Branda, F.; Marozzo, F.; Talia, D.; Discovering Travelers' Purchasing Behavior from Public Transport Data. In Proceedings of the 6th International Conference on machine Learning, Optimization and Data science-LOD 2020, Siena, Italy, 19–23 July 2020; pp. 702–713.

Discovering Travelers' Purchasing Behavior from Public Transport Data

Francesco Branda[1], Fabrizio Marozzo[1], and Domenico Talia[1]

DIMES, University of Calabria, Italy
{fbranda,fmarozzo,talia}@dimes.unical.it

Abstract. 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 adopted a dynamic pricing strategy, providing low-cost transport services 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. Starting from a set of 3.23 million user-generated event logs of a bus ticketing platform, the methodology shows the correlation rules between travel features and the purchase of a ticket. Such rules are then used to train a machine learning model for predicting whether a user will buy or not a ticket. The results obtained by this study reveal 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. The methodology reaches an accuracy of 93% in forecasting the purchase of a ticket, showing the effectiveness of the proposed approach and the reliability of results.

Keywords: Public Transport, Bus, Travelers' Buying Behaviour, Ticketing Platform, Machine Learning, Dynamic pricing

1 Introduction

The long-distance bus industry has traditionally been slow to evolve and it is quite resistant to change. While countries like UK, Sweden and Norway liberalized their coach transport market beyond high-speed rail a long time ago, other important markets like France, Italy, and Germany opened up recently [7]. A turning point has occurred in 2015 when FlixBus entered the European market, significantly increasing the supply of interregional buses and practicing aggressive pricing policies, to which many other local operators decided to adapt [6]. Therefore, thanks to relatively low cost, rapidly increasing convenience and routing flexibility, the bus transportation offers an added value to passengers over airlines and trains in the last years. In particular, the far-away locations of airports and the strict security procedures have made flying slower and tedious, are pushing more travelers to avoid airlines on short-to-medium distances. The train, on the other hand, appears to be rather expensive in many countries and though faster than the bus, it is also potentially more vulnerable to delays and missed connections. The rise in competition, the greater attention to customer experience and the possibility of offering long-distance journeys, lead transport companies to use intelligent tools to plan and manage their mobility offer in a dynamic and adaptive manner.

This paper presents a methodology, called *DA4PT* (*Data Analytics for Public Transport*), aimed at discovering the factors that influence travelers' behaviour in ticket purchasing. In particular, DA4PT uses Web scraping techniques and process mining algorithms to understand behaviours of users while searching and booking bus tickets. Starting from a set of user-generated event logs of a bus ticketing platform, the methodology shows the correlation rules between travel attributes and the purchase of a ticket. Then, such rules are used to train a machine learning model for predicting whether a user will buy or not a ticket.

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 1st, 2018 to October 20st, 2019. The results obtained by this study reveal that factors such as occupancy rate, fares of tickets, and number of days passed from booking to departure, have significant influence on traveler's buying decisions. We experimentally evaluated the accuracy of our methodology comparing some of the most relevant machine learning algorithms used in the literature [15]. Among them, Random Forest proved to be the best classification algorithm with an accuracy of 93% and low variance in results than other algorithms in the demand forecasting domain.

Compared to the state of the art, the presented methodology analyzes the user behavior on a bus ticketing platform for understanding if a user will buy a ticket after visiting the website and the main factors that influence her/his decision. The model obtained could be used to allow bus platforms to switch to (or improve) dynamic pricing strategies that maximize the percentage of occupation of a bus, the number of tickets sold, and the total revenue.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 outlines the main concepts and goals of our analysis. Section 4 presents the proposed methodology. Section 5 illustrates the case study. Section 6 concludes the paper.

2 Related work

Several approaches concerning demand forecasting have been proposed in the literature. In this section we briefly review some of the most representative related work in the area of demand forecasting, discussing differences and similarities with the methodology we designed.

Liu et al. [9] proposed a multi-factor travel prediction framework, which fuses complex factors of the market situation and individual characteristics of customers, to predict airline customers' personalized travel demands. With respect to our work, this is not focused on bus travels, however it could be applied to predict travel demands on the basis of the factors that influence travelers' buying decisions.

