SUNS Zadanie 2

Marek Dráb 97757

Načítanie dát, normalizácia a škálovanie

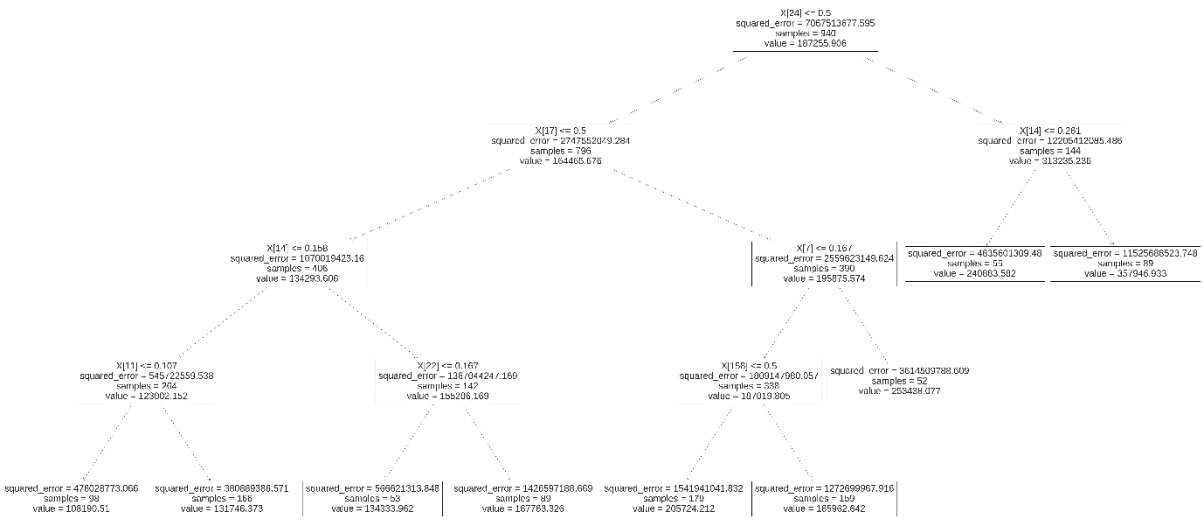
Boli použité súbory train_dummy.csv a test_dummy.csv. Keďže na nich už bol použitý one hot encoding, stačí ich škálovať. Na škálovanie bol použitý MinMaxScaler.

Trénovanie

Rozhodovací strom

Zobrazenie jedného z výsledkov

- max_depth = 4
- min_samples_leaf = 50



Obrázok 1 Rozhodovací strom

Najlepší výsledok

Parametre na výber:

- max_depth: <2,10>
- min_samples_leaf: <2,10>
- criterion:
  - squared_error
  - friedman_mse
  - absolute_error
  - poisson

Najlepšie parametre podľa GridSearchCV:

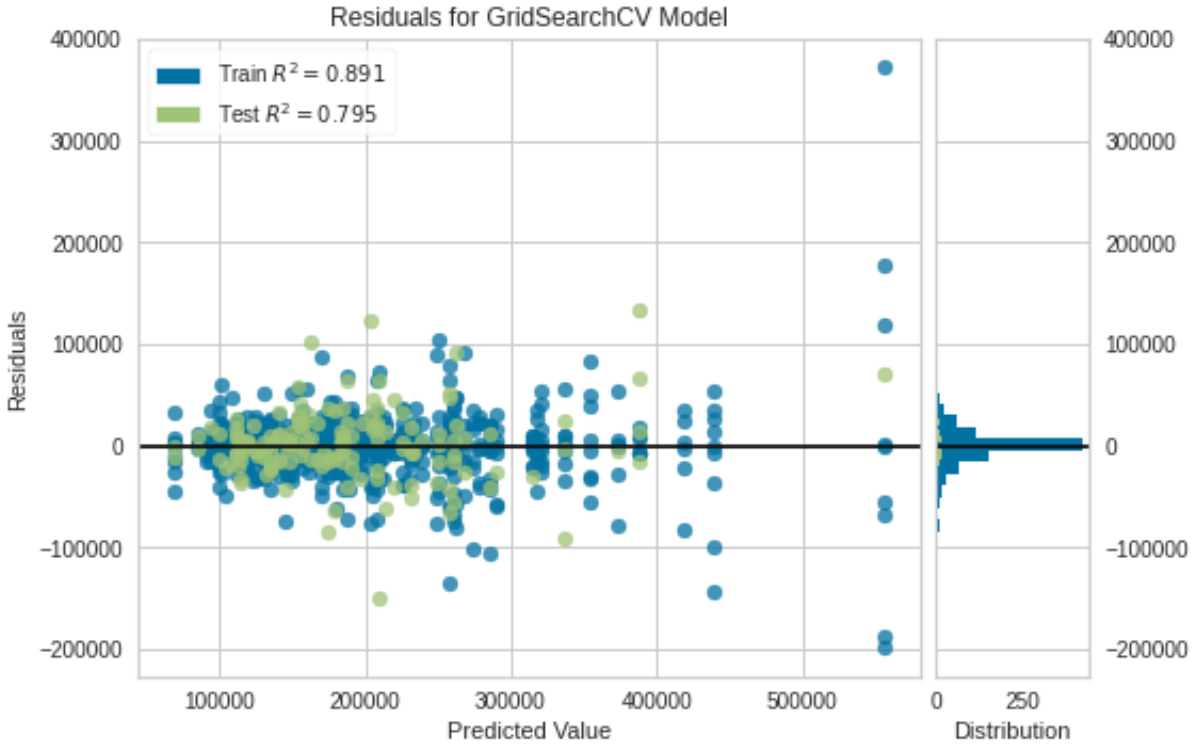
- criterion = absolute_error
- max_depth = 9

- min_samples_leaf = 6

R2 = 0.79452

MSE = 1222094176.6931818

Reziduály



Obrázok 2 Reziduály rozhodovacieho stromu

SVM

Parametre a trénovanie

Nastavenie parametrov:

