Automatic classification of accommodations based on UGC through Support Vector Machine

Dra. Eva Martín Fuentes - eva.martin@udl.cat
Dr. Cesar Fernandez
Dr. Carles Mateu
Dra. Estela Mariné Roig

Universitat de Lleida

Universitat de Lleida

LLEIDA TECH
ARTIFICIAL INTELLIGENCE & OPTIMIZATION CONGRESS

International Journal of Hospitality Management 69 (2018) 75–83

Contents lists available at ScienceDirect

International Journal of Hospitality Management

journal homepage: www.elsevier.com/locate/ijhm

Discussion paper

Modelling a grading scheme for peer-to-peer accommodation: Stars for Airbnb

Eva Martin-Fuentes[a,*], Cesar Fernandez[b], Carles Mateu[b], Estela Marine-Roig[a]

[a] Department of Business Administration, University of Lleida, C/Jaume II, 73, 25001, LLEIDA, Spain

[b] INSPIRES Research Institute, University of Lleida, C/Jaume II, 69, 25001, LLEIDA, Spain

ARTICLE INFO

Keywords:
Airbnb
Hotel classification system
Support vector machine
Big data
Peer-to-peer accommodation platform

ABSTRACT

This study aims, firstly, to determine whether hotel categories worldwide can be inferred from features that are not taken into account by the institutions in charge of assigning such categories and, if so, to create a model to classify the properties offered by P2P accommodation platforms, similar to grading scheme categories for hotels, thus preventing opportunistic behaviours of information asymmetry and information overload. The characteristics of 33,000 hotels around the world and 18,000,000 reviews from Booking.com were collected automatically and, using the Support Vector Machine classification technique, we trained a model to assign a category to a given hotel. The results suggest that a hotel classification can usually be inferred by different criteria (number of reviews, price, score, and users' wish lists) that have nothing to do with the official criteria. Moreover, room prices are the most important feature for predicting the hotel category, followed by cleanliness and location.

Martin-Fuentes, E., Fernandez, C., Mateu, C., & Marine-Roig, E. (2018). Modelling a grading scheme for peer-to-peer accommodation: Stars for Airbnb. *International Journal of Hospitality Management*, 69, 75-83.

Research aim

- To predict the hotel category through UGC.
- To create a model that would allow the properties offered by P2P accommodation platforms to be classified.

Introduction

- Social Media contribution to the rise of User-Generated Content (UGC)
- Valuable source of information:
  - Other travellers
  - Hotel
  - Researchers

(Buhalis & Law, 2008; Kwok et al, 2017; Cantallops & Salvi, 2014)

Introduction

- Huge amount of reviews for the same product.
- Positive and negative.
- Information overload complicates the decision-making process.

(De Ascaniis & Gretzel, 2012; O' Connor, 2010, Fang et al, 2016; Marine-Roig, 2017)

Introduction

- To simplify this process, in the case of the hospitality industry, categories can be a filter.

(Manes & Tchetchik, 2018)

Literature

- The hotel rating mechanism does not follow the same pattern in the world.
- Official and unofficial systems.

(Zhan-Qing & Liu, 1993)

Literature

- Most common stars from 1 to 5.
- Others: Diamonds, crowns, characters, etc.

Literature

- The Hotrec Association (Hotels, Restaurants & Cafes in Europe) is trying to unify the criteria for assigning categories in different European countries.
- It's not easy.

(Hotrec, 2015)

Literature

- Consumers sometimes use categories to choose hotels.
- Useful in reducing the adverse effects of information asymmetry.
- It occurs when one of the two parts of the buying and selling process does not have the same information as the other.

(Akerlof, 1970; Fang et al, 2016; Nicolau & Sellers, 2010; Öğüt & Onur Tas, 2012)

Literature

- With the Internet, the accommodation sector is undergoing a revolution.
- With sharing economy platforms acting as intermediaries: Couchsurfing, HomeExchange, Airbnb, HomeAway.

(Cheng, 2016; Sigala, 2015)

Literature

- Barrier to using collaborative hosting platforms: lack of trust.
- Lack of trust is related to information asymmetry.

(Tussyadiah & Pesonen, 2016)

Literature

- The current hotel classification system does not take UGC into account.
- Interest in bringing reviews and hotel stars closer together.

