Predicting Airbnb Listing Prices

Sherry Duong
May 15th, 2020

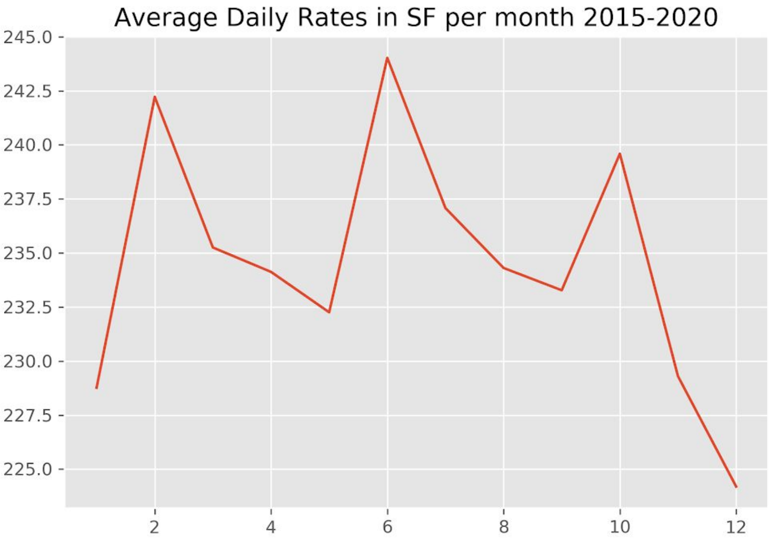
The Data

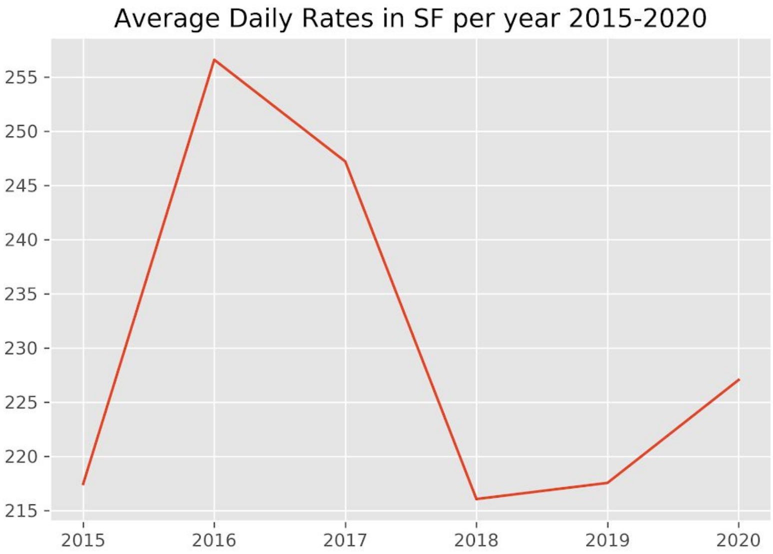
- 52 Datasets comprised of information for every Airbnb listing available in San Francisco per month, for 52 months between Sep 2015 & April 2020
- Total Data:
  - 395,202 entries
  - 92 columns
- Null Values: 2842171
- Missing Monthly Data:
  - 2015 only had Sep, Nov, and Dec
  - 2016 missing Jan and Mar data
  - 2018 missing June data

Data Cleaning

- Removed all columns with more than 70% of data being null
- Converted the "Last Scraped" date to date format, and engineered additional date features to indicate year, month-year, month, dayofweek, and day
- Converted columns related to currency (price, extra_people, security_deposit and cleaning_fee] from string to float, removed '$'
- Converted Categorical to dummies

