Airbnb Price Prediction Using Machine Learning

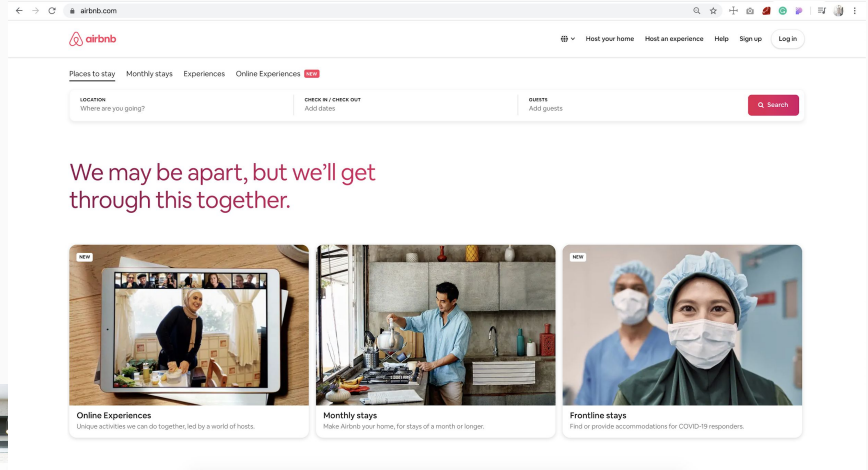
A case study using data from the City of Berlin, Germany

Overview

- Introduction
- Objectives
- Dataset / Features
- Exploratory Data Analysis
- Data modeling
  - Feature Engineering
  - Training a Linear and Ridge Regression Models
  - Training an XGBoost Regressor
  - Results
- Outlook

Introduction

Airbnb is a home-sharing platform that allows **homeowners** and **renters** (hosts) to post their properties (listings) so that guests can rent them out and stay in them.



Research Question

How to predict the price of a listing that is optimal for both hosts and renters using a wide range of data points?

Which approach will be the best?

Objectives

Analyze Airbnb listings in the city of Berlin to better understand how different features can be used to predict the prices.

Build a robust machine learning model that can uses a wide range of data points to predict optimal prices.

Experiment with linear and ridge regressions, gradient boosting framework and grid search to improve the accuracy.

Dataset

- Data comes from Inside Airbnb website[1]
- The dataset collected on March 17th, 2020
- The dataset contains 25165 detailed listings data of Airbnb listings in Berlin with rental features, such as bedrooms, location, house type, cancellation policy, geographic location, price, amenities, and number of reviews.

[1] *Source:* http://insideairbnb.com/

Dataset

```
RangeIndex: 25164 entries, 0 to 25163
Data columns (total 20 columns):
summary                         23915 non-null object
neighbourhood_group_cleansed    25164 non-null object
latitude                        25164 non-null float64
longitude                       25164 non-null float64
property_type                   25164 non-null object
room_type                       25164 non-null object
accommodates                    25164 non-null int64
bathrooms                       25146 non-null float64
bedrooms                        25131 non-null float64
beds                            24951 non-null float64
amenities                       25164 non-null object
price                           25164 non-null object
security_deposit                15468 non-null object
cleaning_fee                    17725 non-null object
minimum_nights                  25164 non-null int64
number_of_reviews               25164 non-null int64
review_scores_rating            20137 non-null float64
instant_bookable                25164 non-null object
cancellation_policy             25164 non-null object
reviews_per_month               20636 non-null float64
dtypes: float64(7), int64(3), object(10)
```

