Prediction Intervals For Real Estate Price Prediction

Moritz Beck, Rainer Göb

Julius-Maximilians Universität Würzburg

ENBIS-21, 14 September 2021

Table of Contents

German real estate market

real estate prices

- The real estate price index increases by 28.1 points from 2015-2019 [Statistisches Bundesamt(2020)]
- prices have risen in both rural and urban regions.

real estate platforms

- leading platform immoscout24 by Scout24 AG
- monthly users: 20 million

automatic real estate price estimation

definition

- prediction of the value of a property, which exceeds the accuracy of a simple calculation

$$\text{average price per } m^2 \text{ of the region} \times \text{area of the property in } m^2$$

leading plattform in US

- Zillow offers automated real estate price estimates in the business-to-customer area
- Zillow median percentage error of prediction: 7.3 percent

empirical setting

observations

- N real estate objects $i = 1, ..., N$
- each i associated with a price Y_i and feature vector $\boldsymbol{x}_i$
- features include information about size, location, etc.

objective

- predict appropriate characteristics of random price Y conditional on features $\boldsymbol{X} = \boldsymbol{x}$

point versus interval prediction

- point prediction gives no information about prediction uncertainty

- predict price interval $I(\boldsymbol{x})$ as a function of feature vector $\boldsymbol{x}$

- wide interval $\Longrightarrow$ high uncertainty
  narrow interval $\Longrightarrow$ low uncertainty

Example: visual representation of prediction intervals

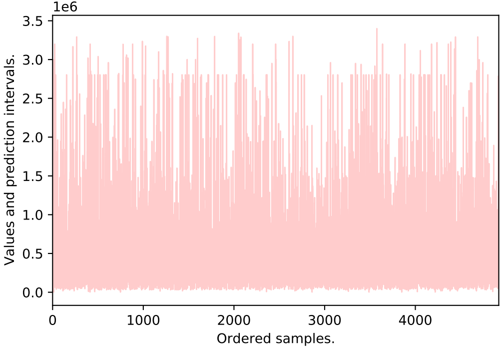

Figure: prediction intervals for Berlin by Random Forest

reliability constraint for prediction interval

- $\mathrm{P}\Big(Y \in I(\boldsymbol{x}) | \boldsymbol{X} = \boldsymbol{x}\Big) \overset{!}{\geq} \gamma$
- price covered by interval with a probability at least γ
- popular choices for γ: $\gamma = 0.90$ or $\gamma = 0.95$

Table of Contents

real estate price prediction: point prediction

- most papers focus on point prediction
- best performance: random Forest Algorithm (RF) by Breiman (2001)
- e. g., Ravikumar (2017), Zhou et al. (2019), Alfaro-Navarro et al. (2020)

real estate price prediction: interval prediction

- interval prediction nearly not considered
- some elementary quantile regression (linear), e. g., Garcia et al. (2019)
- no applications of advanced methods like Support Vector Quantile Regression, Quantile Gradient Boosting, Quantile Random Forest, Quantile KNN

Table of Contents

1) quantile definition by CDF F_Y

- τ-quantile q_τ (quantile of level τ) $=$ smallest y such that $F_Y(y) \geq \tau$

2) quantiles as solution of optimisation problem

- quantile loss function: $\rho_\tau(y) = y(\tau - 1_{(y<0)})$
- τ-quantile q_τ minimises expected quantile loss

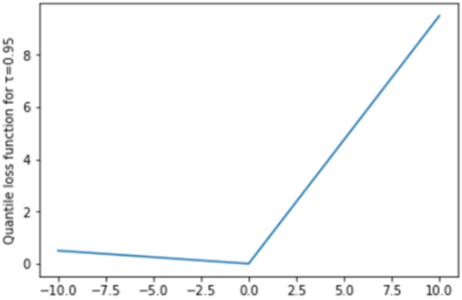

Figure: quantile loss

Table of Contents

quantile regression for interval prediction

- quantile regression: predict τ-quantile $q_\tau(\boldsymbol{x})$ from features $\boldsymbol{x}$
- build prediction interval $I(\boldsymbol{x})$ from quantiles

quantile regression

- consider parametric quantile function $q_{\tau,\theta}(\boldsymbol{x})$
- τ = quantile level, θ = fit parameter
- observe paired training data $(Y_i, \boldsymbol{x}_i)$, $i = 1, ..., N$
- learn quantile function $q_{\tau,\theta}(\boldsymbol{x})$ from training data

1) Quantile regression by minimising quantile loss

- sample quantile loss = average quantile loss over training data
- $\widehat{\rho}_\tau(\theta) = \frac{1}{N}\sum_{i=1}^{N} \rho_\tau(y_i - q_{\tau,\theta}(\boldsymbol{x}_i))$
- minimise $\widehat{\rho}_\tau(\theta)$ in θ to obtain argmin θ_0
- obtain predictor $\widehat{q}_\tau(\boldsymbol{x}) = q_{\tau,\theta_0}(\boldsymbol{x})$

2) Quantile regression by empirical CDF

- fit parameter = CDF $F_Y(\cdot|\boldsymbol{X} = \boldsymbol{x})$
- learn $F_Y(\cdot|\boldsymbol{X} = \boldsymbol{x})$ by empirical CDF $\widehat{F}(\cdot|\boldsymbol{X} = \boldsymbol{x})$
- predict τ-quantile by quantile of empirical CDF

