Machine Learning Engineer Nanodegree

Capstone Project

Joaee Chew
September 14, 2017

I. Definition

(approx. 1-2 pages)

Project Overview

AirBnb is a popular home-sharing platform, enabling people all over the world to lease or rent short-term lodgings. For potential hosts, this is a potentially lucrative option spare rooms or even beds. However, it is difficult for new hosts to ascertain how much their property could earn them and whether it will be a worthy investment. Furtherm their beloved penthouse adorned with designer furniture, compared to your run-of-the-mill IKEA filled apartment?

This project seeks to solve the problem by building a predictive model for the potential earnings of Airbnb listings, taking into account text descriptions to capture a rich a While there are available commercial tools such as AirSorted (http://www.airsorted.uk/) that predict Airbnb value, these models are generic with minimal features used and

In addition, this project will implement an interpretability layer, using the TreeInterpreter (https://github.com/andosa/treeinterpreter) package to decompose predictions into provides greater transparency and assurance to hosts that the nuances of their unique home has been captured by the model.

To build this model, I use the dataset provided by Inside Airbnb (http://insideairbnb.com/get-the-data.html), where publicly available information about a city's Airbnb's list independent, non-commercial use. This includes details about the listing such as no. of rooms, guests available, description, location as well as information about the yiel

Specifically, I will use the detailed listings information for London listings active from 3rd October 2016 – 4th March 2017. An active listing is defined as a property that has time period. After data cleaning, the dataset has 9722 rows and 16 columns. Importantly, the dataset contains the listings description as well as headers in text format tha

Problem Statement

To build this model, the concept of yield is used as a proxy for potential future earnings. Yield is defined as the amount of revenue that a property will earn over a year. The 'San Francisco Model' based on price, average length of stay and review rate. A review rate of 50% (http://insideairbnb.com/about.html) to convert reviews to estimated b

> **This project will predict the yield on prospective AirBnb listings in London and test whether text mining can lead to an increase in predictive accuracy.**

The intended workflow for this project is as follows:

1. Read, clean and preprocess the data:
   a. Remove unnecessary features. This includes data that does not help with predicting yield, such as listing URL or scrape ID.
   b. Remove 'leaking' features. This includes data that should not be used to predict yield, such as ratings (which are not present at the point of listing), or the no. of oth (model should be built on the property only and not the host).
   c. Removing incomplete listings (e.g. listings with no price, or with availability less than 50% which indicates part-time listings that are not comparable).
   d. Impute missing values (e.g. using the most frequent or median where appropriate).
2. Build the model. I will use three models of increasing complexity, both to test for increase in predictive power but also to act as challenger models to the final model.
   a. I will begin with a simple linear regression to establish a benchmark accuracy.
   b. I propose the use of a decision tree as a next step.
   c. Finally, I will use a random forest ensemble model as the final model.
3. Feature engineering. This will be done iteratively with step 3, with new features tested for improvement in model accuracy. Potential ideas include:
   a. Using natural language techniques to extract features from the description. This involves the following NLP pipeline:
   i. Text data preparation (creating corpus, word stemming, creating document term matrix, removing sparse terms etc.)
   ii. Selection of features using bag-of-words and/or TF-IDF
   iii.Use topic modelling to discover any natural topics in the listing descriptions
   b. Bucketing categorical features to reduce dimensionality (eg. Whole apartment vs. non-whole apartment instead of entire class size).
4. Test the model. Here I will print final accuracy measures and produce insights such as feature importance.
5. Add interpretation layer to the final model. As a final step of the project, I will add an interpretation layer. This decomposes each prediction into feature contribution co and helps improve transparency into model predictions.

Metrics

Mean squared error is used as the measure for model accuracy. This measures the average of the square of the errors between the actual yield and predicted yield. The fo

MSE is a commonly used error score for regression models, and allows for an intuitive measure of error i.e. an MSE of 10000 indicates the yield model is off by £100 (squa negative, and values closer to zero are better.

II. Analysis

(approx. 2-4 pages)

Setting up

Read data

```
Dataset has 53904 rows, 94 columns.

//anaconda/lib/python3.6/site-packages/IPython/core/interactiveshell.py:2717: DtypeWarning: Columns (88) have mixed types. Spe
t low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

Data Exploration

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input file data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics abou abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of writing this section:

- *If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?*
- *If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?*
- *If a dataset is **not** present for this problem, has discussion been made about the input space or input data for your problem?*
- *Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)*

The data set as scraped by Inside AirBnb is extremely comprehensive, containing 53904 rows and 94 columns. However, many of the data collected are unnecessary met review data that causes leaking (e.g. review_scores), or has low relevance to the yield model (e.g. security_deposit, is_location_exact, notes).

To keep the model manageable and effective, I will focus on the following features:

1. 'name': Listing header text for text mining.
2. 'description': Listing description text for text mining.
3. 'property_type': Type of property e.g. apartment, house etc.
4. 'room_type': Type of room e.g. private, shared etc.
5. 'accommodates': No. of people the listing can accommodate.
6. 'bathrooms': No. of bathrooms.
7. 'bedrooms': No. of bedrooms.
8. 'beds': No. of beds.
9. 'square_feet': Size of listing.
10. 'price': Price of listing.
11. 'cleaning_fee': Cleaning fee.
12. 'guests_included': No. of guests included in the base price.
13. 'extra_people': Price for extra guests not included in the base price.
14. 'minimum_nights': Minimum no. of nights required for booking.
15. 'availability_365': No. of nights the listing is available for booking. (Note: This is a 'leaking' feature as listings already booked and leaks the popularity. However, this needs to be used for filtering part-time listings and will be remov
16. 'reviews_per_month': Average no. of reviews listing receives per month. Used to calculate yield.
17. 'latitude': Latitude of listing.
18. 'longitude': Longitude of listing.

A sample of the reduced dataset along with summary statistics is shown below.

