 12000 veri de train verisi olarak kullanıldı. %80'lik veriyi de kendi içinde %70-%30 olarak tekrar ayırdık.
- Mevcut veri setimiz dengesiz (imbalanced) olduğundan undersampling yaparak kıymetli verileri kaybetmek yerine oversampling işlemi ile modelin başarısı artırıldı.
- Kategorik değişkenler için One-Hot Encoding işlemi uygulandı.

Modelin Tahminlemesi (Model Prediction)

Model tahminlemesinde DecisionTree, RandomForestClassifier, Logistic Regression, CatBoost Classifier, XGBoost Classifier modellerini denedik. En doğru tahminlemeyi CatBoost classifier'ın yaptığını gördük ve işlemlerimize bunun üzerinden devam ettik.

Bize verilen 15.000 satırlık veride modelleme yaptığımızda. Recall değerimiz %83, f1-score'u %77 olarak tespit ettik. Recall değerinin %83 olması ile veri setinde zaten çok az miktarda bulunan 1'lerin yani nitelikli kaçakların sayısını tespit etmede yüksek bir oran yakaladık.

Feature Importance:

CatBoost'un feature importance metodu kullanılarak çizdirilen grafikteki feature importance verileri şekildeki gibidir.

İlk iki sıradaki %1-3 ve %3-5 ile gösterilen değerler sayaç modellerinin gruplanmış halidir. Onlardan sonra gelen SOKAK_RISK_SKORU veride bize verilmiş olan sütunken bir altındaki ABONELIK_SURESI bizim oluşturduğumuz yeni bir sütundur ve modeldeki etkisi yüksektir.

Feature importance grafiğine göre modelimizin herhangi bir feature'a bağımlı olmadığı görülmüştür

ROC Curve:

ROC Curve eğrisinin 90 dereceye yaklaşıyor olması, eğrinin altında kalan alanın maksimuma yakın olması modelimizin başarılı olduğunu göstermektedir.

Modelin Validasyon Sonuçları

Modelin hiç görmediği veri seti ile tahminleme sonuçları göz önüne alındığında 0.5 threshold değeri ile recall oranında 0.08 , f1-score'unda 0.02'lik bir düşüş olduğu görüldü. Bu tarz bir düşüşü zaten öngörüyorduk.

Elimizdeki veri dengesiz (imbalanced) olduğu ve nitelikli kaçağı tespit etmek önemli olduğu için recall değerinin yükseltilmesi büyük önem arz etmektedir. Çünkü recall değerindeki her bir artış bizim modelimizin daha iyi oranda nitelikli kaçak tahmin etmesi anlamına gelecektir. Bunun için threshold değeri 0.15'e düşürülerek yeni bir recall ve f1-score elde ettiğimizde, threshold'un 0.5 olduğu esas modelimize göre f1-score'dan ciddi ölçüde feragat ettiğimiz görülmüştür. Bizden istenen f1-score'un maksimize edilmesi olduğu için bir önceki tabloda gösterdiğimiz değerler modelimizin sonuçlarıdır.

Modelin recall değerlerine baktığımızda 0 ve 1'leri %91 oranında tahmin ettiğini gördük. Threshold'u 0.15'e kadar düşürmemize rağmen model Sıfırları hala çok iyi tahminliyor. Modelin sıfırları çok iyi tahminlemesinde bu modeli en çok etkileyen parametrelerden biri olan SOKAK_risk skoru 0 olan yaklaşık 11bin abonenin %98'inin kaçak kullanmıyor olması söylenebilir.

.