- kernel:
  - rbf
  - linear
- gamma: od 10e-2 do 10e3
- C: od 10 do 10e5

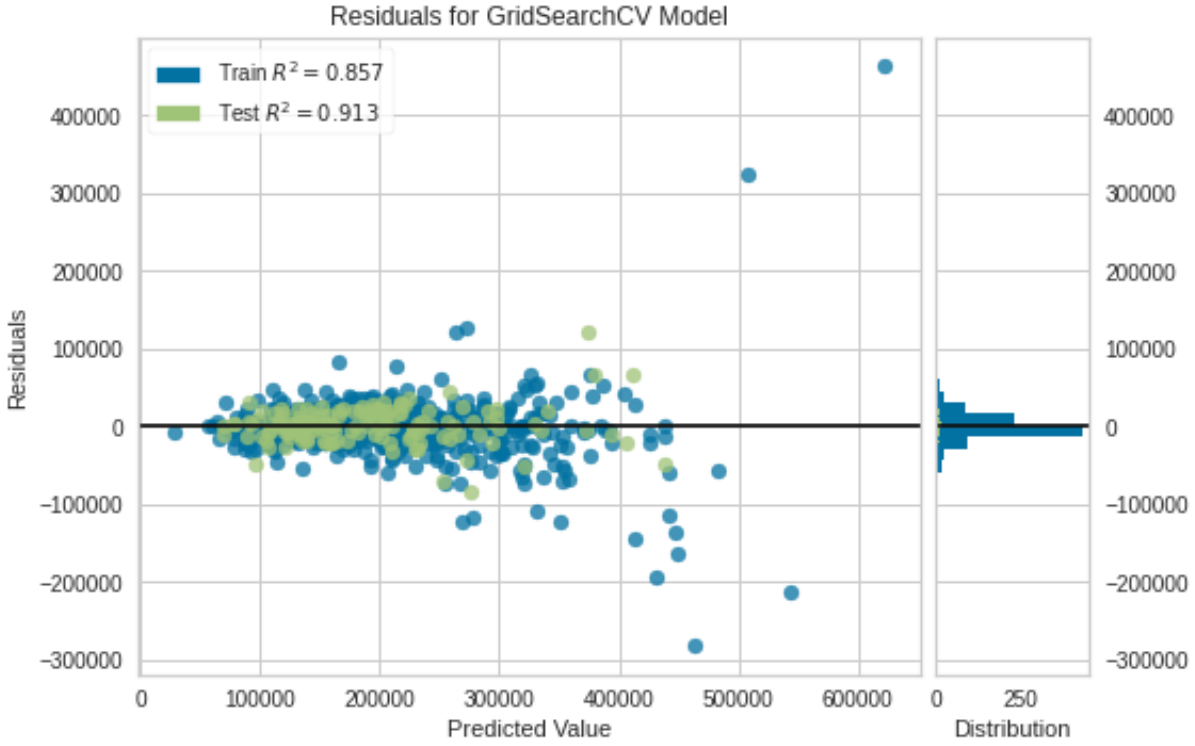
Najlepšie parametre podľa GridSearchCV:

- C = 10000
- gamma = 1000
- kernel = linear

R2 = 0.91307

MSE = 517037861.708311

Reziduály



Obrázok 3 Reziduály SVM

RandomForestRegressor

Parametre a trénovanie

Nastavenie parametrov:

- max_depth: od 1 do 10
- n_estimators: od 10 do 200 s krokom 20
- max_features: sqrt, log2
- criterion:
    - squared_error
    - absolute_error
    - poisson

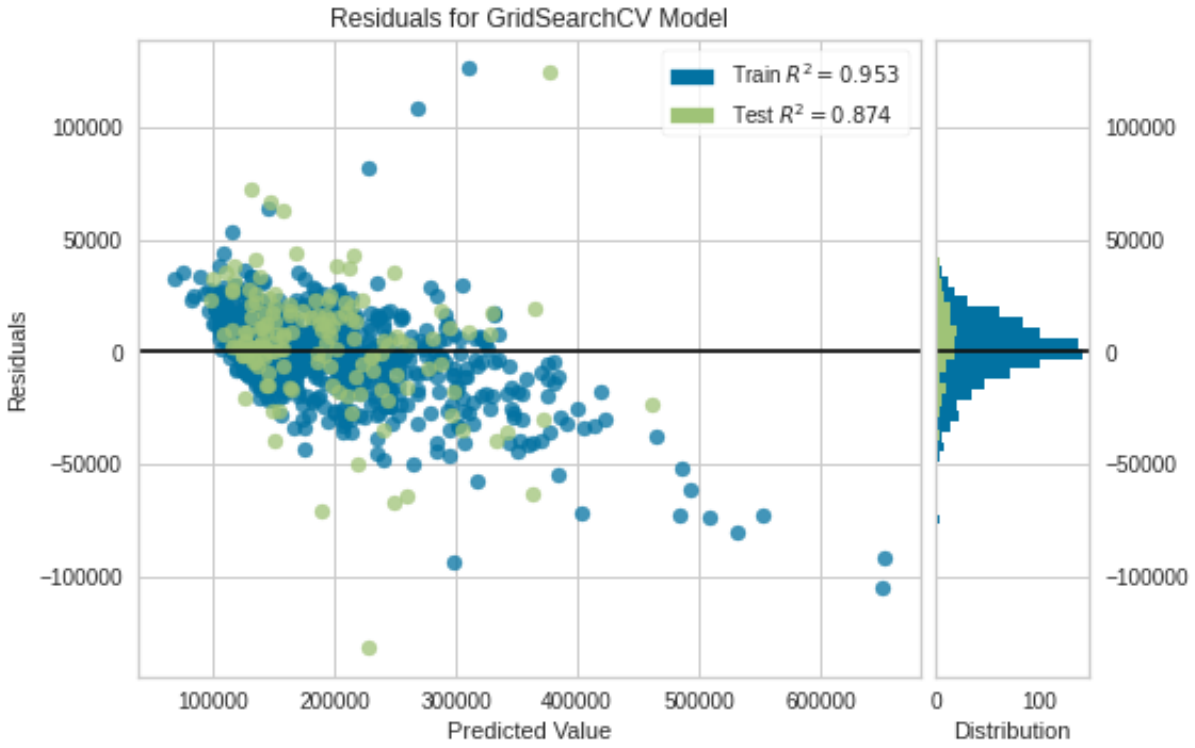
Najlepšie parametre podľa GridSearchCV:

- max_depth = 9
- n_estimators = 190
- max_features = sqrt
- criterion = squared_error

R2 = 0.87446

MSE = 746669833.2717698

Reziduály



Obrázok 4 Reziduály RandomForestRegressoru

Porovnanie

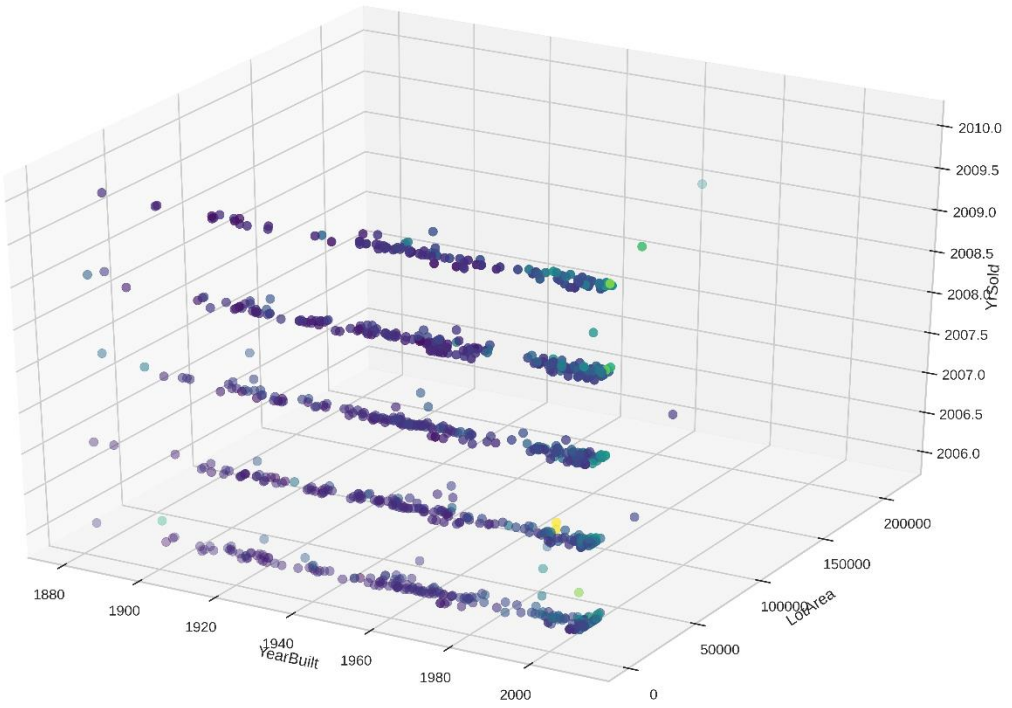
Z grafov reziduálov je možné vidieť najlepšie hodnoty pri SVM. Väčšina hodnôt sa drží blízko osi x, až na pár prípadov a jednu hodnotu nad 100000. Pri ostatných dvoch modeloch sú hodnoty viac roztrúsené v rámci daného intervalu.

