

Paper Structure

Contents

1	Introduction	2
2	Methods	2
2.1	Preprocessing	2
2.1.1	Feature Engineering	2
2.1.2	Feature Selection	2
2.1.3	Price Distribution	2
2.2	Models	2
2.2.1	Classical Models	2
2.2.2	Neural Network	2
3	Results	3
3.1	Predictive Performance	3
3.2	Explanations and Interpretation	3
4	Conclusion	3
5	Appendix	3
6	References	4

1 Introduction

2 Methods

2.1 Preprocessing

2.1.1 Feature Engineering

Images

- Discuss if figure of cnn examples can be moved to appendix
- Discuss if first two sections about webscraping and preprocessing can be moved to appendix

Reviews

- Description of Sentiment Analysis, stating procedure and results and including **Figure** with Wordcloud, either only English Words or Side-by-Side Wordclouds of English and Norwegian Words
- In addition: Language Detection to include the *number of different languages* and the *fraction of norwegian languages* and Analyzing the reviews lengths to include the *median review length*
- Since there are multiple reviews per apartment the results for each review were averaged for each apartment separately.

2.1.2 Feature Selection

2.1.3 Price Distribution

- Discuss if figure of price distribution can be moved to appendix

2.2 Models

2.2.1 Classical Models

2.2.2 Neural Network

- Discuss if figure of dropout impact can be moved to appendix

- Discuss if second section about further hyperparameters can be moved to appendix

3 Results

3.1 Predictive Performance

- **Figure** of performance comparison between selected classical models and neural network for given feature selector (e.g. RFE) and different number of selected features
- Interpret Differences in Training and Validation Performance between different models
- Interpret Differences in Performance for different number of selected features
- Compare Performance on Validation Set with Performance on Test Set for the best model of each class by means of a table
⇒ Models whose hyperparameters were tuned on validation set generalize worse to test set, e.g. `HistGradientBoosting`, `RandomForest` and `Ridge`
- Include average predictions of top 2/3/4/5 models, where models are selected based on validation set performance and Test Set predictions are averaged
- Potentially mention which models contributed to predictions on new, unseen dataset from challenge (only in presentation)

3.2 Explanations and Interpretation

- Discuss if coefficient plot can be moved to appendix

4 Conclusion

5 Appendix

- include link to repository with codebase to reproduce all findings

- include images of: cnn examples, price distribution, dropout impact, coefficient plot
- include webscraping and preprocessing of images
- include second section of mlp

6 References