Listing Price Over Time

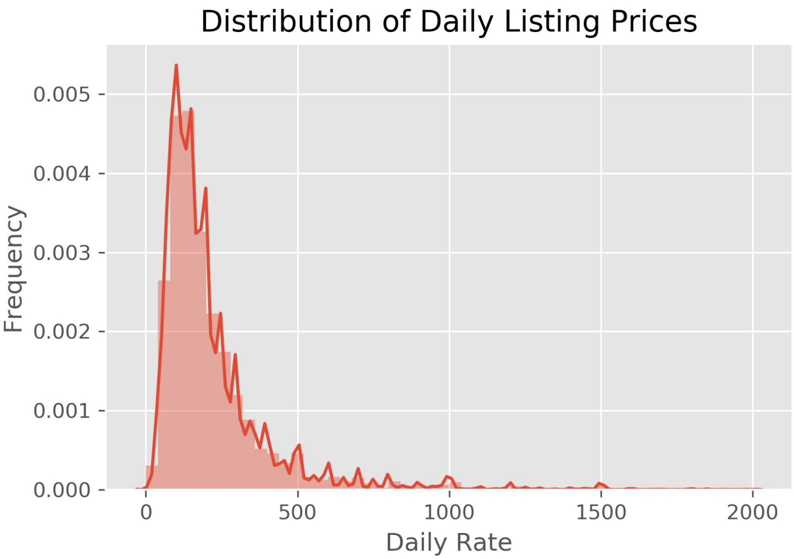


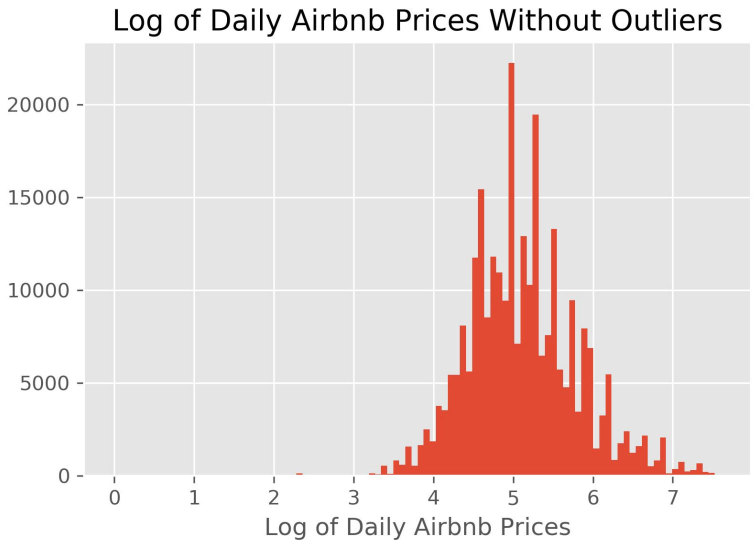


Clear seasonality between months.

Spike in 2016 - 2017 due to rent increase in major US metro areas including San Francisco.

Distribution of Listing Prices





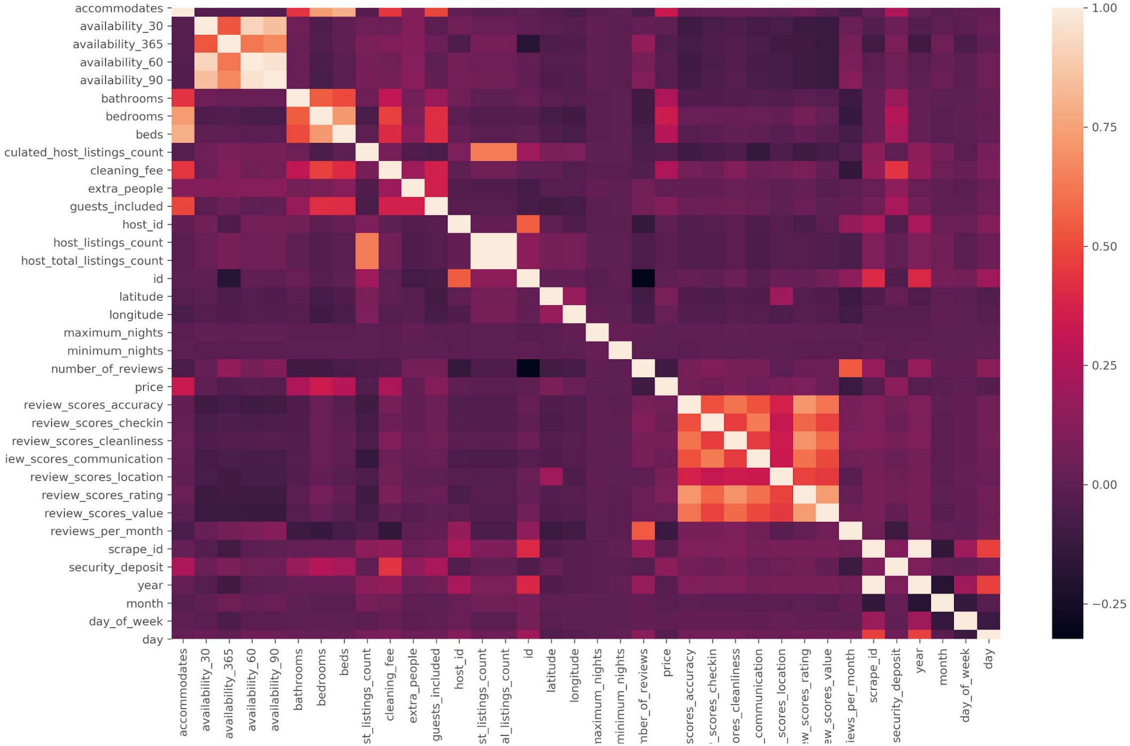
Daily listing prices is very right skewed, with clear outliers (price = 0 or price > 2000) present.