Szopiński and Nowacki [14] discovered that flight duration affect the price dispersion of airline tickets and the price dispersion increases as the date of departure approaches. Similarly, our work discovered that the low cost of a ticket can result in more sales. Therefore, it is advisable for a bus company to adopt a dynamic price strategy.

Other studies have attempted to predict the demand for transport services on the basis of price elasticity, i.e. a measure used to understand how demand can be affected by changes in price. Mumbower et al. [11] estimated the change in flight prices by using factors such as departure day of the week, time of departure, and date of booking. In particular, a linear regression method has been used to predict the number of bookings for a specific flight by date of departure, route and number of days before the departure date. Escobari [5] studied that consumers become more price sensitive as time to departure approaches and the number of active consumers increases closer to departure. These two papers cover only the Step 3 of our methodology and may be considered as alternative method for defining the correlation between the factors that influence travelers' buying decisions.

Abdelghany and Guzhva [1] used a time-series modelling approach for airport short-term demand forecasting. The model assesses how various external factors such as seasonality, fuel price, airline strategies, incidents and financial conditions, affect airport activity levels. In [17] Yeboah et al. developed an explanatory model of pre-travel information-seeking behaviours in a British urban environment, using binomial logistic regression. The considered factors include socio-demographics, trip context, frequency of public transport use, used information sources, and smartphone ownership and use. The two models proposed can be integrated in the methodology we designed.

3 Problem definition

Let $D = \{d_1, d_2, ... \}$ be a dataset collecting trip instances of a bus company, where each d_i is a tuple described by the following features: *trip itinerary* identifier; *origin* and *destination* cities; *booking* and *departure* date; *fare* of a ticket; number of *bus seats.*

Let $EL = \{e_1, e_2, ...\}$ be a set of event logs generated by users of the bus ticketing platform, where an event e_i is a tuple defined by the following fields: *cookie* of the user; description of the user *action*; *timestamp*; *trip itinerary* identifier; number of *bus seats* required by the user. For instance, a single event e_i may be: (*i*) find a trip, or (*ii*) calculate fare of a ticket for a given trip, or (*iii*) select a seat on the bus, or *(iv)* pay the booked trip. Some users finalize their search by purchasing the ticket (*purchased*) while others abandon the platform without purchasing the ticket (*abandoned*).

The main goal of this work is to infer patterns and trends about users behaviour for training a machine learning model that can predict whether a user will buy a ticket or not. More specifically, the data analysis we carried out aims at achieving the following goals:

1. Discover a set F of *factors*, $F = \{F_1,...,F_J\}$, where a *factor* F_j influence travelers' purchasing behaviour;
2. Train a machine learning model on the basis of F, to predict whether or not a user will buy a ticket.

4 Proposed methodology

As shown in Figure 1, the proposed methodology consists of four main steps:

1) Data collection through *Web scraping* techniques;
2) Pre-processing of event data and execution of *process mining* algorithms on event data;
3) Identification of *the main factors* influencing traveler's purchasing behaviour;
4) *Data analysis* and *machine learning* for purchase prediction.

For each step, a formal description and a use case are illustrated in the following sections.

4.1 Steps 1-2: Web scraping and process mining

The first two steps aim at defining the event logs EL used for understanding the behaviour of users while searching and booking bus trips. Specifically, during step 1, data collection is carried out by using Web scraping techniques (i.e., a set of techniques used to automatically extract information from a website), to know all the interactions of a user with the bus ticketing portal, for instance whether a user buys or not a ticket, or in which step of the buying decision process s/he leaves the platform.

Step 2 is aimed at exploiting process mining algorithms for learning and support in the (re)design of purchasing processes by automatically discovering models that explain the events registered in log traces provided as input [4]. The set of event logs EL is pre-processed in order to clean, select and transform data, making it suitable for analysis. In particular, we first clean the collected data for identifying the most compliant model with the event logs. Then, we proceed by selecting only the events that end successfully and unsuccessfully (e.g., the

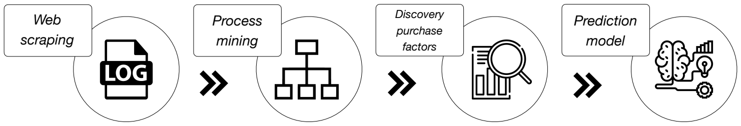


Fig. 1. The main steps of the DA4PT methodology.

purchase of a ticket and the abandonment of the platform by users, respectively). Finally, we transform data by keeping one event per user in a day.