```
Dataset has 53904 rows, 18 columns.
```

	description	property_type	room_type	accommodates	bathrooms	bedrooms	beds	square_feet	price	cleaning_fee	guests_included	extra_
id												
15896822	My place is close to TK Max, John Lewis, Marks...	Apartment	Private room	1	1.0	1.0	1.0	NaN	$23.00	NaN	1	$8.00
4836957	This lovely spacious double bedroom is set in ...	Apartment	Private room	2	1.0	1.0	1.0	NaN	$50.00	NaN	1	$0.00
13355982	Spacious double bedroom, because of the light,...	Apartment	Private room	2	1.0	1.0	1.0	NaN	$24.00	NaN	1	$0.00
13472704	My place is good for couples, solo adventurers...	House	Private room	2	1.5	1.0	1.0	NaN	$50.00	NaN	1	$0.00
17430865	very new decorated beautiful room and very com...	House	Private room	1	1.0	1.0	1.0	NaN	$25.00	NaN	1	$0.00

```
Dataset has 53904 rows, 18 columns.
```

	accommodates	bathrooms	bedrooms	beds	square_feet	price	cleaning_fee	guests_included	extra_people	minimum_ni
count	53904.000000	53644.000000	53811.000000	53731.000000	582.000000	53904.000000	32483.000000	53904.000000	53904.000000	53904.00000
mean	3.036676	1.262751	1.353980	1.708027	577.508591	96.099622	37.115907	1.407428	6.686238	3.285229
std	1.907429	0.547699	0.841912	1.201165	726.154243	117.641082	33.961914	1.040308	12.705758	28.536837
min	1.000000	0.000000	0.000000	0.000000	0.000000	8.000000	3.000000	1.000000	0.000000	1.000000
25%	2.000000	1.000000	1.000000	1.000000	108.000000	42.000000	15.000000	1.000000	0.000000	1.000000
50%	2.000000	1.000000	1.000000	1.000000	484.000000	70.000000	30.000000	1.000000	0.000000	2.000000
75%	4.000000	1.500000	2.000000	2.000000	819.500000	119.000000	50.000000	1.000000	10.000000	3.000000
max	16.000000	8.000000	10.000000	16.000000	10710.000000	7000.000000	517.000000	16.000000	240.000000	5000.000000

The summary statistics reveal certain interesting insights about the dataset, and these are explored individually below.

Removing part-time listings

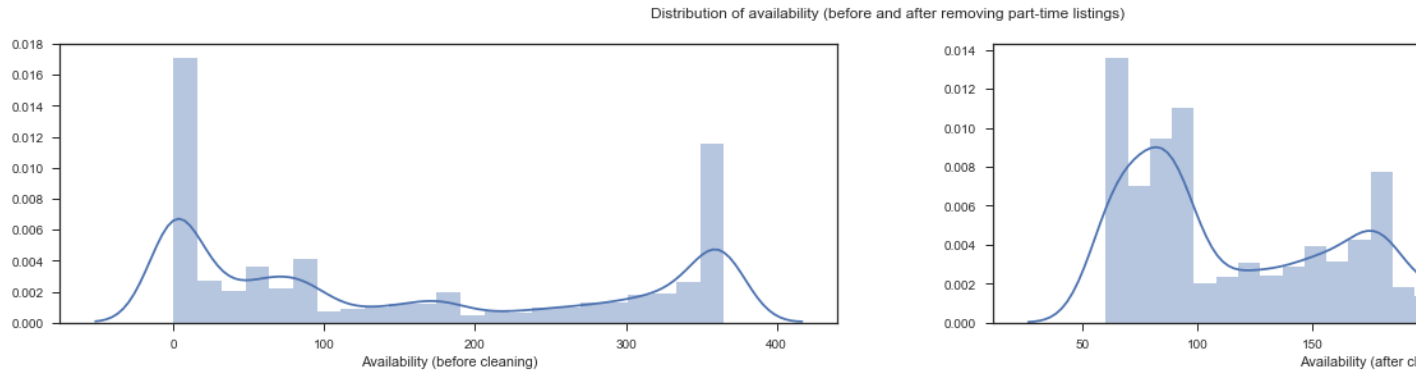
An Airbnb host can setup a calendar for their listing so that it is only available for a few days or weeks a year. Other listings are available all year round (except for when it i use Inside AirBnb's definition of highly available being >60 days a year.

I will remove part-time listings by removing availability <60 days a year, and also newly-listed listings with availability >300 days (using the other tail end as approximation)

The below distribution curve shows a bi-modal distribution for availability, demonstrating two concentrations of listing types. The lower availability peaks below 50 days in listing, while the higher availability peaks at almost at the maximum 365 indicating a dedicated rental property.

Once used to clean the data, I will remove availability as it is a leaking feature. It reveals how many bookings there will be in the coming year.

```
Dataset has 18502 rows, 18 columns.
Dataset has 18502 rows, 17 columns.
```

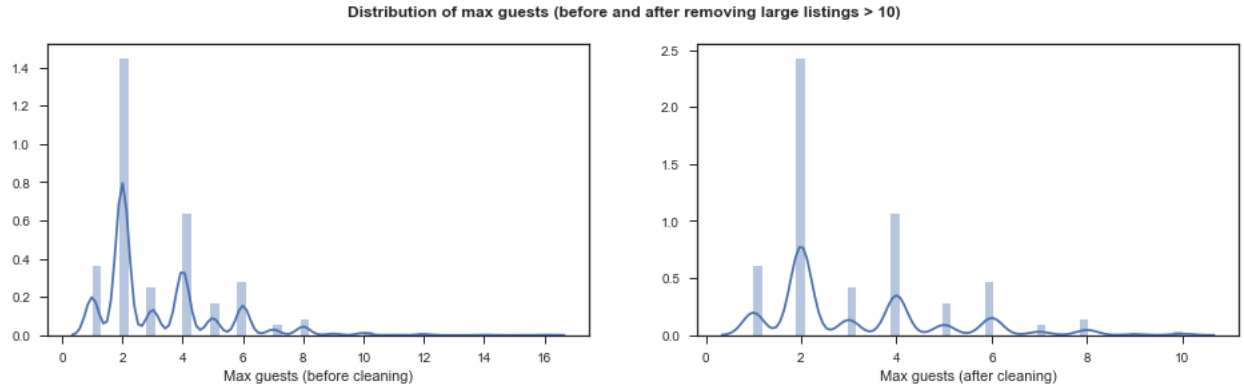


Removing large houses

For this model, I will remove extremely large rentals accommodating more than 10. The no. of persons a listing can accommodate will directly impact the calculation of yie
influence this number, this is less within their control and more an indicator of size. It is also unlikely that we have enough data to make tailored predictions for such listing

```
You dropped 116 rows.
Dataset has 18386 rows, 17 columns.