The distribution looks a lot more normal, and this should help with the model.

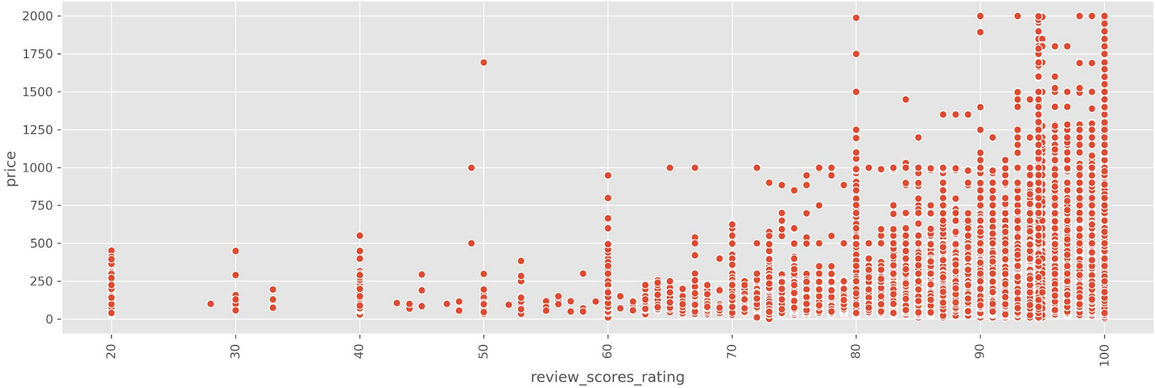
EDA: Numeric Features

Most important features (based on Correlation plot with correlation over 1%):

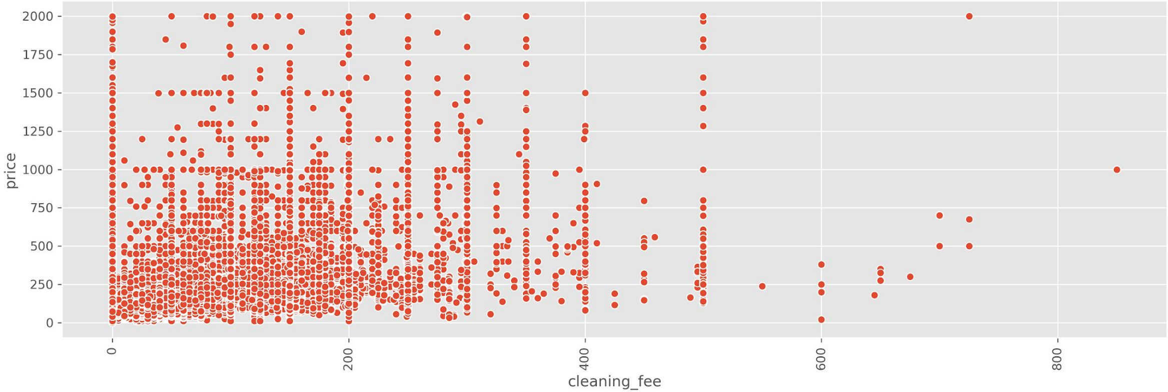
Accommodates, bathrooms, bedrooms, beds, cleaning_fee, security_deposit, review_scores_rating, review_scores_cleanliness, review_scores_location, review_scores_accuracy, review_scores_communication, review_scores_checkin, review_scores_value, extra_people, month, year, number_of_reviews



