UNIVERSITÀ
DELLA CALABRIA

Discovering Travelers' Purchasing Behavior from Public Transport Data

Francesco Branda, *Fabrizio Marozzo*, *Domenico Talia*

DIMES, University of Calabria, Italy

The Sixth International Conference on Machine Learning, Optimization, and Data Science (LOD 2020)

Certosa di Pontignano, Siena – Tuscany (Italy)

July 19-23, 2020

Motivations and goal

- In recent years, the demand for **collective mobility services** is characterized by a **significant growth.**

- The long-distance coach market has undergone an important change in Europe since **FlixBus** provided **low-cost bus services** with a **dynamic pricing strategy** and an **efficient** and **fast information system**.

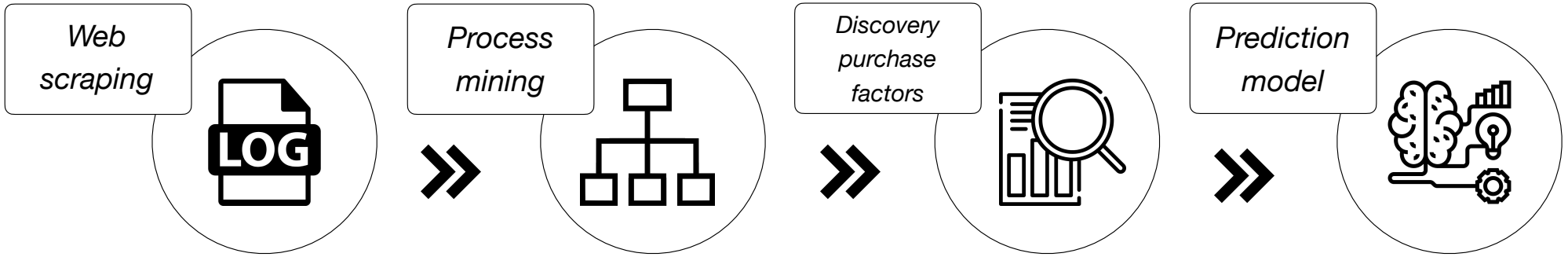
- This paper presents a **methodology**, called **DA4PT** (**Data Analytics for Public Transport**), aimed at discovering **the factors** that **influence travelers** in **booking** and **purchasing** a **bus ticket**.

Outline

- Proposed methodology (DA4PT)
- Case study
- Evaluations
- Conclusions

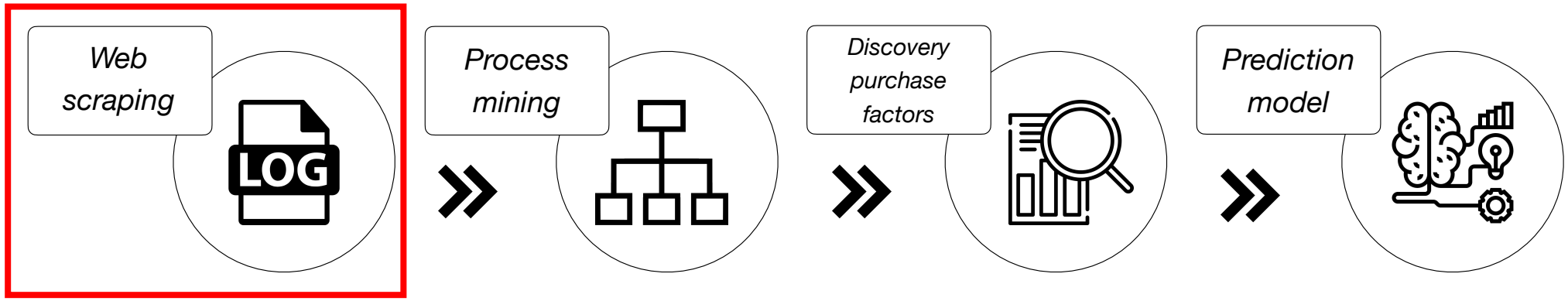
Proposed methodology (DA4PT)

