BI2023_gr37_12238682_11711533

Philipp Rettig, 11711533, Person B[*]
Vienna University of Technology
Vienna, Austria
e117115332@student.tuwien.ac.at

Vladimir Panin, 12238682, Person A
Vienna University of Technology
Vienna, Austria
e12238682@student.tuwien.ac.at

CCS Concepts: • **Data Mining**; • **Data Analytics**; • **Business Intelligence**; • **Machine Learning**;

Keywords: Data Mining, Business Intelligence

ACM Reference Format:
Philipp Rettig, 11711533, Person B and Vladimir Panin, 12238682, Person A. 2024. BI2023_gr37_12238682_11711533. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Business Understanding

In this class assignment, the objective is to address a data analytics problem using the CRISP-DM Process. The task involves navigating through the data analysis process within the constraints of a simulated, rather than a real-world, scenario. However, a general understanding for the CRISP-DM Process is to be obtained.

1.1 Scenario

The dataset deals with AirBnB data from Berlin in July 2021. The dataset was taken from "About Inside Airbnb", which is as described by themselves a "mission driven project that provides data and advocacy about AirBnB's impact on residential communities". The data is obtained by scraping the website [1]. An imagined scenario is that an AirBnB provider wants to set an appropriate price for the apartment being offered on the website.

1.2 Business Objectives

The business objective is to be able to predict the price based on the information given about this apartment. It has to be considered that for higher prices, the apartment could be available more frequently. This trade-off has to be considered.

[*]Both authors contributed equally to this research.
[1]http://insideairbnb.com/about/

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.3 Business Success Criteria

The Business Success Criteria are maximizing the annual revenue and find the ideal price. The goal might therefore lead to finding a balance between the price and the average vacant time. The goal therefore can considered to be minimizing the RMSE value in price predictions.

1.4 Data Mining Goals

The data mining goals are, first of all, to predict the price based on the information given. Furthermore, variables which have an influence on the price should be identified. In our case these can be based on different groups, e.g. districts, which have higher or lower prices or another possibility, listings that are vacant more often than others. Finally, these goals should help in finding the right balance for predicting a price, especially considering availability and location.

1.5 Data Mining Success Criteria

The success criteria for the data mining goals encompass several key aspects. Firstly, to measure the effectiveness of the model that is going to be build in the process of this assignment, we will utilize metrics like Mean Absolute Error or Root Mean Squared Error. Since standalone values of these metrics will not be very meaningful, we aim to compare them against other implementations of price predictions on the same dataset [2] and use this as a benchmark for our own implementation. All in all, the model should be able to give out fairly accurate price predictions and maybe could help us understand reasons or important variables for price differences.

1.6 AI risk aspects

One possible AI risk aspect might be concerning bias and fairness. As the names of hosts are provided in the dataset, the result might discriminate on the basis of names. A possible implication might be that non "German-sounding" names could get attributed a lower price or could implicate a less wealthy district. Additionally, when using AI in housing data, there's a risk that the system might unintentionally favor certain areas. For instance, it could perpetuate biases related to neighborhoods or historical socioeconomic factors. Another concern is that the AI might rely too much on outdated information, missing out on current trends or changes in the housing market, which especially in bigger urban areas

[2]https://www.kaggle.com/code/lennarthaupts/airbnb-prices-in-berlin/notebook

change rapidly. This could lead to Historical Bias. To address these risks, it's important to evaluate and update the produced regularly. Finally, another possible bias might be the Omitted Variable Bias as possible variables like for example, the average rating of a host, the host's age is not considered.

2 Data Understanding

This section outlines a comprehensive data analysis process, covering attribute types, statistical properties, data quality considerations (such as missing values, distributions, outliers, and provenance), visual exploration, ethical sensitivity assessment, identification of potential biases, and the formulation of necessary actions in data preparation based on the analysis results.

