Capstone Project I

Problem Statement

What's the likelihood of an Airbnb listing reservation to be cancelled?

Along with the price and ratings shown in their listings, an Airbnb host's likelihood of cancelling should also be readily available for prospective guests to see.As part of the "share economy", Airbnb hosts become the key determinants of the guest's experience with the company. This experience can determine whether or not the guest will continue to use the app in the future.

This can become an incentive for the host to cancel on guests less (right now, they only get a $50/$100 penalty when they cancel).

Data Wrangling Report

In this project, I focus on the Airbnb market in London, the second largest Airbnb city outside the US.

Most of the datasets used in this project were all found on insideairbnb.com. This website "is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world."

The following are the csv files found on the Inside Airbnb website:

- listings.csv
    - information and metrics for listings in London
- reviews.csv
    - reviews left by the guests on the hosts' listings
- neighbourhoods.csv
    - Lists all neighborhoods and neighborhood groups in London

I also want to see the effect that the Airbnb listing's neighborhood's crime rate has on our study so I also make use of the London crime report posted in the 'London Datastore' website. This csv contains all the crime rates in all London boroughs. The data dates back from 1999.

`listings.csv`
columns:

id	room_type
name	price
host_id	minimum_nights
host_name	number_of_reviews
neighbourhood_group	last_review
neighbourhood	reviews_per_month
latitude	calculated_host_listings_count
longitude	availability_365

The listings data contains 80,767 rows and 16 columns. Each row represents a unique listing in London.

This is one of our main datasets as it gives us a lot of information about the listings. When we take the minimum of `number_of_reviews`, we get 0. This is relevant in this study because we get the information on the cancellations through the automated postings listed on the reviews section. We cannot simply set the number of cancellations in a listing as 0 because of 0 reviews; this could also mean that the listing is inactive. We determine inactivity by checking those airbnbs' `availability_365` and see if any of them have 365 days available. Once the inactive listings are determined, they are dropped from the listings DataFrame.

After dropping the 2,693 inactive listings (3.33% of the listings), we are left with 78,074 listings.

reviews.csv
columns:

listing_id
id
date
reviewer_id
reviewer_name
comments

As previously mentioned, we track cancellations made by the listings by looking at the `comments`. When a host cancels on a guest, an automated review is posted on their profile, indicating that they canceled on the guest. This review cannot be deleted. This reviews data contains 1,249,466 rows and 6 columns--where each row represents a comment written on a specific listing's page. When we apply .info() to the DataFrame, we see that there are 407 null values (i.e. empty reviews) on the "comments" column. Since this is only .03% of the reviews, we can drop these rows using .dropna().

neighbourhoods.csv

The neighborhood data is simple and just gives us the 33 neighborhoods in London. There are no null values or outliers.

crimerates.csv

Code
Borough
Year
Offences
Rate
Number_of_offences

In this dataset, we are only interested in:

- Overall crime rate of each neighborhood (categorized as "All recorded offences" under the `Offences` column), so we drop the other offences
- Neighborhoods included in the neighbourhood DataFrame. Places like "Inner London","England and Wales", "Met Police Area", "Outer London", and "Heathrow" are included in the `Borough` column despite not being an official borough (i.e. neighborhood) in London and thus not included in our neighbourhood DataFrame. We simply ignore this and drop it from our data.
- The crime rates from 2011 to present since the earliest data we have on our Airbnb data is from 2011. We do this by dropping all the crime rates from the years prior. We use: `all_crimes[(all_crimes['Year'])>=201112]`

Creating the Main DataFrame

Using all these datasets, we create a main DataFrame where each row represents a unique Airbnb listing. Besides merging the DataFrames, new columns are created to provide better understanding of our data.

In the `comments` tab of the reviews data, there were two versions of the automated postings: "The host canceled my reservation n days before arrival" and "The reservation was canceled n days before arrival. This is an automated posting. " Upon further review, we see that the automated posting template changed sometime in 2012. We create a new DataFrame only containing listings with automated postings in their comments, and apply `.value_counts()` on the `comments` tab. This will give us information on how many times reservations for each listing were cancelled before. We add this in our main DataFrame as a new tab, `num_cancellations.`

We also add a `rate` tab to the DataFrame. The rate represents the overall crime rate of the neighborhood that the listing belongs in.

