Predicting City Pass use among low-income citizens of Amsterdam

In collaboration with the Municipality of Amsterdam

Wenjin "Sissi" Cao

Daily supervisor: Xanne Cooke
First examiner: Dr. Krista Overvliet
Second examiner: Dr. Chris Janssen

Master thesis

Utrecht University

Gemeente Amsterdam

Artificial Intelligence
Utrecht University
The Netherlands
2019

Abstract

The City Pass is one of several Poverty Reduction programmes from the Municipality of Amsterdam. It enables low-income citizens in Amsterdam to partake in a wide range of activities, either for free or with a discount. These include cultural and sport locations. This study investigates how well City Pass use can be predicted and understood with machine learning techniques, with a focus on interpretability. Interpretability includes insights such as feature importance from the supervised machine learning models. This can be valuable in creating more understanding of City Pass user behaviour. City Pass use encompasses unique use, as well as cultural and sport participation. Unique use refers to whether an owner of a City Pass actively uses it. Participation refers to having visited a type of location as outcome. Three existing supervised machine learning models and an unsupervised machine learning model were implemented for this task. Data included user level information such as demographic data, as well as neighbourhood data, information about the locations, and additionally travel distance. The obtained results show that the supervised models generally perform well on predicting unique use, and visiting different cultural and sport locations. The models rely on a mix of aforementioned feature types, each varying in effect depending on the outcome. Based on these results, it can be concluded that machine learning can be an interesting tool in uncovering the underlying contribution of various factors in behaviour.

Acknowledgements

While writing this thesis, I have received a great deal of support from various people. I would first like to thank my supervisor at Utrecht University, Dr. Krista Overvliet, for her constructive guidance and expertise throughout my writing process. I would also like to thank my supervisor Xanne Cooke at the Municipality of Amsterdam for the inspiration and helping me navigate through my internship. Also many thanks to all my colleagues at the Municipality of Amsterdam for their enthusiasm and help. Finally, I would like to express gratitude to all my dear friends and family who have helped and supported me.

Contents

1 Introduction **5**
- 1.1 Poverty Reduction . . . 5
- 1.2 City Pass . . . 5
- 1.3 Cultural and sport participation . . . 6
- 1.4 Theoretical and practical relevance . . . 8
- 1.5 Goal and research questions . . . 9
- 1.6 Machine learning . . . 9

2 Methodology **12**
- 2.1 Data and procedure . . . 12
  - 2.1.1 Privacy . . . 13
- 2.2 Preprocessing . . . 14
  - 2.2.1 User data . . . 14
  - 2.2.2 Transaction data . . . 16
  - 2.2.3 Neighbourhood data . . . 17
  - 2.2.4 Travel distance . . . 18
  - 2.2.5 Online ratings . . . 18
  - 2.2.6 Final prediction data . . . 18
  - 2.2.7 Final clustering data . . . 19
- 2.3 Models . . . 21
  - 2.3.1 Logistic Regression . . . 23
  - 2.3.2 Random Forest . . . 24
  - 2.3.3 XGBoost . . . 24
  - 2.3.4 K-prototypes . . . 25

3 Results **27**
- 3.1 Prediction: unique use of City Pass . . . 27
- 3.2 Prediction: cultural and sport participation . . . 32
  - 3.2.1 Museum . . . 34
  - 3.2.2 Cinema . . . 37
  - 3.2.3 Theatre . . . 40
  - 3.2.4 Sport . . . 43
- 3.3 Clustering . . . 46
  - 3.3.1 Subgroups unique use of City Pass . . . 46
  - 3.3.2 Subgroups cultural and sport participation . . . 47

4 Discussion **50**
- 4.1 Prediction performance . . . 50
- 4.2 Feature importance . . . 51
  - 4.2.1 Top features . . . 51