Initial Dataset Description

The price distribution after dropping outliers

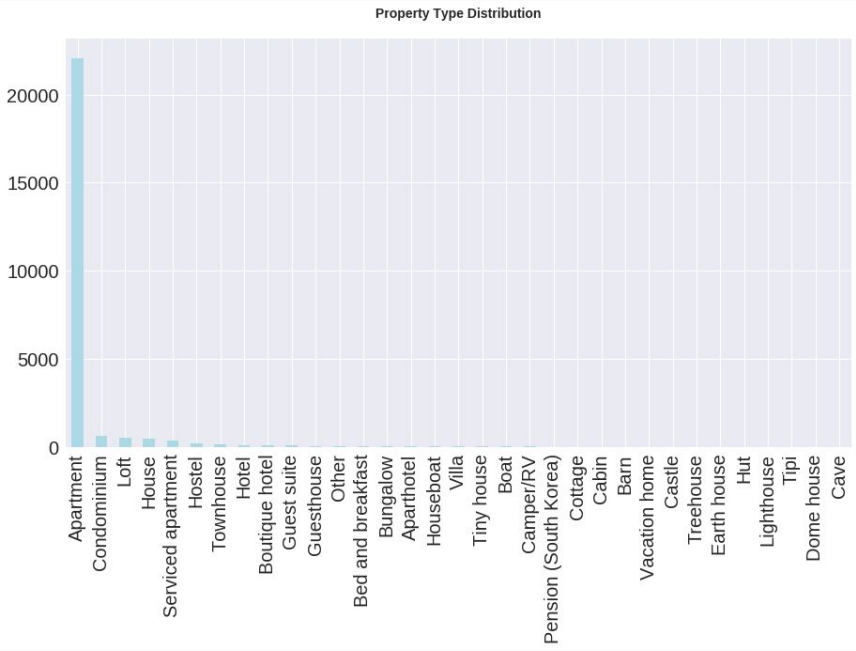
	summary	neighbourhood_group_cleansed	latitude	longitude	property_type	room_type	accommodates	bathrooms	bedrooms	beds	amenities
0	This beautiful first floor apartment is situa...	Pankow	52.53500	13.41758	Apartment	Entire home/apt	4	1.0	1.0	2.0	{Internet,Wifi,Kitchen,"Buzzer/wireless interc...
1	First of all: I prefer short-notice bookings. ...	Tempelhof - Schöneberg	52.49885	13.34906	Apartment	Private room	1	1.0	1.0	1.0	{Internet,Wifi,"Pets live on this property",Ca...
2	NaN	Friedrichshain-Kreuzberg	52.51171	13.45477	Loft	Entire home/apt	2	1.0	1.0	1.0	{TV,"Cable TV",Internet,Wifi,"Air conditioning...
3	Cozy and large room in the beautiful district ...	Pankow	52.54316	13.41509	Apartment	Private room	2	1.0	1.0	2.0	{Wifi,Heating,"Family/kid friendly",Essentials...
4	4 bedroom with very large windows and outstand...	Pankow	52.53303	13.41605	Apartment	Entire home/apt	7	2.5	4.0	7.0	{TV,"Cable TV",Internet,Wifi,Kitchen,"Paid par...

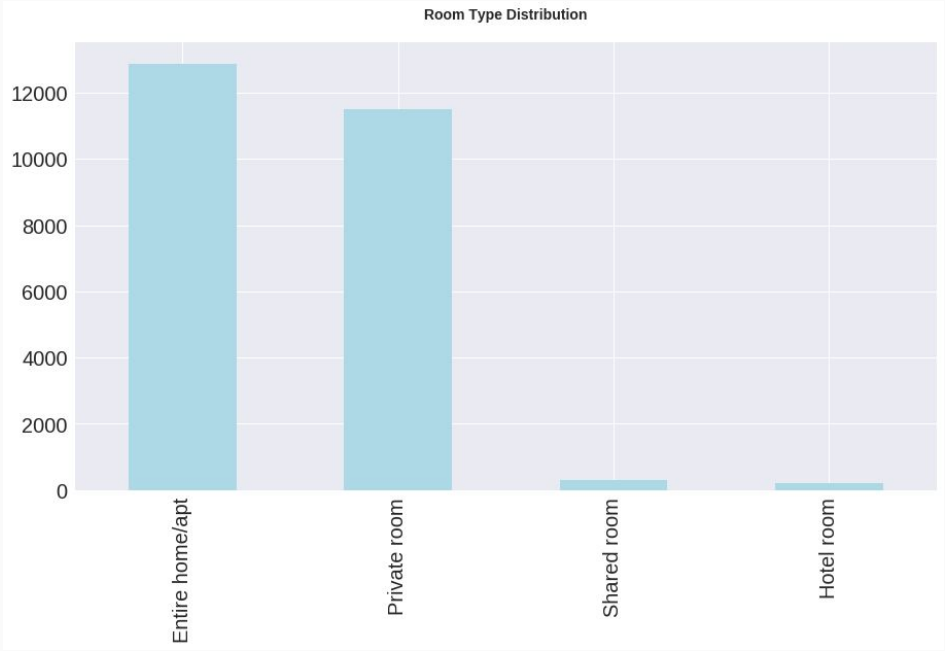
price	security_deposit	cleaning_fee	minimum_nights	number_of_reviews	review_scores_rating	instant_bookable	cancellation_policy	reviews_per_month
$90.00	$300.00	$100.00	62	145	93.0	f	strict_14_with_grace_period	1.11
$28.00	$250.00	$30.00	7	27	89.0	f	strict_14_with_grace_period	0.34
$125.00	$0.00	$39.00	3	133	99.0	f	moderate	1.08
$33.00	$0.00	$0.00	1	292	97.0	f	moderate	2.27
$180.00	$400.00	$80.00	6	8	100.0	f	strict_14_with_grace_period	0.14

