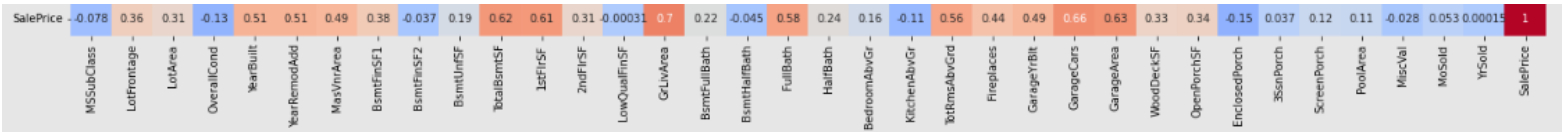
I-SUNS: Zadanie č.2

Neurónové siete

Tomáš Minárik

Príprava dát

- Najprv sme načítali dáta train.csv, train_dummy.csv a test_dummy.csv.
- Z dát načítaných z train.csv sme vytvorili korelačnú maticu.



- Z korelačnej matice sme zistili že s cenou najviac koreluje GrLivArea, GarageCars a GarageArea.

```
YearBuilt :  0.51
YearRemodAdd :  0.51
TotalBsmtSF :  0.62
1stFlrSF :  0.61
GrLivArea :  0.7
FullBath :  0.58
TotRmsAbvGrd :  0.56
GarageCars :  0.66
GarageArea :  0.63
SalePrice :  1.0
```

- Ďalej pracujeme s dátami z train_dummy.csv a test_dummy.csv.
- Dáta sme škálovali pomocou StandardScalera.

```
MSSubClass       55.127660
LotFrontage      70.706383
LotArea       10257.188298
OverallCond       5.579787
YearBuilt      1972.154255
YearRemodAdd   1985.639362
MasVnrArea      111.120213
BsmtFinSF1      451.493617
BsmtFinSF2       49.026596
BsmtUnfSF       601.948936
dtype: float64
```

```
MSSubClass     1.096050e-16
LotFrontage   -2.947996e-16
LotArea       -3.401534e-17
OverallCond    1.908639e-16
YearBuilt      1.557147e-15
YearRemodAdd   7.294401e-16
MasVnrArea     2.078715e-17
BsmtFinSF1     4.157431e-17
BsmtFinSF2     0.000000e+00
BsmtUnfSF      8.314862e-17
dtype: float64
```

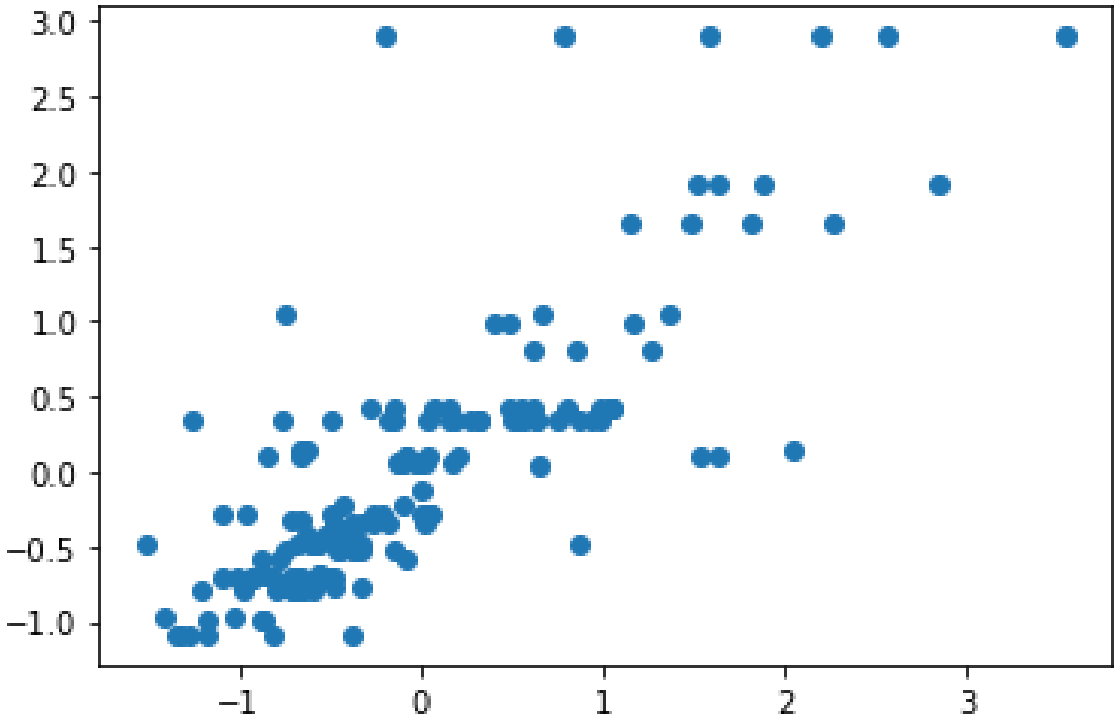
Trénovanie

1. Rozhodovací strom

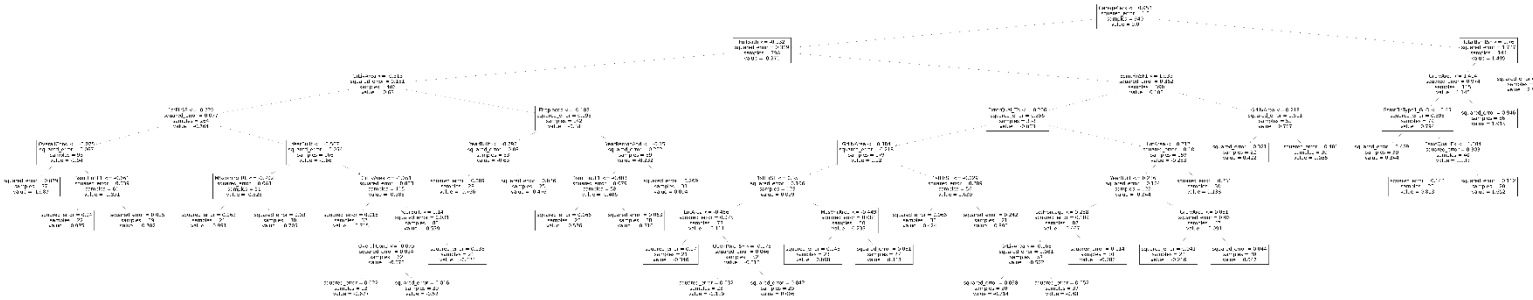
- Parametre použité na trénovanie

```
Fitting 5 folds for each of 81 candidates, totalling 405 fits
GridSearchCV(cv=5, estimator=DecisionTreeRegressor(),
             param_grid={'max_depth': [2, 6, 8], 'max_leaf_nodes': [5, 20, 100],
                         'min_samples_leaf': [20, 40, 100],
                         'min_samples_split': [10, 20, 40]},
             scoring='r2', verbose=1)
```

- Najlepšie výsledky dosiahol strom s hyperparametrami:
  - max_depth=8
  - max_leaf_nodes=100
  - min_samples_leaf=20
  - min_samples_split=20
- Tento strom dosiahol skóre 0.708 na trénovacích dátach.
- Na testovacích dátach dosiahol skóre 0.603 a MSE 0.334



- Na grafe môžeme vidieť že strom dosahuje pomerne dobré výsledky pri nižších hodnotách keďže tie obsahovali najviac záznamov a najhoršie výsledky pri vysokých hodnotách.