- The **methodology** consists of four steps:
  1. Web scraping
  2. Process Mining
  3. Discovery purchase factors
  4. Prediction model



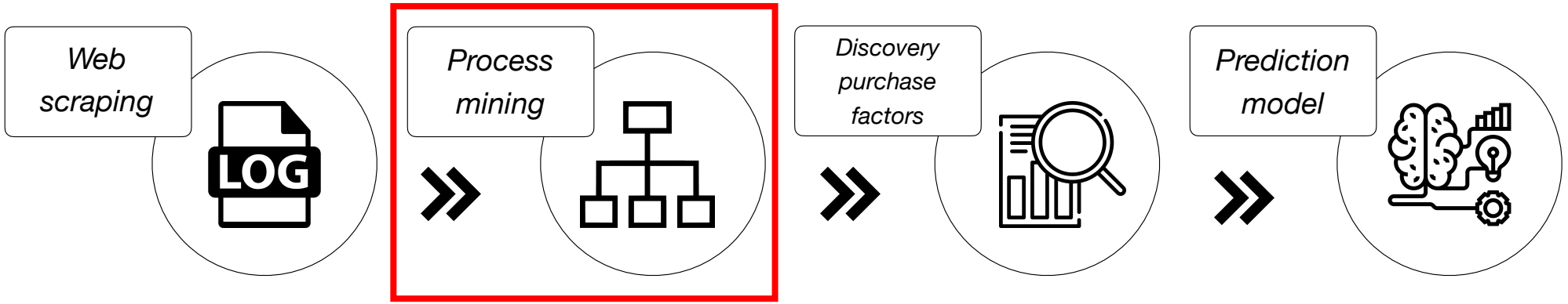
Proposed methodology – Step 1/4

1. **Web scraping**: Data collection is carried out by using *Web scraping techniques* to capture the **interactions** of users with a **bus booking platform** (e.g., whether a user *buys or not a ticket*, or in which step of the buying decision process she/he *leaves the platform*).



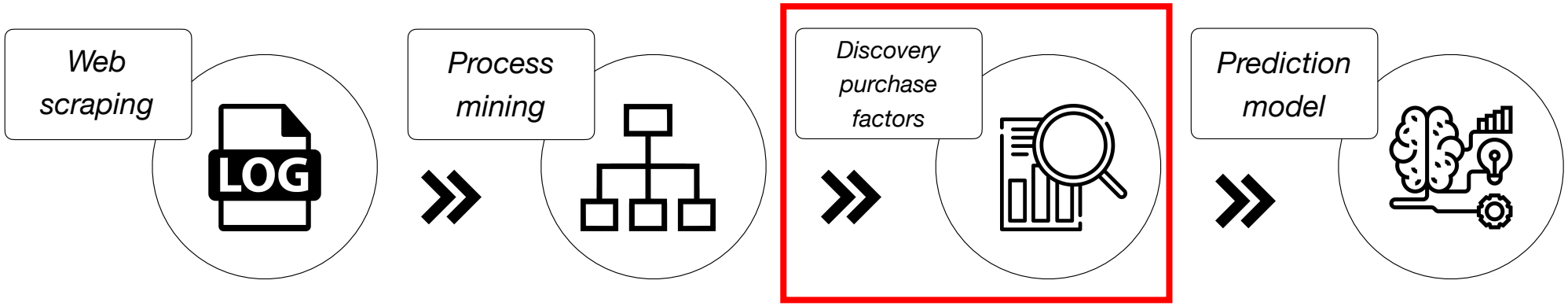
Proposed methodology – Step 2/4

2. **Process Mining** algorithms is applied with the aim of **identifying trends** and **human patterns**, and **understanding behaviours** of users while searching and booking bus trips.