The output of step 2 is the dataset $\hat{D} = \{\hat{d}_1, \hat{d}_2, ... \}$, where $\hat{d}_i$ is a tuple $\langle u_i, \{e_{i1}, e_{i2}, ..., e_{ik}\}\rangle$ in which e_{ij} is the j^{th} event generated by user u_i.

4.2 Step 3: Discovery of purchase factors

The goal of step 3 is to identify the key factors that push a user to buy a ticket. Specifically, we perform exploratory factor analysis for reducing a large set of attributes to a more coherent number which can explain travelers' purchasing behaviour. Then, we apply the correlation analysis to define the conditions that tend to occur simultaneously, or the patterns that recur in certain conditions.

In particular, the goal of our analysis is to generate correlation rules like $f \rightarrow e_{purchased}$ (if factor $f \in F$ occurs, then it is likely that also event $e_{purchased}$ occurs). The correlations between an attribute and the class attribute (*purchased* or (*abandoned*) is evaluated using the Pearson's correlation coefficient [12]. The values of the Pearson's correlation can be in the range [-1,1], where the value of 1 represents a strong linear relationship, 0 no linear correlation, while -1 corresponds to a negative linear correlation.

Below we report the meaning of the attributes that we have added into the dataset $\hat{D}$ before performing exploratory factor analysis. The value of these attributes has been calculated from other attributes in D: *i*) *Days before departure* (*DBD*), by calculating the difference between booking and departure date; *ii*) *Booking day of the week* (*BDOW*), by extracting the day from a booking date; *iii*) *Occupancy rate for a bus* (*OCCR*), by evaluating the number of required bus seats per passenger; *iv*) *Fare of a ticket* (*HMLF*), by dividing the price of each trip itinerary into three bands (high, medium, and low).

4.3 Step 4: Prediction model

After defining the purchase factors, we start the process of learning by analyzing the dataset $\hat{D}$ in order to train a model capable of automatically learning whether or not a user will finalize a purchase. In particular, the model has been trained on information which depends on the route, departure date and date of booking (e.g., ticket fare, occupancy rate of a bus).

The accuracy of our approach is evaluated comparing the most relevant machine learning algorithms used in the literature (Naïve Bayes [10], Logistic Regression [16], Decision Tree [13], Random Forest [3]). The performance of the machine learning models has been evaluated through a confusion matrix. Specifically, tickets that are correctly predicted as *purchased* are counted as True Positive (*TP*), whereas tickets that are predicted as *purchased* but are actually *abandoned* are counted as False Positive (*FP*). Similarly, tickets that are correctly predicted as *abandoned* are counted as True Negative (*TN*), whereas tickets that are predicted as *abandoned* but are actually *purchased* are counted as False Negative (*FN*). Starting from the confusion matrix we can compute metrics such as *accuracy*, *precision*, *recall* and *F1-score*.

It is worth noticing that our dataset is unbalanced because the two classes, *purchased* and *abandoned*, are not equally represented. In particular, there is a high percentage of users who visit the bus website without buying any tickets, and a low percentage who instead purchases tickets. In order to get accurate prediction models and correctly evaluate them, we need to use balanced *training sets* and *test sets* in which half the logs lead to the purchase of a ticket and half to abandonment. We used to this purpose the random under-sampling algorithm [8], which balances class distribution through random discarding of major class tuples as described in [2].

The machine learning algorithms have been implemented in Python using the library *sklearn* for producing the confusion matrix and the resulting measures, and its library *imblearn* that has been used to deal with the class-imbalance problem.

5 A case study

This section reports the results obtained by the analysis of event logs of an Italian bus ticketing platform. Specifically, we extracted more than 3 million of event logs about trip itineraries and price tickets of an Italian bus company, collected from August 1st, 2018 to October 20st, 2019. The total size of the final dataset $\hat{D}$ is about 700 MB. Data have been analyzed using the DA4PT methodology to discover the main factors influencing customers in purchasing bus tickets and, based on those factors, to train a machine learning model for predicting whether or not a customer will buy a ticket.