(Blomberg-Nygard and Anderson, 2016)

Methodology

Booking.com Data

Region	Countries	Destinations	Hotels	Reviews
EUR	17	168	14,395	11,097,703
AME	23	122	7,022	3,285,925
ASP	16	109	10,448	3,559,306
MEA	9	44	1,179	767,947
Total	65	443	33,044	18,710,881

Methodology

Data collection:

Automatic

Webscraping Technique

Python

Simulates a user's navigation (clicks and selections)

Methodology

Booking.com	Airbnb
Value for money	Value for money
Cleanliness	Cleanliness
Location	Location
Services	Check-in
Comfort	Communication
Staff	Precision
Price	Price
Number of reviews	Number of reviews
Wishlist	Wishlist

Methodology

Booking.com	Airbnb	
Value for money	Value for money	✓
Cleanliness	Cleanliness	✓
Location	Location	✓
Services	Check-in	✗
Comfort	Communication	✗
Staff	Precision	✗
Price	Price	✓
Number of reviews	Number of reviews	✓
Wishlist	Wishlist	✓

Methodology

Support Vector Machine (SVM)

A dataset is often not linearly separable.

We use Radial Basis Function kernels.

Methodology

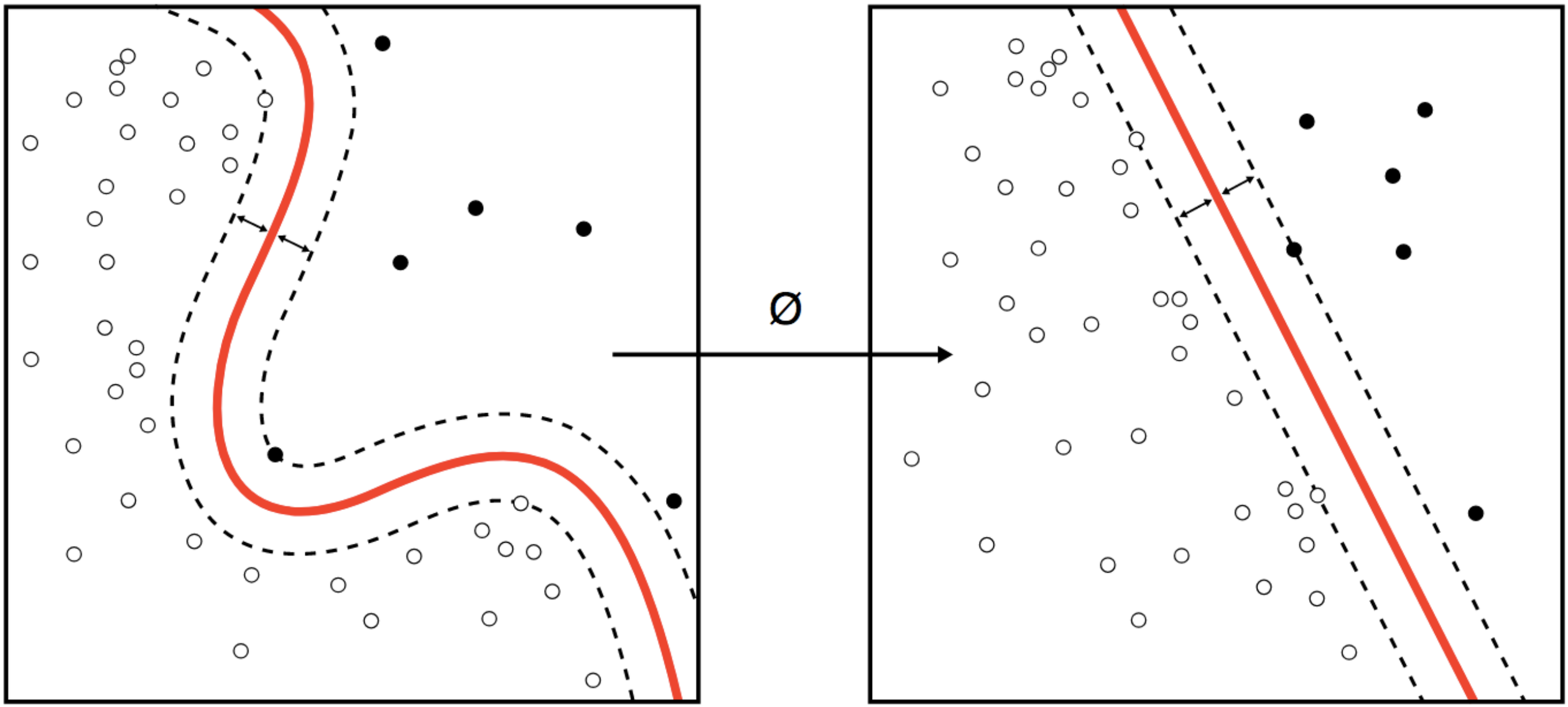

Imagen: Wikipedia

Methodology

- Supervised machine learning system.
- Create a model with a sample of data (training).
- Train a model to assign a category to any hotel.

Methodology

- Assigns a vote to the class that has been assigned in the verification phase.
- It is then checked against the rest of the data.
- Accuracy: ratio of correctly ranked hotels.

(Hsu & Lin, 2002)

Results

AME

Category	Test (n)	Accuracy SVM	Ratio SVM
1	500	0	0.00
2	1000	686	0.69
3	1000	442	0.44
4	1000	256	0.26
5	1000	876	0.88

Results

ASP

Category	Test (n)	Accuracy SVM	Ratio SVM
1	1000	28	0.03
2	1000	741	0.74
3	1000	363	0.36
4	1000	314	0.31
5	1000	841	0.84

Results

EUR

Category	Test (n)	Accuracy SVM	Ratio SVM
1	1000	12	0.01
2	1000	729	0.73
3	1000	452	0.45
4	1000	371	0.37
5	1000	837	0.84

Results

MEA

Category	Test (n)	Accuracy SVM	Ratio SVM
1	200	74	0.37
2	200	2	0.01
3	200	151	0.76
4	200	119	0.60
5	200	153	0.77

Results

