Prediction-Based-Marketing-Tool (PBMT) for Bike Sharing Provider

Timur Levent Sereflioglu
Consumer Analytics
University of Siegen
Human-Computer-Interaction
1075805

ABSTRACT

UPDATED—10 February 2020. In this model, we try to predict exactly what the next location target of the bike sharing user could be. Data from different data records are merged and edited in the following model and prepared for the modeling of a neural network class classification model (CCM). The benefit could be to allow the user to receive targeted advertising and offers that are tailored to the interests or concerns of the user. The overall goal is to predict the user's next location goal using the record of date, day of the week, starting location, and weather. Although this model is still rudimentary, it can certainly have potential for further development. This work is therefore to be understood as a prototype idea that can be developed and further designed using human-computer interaction methods, e.g. through design thinking and user experience design (UXD) approaches.

Author Keywords

bike-sharing; consumer-analytics; data-science; neuronal-networks, multi-class-classification; predictive-analytics; predictive-marketing.

CSS Concepts

Architectures; *High Relevance*

- Other architectures
- Neural networks

Computer systems organization; *Medium Relevance*

- Real-time systems
- Real-time operating systems

text: https://dl.acm.org/ccs/ccs_flat.cfm

INTRODUCTION - CONTEXT

There are now 1.3 billion vehicles in the world, and the number is increasing every day (STATISTA, 2017). This development has a particularly problematic impact on the infrastructures of large cities and metropolises. Infrastructure is not developing at the same speed, and there are limits that a city's infrastructure cannot overcome without new ideas. In the western world in particular, people's income is growing and the car is becoming more attractive and affordable while the importance of pedestrian and bicycle traffic is decreasing. This development is exacerbated by the strong urbanization that leads to the rapid growth of cities. As a result, the road network of many cities is congested, the population suffers from traffic jams, bad air and noise. And since the road network in many cities is not geared towards so much car traffic, the average speed in city centers on weekdays is often below 20 km/h (Accelerating Urban Logistics, 2019). In Germany, 62% of the traffic routes are covered by car, 32% on foot and only 15% by bicycle (Nobis & Kuhnimhof , 2018). However, there are many advantages to using bicycles in cities, so you are often faster on your short journeys than by car and do not cause any emissions. Bike sharing is therefore a watchful concept, the greatest development can be observed in Asia.

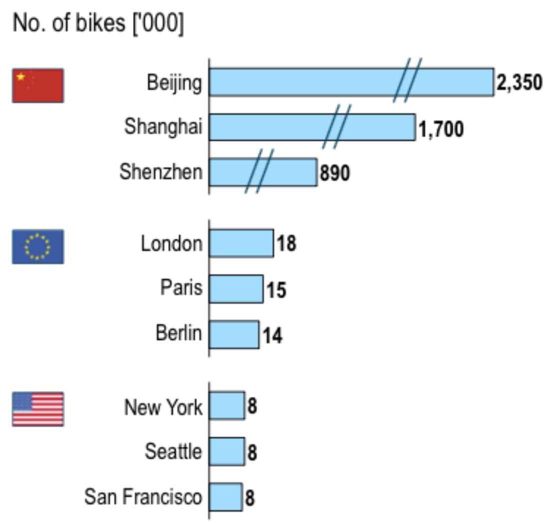

Figure 1: Number of bikes in China, Europe and USA

Source: Press Research; Roland Berger

Worldwide there are more than 10 million bicycles for rent in 71 countries (Roland Berger , 2018).

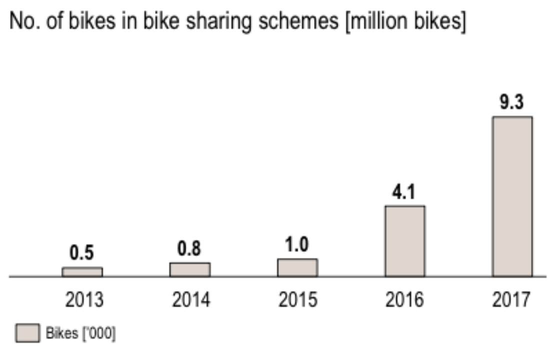

Figure 2: Number of bike sharing schemes

Source: Press Research; Roland Berger

Even if this can significantly improve the traffic situation, as could be observed in the 100 largest Chinese cities, it also brings with it some problems, which we will not go into further in this paper. In this study, we will concentrate on the possibilities that bike sharing and its use offer. For this purpose, we have several data sets from the Cologne Transport Association (KVB rad) and data from the German Weather Service available to us as part of a university project called "*Consumer Analytics*".