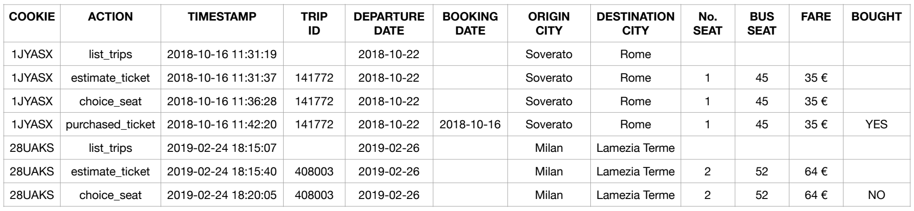
5.1 Steps 1-2: Web scraping and process mining

At *Step 1*, the set of event logs EL is composed by all interactions of the users with the bus ticketing platform. In particular, the buying decision process of a user is described by four types of event:

- *list_trips*, to find the routes between the 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;
- *purchased_ticket*, to confirm the payment of the booked trip.

Specifically, each event $e_i \in EL$ is defined as shown in Figure 2. For example, the user with ID *1JYASX* queried the system asking for the list of trips of the route Soverato-Rome (*line 1*). Then, he estimated the cost of the trip (*line 2*), and selected a seat on the bus (*line 3*). Finally, he paid for the ticket (*line 4*). The user with ID *28UAKS* logged on to the bus ticketing platform to estimate the cost of the route Milan-Lamezia Terme (*line 5-6*), but hasn't finalized the purchase of the ticket (*line 7*).

At *Step 2*, after having defined the set of event logs EL, we applied process mining algorithms with the aim of identifying trends and human patterns, and understanding behaviours of users while searching and booking bus trips.



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

Fig. 2. Example of the log traces extracted from the bus ticketing Web portal.

Figure 3 shows the navigation paths corresponding to those produced by human navigation on the bus ticketing portal. There are three type of paths: the green and red paths related to events that end with the purchase of a ticket (*purchased*) and the abandonment of the platform (*abandoned*) respectively, whereas the blue paths are related to redundant events that generate loops. The percentage present on the edges describes the users who leave a state to reach a previous/next state, or a terminal state (*abandoned* or *purchased*). For example, 100% of the users of *Start* state are distributed as follows: 70% look for a trip, then 14% continue browsing estimating the cost of the chosen trip, and 16% leave the platform without buying a ticket. We cleaned collected data by removing for each user all blue paths. Then, we selected only paths related to *purchased* and *abandoned* events. Finally, we transformed data by keeping the last event of the booking life cycle per user. At the end of this step, we built the final dataset $\hat{D}$ that results composed of about 300,000 tuples, each one containing the *purchased* or *abandoned* event by a single user.

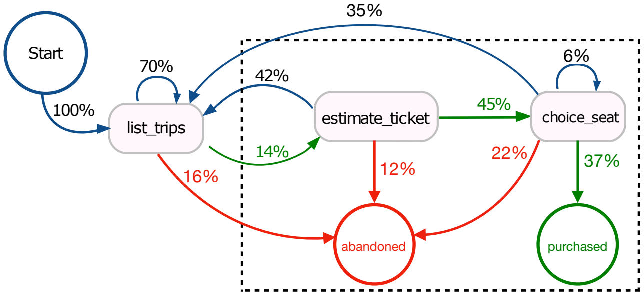


Fig. 3. Process mining algorithms applied to user event logs.