The sample of the dataset after cleaning and dropping columns

Exploratory Data Analysis

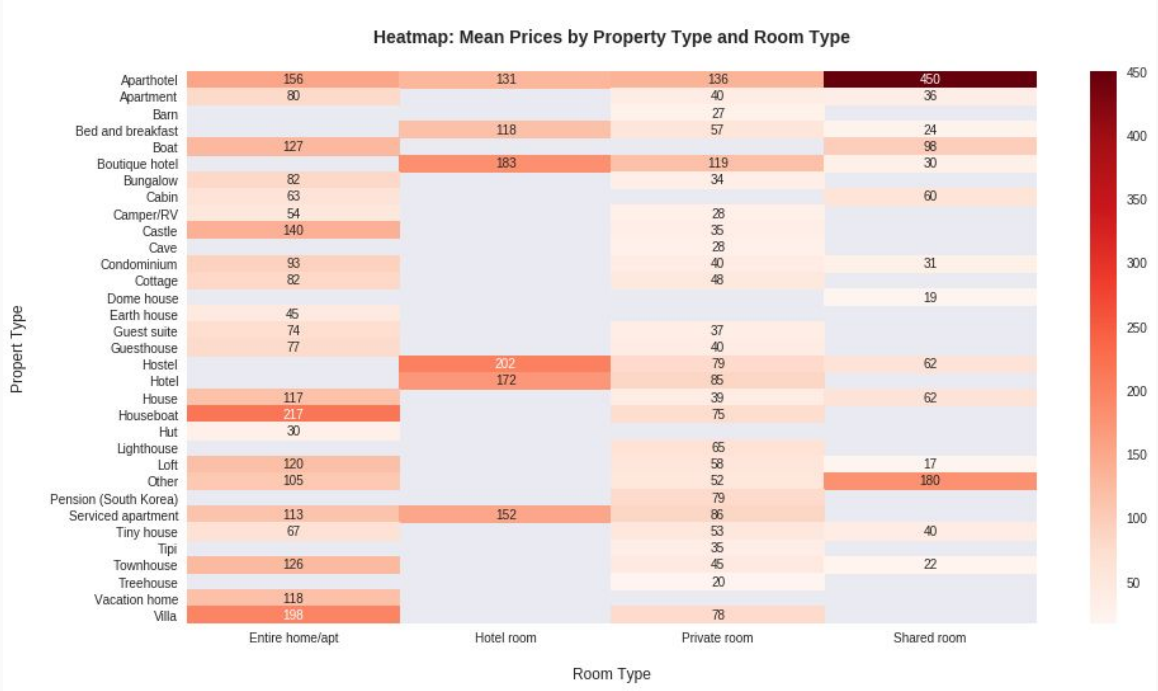
Room Type and Property Type Distribution