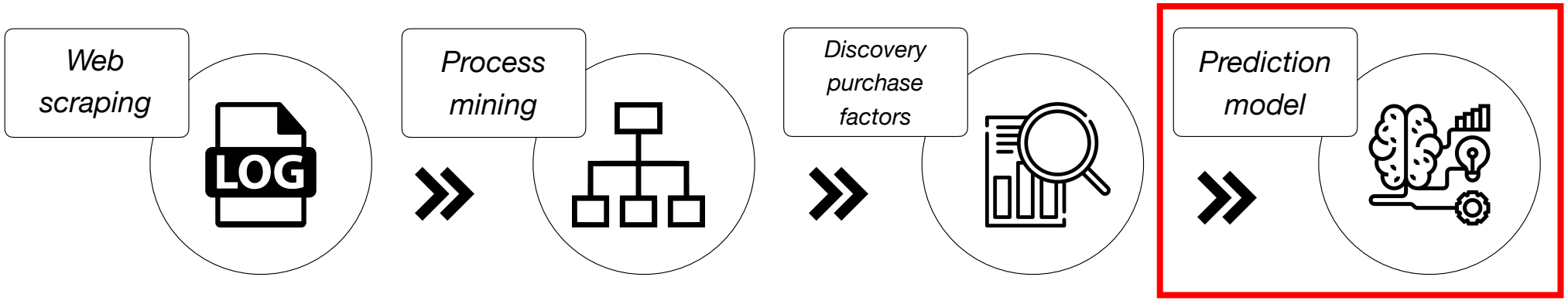
Proposed methodology – Step 3/4

3. **Discovery purchase factors**, i.e. **identify** the **key factors** that push a user to **buy a ticket**. The correlations between an attribute and the class attribute (purchased or abandoned) is evaluated using the Pearson's correlation coefficient.



Proposed methodology – Step 4/4

- **Prediction model**: A **model** capable of automatically **learning** whether or not a user will **finalize a purchase**. In particular, the model has been trained on information that depends on the route, departure date and date of booking (e.g., ticket fare, occupancy rate of a bus).



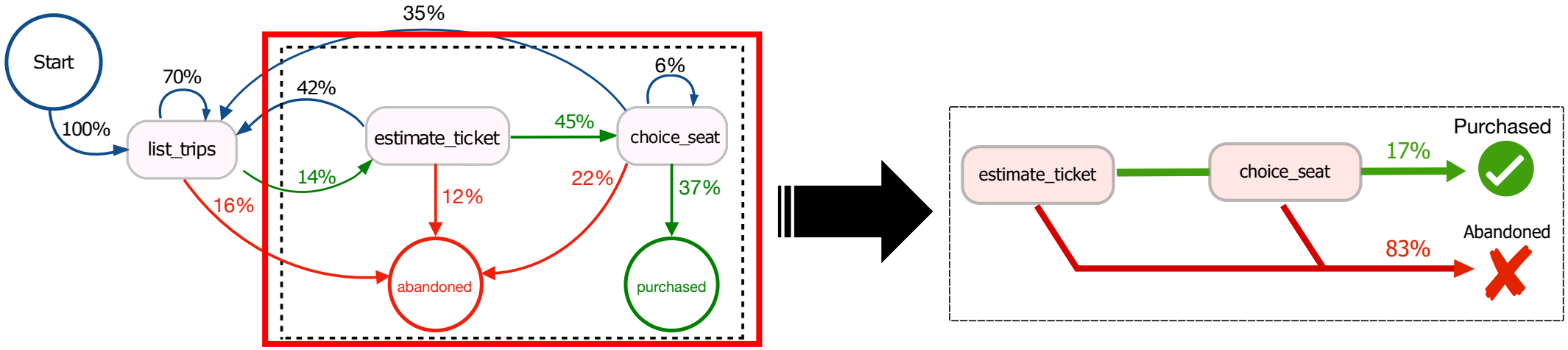
Case study (Web scraping)