4.2.2 Neighbourhood . . . 53
4.2.3 Demographic . . . 53
4.2.4 Location . . . 54
4.2.5 Multicolinearity . . . 55
4.3 Clustering for subgroups . . . 55
4.4 Limitations . . . 56
4.4.1 Data . . . 56
4.4.2 Models and analysis . . . 56
4.5 Future research . . . 57

5 Conclusion **59**

References **60**

Appendices **65**

A Feature importance scores **65**

B Learning curves **70**

1 Introduction

1.1 Poverty Reduction

Poverty is a long-standing global issue (United Nations, 2019), also present in developed Western countries such as the Netherlands. Poverty can have profound consequences, which include social exclusion, decreased participation in cultural and sports activities, struggles affording food, housing, education, and health care. Yoshikawa et al. (2012) and Michon and Slot (2014) mentioned there is a direct relationship between poverty and decreased physical and mental health. Therefore, poverty is an important focus in government policies, which try to address this problem (European Commission, 2014). Many municipalities in the Netherlands have special programmes targeted at reducing the burden of poverty among low-income citizens (Rijksoverheid, 2019). The municipality involved in this study is the Municipality of Amsterdam.

The Municipality of Amsterdam has eleven main Poverty Reduction programmes as part of their social services. They are primarily focused on increasing social participation among their citizens, the main four target groups being: low-income households with children, low-income pensioners, low-income adults and chronically ill citizens (Onderzoek, Informatie en Statistiek, 2017). Each year, the Municipality of Amsterdam sends out application forms to citizens who are considered eligible for any of the above mentioned programmes. These programmes include services such as Free Public Transport for the elderly, Schoolchildren's Allowance, PC Provision, Individual Income Support and the so called *City Pass*, also known as Stadspas in Dutch. For this study, the City Pass is of specific interest because it is a programme that all low-income citizens can freely apply for. It is a pass that enables citizens in Amsterdam to partake in a wide range of activities, either for free or with a discount. These activities are categorised into cinema, museum, sport & swimming, theatre and other. The City Pass is targeted at low-income citizens of all ages, and a smaller subset is available for regular non-low-income pensioners as well (Gemeente Amsterdam, 2019).

1.2 City Pass

Recent statistics on the Poverty Reduction programmes show varying degrees of participation rates for the programmes. Participation rate can be defined by the number of households who have been assigned to a program, divided by the number of eligible citizens for a programme. 'Eligible citizens' refers to citizens who can apply for the programme, but have not necessarily done so. For the City Pass, the range was 68% between 2015 and 2017 (Onderzoek, Informatie en Statistiek, 2017). The various programmes are meant to supplement each other. For example, the Child Coupons, a financial contribution from the municipality meant for spending on clothes and toys, can only be used if a citizen also owns a City Pass. According to the latest statistics

from December 2018, there are 118.747 low-income citizens in possession of the City Pass. This is excluding non-low income 65+ users. There are 63.122 unique active users in this group, meaning only 55 % of the total City Pass users have actively used the City Pass in 2018. There were a total of 370 different organizations involved in facilitating City Pass activities (Gemeente Amsterdam, 2018). The unique use rate is not as high as desired, and some activities might be under attended or not quite well tailored towards the users. Therefore, the Municipality of Amsterdam is interested in how unique use rate for the City Pass can be increased, and whether the use of the City Pass can be predicted and perhaps understood. Increasing this rate is important because the City Pass can offer improvements to the quality of life of low-income citizens. City Pass use encompasses unique use rate: who does and does not actively use the City Pass, as well as participation in different types of activities or locations. This study will investigate these activities on the level of City Pass partners (i.e. locations), as well as on the category level. The categories are museum, cinema, theatre and sport. The question arises if there is a pattern in which type of City Pass activities citizens tend to partake, and how well this can be predicted. Furthermore, these insights in prediction can help gain understanding about which factors influence participation in various City Pass locations. These insights can perhaps help improve the service of the City Pass.