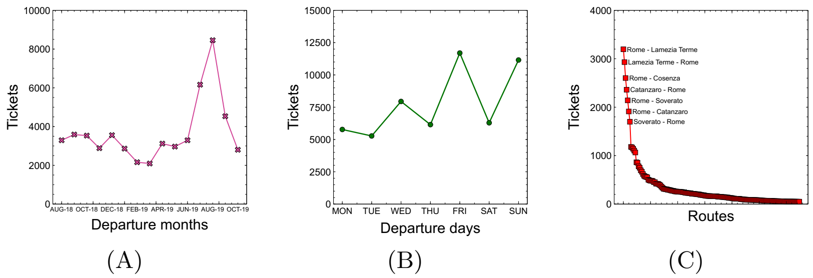
To analyze the behaviour of a user who searched the possible routes between two localities in a given date (*list_trips*), we focused on all the events he/she generates when chooses one of the routes. As shown in the dashed rectangle in Figure 3, we focused on the events *estimate_ticket*, *choice_seat*, *purchased* and *abandoned*. In this range, only 17% of users purchase a ticket (45% choose a seat, of which only 37% purchase) while 83% abandon the platform without buying. In the next section we study what are the main factors that lead a user to the purchase of a ticket.

5.2 Step 3: Discovery of purchase factors

In the following we present the main results of the exploratory factor and correlation analysis discussed above.

First, we describe some statistical indications of travelers' purchasing behaviour obtained from several attributes included in D. In particular, Figure 4 reports the number of purchased tickets considering *departure months*, *departure weekdays*, and *routes* attributes.

Specifically, Figure 4(A) suggests that most people travel in spring and summer, with a significant drop in passengers in autumn and winter, except for the Christmas holidays. Figure 4(B) indicates that the number of trips in the weekend is higher than on working days, whereas Figure 4(C) shows how some routes are more popular than others. For example, the number of purchased tickets for the route <Rome - Lamezia Terme> is about 3,400, because Calabria is a tourist destination and many workers/students live outside the region.



(A) (B) (C)

Fig. 4. No. of purchased tickets considering A) *departure month*, (B) *departure day of the week*, and (C) *route* attributes.

Another result of our analysis was discovering the correlation between the four derived attributes and the class attribute (*purchased* or (*abandoned*) as described in Section 4.2. Specifically, for each derived attribute, we measure the numbers and the percentage of purchased tickets and the correlation.

Starting from the *DBD* attribute, Figure 5(A) shows the number of purchased tickets versus the number of days before departure. It clearly shows that a few days before departure, users buy more frequently. For example, about 30K tickets have been purchased on the platform between 0 and 9 days before the trip, 20K between 10 and 20 days before, and so on. By observing the trend line over histogram in Figure 5(B), it can be noted that the percentage of purchasing a ticket is pretty high a few days passed from booking to departure, then there is a decreasing when the date of departure is far away. For example, 22% of users complete the purchase of the ticket between 0 and 19 days before the trip, 10% between 20 and 29 days before, and so on.