- The proposed methodology has been applied on a **dataset** composed by **3.23 million event** logs of an **Italian bus ticketing platform**, collected from August 2018 to October 2019.
- A user interacts with the platform generating **4 types of events**:
  - **list_trips**, to find the routes between an origin and destination locations;
  - **estimate_ticket**, to determine the itinerary cost on the basis of the route select by user;
  - **choice_seat**, to find available seats on the bus chosen;

COOKIE	ACTION	TIMESTAMP	TRIP ID	DEPARTURE DATE	BOOKING DATE	ORIGIN CITY	DESTINATION CITY	No. SEAT	BUS SEAT	FARE	BOUGHT
1JYASX	list_trips	2018-10-16 11:31:19		2018-10-22		Soverato	Rome				
1JYASX	estimate_ticket	2018-10-16 11:31:37	141772	2018-10-22		Soverato	Rome	1	45	35 €	
1JYASX	choice_seat	2018-10-16 11:36:28	141772	2018-10-22		Soverato	Rome	1	45	35 €	
1JYASX	purchased_ticket	2018-10-16 11:42:20	141772	2018-10-22	2018-10-16	Soverato	Rome	1	45	35 €	YES
28UAKS	list_trips	2019-02-24 18:15:07		2019-02-26		Milan	Lamezia Terme				
28UAKS	estimate_ticket	2019-02-24 18:15:40	408003	2019-02-26		Milan	Lamezia Terme	2	52	64 €	
28UAKS	choice_seat	2019-02-24 18:20:05	408003	2019-02-26		Milan	Lamezia Terme	2	52	64 €	NO

Case study (Process mining)

- The figure shows the **navigation paths** corresponding to those produced by users on the bus ticketing platform. The **green paths** end with the purchase of a ticket (**purchased**), while the **red paths** end with the **abandonment** of the platform (abandoned).
- We focused on all the events generated users after they have chosen a route (**estimate ticket**). In this range, only **17%** of users **purchase a ticket**, while **83% abandon the platform** without buying.

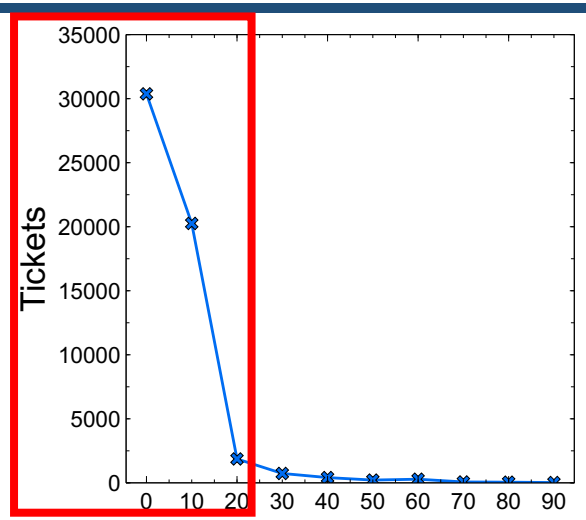


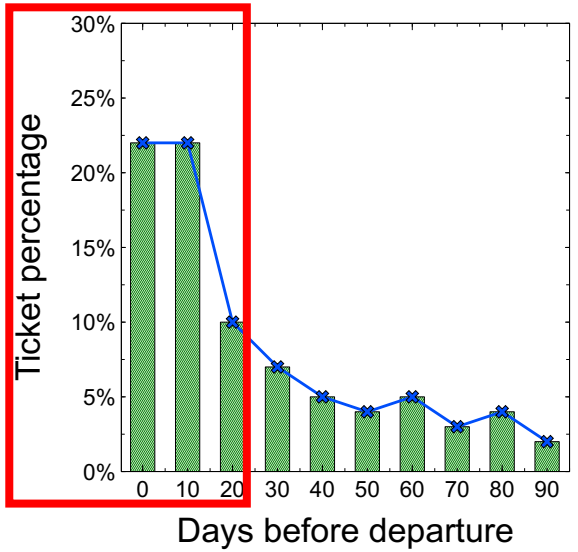
Case study (Discovery purchase factors)