2.1 Attribute types and their semantics

The given attribute types and their semantics are described via :

1. id has a nominal attribute type: The Listening id of the housing offer
2. name has a nominal attribute type: The title of the offer
3. host_id has a nominal attribute type: The ID of the host
4. host_name has as nominal attribute type: The first name of the host
5. neighbourhood_group has a nominal attribute type: The district of the housing offered
6. neighbourhood has a nominal attribute type: A more precise description of the location inside the district
7. latitude has a ratio attribute type (although it could be debated if a true zero point exists): Latitude of the hosing offer
8. longitude has a ratio attribute type: Longitude of the housing offer
9. room_type has a nominal attribute type: Type of the housing offer
10. price has a ratio attribute type: Price of the housing offer per night
11. minimum_nights has a ratio attribute type: Amount of minimum nights to book this housing offer
12. number_of_reviews has a ratio attribute type: Number of total reviews per month for this housing offer
13. last_review has an interval attribute type: Last review for this housing offer
14. reviews_per_month has a ratio attribute type: Number of reviews per month for this housing offer
15. calculated_host_listings_count has a ratio attribute type: Amount of listenings per for host
16. availability_365 has a ratio attribute type: Amount of available days in a year

2.2 Statistical properties describing the dataset including correlations

For the numerical variables the statistics for the quantiles, min and max, the count and the standard deviation are shown in the following three tables: 5a, 5b and 5c. It is to note, that prices can get quite high up to 8000 €. It has to be checked, whether these offers are plausible. Furthermore, there is at least one room that is vacant for the whole year, 365 days. Also, for the maximum minimum_nights, there is an instance with 1124, which is probably an error, as this would be more of a rental offer with a time span of more than three years. This will be investigated further in the data preparation steps. The other variables based on the summary seem to follow a reasonable distribution.

2.3 Data Quality

There is comparably few missing data. In the dataset there are 19095 rows and 16 columns. The following table shows the missing data 1. From the table we can derive, that for a sizeable portion of the instances the information about the reviews is lacking. Hence, this might make the predictions more difficult, as the variables "last_review" and "reviews_per _month".

Variable	**Number of NaNs**
last_review	4155
reviews_per_month	4155
name	30
host_name	12

Table 1. Number of NaNs in Each Variable

The outliers were discussed in the section 2.2

2.4 Visual exploration

During the exploration of the data, many insightful plots were created, of which a handful are going to be shown and explored. Figure 1 shows the median prices for the various neighborhoods. The x-axis represents the neighborhoods, while the y-axis indicates the mean price scale. It is evident that Neukölln and Reinickendorf offer the most affordable accommodations, whereas Charlottenburg-Wilmersdorf and Mitte host the highest-priced listenings. Additionally, it is worth mentioning the significant price variability observed in the Spandau and Marzahn-Hellersdorf neighborhoods.

Figure 2 shows the correlation matrix for the numerical variables. One can observe that there is a high positive correlation for the three review variables, which was to be expected. However, the price doesn't show high correlation with any of the numerical variables. The highest correlation can be observed with the availability_365 variable.

Finally, it can be pointed out that some numerical variables follow a right skewed distribution. An example is given by