After all the data wrangling, we come up with a main DataFrame with the following columns:

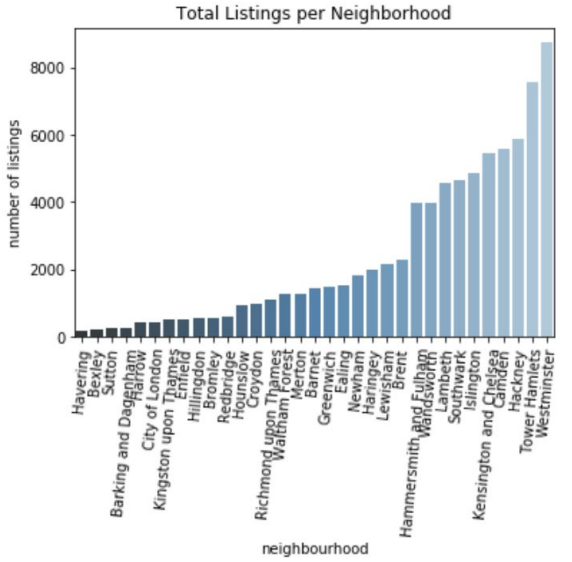
neighbourhood	days booked
room_type	num_cancellations
price	Year
minimum_nights	Rate
number_of_reviews	Number_of_offences
calculated_host_listings_count	

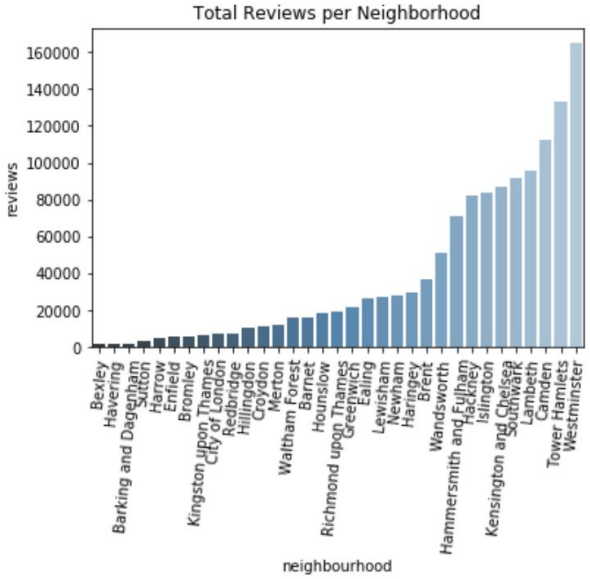
Exploratory Data Analysis

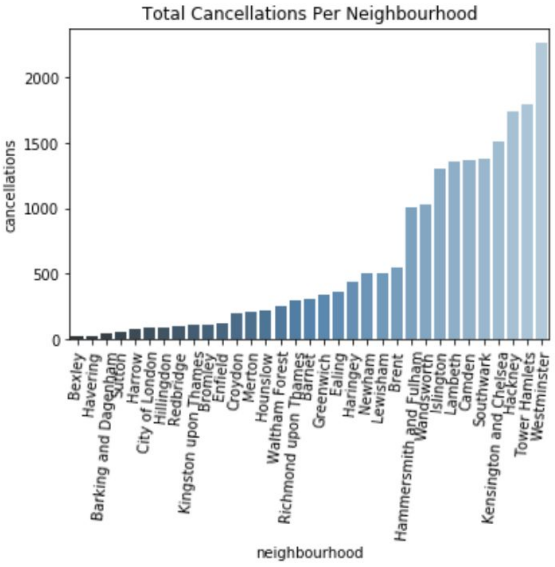
Next step is to ask questions and look into possible trends between the features of the Airbnb listings.