<matplotlib.axes._subplots.AxesSubplot at 0x131360e10>
```

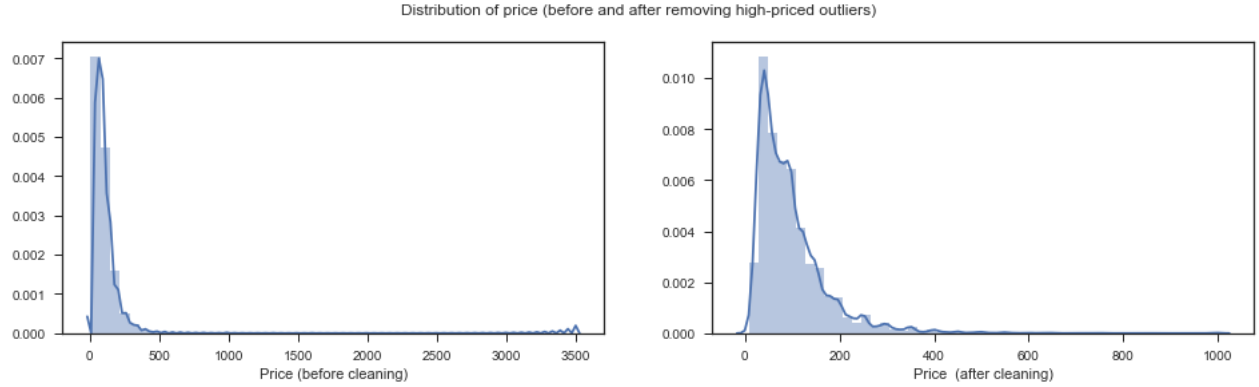


Removing high-priced rentals

The below distribution curve for price is right-skewed with a long tail of low frequency high-priced rentals. For this model, I will remove extremely high priced rentals above
is also unlikely that we have enough data to make tailored predictions for such bespoke listings

```
You dropped 8 rows.
Dataset has 18378 rows, 17 columns.

<matplotlib.axes._subplots.AxesSubplot at 0x124c365c0>
```



Changing rare category types into 'other' bucket

For certain categories, we need to have a big enough sample set to ensure that it is statistically significant, i.e. the category is common enough to isolate correlation betw
benefit of reducing dimensionality. For property types, I create an 'Others' category for all except for the most common ones - Apartment, House and Bed & Breakfast.

```
Apartment             13478
House                  4060
Bed & Breakfast         360
Townhouse               103
Loft                     85
Other                    83
Guesthouse               37
Serviced apartment       33
Dorm                     33
Boat                     32
Bungalow                 19
Condominium              18
Cabin                    16
Boutique hotel           10
Hostel                    4
Villa                     3
Chalet                    2
Lighthouse                1
Yurt                      1
Name: property_type, dtype: int64
```

```
//anaconda/lib/python3.6/site-packages/pandas/core/indexing.py:179: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

```
//anaconda/lib/python3.6/site-packages/pandas/core/indexing.py:179: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
  self._setitem_with_indexer(indexer, value)
```

Calculation of yield

As disuccsed earlier, **Yield** is defined as the amount of revenue that a property will earn over a year. This is calculated as follows:

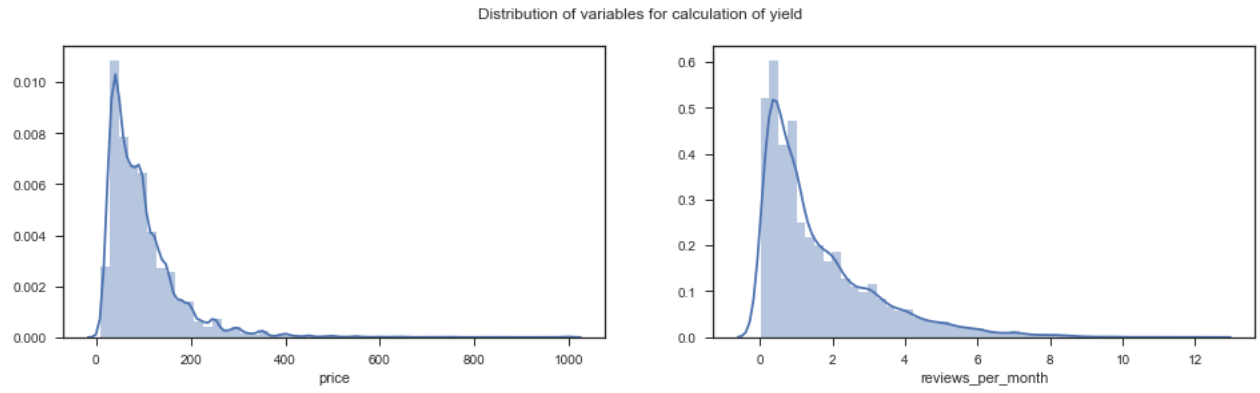
> AVERAGE LENGTH OF STAY X PRICE X NO. OF REVIEWS/MTH X REVIEW RATE X 12 MONTHS

Using Inside AirBnb's San Franciso Model, an average length of stay in London is 3 nights with a review rate of 50%. Note that this is fairly conservative, and yield in realit