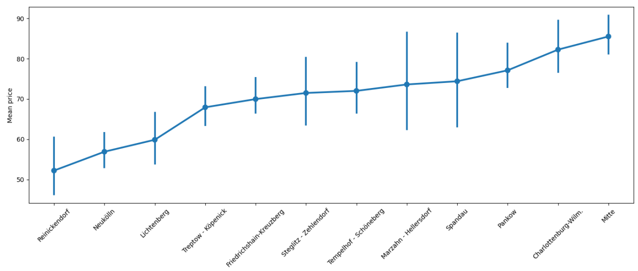

Figure 1. Mean price by neighborhood

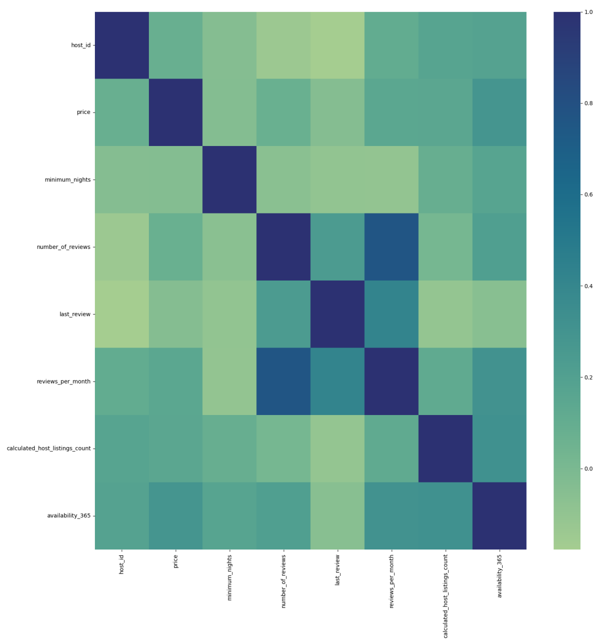

Figure 2. Correlation matrix

the variable "availability_365" following plot 3. The skewness will be further analyzed in the data preparation section.

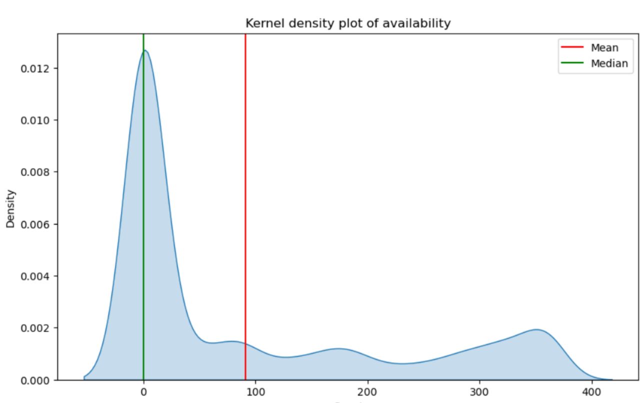

Figure 3. Kernel Density for variable availability_365

Table 2. Frequency of Neighborhoods in Berlin

Neighborhood	Count
Friedrichshain-Kreuzberg	4197
Mitte	4173
Pankow	2981
Neukölln	2608
Charlottenburg-Wilm.	1567
Tempelhof - Schöneberg	1371
Treptow - Köpenick	620
Lichtenberg	612
Steglitz - Zehlendorf	415
Reinickendorf	271
Marzahn - Hellersdorf	142
Spandau	138

2.5 Ethically sensitive information

Ethically sensitive information in this dataset, could be the distribution of names of the host. It is to say that the top 10 most occuring feature hosts are Anna, Michael, Julia, David, Baharbin, Daniel, Flo , Martin, Laura and Jan. Therefore, the classes with not classically German names could use a special sampling strategy. Apart from that, not ethically sensitive information is stored.

2.6 Risk and Bias

It has to be pointed out that some districts have much more entries that other one's. This can be seen from the counts of the different neighbourhoods 2. While Friedrichshain-Kreuzberg and Mitte have the most entries, different neighborhoods have much less offers. Therefore, this could skew the model and appropriate techniques such as under - or oversampling could be used.

An expert could answer the questions whether different housing options like hotels for example also have the same distribution around the neighborhoods. However, it would also make sense to compare to the population of the districts and the youth population, to obtain insights about the distribution itself.

2.7 Actions in data preparation

As stated before in the section 2.2 a few variables had unreasonable ranges. These outliers have to be further investigated and dealt with by for example dropping or imputing with reasonable values. It has to be considered that machine learning methods usually can't deal with categories, Therefore they will be converted to a one hot encoded representation. Possibly, connections between the "host_name" and the price could be found. Additionally, the variables have to be converted to reasonable types. Going on, missing values have to be dealt with by for example imputation.

3 Data preparation

This section comprises the data preparation phase, including analysis of the possible and necessary steps as well as incorporation of additional external data sources.

3.1 Necessary actions

Using the information gathered in the data understanding section, a few necessary actions have to be performed.

