# Regression for Pricing of LEGO Sets

Phillip Busko
&
Sorin Luca

# The LEGO Dataset

Data Source

- brickset.com
- kaggle.com/lego-database

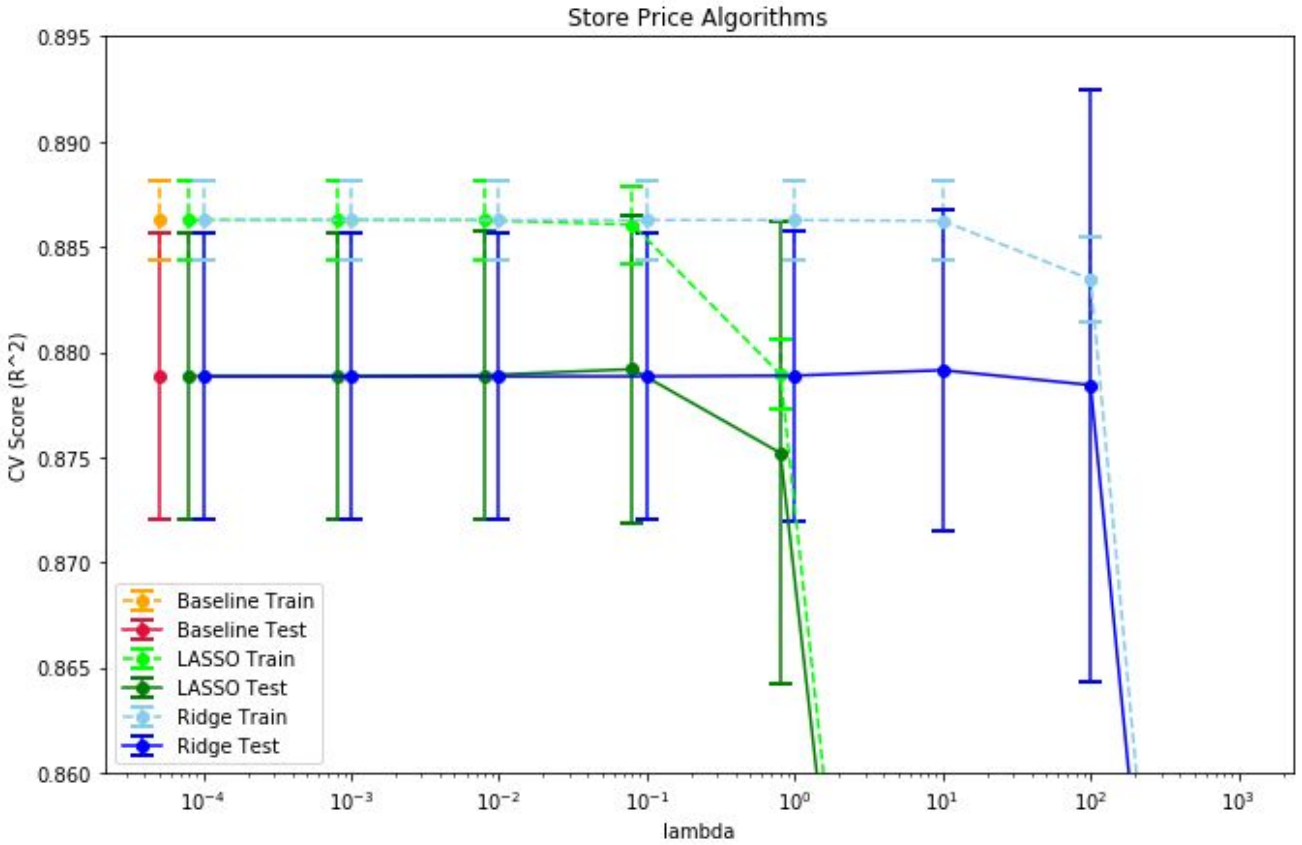
| Identifiers |
| --- |
| Set No # |
| Name |

| Targets |
| --- |
| Store Price |
| Used Price |
| Popularity |

| Features |
| --- |
| Year |
| Theme |
| Total Parts |
| Different Parts |
| Different Colors |
| Primary Color |
| Secondary Color |