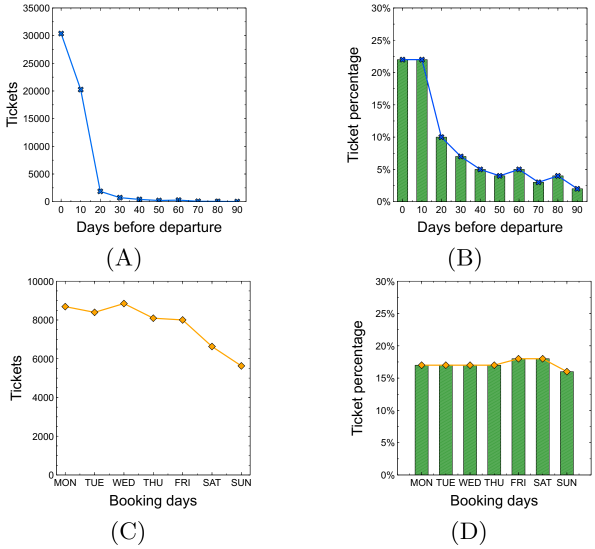


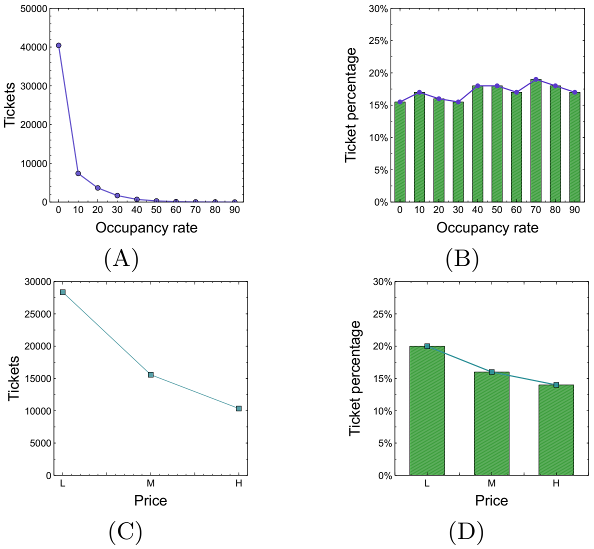
Fig. 5. No. and percentage of purchased tickets considering the *days before departure* (*DBD*) and the *booking day of the week* (*BDOW*).

Considering the *BDOW* attribute, Figure 5(C) shows the days of the week when users prefer to book a ticket. 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. Histogram in Figure 5(D) shows that the probability of purchasing a ticket is slightly higher on Fridays and Saturdays compared to other days of the week.

We also evaluated how the occupancy rate (*OCCR*) attribute influences the buying behaviour of the users. As shown in Figure 6(A), the tickets are mostly bought when the percentage of available seats is between 10% and 30%, whereas the trend line shown in Figure 6(B), describes that the probability of purchasing a ticket lightly increases when the bus seats are running out. Note that several buses do not reach the full occupancy because many tickets are not bought on the platform, and then are not registered in the event logs.

Finally, we show the impact that the *HML* attribute had on users' purchasing choices. In particular, for each trip itinerary we have divided the price into high, medium and low, discovering that most users are pushed to buy a ticket when the price is low (Figure 6(C)). In fact, the probability of buying a ticket in low range (about 20%) is much higher than buying it in medium and high ranges (about 15% and 13% respectively), as shown in Figure 6(D).

To define the potential purchase factors, a correlation analysis was executed. The *DBD* (days before departure) attribute have the highest correlation coeffi-



(A) (B)

(C) (D)

Fig. 6. No. and percentage of purchased tickets considering the *occupancy rate* for a bus (*OCCR*) the fare of a ticket *high, medium, and low* (*HML*).

cient (r) with a value of 0.86. The other attributes also have a high correlation with the class attribute: r=0.74 for *BDOW* (booking day of the week) and *OCCR* (occupancy rate) and, r=0.68 for *HML* (fare of a ticket).

5.3 Step 4: Prediction model

Before running the learning algorithms, we used the random under-sampling algorithm to balance class distribution in $\hat{D}$. In our case, we have a total of 247,525 samples: 42,995 *purchased*, and 204,530 *abandoned*.

The following parameters have been used for the evaluation tests: *i*) target dataset $\hat{D}$, *ii*) purchase factors, as described in Section 5.2, and *iii*) number of routes considered. As performance indicators we used the accuracy and *weighted-average* F1-score. The goal is to maximize accuracy with balanced values of F1-score. Moreover, to measure the quality of a classifier with respect to a given class, for each algorithm we evaluated the *purchased* recall (R_p) and *abandoned* recall (R_a).

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

Table 1. Performance evaluation.

Table 1 summarizes the results obtained by the four machine learning algorithms we used. Specifically, Random Forest proved to be the best classification model with R_p=0.95 and R_a=0.85, showing better accuracy and low variance in results than other algorithms. A high value of accuracy is also obtained by Decision Tree with R_p=0.87 and R_a=0.84, showing good robustness and stability.

A similar value of accuracy is observed for Naïve Bayes and Logistic Regression (R_p= 0.56 and R_a=0.65), but Naïve Bayes is less accurate on the *purchased* class than *abandoned* class (R_p=0.38 and R_a=0.84).

Figure 7(A) shows a time plot of the collected tickets data, in which the accuracy performance of the four machine learning algorithms is plotted versus the number of routes. The trend is quite evident: 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 Naïve Bayes (0.52-0.59).

Figure 7(B) shows the number of tickets correctly predicted related to the *purchased* class. Also in this case, the accuracy of Random Forest is the highest in all routes considered, confirming its very good prediction performance with respect to the other algorithms in the demand forecasting domain. Please notice that for both examples, we considered the first thirty routes based on the number of purchased tickets.

