İSTANBUL VERİ BİLİMİ BOOTCAMP

HOUSE PRİCE VERİ SETİNDEKİ DEĞİŞKENLERİN İNCELENMESİ

TAKIM – 2

KODLUYORUZ

İSTANBUL BÜYÜKŞEHİR BELEDİYESİ

İPA İstanbul Planlama Ajansı

VERİ LABORATUVARI

MsSubclass: Kategorik veri, kategoriler bazında tekrardan düzenleme yapılabilir.

MsZone: En çok RL kategorisinden var. Azınlıkta olan kategorilerin etkisine bakılabilir ve

RL ve diğerleri olarak sınıflayabiliriz.

LotFrontage: 259 missing values. boş değerleri ortalama ile doldurabiliriz/ silebiliriz. Yüksek korelasyonlu değişkenlerle boş değerler doldurulabilir.Uç değerler çıkarılabilir. Sokağa sıfır olan 40-50 ev var. 300'den fazla olan

1 ev var: yanlış girilmiş olabilir veya uzak müstakil ev olabilir. Bu çıkarılabilir.

(Bu değer çıkınca normal dağılıma yaklaşıyor olabilir.)

LotArea: Uç değerler incelenebilr.Metrekarenin fiyata etkisi çok olduğundna uç değerleri

elemek yerine metrekareye göre gruplayıp değerlendirilebilir.

Street: 1454 pave, 6 gravel olduğu için gravel kurulacak modele anlamlı değer katmayabilir.

Alley: missing değerler = 93% ulaşım yolu yok anlamında;yani boş değer değil.

Highly correlated with sales prices.

LotShape: çoğunlukla düz ve düze yakın şekiller tercih edilmiş.

ikili gruplar halinde(reg+IR1) (IR2+IR3) karşılaştırılabilir.

LandContour: high corr. with neighbourhood. 4 büyük değer var.

Missing val yok. Çoğu ev düzlük alanda. Eğim artıkça evin sıklığı azalıyor.

*****Utilities*****: 1 unique. NoSeWa (1) değerini çıkarmak mantıklı olabilir.(Bu değer yanlış girilmişte olabilir.)

Başka değerlerden olmadığından kategorileştirme yeniden yapılabilir.

Lot config: 5 distinct. Çoğunluk "Inside"; "Corner".

>Nan değerler no pool olarak değiştirilebilir.

LandSlope: 3 distinct. High corr. with neightbour. No missing val.

Yüksek korelasyonlu değişkenlere bakılabilir.

Neighbourhoor: 25 distinct. eşit dağılım var.

Cond.1:

Cond.2: İki koşulu kendi aralarında karşılaştırma VEYA ikisini birleştirme.(HOCAYA SORALIM)

BuildingType: Single family detached en fazla.

Outlier değerler de önemli. Az olan kategorilerde evler daha pahalı olabilir.

Korelasyonlara bakılabilir.

HouseStyle:Dengeli bir dağılım var. Kategoriler arasında normal dağılıma yaklaştırılabilir.

OverallQual: Malzeme kalitesi. Normale yakın dağılım, sayısal veriler.

No missing value ve fazla değişkenle high correlation. Sale price ile pozitif korelasyon

Genel olarak ortalam malzeme kalitesi kullanılmıştır.

OverallCond.:Korelasyonlu olduğu alanlar var.

YearBuilt: Yapım yılı. 1880 öncesinden 2006'ya kadar yapılan evler var.

Model sola çarpık model, öncelikle normal dağılıma yaklaştır. Sapmalar geniş. Normal dağılıma yaklaştırınca 1880 ler dışarıda kalır. yüksek korelasyonlu ilişkileri modelde kullanmakta fayda var. 3 ayrı tepe noktası var gibi=3 kategorik veri=3 normal dağılımlı grup.

>Değişkenimizi kategorik veriye dönüştürebiliriz.

YearRemodAdd: Binada yapılan tadilatların tarihi. No missing val çünkü tadilat yoksa yapım tarihi girilir. Tadilat yapılan yıl ve construction year arasındaki fark alınabilir.

Farkı sıfır çıkanların restore edilmediği görülmüş olur.

1950 yılındaki tadilat sayısındaki artış önceki yıllardaki evlerin restoresi 1950de yapılmış olabilir ya da yasalarla alaklı olabilir. bu yıldaki artış incelenebilir.

RoofStyle: 6 distinct var. Gable ve hip çoğunluğu oluşturuyor. Kategorik veri old için tip-sayı karşılaştırmasını yapabiliriz. Bir değerde baskın olduğundan modele eklenmeyebilir.

RoofMtl: Bir değişken çok baskın olduğundan modelde kullanılmayabilir.

Exterior1st: 5 değişken de normale yakın dağılmış. Az sayıda olanlar Other values olarak gösterilmiş/ bir gruba alınmış.

Exterior2: Exterior1st ile aynı

MasVnrType: Sola çarpık dağılım. 2 kategoriye ayırıp duvarın desenli veya desensiz olduğunu ayırabiliriz.