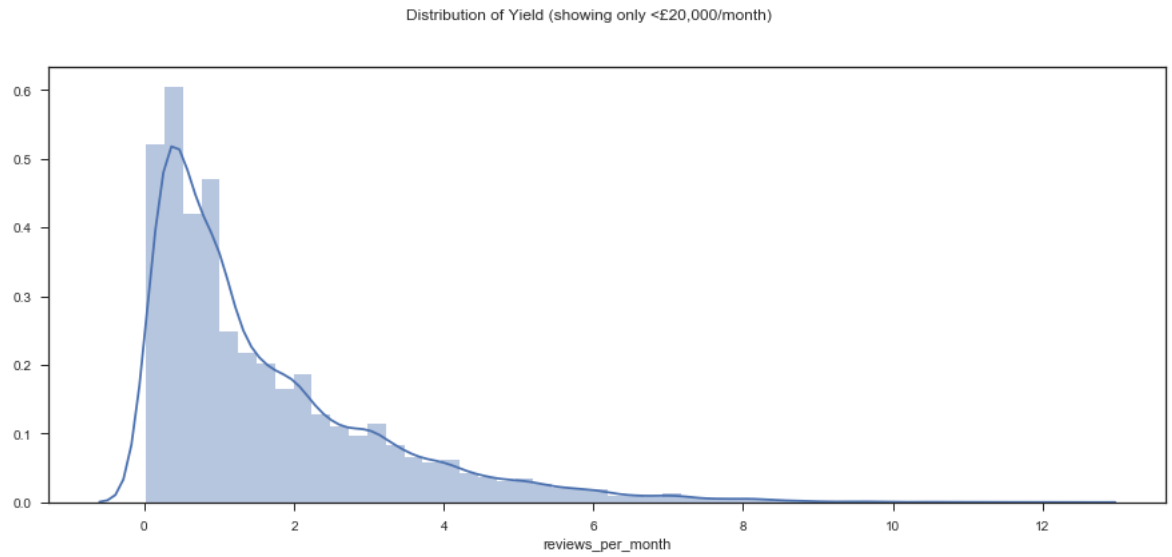
Analysing the distributions for price, reviews and resulting yield makes intuitive sense. The bulk of listings are less than £100/night probably for a private room, and receive weekend. This indicates occupancy rates of 10-50%, earning the average AirBnb host £15,739 per listing in a year, going up to £21,480 for upper quartile listings.

Sense checking against the median residential rents in Central London (https://www.london.gov.uk/what-we-do/housing-and-land/renting/london-rents-map,) of £7,800 al

```
<matplotlib.axes._subplots.AxesSubplot at 0x12535d898>
```




```
Dataset has 18378 rows, 17 columns.
```




```
count      9722.000000
mean      15739.358560
std       17900.206348
min          82.080000
25%        3830.400000
50%        9288.000000
75%       21480.660000
max      250560.000000
Name: yield, dtype: float64
```

Missing values

A preview of the dataset above shows that there are missing values in some of the features such as 'beds', 'price' and 'reviews_per_month'. Depending on the nature of th differently.

In 'square_feet', almost 90% of the rows are missing and I will choose to remove the feature completely

As 'yield' is an important target variable, I will completely remove any rows that have this datum missing. This will also take care of 'price' and 'reviews_per_month' as thos 'yield'.

For the remaining features with a small number of missing values, I impute the categorical features with the most frequent category and numericla features with the media

```
Dataset has 9722 rows, 16 columns.
```

```
False
```

Topic modelling (Description)

For textual information, we use a Natural Language Processing pipeline to convert the corpus into a Document-Term-Matrix, whereby each listing (document) consists of a this in place, we can use Latent Dirichlet Allocation (LDA) to discover topics inherent in the corpus, classify the corpus according to the learned topics and use them as fea

> LDA is a generative Bayesian inference model that associates each document with a probability distribution over topics, where topics are probability distributions over v large volumes of text and is a more human interpretable method of topic modelling.

In this project, I model the listing descriptions using LDA to explore emerging topics. The listing description field contains the bulk of the textual information found in a listi information that is not captured elsewhere. What emergent topics might we see from the rich text describing the history of a home, a host's beloved neighbourhood or it's

```
Dataset has 9722 rows, 15 columns.
```

	description
id	
14912894	Lovely bright and sunny room overlooking the g...
14296637	My place is situated about 10-15mins walking d...
6222799	Sunny ensuite room in the newly converted loft...
7327883	Surbiton is very convenient for getting into ...
7515814	Our one bedroom apartment is a comfortable spa...

```
CPU times: user 6.55 s, sys: 82.2 ms, total: 6.63 s
Wall time: 6.71 s
```