3.1.1 Missing values. Using the information layed out in section 2.3, we decided to remove the entire columns for name and host_name, taking into consideration the high number of unique values in them. After some further investigation, we figured that the missing values in last_review and reviews_per_month always occurred when a listing had zero reviews. Thus, the strategy was not to drop the missing values, but to impute them with a numerical value. For the reviews_per_month we replaced the NaNs with zeros and for the last_review we first transformed the date to "days since last review" and then replaced the NaNs with the highest value.

3.1.2 Outliers. Some listings in the dataset did show a price of zero. We considered this an error in the data and removed the corresponding rows. Other outliers in this column included listings with a price of 4000 or 8000. Such high values are not very frequent, but still we thought it reasonable that high valued listing like these could exist and decided to keep them unchanged. The minimum nights attribute includes some high values up to 1124 nights for a few listings, see figure 4, which could be valid, since there probably exists some hosts who focus on long term rentals. However, for these few values to not have a strong impact on the models performance, we decided to only focus on short term rentals and removed all rows with "minimum nights" > 30.

3.1.3 One hot encoding. For our machine learning model to be able to handle the categorical input features, some variables are one to n encoded. These comprises the attributes neighbourhood group and room type. It has to be noted that this procedure increases the dimensionality of our dataset by the number of categories -1. Considering the high number of records (19000) this should have a sizeable influence.

3.1.4 Attribute removal. As a final necessary step we considered to remove some attributes, which might not be that important for the data mining goal or simply not useful for the model:

- id and host_id: not useful for the model, only unique values
- name and host_name: too many unique values and difficult to group them further
- latitude and longitude: instead we create a new feature (see section 3.2)

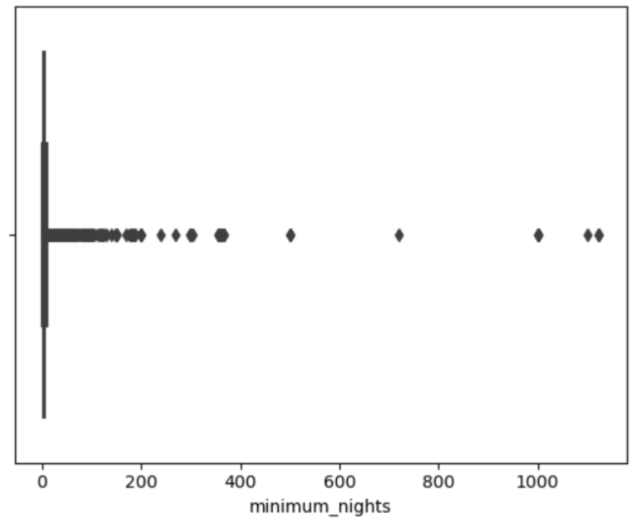

Figure 4. Boxplot minimum nights

- neighbourhood: too many categories (138), instead use the neigbhourhood group

3.2 Derived attributes

As mentioned in section 3.1.1 the attribute "last_review" is transformed to a new attribute, which represents the days since the last review instead of the date of the last review. The numerical representation should aid in better understandibility for the machine learning model. Additionaly, the latitude and longitude information is used to create a new attribute, which represents the distance to the berlin city center. The chosen coordinates are subjective: (52.51638889, 13.37888889). The distance is calculated using the hamming function and the resulting values are given in kilometers.

Other potentials include the transformation of the host_name to a binary category indicating, whether the name is German sounding or not. We speculate that this might influence some AirBnB guests on choosing a listing. However, as this would take a very long time to manually annotate, we decided not to perform these steps.

3.3 External data sources

The following section will discuss options about additional data sources and attributes.

To enhance the dataset one option is to consider **demographics** about the neighbourhood, like population size and density, average income, proximities to public transport or age distribution. These could provide some further insights into the different regions and why prices differ.

Especially useful for airbnb listings are probably **events** happening in the near area. Hence, the data could be enriched with information about local events, to provide better understanding of price differences at different times.

As already mentioned, it would be ideal to make use of the host name and classify it as German sounding or not. Ideally, the data could be enriched with an extensive list comprising **German names**, which could be used to perform the aforementioned binary classification.

