**Data Science Project Protocol**

*Author(s):*

Daniel Oren

# Introduction

In the following project I’ll try to answer the question:

Can you expect your AirBnb property in London to be occupied most days of a specific month you're interested in?

The prediction result (model outcome) will be binary: yes / no, therefore I’ll use classification models for receiving this type of result.

Furthermore, if indeed I’ll manage to give such a prediction, how accurate will it be...

I’ll use Kaggle London AirBnb dataset for the project. The dataset is pretty good, has many features, but at the same time lack in data which I find very important for such predictions as monthly weather, holidays days per each month, as such features are very likely to impact weather your property will have better chance of being occupied most month,

At the feature selection part we will be able to see which features have a larger effect on the models outcome result

The additions of weather and holidays features might give us interesting insight and better predictions about booking results among customers, as well as we might find that given features in the data affect the booking results much more than we expected, for example: special amenities given in the property such as Coffee machine, Private entrance, working station etc…

A main problem I had to deal with is the large scale dataset (compared to my computer abilities), but I manage to solve it by not using part of the hyperparameters at the tuning process of the models, as well as not using “heavy” models such as SVM classifier.

At the end of the process, I manage to receive a very good predictor for my project question, as can be seen in the conclusion part.

Methodology (Project design)

# Data

Here you have to describe how you plan to manipulate the data. For this you have to answer to the following questions:

### Data that will be used

London Airbnb Data

Data scraped from Airbnb by Inside Airbnb project: The datasets were scraped on November 05th, 2019 and contain detailed listings data, review data and calendar data of current Airbnb listings in London.

### Data sources

<https://www.kaggle.com/labdmitriy/airbnb?select=listings.csv>

### External data sources to enrich my data

* England holidays 2019: <https://www.gov.uk/bank-holidays>
* London monthly average weather 2019: <https://en.climate-data.org/europe/united-kingdom/england/london-1/>

### Data for external validation

Test part of the data (performed before the cross validation)

### Outcome variable

**is\_occupied\_most\_month** : Categorical binary: will the apartment be occupied most month (more than 15 days): 1- yes ; 0 - no

### Confounder variables that may affect the outcome

**occupied\_days\_in\_month** : the target variable will be constructed from it, so obviously will be highly correlated, however will be deleted once the label is created

### Possible source of bias in our data

Not in my data case

### Data exploration strategy

Checking for:

## Target variable analysis

* Quick look with SweetViz
* Variables correlations (numeric - numeric, categorical - categorical)
* Outliers removal or log transformation
* Missing values: imputation transformation into categorical

### Techniques to enrich the data

## Cleaning and filling price related data

* Features transferred into binary
* Impact coding - Bin counting
* Feature Extraction (Distance to Centroid of London, Lodging Size, Lodging Amenities)

### Deal with outliers

Removal or transformation or log transformation:

#### Checking for Distribution and correlation changes before and after outliers removal

* Checking for differences in the distribution: We will compare if there are differences between the distribution of the variable when having the outliers and when the outliers are removed. We do so using the Kolmogorov–Smirnov statistic. This is a non-parametric statistic that can be used on variables without outliers
* The second test we have to make with the outliers is to check if they change the correlation between the variable and the outcome variable. A statistical method for checking if two correlation are significantly different is Fisher r-to-z transformation

Case: none of these happens or only one of them: we will remove the outliers

Case both happens: we will make log transformation on them to reduce the outliers effect

### Deal with missing values

* Checking for differences in the distribution: We will compare if there are differences between the distribution of the variable when having missing + not missing values in front of having the missing values removed. We do so using the Kolmogorov–Smirnov statistic.

Case changed: we can’t perform imputation: will be transformed into categorical

Case not changed: we can do imputation: make imputation with fancyimpute MICE

# 

# Models

How I plan to develop my models.

### Pres teps

1. Converting categorical variable into dummy / indicator variables
2. Feature Selection - select what will be the features I’ll use for training and creating the models. That will be done with the help of the following algorithms:

* Univariate Analysis: Filter based
  + Correlation (Spearman)
  + Chi-Square
* Multivariable Analysis:
  + Wrapper-based
    - Recursive Feature Elimination
  + Embedded
    - LASSO (L1 penalization)
    - Random Forest
    - Gradient Boosting
    - SVM

### Dividing the data

Training, validation, test - proportions, techniques:

Splitting and Scaling the Data into:

* Train set = 70%
* Test set = 30%

The Train set will be splitted by **Cross Validation** (StratifiedKFold to keep proportion of features and labels in each iteration), K-fold = 4 times (dealing with large dataset)

### Balance data

The data has no special need for balancing, will only make sure during the cross validation to use the StratifiedKFold.

### Stratify/subsample the data

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html>

StratifiedKFold:

Stratified K-Folds cross-validator.

Provides train/test indices to split data in train/test sets.

This cross-validation object is a variation of KFold that returns stratified folds. The folds are made by preserving the percentage of samples for each class.