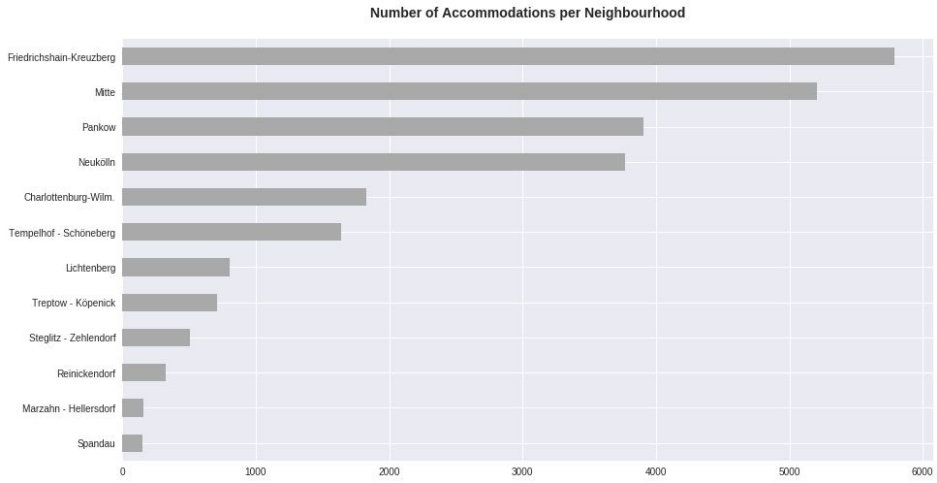
Exploratory Data Analysis

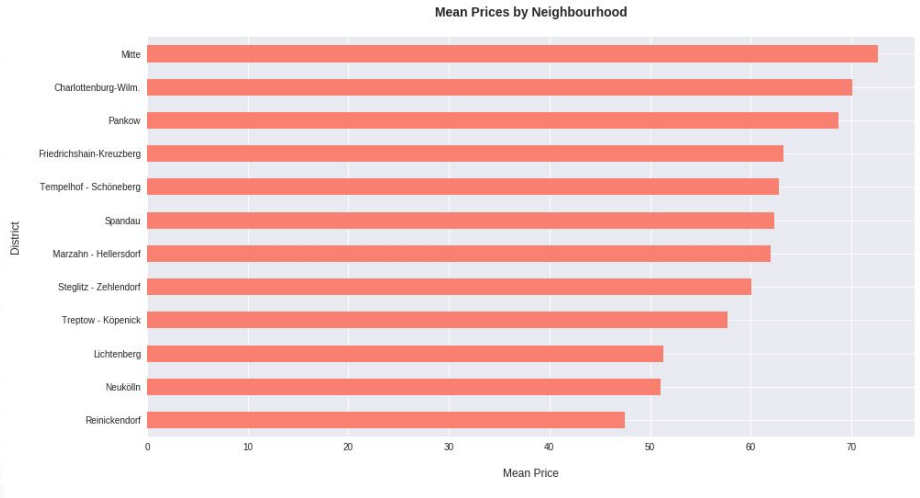
Room Type and Property Type Distribution



