2022

SAN FRANCISCO RENTAL PRICES

Tadashi Mori

TABLE OF CONTENTS

INTRODUCTION

Inn the Neighborhood is an online platform that allows people to rent out their properties for short stays. At the moment, only 2% of people who come to the site interested in renting out their homes start to use it.

The product manager would like to increase this. They want to develop an application to help people estimate how much they could earn renting out their living space. They hope that this would make people more likely to sign up.

Customer Request

Develop a way to predict how much someone could earn from renting their property that could power the application?

Initial KPI request

Avoid estimating prices that are more than 25 dollars off of the actual price, as this may discourage people.

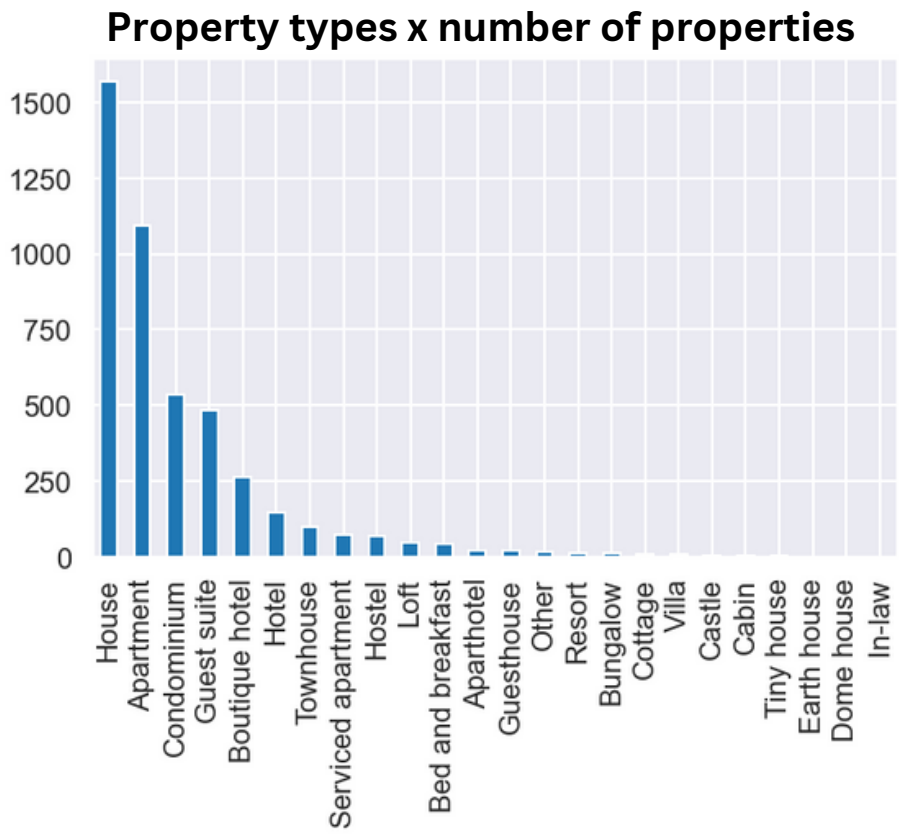
Dataset

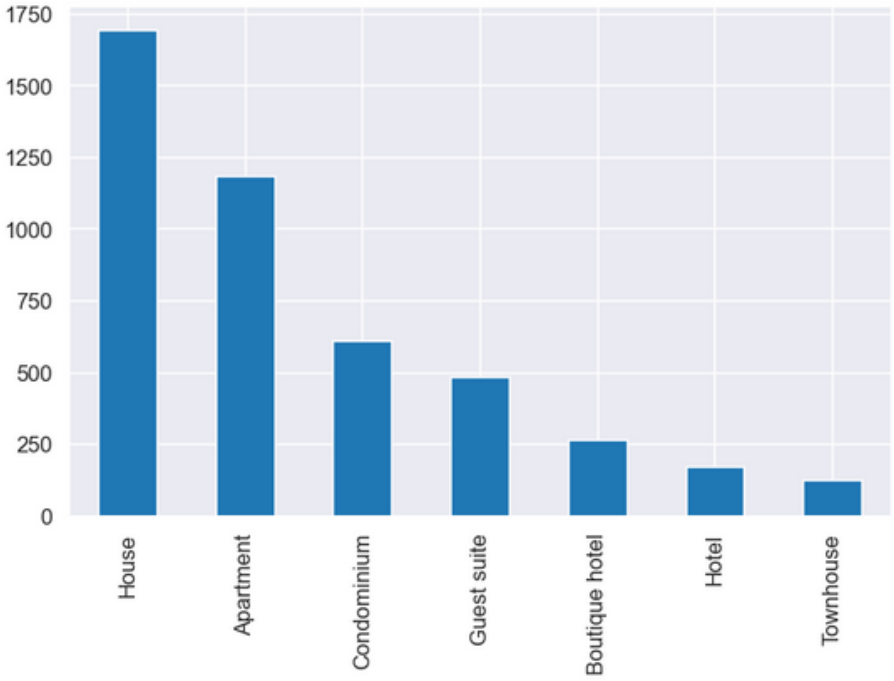
The dataset with last month's rental prices was provided by the customer to use in our analysis.