### Techniques applied to get model outcome

Classification

* cross-validation usage + plotting learning curves as they are a good way to see the **overfitting** effect on the training set and the effect of the training size on the accuracy

### Measures use to train and evaluate your models

ROC AUC score, as it’s better for data that its label is not best balanced (in my case it’s not badly balanced but not close to 50% - 50% as well....)

* Usage of ensembling for getting best results: I chose a voting classifier to combine the predictions coming from the 2 classifiers. I preferred to pass the argument "soft" to the voting parameter to take into account the probability of each vote.

# Deployment of your model

### QA of the project

At this project no QA will be made

### Final user of the predictions

Company or private person that would like to see the possibility of his apartment being occupied most month, based on different features

### Prediction presented to the final user

Binary: 1 - occupied most month ; 0 - not occupied most month

### Final user way to be trained to use and interpret the prediction

The user will insert the features corresponding the model and will gain an insight if he can expect the apartment to be occupied most month or not

### On which platform the predictions will be deployed

The model will not be deployed to production

### Model updates frequent

Once a year according the new booking data

### What will happen in cases where the model return a null prediction (eg. incomplete data)

The data will need to be checked, if missing or outliers values gives these results

### Measurements used to evaluate if the prediction is decaying

ROC AUC score starts to reduce

Results

Presenting the main results of all the processes.

### The final amount of data used (total, train, test, etc)

The dataset has 1020816 rows and 113 columns.

After features selection:

The dataset has 1020600 rows and 48 columns.

Test = 30%

Train = 70%

* Cross validation (K-Fold = 4): each iteration: train = 75% ; dev = 25%

### The amount of outliers and the way of treating them

* Variables number with outliers: 40
* Total outliers number: 6,029,064

Used methods for treating outliers:

Removal or transformation or log transformation:

#### Checking for Distribution and correlation changes before and after outliers removal

* Checking for differences in the distribution: using the Kolmogorov–Smirnov statistic.
* Check if the correlation changed between the variable with outliers and without outliers: using Fisher r-to-z transformation

Case:

* None of these happens or only one of them: we will remove the outliers
* Both happens: we will make log transformation on them to reduce the outliers effect

### The amount of missing values and the methods used for imputing them

* Variables number with outliers: 60
* Total outliers number: 21,564,828

Checking for differences in the distribution: using the Kolmogorov–Smirnov statistic.

Case:

* Changed: we can’t perform imputation: will be transformed into categorical
* Not changed: we can do imputation: make imputation with fancyimpute MICE

### The methods used to transform the data and to generate new features

**Data transformed**

Transforming numeric data with outliers that couldn’t be removed using log

### **Generate new features**

Feature Extraction

* Distance to Centroid of London - Location is always an important factor in lodging services. To make it more descriptive, I decided to calculate each accommodation's distance to the so-called centroid of London instead of just relying on the neighbourhoods or areas.
* Lodging Size - An important piece of information that might help predicting occupation during the month is size. Since the column square\_feet was heavily filled with null values, I dropped it. I’ll use the column description to reveals any information about size instead
* Lodging Amenities - I'm interested in what amenities hosts offer their guests, and in order to enrich our prediction, whether we can determine what some of the more special and/or rare amenities might be that make a property more desirable

Conclusion

As I’m working for a british travel industry company (Culture Trip), I was thinking of the AirBnb dataset as interesting data to create a project on.

Two of the biggest challenges I had to deal with were especially from 2 aspects data cleaning and large scale data:

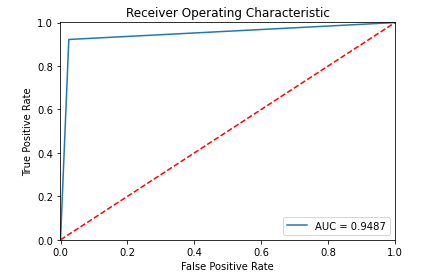
* Cleaning: many variables had outliers which couldn’t be removed, and required more special treatment, as well as variables with missing values that I couldn’t perform imputation on, and also required more special treatments (as described earlier in this document).
* Large scale data: the large scale of the data made some of the processes run for very long time, somthomes so long I had to remove part of the process, for example at the hyperparameter tuning: the trees hyperparameters tuning were so long (more than 1 day) that I tune only the depth parameter. Another example was the SVC classifier that I didn’t use because of the same reason.

Note: In addition I had some challenges with the data itself: low correlation between variables, and low voting in the feature selection part. I deal with both by reducing the cutoff value:

* Correlations: I looked for lower correlation values
* Feature selection: I reduced the voting cutoff to 2 (usually we look for voting result >= 4)
* P-value: as default P-Value for this data set of 0.05 tends to be too high (make easier variable to be missing NOT at random for example), I'll lowered it to 0.001 value, so it will be "less sensitive"

My final prediction has been done on the X-test data, and then comparing the result to y\_test, checking the AUC ROC result:

I received AUC = 0.9487



The model is limited to the London area, and as based on 2019 data, therefore might be needed to get updated once a year, to fit the updated annual data.

The model might be also sensitive to dramatic changes, as for example dramatic climate changes or dramatic changes in property booking due to unseen reasons...