Lastly, a more extensive analysis using **additional months** would be preferred. This would allow use to consider temporal trends and perform monthly comparisons.

3.4 Other preprocessing steps

In this section we outline some additional preprocessing steps we considered, but didn't deem imperative for a successful machine learning model.

3.4.1 Scaling. Most of the numerical values in the dataset, are simply count based, which means they are in the same range of magnitude. However, the attributes "distance to city center" and "reviews per month", are not simply counts, but feature a different range that might not be easily comparable. To avoid for an explanatory variable to have an unreasonable impact in the model, we decided to align the scale for each using a min max scaler. It has to be noted though that we will lose some sense of interpretability using this procedure.

3.4.2 Log transformation. By utilizing a kernel density plot we did get some idea of the distribution of the numerical attributes. Most of them did show a slightly right skewed distribution. However, the median and mean seem to not deviate too much in most cases. Therefore, we are not sure if log transforming these attributes will have a significant impact in the performance of the model, especially considering that we will also lose some sense of explainability. After log transforming the number of reviews and reviews per month, the distribution seems to be more normally distributed. Eventually , we decided against transforming the attributes, but keep this option in mind in case the results are not satisfactory.

4 Modeling

4.1 Data mining algorithms

In selecting the most suitable data mining algorithm for the experiments a possible flowchart for selecting an appropriate algporithm was chosen according to the documentation by scikit.learn.[3] Considering the nature of the regression task at hand and the number of samples and an assumption of the strong importance of certain features Linear Regression was chosen as an appropriate choice. In this case three algorithms were tested: Elastic Net, Ridge Regression and Lasso Regression. It can be pointed out, that Linear Regression allows the quantification of the importance of variables and therefore analyzing the importance or the lack of importance of certain features. Hence, landlords can identify important features for selecting real-estates for renting it on AirBnB.

4.2 Hyper-parameters

In the context of Elastic Net, Ridge, and Lasso Regression, key hyper-parameters can have significant influence towards model performance. For Elastic Net, the primary hyper-parameters include alpha, which controls the weight of the penalty term, and the l1_ratio, determining the balance between L1 and L2 regularization. The choice of alpha values, representing the regularization strength, is crucial as it regulates the trade-off between under fit and over fit A grid search is often employed to explore a range of alpha values. The selection of these hyper-parameters should was tested using cross-validation to strike a balance between model simplicity and predictive performance. Regularization techniques like Ridge and Lasso, which were also used, also included the penalty weight factor alpha. In our case for alpha the values [0.1, 0.3, 0.5, 0.7, 1.0] were provided and for the l1_ratio the values [0.1, 0.5, 0.9] in the Grid Search.

4.3 Train / Validation / Test Set Split

The provided code segment adeptly divides the dataset into training, validation, and test sets, with 70% allocated for training, 15% for validation, and the remaining 15% for testing. This three-way split serves to train the model on a substantial portion of the data, fine-tune its hyper parameters using the validation set, and ultimately evaluate its performance on a separate test set. This split was chosen as it is widely used. Additionally, other train / validation / test splits could be evaluated, however this wasn't provided in the task description. For reproducibility a random state of 42 was provided.

4.4 Training of the model & Documentation of all parameter settings

After training the model using the different hyper-parameters considering the alpha values [0.1, 0.3, 0.5, 0.7, 1.0] and the l1_ratio values [0.1, 0.5, 0.9] cross-validation was performed using the Grid Search. The visualization of the training results for the Elastic is given by 5 and for Ridge and Lasso via 6.

4.5 Performance metrics

For the evaluation of the performance the RMSE value was used. The RMSE values were reported for the three methods for the train and test data and the most important variables were provided. The results for the different RMSE values are given by the table 6 and the most important features are given via table**??**. It can be highlighted, that even though the RMSE values can be regarded as quite close, the three most important features differ. Generalizing, it can be said that the room type plays a crucial role for the price.