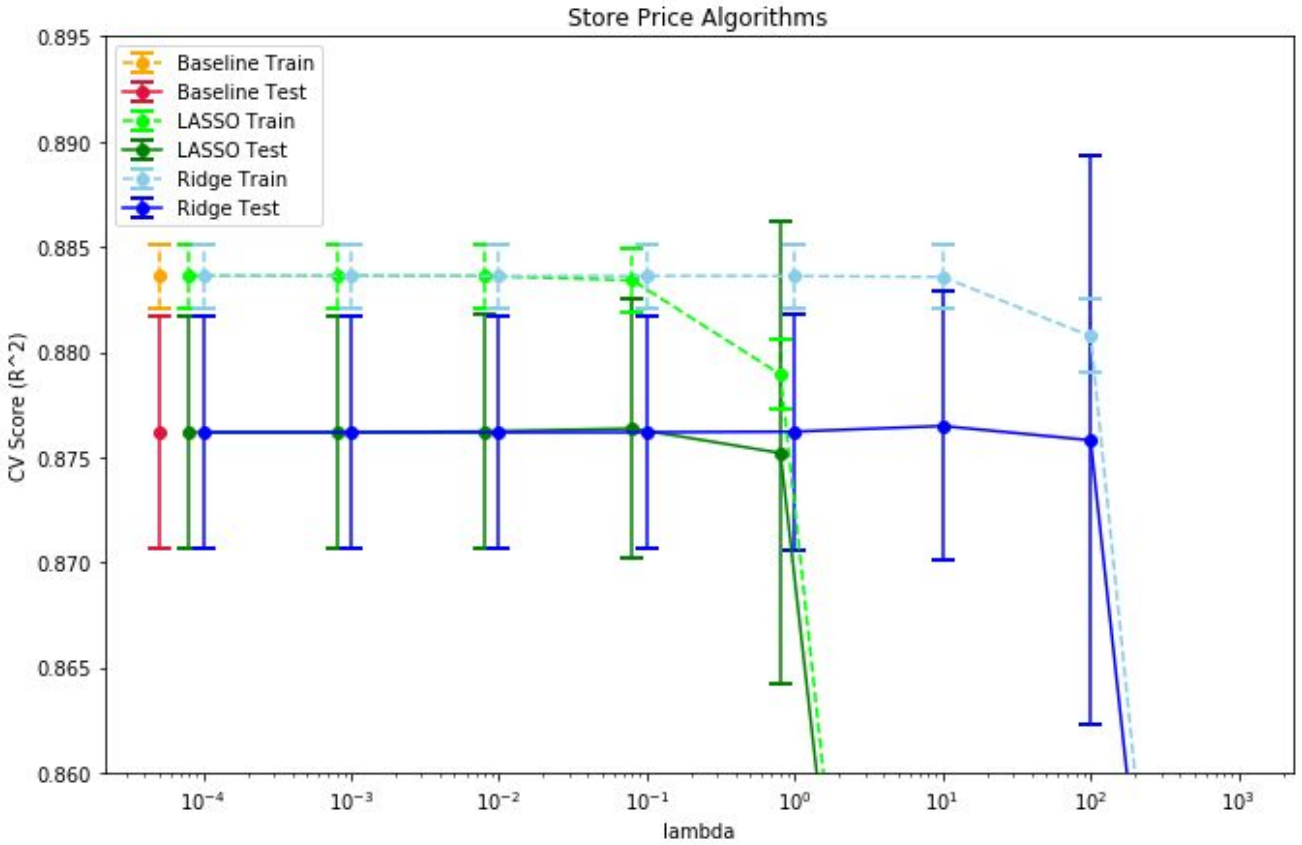

# Linear Regression First Pass

| Feature | VIF |
|---|---|
| Year | 19.2 |
| Theme | ~1.9 |
| Total Parts | 4.9 |
| Different Parts | 14.6 |
| Different Colors | 16.7 |
| Primary Color | ~1.3 |
| Secondary Color | ~1.4 |



