Paper Structure

Contents

1 Introduction

2 Methods

2.1 Preprocessing

2.1.1 Feature Engineering

Images

- Description of comparing custom CNN with pretrained `ResNet`
- Explain procedure of scraping images from website, transforming images and adding custom layer with other weights frozen
- Describe Results and Predictive Performance solely by images
- Put Performance into context by showing **Figure** of sample images with true and predicted price

Reviews

- Description of Sentiment Analysis, stating procedure and results and including **Figure** with Wordcloud
- Mention other reviews features, e.g. language detector

Others

- Optionally mention all other features that we added to the dataset

2.1.2 Manual Feature Selection

- Describe process of combining features from images, reviews and numeric features into one dataframe
- Describe decision making which features were kept in the model
- Mention distinction between a selected subset and fitting a regression model with (almost) all features as hints for important features

2.1.3 Encoding and Automatic Feature Selection

- One-Hot Encoding of Categorical Variables, Standardization of Numeric Variables
- Experimenting with different ways of algorithm-based feature selection / dimensionality reduction
- Focus on `PCA` as most theoretically supported procedure and `RFE` as procedure we chose
- PCA has advantage of reducing dimensionality and simultaneously producing *uncorrelated* features, disadvantage of producing linear combinations of original features, harder to interpret
- RFE showed best performance and selects subset of original features, can be immediately interpreted as potentially most important features

2.2 Models

2.2.1 Classical Models

- serve as benchmark models to better evaluate performance of custom neural network
- selected with increasing degrees of complexity
- focus on 4-5 models, e.g. Linear Regression, Ridge Regression, Random Forest, HistGradientBoosting
- Describe Model Fitting process and hyperparameter tuning with Randomized Search Cross Validation

2.2.2 Neural Network

- Explain Architecture of Neural Network
- **Figure** with diagram of `Linear Block`
- Explain Choice of Model Parameters and Hyperparameters and emphasize which components were most important
- **Figure** with impact of `Dropout`

2.2.3 Price Distribution

- **Figure** of Side-by-Side Histograms of Price and Log-Price Distribution
- Price Distribution right-skewed with some very large outliers
- Explain benefits of normally distributed dependent variable, particular for models with distributional assumptions (e.g. Linear Regression vs. Neural Network)
- State that all classical Machine Learning Models benefitted from log transformation
- Briefly discuss why we did not transform the price variable for the Neural Network

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

3.2 Explainability / Interpretability

3.2.1 Feature Importance

- **Figure** of Coefficient Plot for Linear Regression with e.g. 25 selected features
- Interpret Figure

3.2.2 Sensitivity of Neural Network Performance on Outliers

- State shortcomings of Neural Net to predict prices in the tails of the distribution, error metrics thus largely impacted by outliers
- State (maybe with a table) drastic increase in predictive performance when excluding largest quantiles of price distribution from dataset
- Discuss if the task itself is theoretically feasible for any kind of model
- **Figure** of latent space representation
- Discuss that the data is not expressive enough to capture all features that determine the price in reality, particularly for apartments with very high prices, that do not differ from lower priced apartments based on their feature set
- Question underlying assumptions that all apartments are reasonably priced, difficult to detect overpriced listings that bias model predictions

4 Conclusion

5 Appendix

- include link to repository with codebase to reproduce all findings

6 References