Reviews & Fees

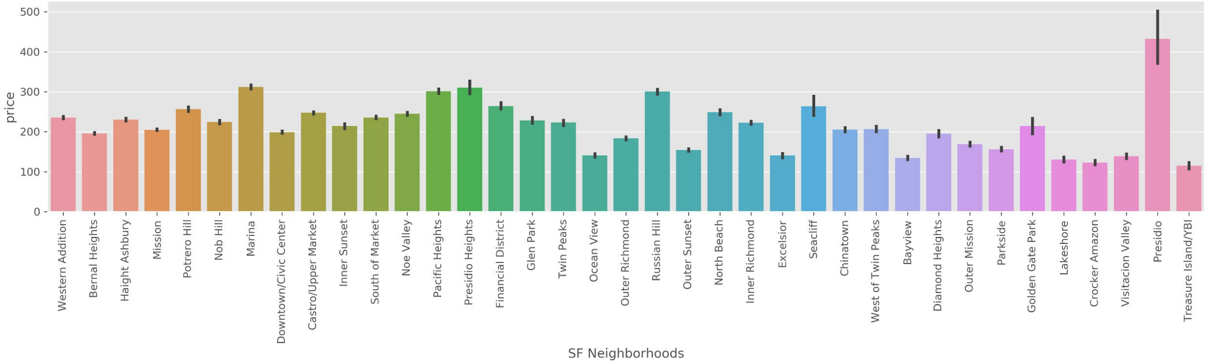


Reviews have a clear positive correlation on price.

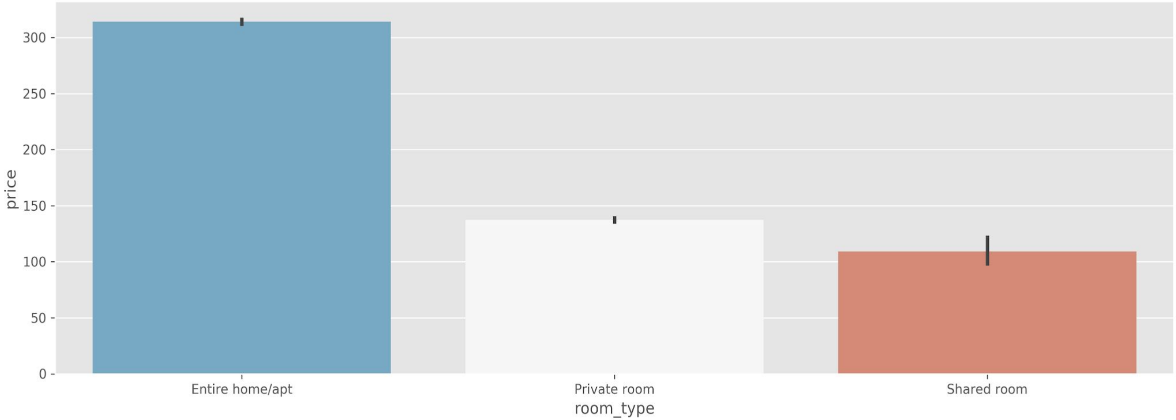


Fees have a less clear correlation, may be useful to convert to “Y” or “N”.