Redukcia dimenzií

Redukcia bolo vykonaná pomocou PCA.

Príznaky z pôvodnej databázy

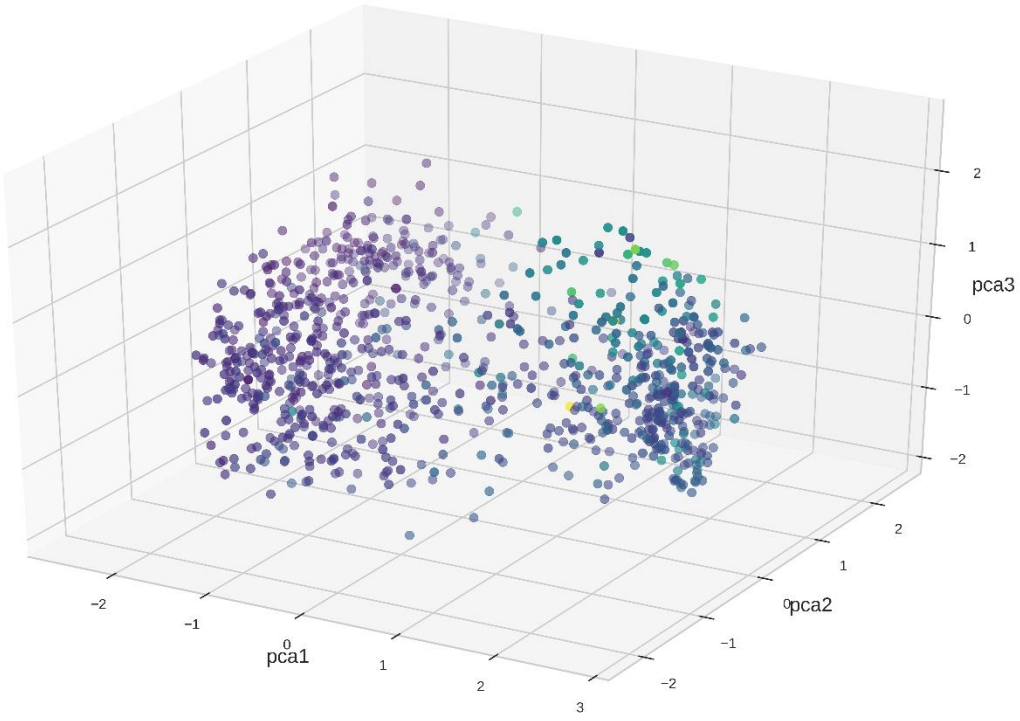
- YearBuilt
- LotArea
- YrSold



Obrázok 5 Graf závislosti ceny od roku stavby, veľkosti pozemku a roku predaja

PCA

Príznaky z Obrázku 5 po redukcii dimenzie.



Obrázok 6 Graf redukovaných dimenzií

Trénovanie na redukovaných dimenziách

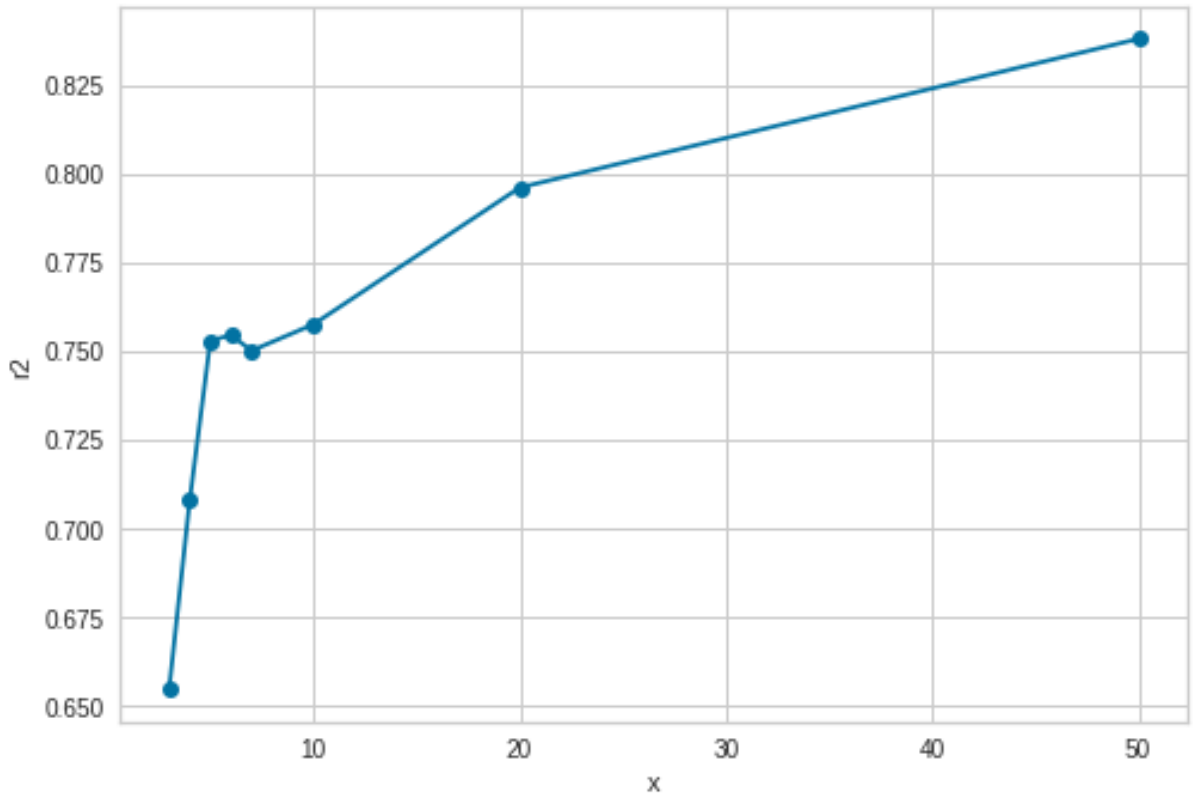
Z predošlých trénovaní bolo zistené, že najúspešnejším modelom bolo SVM. To bolo použité aj na ďalšie trénovanie na množinách so zredukovanými dimenziami.