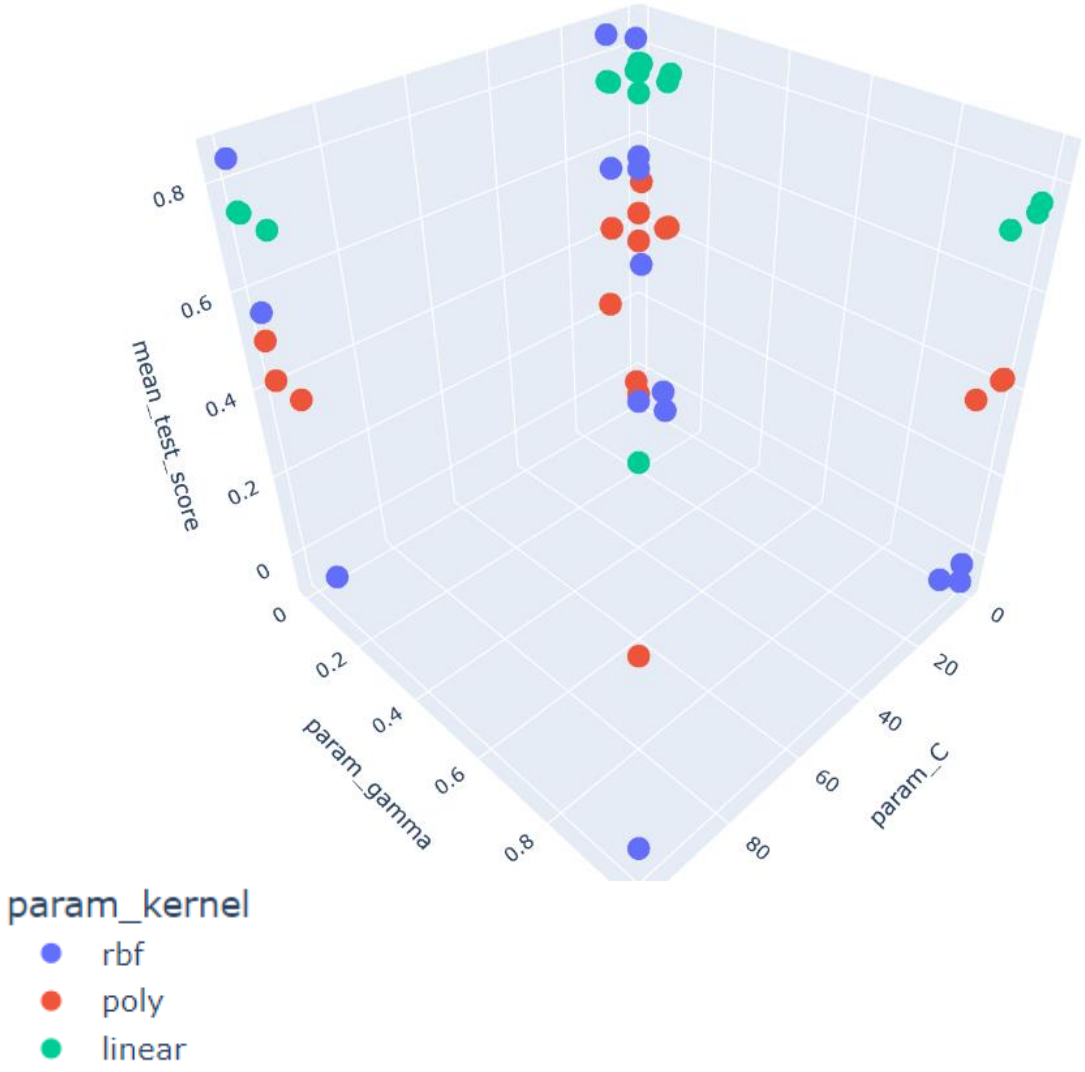


2. SVM

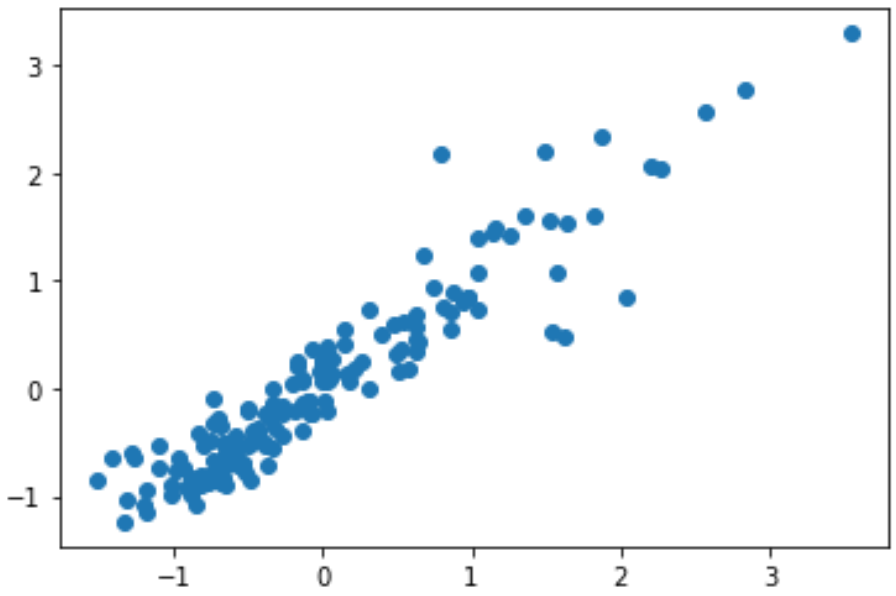
- Parametre použité na trénovanie.

```
{'C': [0.1,1, 10, 100], 'gamma': [1,0.1,0.01,0.001],'kernel': ['rbf', 'poly', 'linear']}
```

- Výsledky gridsearchu.



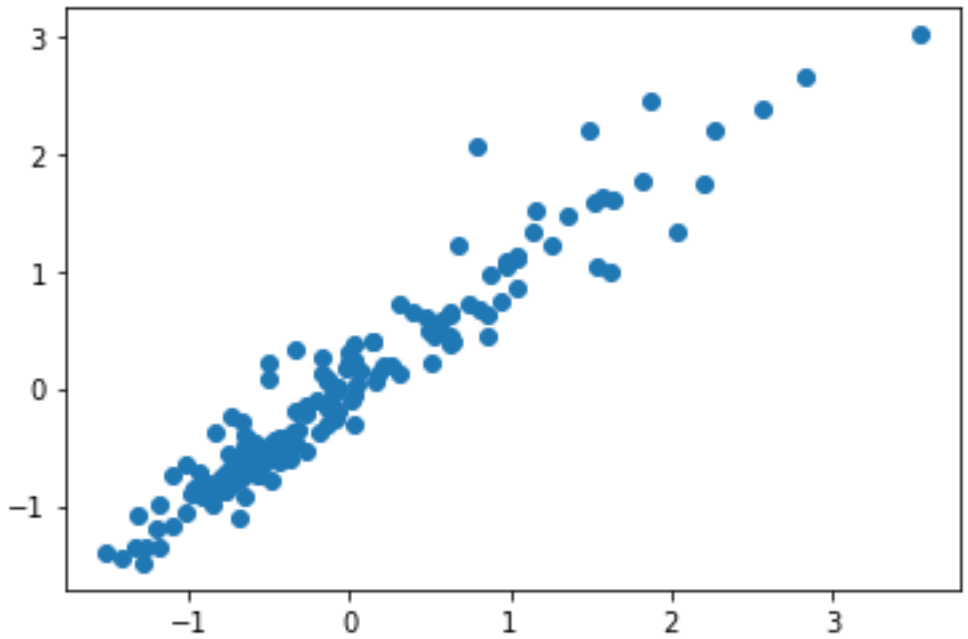
- Najlepšie výsledky dosiahol model s hyperparametrami
  - C=10
  - Gamma=0.001
  - kernel=rbf
- Tento model dosiahol skóre 0.847 na trénovacích dátach.
- Na testovacích dátach dosiahol skóre 0.881 a MSE 0.1.



- Vo výsledkoch gridsearchu si môžeme všimnúť že pre kernel=linear je skóre vždy okolo 0.75.
- Pri cross validácií kernel=linear vždy dosahoval v jednom prípade skóre 0.897 a pre to sme sa rozhodli vyskúšať aj tento model na testovacích dátach.

```
[CV 1/3] END .....C=0.1, gamma=1, kernel=linear;, score=0.897 total time=   1.1s
[CV 2/3] END .....C=0.1, gamma=1, kernel=linear;, score=0.610 total time=   0.7s
[CV 3/3] END .....C=0.1, gamma=1, kernel=linear;, score=0.795 total time=   0.9s
```

- Tento model na testovacích dátach dosiahol skóre až 0.921 a MSE 0.066.