PROPERTY CATEGORIES

There are SDGs and 24 property types. While they are all important, some categories has a small sample in the dataset, increasing the variance of the price predictions, consequently making the predictions worse.

To improve the predictions, it's necessary to have a bigger sample or to cluster the categories into less categories for the model prediction. I clustered property types with less than 100 samples by finding the best fit (best MSE) to the linear regression of bedrooms x price





OUTCOMES

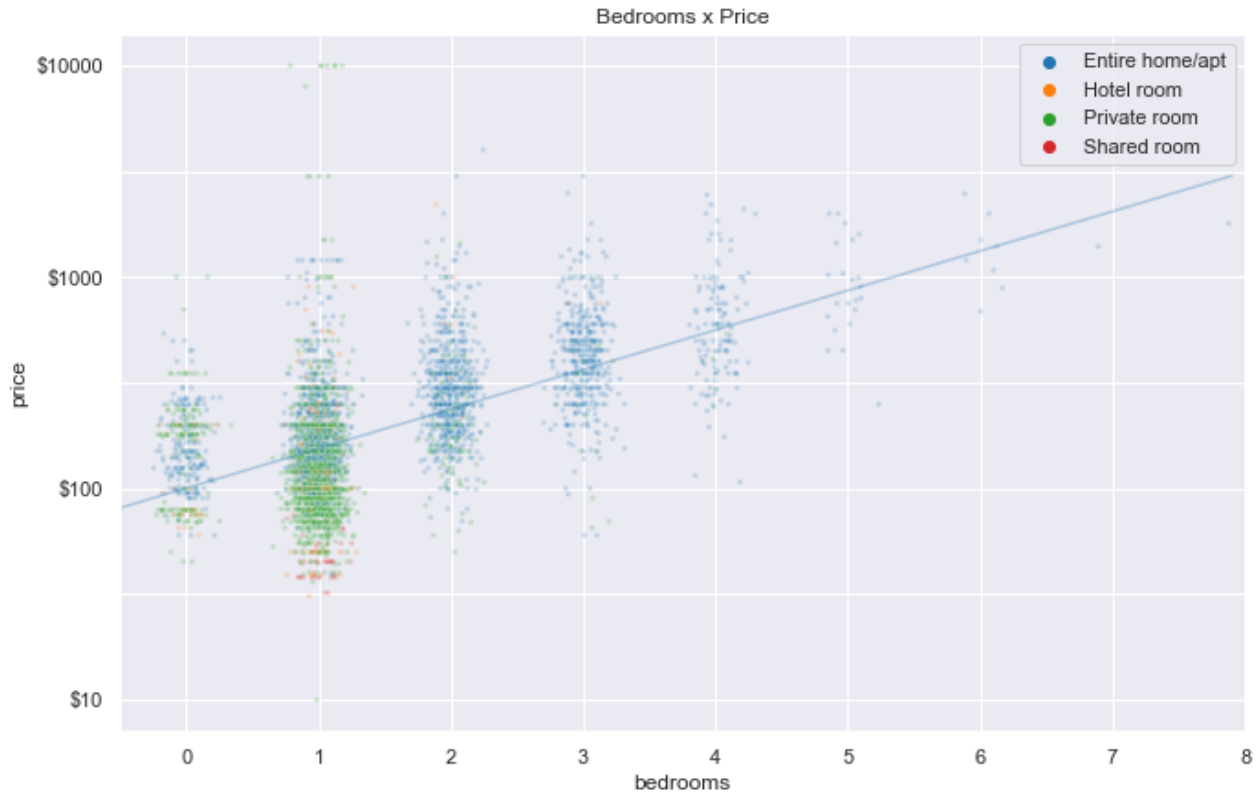
TWO MODELS - LINEAR REGRESSION AND GRADIENT BOOST REGRESSOR

I'm using linear regression as a baseline model because we can see strong to moderate relationship between some features and the target variable.
The comparison model I am choosing is the Gradient Boosting Regressor model because it is easy to interpret and it generally provides better accuracy than the linear regression.