Categorical Features

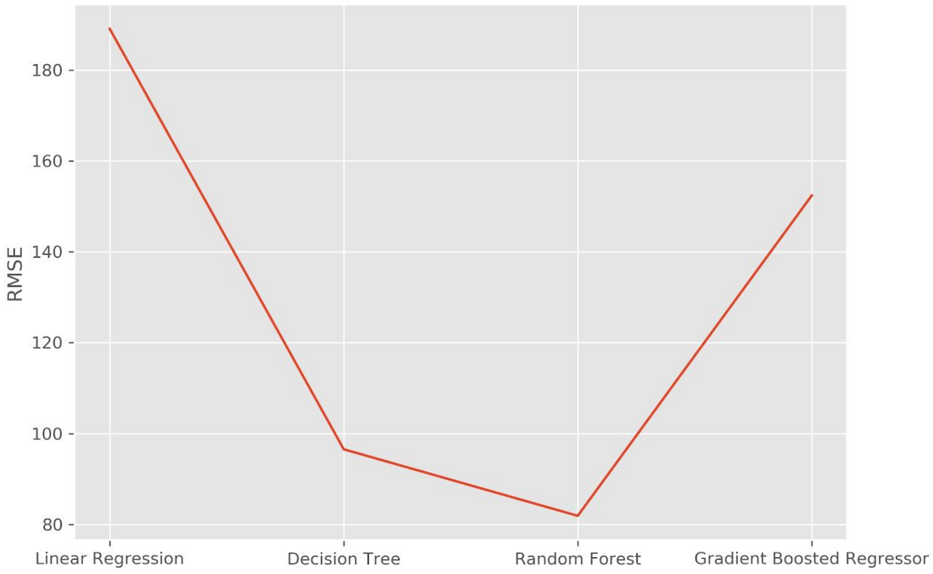


Neighborhoods & Property types seem to have an impact on price, though not as clear.



Very obvious price difference between room types.

Baseline Models



Baseline Using the Average (Not pictured):
-Took the average of daily listing prices as the predicted.
-RMSE: 207.59, which is not good considering the average listing price is $234.

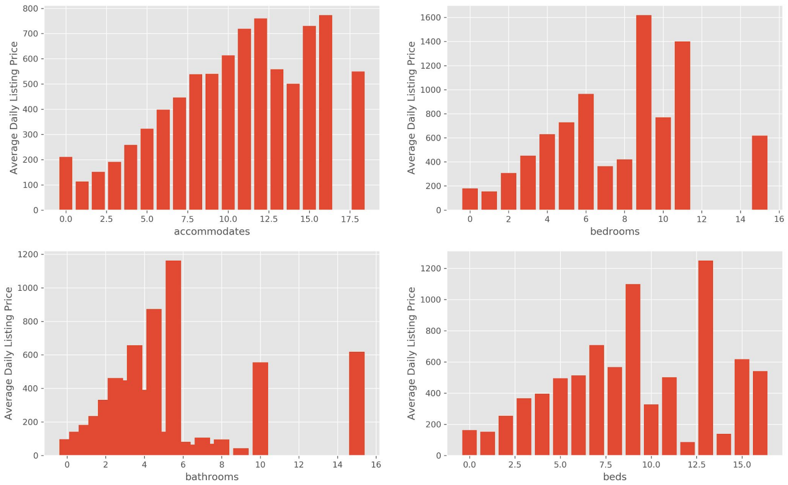
Baselines using features selected from EDA:
-Linear Regression: RMSE: $189.06
-Decision Tree: RMSE: $96.53
-Random Forest: RMSE: $81.89
-Gradient Boosted: RMSE: $152.38
Will proceed with Random Forest Estimator for future iterations.

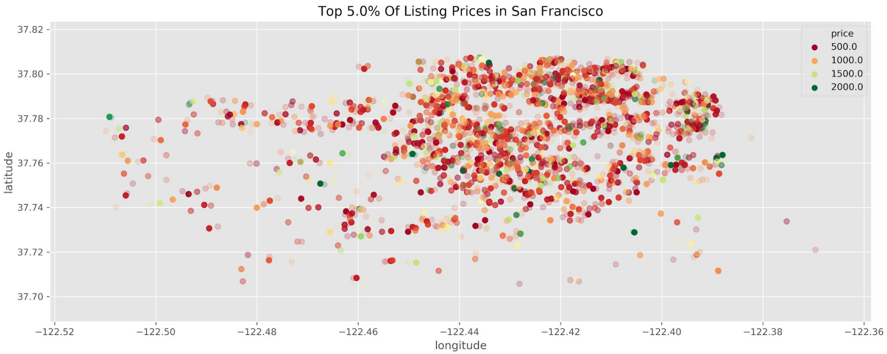
Features selected for Model Selection based on EDA
'accommodates','bathrooms','bed_type','bedrooms', 'beds','cleaning_fee','extra_people', 'host_response_time', 'neighbourhood_cleansed', 'property_type', 'review_scores_cleanliness', 'review_scores_rating', 'room_type', 'security_deposit', 'year', 'month', 'day_of_week'

Feature Engineering: What did not work

1. Converting columns related to Fees from numerical to dummy variables of “Was the fee present?”
2. Categorical feature to capture whether or not the listing “accommodations” was over the “threshold” for diminishing returns.
3. Neighborhoods:
   -Creating regions of neighborhoods
   -Creating a category of whether or not the neighborhood was in the “city center”
4. Natural Language Process:
   -Creating word vectors based on the listing “summary” from hosts

All of the above changes made resulted in a higher RMSE.