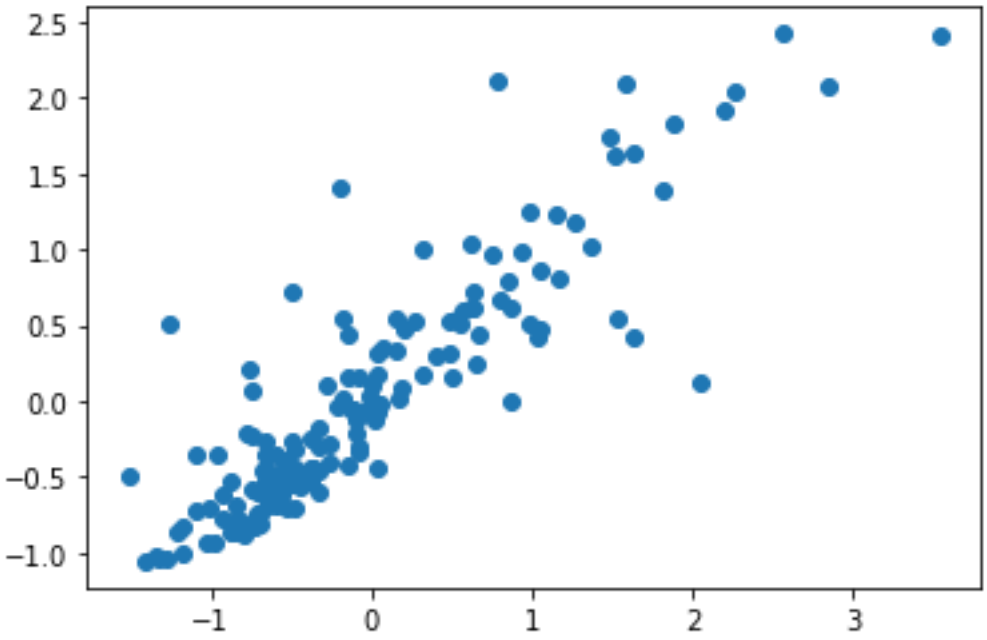


3. RandomForest

- Na trénovanie boli použité rovnaké parametre ako pri rozhodovacom strome len s pridaním n_estimators.

```
{'n_estimators' : [20, 40, 60, 80, 100],
 "min_samples_split": [10, 20, 40],
 "max_depth": [2, 6, 8],
 "min_samples_leaf": [20, 40, 100],
 "max_leaf_nodes": [5, 20, 100],
 }
```

- Najlepšie výsledky dosiahol model s hyperparametrami:
  - max_depth=8
  - max_leaf_nodes=100
  - min_samples_leaf=20
  - min_samples_split=10
  - n_estimators=100
- Tento model dosiahol skóre 0.759 na trénovacích dátach.
- Na testovacích dátach dosiahol skóre 0.773 a MSE 0.191.



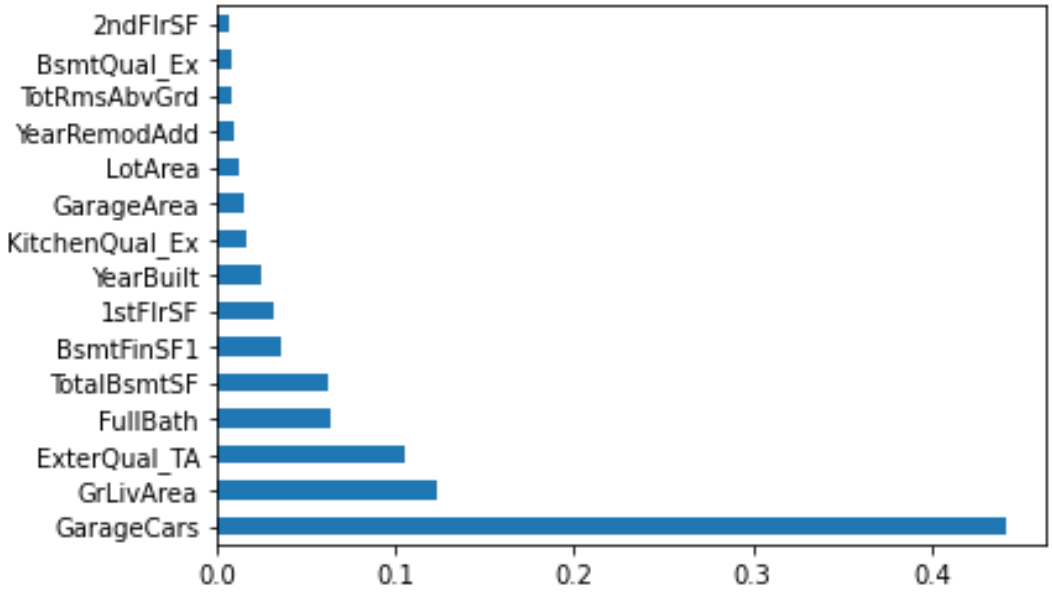
- Najdôležitejšie vstupné parametre.

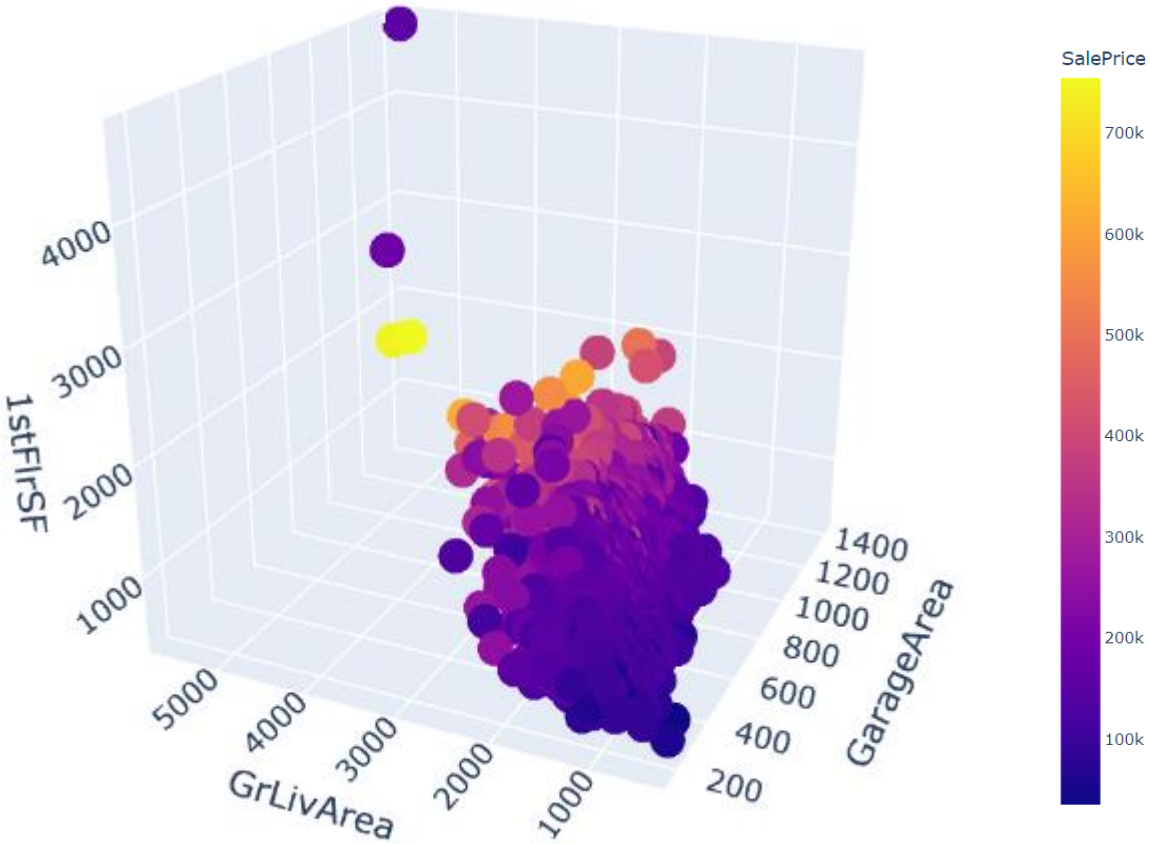
4. Záver

- Najlepší model bol SVM, ktorý pre kernel=rbf dosiahol skóre 0.881 a pre kernel=linear až 0.921.