TWO METRICS - R^2 AND RMSE

IFor the evaluation, I am choosing R^2 and RMSE (Root Mean Squared Error). R^2 measures how well the model fits dependent variables (i.e. features).

RMSE measures how much your predicted results deviates from the actual number.



OUTCOMES

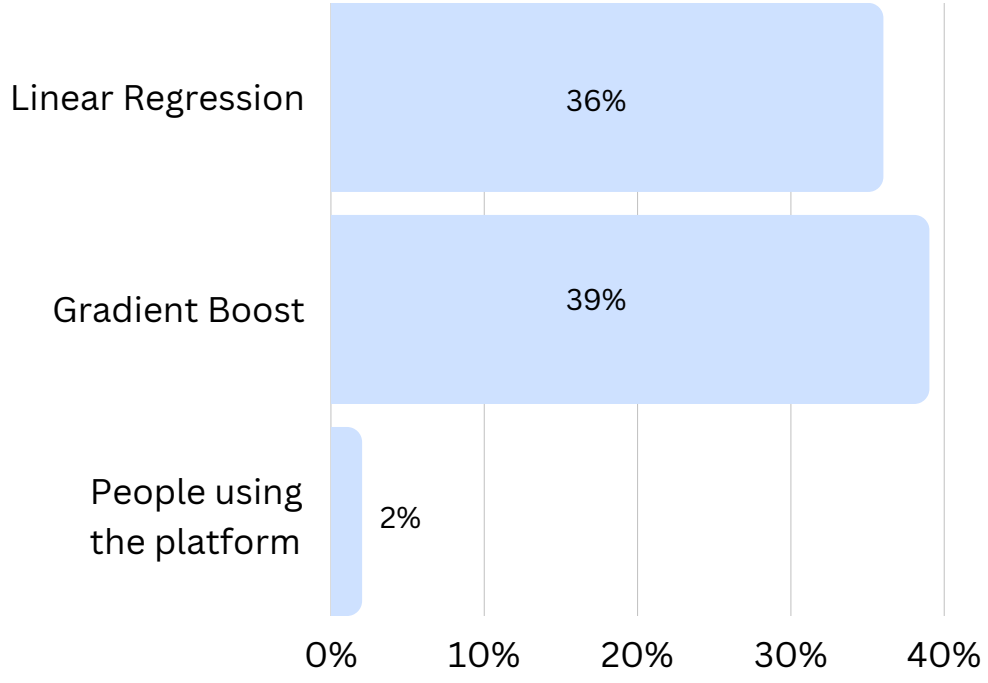
WINNER - GRADIENT BOOSTING REGRESSOR

Model	R^2	RMSE
Linear Regression	0.55	0.65
Gradient Boost Regressor	0.61	0.61

The R squared of the Linear Regression, and the Gradient Boosting Regressor model is 0.55 and 0.61, meaning the Gradient Boosting Regressor model fits the features better.

The RMSE of the Linear Regression, and the Decision Tree Regression model is 0.65 and 0.61, meaning the Decision Tree Regression model has less error in predicting values.