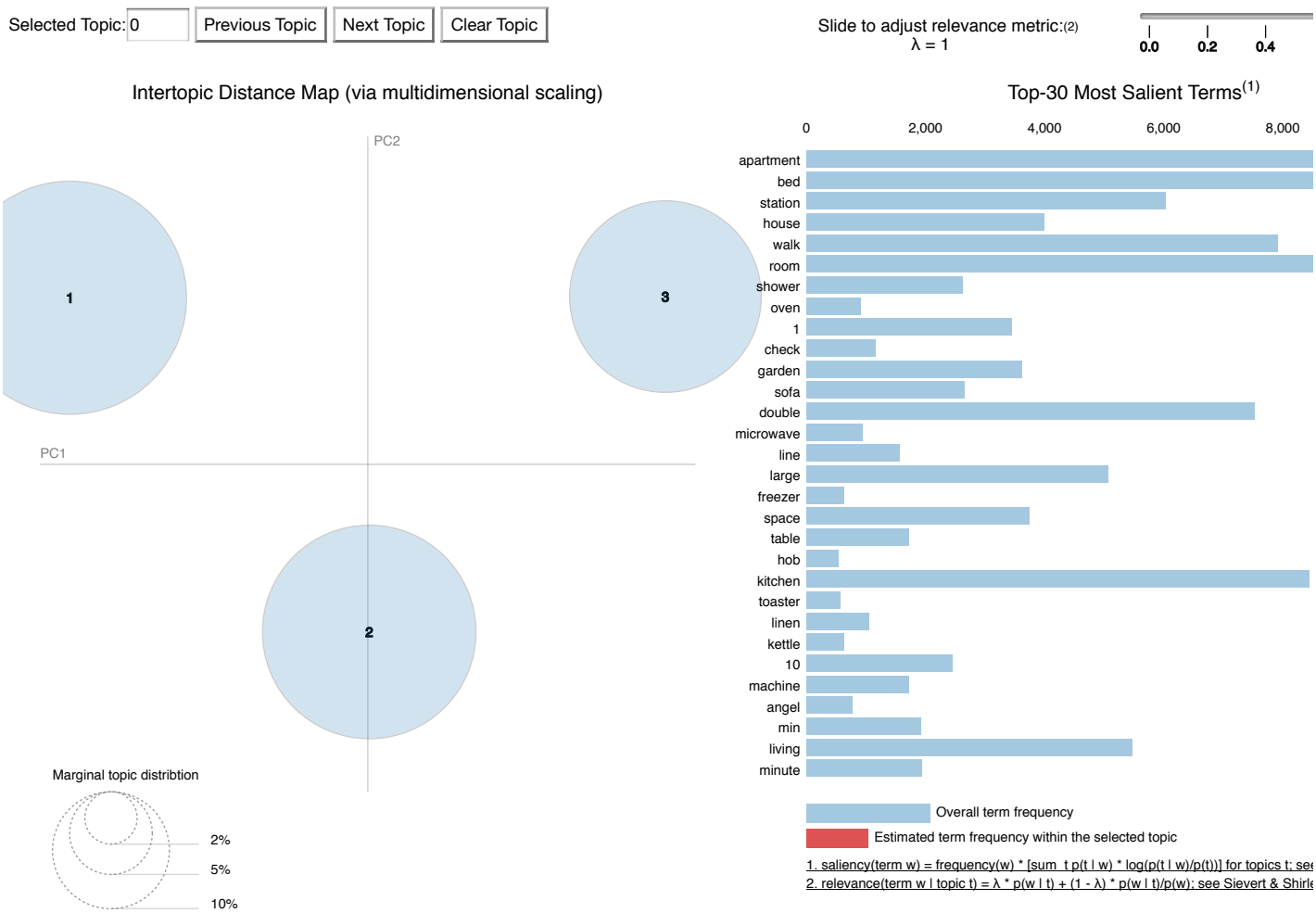
```
CPU times: user 24.3 s, sys: 39 ms, total: 24.3 s
Wall time: 24.3 s
```

Exploratory Visualization

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

- *Have you visualized a relevant characteristic or feature about the dataset or input data?*
- *Is the visualization thoroughly analyzed and discussed?*
- *If a plot is provided, are the axes, title, and datum clearly defined?*

```
CPU times: user 15.6 s, sys: 123 ms, total: 15.7 s
Wall time: 16.1 s
```



I chose to cluster the document terms into 3 topics as it was the largest number of topics that had no overlaps with very distinct topics. The separate topics also makes s

Based on the top terms in each topic, the topics for listings could be described as:

- Topic 1 - Apartments: This seems to focus on modern apartments, with words such as 'kitchen', 'bed', 'modern and 'space' standing out. This might appeal to young with modern facilities and furnishings.
- Topic 2 - Location: This focuses on location with words such as 'location', 'business', 'station' and 'central'. This is likely to appeal to business travellers who value cc links.
- Topic 3 - Houses: This focuses on words associated with more traditional housing such as 'house', 'quiet', 'welcome' and 'clean'. This might appeal to families who w safety of a residential neighbourhood.

Algorithms and Techniques

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the char domain. Questions to ask yourself when writing this section:

- *Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*
- *Are the techniques to be used thoroughly discussed and justified?*
- *Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

I intend to use 3 regression algorithms:

1. Linear Regression

To establish a baseline model

2. Decision Tree

This is a more complex model that can capture nonlinear relationships in the dataset whilst still retaining interpretability. This is also robust to missing values, and has the a with feature selection based on feature importance.

3. Random Forest

This is the most complex ensemble model built from decision trees, and should be able to provide additional accuracy. While it sacrifices interpretability, sense checking c model whilst added an interpretation layer using Tree Interpreter allows model users to break down individual predictions.

Benchmark

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behi not an established result) should be discussed. Questions to ask yourself when writing this section:

- *Has some result or value been provided that acts as a benchmark for measuring performance?*
- *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

To my knowledge, there are no existing Airbnb pricing models released to the public. However, I will use a hold-out set to test my model, and the mean squared error will b model.

I will also build challenger models, and in particular start with a linear regression model to act as the baseline benchmark. A linear regression model is chosen as it is simp be easily interrogated. The methodology for building this model can be found below in the methodology section.

After building the model, an MSE of 254474109 with variance of 0.18 is established.

III. Methodology

(approx. 3-5 pages)

Data Preprocessing

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteri will be addressed and corrected here. Questions to ask yourself when writing this section:

- *If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?*
- *Based on the **Data Exploration** section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?*
- *If no preprocessing is needed, has it been made clear why?*

Before performing any machine learning, the dataset needs to be pre-processed. The following steps were taken for this project and discussed in the 'Data Exploration' se

1. Converting data types (e.g. string into integers)
2. Removing part-time listings
3. Removing large houses
4. Removing high-priced rentals
5. Changing minor category times into ‘Other’
6. Calculation of yield
7. Topic modelling

In addition to those steps already taken, missing values need to be imputed, features have to be selected and categorical features have to be encoded.

	accommodates	bathrooms	bedrooms	beds	price	guests_included	extra_people	minimum_nights	reviews_per_month	latit
count	9722.000000	9722.000000	9722.000000	9722.000000	9722.000000	9722.000000	9722.000000	9722.000000	9722.000000	9722.000
mean	3.470582	1.267846	1.431393	1.894260	141.301893	1.681033	9.123637	3.473668	1.631914	51.50993
std	1.867488	0.517121	0.846198	1.212666	95.602315	1.235965	14.000453	10.015233	1.521587	0.039996
min	1.000000	0.000000	0.000000	1.000000	17.000000	1.000000	0.000000	1.000000	0.020000	51.31956
25%	2.000000	1.000000	1.000000	1.000000	70.000000	1.000000	0.000000	1.000000	0.490000	51.48927
50%	3.000000	1.000000	1.000000	2.000000	121.000000	1.000000	5.000000	2.000000	1.080000	51.51378
75%	4.000000	1.500000	2.000000	2.000000	180.000000	2.000000	15.000000	3.000000	2.300000	51.53312
max	10.000000	8.000000	7.000000	10.000000	1286.000000	10.000000	230.000000	300.000000	10.530000	51.66364

Feature selection

I also remove features that are not necessary for the machine learning phase.