Redukcia dimenzie

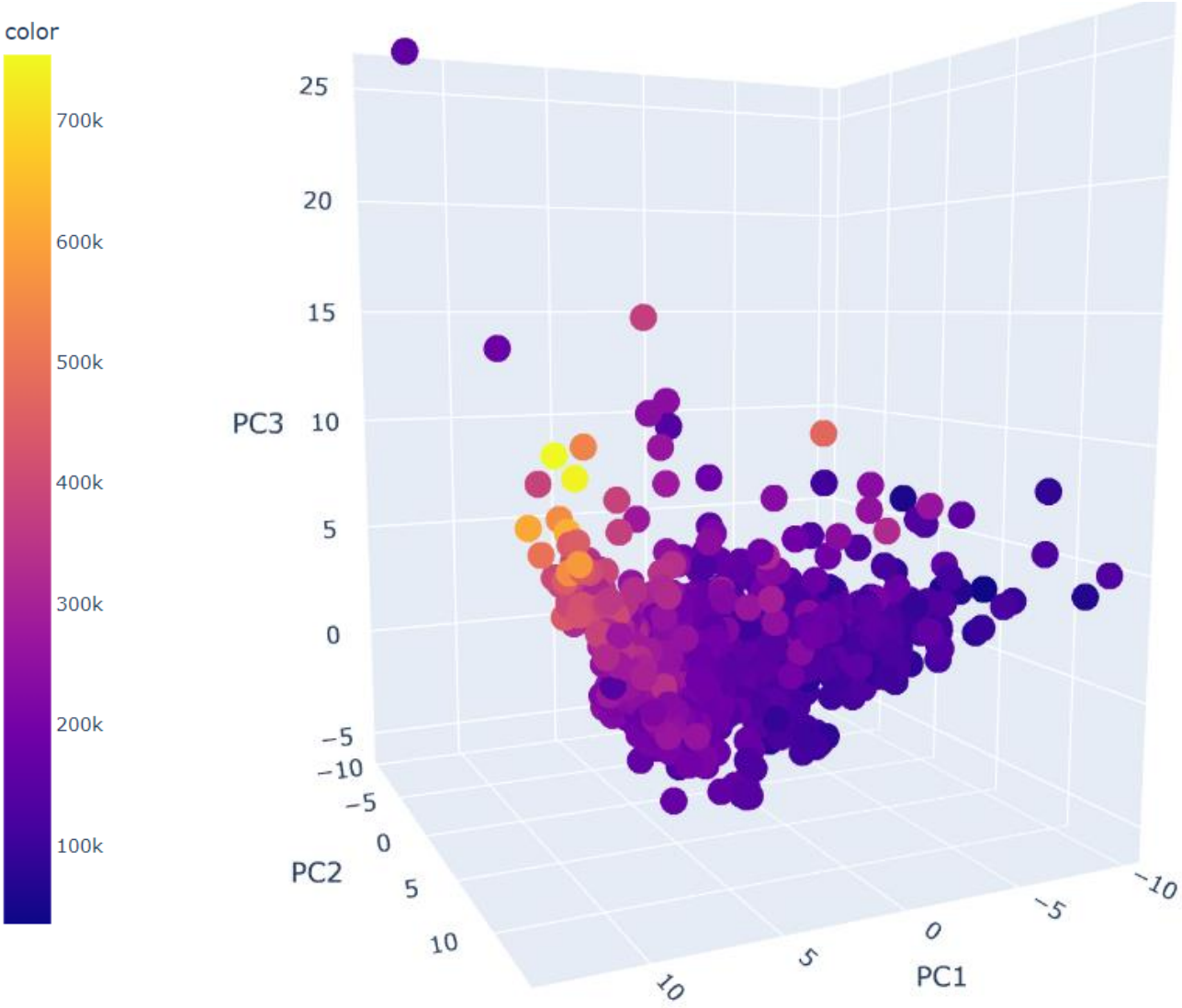
1. Závislosť ceny od GrLivArea, GarageArea a 1stFlrSF





- Keďže parametre do grafu sme vyberali podľa korelačnej matice tak môžeme jasne vidieť ako jedným smerom narastá cena.
- Môžeme vidieť 2 výnimky ktoré majú najväčšiu rozlohu ale nemajú vysokú cenu. Ide o záznamy ktoré v danom čase ešte neboli dostavané.

2. Redukcia na 3 dimenzie



- Aj napriek tomu že sme redukovali 255 stĺpcov na 3 tak stále môžeme vidieť ako rastie cena.

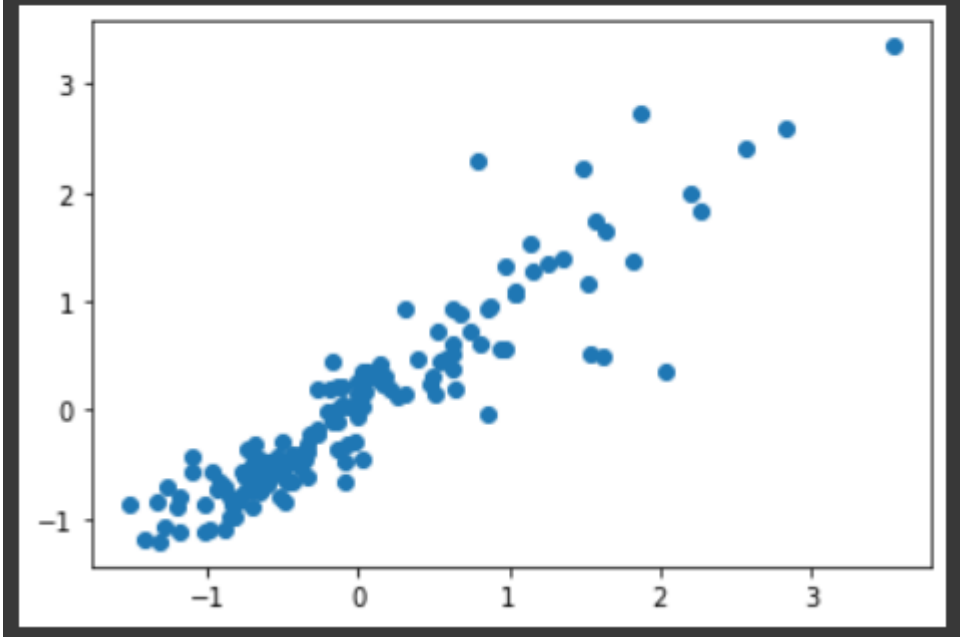
Redukcia na x dimenzií

- Najprv vybrali stĺpce ktorých korelácia s cenou bola abs(corr) >= 0.1.
- Z 255 stĺpcov nám zostalo 121.
- Keďže najlepšie dopadlo SVM budeme používať model kde kernel=rbf aj kernel=linear.

1. X=5

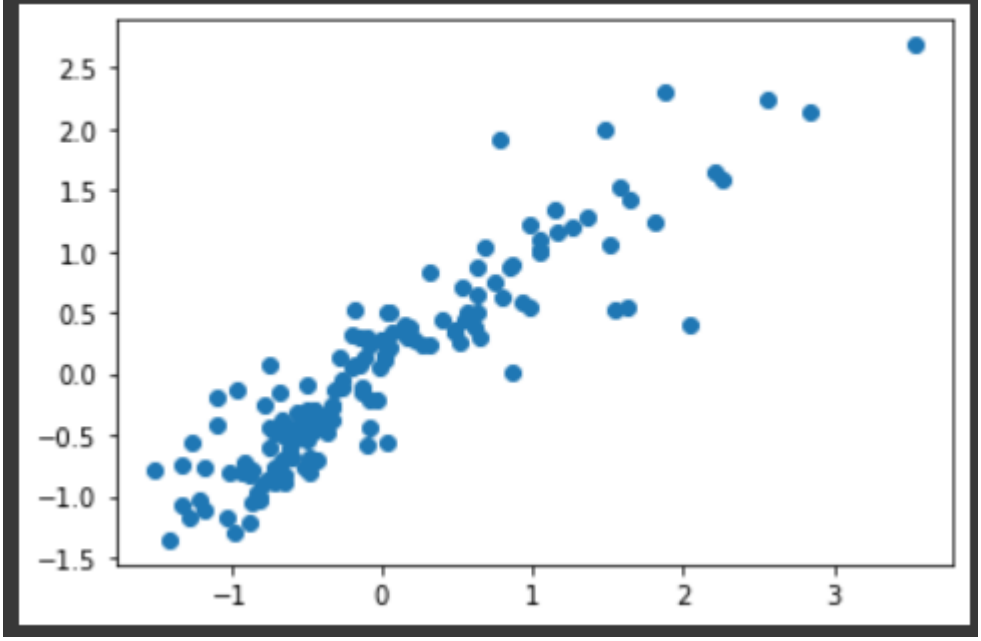
- *Rbf*

```
MSE:  0.11755540674942368
R2:  0.8603062651757091
```