EVALUATING THE MODELS BY OUR KPI



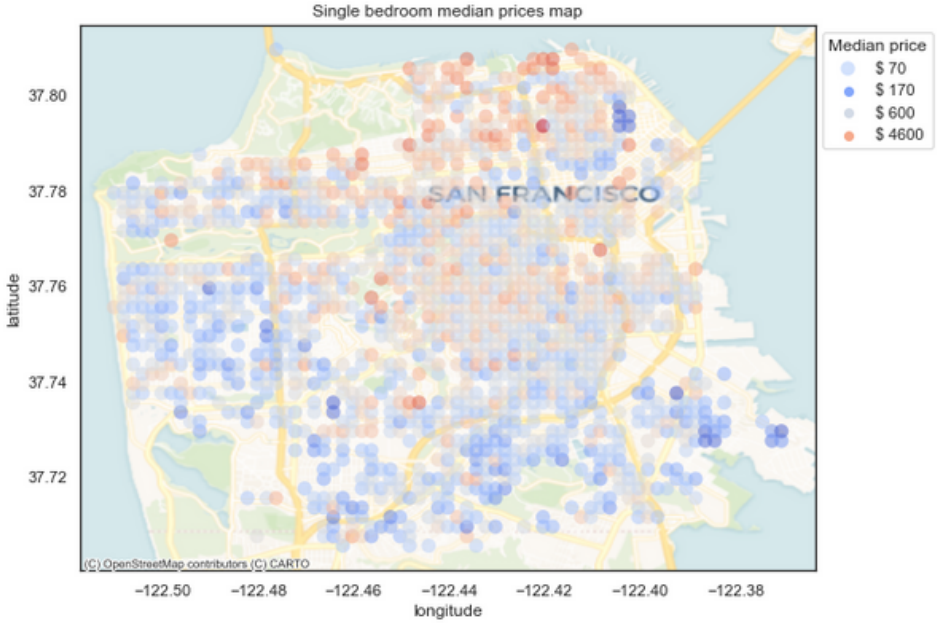
Our KPI

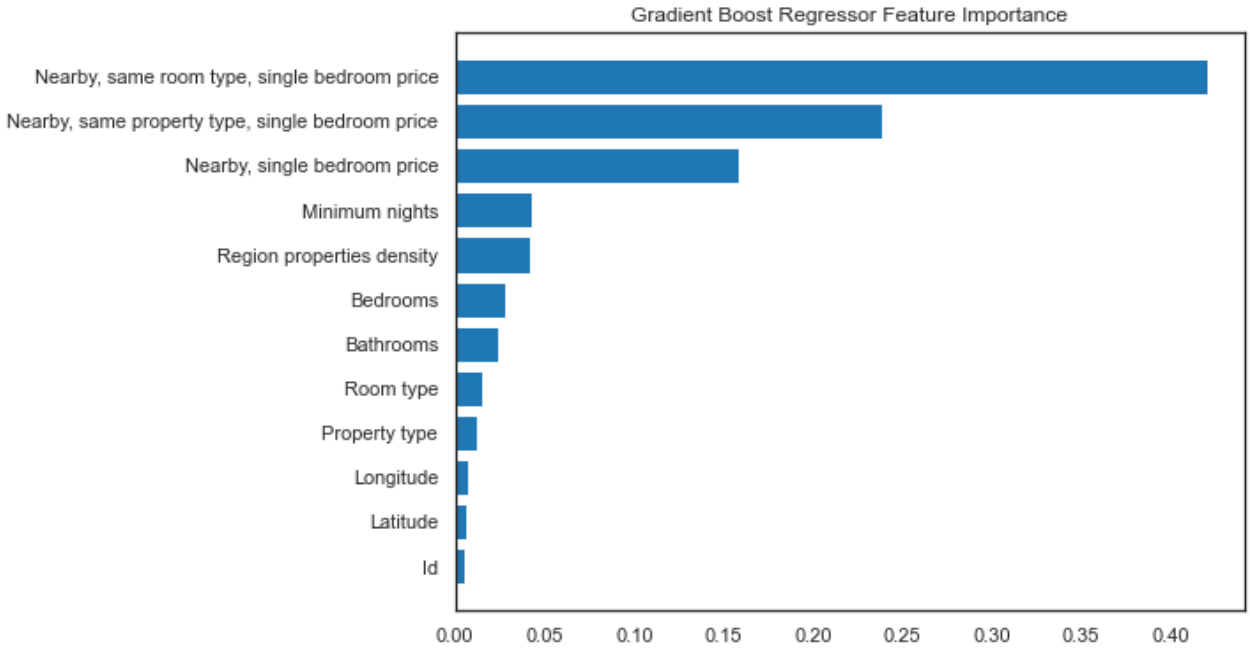
Percentage of prediction which the predicted price is not more than $25 off of the actual price. The higher the KPI the better.

OUTCOMES

FEATURE IMPORTANCE

One engineered feature for our rental prices prediction model is the average or median price for a single room in the same region. This indicator gives us a map of how "valuable" is each region, helping the model predict the rental prices and it was one most important features.





