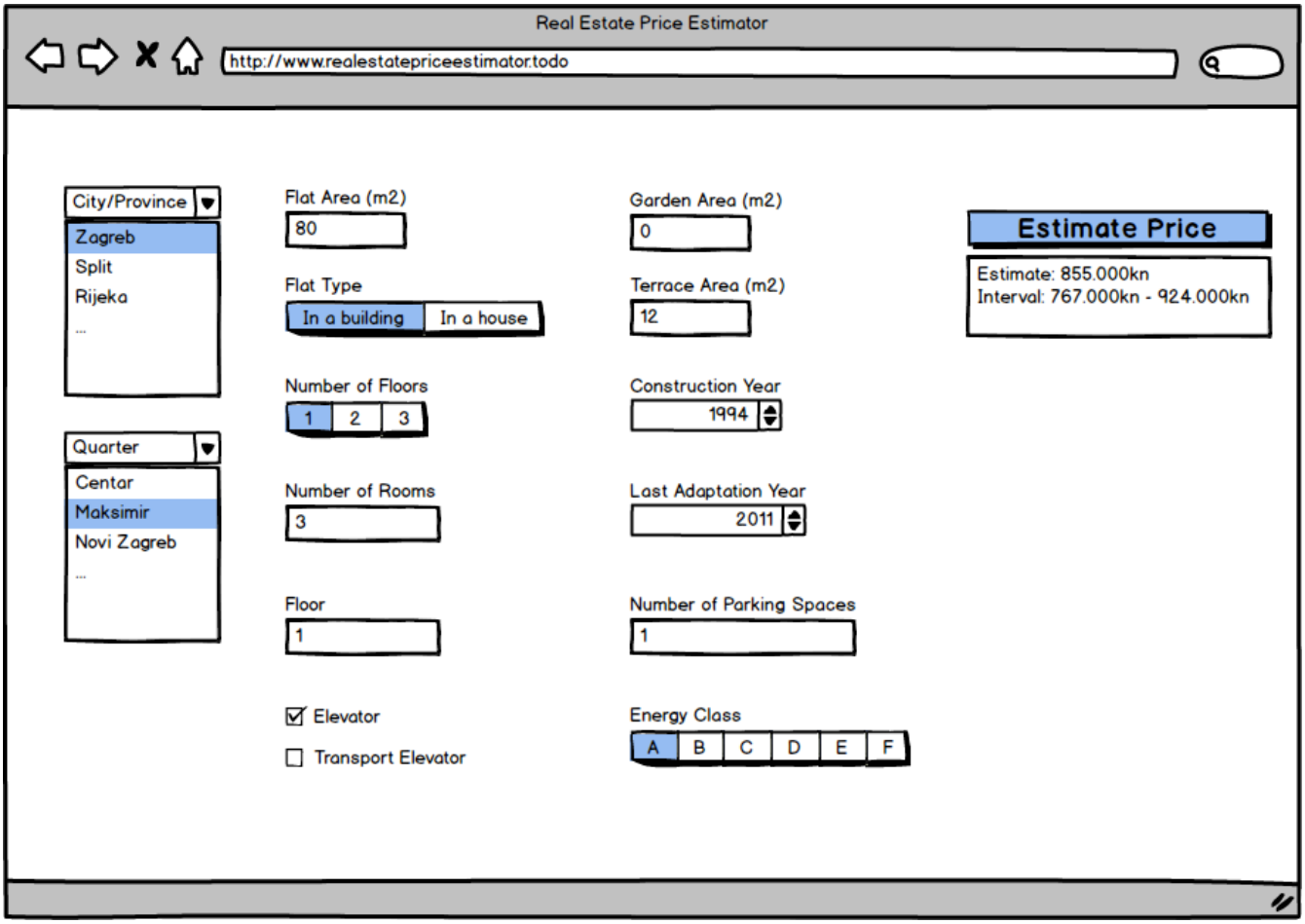
Real Estate Price Modelling

Danijel Kopčinović



Problem

One of the most important questions in real estate business is how to estimate the current market value of a property. It obviously depends on some variables like area, location etc. but which of these variables influence the price the most and how to assess each variable's influence on the price?

