Price Predictions on Airbnb Accomodations in Oslo, Norway

Marei Freitag, Joel Beck

Georg-August-University of Göttingen

21.02.2022

Table of Contents

Introduction

Aims of this work:

- Establish a deep learning approach to predict the price of an accomodation per night
- Focus on explainability and interpretability

→ Underlying Data: provided by Airbnb, contains various information about the listings in Oslo, Norway

Table of Contents

Feature Engineering: Images

- Use transfer learning on a pretrained CNN (ResNet18)
- But have to be sure, the CNN is able to generalize: added Fully Connected Network at the end containing three layers and `ReLU` activation functions
- Also implemented CNN manually as a benchmark model to compare the results

Results:

- pretrained `ResNet18` achieved a Mean Absolute Error of 579 NOK (approx. 58 Euros) on the Validation Set
- But *correlation* of the CNN predictions with the true price was 0.41

Image Predictions

Beispiel Bilder der Predictions!

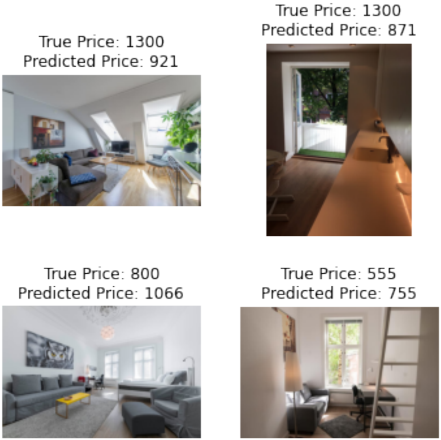

Figure: CNN example predictions

Feature Engineering: Reviews

- Language: Detect language of each review
- Sentiment analysis: Get the sentiment of each review

New Features per listing:

1. Number of reviews
2. Median review length
3. Number of different languages of the reviews as well as a list of the different languages
4. Fraction of Norwegian and English reviews
5. Ratio of negative reviews to the total number of reviews

Feature Engineering: Reviews

Figure: Wordclouds in English and Norwegian

Feature Selection & Data Cleaning

Feature Selection:

1. Manually selected features based on three criteria
2. Analyzed results of different feature selection algorithms

Data Cleaning:

- Converting data types
- Splitting text-based variables into more convenient numeric or boolean features
- Aggregating rare categories of categorical variables into one larger *Other* group to stabilize estimation
- One-Hot Encoding of categorial variables and standardization of numerical variables

Table of Contents

Classical Models

1. **Linear Regression**: simple, well understood in terms of underlying theory and highly interpretable.
2. **Ridge Regression**: still very interpretable with a closed form analytical solution, adds one hyperparameter to the equation
3. **Random Forest**: very flexible model with many hyperparameters determining e.g. the number of regression trees and the tree depth. Can be applied to many contexts and often works 'out of the box'.
4. **Histogram-Based Gradient Boosting**: modern and fast tree-based gradient boosting algorithm.Comes with a large number of tunable hyperparameters, some of them similar to the Random Forest parameters, some of them more specific to the *Boosting* instead of the *Bagging* approach such as the learning rate.

Neural Network

explain Hyperparameter and Architecture

Model Architecture

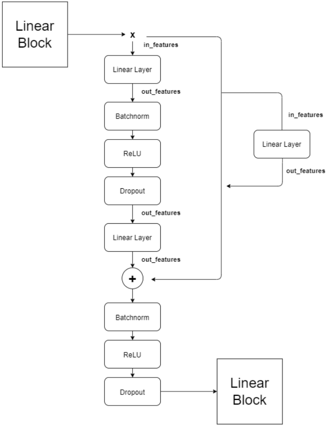

Figure: Architecture of each Block in the Fully Connected Neural Network

Table of Contents

Test - Slide 1

Table of Contents

Test - Slide 2

Thanks for listening!

Questions?