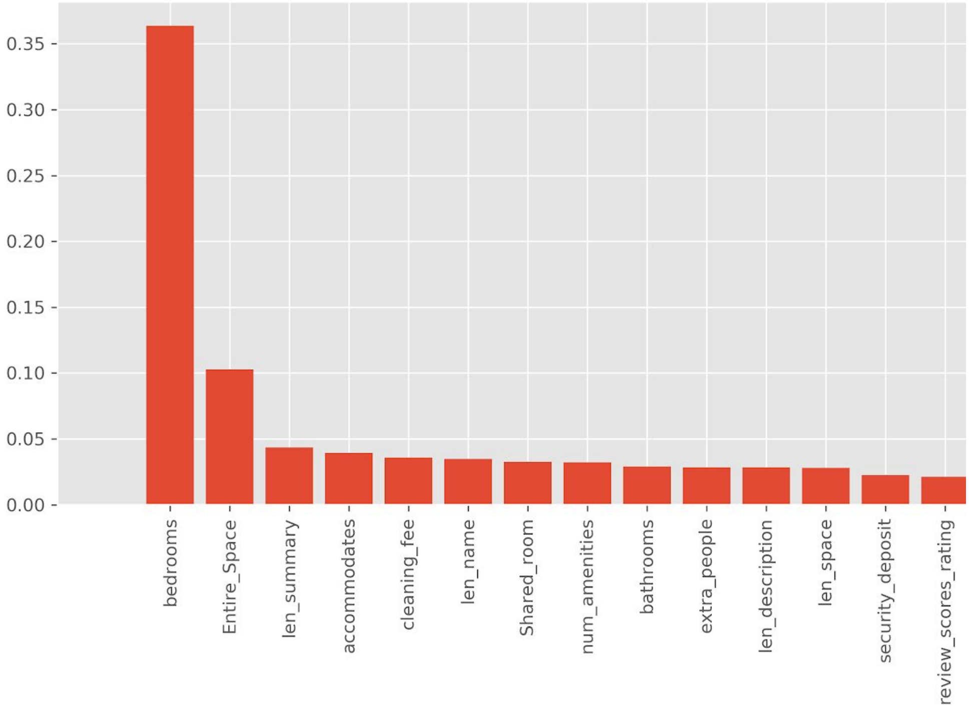


Feature Engineering: What Worked

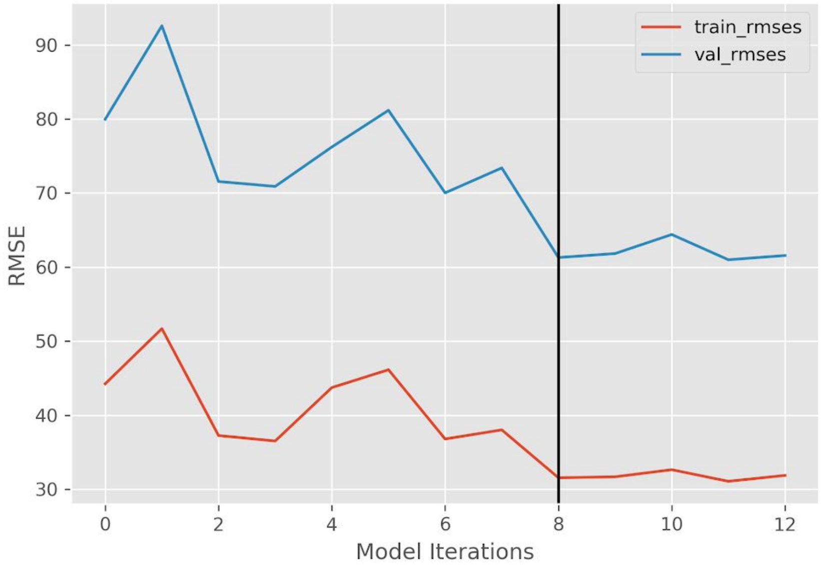
1. Creating feature to capture of number of "Amenities" listed by the host
   ->Reduced RMSE by $10
2. Creating feature to capture the length of the listing summary, name, and description (in characters) provided by the host.
   ->Reduced RMSE by $10
3. Creating features to capture whether the listing was the Entire House/Apartment, Shared room, House, or Apartment.
   ->No change to RMSE, but reduced model complexity.