- The goal is to study the **correlations** between a **purchase factor** and the **class attribute** (**purchased** or **abandoned**).
- We focused our attention on four purchase factors:
  1. **Days before departure** (**DBD**), by calculating the difference between booking and departure date;
  2. **Booking day of the week** (**BDOW**), by extracting the day from a booking date;
  3. **Occupancy rate for a bus** (**OCCR**), by evaluating the number of required bus seats per passenger;
  4. **Fare of a ticket** (**HMLF**), by dividing the price of each trip itinerary into three bands (high, medium, and low).
- We measure the **numbers** and the **percentage of purchased tickets** and the **correlation** on the basis of these factors.

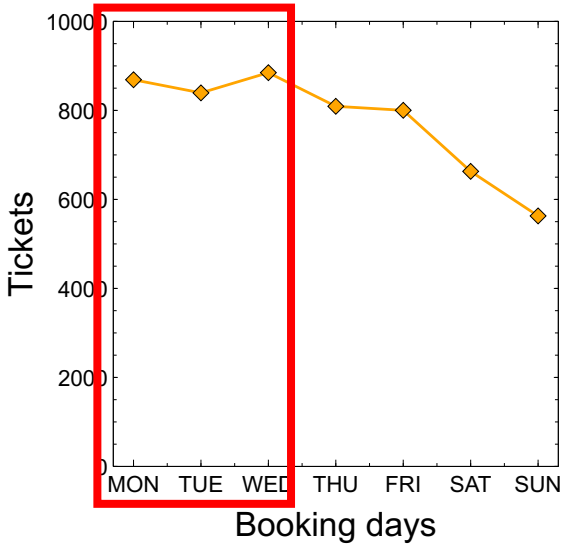
Case study (Discovery purchase factors)

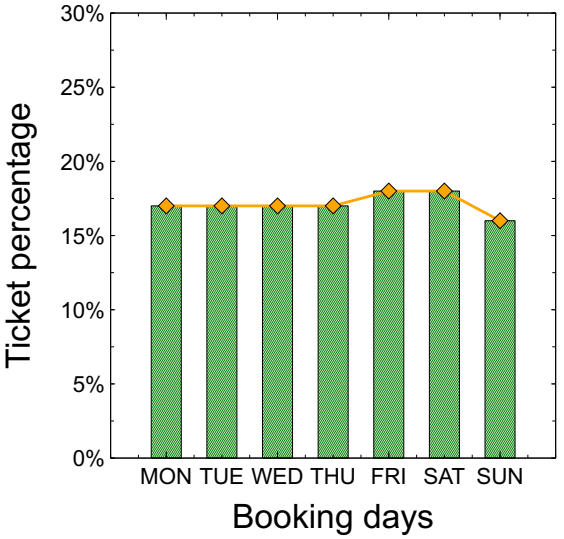
- **Days before departure** (DBD): *few days* before departure, users buy more frequently.





- **Booking day of the week** (BDOW): In the first three days of the week (MON, TUE, WED) most tickets are sold, while in the other days the number of tickets sold drops drastically.