'Price' and 'reviews_per_month' are removed as they are used to calculate 'yield' and would have a strong correlation with 'yield'.

```
Dataset has 9722 rows, 13 columns.
```

Dummy encoding

Finally, categorical features need to be encoded to be treated properly by the machine learning algorithms. As the algorithms require numerical features only, the different
However, the categorical features present do not have any order inherent in them (eg. apartment vs. house) and thus needs to be encoded where 1 is given if the category

```
Dataset has 9722 rows, 19 columns.
```

Implementation

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abunda
carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

- *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
- *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
- *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

Once the dataset has been processed it is now ready to for the models to be built.

1. As a first step, the target variable 'yield' has to be seperated from the response variables to from the X and y data sets.
2. Both X and y data sets and respectively split into a training set and a testing set. I chose to use 30% of the data set for the testing set. As part of the splitting process
ensure there is no inherent or hidden order to the data set and that both the testing and training sets are a fair representative of the complete data set.
3. In the training phase, the training set is fitted to the 3 algorithms set out in the 'Algorithms and Techniques' section. During this process, the models are tuned using s
This allows us to set out a grid of parameters to iteratively train and set the models on and ultimately choose the best parameters. As tree-based algorithms are being
are the max no. of features used, the max depth of trees, the minimum sample split and the no. of trees grown (for random forest).
4. Once the 3 models have been fit, they are scored on the testing set. To do this, we predict the 'yield' variable using the trained models on the response data set. The
against the actual 'yield' values in the test set using the mean squared error (MSE) as set out in the 'Metrics' section. Variance is also measured alongside MSE to hav
models.
5. After each model is trained and tested, the results are compared to check that more complicated models do indeed produce a better accuracy.

	accommodates	bathrooms	bedrooms	beds	guests_included	extra_people	minimum_nights	latitude	longitude	yield	property_type_Ap
id											
14912894	2.0	1.0	1.0	1.0	1.0	9.0	2.0	51.431117	-0.309713	2595.60	1
14296637	1.0	1.0	1.0	1.0	1.0	10.0	1.0	51.421323	-0.278367	807.84	1
6222799	3.0	1.0	1.0	1.0	1.0	10.0	1.0	51.419415	-0.288245	5113.44	0
7327883	2.0	1.0	1.0	1.0	1.0	10.0	1.0	51.390646	-0.310919	972.00	1
7515814	2.0	1.0	1.0	1.0	1.0	0.0	2.0	51.413554	-0.296376	12947.04	1

Linear regression model

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

```
Mean squared error: 254474109.10
Variance score: 0.18
```

Decision tree model

```
DecisionTreeRegressor(criterion='mse', max_depth=5, max_features=None,
           max_leaf_nodes=None, min_impurity_decrease=0.0,
           min_impurity_split=None, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           presort=False, random_state=42, splitter='best')
```

```
Mean squared error: 253282661.10
Variance score: 0.18
```

Random forest model

```
CPU times: user 2.13 s, sys: 12.3 ms, total: 2.14 s
Wall time: 2.15 s
```

```
Mean squared error: 212548432.31
Variance score: 0.31
```

Refinement

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusti acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as ne writing this section:

- *Has an initial solution been found and clearly reported?*
- *Is the process of improvement clearly documented, such as what techniques were used?*
- *Are intermediate and final solutions clearly reported as the process is improved?*

In the 'Problem Statement' section, it was first highlighted that one of the key aspects of this project is to test the predictive value of textual information in AirBnb listings. made is to include the earlier topics that have been modelled using LDA (explained in the 'Data Exploration' section). Each listing is classified into one topic based on prob model. This topic is then added in as a few feature.

Before this, the tuned random forest model had the best result with the lowest MSE of 212548432 and the highest variance of 0.31. After including the description topics, variance increases further to 0.32. This validates the hypothesis that there is valuable information as modelling description text as topics increases the model's predictive

```
Dataset has 9722 rows, 22 columns.
```

```
CPU times: user 2.31 s, sys: 9.99 ms, total: 2.32 s
Wall time: 2.33 s

RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=10,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_decrease=0.0, min_impurity_split=None,
           min_samples_leaf=1, min_samples_split=4,
           min_weight_fraction_leaf=0.0, n_estimators=150, n_jobs=1,
           oob_score=False, random_state=42, verbose=0, warm_start=False)
```

```
Mean squared error: 211329730.05
Variance score: 0.32
```

Tuning the final model

After confirming that including NLP features helps to improve accuracy, it is included in the final model. The parameters for the final Random Forest model is then tuned us

A parameters grid of hyperparameters are created for GridSearchCV to iteratively use and test for the best cross-validated accuracy. The paramters being tuned are: Num and the minimum samples split. After tuning, the best parameters are a max_depth of 15, minimum samples split of 10 and number of estimators 200.