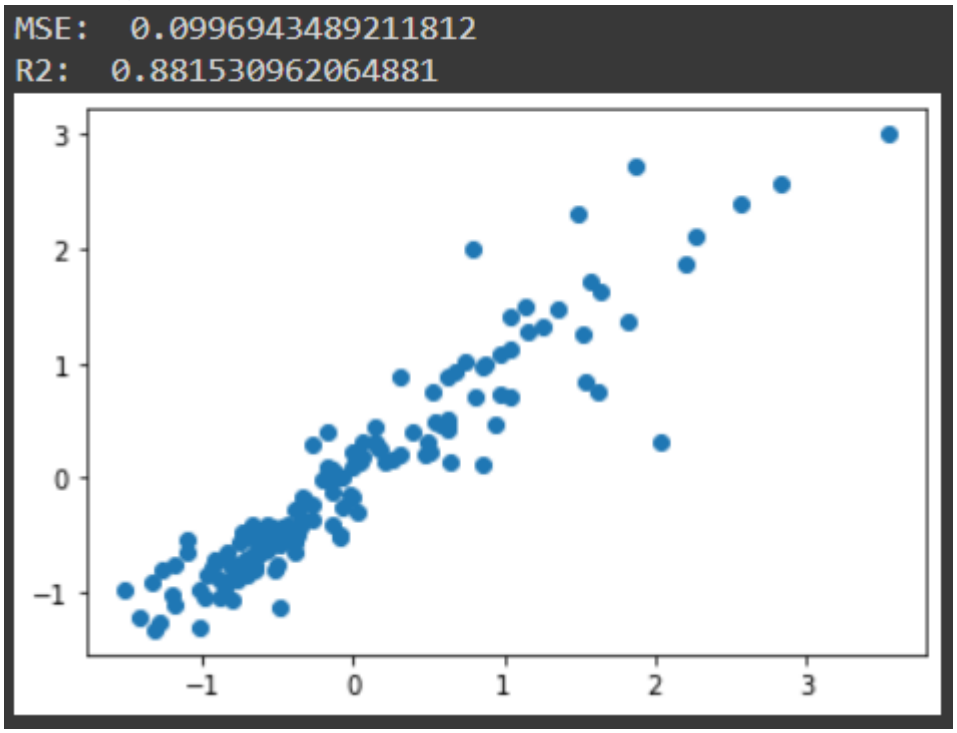
- *Linear*

```
MSE:  0.13863864479503654
R2:  0.8352525790355292
```