Parametre boli:

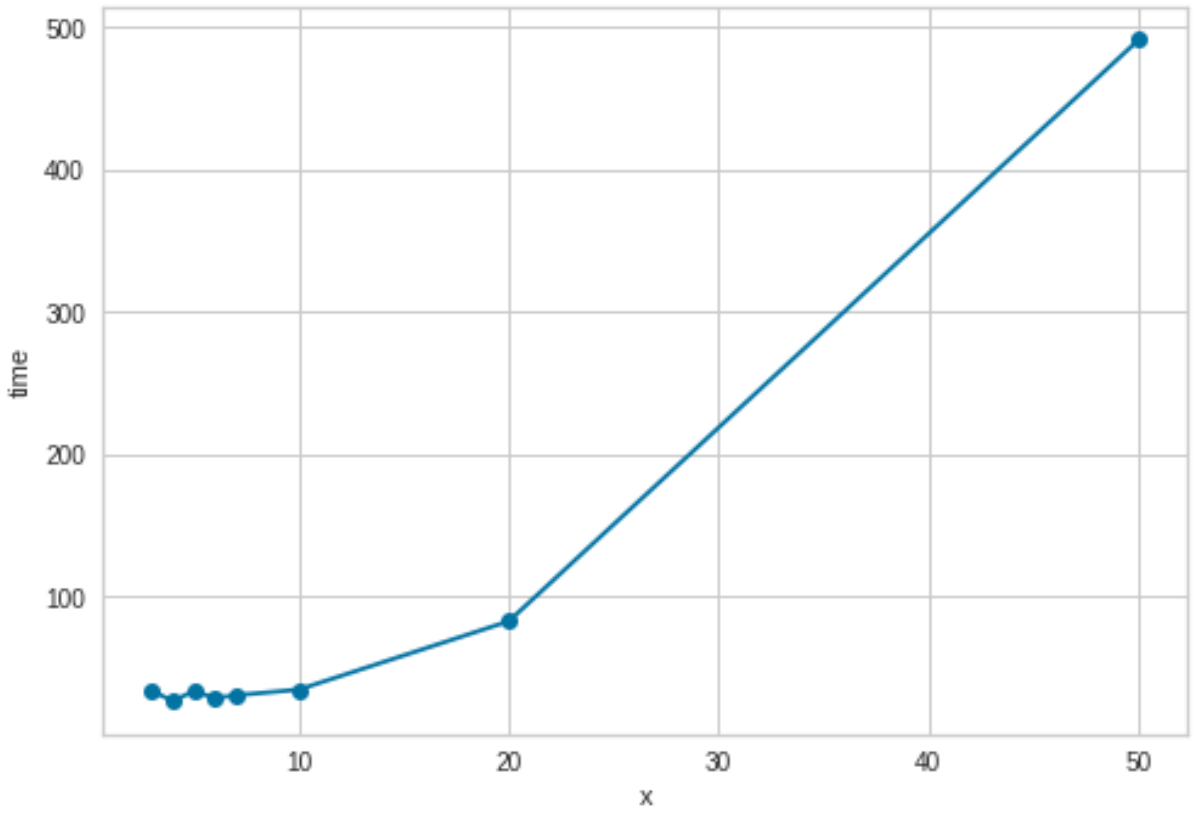
- kernel:
  - rbf
  - linear
- gamma: od 10e-3 do 10e3
- C: od 10 do 10e6

Tabuľka 1 Výsledky trénovaní SVM na redukovaných dimenziách

Počet dimenzií	Kernel	Gamma	C	R2	Čas
3	rbf	0.1	1000000	0.65483	0:34
4	rbf	0.1	1000000	0.7081	0:27
5	rbf	0.1	100000	0.75301	0:34
6	rbf	0.1	100000	0.75456	0:29
7	rbf	0.1	100000	0.75018	0:31
10	rbf	0.1	100000	0.75761	0:35
20	rbf	0.1	100000	0.79614	1:23
50	rbf	0.01	1000000	0.83085	8:11