# Linear Regression Second Pass

| Feature | VIF |
|---|---|
| ~~Year~~ | - |
| Theme | ~1.7 |
| Total Parts | 4.4 |
| Different Parts | 6.9 |
| ~~Different Colors~~ | - |
| Primary Color | ~1.1 |
| Secondary Color | ~1.2 |



# Linear Regression Final Results

Second Pass Winner

- Ridge with $\lambda$=10
- CV Train $R^2$-adj: 0.882
- CV Test $R^2$-adj: 0.875
- Full Test $R^2$-adj: 0.830

Final Model

store_price =

+ 33.8 • total_parts

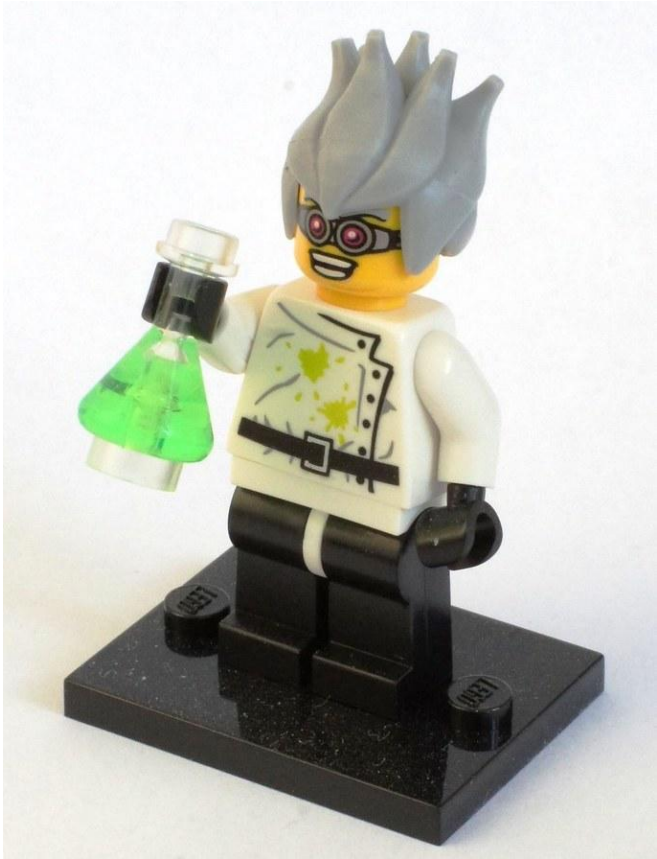
+ 8.95 • different_parts

+ 0.76 • ~theme

- 0.23 • ~primary_color

- 0.37 • ~secondary_color

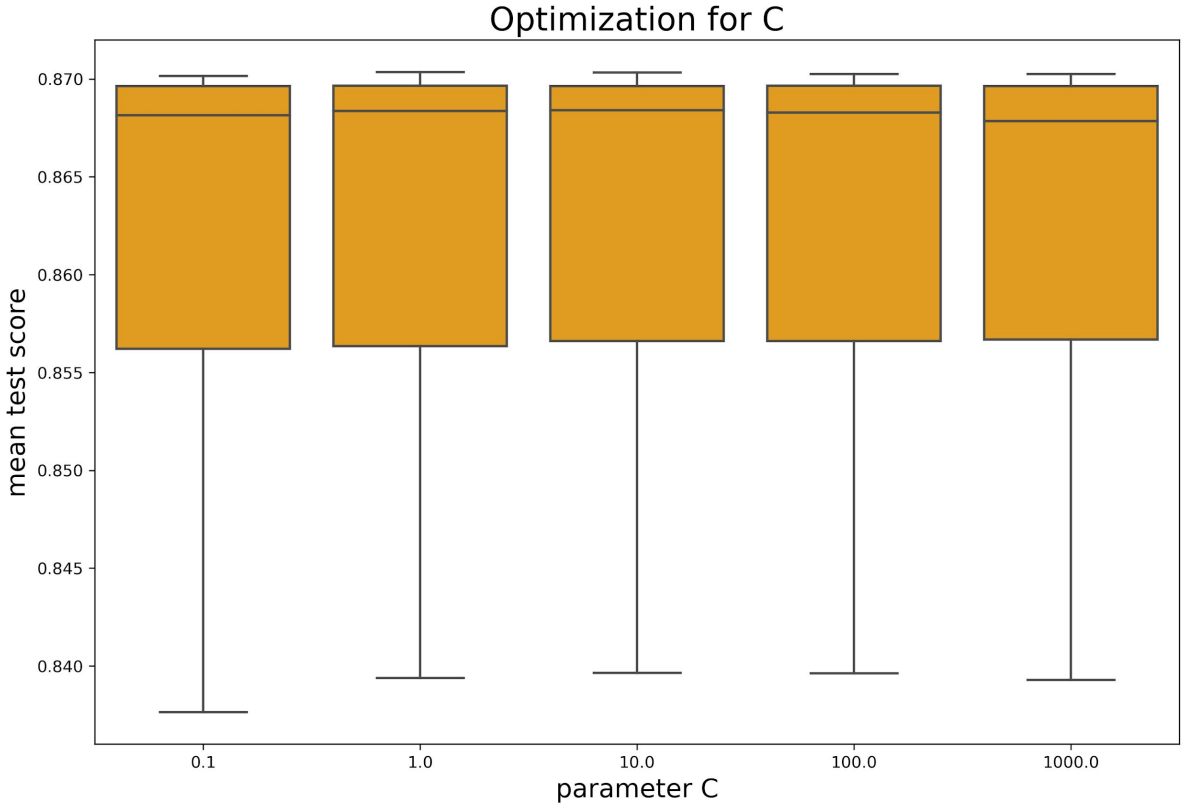
- 38.6



# Support Vector Regression (linear kernel)

- $R^2$ train 0.875
- $R^2$ test 0.867

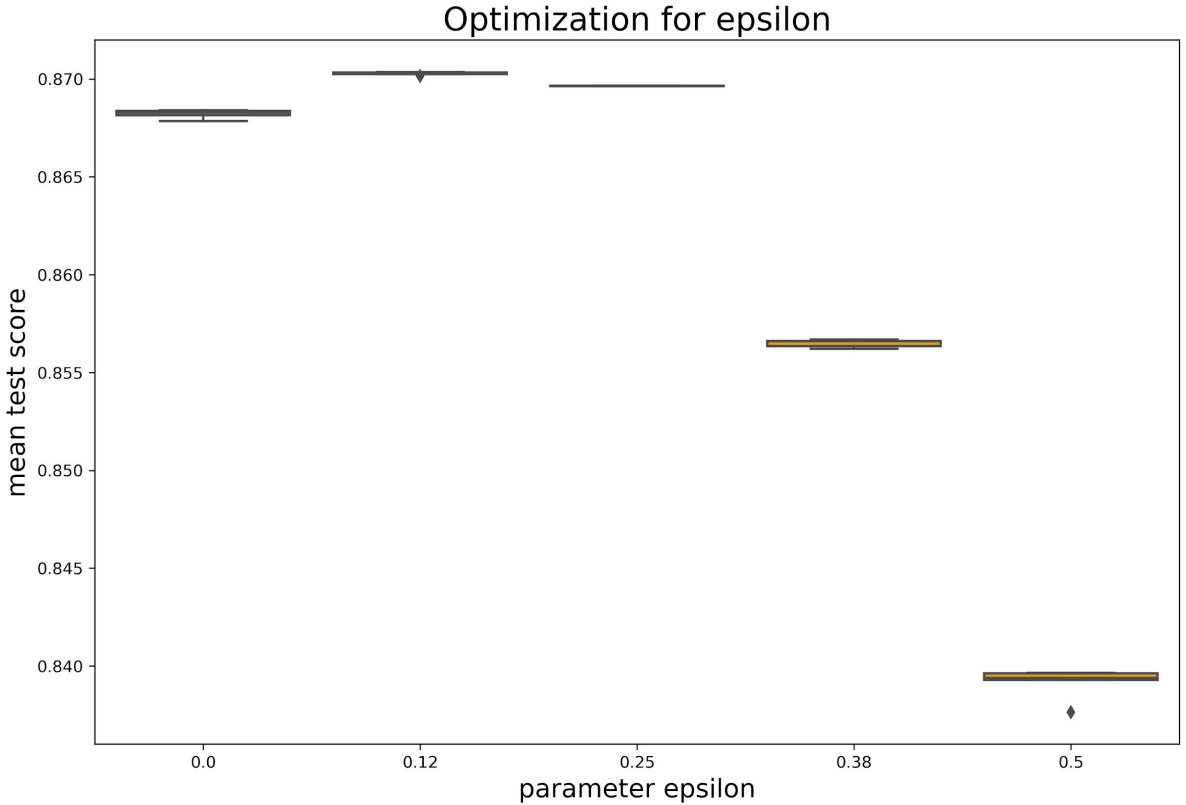
| | |
|---|---|
| C | 1 |
| epsilon | 0.125 |



