UNIVERSITÀ
DEGLI STUDI
DI MILANO

LA STATALE

Department od Economics, Management and Quantitative Methods

Università degli Studi di Milano

Data science and economics

Statistical Learning Project

Analysis and prediction models for New York City Airbnb

Supervised Learning Report

Author:

Andrea Ierardi

Academic year 2019-2020

Contents

1

Abstract

The aim of the project is to analyse and develop prediction models for the data from the ”New York City Airbnb Open Data” from a Kaggle competition.
In particular, one part is focused on the developing predictive model to forecast house prices using Supervised Learning technics:

- Linear Regression
- Decision Tree
- Random Forest
- Ranger Random Forest
- Neural Newtworks (?)

The second part is focused on the clusterisation of the data using Unsupervised Learning technics: Cluster Analysis and K-means algorithm.

- Cluster Analysis
- K-means Algorithm

A crucial part of the project was to the tuning and finding of the hyperparameters of the different models in order to get the best fit.

2

Problem Definition and Algorithm

2.1 The main Goal

The project is focused on a AirBnB user point of view. It is possible to image an application in which the user has a lot of information depending on the objective. For a landlord point of view, giving the information about the longitude and the latitude of the house, he/she will have as output the estimated price given based also on the neighbourhood group in New York City. For a guest point of view, he/she will give in input information about the price range and the type of room to get as output the most suitable houses for him or predict the position of one of that.
For the Unsupervised part, the goal is to define cluster of price in which is possible to split out the houses, firstly in the entire city, secondly filtering for Neighborhood or room type.

2.2 Algorithms

2.2.1 Linear Regression

Linear regression is a linear approach to modeling the relationship between a dependent variable and one or more independent variables. Linear regression should be suitable, since there could be a linear relationship between the position and the price. In the city center there will be the expensive houses, while in the outskirts there will be the cheaper ones.

2.2.2 Decison Trees

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making.

2.2.3 Random Forest

Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees.

2.2.4 Ranger Random Forest

Ranger is a fast implementation of random forests or recursive partitioning, particularly suited for high dimensional data.

3

Experimental Evaluation

3.1 Methodology

3.1.1 Data exploration and visualisation

3.1.2 Data cleaning and pre-processing

3.2 Results

3.3 Discussion

4

Conclusion

The dataset used in this project is one of a Kaggle competiotion and is called the New York City Airbnb Open Data. It contains 48.000 data points for each different column. The dataset has different columns:

- id
- **name**: name of the listing
- **host_id**
- **host_name**
- **neighbourhood_group**: location

- **neighbourhood**: area
- **latitude**: coordinates
- **longitude**: coordinates
- **room_type**: space type
- **price**: in dollars
- **minimum_nights**: amount of nights minimum
- **number_of_reviews**: number of reviews
- **last_review**: latest review
- **reviews_per_month**: number of reviews per month
- **calculated_host_listings_count**: amount of listing per host
- **availability_365**: number of days when listing is available for booking

We select just 5 of this feature from the dataset since we denotes them as the most important for the price of a house: latitude, longitude, room type, neighbourhood and the price itself to compare the prediction during the tests.

```
     price
Min.   :    0.0
1st Qu.:   69.0
Median :  106.0
Mean   :  152.7
3rd Qu.:  175.0
Max.   :10000.0
```

Figure 1: Price summary

From the Figure 1 is possible to see that there are some outliers in the dataset that can not be removed, since we have to take in account that could be present luxury houses. In fact, no outlier was removed from the dataset. Also, there were no null and missing value apart from the reviews_per_month column that we do not take in account in the project.

4.1 Data Pre-processing

The dataset has more or less 48.000 data points for each column, so an important part of this work was the pre-processing since it is large. Since the dataset is very large, running the different methods is time consuming. Scaling the numerical data is a key point to get better performance during the fitting process of the model. We starting scaling numerical data from -1 to 1, using pre installed R functions while for the categorical data we identify each label with a different integer number.
The categorical features from the selected are: neighbourhood and room type.
The numerical features are: latitude, longitude and price.

4.2 Decision Tree

Decision tree are one of the most used model in the Machine Learning world since are very familiar to human users and can be easily plotted.

4.3 Random Forest

Random Forest is an ensemble method which use a combination of decision tree to get the prediction.

4.3.1 Ranger Random Forest

Ranger Random Forest is a computationally light model which results are very close the classical Random Forest.

4.4 Linear Regression

5

Results

6

Conclusion

From the result the method with the most higher accuracy is the Random Forest method... while the worst are ....
Moreover, Random Forest method is also the worst in term of computation time for the tuning part since it takes for a configuration with 4 core, more or less 1 hour to tune the parameters.

7

Appendix

8

Tabelle e grafici

Here a few examples of tables and graphs.