PROBLEM

In this work, we deal with the question of how this data can be used to solve the problems associated with bike sharing, but also how this data can be used to create new offers. Marketing is often an interesting topic in this context. Predictive analytics for marketing has only been a big issue since some time, because the computing power of computers has increased, the data has become more accessible and the software are more user-friendly. Today, after almost 30 years of marketing, the term ***"predictive analytics"*** is a catchphrase in the marketing industry.

This approach is not new, it is known from **"others also bought..."** suggestions. In our concept, however, we try to tell the drivers beforehand which *"point of interest"* they will go to next. In terms of marketing technology, this is intended to offer a new possibility of monetization. After all, sharing providers such as **obike** or e-scouter **coup,** made headlines because of their bankruptcy. The big bike sharing pioneer **ofo** is reportedly also at risk of bankruptcy if new money does not flow into the business quickly enough (Focus Money Online, 2019). Our **Prediction-Based Marketing Tool (PBMT)** could represent an additional source of income.

SOLUTION

How can we make our data accessible for marketing concepts? Conventional prediction models often take into account the individual user, which means that data about interests or preferences must first be collected. This then leads to individual predictions in the form of ***"offers that might interest you"***. In our approach, we use data from all users and draw conclusions about the person without having previously collected a single data point about an unique person.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 101076 entries, 0 to 101075
Data columns (total 8 columns):
id                 101076 non-null object
bike               101076 non-null int64
start_datetime     101076 non-null object
end_datetime       101076 non-null object
start_lat          101076 non-null float64
start_lng          101076 non-null float64
end_lat            101076 non-null float64
end_lng            101076 non-null float64
dtypes: float64(4), int64(1), object(3)
memory usage: 6.2+ MB
```

Table 1: Overview of the values of the bike sharing data

In order to develop a neural network that gives us the output we expect, we have to make a selection of the data and variables. We chose the starting coordinates, the starting time, the date of the trip and the weather conditions of the trip. Further data that we need for the output are the data about the point of interest, whose variable are the title and the category of the coordinates. These variables represent the variables of the learning algorithm.

Table 2 shows the data frame of the x-axis, which we use for our **multi-class classification model**. To achieve our results, we use the **keras deap learning** library. The following figure *(Figure 1*) shows the marketing-relevant POIs, we have filtered them according to commercial use. Of the original 69 categories, 42 still remain.

```
array(['Service', 'Hotel', 'Public Transport', 'Food & Drink',
       'Petrol Station', 'Restaurant', 'Clothing & Accessories', 'Shop',
       'Theatre, Music & Culture', 'Bar/Pub', 'Book Shop', 'Coffee/Tea',
       'Dance or Nightclub', 'Going Out', 'Sport Facility/Venue',
       'Snacks/Fast food', 'Outdoor Sports', 'Shopping Centre',
       'Sights & Museums', 'Hospital or Healthcare Facility',
       'DIY/garden centre', 'Business & Services', 'Electronics',
       'Railway Station', 'Post Office', 'Car Dealer/Repair',
       'Eat & Drink', 'Communications/Media', 'Car Hire', 'Museum',
       'Taxi Stand', 'Transport', 'Wine & Spirits', 'EV Charging Station',
       'Casino', 'Travel Agency', 'Department Store',
       'Kiosk/24-7/Convenience Store', 'Theme Park', 'Hostel', 'Zoo',
       'Cinema'], dtype=object)
```

Figure 3: Classification array of all categories after filtering

It is also important to explain how the POI's could be assigned to the ID's of bike sharing data. The assigned data do not correspond to the exact goals of the bike sharing users, they were assigned with a euclidean distance calculation and are therefore the closest points.