Exploratory Data Analysis

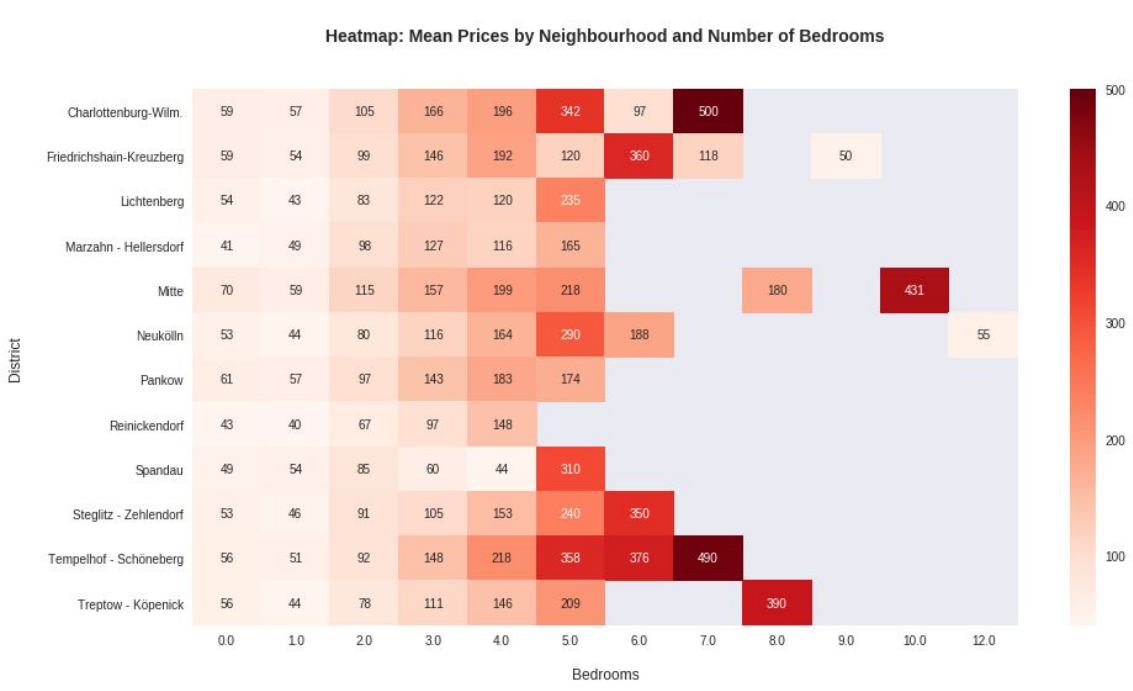
Neighborhood Analysis





Exploratory Data Analysis

Neighborhood Analysis

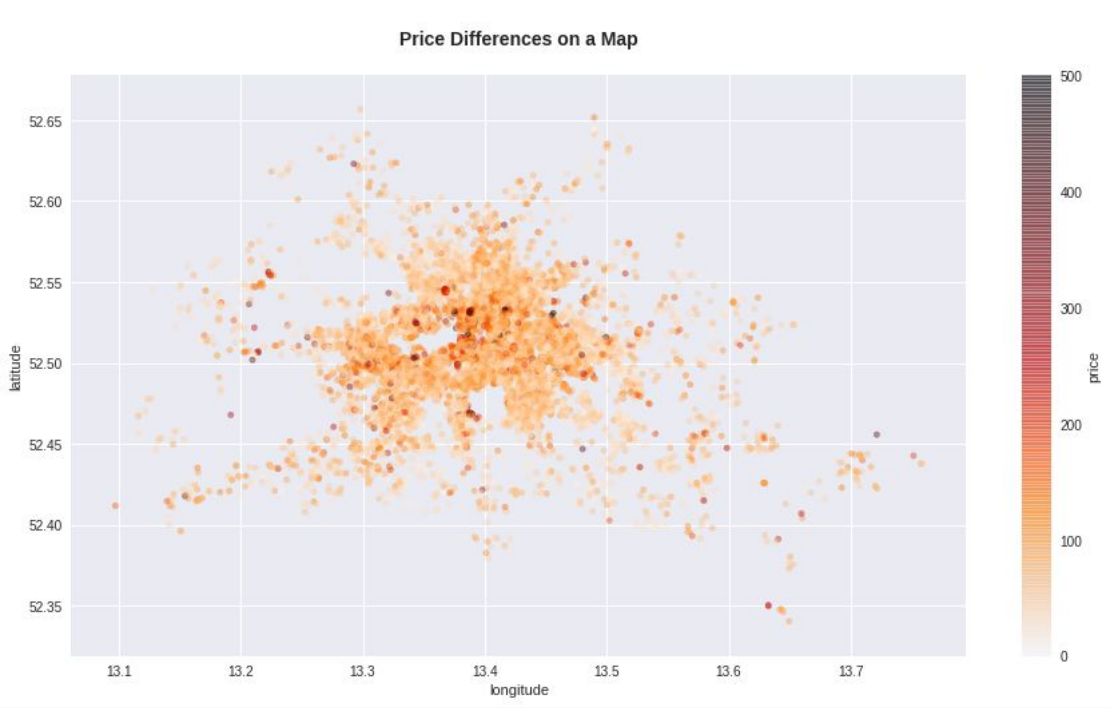


Exploratory Data Analysis

Neighborhood Cancellation Policy Analysis

Exploratory Data Analysis

Price Differences by Location

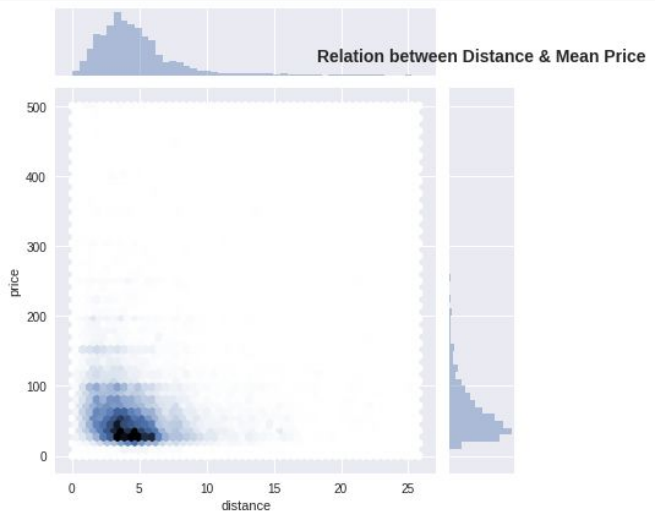


Exploratory Data Analysis

Number of Reviews by Price

Data Modeling

Feature Engineering



Data Modeling

Feature Engineering

- The room type (e.g., entire home)
- The number of people the property accommodates
- The number of bathrooms, bedrooms and beds
- The presence or absence of a wide range of amenities
- The price of the listing
- The cleaning fee and security deposit
- The distance to the center of Berlin
- The number of minimum nights per stay
- Total number of reviews
- Total review rating
- Whether the property is instantly bookable
- The type of cancellation policy
- The number of reviews per month

	price	room_type	number_of_reviews	instant_bookable	review_scores_rating	beds	bedrooms	bathrooms	accommodates	amenities	cancellation_policy	reviews_per_month	cleaning_fee	security_deposit	minimum_nights	distance
0	90.0	Entire home/apt	145	f	93.0	2.0	1.0	1.0	4	14	strict_14_with_grace_period	1.11	100.0	300.0	62	1.422238
1	28.0	Private room	27	f	89.0	1.0	1.0	1.0	1	30	strict_14_with_grace_period	0.34	30.0	250.0	7	5.115770
2	125.0	Entire home/apt	133	f	99.0	1.0	1.0	1.0	2	16	moderate	1.08	39.0	0.0	3	3.003533
3	33.0	Private room	292	f	97.0	2.0	1.0	1.0	2	16	moderate	2.27	0.0	0.0	1	2.310025
4	180.0	Entire home/apt	8	f	100.0	7.0	4.0	2.5	7	27	strict_14_with_grace_period	0.14	80.0	400.0	6	1.190529