Machine Learning Models For Quantile Regression

Models estimating quantiles over customised loss function

- linear quantile regression
- support vector quantile Regression
- quantile gradient boosting

Models estimating empirical distribution

- Random Forest
- KNN

Stacking method

- use linear combination of methods above
- weights $\longrightarrow$ minimise penalised quantile loss

Goodness Of Fit

- $\widehat{q}_\tau(\boldsymbol{x})$ = predictor of the τ-quantile
- $R^1(\tau)$ score defined by

$$1 - \frac{\text{Sum of quantile losses of full model}}{\text{Sum of quantile losses for level model without regressors}}$$

- high $R^1(\tau)$ score => good fit of the τ-quantile
- low $R^{1(\tau)}$ score => bad fit of the τ-quantile

Table of Contents

Real Life Dataset

- 270.000 real estate objects collected from German platform Immoscout using web scraping
- different cities with between 150 and 15.000 properties each
- different feature types:
  1. numeric: size in m^2
  2. categorical: house type
  3. text: location description
- Use of state of the art methods to
  - select features
  - convert categorical and text features into numeric ones
- Standard preprocessing (standardization to mean 0 and variance 1, outlier detection, etc)

Experiment: Performance through Quantile Loss

Setting

- 90 percent prediction intervals: estimate 0.05 and 0.95 quantiles
- Algorithms:
  1. Quantile Random Forest
  2. Quantile KNN
  3. Quantile Gradient Boosting
  4. Quantile Stacking (combination of 1-3)

Strategy

- Fit one model per city
- 70 percent of the data for training and rest for testing
- Evaluate total quantile loss

Mean quantile loss ($\tau = 0.95$)

	Stacking	Random Forest	KNN	CatBoost
95% confidence interval lower bound	1.0901	0.9722	0.6995	0.7710
mean	1.0949	0.9796	0.7058	0.7759
95% confidence interval upper bound	1.0997	0.9870	0.7121	0.7808
standard deviation	0.7003	1.0806	0.9211	0.7173
test set size	81410			

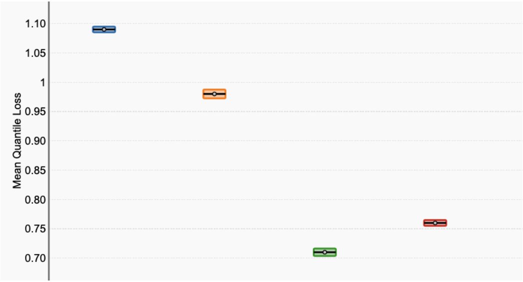

Figure: Mean quantile loss ($\tau = 0.95$)

Mean quantile loss ($\tau = 0.05$)

	Stacking	Random Forest	KNN	CatBoost
95% confidence interval lower bound	0.7693	0.6529	0.5358	0.5797
mean	0.7754	0.6586	0.5413	0.5853
95% confidence interval upper bound	0.7815	0.6643	0.5468	0.5909
standard deviation	0.8866	0.8332	0.8001	0.8215
test set size	81410			

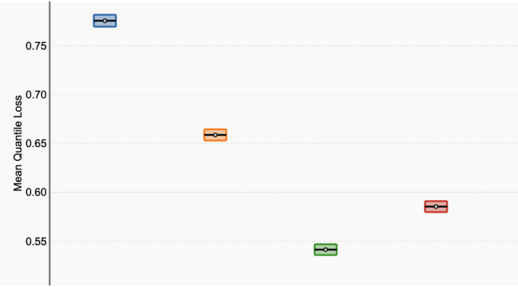

Figure: Mean quantile loss ($\tau = 0.95$)

Conclusion

- best method: KNN Quantile Regression
- no gains from using stacking ensemble

Thanks for your interest!

- moritz.beck@uni-wuerzburg.de
- goeb@mathematik.uni-wuerzburg.de

Bibliography

- Statistisches Bundesamt. (27.3.2020). Development of house prices in Germany in the years from 2000 to 2019, Retrieved 28.02.2021, `https://de.statista.com/statistik/daten/studie/70265/umfrage/haeuserpreisindex-in-deutschland-seit-2000/`
- Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). `https://doi.org/10.1023/A:1010933404324`
- Real Estate Price Prediction Using Machine Learning. Masterthesis (2017). `http://norma.ncirl.ie/3096/1/aswinsivamravikumar.pdf`
- Alfaro-Navarro, Cano, Alfaro-Cortes, Garcia, Gamez, Larraz, A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems. Complexity, vol. 2020, 2020.`https://doi.org/10.1155/2020/5287263`
- Zhou, Xiaolu , Tong, Weitian and Li, Dongying. (2019). Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual Information and Deep Learning. ISPRS International Journal of Geo-Information. `https://doi.org/10.3390/ijgi8080349`
- Garcia, Raul Tomas and Lopez, Maria Francisca and Perez Sanchez, Raul and Marti-Ciriquian, Pablo and Perez Sanchez, Juan Carlos. (2019). Determinants of the Price of Housing in the Province of Alicante (Spain): Analysis Using Quantile Regression. Sustainability. 11. 437. 10.3390/su11020437.