# Support Vector Regression (linear kernel)

- $R^2$ train 0.875
- $R^2$ test 0.867

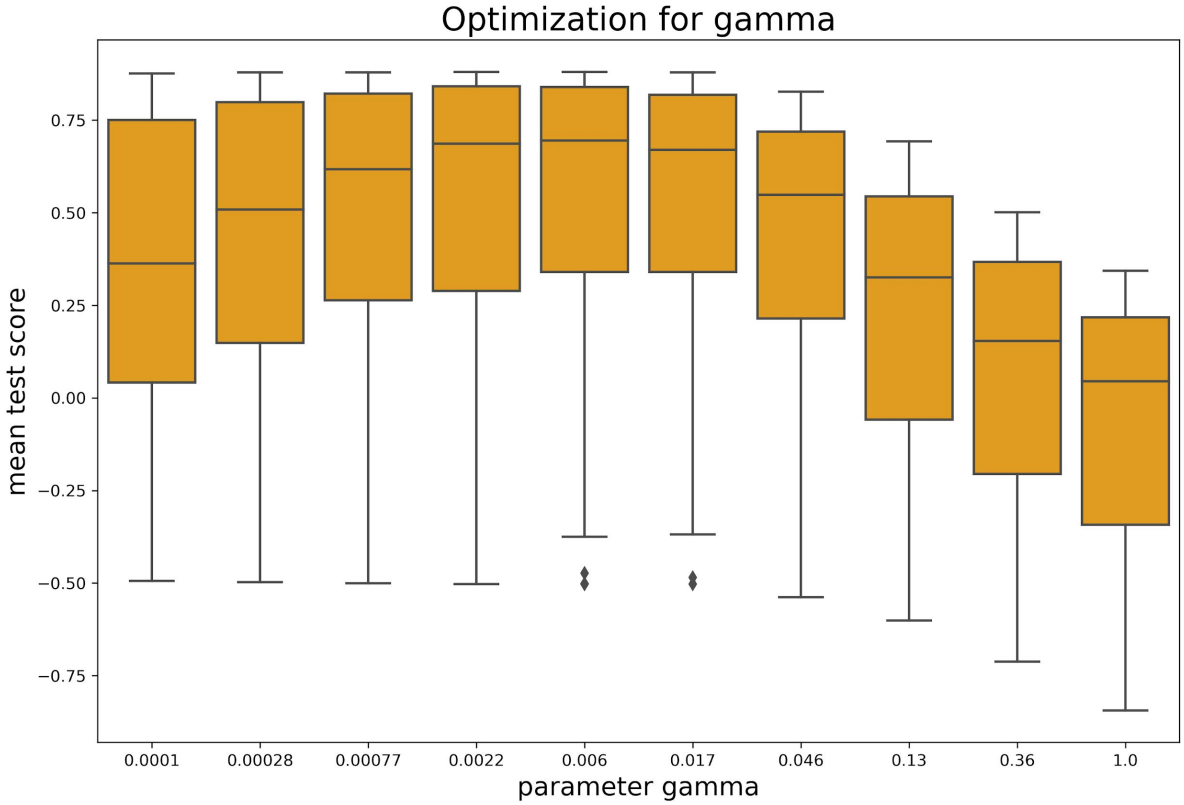
| | |
|---|---|
| C | 1 |
| epsilon | 0.125 |