Obrázok 7 Závislosť R2 od veľkosti dimenzie

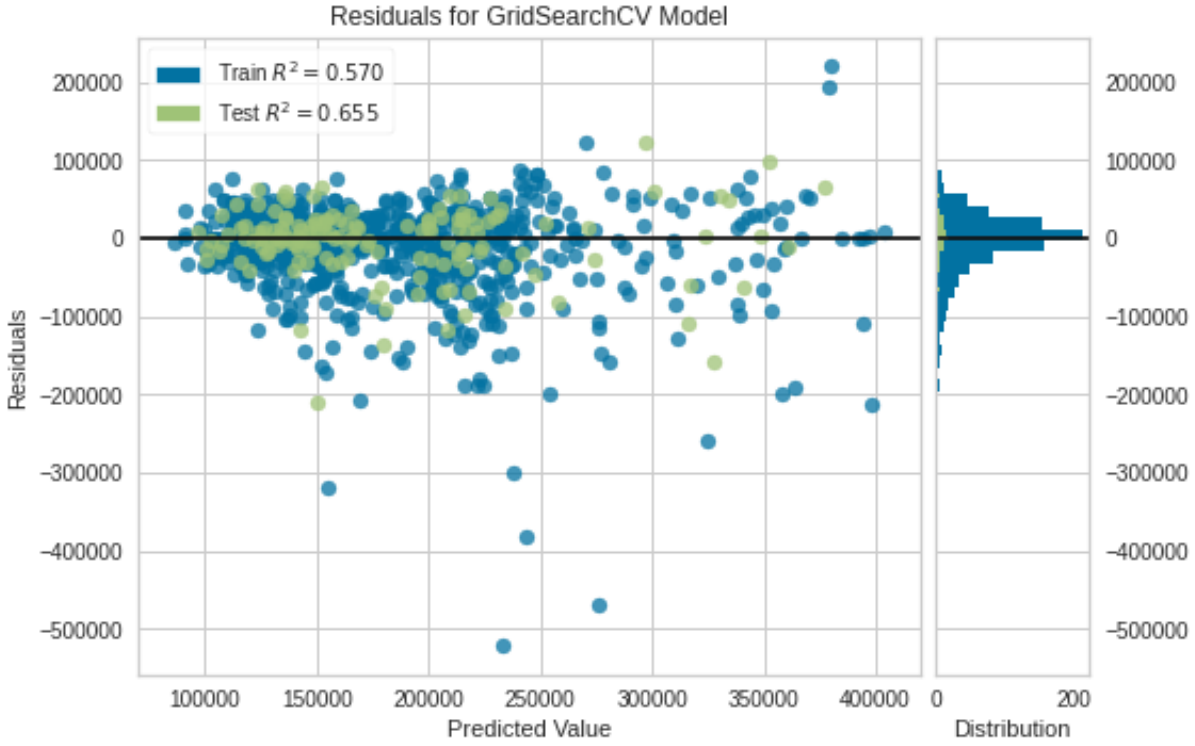


Obrázok 8 Závislosť času od veľkosti dimenzie

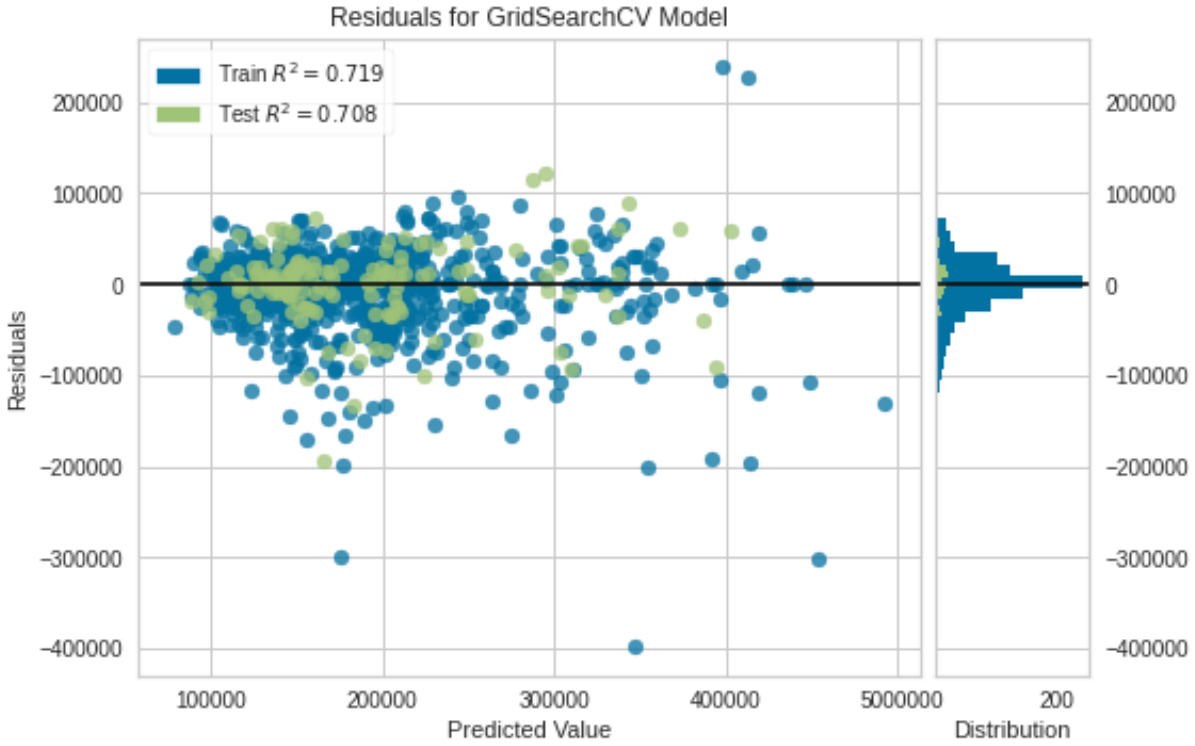
Z grafu úspešnosti v závislosti od veľkosti dimenzie je možné vidieť, že výrazne stúpla už pri zmene z 3 na 5 dimenzií. Čas pri týchto dvoch hodnotách bol veľmi podobný (zaokrúhlené na s). Taktiež je možné vidieť, že úspešnosť stúpa s počtom dimenzií. Podľa grafu sa dá povedať, že úspešnosť by sa mala blížiť hodnote 1, resp. v našom prípade k hodnote 0.91, ktorú sme dostali pri trénovaní SVM.

Čas rastie exponenciálne. Pri zväčšení dimenzie z 20 na 50, čo je 2,5 násobok, vzrástol takmer 6 násobne.