We will answer this question for one example, using data taken from real, existing flat selling ads to predict the flat prices. This model can then be used to make good price estimates for flats based on parameters like flat size, location, number of rooms etc.

Data

The data used to create the model consists of:

- Price, the variable that we want to predict related to the other variables.

- 16 other variables, related to different aspects of the flat in the ad (town, quarter, flat area, number of rooms, floor, construction year, last adaptation year...). Some variables exist in the original ads and some we created to make the analysis and modelling easier.

We had a total of around 7500 ads - data entries.

```
## 'data.frame':    7417 obs. of  17 variables:
##  $ cijena                  : num  1154494 2866621 867530 2640309 3168371 ...
##  $ stambena_povrsina       : num  95.7 140 70 175.9 220 ...
##  $ tip_stana               : Factor w/ 2 levels "u kući","u stambenoj zgradi": 2 2 2 2 2 2 2 2 2 2 .
##  $ broj_etaza              : num  1 1 2 2 2 1 2 1 1 1 ...
##  $ broj_soba               : num  5 5 4 5 5 5 5 3 5 1 ...
##  $ kat                     : num  4 5 12 4 6 4 5 5 3 4 ...
##  $ lift                    : num  0 0 1 0 0 0 0 0 0 1 ...
##  $ teretni_lift            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ povrsina_vrta           : num  50 0 0 57.9 0 ...
##  $ povrsina_balkona        : num  0 2.5 13.5 11.6 0 ...
##  $ povrsina_terase         : num  0 0 0 12.7 0 ...
##  $ godina_izgradnje        : num  2018 0 0 1931 0 ...
##  $ godina_zadnje_adaptacije: num  2018 2013 2016 2009 2016 ...
##  $ novogradnja             : num  1 0 0 0 0 0 1 1 0 1 ...
##  $ broj_parkirnih_mjesta   : num  1 0 0 2 0 0 0 1 1 1 ...
##  $ energetski_razred       : num  8 6 5 2 0 5 5 6 5 5 ...
##  $ grad_opcina_naselje     : Factor w/ 550 levels "Bakar Bakar-dio",..: 416 45 189 45 45 45 249 61 78
```

The initial part of the analysis and modelling process also includes cleaning the data, estimating the missing values and scaling all the values to similar scale (0-1) because different scales can badly influence the modelling algorithms.

Analysis

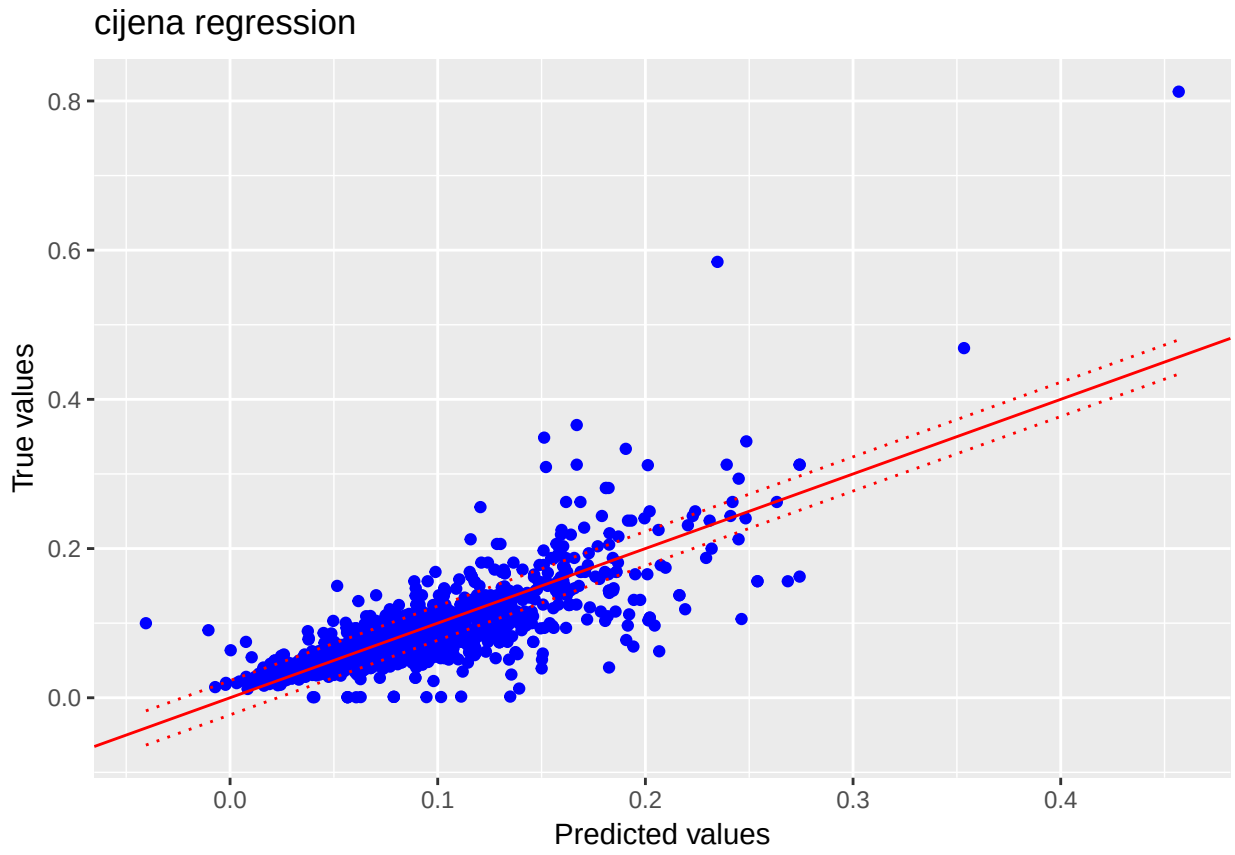
First part of the analysis calculates correlation between our target variable (price) and the other variables.