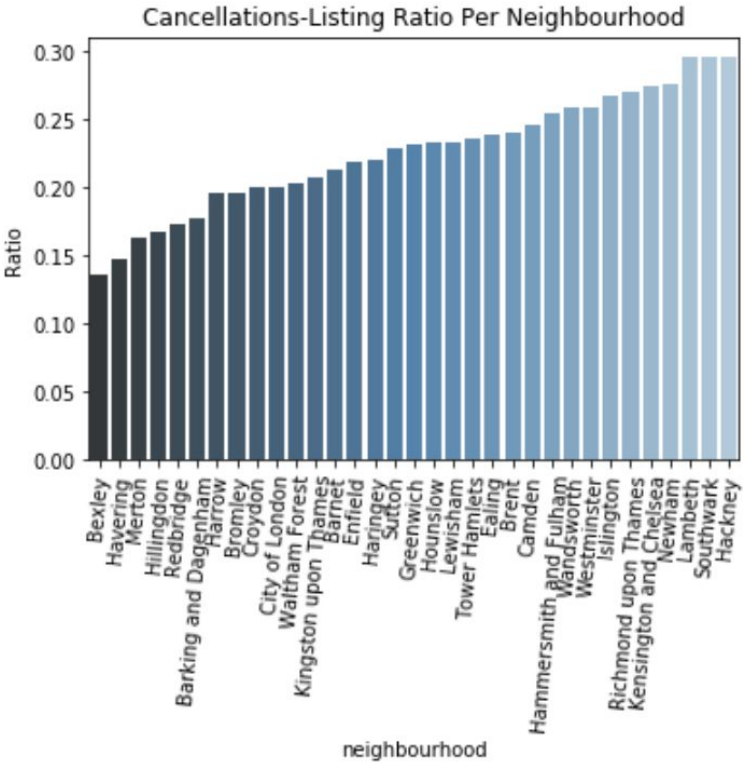
The following barplots were created to see the distribution of listings, reviews and cancellations across the London neighborhoods.

Which neighborhoods have the most Airbnb listings? Which neighborhoods have the most Airbnb reviews?Which neighborhoods have the most Airbnb cancellations?





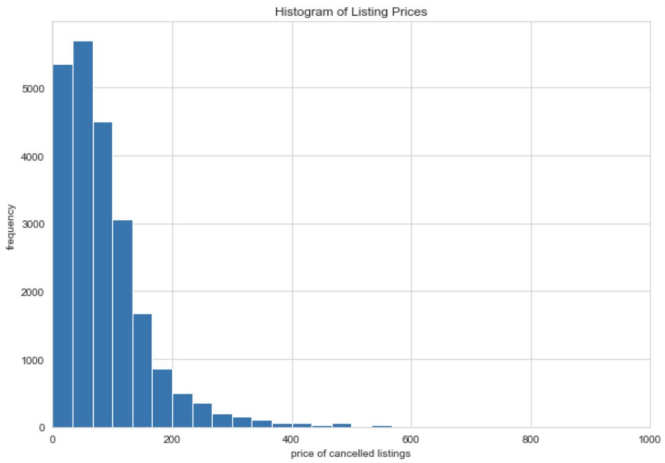




Westminster and Tower Hamlets are the top two neighborhoods with the highest frequency in all features--could this mean that listings in their neighborhoods have the highest likelihood of cancelling or is it just because they also have the most number of listings? This question is addressed by creating a barplot that shows the ratio of the number of cancellations to the number of listings per neighborhood.

This barplot shows us that 1) the distribution is no longer as spread out and 2) Westminster does not have the highest ratio.

Does price correlate with these cancellations?

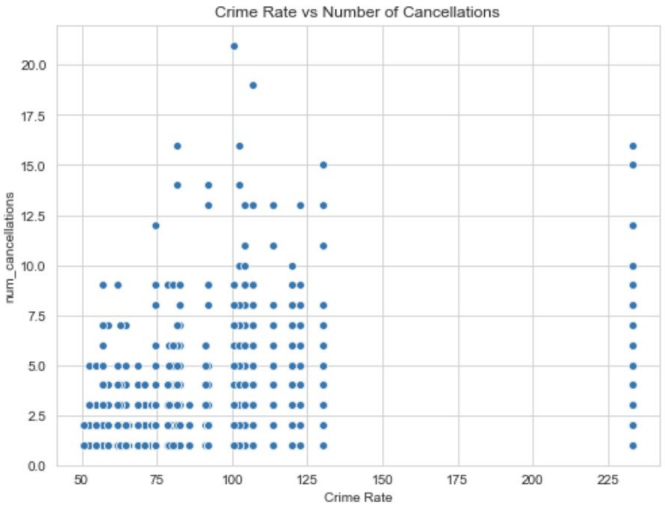


A histogram of the price of each cancelled listing reservation is created to study this relationship.