8.1 Tabelle

Codice	*CdL*	*Lotto*	$T_{setup/lotto}$	$T_{lav/pezzo}$	$T_{proc/pezzo}$	*Quantità*	T_{tot}
100	4	250	25	0,5	0,6	1	0,6
111	2	250	20	2	2,08	1	2,08
111	3	250	15	1,5	1,56	1	1,56
112	2	250	20	2,5	2,58	1	2,58
112	3	250	15	2	2,06	1	2,06
113	3	500	15	1	1,03	2	2,06
120	1	50	30	2	2,6	0,1	0,26
121	1	25	30	3	4,2	0,1	0,42
121	1	25	30	2,5	3,7	0,1	0,37

8.1.1 Altra tabella

8.2 Grafici

9

Altro

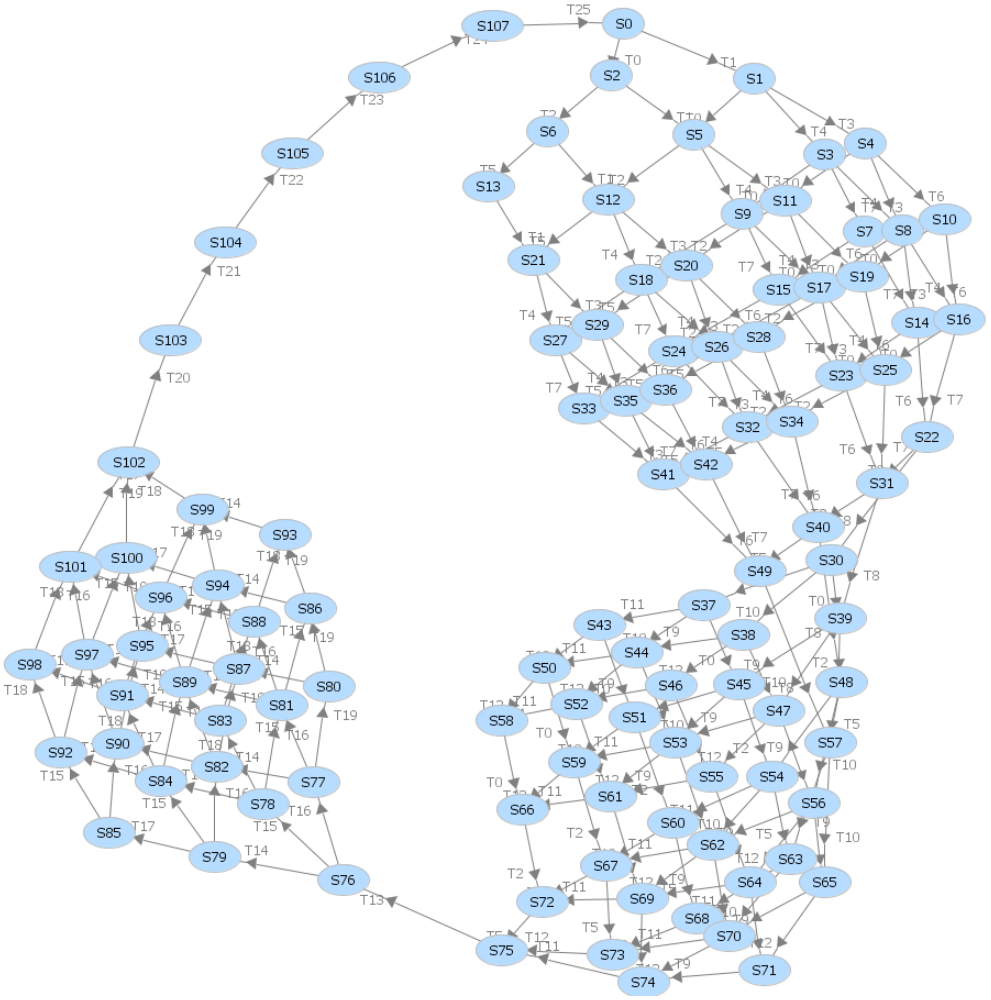

Figure 2: Didascalia.

9.1 Footnote

You can create a footnote like this.[1]

[1] I created a footnote.

References

[1] Giusti, Santochi, *Tecnologia Meccanica e Studi di Fabbricazione*. Casa Editrice Ambrosiana, Seconda Edizione

[2] Mechteacher, *Knuckle Joint – Introduction, Parts and Applications*, http://mechteacher.com/knuckle-joint/

[3] Totalmateria, *G32NiCrMo8*, http://www.totalmateria.com

[4] Sandvik Coromant,*Catalogo generale 2018*, http://www.coromant.sandvik.com/it

[5] Norme UNI, Ente nazionale italiano di unificazione