V1	V2	indicator_value
cijena	stambena_povrsina	0.7035016
cijena	tip_stana	0.08564247
cijena	broj_etaza	0.1198698
cijena	broj_soba	0.4983864
cijena	kat	0.03967443
cijena	lift	0.02638398
cijena	teretni_lift	-0.037232
cijena	povrsina_vrta	0.08103382
cijena	povrsina_balkona	0.07187488
cijena	povrsina_terase	0.1448048
cijena	godina_izgradnje	0.07485716
cijena	godina_zadnje_adaptacije	0.06487896
cijena	novogradnja	0.1549361
cijena	broj_parkirnih_mjesta	0.009192471
cijena	energetski_razred	0.07751353
cijena	grad_opcina_naselje	0.5297756

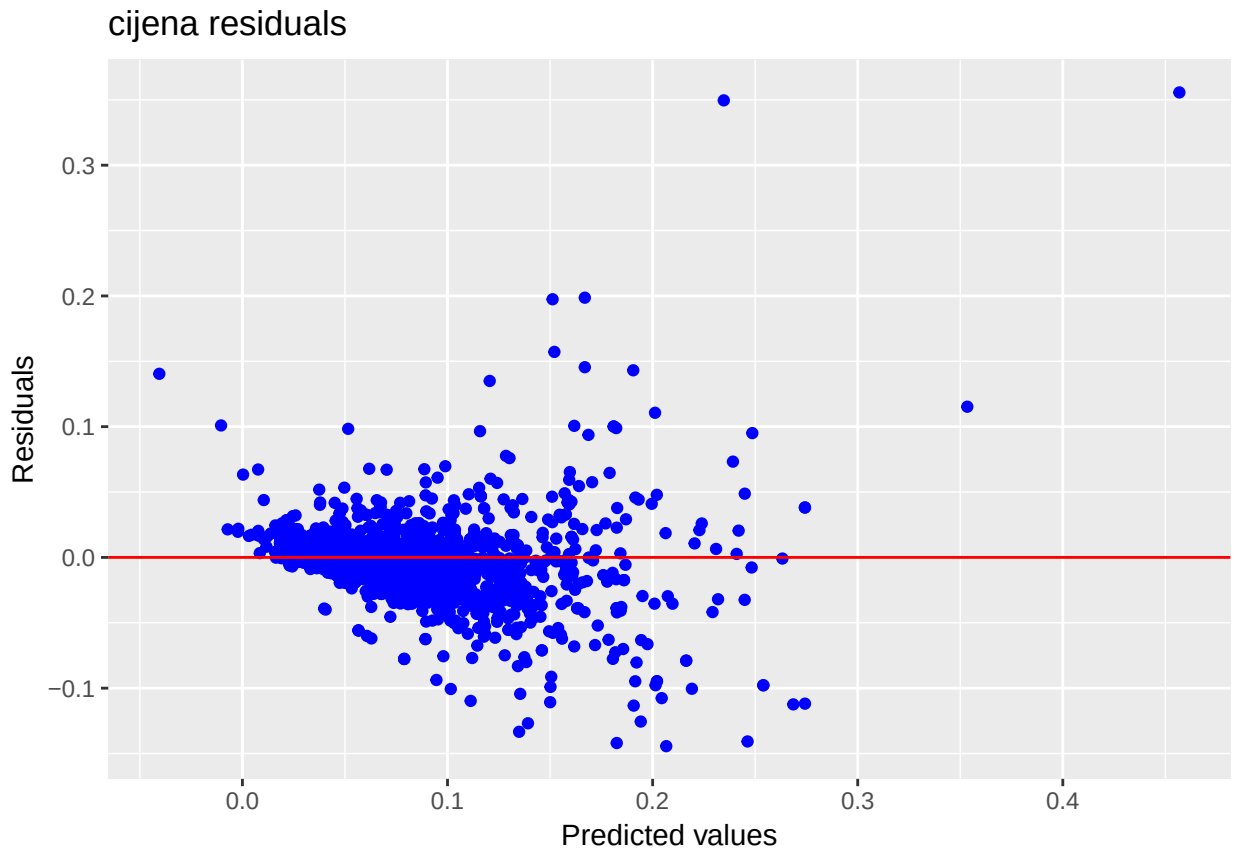
From the calculated correlations, we decide to keep only the most correlated variables: grad_opcina_naselje, stambena_povrsina, broj_soba, novogradnja, povrsina_terase, broj_etaza. A bit surprisingly, floor (kat),