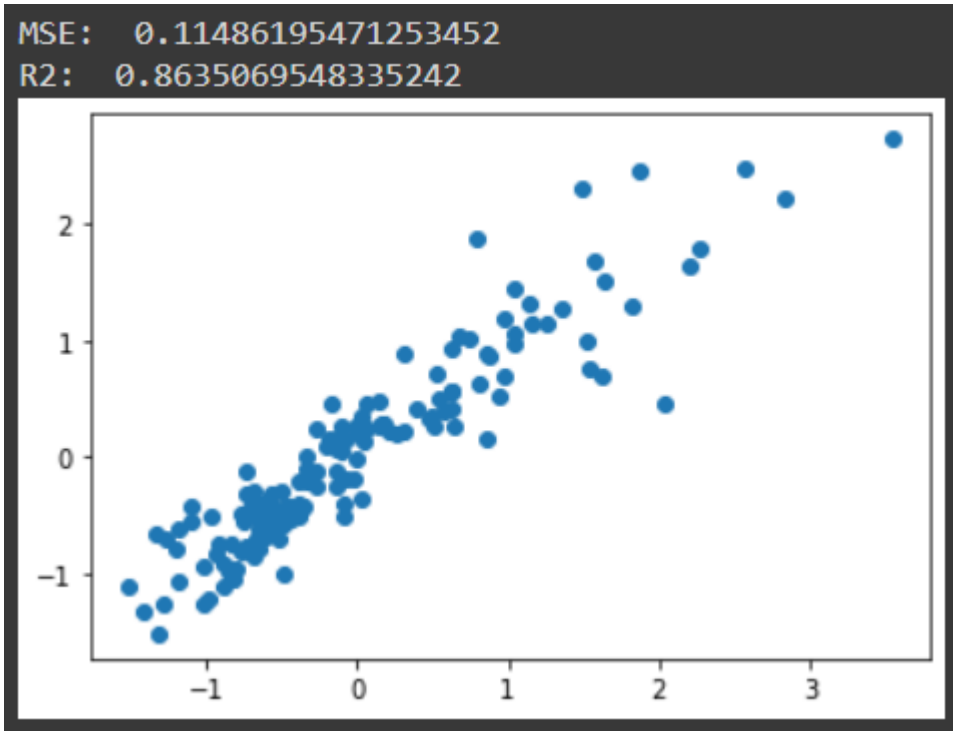


2. X=10

- *Rbf*

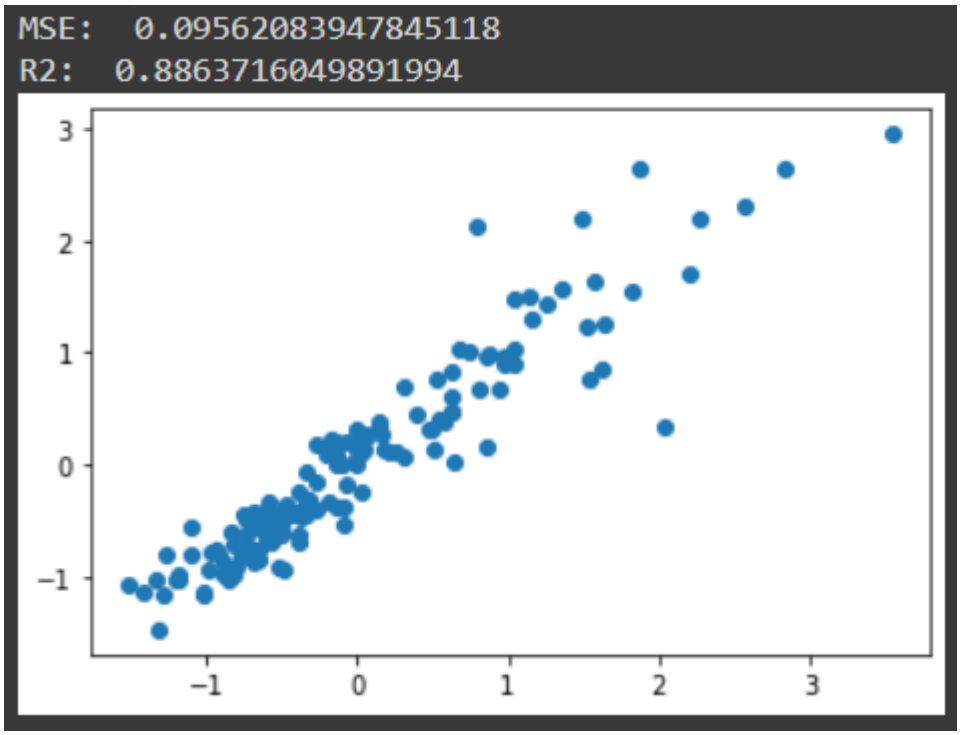


- *Linear*

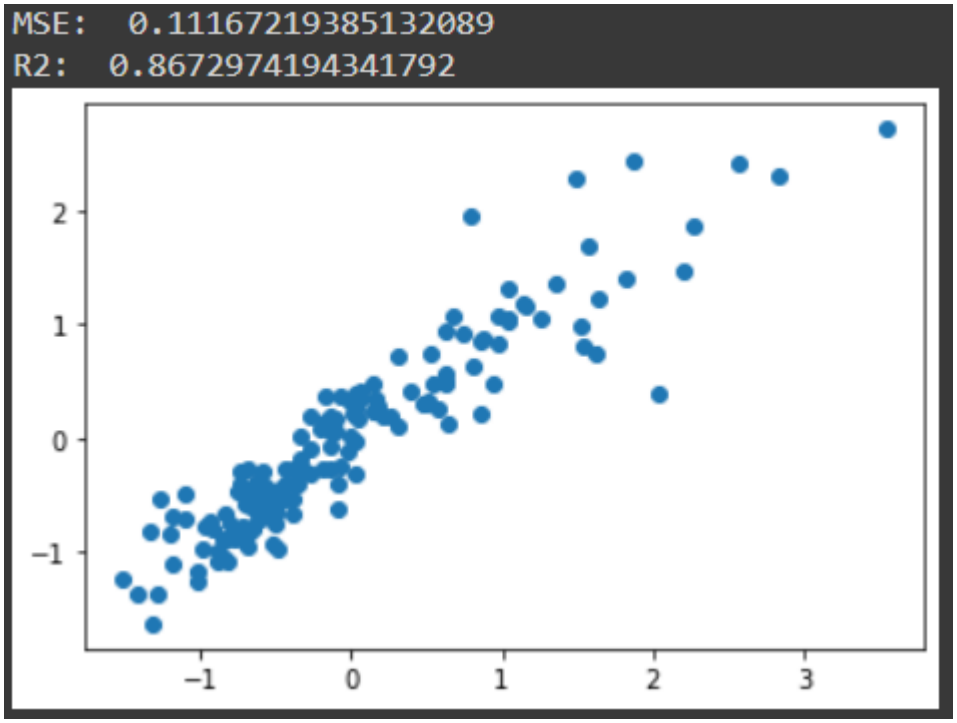


3. X=25

- *Rbf*



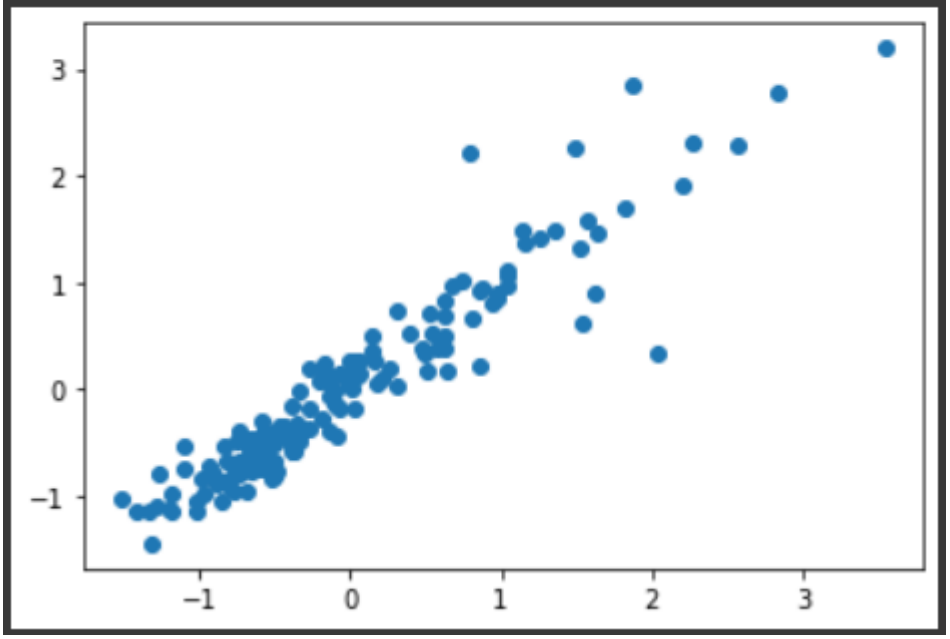
- *Linear*



4. X=50

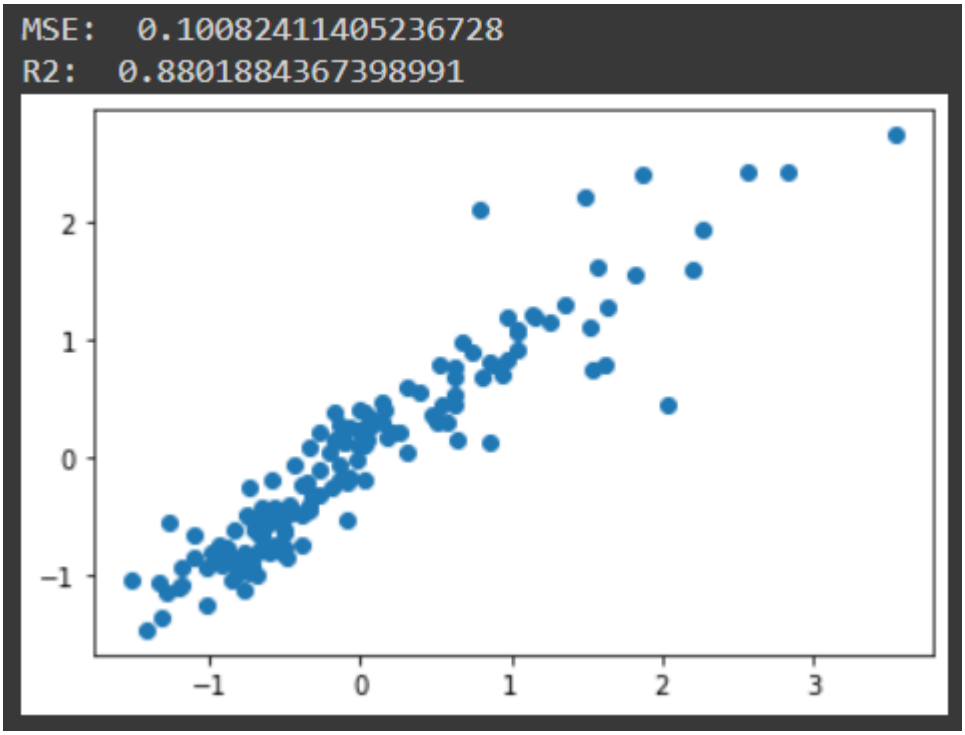
- *Rbf*

```
MSE:  0.09223122189335324
R2:  0.8903995638316007
```



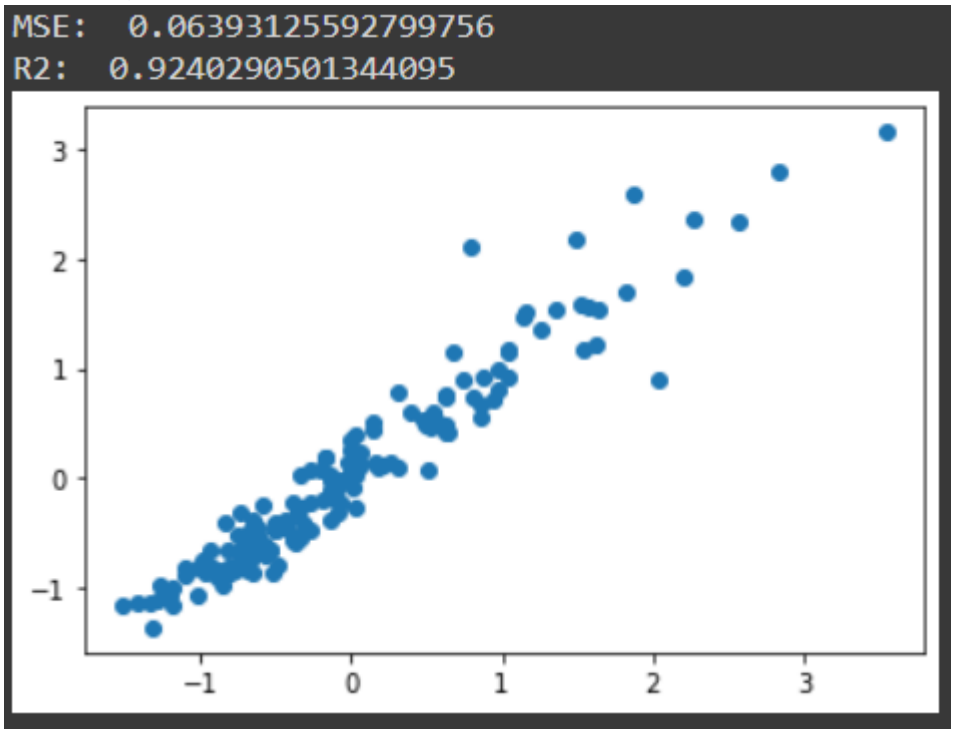
- *Linear*

```
MSE:  0.10082411405236728
R2:  0.8801884367398991
```