- Budget (1 & 2* hotels)
- Lower-middle range (3* hotels)
- Upper-middle range (4* hotels)
- Superior (5* hotels)

Results

AME

Category	Test (n)	Accuracy SVM	Ratio SVM
Budget	1500	1,113	0.74
Mid-Low Range	2000	1,642	0.82
Mid-High Range	2000	1,095	0.55
Superior	2000	1,658	0.83

Results

ASP

Category	Test (n)	Accuracy SVM	Ratio SVM
Budget	2000	1,660	0.83
Mid-Low Range	2000	1,657	0.83
Mid-High Range	2000	1,092	0.83
Superior	2000	1,686	0.84

Results

EUR

Category	Test (n)	Accuracy SVM	Ratio SVM
Budget	2000	1,623	0.81
Mid-Low Range	2000	1,716	0.86
Mid-High Range	2000	1,196	0.60
Superior	2000	1,724	0.86

Results

MEA

Category	Test (n)	Accuracy SVM	Ratio SVM
Budget	400	110	0.28
Mid-Low Range	400	308	0.77
Mid-High Range	400	346	0.87
Superior	400	355	0.89

Discussion

- Hotel ranking can be inferred quite accurately through other criteria such as reviews scores.
- By predicting it, old-fashioned criteria can be avoided.

(Torres et al., 2014)

Discussion

- The highest accuracy is obtained with 5* hotels. This category has some uniformity at the international level.
- The lowest in 1* hotels (AME, ASP and EUR), and 2* (MEA). Users do not perceive differences between these categories.

(Minazzi, 2010)

Discussion

- Trust is vital on these platforms.
- General, comprehensive information.
- Not just for Airbnb.

(Gretzel and Yoo, 2008; Vermeulen and Seegers, 2009)

Conclusions

- User's point of view matches.
- Different systems converge.
- It validates the official classification system: fewer audits and bureaucracy.
- Problem: Properties with no reviews.

Conclusions

- This validation allows for the creation of an international classification system.
- Any type of accommodation.
- Comparison tool.
- Any product or service rated and rated by users.

Automatic classification of accommodations based on UGC through Support Vector Machine

Thank you

Universitat de Lleida

LLEIDA TECH
ARTIFICIAL INTELLIGENCE & OPTIMIZATION CONGRESS