# Support Vector Regression (RBF kernel)

- $R^2$ train 0.897
- $R^2$ test 0.896

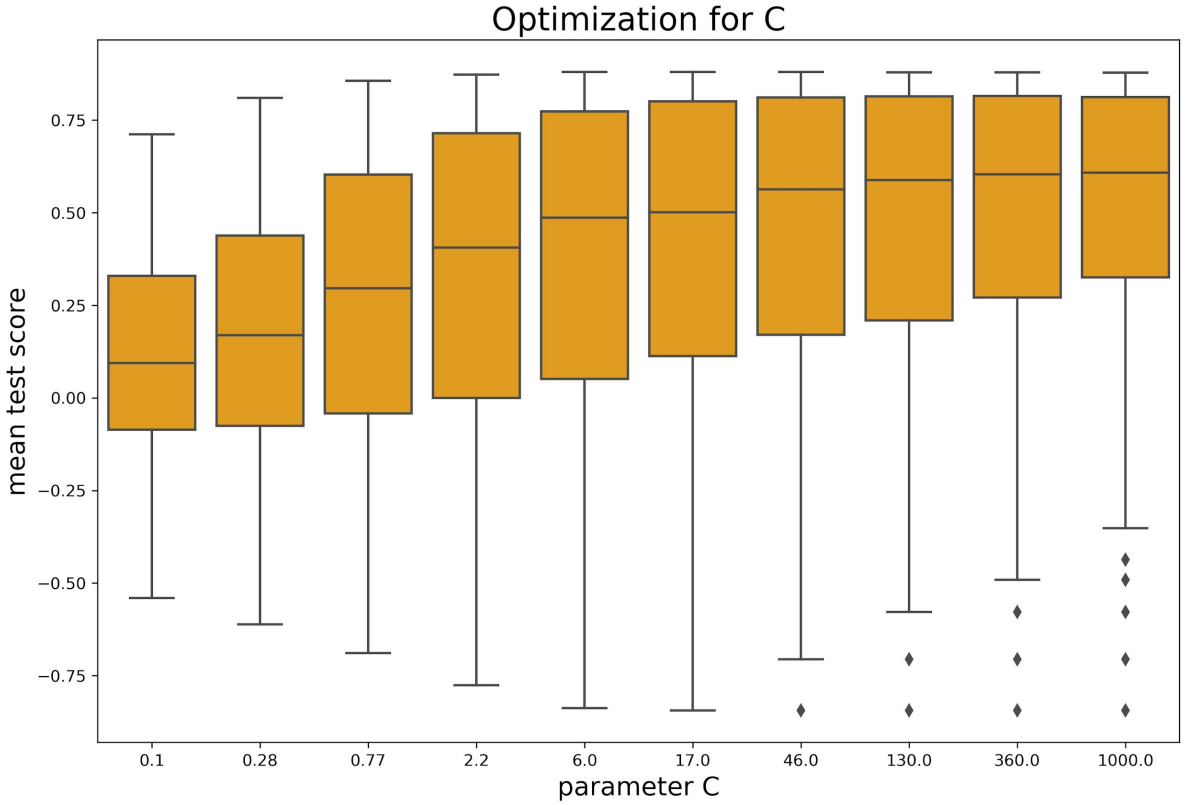
| | |
|---|---|
| C | 16.7 |
| epsilon | 0.05 |
| gamma | 0.002 |