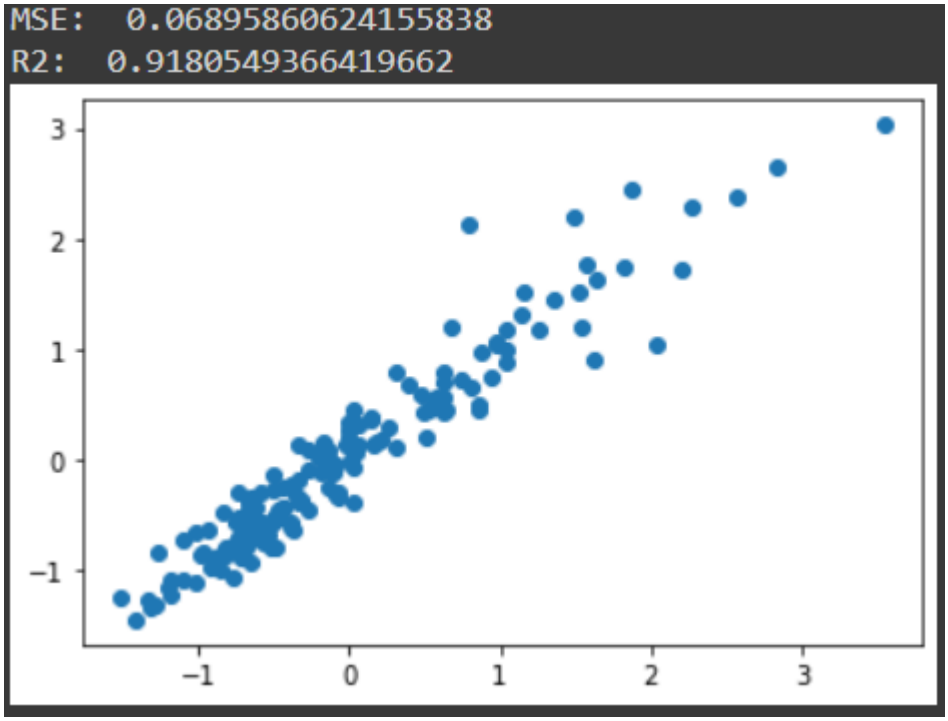


5. X=100

- *Rbf*



- *Linear*

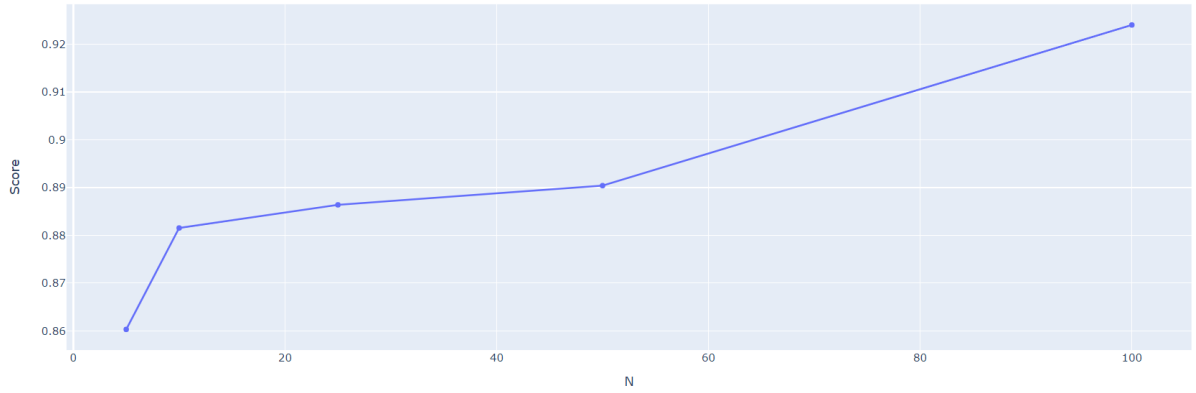


Zhrnutie

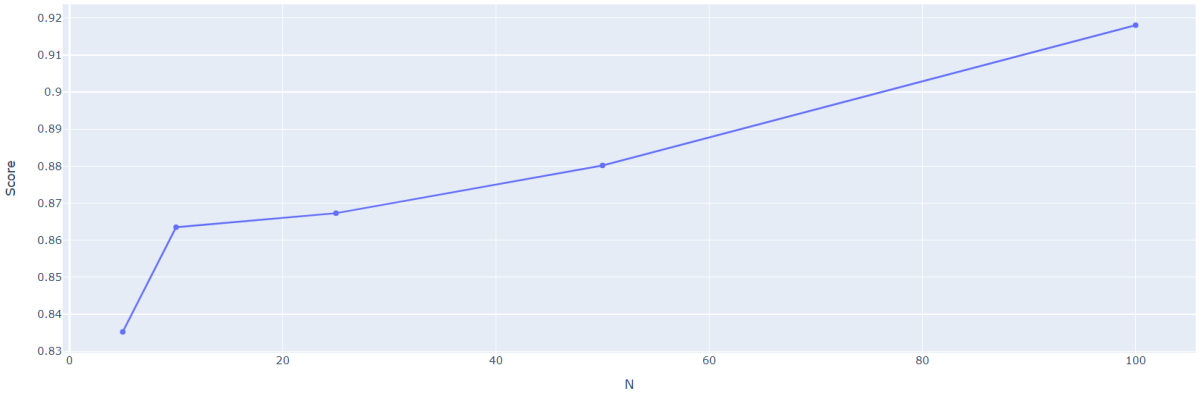
Kernel	N	R2	Time
Rbf	5	0.860306	0.119794
Linear	5	0.835253	0.076579
Rbf	10	0.881531	0.124930
Linear	10	0.863507	0.084409
Rbf	25	0.886372	0.150794
Linear	25	0.867297	0.162252
Rbf	50	0.890400	0.139008
Linear	50	0.880188	0.209399
Rbf	100	0.924029	0.136797
Linear	100	0.918055	0.352890

Vzťah skóre a n

- *Rbf*



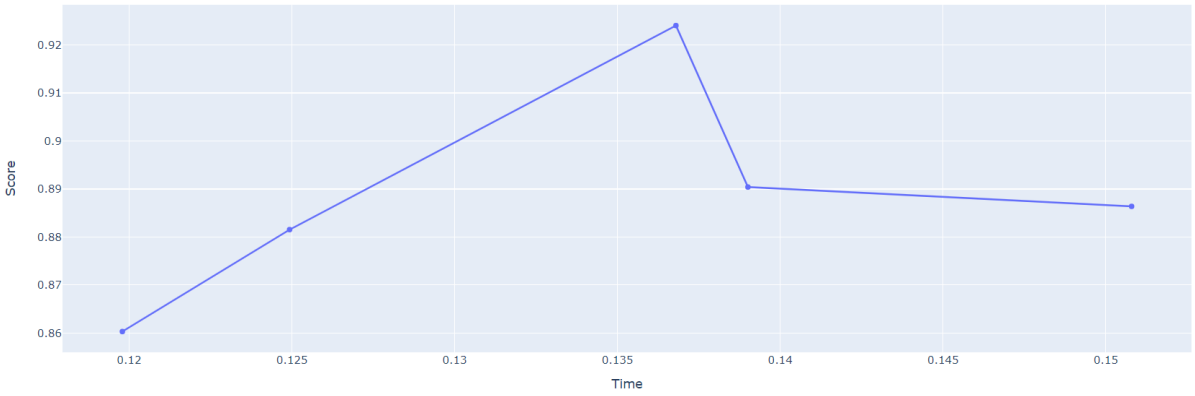
- *Linear*



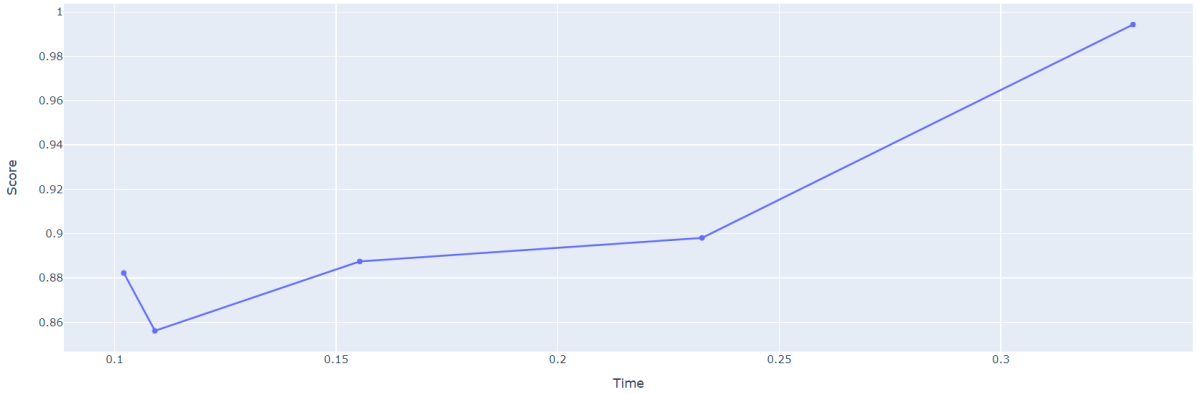
Pre obidva modely je vývoj skóre vzhľadom na veľkosť n približne rovnaký.