The histogram shows a negative correlation between the two variables.

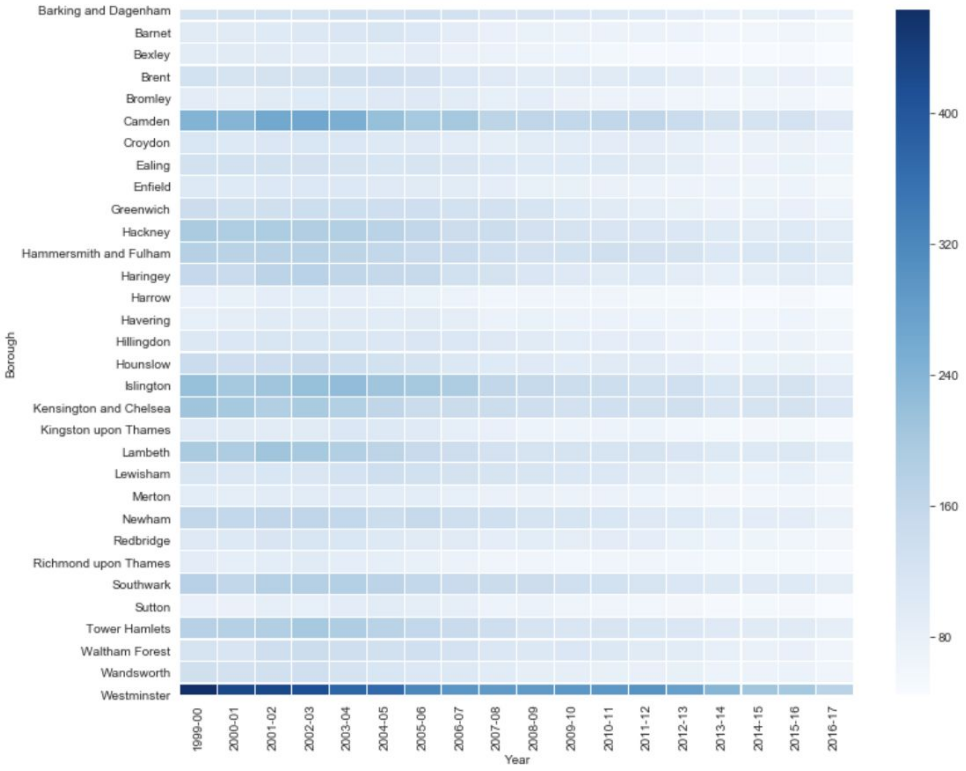
Does the neighborhood crime rate correlate with the cancellations in the area?

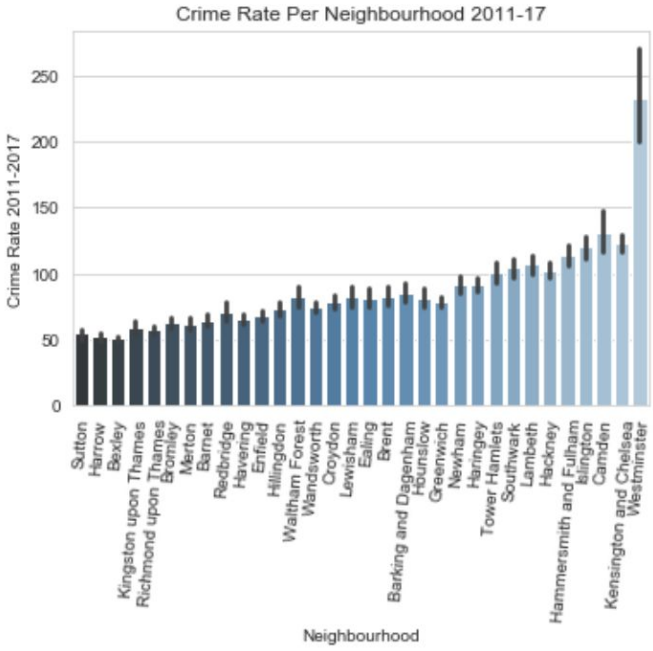
The relationship between the two seems random. We can confirm if the correlation is indeed insignificant by using statistical techniques later.