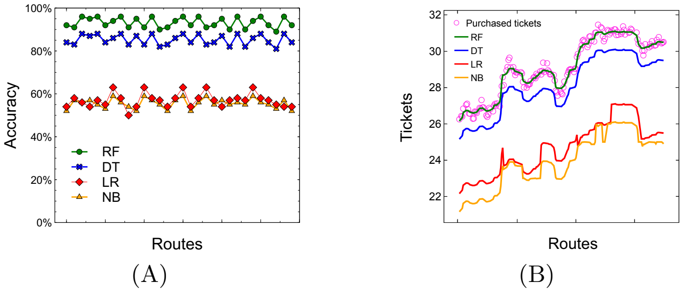


Fig. 7. Comparison among Naïve Bayes (NB), Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF).

6 Conclusions

This paper proposes a methodology (DA4PT) that through Web scraping techniques and process mining algorithms allows to discover the factors that influence the behaviour of bus travelers in ticket booking and to learn a model for predicting ticket purchasing.

DA4PT has been validated through a real case study based on 3.23 million event logs of an Italian bus ticketing platform, collected from August 1st, 2018 to October 20st, 2019. 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. We experimentally evaluated the accuracy of our methodology and Random Forest proved to be the best classification algorithm, showing an accuracy of 93% and a low variance.

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 fundamental for the growth of business and organization systems.

References

1. Abdelghany, A., Guzhva, V.: A time-series modelling approach for airport short-term demand forecasting. Journal of Airport Management **5**(1), 72–87 (2010)
2. Belcastro, L., Marozzo, F., Talia, D., Trunfio, P.: Using scalable data mining for predicting flight delays. ACM Transactions on Intelligent Systems and Technology **8**(1), 5:1–5:20 (2016)
3. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
4. Diamantini, C., Genga, L., Marozzo, F., Potena, D., Trunfio, P.: Discovering mobility patterns of instagram users through process mining techniques. In: IEEE International Conference on Information Reuse and Integration. pp. 485–492 (2017)
5. Escobari, D.: Estimating dynamic demand for airlines. Economics Letters **124**(1), 26–29 (2014)
6. Gremm, C.: Impacts of the german interurban bus market deregulation on regional railway services (2017)
7. Grimaldi, R., Augustin, K., Beria, P., et al.: Intercity coach liberalisation. the cases of germany and italy. In: World Conference on Transport Research-WCTR 2016. pp. 474–490. Elsevier BV (2017)
8. Kotsiantis, S., Kanellopoulos, D., Pintelas, P., et al.: Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering **30**(1), 25–36 (2006)
9. Liu, J., Liu, B., Liu, Y., Chen, H., Feng, L., Xiong, H., Huang, Y.: Personalized air travel prediction: A multi-factor perspective. ACM Transactions on Intelligent Systems and Technology (TIST) **9**(3), 1–26 (2017)
10. Maron, M.E.: Automatic indexing: an experimental inquiry. Journal of the ACM (JACM) **8**(3), 404–417 (1961)
11. Mumbower, S., Garrow, L.A., Higgins, M.J.: Estimating flight-level price elasticities using online airline data. Transportation Research Part A: Policy and Practice **66**, 196–212 (2014)
12. Pearson, K.: Determination of the coefficient of correlation. Science **30**(757), 23–25 (1909)
13. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics **21**(3), 660–674 (1991)
14. Szopiński, T., Nowacki, R.: The influence of purchase date and flight duration over the dispersion of airline ticket prices. Contemporary economics **9**(3), 253–366 (2015)
15. Talia, D., Trunfio, P., Marozzo, F.: Data Analysis in the Cloud: Models, Techniques and Applications (2015). https://doi.org/10.1016/C2014-0-02172-7
16. Walker, S.H., Duncan, D.B.: Estimation of the probability of an event as a function of several independent variables. Biometrika **54**(1-2), 167–179 (1967)
17. Yeboah, G., Cottrill, C.D., Nelson, J.D., Corsar, D., Markovic, M., Edwards, P.: Understanding factors influencing public transport passengers' pre-travel information-seeking behaviour. Public Transport **11**(1), 135–158 (2019)