```
df1 = data1
df2 = data2

df1['point'] = [(x, y) for x,y in zip(df1['lat'], df1['lng'])]
df2['point'] = [(x, y) for x,y in zip(df2['end_lat'], df2['end_lng'])]

df2['closest'] = [closest_point(x, list(df1['point'])) for x in df2['point']]
df2['title'] = [match_value(df1, 'point', x, 'title') for x in df2['closest']]
```

Figure 4: For-loop euclidean distance (jupyter notebook)

It is therefore quite possible that the assigned POIs were not approached at all. The fact that these POIs are not exact does not have a major impact on the result of our model, despite everything, in our attempt we have proven the theoretical functionality of our learning algorithm. *Figure 4* shows the programming code for the distance calculation and assignment to the final coordinates of the bike sharing users.

Our y-axis is the variable of the POI data, which specifies the exact title of the location and also bears the name *"title"*. The *table 2* shows a list of the POI titles, with 2089 data points. This class list represents also the y-axis, which we will convert into an array.

```
0                  AMERON Koln Hotel Regent
1                        Eis Café la Michel
2        Dr. med. Hayriye Öznur Ünal-Maelger
3                       Dost Döner Pizzeria
4                        Marietta Schlieker
                           ...
2084                 Heilerin Wiebke Nimmer
2085                            Baolian Han
2086                 Weidenpesch Mollwitzstr.
2087                         Düxer Schützen
2088                      Bogdan Paul Dubas
Name: title, Length: 2089, dtype: object
```

Table 2: List the title of the POI's (y-axis)

In the final step, we use a sequential learning model that uses classifiers to sequentially limit the number of competitions, while maintaining the true outcome in the candidate set with high probability. Due to the large number of data points, our model has an output of 481 nodes and 1671 hidden layer nodes, which are calculated with a softmax activation function.

	bike	day_of_week_x	day	month	start_time_h	start_lat	start_lng	temperature	wind	sunshine	precipitation
0	22196	1	1	7	14	50.927955	6.932473	24.6	5.4	47.0	0.0
1	21596	3	31	7	7	50.942062	7.001396	17.3	2.3	0.0	0.0
2	21596	3	31	7	7	50.942062	7.001396	17.3	2.3	0.0	0.0
3	21596	3	31	7	7	50.942062	7.001396	17.3	2.3	0.0	0.0
4	21841	5	19	7	18	50.915899	6.942514	24.7	2.4	13.0	0.0

Table 3: Random columns and rows of the x-axis

FINDINGS

Our model gets an accuracy of 0.7037 which means that the forecast will be 70% accurate. In a test sample, we selected five random data series, as can be seen in the *table 5*. The coordinates on the left are the location data of the POIs and the coordinates on the right are the data of the final coordinates from the bike sharing data set. For our marketing tool, we should be able to predict the correct title of the user location target. In our example, we have five random users, who we know at what time they start, where they start from and what the weather conditions currently is, so we simulate that this scenario takes place in real time. We only do not know those in the *table 5* shown.

Before we move on to the results, we should give some data about the **Root Mean Square Error** (RMSE). In our model, we have an RMSE of the test file of 0.03 and an RMSE of the training file also of 0.03, the similarity of these two values means that our model is neither over fitted or under fitted *(Figure 5)*.

lat	lng	title	category	end_lat	end_lng
50.94333	6.95757	McDonald's	Snacks/Fast food	50.927979	6.942549
50.95038	6.91339	Street Food Festival	Theatre, Music & Culture	50.953099	6.937660
50.94036	6.89899	Hermes PaketShop	Post Office	50.966913	6.965007
50.92038	6.96898	Plazaa	Business & Services	50.920055	6.967458
50.95105	6.91741	PENNY	Food & Drink	50.951083	6.917523

Table 4: Random sample from the data frame

```
Train RMSE: 0.03
Test RMSE: 0.03
-----------------
```

Figure 5: RMSE of the test train split

In the last step of our model we have the prediction printed out and compare it with the data from the sample. *Table 5* shows five predictions, each belonging to the individual example user. To remember, we have a 70% chance that the output is correct.