[3]https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

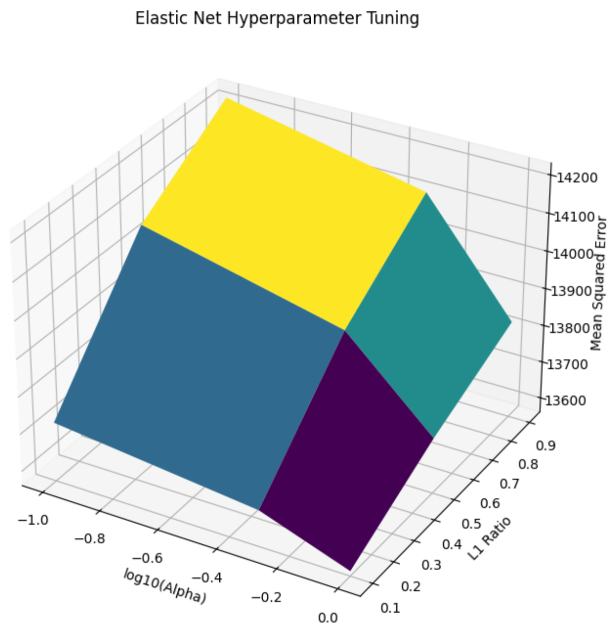

Figure 5. Elastic Net Hyper-parameter tuning

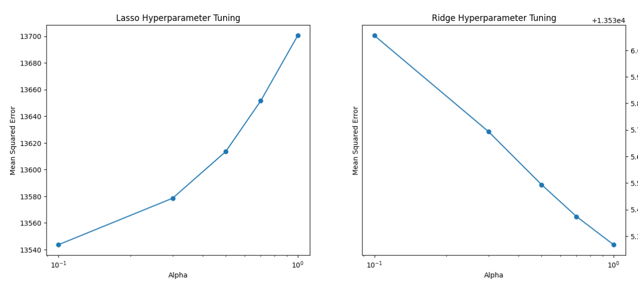

Figure 6. Lasso Ridge Hyper-parameter tuning

5 Evaluation

5.1 Final model

The final model chosen for the task of predicting prices of AirBnb listings is ridge regression. This decision was based on its superior performance, albeit marginally. Additionally, ridge regression is known for its robustness against co-linearity.

The parameter optimized ridge regression was initially trained on the split training set using cross validation and evaluated on the test and train set. The calculated RMSE for the test set is 161.67, whereas the RMSE on the train set is approximately 116. This might indicate a reasonable level of model generalization. Even when retraining the model with the identical alpha parameter using the full train and validation split, the performance on the test set stays the same. The reported RMSE is only slightly higher at 161.76, further showing the model's robustness and consistency. However, the standard deviation with the same value of

Table 3. Price segmented RMSE values

Segment	Ridge RMSE	Baseline RMSE	Price Range
Low price	29.10	43.44	0 - 40 €
Mid price	30.43	19.62	40 - 70 €
High price	280.02	285.21	70 - 8000 €

161.7 might express a relatively high error in terms of price predictions, considering most listings are priced below 300 euros.

5.2 Performance comparison

To further evaluate the performance of the chosen model, comparisons with benchmark models and trivial baseline models have to be considered.

5.2.1 Benchmark. To the best of our knowledge, no literature or solid internet publications exist on the same task and dataset. However, the provider of the dataset hosts a separate analysis notebook on kaggle, which will be considered as a benchmark [4]. The reported RMSE for the same price prediction task is about 34 and the chosen method is a random forest regressor. The value is quite a drastic difference to our performance of around 161. Possible reasons for the high variation include distinct features as well as differences in outlier handling and scaling of the data. Especially interesting is the chose in the benchmark model to exclude all listings with a price higher than 300 euros.

5.2.2 Baseline. For the chose of a baseline a simple model was chosen, which always predicts the mean. Applying the baseline on the test set, the calculated RMSE is 165.32. Thus, the ridge regression performs slightly better than a simple baseline.

