Analiza podataka

- ➔ 18 atributa ima NA vrednosti u celom DataSetu-u.
- ➔ **IsBadBuy je IZLAZ** : da li je kupovina auta pametan izbor za dalju preprodaju.

Izdvajanje atributa:

- **PurchDate** (godina kada je auto stavljen na aukciju) ima godinu u sebi, izdvojili smo godine u poseban atribut jer su veoma bitne za odluku. Pravimo atribut **PurchYear .** Postoje 2 godine 2009 I 2010 god aukcija, I relativno jednak broj automobile u obe godine.
- **Auction** se odnosi na dobavljaca automobila I ima nedostajuce vrednosti I to 3815. Dakle preko pola ima NA I jos oko 1598 auta ima dobavljaca OTHER, **sto ovaj atribut cini neprakticnim za dalju analizu.**
- **VehYear** predstavlja godinu proizvodnje automobila a **VehicleAge** predstavlja starost automobile u godinama. **Ova dva atributa su visoko korelisana, te izbacujemo jednog od njih.**
- **Make** predstavlja marku automobila, I imamo 32 marke automobila. Pored modela automobila imamo I **model** automobila I to 632 modela. To je previse, te zelimo da ovo svedemo na manji broj automobila da ne dodje do pretreniranosti. Vidimo da u **okviru naziva modela automobila postoji I oznaka za vrstu pogona I broj cilindara na motoru.** Postoje 3 vrste oznaka – AWD, 4WD I 2WD za vrstu pogona kao I V8, V6, 6C, I6 za broj cilindara. Shodno tome, kreirali smo jos 2 atributa **DriveType I NumCyls** .
- **Model** koji sadrzi u sebi vec ova 2 podatka (pogon I broj cilindara), filtriramo I ostavljamo samo naziv modela. Smanjujemo broj vrednosti za model sa prethodno 632 na samo 185.
- **Trim** predstavlja nivo kvaliteta opreme automobila. Kako je rec o kategorickom atributu popunjavamo nedostajuce vrednosti **onom vrednoscu koja se NAJCESCE pojavljuje I za to se koristi ova funkcija mode().**
- **SubModel** atribut ima 465 jedinstvenih vrednosti I zbog tog velikog broja atribut se izostvalja iz dalje analize.
- **Color I Transmission** su samo plotovani I iskazan je neki odnos izmedju njih.
- **WheelTypeID** se odnosi na to da li tocak ima **poklopac ili felnu** I ima vrednosti 1 I 2 a **WheelType** se odnosi na samu vrstu tocka. Kako se prvi atribut odnosi iskljucivo na ID **izostavlja se iz dalje analize (WheelTypeID).** *2 je Covers a 1 je Alloy* bilo
  Relativno slican odnos ove dve vrednosti 3376 I 3083. **Kako alloy ima vecu frekvenciju NA se popunjavaju tom vrednoscu.** Dakle, 339 Na vredosti se popunjava alloy (3376 > 3083).
- **VehOdo** se odnosi na predjenu kilometrazu a **Nationality** se odnosi na zemlju porekla automobila, I ima JEDNU NA vrednost koju popunjavamo najfrekventnijom nacionalnoscu, a to je AMERICAN.
- **Crosstab-ovajem MAKE (MARKA) I NATIONALITY (ZEMLJA POREKLA) vidimo da su ova dva atributa VISOKO KORELISANA I jednog mozemo da izbacimo iz dalje analize.**
- **Size I TopThreeAmerican** atributa se dalje analiziraju. Size ima 1 NA vrednost I pripada marki JEEP, pa gledamo koja je najfrekventnija vrednost ovog atributa za JEEP marku I vidimo da je to **Small SUV** I na mesto NA stavljamo upravo to.

- **TopThreeAmerican** atribut ima 1 NA vrednost marke JEEP I model PATRIOT. Opet, crosstabovanjem vidimo da je najcesci za TopThreeAmerican za JEEP upravo CHRYSLER.
- **Medjutim MARKA I TOPTHREEAMERICAN su visoko korelisana te jedan izbacujemo iz dalje analize.**
- Dalje analiziraju kretanje cena po razlicitim osnovama.
- Sve NA vrednosti popunjavamo **medijanom tog atributa (kolone).**
- **PRIMEUNIT I AUCGUART** imaju veliki broj NA vrednosti te se ova dva atributa izostavljaju iz dalje analize.
- **BYRNO I VNZIP1** se odnose na sifre I kodove pa se takodje izostavljaju.
- **VNST** se odnosi na zemlju porekla kupca.
- **Plotovanjem IsBadBuy I VehichleAge vidimo da sto je auto stariji to je kupovina bila gora odluka (logicno).** *Na ovom grafiku se uočava da starost automobila utiče na to da bude klasifikovan kao nekvalitetan. Što je automobil mlađi manja je verovatnoća da će se klasifikovati kao nekvalitetan.*
- **Dakle, izostavljaju se atributi**

```
data = data.drop(columns=['RefId', 'PurchDate', 'Auction', 'VehYear', 'WheelTypeID', 'Nationality', 'TopThreeAmericanName',
                          'PRIMEUNIT', 'AUCGUART', 'BYRNO', 'VNZIP1', 'VNST', 'SubModel', 'Trim'])
```

Ubacujemo dummie vrednosti za Kategoricke atribute (binarne I nebinarne).

```
df_dummies = pd.get_dummies(data, columns = ['Make', 'Model', 'Color', 'Size', 'DriveType'])

pd.set_option('display.max_columns', 200)

df_dummies
```

```
df_dummies = df_dummies.replace({'Transmission' : {'MANUAL':0, 'AUTO':1}})

df_dummies = df_dummies.replace({'WheelType' : {'Covers':0, 'Alloy':1}})
```

Promena tipa izlaza – IsBadBuy

Posto je int izlaz menjamo ga u bool atribut.

```
df_dummies['IsBadBuy'] = df_dummies['IsBadBuy'].astype('bool')
```

Dakle sada nije 1 I 0 vec True I False.