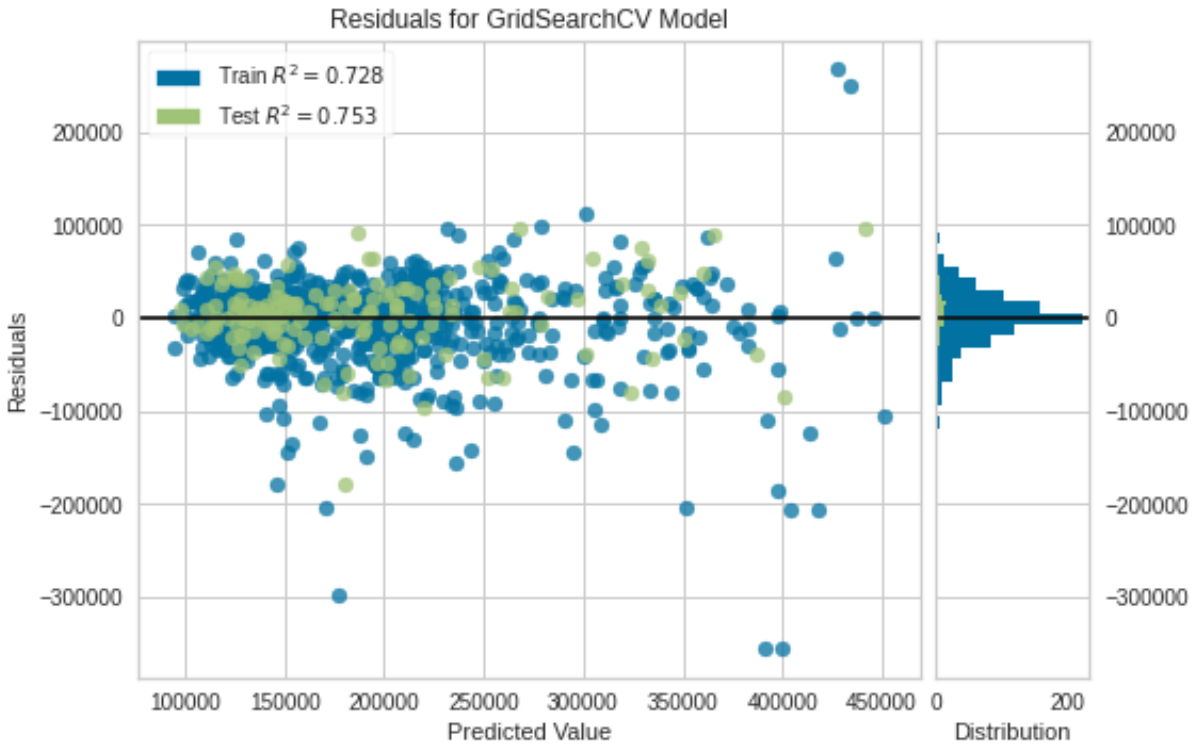
Reziduály pre zredukované dáta



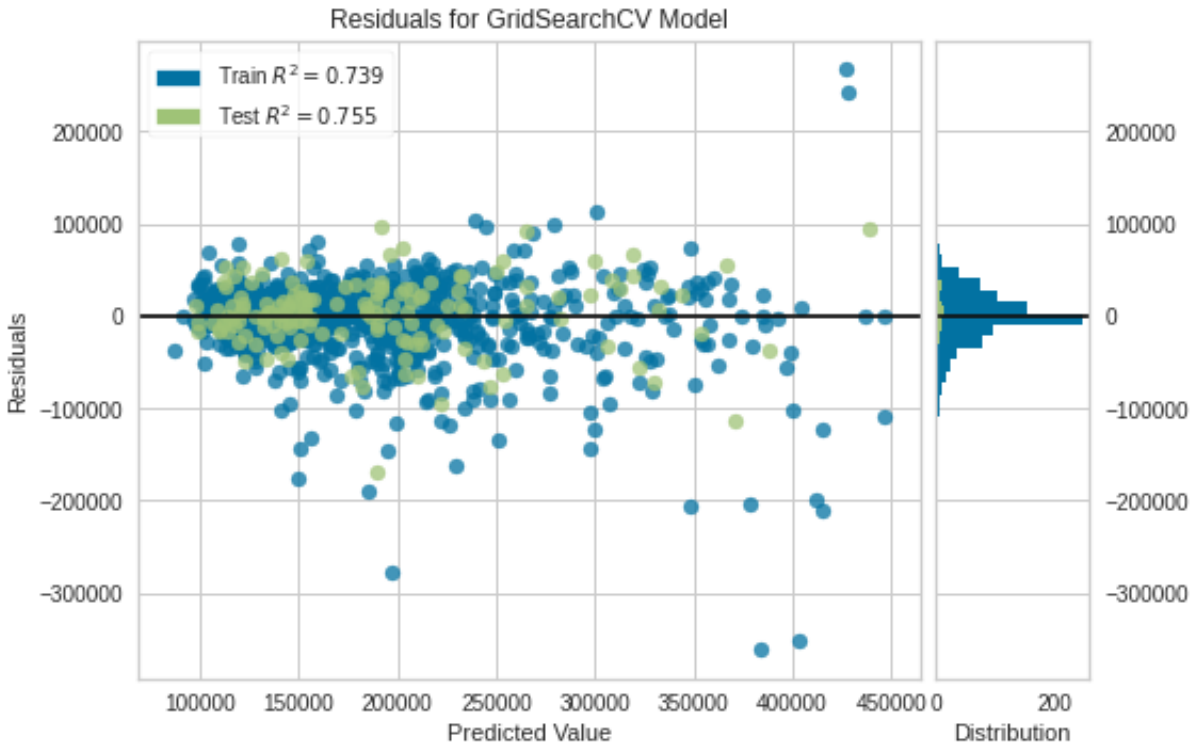
Obrázok 9 Reziduály 3 dimenzií



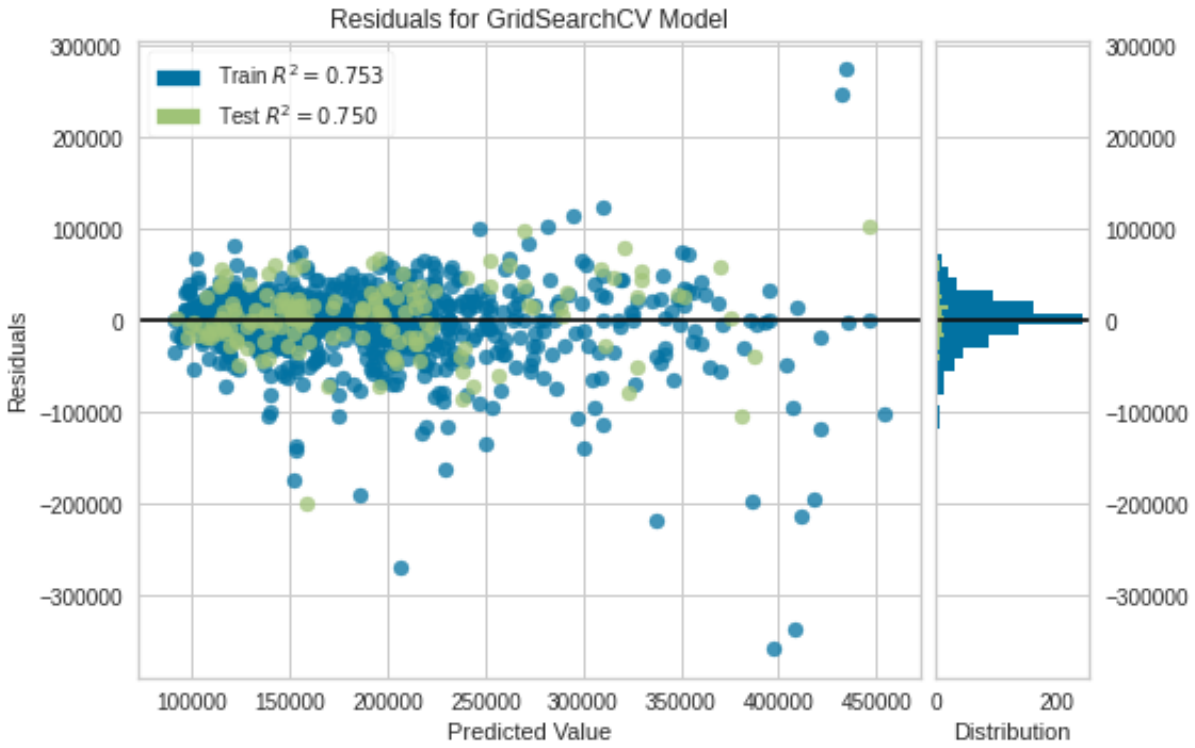
Obrázok 10 Reziduály 4 dimenzií



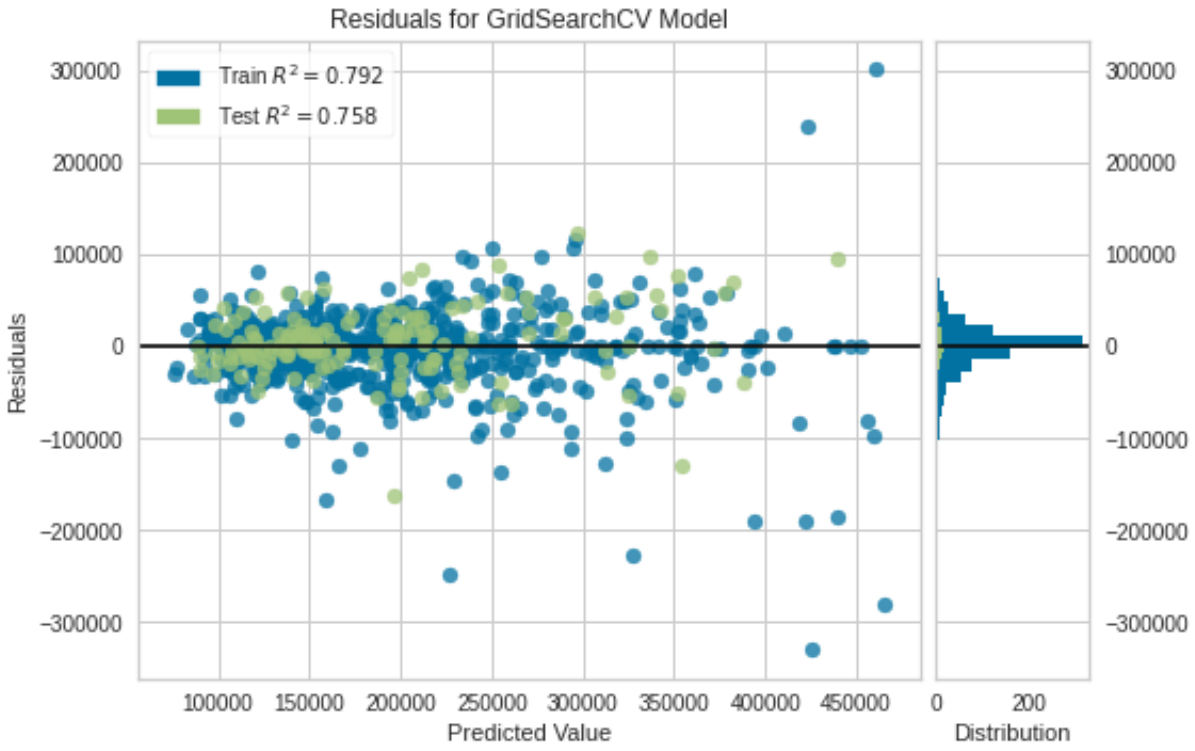
Obrázok 11 Reziduály 5 dimenzií



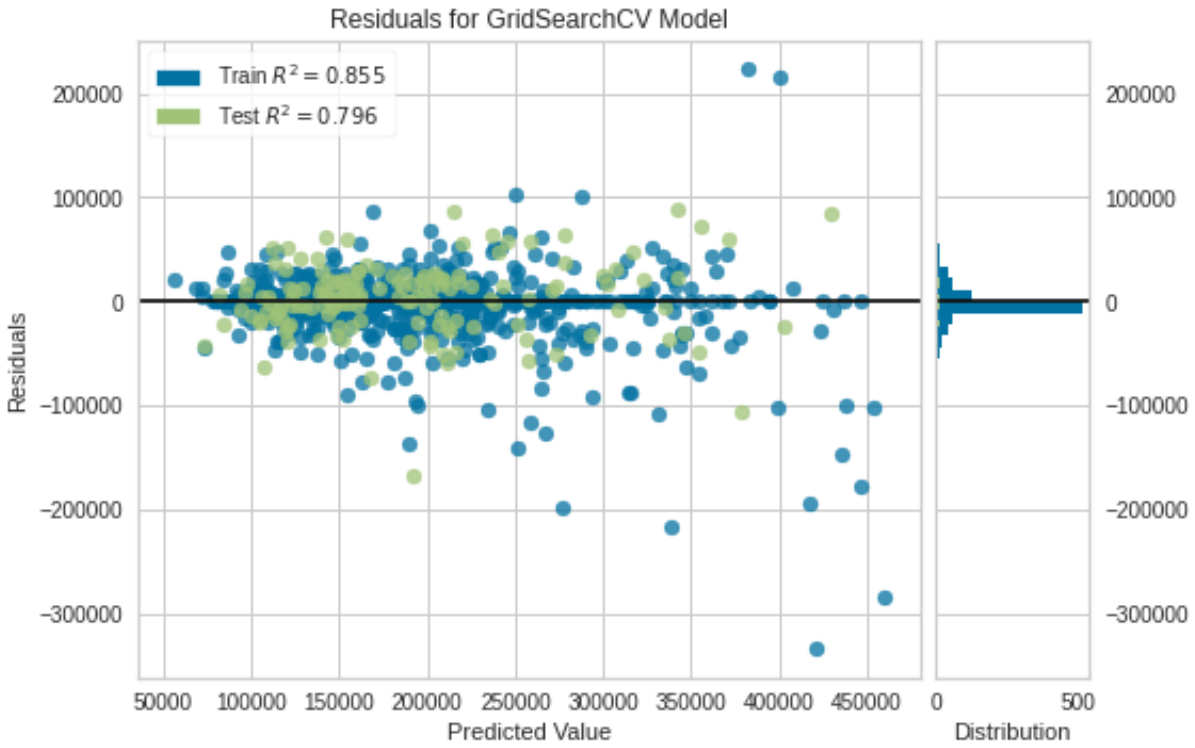
Obrázok 12 Reziduály 6 dimenzií