OUTCOMES

EVALUATE BY BUSINESS CRITERIA

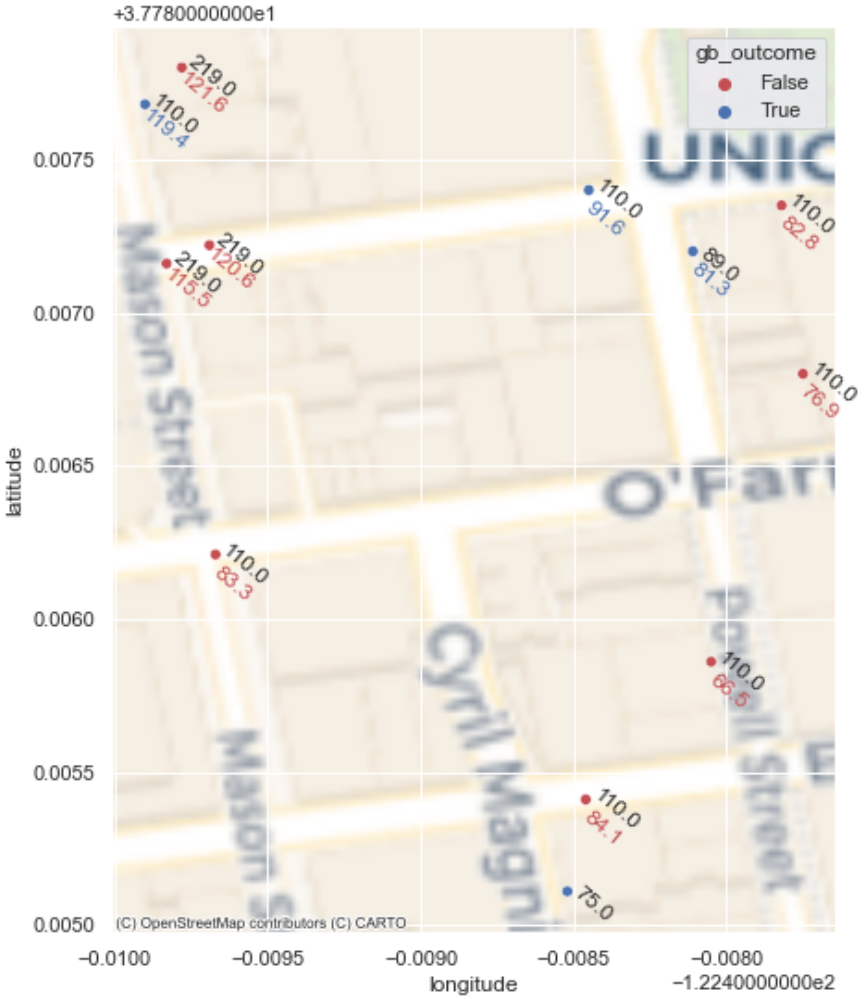
Only 39% of the predictions are in the error range suggested by the business.

39%

Predictions within +-25$

Properties in the same block and with the same characteristics can have very discrepant prices, making it impossible to predict the price within +- 25$ with the given information

Properties with same characteristics: Location and price



Gradient Boost regressor predictions in blue (within $25) and red (more than $25 error)

RECOMMENDATIONS

To help the user predict the price, we can deploy this GradientBoostingRegressor model into production. By implementing this model, about 39% of the users will be encouraged to rent the property. The model should be deployed as a web services in the home page of the website.

The calculator should lead to the registering page, auto-completing the fields filled before to make it as friction-less as possible. The deployed model should not make the forms more complicated than the previous forms.

To better evaluate whether this model can really encourage more people to rent their property, I would also recommend A/B testing about using this model to compare two groups of users.

As we have only 39% of accuracy by our KPI, make staged releases to a small portion of the users (Canary deployment) to better evaluate the outcome of the recommending system in the users behavior before releasing to everyone.

The percentage of users in the test set will depend on the website traffic. Ideally we should have at least 2000 users in the test group to be able to have a more precise confidence interval.

- **Deploy on home page** - Guarantee that the user will see the calculator
- **Friction-less forms** - Make the forms completion friction-less.
- **A/B testing** - Canary deployment for at least 2000 users.

5%

Even with only 5% conversion rate among the encouraged users, the amount of people joining the website will increase.

NEXT STEPS

- Collect more data, e.g. property area, property age, property internet access, garage, prices seasonality, etc to improve the price prediction system.
- Increase sampling size to have a considerable amount of information on each property type and room type. Ideally we should have a few hundreds samples of each possible property and room types combination.
- The calculator could show a price range according to the region variance instead of a single prediction.
- Show similar properties to compare intangible values such as brand, design and how well maintained is the place.

Contact

Tadashi Mori

tadashiorigami.github.io/

linkedin.com/in/tadashi-mori/

medium.com/@tadashi-mori

tadashimori.br@gmail.com