construction year (godina_izgradnje) or energy class (energetski_razred) are very little correlated with the price.

We build our model using the neural network infrastructure, predict the prices on a test set and compare the predictions with the real values.



The predictions and the true values lay approximately on the $y = x$ line which is good (predictions are similar, close to the true values).



The residuals (differences between true values and predictions) are similarly scattered above and below the $y = 0$ line which is good (there is no clear grouping of residuals which would indicate that our predictions are somehow biased).

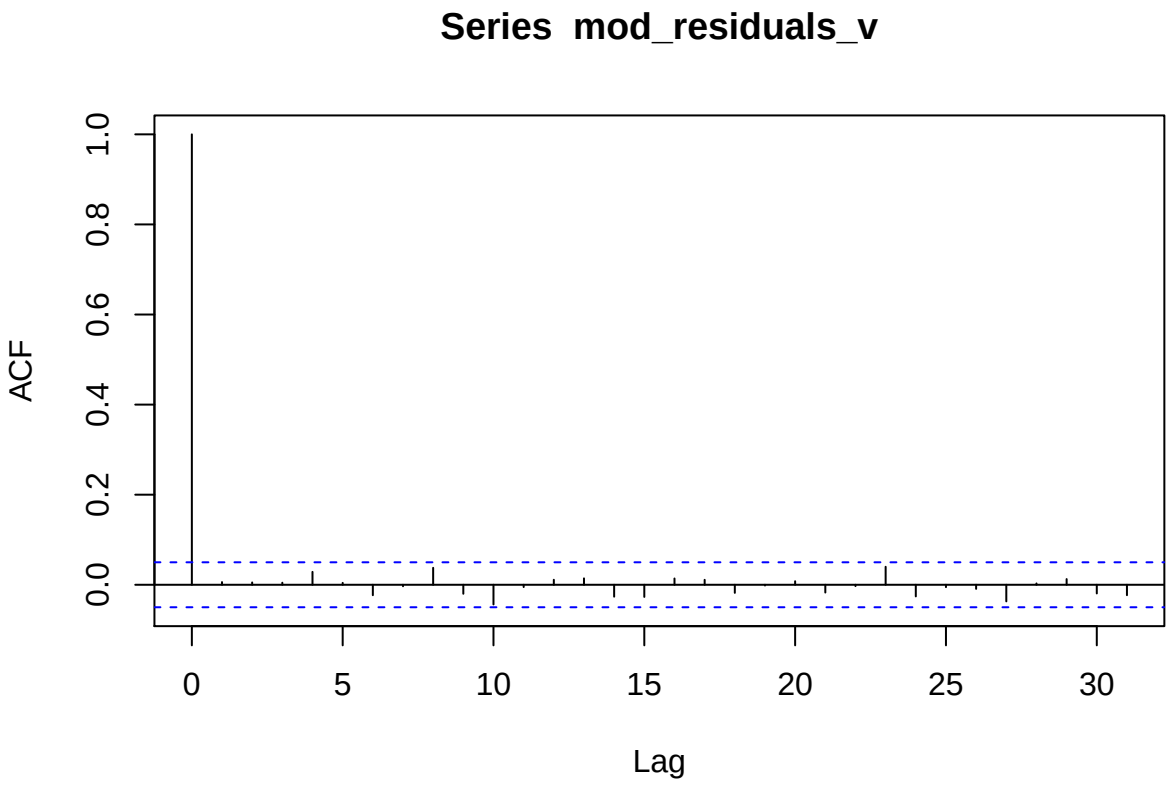
We also check the goodness of predictions by calculating different test statistics.

