**23.09.2020**

**EDA ( Exploratory Data Analysis ) Processes**

* Missing Values
* Outliers
* Feature Engineering
  + Extraction:
    - Mevcut sütunlardan 2-3 tanesini birleştirip, yeni sütunlar oluşturma
  + Transformation:
    - Bir sütunun bilgilerini alıp, başka sütunlar oluşturma

**COURSE NOTES:**

* Sütun isimlerini bir listede tutularak, indeksleme yapılırken lazım olduğunda çağırıp isim kopyalamasında kullanılabilir. col=list( df.columns)
* Yüksek Nan içeren (Missing Values) sütunlardan başlayıp onlara odaklanmak gerekir. ilk etapta %35 üzeri Nan içerenlere EDA faaliyeti yapmak üzere odaklanılır.
* %90 dan fazla Nan içeren sütunlar drop edilir. Normalde, profesyonel ticari projelerde %50 den fazla missing value varsa drop edilir.
* Çalışılan sütunun ismi Markdown yapılarak, o başlık altında söz konusu sütunda çalışılır. Gerekli yerlere hatırlatma notları eklenir.
* Aynı bilgileri tutan iki sütun varsa ikisinde de cleaning yapılır ve sütun değerleri düzgün dağılım gösteren sütun alınır, 2. sütun atılır. veya 2. sütundaki data, missing value ları doldurmak için kullanılır.
* Sonuca(fiyat vs) etkisi olmayacağı düşünülen sütunlar, %100 dolu olsada drop edilir.
* Aynı işlemleri çok sayıda yapacaksak jenerik hale dönüştürüp, fonksiyonu yazılır
* Nan sayısı az ise (gerekirse ilgili sutun gruplanıp) mode kullanılır.
* Nan sayısı fazla ise ffill/bfill kullanmak daha uygun doldurma metodu olabilir. Ffill yapılınca hala Nan varsa, ardından birde bfill yapılır.
* Outlier sayısı fazla ise median kullanılarak filling yapılmalıdır.
* ML ye sokunca %95 üzeri değer çıkıyorsa Validation, Cross Validation yapılır. Overfitting, underfitting, data leakage var mı bakılır.
* Üzerinde çalışılan (Feature Engineering) orijinal kirli sütunlar en son drop edilir. Extraction ve transformation dan sonra elde edilen temiz sütunlar, dataframe e yeni sutun olarak tanımlanır.
* ÖNEMLİ!!!

Data Cleaning işlemi 1 .notebook ta yapılır, sonunda dosya precleaned.csv olarak kaydedilir

Handling Mising Value & Outlier için 2.notebook gerçekleştirilir, precleaned.csv üzerinde çalışılır.

Filling işlemleri normalde 2. Notebookta yapılır ancak basit ve ayrıntı çalışma gerektirmeyen sütunlar için filling 1. Notebookta yapılabilir.

* str.contains(), str.extract(), str.replace(), regex ve str.get\_dummies() etkili kullanılmalıdır. Basit olan işlemlerde regex yerine replace tercih edilmelidir.
* Pandas Serials lara replace yaparken regex attribute default olarak True dur, ancak DataFrame e replace yaparken regex atribute default olarak False tur. Bu nedenle DataFrame e replace yaparken replace içine regex=True eklenmeli.
* Bir sütunu temizleyip yeni sütuna atarken, sütun isiminde küçük harf ve \_ kullanılabilir.
* Eğer çok kritik bir sütun değerleri karışık ve anlamsız gelmişse, hatalı script edilmiştir. imkan varsa işverenden bu attribute ile ilgili verileri tekrardan düzgün şekilde göndermesi istenmelidir. Eğer database den SQL ile verileri kendiniz çekti iseniz, bu sütunu bir kez daha çekmelisiniz.
* Temizlenen kolonların içeriği numerikse, pd.to\_numeric(...) ile dönüştürülmeli, tarihse, pd.to\_datetime(...) yapılmalı
* get.dummies() yapma süreci:
  + Nan ---> [ ] yap, Nan\_to\_list(df, col)
  + lists içinden verileri “|“ ile ayrılmış vaziyette str formatta dışarı çıkari, insert\_sep(df,col)
  + [ ] -----> np.nan yap, list\_to\_Nan(df,col)
  + fill\_most( ) veya uygun başka metod ile Nan ları doldur.
  + df[col].str.get\_dummies() yap.
  + .add\_prefix("item\_") ile Series ların indexlerine, df lerin columns adlarına ön ad konulabilir. get dummies sonrası yeni oluşan columns lara hatırlatıcı prefix eklenir.
* Feature Importance/Feature Evaluation nedir?