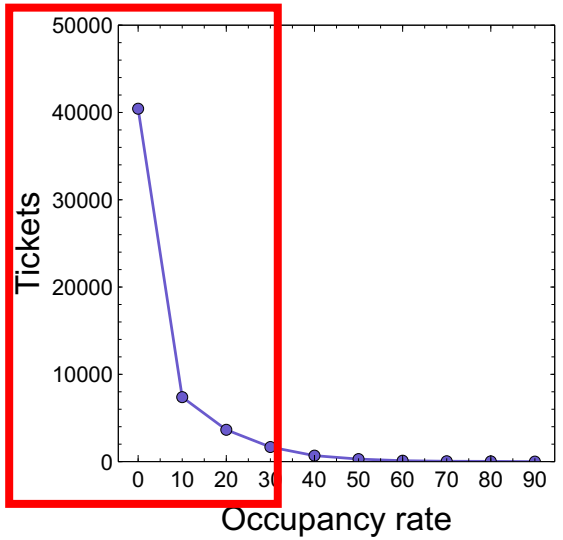


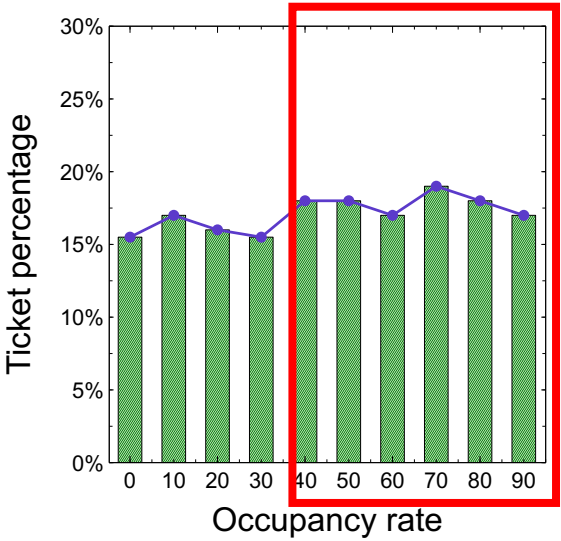


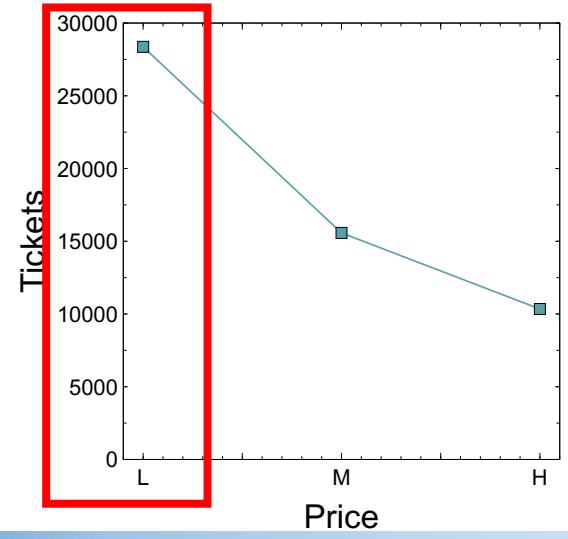
Case study (Discovery purchase factors)

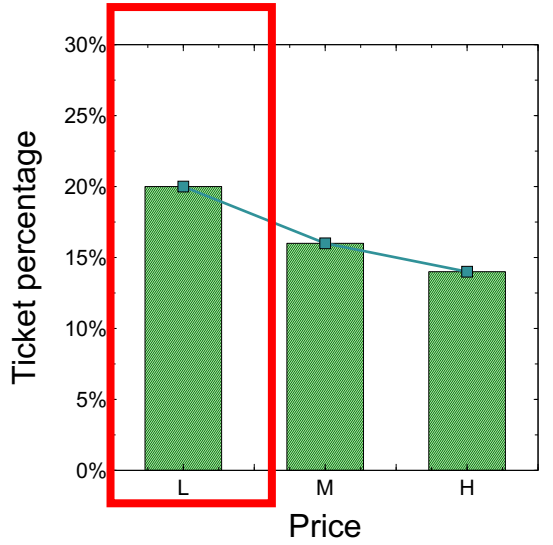
- **Occupancy rate for a bus** (**OCCR**): the tickets are mostly bought when the percentage of available seats is between 10% and 30%, whereas the probability of purchasing a ticket lightly increases when the bus seats are running out.