Ljung-Box, the p-value indicating that the residuals are normally distributed (closer to 1 is better):

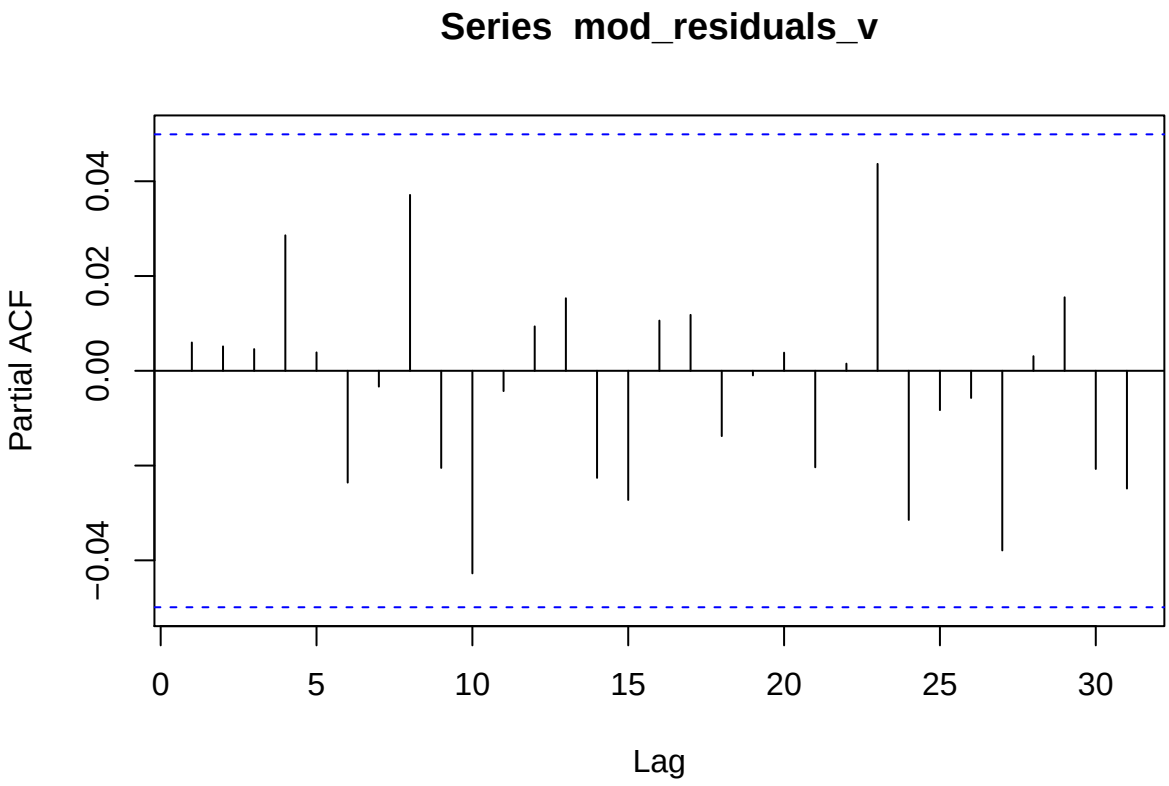
```
## [1] 0.6208083
```

AIC, Akaike information criterion, indicating the error size (less is better):

```
## [1] -2835.102
```