The final MSE lowers to 200185564 with a variance score of 0.33. This shows that the model is fairly complex, with a large number of trees and depth.

```
CPU times: user 34min 33s, sys: 31.8 s, total: 35min 5s
Wall time: 38min 26s
```

```
Mean squared error: 200185564.28
Variance score: 0.33
Tuned Model Parameters: {'bootstrap': True, 'criterion': 'mse', 'max_depth': 15, 'max_features': 'auto', 'min_samples_split':
```

IV. Results

(approx. 2-3 pages)

Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chos
should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model's solution is affected (th
to ask yourself when writing this section:

- *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*
- *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*
- *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*
- *Can results found from the model be trusted?*

To test the robustness of the final model, sensitivity analysis is done by making adjustments to the inputs and evaluating the size of changes in the outputs. The adjustmer

1. Changing the random state: This will cause different data points to be randomly selected during the train/testing split and during the random forest training process. B
we can ensure that model training is not a fluke and can be replicated on different states. Changing the state from 42 to 49, the MSE remains in the same ballpark at 2
same.
2. Changing the split proportions between training and test set: By changing the amount of data split between training and testing, we can ensure that the model retains
testing data. A fairly conservative split of 30% for the test set had been chosen initially. By reducing this to 10%, the MSE drops to 205098824 which is to be expecte
radicially different and the model is fairly robust.
3. Changing input data: By manually tweaking some of the input data, it is possible to see how senstive the model is. If a small change in inputs lead to an unrealistic ch
unlikely to be robust. As an example, I take a random data point in the test data set and change the 'accommodates' feature which is one of the most significant featu
accommodates 4 persons and the model predicts a yield of £17,400 annually. Adjusting it to accommodate for 2 reduces the predicted yield to £13,377 while increasi
it to £18,106. This change is in line with expectations, as a property that can accommodate less guests should have a smaller yield and vice-versa. Importantly, increa
smaller increase in yield than a decrease in yield for decreasing the number of guests. This makes sense as increasing the number of guests without changing any oth
needs to share of the the listing facilities and hence have less yield per guest.

```
Mean squared error: 211329730.05
Variance score: 0.32
```

```
Mean squared error: 205098824.71
Variance score: 0.32
```

```
accommodates                     4.000000
bathrooms                        1.000000
bedrooms                         2.000000
beds                             3.000000
guests_included                  2.000000
extra_people                     5.000000
minimum_nights                   2.000000
latitude                        51.467320
longitude                       -0.167568
property_type_Apartment          1.000000
property_type_Bed & Breakfast    0.000000
property_type_House              0.000000
property_type_Other              0.000000
room_type_Entire home/apt        1.000000
room_type_Private room           0.000000
room_type_Shared room            0.000000
bed_type_Non-Real Bed            0.000000
bed_type_Real Bed                1.000000
topics_description_Budget        1.000000
topics_description_Business      0.000000
topics_description_Luxury        0.000000
Name: 1400514, dtype: float64
```

```
//anaconda/lib/python3.6/site-packages/ipykernel/__main__.py:2: FutureWarning: reshape is deprecated and will raise in a subse
s.reshape(...) instead
  from ipykernel import kernelapp as app

[ 17400.42410558]
[ 13377.87392587]

//anaconda/lib/python3.6/site-packages/ipykernel/__main__.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
//anaconda/lib/python3.6/site-packages/ipykernel/__main__.py:5: FutureWarning: reshape is deprecated and will raise in a subse
s.reshape(...) instead

[ 18106.0007421]

//anaconda/lib/python3.6/site-packages/ipykernel/__main__.py:7: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
//anaconda/lib/python3.6/site-packages/ipykernel/__main__.py:8: FutureWarning: reshape is deprecated and will raise in a subse
s.reshape(...) instead
```

Another method of evaluating the model is to look at what the model is saying and sense checking the results. By looking at feature importance, we are able to determine important features and whether that aligns with what might be expected. Features that are surprising could be further explored to check if they are genuine anomalies or if

The top features are latitude (25%), accomodates (22%) and longtitude (19%) accounting for 66% of model importance. Longitude and latitude makes intuitive sense, and it's all about 'Location, location, location!'. 'Accomodates' is the number of people a listing can accommodate which also makes sense as that is a direct influence on pric that an extra sofa-bed would be well worth the investment?

Interestingly, the topics modelled from the listing descriptions have a combined importance of 2.5%. While this seems like a small number, it is higher than property type (i 1.3%. This further validates the hypothesis that there is valuable information in the descriptions and tell you more about a listing than whether it is a house or apartment.

The features importances produced by the model seem to make sense, and further builds confidence in the final model built.

	importance %	feature
7	25.097217	latitude
0	21.785352	accommodates
8	19.507045	longitude
6	7.704549	minimum_nights
5	6.343909	extra_people
1	4.201948	bathrooms
4	3.822254	guests_included
2	2.952059	bedrooms
13	2.080424	room_type_Entire home/apt
3	1.888553	beds
18	1.013714	topics_description_Business
16	0.969459	topics_description_Budget
17	0.641166	topics_description_Luxury
12	0.472587	property_type_Other
11	0.436129	property_type_House
9	0.435698	property_type_Apartment
14	0.397654	room_type_Private room
10	0.154542	property_type_Bed & Breakfast
15	0.095740	room_type_Shared room

Justification

In this section, your model's final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- *Are the final results found stronger than the benchmark result reported earlier?*
- *Have you thoroughly analyzed and discussed the final solution?*
- *Is the final solution significant enough to have solved the problem?*

The final MSE is 200185564 with a variance score of 0.33, which is a significantly better result compared to the baseline model with an MSE of 254474109 and variance of rigorously tested, cross-validated and evaluated with different inputs and been proven to be a robust solution.

This justifies the use of the final model, as it is robust and a more accurate predictor of AirBnb yield compared to a simple linear regression model. It has a lower error rate a yield that is closer to the actual yield.