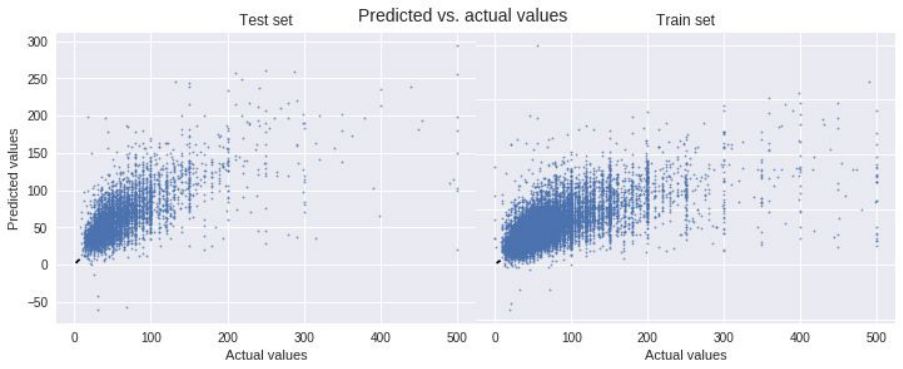
Data Modeling

Training a Linear and Ridge Regression Models

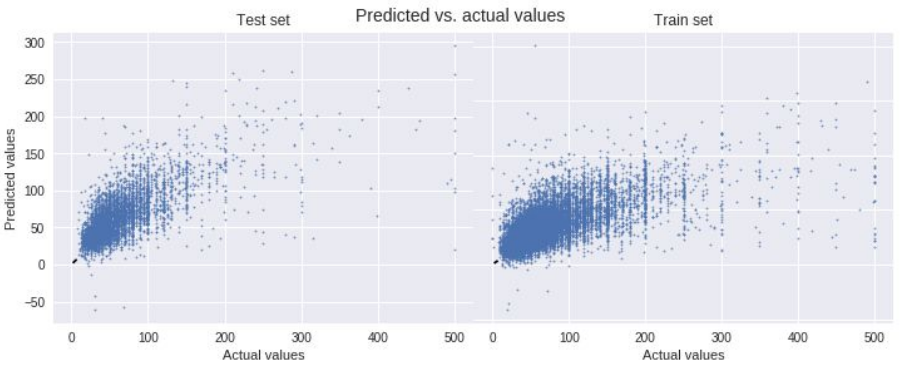
- Linear Regression: R^2 value of **0.45996.**
- Ridge Regression: R^2 value of **0.45995.**

Splitting and Scaling the Data

- Split the training and testing set with a test size of 0.2.



Linear Regression

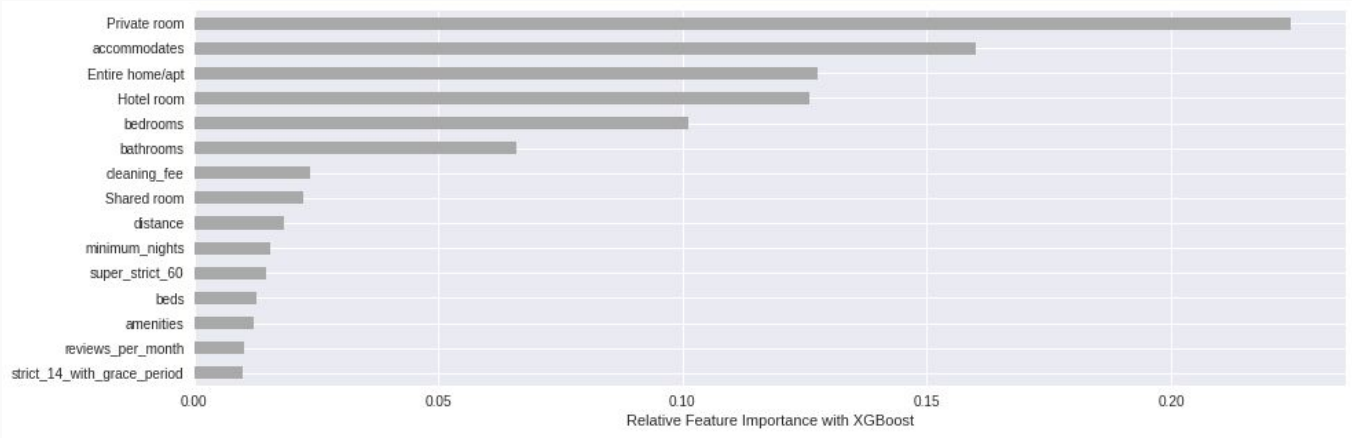


Ridge Regression

Data Modeling

Training an XGBoost Regressor

- XGBoost Regressor: R^2 value of **0.54543** and an RMSE value of **33.48601.**



Outlook

- Models need more tuning
- Add other cities to improve generalizability
- Advanced Feature Engineering
- Use Neural Networks

Any Questions?