MasVnrArea: bir önceki değişkene bakılıdığında gördüğümüz kaplanmayan değerler bu tabloda zeros olarak karşımıza çıkmıştır. Missing value olanlar en çok olan değerle doldurulabilir.

ExterQual: Dış yalıtımda kullanılan malzeme. Katagorik, no missing val, çoğunluk TA ve Gd alanında, sonrasında Ex ve Fa.

ExterCond:

Foundation: kategorik veri

BsmtQual: Missing values=NA(no basement) TA ve Gd birbirine en yakın ve verinin çoğunu oluşturuyor. Ev satışında bodrumun yüksekliğinin etkisi var. Sayısal veriyi kategorikleştirmiş

BsmtCond: Missing values=NA(no basement)

BsmtExposure: en çok no verisi var, düzgün anlaşılmaıs için tekrar kategorileştirilebilir

na = no basement

BsmntFinType1: na = no basement

BsmntFinSF1: Zero fazla => no basement + no exposure toplamı olabilir.Sıfırların etkinliği düzenlenebilir(-0.95 düzenlemesi).saleprice ile high corelation.

İstatiksel yöntemlerle normal dağılıma yaklaştırılıp modele koyulabilir.

BsmntFinSF2: İkinci bodrum genelde yok

Heating: Boş gözlem yok. Kategorik. GasA, yani bir veri, baskın. Modele katmak anlamlı değil.

Foundation,catralair, lowqualfinsf => korelasyonlu olduğu değerler.

HeatingQC: Isıtmada bir kategori baskın olduğu için bu çok önemli değil. Kategorik veri.

CentralAir: Boolean. Modele katmaya gerek yok.

Electrical:Kategorik, bir veri baskın.

1stFlrSF: Sayısal alan, boş gözlem yok, normale yakın distribution. 3000 ve 4000 civarında

birkaç tane outlier var. Yanlış girilip girilmediğini araştırabiliriz.

2stFlrSF: 0 verisi fazla. Yaklaşık 250-1300 arası normal dağılım. Eğer sıfır verisi 2. katın

olmadığını belirtiyorsa bu veriyi çıkarabiliriz. Kategorik veriye çevrilebilir.

LowQualFinSF: 98% boş değer.

GrLivArea: mean-median yaklaşık=>normal dağılımlı değişken, boş gözlem yok.

Çoğu değşikenle high corr.=>modele katılabilir.

BsmtFullBath: saleprice ile yüksek korelasyon. sayısal değer.

BsmtHalfBath: " " " "

FullBath:

HalfBath:

Bedroom: sayısal alan. boş değer yok.

KitchenAbvGr: sayısal alan, genelde 1 mutfak var; baskın.

Hiç mutfağı olmayan değer yanlış girilmiş olabilir. İki mutfağın fiyata etkisi var mı

bakılabilir. Modele katkısı yok denebilir

KitchenQual: Saleprice ile korelasyon var, modele katılabilir. boş gözlem yok. Good ve Avr baskın

TotRmsAbvGrd: Normal dağılım, boş gözlem yok, modele katılabilir.

Functional: . kat çok baskın modele katmak anlamsız

Fireplaces: boş gözlem yok, iki değer baskın, 1 ve 2'yi birleştirip:

şömine VAR(1), YOK(0) yapılabilir. iki tane high korelasyon.

GarageType: missing values= garaj yok, high corr. 1 tane.

GarageYrBlt: missing values = garaj yok, yıl aralığına bölerek kategorileştirilebilir.

0-1940;1940-1980;1980-2000;2000-ve sonrası

GarageFinish: missing values=garaj yok, kategorileştirip modele eklenebilir

GarageCars: 0 = garaj yok, kategorik, saleprice ile corr. var

GarageArea: missing values= garaj yok, saleprice ile corr. var. kategoriye bölebiliriz.

100-400; 400-700;700-1000;1000-ve sonrası

GarageQual/Cond:-

PavedDrive:-

WoodDeck: deck olmayanlar çoğunlukta olduğu için YOK ve VAR olarak kategorileştirebiliriz veya

var ve yok olanların fiyata etkisine bakabilriz.

OpenPorch: aynı

enclosedporch: aynı

MoSold: korelasyon yok modele eklemeye gerek yok

SONUÇ: *Değişkenlerimiz incelendiğinde görülmüştür ki; sayısal değişkenlerimiz genellikle normal şekilde dağılmışlardır. Normal dağılmayan sayısal değişkenlerimiz için normal dağılıma yakınlaştırılabileceğimizi veya kategorileştirme yoluna gidilebileceğimiz gözlemlenmiştir.*

Kategorik verilere bakıldığında ise genellikle bir değerin baskın olduğu görülmüş ve bağımlı değişkenimizi çok etkilemeyeceği için model aşamasında kullanılmaması gerektiği sonucuna varılmıştır. Kategorik verilerimizdeki boş gözlemlerin, bir kategoriye karşılık geldiği veri setimizin açıklamasında verilmiş ve buna göre yeniden düzenleme yapabileceğimiz fark edilmiştir.