All of the above changes appeared in the updated top features in the feature importance graphs.

Final performance after finalizing features:
-RMSE $61.79



Final Model

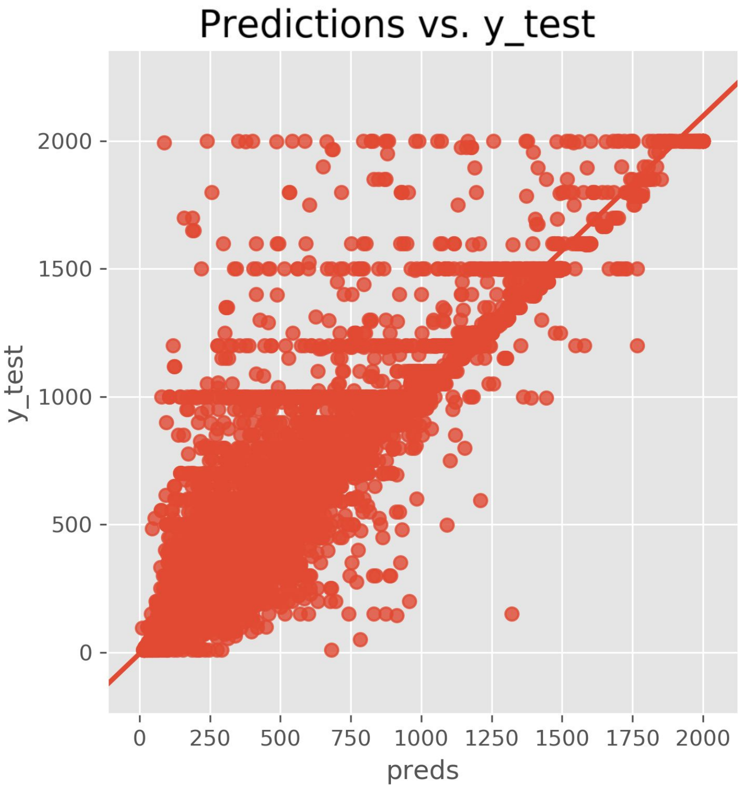


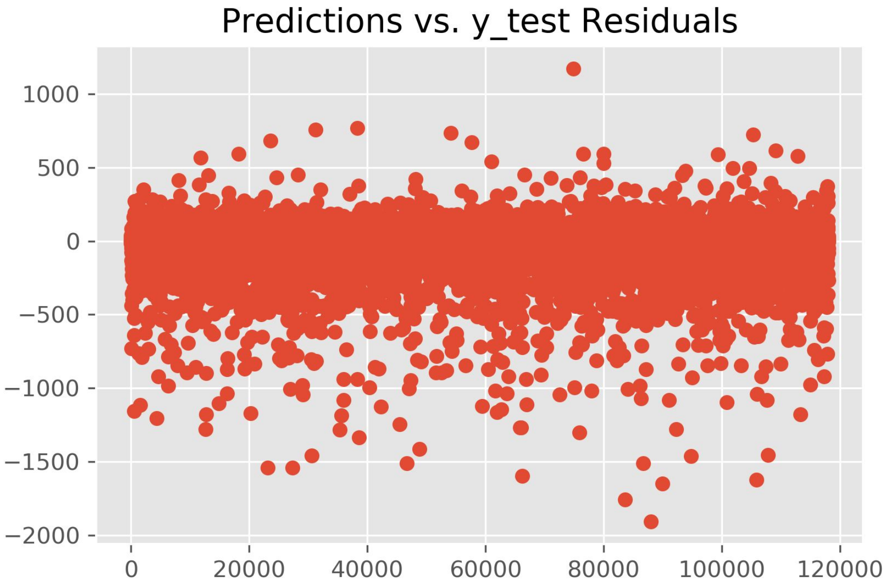
Final Features: 73 (20 Main Features + Dummies)
-accommodates','bathrooms', 'bed_type','bedrooms', 'beds','cleaning_fee', 'extra_people', 'num_amenities', 'neighbourhood_cleansed', 'review_scores_cleanliness', 'len_space', 'len_summary', 'len_description', 'len_name', 'review_scores_rating', 'security_deposit', 'month', 'House', 'Apartment', 'Entire_Space','Shared_room', ‘availability_60’

-Tried to tune model with GridSearchCV, ran out of memory
-RandomForest with 400 estimators and max_features = 20

Final performance after tuning:
-RMSE $59.32

Performance on Test Data





Test Data Performed Better Than Cross-Val Data
-RMSE on Unseen Test Data: $56.68
-From this, it appears the model is not great at predicting prices above $1000 per listing, which was roughly 1% of the dataset. For the future, I would consider removing these.