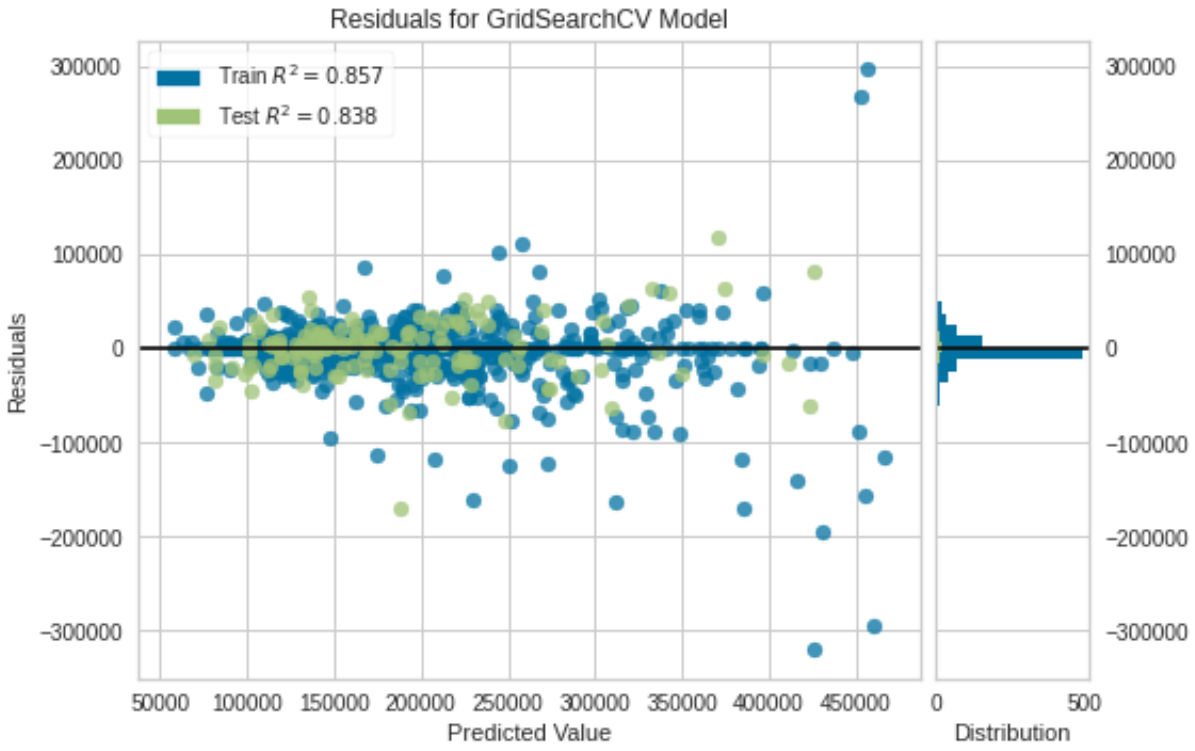
Obrázok 13 Reziduály 7 dimenzií



Obrázok 14 Reziduály 10 dimenzií



Obrázok 15 Reziduály 20 dimenzií



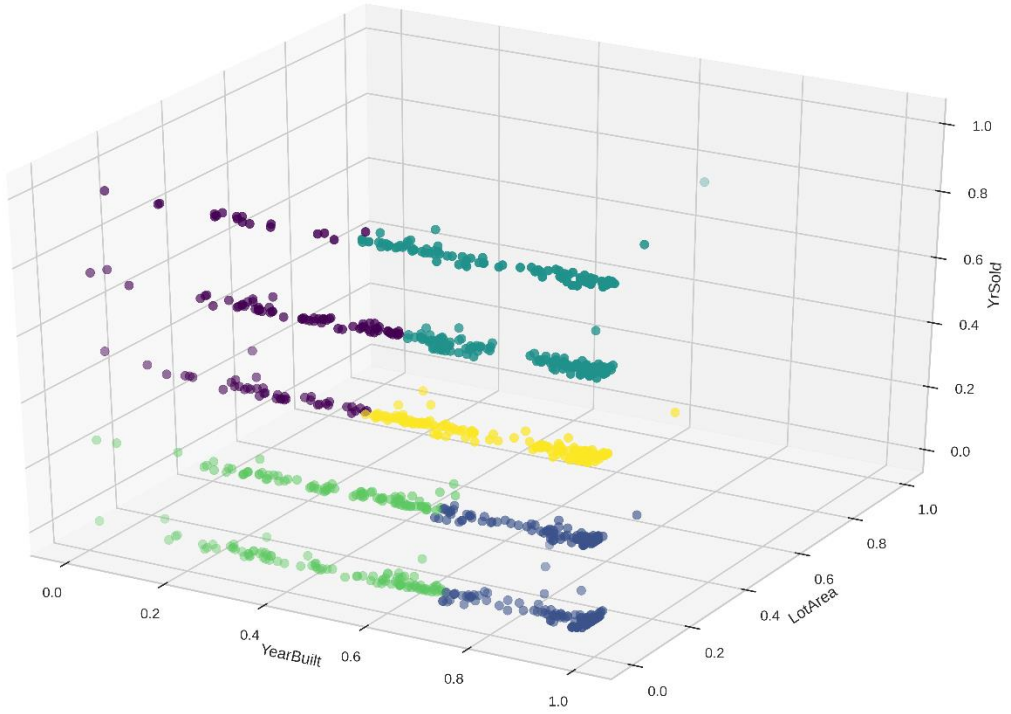
Obrázok 16 Reziduály 50 dimenzií

Porovnanie

Z grafov je možné vidieť, ako sa zvyšovala hustota hodnôt v intervale -100000 až 100000 úmerne k zvýšeniu dimenzie.