- **Fare of a ticket** (**HMLF**): most users are pushed to buy a ticket when the price is low.



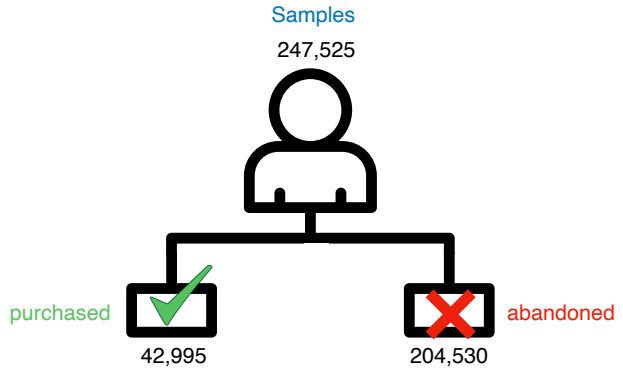






Case study (Prediction model)

- The goal is to define a **model** capable of automatically **learning** whether or not a user will **finalize a purchase**.
- Before running the learning algorithms, we used a random **under-sampling algorithm** to balance **class distribution**.



- The performance of the machine learning models has been evaluated through a **confusion matrix**.

	Purchased (predicted)	Abandoned (predicted)
Purchased (actual)	True Positive (TP)	False Negative (FN)
Abandoned (actual)	False Positive (FP)	True Negative (TN)

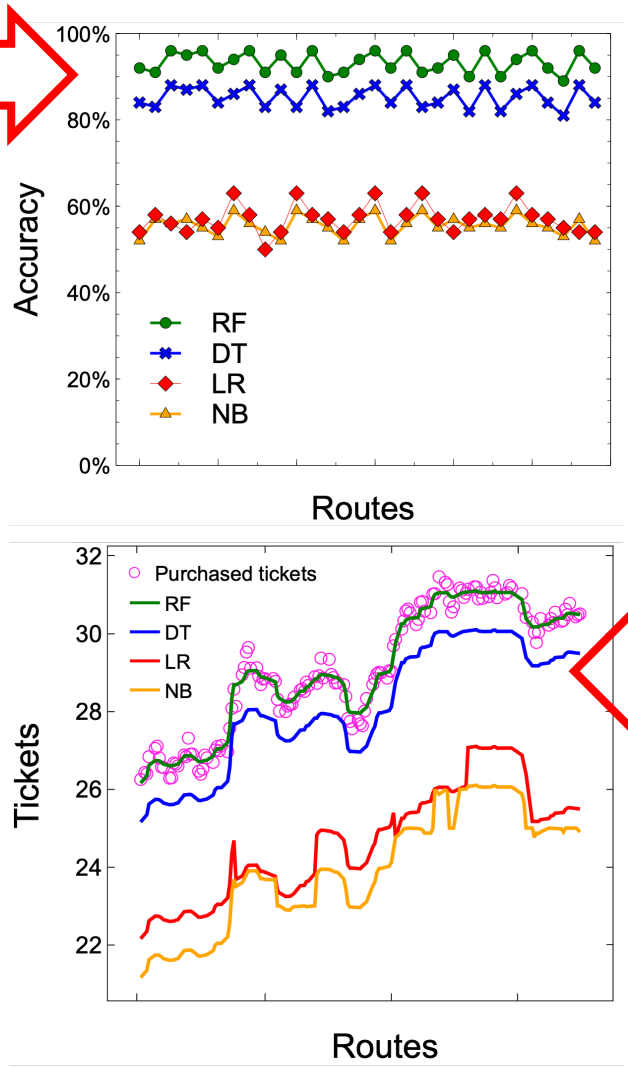
- For each algorithm, we evaluated the *purchased* recall (*Rp*) and *abandoned* recall (*Ra*) to measure the quality of a classifier with respect to a given class.