Conclusion & Caveats

- The final RMSE was still about 25% of the mean average listing price
- Based on this model, the most important features for determining price of listing in San Francisco are: number of bedrooms, whether or not the listing is private space or shared, the length of the summary & name of listing, the number of people accommodated, and if there are extra fees associated.
- Data Source: InsideAirbnb.com.
    - Listing Prices are set by the host and may not reflect the final price paid by the tenant. Thus, it would make sense that hosts that spend more time writing summaries and noting amenities would aim for a higher listing price. It cannot be confirmed whether or not the attempts were successful and tenants actually paid this much.

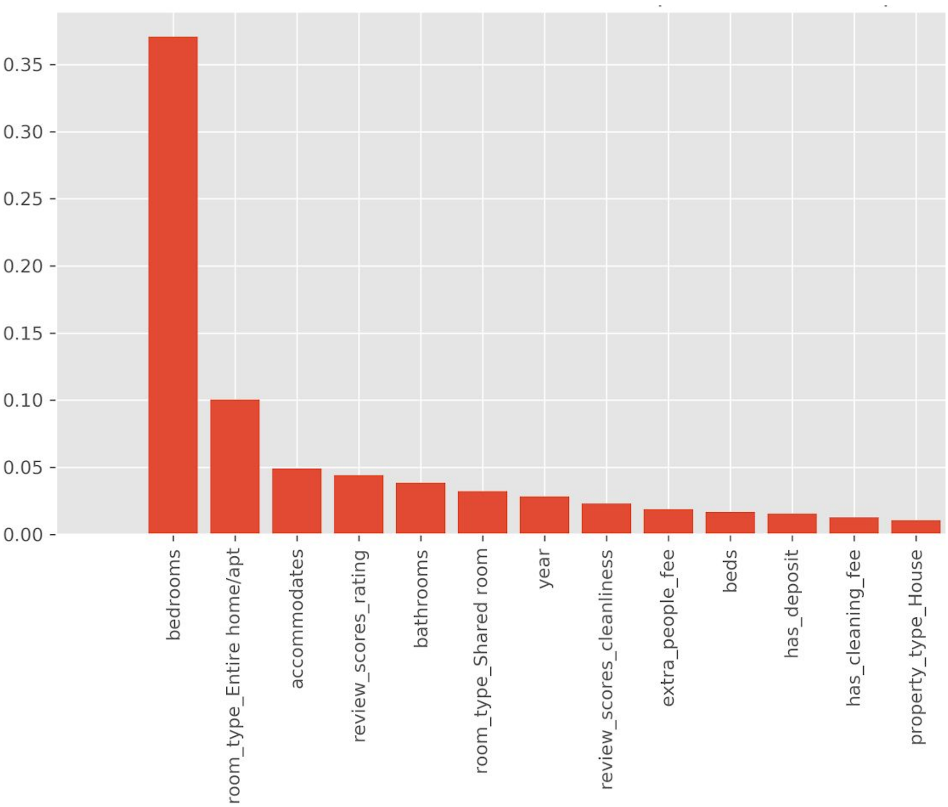
Goals for Future Analysis

- Compare the listing prices to actual value of the property according to Zillow.
- Look into specific amenities and if they are more important
- Inferential Model with Linear Regression
- Dive Deeper into Neural Network and deep learning model
- Improve the RMSE even further

Thank You!

Questions?

Extra: Feature Importance from Base Model Random Forest



Most Important Features

1. Number of Bedrooms
2. Privacy: Whether or not the listing is an entire home/apartment, or shared room
3. Fees: Fees also seem to be important, with all 3 fee categories in the top 20
4. Reviews: Review ratings overall & cleanliness
5. Year, but not Month
6. Was the property a house?
7. Neighborhoods did not make the top 10, but neighborhoods downtown were more important to model than others