```
The title of the first user target is... ["McDonald's"]
The title of the second user target is... ['Uni Caféteria']
The title of the third user target is... ['Hermes PaketShop']
The title of the fours user target is... ['Plazaa']
The title of the fives user target is... ['PENNY']
```

Table 5: Output of the prediction (n=5)

If we compare *table 4* with *table 5*, we see that one prediction is wrong, we predicted the *"Uni Cafeteria"* as the target of the second user but the correct target was *"Street Food Festival"*, so we have 4 out of 5 correct predictions. Whether such a marketing tool can be used sensibly for monetization purposes depends on the benefits of other companies or organizations. Only if a marketing relevance is recognized can this tool also generate money. For this purpose, we have collected a few details that could be interesting for a possible market entry.

title	counts
Hermes PaketShop	90
REWE	39
Netto Marken-Discount	29
Kamps	28
ALDI SÜD	26
O2	19
ARAL	18
dm	18
Automobilgruppe Dirkes GmbH	16
McDonald's	15

Table 6: The most common POIs title that are approached

The most popular POIs for users are *"Herms Paket Shop"*, *"REWE"* and *"Netto Marken-Discount" (Table 6)*. Now you would have to come up with suitable concepts for useful marketing campaigns. The underlying data and the developed model leave some possibilities open and should also be useful for the concept development. The POIs can also be divided into categories, so that campaigns that are suitable for different sectors could be developed. *"Business and Services"* and *"Restaurant"* seem to be the most important business for bike sharing users. *(Table 7).*

category	counts
Business & Services	2894
Restaurant	2651
Hospital or Healthcare Facility	1914
Shop	1752
Service	1421
Clothing & Accessories	1030
Food & Drink	993
Sport Facility/Venue	586
Public Transport	567
Theatre, Music & Culture	429

Table 7: The most common POIs categories that are approached

LIMITATIONS

It is questionable what acceptance the user has towards the marketing of effective offers by the bike sharing provider. Furthermore, a qualitative observation of what the users do at the POI would be recommended; the behavior of the users at the destinations could be decisive for the development of this tool. Do the users shop at the POI or do they possibly work there? These and other differences would have to be established in further studies.

CONCLUSION

The predictive based marketing tool for bike sharing provider could address a financing problem in the industry and could potentially be an additional source of income. Before that, further studies and tests are necessary, however, the fact that the prediction works theoretically has already been proved in this work with PBMT for bike sharing provider.

REFERENCES

[1] Accelerating Urban Logistics. 2019. *Analyse Verkehrsdaten*. (5 May, 2019). Retrieved February 7, 2020 from go.engage.here.com: https://go.engage.here.com/Accelerating-Urban-Logistics.html

[2] Focus Money Online. 2019. *Leihrad-Pionier Ofo steht kurz vor dem Kollaps.* (5 January, 2019). Retrieved February 7, 2020 from focus.de: https://www.focus.de/finanzen/news/unternehmen/kunden-stocksauer-leihrad-pionier-ofo-bediente-sich-an-kunden-kautionen-und-bekommt-nun-die-rechnung_id_10147048.html

[3] Claudia Nobis, Tobias Kuhnimhof. 2018. *Mobilität in Deutschland - MiD Ergebnisbericht*. Retrieved February 7, 2020 from http://www.mobilitaet-in-deutschland.de/pdf/MiD2017_Ergebnisbericht.pdf

[4] Roland Berger. 2018. *Bike Sharing 5.0 - Market insights and outlook.*

[5] STATISTA. 2017. Worldwide car and commercial vehicle inventory. (July, 2017). Retrieved February 7, 2020 from de.statista.com : https://de.statista.com/statistik/daten/studie/244999/umfrage/weltweiter-pkw-und-nutzfahrzeugbestand/

[6] Even-Zohar Y, Roth D. Urbana-Champaign. 2001. *A Sequential Model for Multi-Class Classification*. (20 June, 2001). Retrieved February 7, 2020 from https://arxiv.org/pdf/cs/0106044.pdf