Data Mining Personal Project Report

Sinii Viacheslav

May 3, 2023

Abstract

My client is an Airbnb Management company. When a host does not want to deal with mundane tasks of listing's maintenance, tax payments, etc, they can ask an Airbnb Management company to do all of that for a share of the income.

1 Business Understanding

The main thing that a company wants is money. The income from listings may be separated in several factors: listing's price, amount of listings, frequency with which customers visit the listings, whether the customers return to the same place, whether the customers are interested in the listing in the first place.

In this work I will focus on customer attraction. Since Airbnb Management companies can participate in the listing's page design and maintenance, they are interested in providing the best design and may be willing to pay money for data-driven analysis of important points.

Business Objective. Increase number of users that book company's listings.
Secondary Business Objective. Customer trust is of a great importance for the company. Therefore, the designed system should provide recommendation and insights which, after their implementation on the page, will not mislead the customer. In other words, no click-bate. The customers should leave the place without feeling fooled.
Success Criterion. Newly designed listings increase user attraction by 10%.

1.1 Assess Situation

Resources.

- **Hardware and compute** - Google Colab platform with GPU, a sufficient amount of CPU compute and RAM.
- **Personnel** - No expert support.
- **Software** - Only open-source software is available for data mining.
- **Data** - Listings and reviews scraped from Airbnb.

1.2 Requirements, Assumptions, and Constraints

Requirements.

- Deadline - May 3, 2023.
- Report is presented at the end of the project.
- Either a model is built successfully or the reasons for a failure are presented (based on data properties)

Assumptions.

- The company is able to use the results of the study, to influence the hosts to make changes to the listing description.
- The dataset includes only listings in the Paris city. Thus, the result cannot be straighforwardly transferred to other cities.

Constraints.
No constraints.

1.3 Risks and Contingencies

- High load on other projects leads to the violation of the deadline - work at nights and ask for an extension.
- Data is of poor quality - look for other sources of data.

1.4 Costs and Benefits

Cost - 0$
Benefits. Up to 495$ in a month. The average price of a listing is 120$. The difference in occupied days between 'bad' and 'good' listings is 4. Thus, if we are able to improve 'bad' listings, number of occupied days should grow bringing more revenue.

All calculations can be found in the notebook 'benefit.ipynb'.

1.5 Data Mining Goals

The objective of this research is to build a model which classifies a listing as 'low-rated' or 'high-rated'. Further, we need to interpret the model to highlight the most important features.
Success Criterion. F1-score is ≥ 0.8. Model finishes running in a meaningful time and does not require a lot of resources.

1.6 Project Plan

Phase	Duration	Resources	Risks
Business Understanding	1 week	Business Unit	Objective is vague
Data Understanding	1 week	Data Scientist, Colab	The available data cannot help the business objective / Poor data quality
Business Understanding	1 week	Business Unit, Data Scientist	Objective is not concretisized after review
Data Preparation	2 weeks	Data Engineer, Data Scientist, Colab	Hardware constraints / No data is left after cleaning
Modeling	2 weeks	Machine Learning Engineer, Colab	No correlation between data and target / Low score
Data Preparation	2 weeks	Data Engineer, Data Scientist, Machine Learning Engineer, Colab	The data cannot be improved further
Evaluation	2 weeks	Analyst, Colab	Success criteria are not satisfied / The output is not statistically significant

2 Data Understanding

2.1 Data Collection

Source of data: `https://www.kaggle.com/datasets/mysarahmadbhat/airbnb-listings-reviews`.

2.2 Data description

The data is stored in 'csv' format.

The dataset contains 33 columns with:

- listing-host pairings
- client review scores
- basic information about the listing
- basic informantion about host's profile
- price
- listing's location - coordinates, city, district, neighbourhood, country
- amenities

2.3 Review scores

Almost all review scores are close to the maximum value. Customers are inclined to give good scores even when the apartment is normal, and only awful listings will get low scores. That means two things: first, a low score should be a good indicator of the 'awfulness' of the listing; second, the data is highly imbalanced and requires a careful processing.

