





```
propertylandusetypeid_267.0
heatingorsystemtypeid_24.0
propertycountylandusecode_0100
propertyzoningdesc_LAR3
heatingorsystemtypeid_1.0
propertycountylandusecode_34
fips_6037.0
propertycountylandusecode_122
fips_6037.0
propertylandusetypeid_260.0
propertyzoningdesc_LARD1.5
poolcnt_1.0
heatingorsystemtypeid_6.0
heatingorsystemtypeid_11.0
heatingorsystemtypeid_12.0
propertylandusetypeid_248.0
heatingorsystemtypeid_13.0
heatingorsystemtypeid_18.0
fips_6111.0
heatingorsystemtypeid_2.0
propertyzoningdesc_LAR3
propertyzoningdesc_LARIN
heatingorsystemtypeid_10.0
propertylandusetypeid_264.0
buildingqualitytypeid
heatingorsystemtypeid_20.0
propertycountylandusecode_rare
fips_6059.0
propertylandusetypeid_269.0
propertylandusetypeid_247.0
propertyzoningdesc_rare
unitont_na_flag
propertylandusetypeid_263.0
propertylandusetypeid_31.0
propertyzoningdesc_LAR3
propertycountylandusecode_010C
heatingorsystemtypeid_7.0
propertycountylandusecode_0101
buildingqualitytypeid_na_flag
(37)
```

### Terza Regione 3101

```
In [36]: print_list_info(to_delete[2])

propertylandusetypeid_267.0
heatingorsystemtypeid_24.0
heatingorsystemtypeid_1.0
propertylandusetypeid_265.0
propertylandusetypeid_246.0
propertycountylandusecode_34
propertycountylandusecode_122
fips_6037.0
propertylandusetypeid_260.0
propertyzoningdesc_LARD1.5
poolcnt_1.0
heatingorsystemtypeid_6.0
heatingorsystemtypeid_11.0
heatingorsystemtypeid_12.0
propertylandusetypeid_248.0
heatingorsystemtypeid_13.0
heatingorsystemtypeid_18.0
propertylandusetypeid_275.0
fips_6111.0
propertyzoningdesc_LAR3
propertyzoningdesc_LARIN
heatingorsystemtypeid_10.0
propertylandusetypeid_264.0
heatingorsystemtypeid_20.0
fips_6059.0
propertylandusetypeid_269.0
propertylandusetypeid_247.0
unitont_na_flag
propertylandusetypeid_263.0
propertylandusetypeid_31.0
propertyzoningdesc_LAR3
propertycountylandusecode_010C
roomcnt
fireplacecnt
propertylandusetypeid_261.0
buildingqualitytypeid_na_flag
(36)
```

Sembra che principalmente siano eliminate la maggior parte delle colonne generate dal One-Hot Encoding, come evidenziava anche la prima analisi legate all'allanamento di una foresta

```
In [37]: def remove_column(df, col_names):
         dt_drop(col_names, axis=1, inplace=True)
         return df
```

```
In [38]: dimensionality(y=True)
```

```
X_trainA: (26819, 69)
X_valA:    (9006, 69)
X_testA:  (9085, 69)
y_trainA: (26819, 1)
y_valA:   (9006, 1)
y_testA:  (9085, 1)

X_trainB: (8119, 69)
X_valB:   (2658, 69)
X_testB:  (2606, 69)
y_trainB: (8119, 1)
y_valB:   (2658, 1)
y_testB:  (2606, 1)

X_trainC: (64771, 69)
X_valC:   (21908, 69)
X_testC:  (21876, 69)
y_trainC: (64771, 1)
y_valC:   (21908, 1)
y_testC:  (21876, 1)
```

```
In [39]: for i in range(3):
         for X in [X_train[i], X_val[i], X_test[i]]:
             X = remove_column(X, to_delete[i])
```

```
In [40]: dimensionality(y=True)
```

```
X_trainA: (26819, 55)
X_valA:    (9006, 55)
X_testA:  (9085, 55)
y_trainA: (26819, 1)
y_valA:   (9006, 1)
y_testA:  (9085, 1)

X_trainB: (8119, 32)
X_valB:   (2658, 32)
X_testB:  (2606, 32)
y_trainB: (8119, 1)
y_valB:   (2658, 1)
y_testB:  (2606, 1)

X_trainC: (64771, 33)
X_valC:   (21908, 33)
X_testC:  (21876, 33)
y_trainC: (64771, 1)
y_valC:   (21908, 1)
y_testC:  (21876, 1)
```

### Analisi dei risultati

Il numero di feature selezionate è analogo a quello ottenuto dall'increasing RMSE: una cinquantina di feature per la prima regione e una trentina per la seconda e la terza. Analisi delle feature comuni.

Feature selezionate comuni a tutte e tre le regioni:

```
In [41]: common = set(X_train[0].columns)

         for i in range(1,3):
             common = common.intersection(set(X_train[i].columns))
         common = list(common)

         print_list_info(common)

taxvalusedollarcnt
latitude
strutturetaxvalusedollarcnt
tax_prop
finishedsquarefeet12
rawconsutractiondblock
regionidcity
yearbuilt
bedroomcnt
lotsizeareafeet
propertylandusetypeid_246.0
assessmentyear_2015.0
taxamount
longitude
calculatedfinishedsquarefeet
unitont
living_area_prop
bathroomcnt
neighborhood_mean_price
int_transactiondate
landtaxvalusedollarcnt
calculatedbathnbr
period_mean_price
tax_ratio
regionidzip
(25)
```

Le feature più importanti sono le stesse individuate anche dal ranking della foresta: `longitude` e `latitude` per l'area geografica, `int_transactiondate` per il periodo di vendita e tutte le feature legate alle tasse, che sono probabilmente capaci di considerare in un'unica variabile diversi aspetti dell'abitazione.

Tutte le colonne aggiunte in fase di preparazione (`int_transactiondate`, `period_mean_price`, `neighborhood_mean_price`, `tax_prop` e `living_area_prop`) sono state selezionate.

### Scrittura dei dati

Salvo i dataset su cui è stata fatta la selezione delle feature su una specifica cartella.

```
In [42]: dir_name = 'selezione'

         for i in range(3):
             X_train[i].to_csv( dir_name + f'/{X_train[region_names[i]].csv', index=False)
             X_val [i].to_csv( dir_name + f'/{X_val[ region_names[i]].csv', index=False)
             X_test [i].to_csv( dir_name + f'/{X_test[ region_names[i]].csv', index=False)
             y_train[i].to_csv( dir_name + f'/{y_train[region_names[i]].csv', index=False)
             y_val [i].to_csv( dir_name + f'/{y_val[ region_names[i]].csv', index=False)
             y_test [i].to_csv( dir_name + f'/{y_test[ region_names[i]].csv', index=False)
```