Rapor:

Proje: Formula 1 Veri Analizi

Yazar: Osman Çalışkan

Bu projede Formula 1 yarış verilerini analizi yapıldı ayriyeten 2023 yılı şampiyonun tahmin edilmesi için makine öğrenimi modeli geliştirildi.

Aşağıda veri setinin içerdiği bazı değişken isimleri ve bunların açıklamaları var.

- raceId: Yarış kimliği
- year: Yarış yılı
- round: Yarış sayısı
- date: Yarış tarihi
- time: Yarış başlangıç saati
- circuitId: Pist kimliği
- circuitName: Pist adı
- country: Pist ülkesi
- weather: Hava durumu
- trackTemp: Pist sıcaklığı
- airTemp: Hava sıcaklığı
- humidity: Nem
- windSpeed: Rüzgar hızı
- windDirection: Rüzgar yönü
- grid: Grid pozisyonu
- start: Başlangıç pozisyonu
- finish: Bitiş pozisyonu
- points: Puanlar
- positionOrder: Sıralama

- Başlangıç olarak projemde bir sürü csv dosyası vardı.Bunları nasıl tam anlamıyla birleştirebiliceğimi bilmiyordum.Concat ve Merge fonksiyonlarını kullanarak tüm veriyi birleştirmeyi denedim.Bu şekilde yapınca çoğu veri birbirine girdi ve bazı veriler kayboldu.Bu yüzden sonrasında analiz ediceğim verilerin hangilerinin işime yarıycağını tespit ederek ilerledim.
- Tüm verileri inceledikten sonra içlerinden ‘results’,’stats’,’drivers’,’races’,’constructor,’driver_standings’ veri setlerini seçtim.Bunları merge ederek birleştirdim.
- Verileri incelerken bazı bilgilerin alındığı ‘url’sütununu gördüm ve bunun gereksiz olduğuna karar vererek bunları ortadan kaldırdım.
- Sütun sayısı çok olduğundan bazı sütunlarının isimlerini göremediğimi farkettim ve bunları nasıl görünür hale getirebiliceğimi araştırdım ve pd.get_option(‘display.max_columns’,None) şeklinde bir kod ile karşılaştım ve bunu kullandım
- Sonrasında datasetin içerisinde işime yaramıycağını düşündüğüm sütunları dropladım
- Bazı sütunların isimleri tam anlamlarını ifade etmediğini düşündüğüm için onları rename fonksiyonuyla yeniden isimlendirdim

- Sürücülerin isimleri ve soy isimleri farklı sütünlar halinde ‘forename’ ve ‘surname’ şeklinde verilmişti.Bunları tek bir sütun da toplayarak bu sütunları drop ile birlikte ortadan kaldırdım.
- Dob sütunu altında verilen string şeklindeki tarihleri pd.to_datetime koduyla birlikte date formatına dönüştürdüm
- Sürücülerin yaşlarını tam sayı şeklinde anlayabilmek için from datetime import datetime adı altında bir kütüphane buldum ve ‘dob’ içerisindeki tarihleri kullanarak bunları yaşa çevirdim.Sonrasında bunları ‘age’ adı altında bir sütun haline getirerek df’in içerisine ekledim.Eklemeden önce de yaşların bazıları küsüratlı değer verdiğinden dolayı bunu round fonksiyonuyla yuvarladım
- Constructor csv’sini kullanarak bunun içerisinden ‘name’ ve ‘points’kısımlarını groupby yaparak .en başarılı F1 10 takımının plot grafiğini çizdirdim
- Takımların yarış başına düşen puanlarını hesaplayıp görmek istedim.Bunda constructors csv dosyasının içerisindeki ‘name’,’points’,’raceId’ kısımlarını merge ederek bunları unique değerlerine göre sıraladım ve 100’den fazla yarışa sahip olan takımları göz önüne alarak bunların bir listesini çıkardım.Sonrasında ayriyeten grafiğini çizdirirsem daha iyi gösterileceğini düşünerek görselleştirdim.
- F1 sporuna millet olarak ne kadar düşkün olunduğunu ve hangi milletlerin buna daha çok önem göstererek sürücü yetiştirmeye çalışıldığını merak ederek sürücülerin uluslarına göre bir fig çizdirdim.Fig’i daha önce hiç görmemiştim fakat farklı kodlar incelerken denk geldiğim birşeydi.Bu gösterimde de güzel gözüküceğini düşünerek kullandım.
- En genç 5 sürücüyü buldum ve bunların yaşlarını inceledim
- İncelerken \N şeklinde değerler gördüm ve bunları NaN değerleri haline çevirdim çünkü ilerisinde bazı işlemleri yapmamı engelleyebiliceğini düşündüm.
- Burada en genç pilotları incelerken ve en yaşlılara bakarken bir sorun farkettim.Daha öncesinde 'age',sütunu oluşturduğumda ölen pilotların bilgisi elimde olmadığı için max yaşları çok fazla çıkıyordu .Bunların en son yarış yaptıkları tarihlere bakarak yaşlarını güncelledim fakat emekli olanları ayrıştırmak istemediğim için en son yarışlarını 2015'te yapanların yaşlarını 2015 tarihine göre güncelledim.
- Merge edilmesinden dolayı bazı sürücülerin isimlerinin birden çok tekrar ettiğini gördüm ve 2023 yılında yarışan sürücülerin yaş ortalamasını merak ettiğim bunları unique şeklinde alarak ortalamalarını hesapladım.
- 2023 yılında yarışan sürücülerin yaşlarını ve sayılarını görmek istediğim için bir çubuk grafik çizdirdim
- F1 tarihinde en çok kazanan 10 yarışçıyı merak ettiğim için bunları loc ve groupby yaparak belirli bir değişkene atadım.Sonrasında daha net görebilmek için isimleriyle ve kazandıkları yarış sayısına göre bunları seaborn bar plotu haline getirerek görselleştirdim.
- Kazanmanın yanı sıra F1 tarihinde en çok yarışan 10 yarışçıyı görmek istedim ondan sonrasında ve orda hala yarışmakta olan Fernanda Alonso’nun 1.olduğunu gördüm
- Birden çok yarış pisti olduğunu bildiğim için bu yarış pistlerinde kaç yarışın düzenlediğini görmek için bunların verilerini çektim.Sonrasında da bunları görselleştirmenin güzel gözüküceğini düşünerek Top 10 Grand Prix F1 diye bir barplot oluşturarak bunları görselleştirdim