The next observation is that 1/3 of values are missing, these rows must be dropped leading to a loss of data.

Review scores are numerical columns and cannot indicate exactly what is the problem. Airbnb allows to write reviews elaborating on the score, but the dataset lacks this information - 'reviews.csv'. Thus, it may be problematic to explain customer preferences to the client.

2.4 Price

Contains outliers, needs preprocessing.
For correct use of this column I will need to take into account the following points:

- The date when it was set, so I know which currency conversion value to take.
- Mean income of people in this country/city/district/etc, so I can judge about how adequate the price is.
- Mean price for a listing in this country/city/district/etc.
- Mean expenses on visiting this city (for a day, few days, a week, etc.)

2.5 Images

The customer definitely considers listing's pictures when taking a decision to book the apartment. However, this dataset does not have listing's images.

2.6 Conclusion

The dataset lacks important information such as review texts and images. Thus, it is better to find another source of data.

2.7 A richer dataset

Source: http://insideairbnb.com/get-the-data/.

This dataset contains review texts, listing's images and a much richer information about listing page. Apart from that, the source of data contains datasets for other cities, different time periods and additional tables such as dynamics of listing's price. Thus, it will be the primary source of data.

3 Data Preparation

- Drop uninformative columns.
  Columns deleted: 'scrape_id', 'last_scraped', 'source',
  'host_url', 'picture_url', 'host_thumbnail_url',
  'host_picture_url', 'first_review', 'last_review',
  'license', 'calendar_last_scraped', 'id',
  'listing_url', 'host_id', 'host_neighbourhood', 'host_name'

- Drop rows where 'review_score_*' have nan values.

- Aggregate 'review_score_*' columns and binarize the value with some threshold.
  Columns produced: 'target'.

- Impute missing values with either placeholders or means.
  Placeholders are either empty strings " or 'not specified'.

- Multi-hot-encoding for categorical columns. To avoid explosion of the dataset size - take only 250 most frequent categories.
  Columns modified: 'amenities', 'host_verifications', 'host_response_time', 'property_type', 'room_type', 'neighbourhood_cleansed', 'bathrooms_text'
  Columns produced: a separate column for each category with binary values.

- Preprocessing for text columns - remove stopwords and transform to lower case.
  Columns modified: 'amenities'

- Sentence embedding for text columns.
  Columns modified: 'name', 'description', 'neighborhood_overview', 'host_location', 'host_about'
  Columns produced: columns contatining text embeddings.

- Transform boolean values to binary integers.

4 Modeling

I chose Catboost as it is currently the state of the art for tabular data. Also, it is quite fast and I can prototype quickly. And if this model is not able to predict the target value, then no other will be.

Then, I split the data to the train and test splits, where size of test data is 20% of original dataset size. For a better reproducibility, I set the random state to 0.

Hyperparameters are chosen by 'optuna' tool designed for hyperparameter tuning. It has optimized the target metric F1-score.

The model could achieve only 0.4 F1-score which is two times less then the success criteria. And again, if such a high-capacity model couldn't succeed, than no other model can. This is a strong evidence that with this dataset it is impossible to predict the target, the features or their combinations cannot represent the patterns that are interesting to us.

5 Evaluation

The performance of the model is far from random guessing, however, so we can try to extract some useful insights from the model. I will use SHAP [1] for that.

Examine Figure 1 see how feature values affect model output.

- 'number_of_reviews' - it seems that low number of reviews may either strongly promote class 0 (high-rated) or class 1 (low-rated). This result is not quite well understood.
- 'reviews_per_month' - low value promotes class 1. This may be more of a consequence than a cause. If a listing has a low rating, then less people will live in it and less people will leave reviews.
- '*_nights' - low values promote both classes 0 and 1, so there is no evidence of this feature's importance.
- '*_price' - low prices are an indicator of 'low-rated' listings. This can also be a consequence. If hosts have bad listings, then the only thing they can attract people with is low cost.