Residuals are not correlated (ACF), which is good.



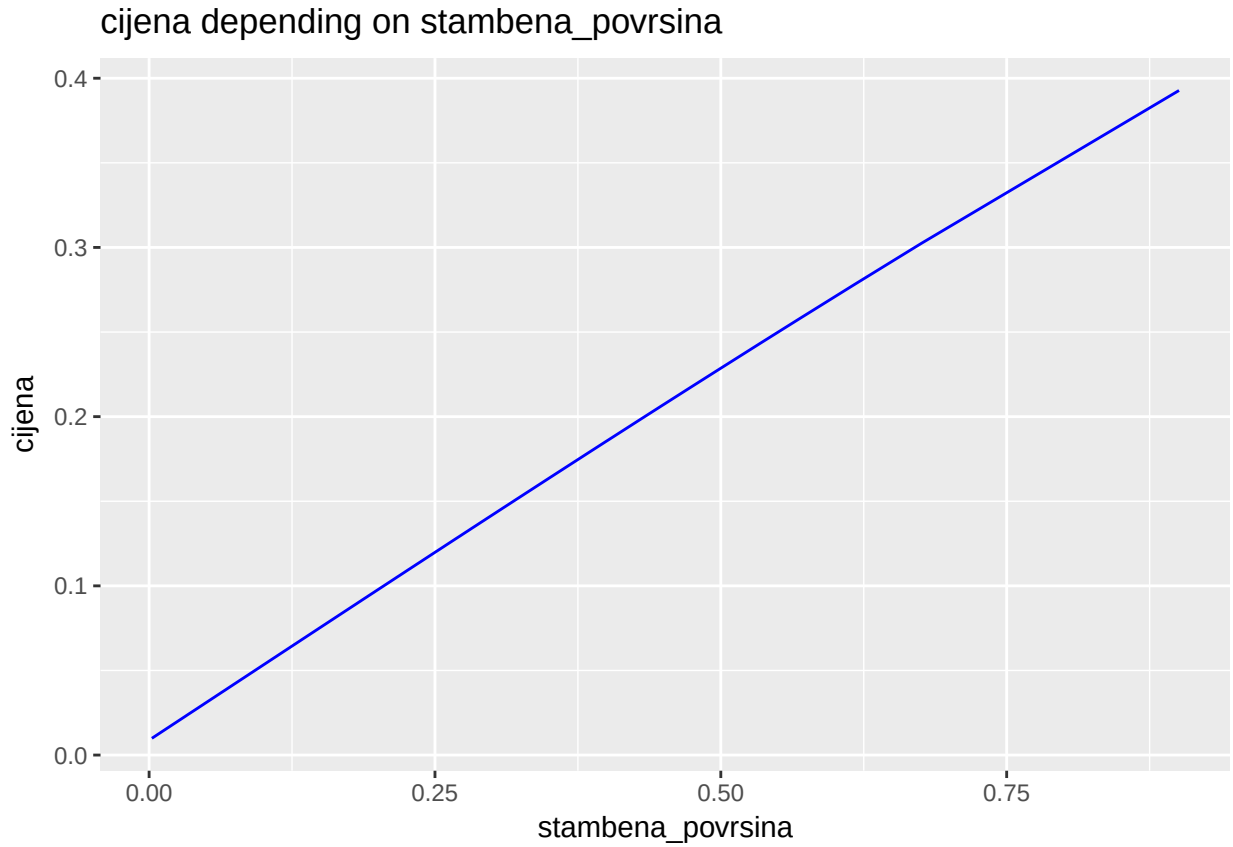
Residuals are not correlated (PACF), which is good.