Projemizin bittiğinde son halinde ML de diyelim ki %96 accuracy/R square çıktı. Bu kesinlikte sonuç vermesi için müşteri tüm kolon bilgilerini girmesi gerekir. ancak bu kadar çok sayıda(belki 100 sütun) kolon bilgisi girmek pek mümkün olmaz. Feature Importance algoritmaları ile fiyata etkileyen en önemli birkaç sütun(6-7 sütun) belirlenir. sadece bu sütunlar ile örneğin %90 accuracy elde edilebilir. Geriye kalan tüm sütunların tahminde etkisi %6 dır. Feature impoortance bu önemli 6-7 sütunu belirlemekle alakalaıdır.

* Referansımız TARGET sütunudur. Target sütunumuz ise projede “price” sütunudur. Bir sütunda feature engineering yaparken, o sütunun target sütununa etkisi onun değerini gösterir. Sütunun değerine göre harcanacak vakitte ona göre ayarlanmalıdır, etki ve önemi az bir sütunda belli başlı yaklaşım ve kabullerle filling yapılarak zaman kazanılmalıdır. Sütun değerli ise çok ayrıntı çalışılabilir.
* TARGET sütununun outlier larında satır drop yapmak tercih edilebilir. Price ı Target olarak belirledi iseniz, onun değerleriyle oynamamalısınız. Ne winsorize, ne logaritma, ne scale uygulanmamalı. 2.notebookta Target sütundaki outlier ları drop etme yoluna gidilir.
* Örneklem sayısının fazla olması çok önem arzeder. Bu yüzden (Target sütun dışında ) satır drop etmek çok tercih edilmez.
* Unsupervised Machine Learning de Target sütunu olmaz. Bunda label lar da yoktur. Bu tarz EDA da tamamen veriyi temizleyip MLA nın kollarına emanet diyorsunuz. Kurs kapsamında yapılan EDA, Supervised Machine Learning e giriyor.
* Sütunda temizlik yaparken her adımda Nan ların sayısı özellikle takip edilmeli, Nan sayısındaki değişiklikler yorumlanıp anlamlandırılarak ilerlemeli.
* Bir sütundaki patternleri eksiksiz tespit etmek için , ayrı ayrı 55lik samples alarak veri incelenir. Farklı patternler keşfedilir. Her farklı pattern için farklı çözüm gerekebilir.
* Bir kolondaki Categorical değişkenlerin sayısının anlamlı şekilde düşürmenin yolları aranmalı. Benzer veya yakın anlama gelen categorik değişkenler birleştirilmelidir. Böylece get\_dummies yapınca daha az kolon ortaya çıkar.
* Tarih ve Finans bilgilerinin missing value larında çok dikkatli olunmalı. Çoğu zaman filling mümkün olmayabilir ve drop yapılır. Bu tarz veriyi varsayımsal filling yapanın, veriyi manuple etme ihtimali yüksektir.

**INSTRUCTOR SPECİAL ADVİCES:**

* Bir Data Scientistin, bu alana özgü nitelikte AWS/DevOps öğrenmesi tavsiye edilir.
* Bir ITcinin 10parmak klavye kullanması beklenir.
* DS’in specific konularında uzmanlaşılırsa daha kolay iş bulunabilir. Mesela:
  + NLP: Natural Language Processing,Text to Speece/Speece to Text işleme gibi
  + Computer Vision: Görüntü işleme
  + Deep Learning te CNN,RNN