Though the model looks at listing's coordinates when making a decision, I cannot see any positional pattern, any grouping of 'low-rated' listings inside some area. See Figure 2.

Figure 3 plot shows that certain combinations of 'number_of_reviews' and 'host_is_superhost' features result in positive contribution to class probability. Namely, if there is a low number of reviews and host is not superhost, then the listing will probably have a low rating.

Impact of 'price' feature is best visible in conjunction with 'number_of_reviews_ltm' (number of reviews in the last 12 month).

Figure 4 shows that if listing is of low price and has a lot of reviews, then it will probably lead to 'low-rated' classification. In contrast, if a listing has high price and a lot of reviews, then it will lead to 'high-rated classification.

On average, a higher number of listed amenities leads to a 'high-rated' prediction (Figure 5).

Figure 6 shows that if there is a high review frequency on a listing with large total number of reviews, then it will be a sign of 'high-rated' listing. In contrast, if high review frequency is on a listing with small total number of reviews, then it will be a sign of 'high-rated' listing. That can probably be due to the experience of listing's host. More reviews are on listings which are on the site for a longer time, meaning that their hosts have more experience or just had time to remove all bad traits.

References

[1] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

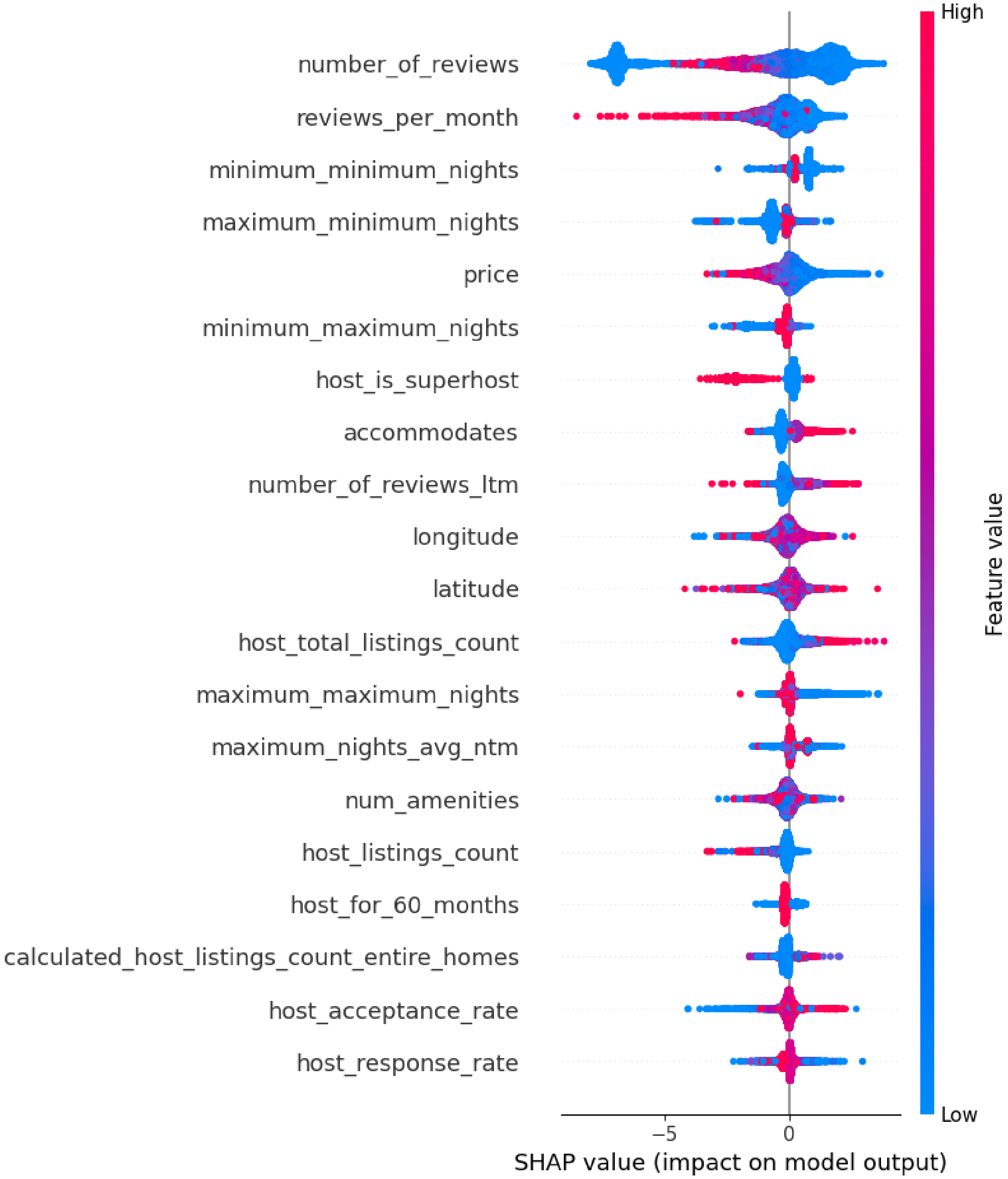

Figure 1: Importance of each feature

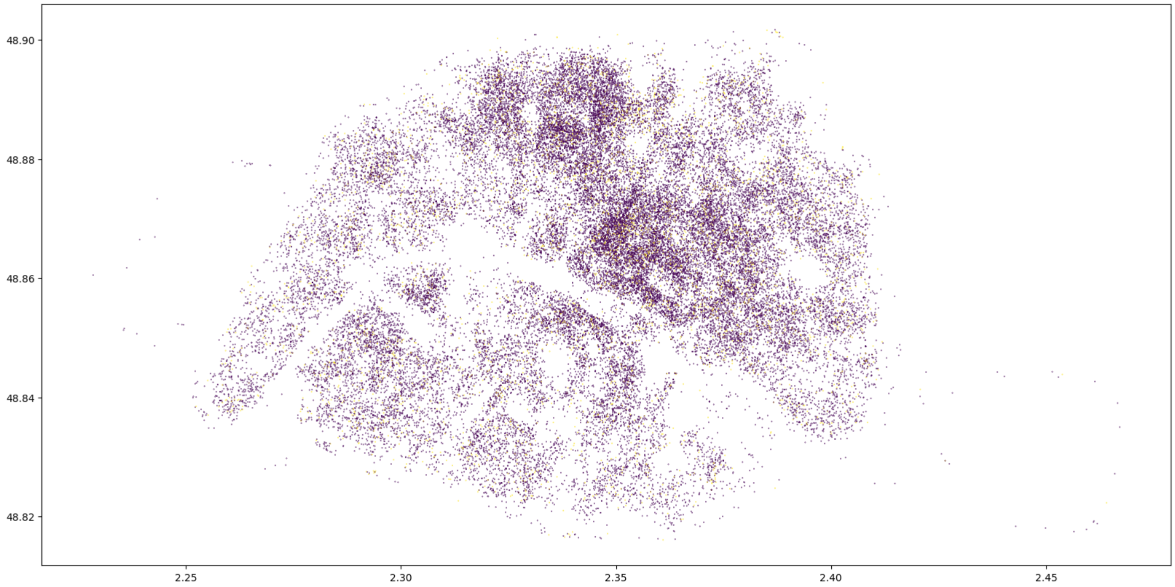

Figure 2: Target value displayed on the city map.