# Support Vector Regression (RBF kernel)

- $R^2$ train 0.897
- $R^2$ test 0.896

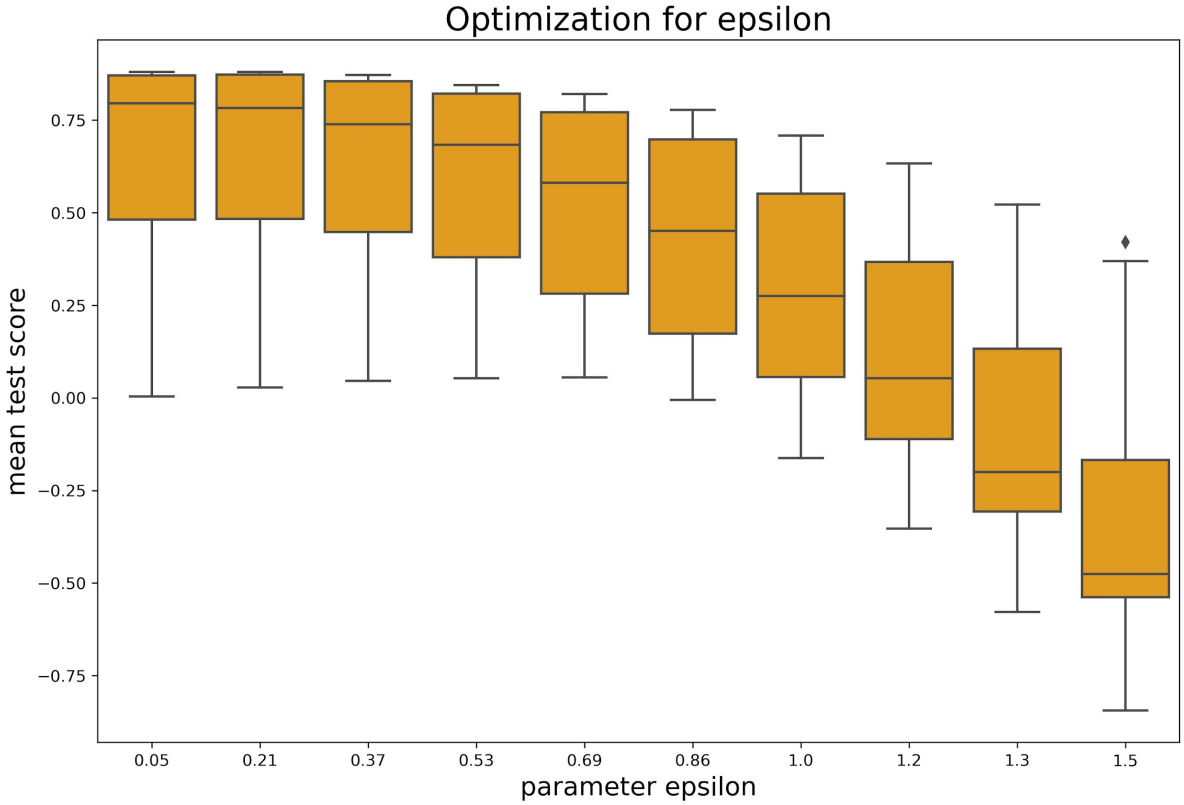
| | |
|---|---|
| C | 16.7 |
| epsilon | 0.05 |
| gamma | 0.002 |



# Support Vector Regression (RBF kernel)

- $R^2$ train 0.897
- $R^2$ test 0.896

| | |
|---|---|
| C | 16.7 |
| epsilon | 0.05 |
| gamma | 0.002 |



# Conclusions

1. The store price of various LEGO sets was modeled by linear (accuracy $R^2$ of 0.83) and support vector regression algorithms (accuracy $R^2$ of 0.89) models.

2. The *total number of parts* and the *different number of parts* were the most important features.

3. The SVR model performed slightly better when the RBF kernel was employed compared to the linear kernel.

4. Future analysis can be done for the used price and popularity of LEGO sets.