R squared, indicating the amount of variability in price that we managed to explain with our model (closer to 1 is better:

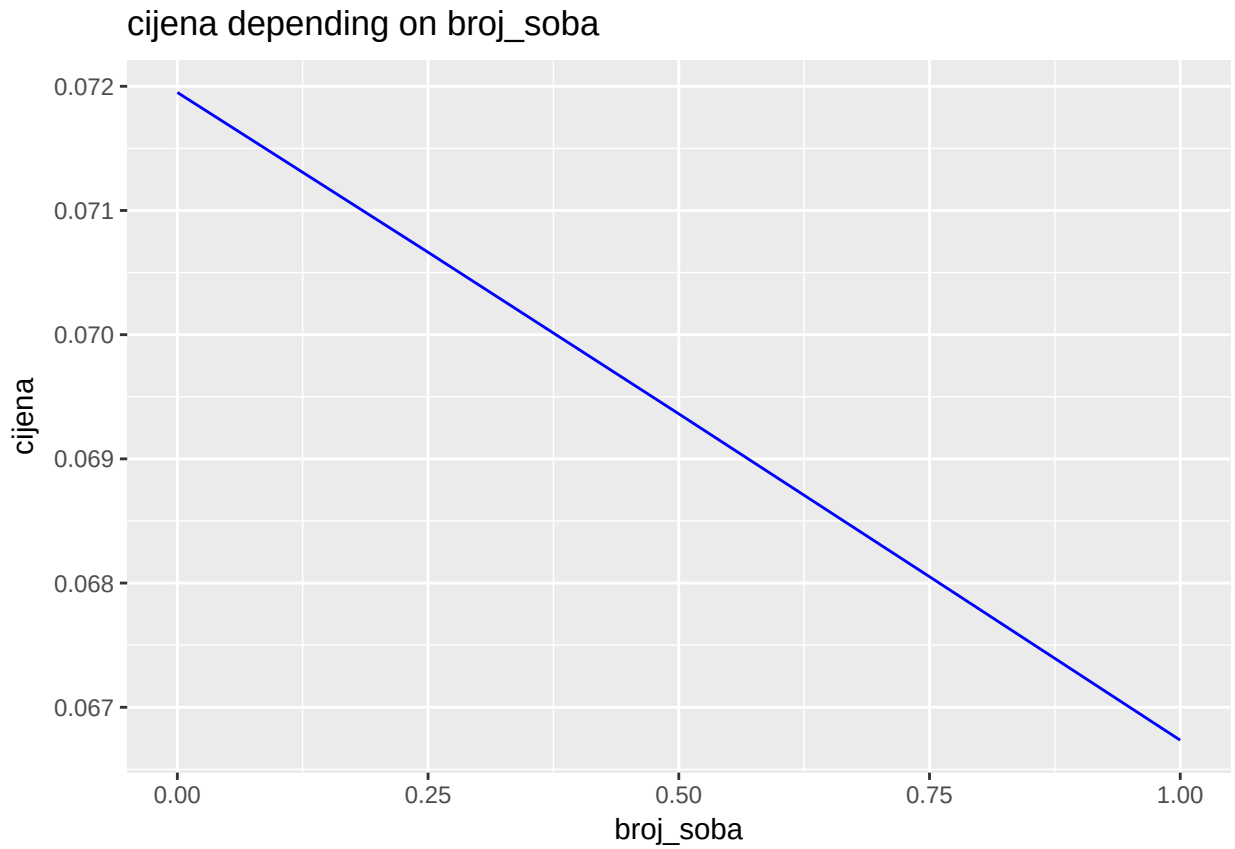
```
## [1] 0.6358102
```

Exact correlation between predictors and predicted variable

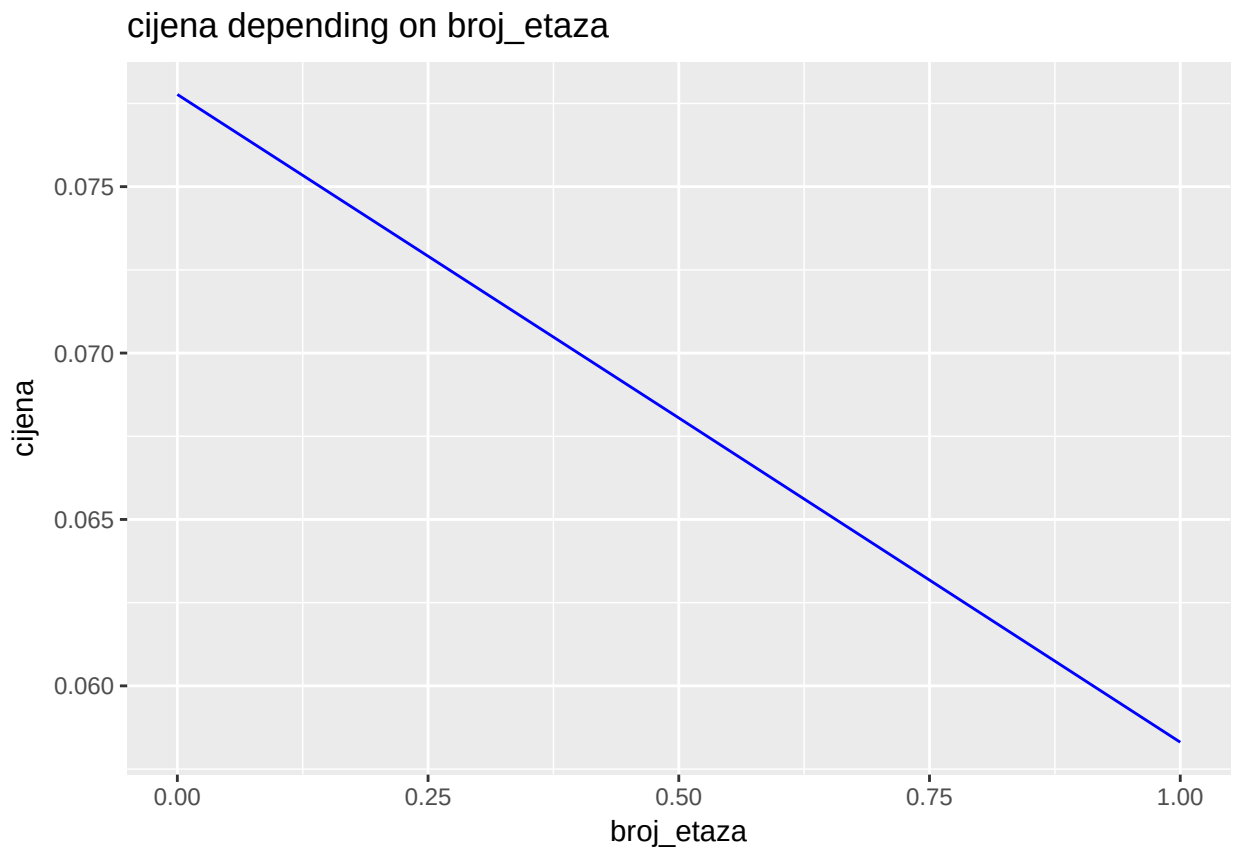
Let's see how **exactly** some of the chosen predictor variables influence the predicted variable price. We keep all other parameters fixed, change only one predictor variable and watch how the price changes.



stambena_povrsina (flat area) influences the price approximately linearly, with relatively high positive factor.



broj_soba (number of rooms) influences the price approximately linearly, with a small negative factor. So the price actually slightly decreases as the number of rooms increases. This is perhaps counter-intuitive, but this is what the data says.



broj_etaza (number of floors) influences the price approximately linearly, with a small negative factor. This is probably expected, but again, this result is supported by concrete data and not only by intuition.

Conclusion

We see that our simple model provides a solid base for correctly predicting the price. It explains more than 60% of the overall variability which is a solid result for the small dataset and the simple model that we've built.

We have also seen the **exact relationship** between some chosen variables and the price. This can be used to predict or manipulate the price by manipulating the predictor variables.

By getting more data and trying out different parameters and algorithms we could (probably :)) greatly improve this already solid model.

Contact and Info

Danijel Kopčinović, IT Market

Mail: danijel.kopcinovic@itmarket.hr

Tel: +385956472127