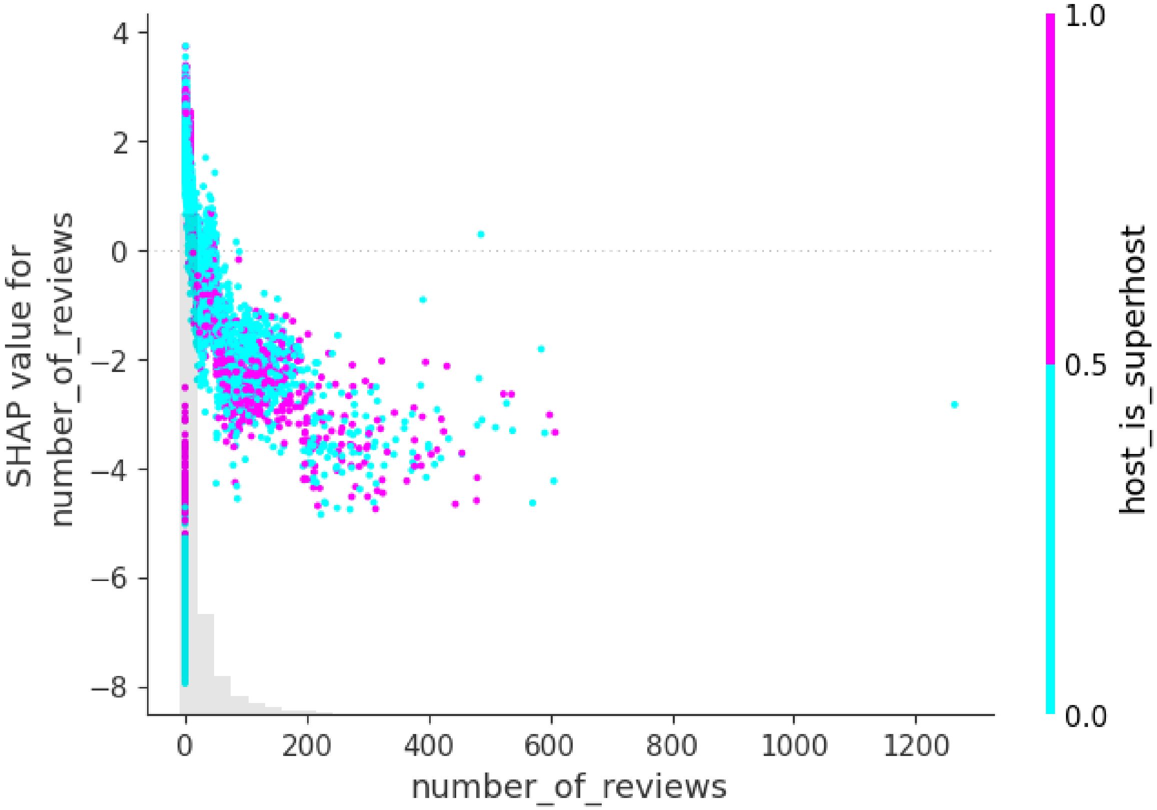

Figure 3: Impact of 'number_of_reviews' feature.