- Yıllar içinde yapılan yarış sayılarının sezondan sezona farklılık gösterip göstermediğini merak ederek bunu görselleştirmek istedim

- Farklı kodlarla inceleme yaparken ve bazı sütunların ne işe yaradığını görebilmek adına araştırma yaparken lng ve lat adı altında bazı sütunlar gördüm.Başkasının bunu kullanarak Map üzerinde görselleştirme yaptığını gördüm.Bundan faydalanarak benzer bir haritada ben oluşturdum ve üzerindeki araba simgeleriyle güzel bir görüntü ortaya çıktı.

- Heatmap oluşturmaya çalışırken bir çok problem yaşadım.Verileri skew etmeye çalışırken bazı veriler uyumsuz olduğu için çok sıkıntı yaşadım.İlk önce bunları droplamaya çalışmıştım aslında fakat tamamiyle de gitmelerini istemiyordum.Sonrasında yapay zekalardan yararlanmaya çalıştım fakat onlarla da çıkmaz bir yola girdim ve hem kod çok uzadı ve veri tiplerinden dolayı bir türlü sorunla karşılaştım ve çalıştıramadım.Sonrasında dataframei başka bir dataframe içerisine kopyaladım.Ondan sonrasında bazı sütunları çıkararak bunları skew ettim ve korealasyonlarını alarak bir heatmap oluşturdum.

- F1 içerisinde overtaking adı altında bir hareket vardır ve bu aslında sollama yapmak demektir.Bu açıdan en kötü olabilicek pistleri belirlemek istedim.Burdan yarışlar ve sonuçlarına bakarak pist isimlerini belirledim ve bunları görselleştirdim.

- İzlediğim yarışlardan hatırladığım bazı pistler hız yapmak açısından çok zordu çünkü mesafeler çok kısaydı ve hız yapılamıyordu.Bunları görselleştirmek istedim ve yarışların tur sürelerine bakarak bunların ortalamasını aldım.Güzel bir görselleştirme ile ortalama hıza göre en kötü pistleri belirledim.

- Yarış galiplerin milliyetlerine göre yıllar içerisindeki değişimi görmek adına bir görselleştirme yaptım.

- Yarışlarda ortalama tur sürelerinin yıllara göre değişip değişmediğini merak ettim.Çünkü her geçen yılda araba teknolojilerinin geliştiğini bildiğim ve takip ettiğim için bunu görselleştirmek istedim.Yıllara ve yarışların tur sürelerine göre bunların ortalamasını alarak bir görselleştirme yaptım.

- 2022 yılı verilerine göre 2023'te kimin şampiyon olacağını tahmin etmek istedim ve bunun için adımlar atmaya başladım.

- 2022 yılındaki fastesLapTime sütunuyla işlem yapmaya çalıştığımda yapamadım ve bunu dakika cinsine çevirdim.Sonrasında bu değerleri de float tipine dönüştürdüm.

- Tahmin için 'raceId','grid','points','laps','fastestLap' ve 'fastestLapTime' sütunlarını kullandım hedef olarak 'positionOrder' belirledim.

- Bunların yanıt etiketlerini kodlayıp eğittikten sonra veriler içerisinde eksikler olduğunu farkettim hata alıp duruyordum.Yapay zekalardan faydalanarak SimpleImputer adı altında bir fonksiyon keşfettim ve bunları tahmin ettirip doldurttum.

- Karar ağacı sınıflandırcısı oluşturarak verilerimi eğittim ve bunları test verileriyle test ettim.Başlangıçta bununla alakalı çok sorun yaşıyordum çünkü ben sürücü adını tahmin etmek istiyordum sonrasında 1 değerini veren şampiyonun adını araştırarak ve yapay zekalardan faydalanrak bir dataframe oluşturarak bulabildim.Bunu da bazı

verileri encoder.inverse_transform(y_pred)sayesinde yaptım.Sonrasında bunları ilk şampiyon adayının indeksini alarak ordan sürücü adayının ismine ulaşarak yaptım.

- İf-else yapısı kullanrak şampiyon olabilicek sürücü olup olmadığını kontrol ettim ve şampiyon olabilicek sürücü adayını Max Verstappen olarak buldum.Şu an ki 2023 yılındaki gidişatına bakılırsa da verim doğru çıkıcak gözüküyor.

Buraya kadar okuduğunuz için teşekkür ederim.Ayriyeten Çağla Hocama da bu yolda bize çok değerli bilgiler verdiği için çok teşekkür ederim.Bu alana her zaman ilgim vardı fakat nerden başlıycağımı bilemiyordum.Kendisine yol gösterdiği için ve bu işi sevdirdiği için çok teşekkür ederim.

Saygılarımla