Which neighborhoods have the highest crime?

The heatmap shows crime has gone down for all of London since 1990. Westminster had the biggest change, although, it still remains to have the highest crime rate.





Which neighborhood currently has the highest crime?

Since the crime rates of all neighborhoods have decreased since 1990, we can just focus on the crime rates from 2011 (earliest Airbnb listing in our data) to 2017(latest crime rate data provided by the csv file).

Inferential Statistics

After studying trends between the Airbnb listings, we apply inferential statistics techniques to see if any of these relationships are statistically significant. We mainly focus on how these features are related to the number of cancellations of the listings.

For most of the variables, we make use of permutation tests to see if their respective correlations with the number of cancellations are statistically significant.

Cancellations vs. Price

Null Hypothesis: There is no significant correlation between the number of cancellations and the price of the listing.

Alternative Hypothesis: There is a significant correlation between the two.

First, we calculate the actual correlation between the number of cancellations and listing prices. The observed correlation is -0.02526.

We test to see if this observed correlation is statistically significant by performing a permutation test. We use np.random.permutation to reorder the listing prices and get their correlation with the number of cancellations each time.

In this case, our p-value is the ratio of the amount of times the absolute value of our correlation replicates was greater or equal to our observed correlation.

```
observed correlation between price and number of cancellations: -0.02526
p-value: 0.0000000000
We reject the null. There is a significant correlation between price and
number of cancellations
```

The p-value we get is really small. We then conclude that the correlation between price and the number of cancellations is statistically significant at the 1% level.

Cancellations vs Popularity (Number of Reviews)

Null Hypothesis: There is no significant correlation between the number of cancellations and the number of reviews a listing gets.

Alternative Hypothesis: There is a significant correlation between the two.

We perform the same test and conclude that their observed correlation of 0.21254 is statistically significant at the 1% level.

```
observed correlation between number of reviews and number of cancellation
s: 0.21254
p-value: 0.0000000000
We reject the null. There is a significant correlation between between th
e number of cancellations and number of reviews
```

Cancellations vs Demand (Days Booked)

Null Hypothesis: There is no significant correlation between the number of cancellations and the number of days the listing is booked.

Alternative Hypothesis: There is a significant correlation between the two.

We perform the same test and conclude that their observed correlation of -0.01100 is statistically significant at the 1% level.

```
observed correlation between number of days booked and number of cancella
tions: -0.01100
p-value: 0.0018000000
We reject the null. There is a significant correlation between the number
of cancellations and the number of days the listing is booked.
```

Cancellations vs Minimum Nights

Null Hypothesis: There is no significant correlation between the number of cancellations and the minimum nights required by a listing.

Alternative Hypothesis: There is a significant correlation between the two.

We perform the same test and conclude that their observed correlation of -0.00473 is **not** statistically significant at the 1% level.

```
observed correlation between minimum nights and number of cancellations:
-0.00473
p-value: 0.1630000000
We fail to reject the null. There is no significant correlation between t
he number of cancellations and the minimum nights required by a listing.
```

Cancellations vs Crime

Null Hypothesis: There is no significant correlation between the number of cancellations and the crime rate of the neighbourhood that the listing belongs in.

Alternative Hypothesis: There is a significant correlation between the two.

We perform the same test and conclude that their observed correlation of 0.01161 is statistically significant at the 1% level.

```
observed correlation between crime rate and number of cancellations: 0.01
161
p-value: 0.0009000000
We reject the null. There is a significant correlation between the number
of cancellations and the crime rate of the neighbourhood that the listing
belongs in.
```

Besides these features, we also want to check if the number of cancellations vary by categories such as neighborhood and room type. Since there are more than two in each category, we cannot perform ttests or permutation tests. Instead, we can use an F-test. In an F-test, we are able to compare the means of various groups and determine if they are equal by looking at their variances. With this technique, we are able to check variations between and within the groups.