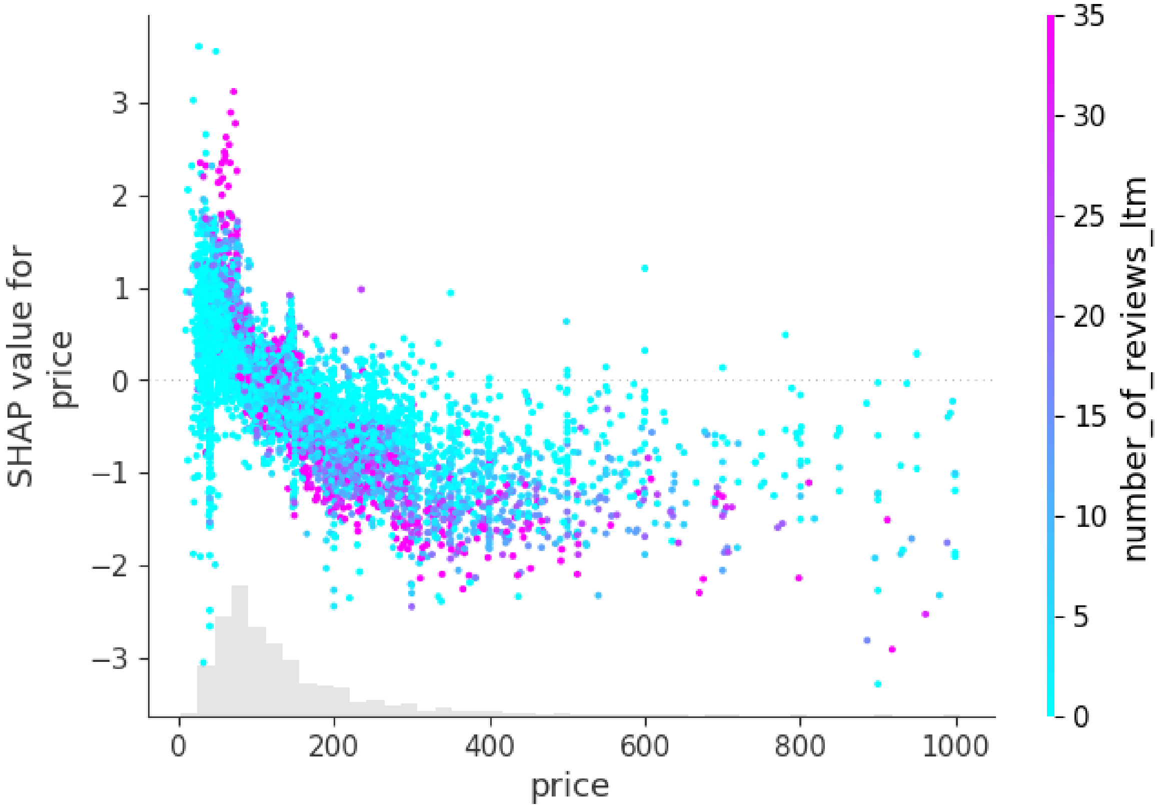

Figure 4: Impact of 'price' feature.