Vzťah skóre a času

- *Rbf*



- *Linear*

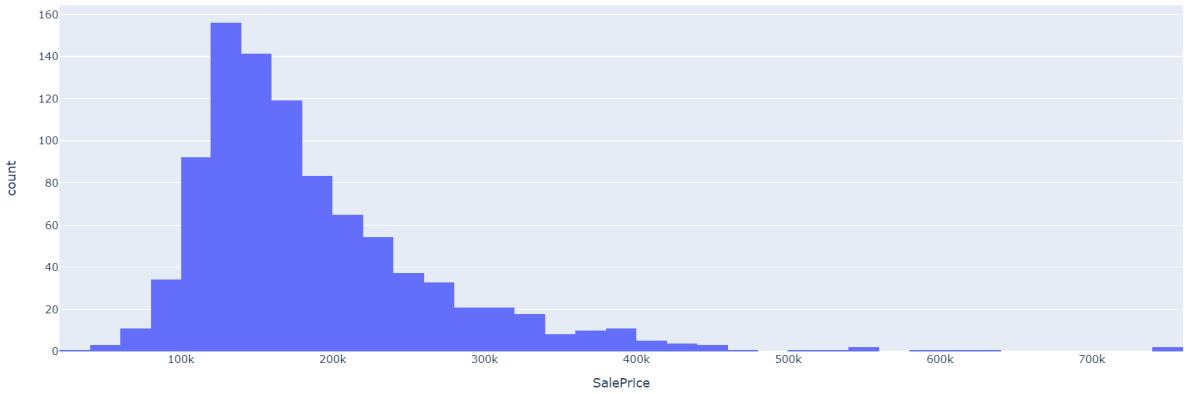


Pri linear môžeme vidieť že čím je vyššie skóre tým rastie aj čas trénovania.

Pri rbf trvá trénovanie približne rovnaký čas keďže rozdiel medzi najdlhším a najkratším trénovaním je len 0.03 sekundy.

EDA

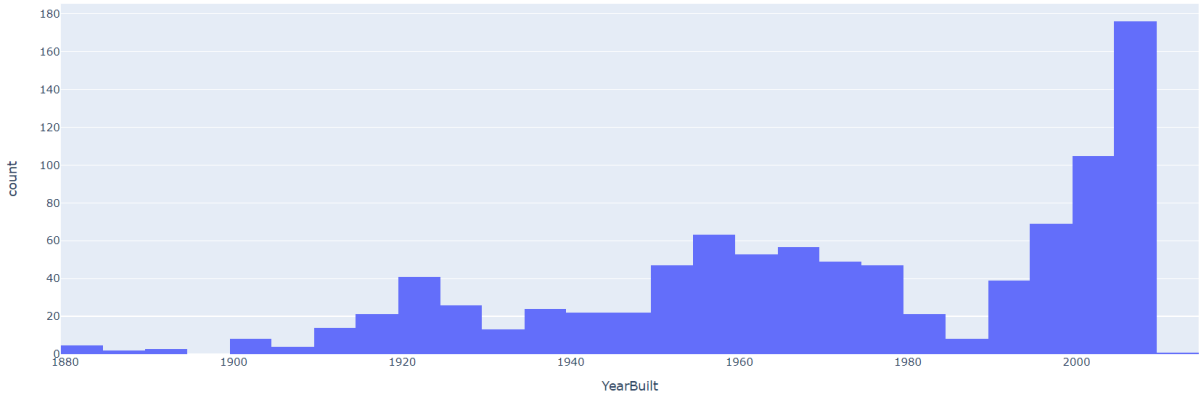
Rozdelenie ceny



Najväčšia časť záznamov sa nachádza v rozsahu od 120k do 139.99k (156)

Len 1.1%(10) záznamov je drahšia ako 500k a 5.2%(49) lacnejšia ako 100k.

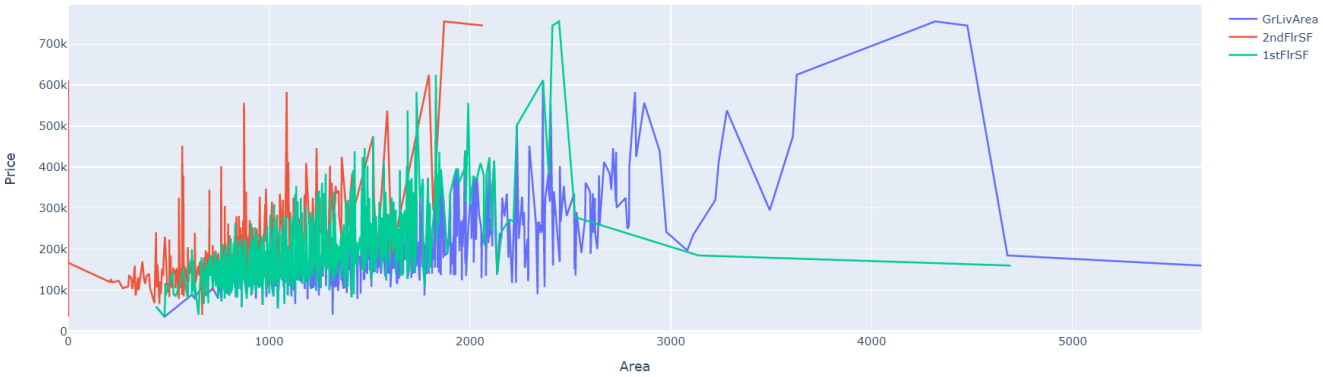
Rozdelenie roku výstavby



Môžeme vidieť že od roku 1985 počet postavených domov rastie.

Máme 10 záznamov, ktoré boli postavené pred rokom 1900.

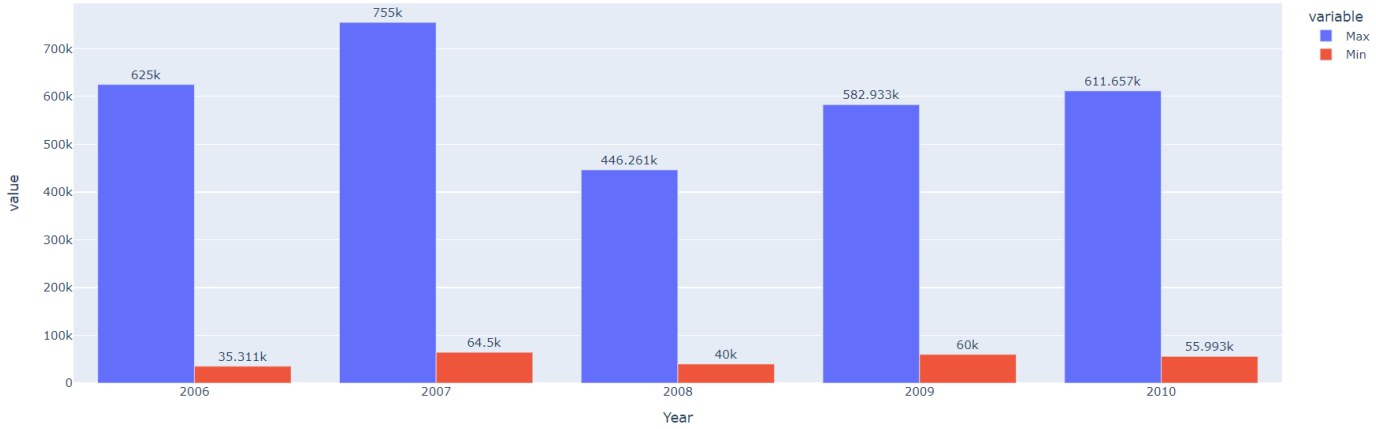
Závislosť ceny od GrLivArea, 1stFlrSF a 2ndFlrSF



Môžeme vidieť že najrýchlejší nárast ceny je pri rozlohe druhého poschodia.

Tak isto môžeme vidieť pád na konci pri GrLivArea a 1stFlSF. Ten je spôsobený dvomi záznamami, ktoré boli spomenuté už pri redukcií dimenzií.

Najdrahší a najlacnejší dom predaný v danom roku



Môžeme vidieť že najdrahší dom bol predaný v roku 2007 a najlacnejší v roku 2006.

Záver

Najúspešnejší model bol SVM s hyperparametrami c=10, gamma=0.001a kernel=rbf. Dáta pre tento model boli najprv za pomoci korelačnej matice zredukované na 121 stĺpcov a následne ešte zredukované pomocou PCA na 100 stĺpcov. Tento model dosiahol skóre 0.924.