Cancellation by Neighborhood?

Null Hypothesis: The average number of cancellations of each neighborhood are not statistically different from each other.

Alternative Hypothesis: The average number of cancellations of each neighborhood are statistically different from each other.

```
f-stat: 4.65329
p-value:0.0000000000
The average number of cancellations of each neighbourhood are statistical
ly different from each other.
```

The f-stat of 4.65 suggests that between-neighbourhood variance is 4.65 time the within-neighborhood variance. The small p-value suggests that the difference is statistically significant.

Cancellation by Room Type?

Null Hypothesis: The average number of cancellations of each room type are not statistically different from each other.

Alternative Hypothesis: The average number of cancellations of each room type are statistically different from each other.

```
f-stat: 11.68846
p-value:0.0000084047
The average number of cancellations of each room type are statistically d
ifferent from each other.
```

The p-value suggests that the differences in the number of cancellations among the room type are also statistically significant.

Findings

After studying trends through data visualization and applying inferential statistics techniques, we conclude that a listing's `number of cancellations` has significant correlation with the following variables:

- `Price`
  - The negative correlation indicates that lower-priced listings are more likely to cancel than those with higher prices
- `Reviews`
  - The positive correlation indicates that listings with more reviews are more likely to cancel
- `Availability`
  - The negative correlation indicates that listings with less bookings (more availability) are more likely to cancel
- `Crime rate`
  - The positive correlation indicates that listings that belong in neighborhoods with higher crime rates are more likely to cancel

We also conclude that the `number of cancellations` vary significantly across `neighborhoods` and `room type`. This tells us that certain neighbourhoods and room types have higher likelihoods in cancelling than others.

In-Depth Analysis Through Machine Learning

After checking which variables are statistically significant to our target variable, `num_of_cancellations`, we proceed to choose which machine learning model best suites our project.

In this study, we try out six different models:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Decision Trees
5. Gradient Boosting
6. Random Forest

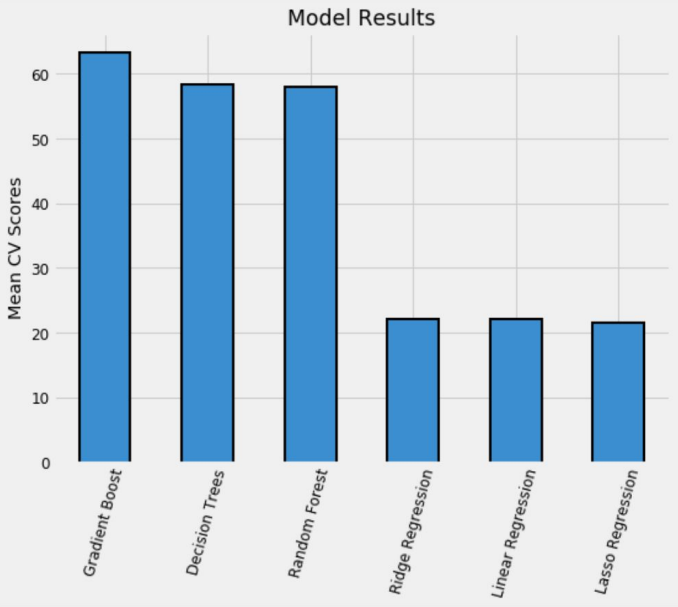
Dealing with Categorical Features
We have two categorical variables in our data frame: `room_type` and `neighbourhood`. Since Scikit Learn rejects categorical features by default, we make sure to create dummies for these two variables.

Training Set and Test Set
We are using 70% of the data as the training set and the remaining 30% as the test set.

Scoring
In this project, we are measuring accuracy by getting the average of 5-fold cross validation scores of each model. Since this is a regression problem, I use R^2 as the scoring metric. The R^2 is the indication of the goodness of fit of a set of predictions to the actual value.