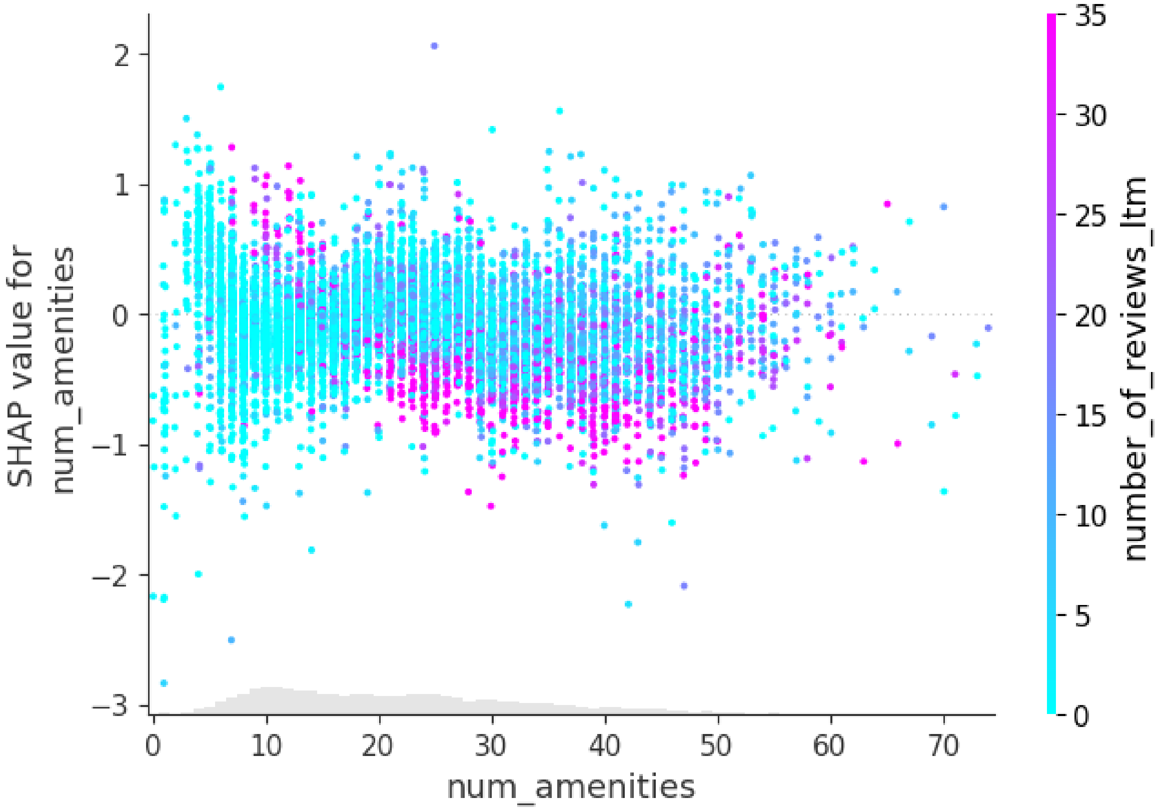

Figure 5: Impact of 'num_amenities' feature.

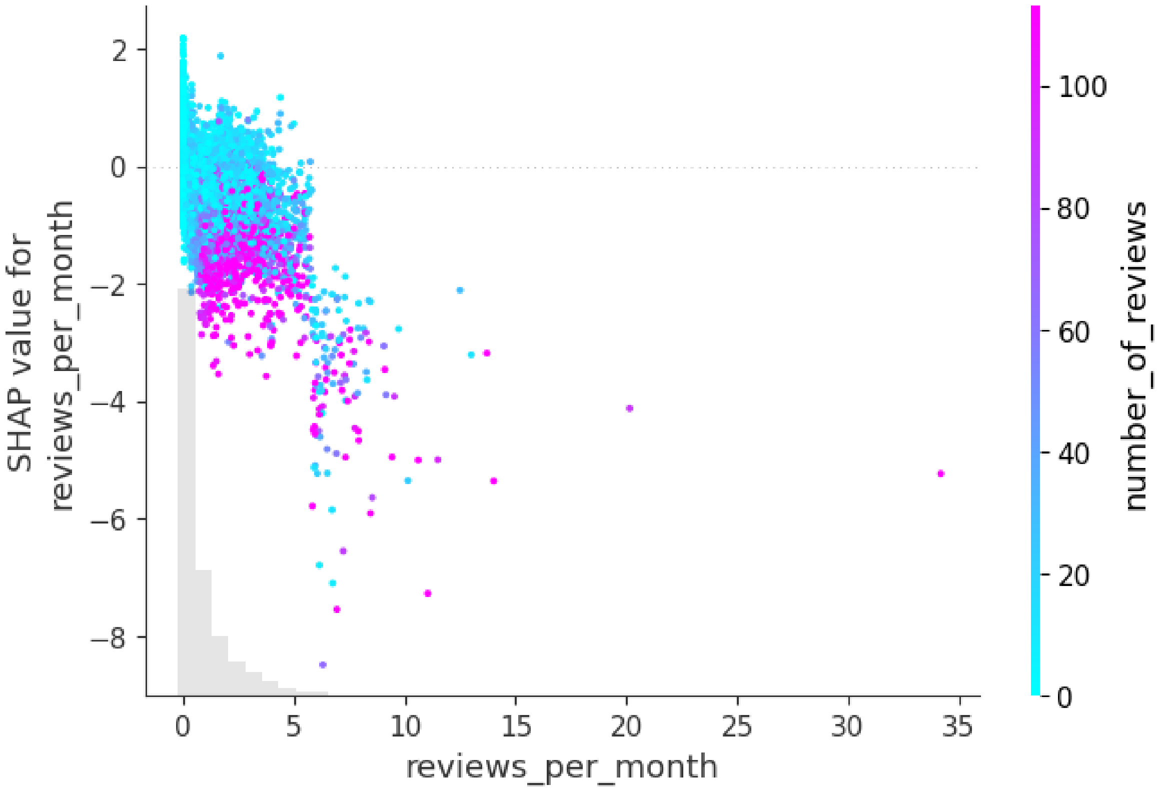

Figure 6: Impact of 'reviews_per_month' feature.