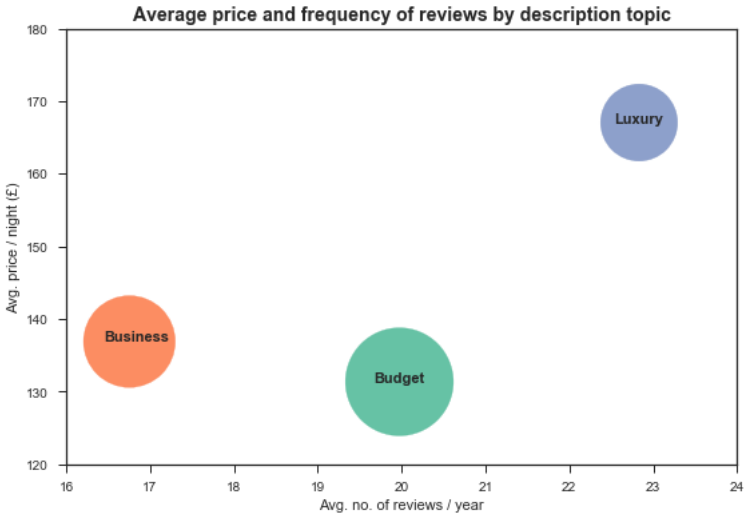
V. Conclusion

(approx. 1-2 pages)

Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonabl characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
- *Is the visualization thoroughly analyzed and discussed?*
- *If a plot is provided, are the axes, title, and datum clearly defined?*



NLP is an important feature of this project, and there are valuable insights produced from topic modelling in addition to improving predictive strength. By plotting a bubble earlier section of this report (Business, Budget and Luxury), it is possible to segment London's Airbnb market into the different topics, each of which demonstrating a diffe are the component factors of yield.

In the below chart, each bubble is plotted at the average of that segment while the size represents the number of listings in that segment. Some of the characteristics of th are slightly more surprising.

In terms of size, the 'Budget' segment has the most number of listings, followed by 'Business' while 'Luxury' has the least number of listings. This makes sense and one w 'Budget' type listings available.

Looking at the average price, 'Luxury' is the clear leader followed by 'Business' and 'Budget' at just £5-10 lower. While the order makes sense, what is surprising is the hu 'Luxury' listings compared to the other 2 segments. This suggests that perhaps AirBnb listings cater to those looking for boutique stays and people in the 'Luxury' segmer experience. On the other hand, business travellers on Airbnb would typically choose Airbnb for its lower price compared to hotels while traditional business travellers with

This trend is also supported by the number of reviews, which is a proxy for the frequency of bookings. There are comparatively very few 'Business' listings, further sugges away from AirBnb. On the other hand, 'Luxury' listings again stand out and prove that the appeal of AirBnb is in the unique experience of staying in a home that is popular quality and luxurious accommodation.

Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You ar whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- *Have you thoroughly summarized the entire process you used for this project?*
- *Were there any interesting aspects of the project?*
- *Were there any difficult aspects of the project?*
- *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

This project has been extremely rewarding to complete, giving me a challenging and interesting problem to explore and design my own solution for. Each stage of the proj points a different part of the analytical toolkit to practice.

In the design phase, the focus was on researching and understanding the problem of predicting yield in AirBnb. What data is publicly available? What research has already assumptions in the problem statement?

This led naturally onto the exploratory data analysis phase once I had gotten my hands on the right data set. To design the right solution, I had to thoroughly understand th understanding what the data points meant, how it was collected, and whether there might have been any unintended biases in the data. This also laid the foundation of ur hypotheses to test and what pre-processing might be necessary.

Iteratively, this also led to the data cleaning and manipulation stage. After understanding the data, I was able to perform the manipulations necessary such as imputing or selection, and calculating yield values.

Once the data has been cleaned and processed, I was able to fit them on the selected machine learning algorithms. The course syllabus was most useful here, and I bega different algorithms and how to interpret the results of each algorithm. By comparing the results I was able to choose a final cross-validated model and fine-tune it accordi

After building the final model, it had to be thoroughly tested and evaluated. This included doing sensitivity analysis the test the robustness of the final solution, and sense produced by the model.

One interesting aspect, and also the most challenging for me, was undertaking natural language processing for the feature engineering stage. I tested various techniques s inverse-document-frequency, latent Dirichlet allocation and also various regexes in the preprocessing stage. Ultimately, LDA proved to be the most powerful method of re able to give more interpretable topics compared to other topic modelling methods.

The unique challenges in topic modelling stemmed from the more qualitive nature of text mining as opposed to working with numerical data. It is a lot more free-form and back to. For example, choosing the number of topics to model greatly affected the results. After testing several topic numbers, I decided to choose a relatively small numb the topics interpretable and also distinct enough to form insightful segments. Labelling the topics was another free-form challenge, though it was thankfully made easier b LDA. Though there is certainly no correct topic label, the word distributions and earlier EDA phase helped to suggest potential topic labels. I tried a few of these labels and they would make sense (e.g. the 'Budget' label should have a proportionaly lower price!). In the end, I was comfortable with the labels I settled on and there was a plausib

Improvement

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your impl and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/c Questions to ask yourself when writing this section:

- *Are there further improvements that could be made on the algorithms or techniques you used in this project?*
- *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
- *If you used your final solution as the new benchmark, do you think an even better solution exists?*

Looking back on the project, there have been numerous learnings and further improvements that could have been made.

As topic modelling was a big part of the project, a potential improvement would be to try and implement other topic modelling algorithms. While I focussed on LDA due to algorithms such as Non-Negative Matrix Factorization that might have also led to good results. I could also have taken a more scientific approach to choosing the number iterate through a range of numbers to choose to optimal number with the highest predictive power.

Ultimately, I am happy with the end results and am convinced that the final solution is better than the benchmark initially established. Nevertheless, there will always be fu

What made this project unique was the attempt to use alternative data sources (in this case, text data) to improve predictive models of AirBnb yield. In a similar vein, AirBr potential features. Attempting this would be quite complicated, requring an understanding of image recognition techniques but also the use of big data processing due to analyse. Nevertheless, that would undoubtedly be an interesting next step to pursue.

Before submitting, ask yourself. . .

- Does the project report you've written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarifica
- Would the intended audience of your project be able to understand your analysis, methods, and results?
- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?
- Is the code that implements your solution easily readable and properly commented?
- Does the code execute without error and produce results similar to those reported?