Case study (Prediction model)

- The table summarizes the results obtained by the four machine learning algorithms we used. Specifically, Random Forest proved to be the best classification model with R=0.93.

- The accuracy of Random Forest stably ranging from 0.91 to 0.96, followed by Decision Tree (0.81-0.88), Logistic Regression (0.50-0.63), and Naive Bayes (0.52-0.59). Also the number of tickets is correctly predicted by Random Forest.

Algorithms	Accuracy	Precision	Recall	F1-score
Naïve Bayes	0.615	0.644	0.615	0.595
Logistic Regression	0.615	0.616	0.615	0.615
Decision Tree	0.864	0.865	0.864	0.864
Random Forest	0.930	0.928	0.930	0.928

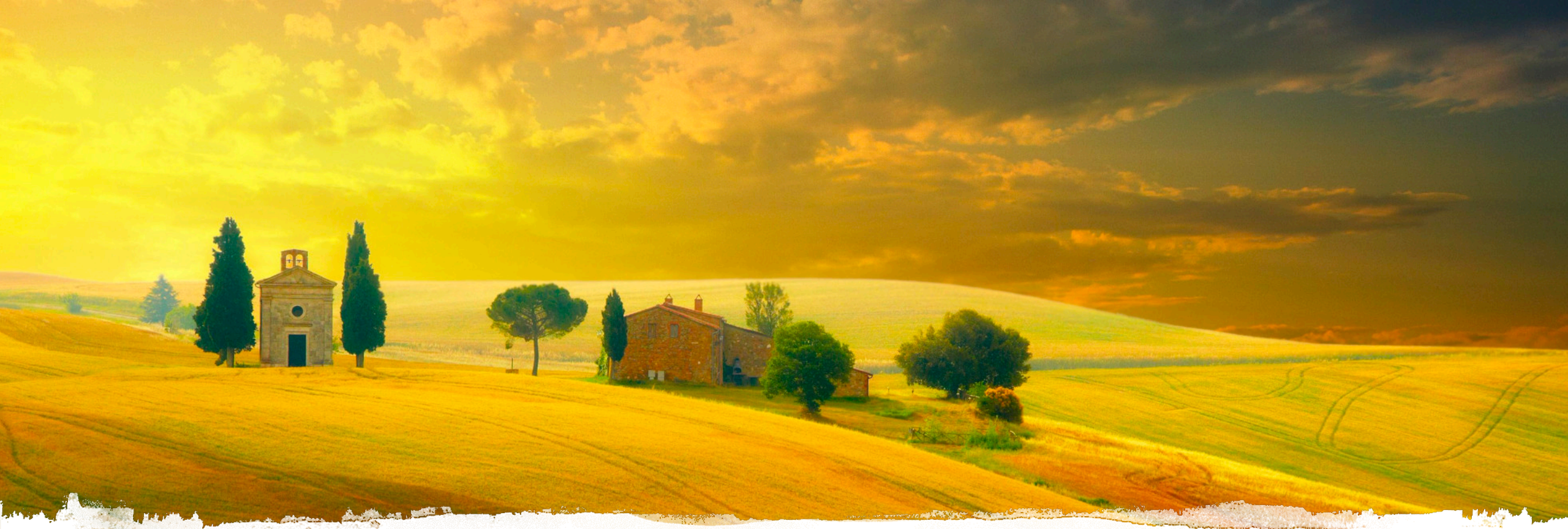


Conclusions

- We proposed a **methodology** for discovering the **factors** that influence the behaviour of bus travelers in ticket booking and to learn a **model** for predicting ticket purchasing.

- The **results** obtained by this study reveals that **factors** such **as occupancy rate**, **fare of a ticket**, and **number of days passed from booking to departure**, have significant influence on traveler's buying decisions.

- Using the methodology discussed in this work, the buying behaviour of **large communities of people** can be **analyzed** for providing **valuable information** and **high-quality knowledge** that are critical for the growth of business and organization systems.



Thank you!