Given the Ridge regression's relative success over the baseline but underperformance compared to the benchmark, it is assumed that the high error rate is influenced by the skewed distribution of the price variable. To investigate this:

5.2.3 Segment wise price comparison. Given the Ridge regression's relative success over the baseline but underperformance compared to the benchmark, it is assumed that the high error rate is influenced by the skewed distribution of the price variable. To further examine this hypothesis, the dataset's price variable was segmented into three distinct ranges. These segments were determined based on the 33rd and 66th percentiles of the price distribution in the test set. Table 3 presents the RMSE values for each of these segments, offering a detailed view of the model's performance across different price levels.

The results support the hypothesis that high prices have an influence on the RMSE values. Unfortunately, this also

[4]https://www.kaggle.com/code/lennarthaupts/airbnb-prices-in-berlin/notebook

Table 4. Neighbourhood group segmented RMSE values

Neighbourhood group	Ridge RMSE
Reinickendorf	42.36
Lichtenberg	56.46
Neukölln	60.07
Spandau	66.91
Tempelhof - Schöneberg	77.53
Friedrichshain-Kreuzberg	81.46
Pankow	85.07
Treptow - Köpenick	92.49
Marzahn - Hellersdorf	91.61
Steglitz - Zehlendorf	102.31
Mitte	351.11

shows that high priced listings will be not possible to predict very reliably.

5.3 Evaluation of success criteria

The primary business objective was to predict the price of Airbnb listings in Berlin, with an emphasis on finding the ideal balance between price and vacancy rate. The success of the process, therefore, hinges on the model's ability to accurately predict prices.

While the model provided slightly better results than the baseline, it fell short compared to the benchmark. However, the model demonstrated a robust ability to predict prices in the low and mid range, with a performance comparable to the benchmark model. The underperformance in the high price segment suggests that the model predictions for luxury listings are less reliable. Nonetheless, it is important to consider that these listings are less common and might not significantly impact overall revenue. Additionally, high priced listing often have unique characteristics that are challenging to capture in a generalized model. In conclusion, it is advised to narrow the model down to only predict prices in the low to mid range and construct a separate model with additional data to accurately predict luxury listings.

5.4 Model bias

In absence of any actually sensitive attribute in the dataset after preprocessing, the neighbourhood group is chosen as a proxy for geographic bias and as a means to investigate potentially skewed performance.

To report on the performance of the different neighbourhoods the following approach was taken:

- Split test data into neighbourhoods
- Predict the prices on the neighbourhood split test set
- Calculate and report the RMSE values

The results in table 4 show that the performance is especially bad for the neighbourhood group Mitte.

As one might suspect, high priced listings will most likely be found in the center of a city. To investigate a possible relationship between the price and the high error regarding the neighbourhood, figure 7 show the RMSE vs the proportion of listings above 300 € in a neighbourhood.

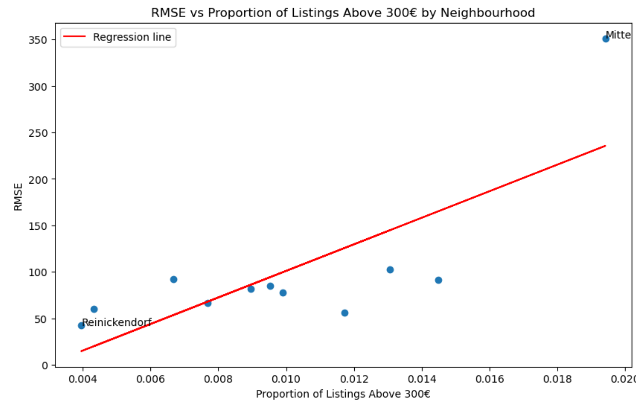

Figure 7. RMSE vs proportion of high priced listings per neighbourhood

In addition to the highest RMSE, the neighborhood group Mitte also has by far the highest proportion of listings above 300€, suggesting the larger error is due to the larger prices. The regression line displays a positive relationship. However, this might be heavily influenced by Mitte, as the other neighborhoods only demonstrate a slight positive correlation. Moreover, it has to be noted that sample size per neighborhood differs dramatically, which might also account for smaller differences in the error values.