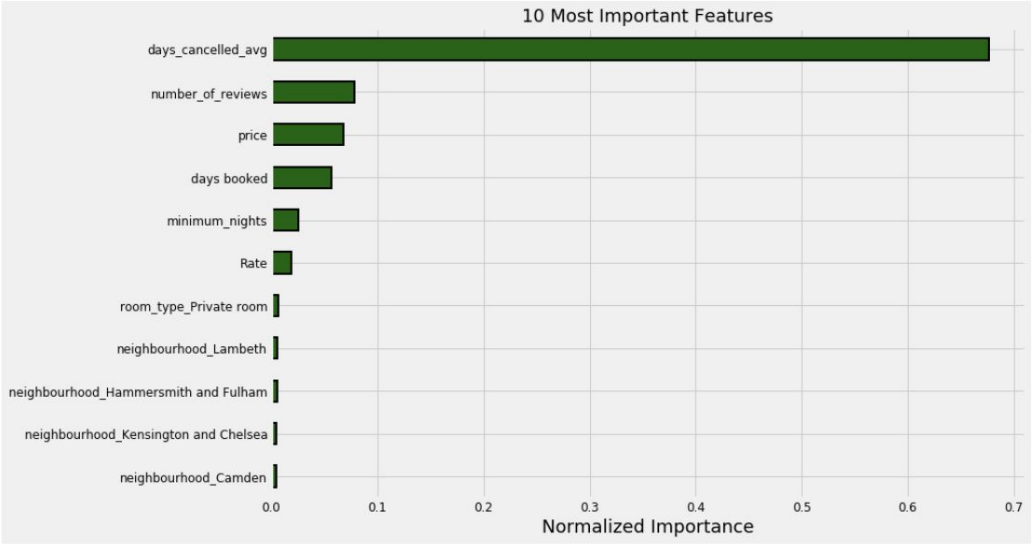
Comparing Model Performance

	cv_mean
Gradient Boost	63.220507
Decision Trees	58.331781
Random Forest	58.023593
Ridge Regression	22.145139
Linear Regression	22.144672
Lasso Regression	21.637382



The Gradient Boosting Regressor performed best while Lasso Regression performed worse. To further optimize the performance of the Gradient Boosting Regressor, we go back into feature selection by determining which features create ‘noise’. Doing so will help us prevent overfitting.

Feature Selection



We see that the non-categorical variable that has the least importance is the crime rate of each neighborhood. We simply drop this from our data frame and fit it to our Gradient Boosting Regressor again to see if the average CV score improves.

After running this, we see that average CV score improves from 63.22% to 63.36%. This indicates that the feature only added noise to our model.

Hyperparameter Tuning

To further optimize our model's performance, we try to tune some of the parameters of the `GradientBoostingRegressor()` function. In this project, I try to find the best values of the following:

- max_depth
- n_estimators
- learning_rate

To do so, we create a list of possible values for each of the parameters for the forloop to search over. We then take the parameter that performs best.

The results are the following:

`n_estimators`

```
{50: 63.54920965957492,
 100: 63.356236092445364,
 200: 62.96948529810927,
 300: 62.38900703879856,
 400: 61.95043209337545,
 500: 61.59295400799691}
```

At an average CV score of 63.55%, our best `n_estimator` is 50.

`max_depth`

```
{3: 63.54920965957492,
 4: 63.29581772212089,
 5: 62.81870742497017,
 6: 62.45862378402221,
 7: 61.81097868920081}
```

At an average CV score of 63.55%, our best `max_depth` is 3.

`learning_rate`

```
{0.0001: 0.6147246537337336,
 0.001: 5.940723336614342,
 0.01: 39.658814557812185,
 0.1: 63.54920965957492}
```

At an average CV score of 63.55%, our best `learning_rate` is 0.1.

After getting the best `n_estimator` of 50, our best CV score has remained at 63.55% when looking for best `max_depth` and `learning_rate` because the best values we found for `max_depth` and `learning_rate` are the GradientBoostingRegressor()'s default values for the parameters.

Despite the feature selection and hyperparameter tuning, the score only improves by a few percentage points. This tells us that we need more features for better prediction.

Results

After selecting the best performing model, and further optimizing it through feature selection and hyperparameter tuning, our final results are:

```
Train set score: 64.60%
Test set score: 61.45%
```

As can be expected, the test set score is less than the train set score. Although, the difference is not much, which indicates that overfitting was not a problem.

Limitations

The model could improve vastly if applied to datasets that provide more features of each listing.