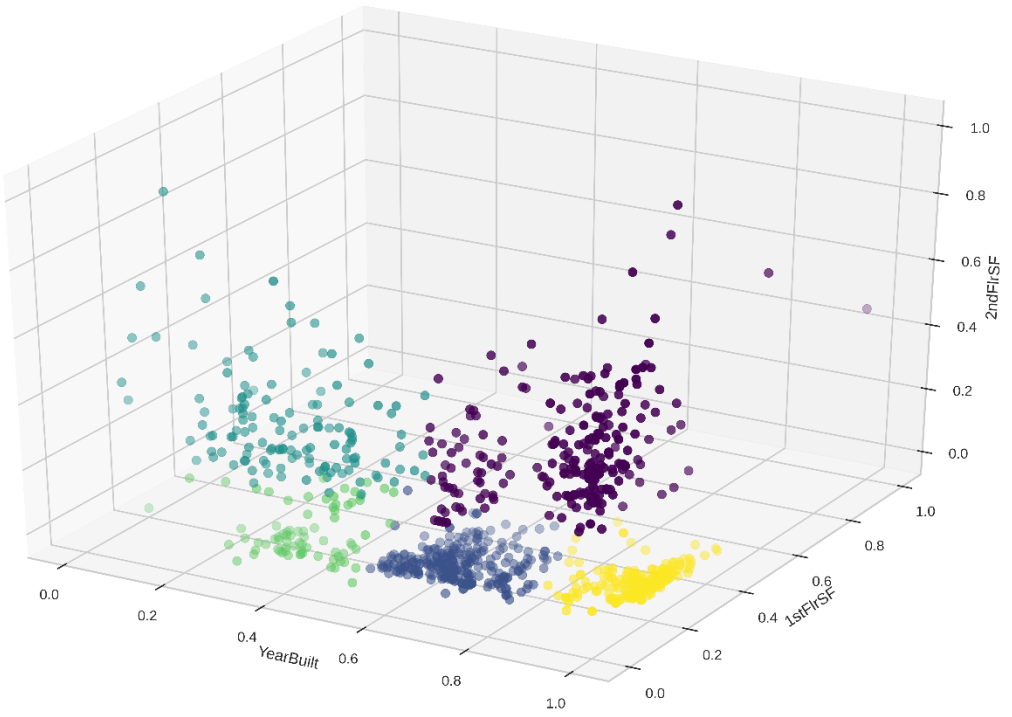
Zhlukovanie dát

- Pomocou kmeans
- 5 clusterov



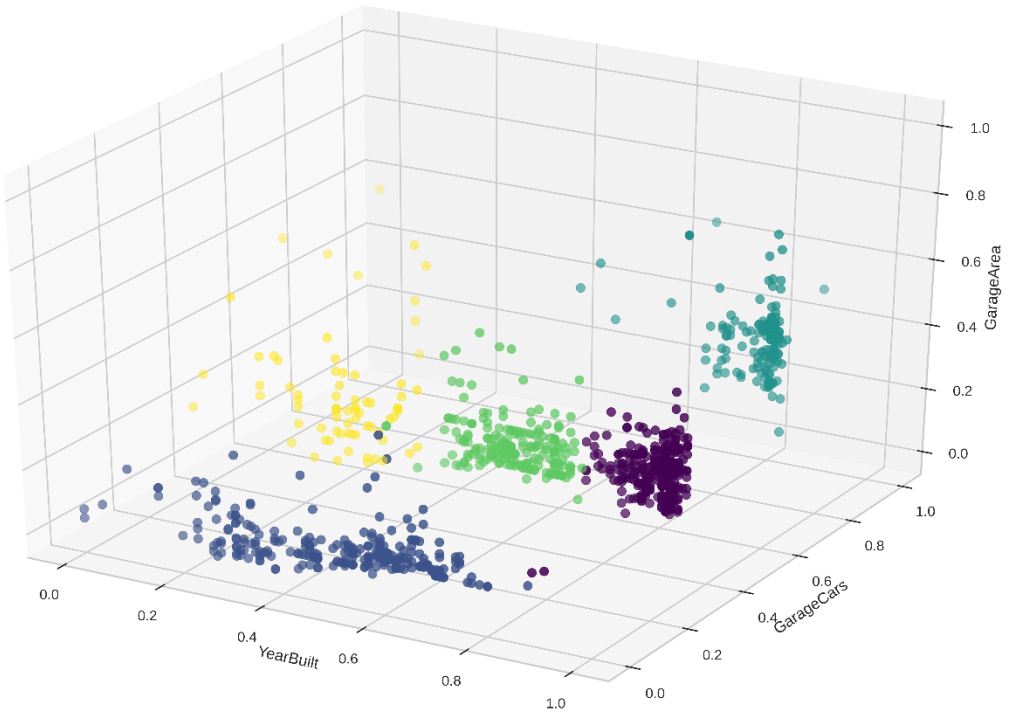
Obrázok 17

Zobrazenie zhlukovaných dát pre YearBuilt, LotArea a YrSold.



Obrázok 18

Zobrazenie zhlukovaných dát pre YearBuilt, 1stFlrSF a 2ndFlrSF.



Obrázok 19

Zobrazenie zhlukovaných dát pre YearBuilt, GarageCars a GarageArea.