6 Deployment

6.1 Business Objectives

Considering the business objectives are concerned with accurately predicting the prices of listings, the usefulness of the model is highly dependent on its performance, quantified by the RMSE. As explained in section 5 the models performance on high priced listings, over 300 €, was generally not satisfactory, even compared to the mean baseline. Thus, the goal for deployment is to only focus on a part of the data space. Specifically, predictions of listings with results over 300 € will be displayed with a warning, stating that the models performance might be inaccurate and that the results should be carefully interpreted.

For further analysis and to properly fulfill the business objectives, it is advised to examine which characteristics are associated with luxury listings. The insights of this analysis could be used to refine the model and adapt it to a broader price range. Furthermore, the price prediction is limited by the available data, which focuses only on the listings in July 2021.

There are several possibilities for improvement of the model:

- Gathering data with a broader time-horizon covering all 12 months
- Accounting for fluctuations on the housing market as market dynamic is constantly changing

6.2 Potential ethical aspects

As far as the risk assessment there is potential for the model to be misused by hosts or guests. By identifying potentially attractive investment opportunities, this can lead to increased rental prices and displacement of local residents and hence to gentrification, which also features a political dimension. Also many apartment buildings therefore might be bought specifically by investors and neighborhoods therefore more divided by income, hence possible result in more gentrification.

In addition to these risks, no other ethical concerns could be identified. However, users must carefully consider the limitations of the model to ensure its responsible and ethical application.

6.3 Aspects to be considered during deployment & Triggers

Monitoring of the price prediction system should enable oversight by humans, in case the predictions deviate from reasonable values. This includes negative or zero predictions, as well as unreasonably high prices. To keep track of inconsistent outcomes, the system will include logging capabilities. After several consecutive incorrect predictions, a warning will be issued to the maintenance team.

6.4 Reproducibility

Reproducibility ensures the consistency of the models and results generated. A critical point is splitting the data into a training, validation and test dataset. For reproducibility issues the term *random_state=42* was added, to set a seed. Apart from generating the test design all the model steps are reproducible.The provided code is submitted in the final hand-in, enabling others to replicate the modeling process and verify the results.

7 Overall findings

Overall the project can be regarded as a success. The prediction for housing prices was fairly accurate for entities under 300 €. This may suggest that for higher listings factors that are not accounted for like floor number, size of the apartment or other softer factors like Instagram-mentions or personal references might be more important. Though this will generally have to be considered in a new setting with more data of high priced listings, as the target variable is highly skewed. However, the overall task of predicting a price for an apartment was fulfilled. The lesson learned was to iterate through the CRISP-DM process and to systematically work through potential risks and biases with diligence.

8 Appendix

Variable	count	mean	std
`latitude`	19095	52.510215	0.032391
`longitude`	19095	13.404654	0.062953
`price`	19095	73.303221	136.249622
`minimum_nights`	19095	9.105944	33.635956
`number_of_reviews`	19095	21.637078	48.670427
`reviews_per_month`	14940	0.718274	1.445272
`host_listings_count`	19095	3.135847	7.773246
`availability_365`	19095	91.271694	127.645330

(a) Part 1

Variable	min	25%	50%
`latitude`	52.340070	52.489710	52.509950
`longitude`	13.097150	13.367160	13.414090
`price`	0	35	52
`minimum_nights`	1	2	3
`number_of_reviews`	0	1	4
`reviews_per_month`	0.01	0.09	0.27
`host_listings_count`	1	1	1
`availability_365`	0	0	0

(b) Part 2

Variable	75%	max
`latitude`	52.533320	52.656110
`longitude`	13.438900	13.757370
`price`	81	8000
`minimum_nights`	5	1124
`number_of_reviews`	17	620
`reviews_per_month`	0.83	94.35
`host_listings_count`	2	76
`availability_365`	175	365

(c) Part 3

Table 5. Numeric Summary

Model	RMSE Train	RMSE Test
Elastic net	116.357137	161.762317
Lasso Regression	116.233843	161.719027
Ridge Regression	116.205063	161.759292

Table 6. Model Performance