1.3 Cultural and sport participation

As mentioned earlier, the City Pass functions as an access point to various cultural and sport activities for low-income citizens of Amsterdam. Concerning cultural participation, work in the Dutch social science domain has shown that different factors can be at play. Demographic factors such as education level can have significant role in cultural participation, especially in *highbrow* activities. This includes museums and theatres, and can be differentiated from lowbrow activities, which includes cinema and popular music concerts. People who have attained a higher level of education are said to show higher participation rates. Women show higher rates compared to men, whereas younger people attend cultural events more frequently (Notten et al., 2015). Notten et al. (2015) mentions that this difference is explained more so by educational level than, for example, income level. Therefore this might be reflected in the City Pass use as well. Nagel (2009) found that differences in education was associated with differences in cultural participation among youth between ages 14-24, although participation is more significantly influenced by participation behaviour of parents rather than by educational level. Some differences were found in cultural participation among youth of different minority ethnic backgrounds (Van Wel et al., 2006), as well as differences in participation within various age groups. Older adults show higher participation rates for highbrow activities compared to lowbrow activities (Toepoel, 2011). A European-wide study shows that income level correlates with participation rates, with the lower-income

group showing lower rates, but proportionally higher rates for lowbrow activities compared to highbrow activities (Eurostat, 2017). Since City Pass is mainly targeted at low-income citizens, differences in income-level might be less obvious. This is also due to the free or discounted nature of the activities, which should form less of an financial barrier. Other factors to consider are environmental ones. Work by Brook (2016) using Logistic Regression to predict cultural participation in London, suggests that neighbourhoods operate as 'opportunity structures' which help enable cultural participation. This means that accessibility of locations can be factors in participation. These are all factors which can be taken into account when considering the inclusion of features in predicting City Pass use.

Regarding sport participation, research in Australia has found that socioeconomic status, as well as accessibility of sport facilities in neighbourhoods, to be a significant factor. Socioeconomic status includes income level, educational background and occupational status (Eime et al., 2015). In the Netherlands, factors such as demographic and socioeconomic ones, specifically age, education and income also account for differing rates in sport participation, as well as travel distance (Hoekman et al., 2016). Shenassa et al. (2006) states that a neighbourhood characteristic such as perceived safety can play a role in sport participation, where men are more likely to participate, and married people were less likely to do so. Allen and Vella (2015) shows this as well, mentioning that demographic, socioeconomic and neighbourhood variables can all be contributing factors in sport participation. Ruseski et al. (2011) found that household type, notably containing children, reduces the likelihood that individuals participate in sport activities.

The current available data on City Pass users includes demographic data, socioeconomic data, and historical data of all activities, the latter kept in a transaction database. Information on which other Poverty Reduction programmes users have applied for is also available. This is particularly interesting if there happens to be a relationship between unique use of the City Pass and the use of other Poverty Reduction programmes. Users who are generally active, might also be more pro-active in other aspects. As such, information on whether a user applied for the following Poverty Reduction programmes were also used: Schoolchildren's Allowance, PC Provision, Free Public Transport and Individual Income Support. While using purely demographic data might be sufficient for predicting City Pass use, it can be subject to many various and possibly complex sociological and psychological factors. The goal is to predict and understand City Pass use with a data-driven approach. This means a wider range of features will be taken into account. It is expected that factors such as age, gender, type of income, educational level, type of household, will play a role in unique use rate. This is expected based on the differences shown in general participation rate, which might extrapolate to unique use of the City Pass as well Onderzoek, Informatie en Statistiek (2017). Such factors might also play a role in what type of activities users participate in, which can be anything from the categories museum, cinema, theatre and sport.

It would be insightful to perform prediction using a combination of user level data

and environmental data. This can be for example data of the neighbourhood an user resides in. Specifically, neighbourhood characteristics such as safety index, ethnic composition, and for example, percentage of families with children might be relevant to take into account. In research on social participation, Piscopo et al. (2017) found that neighbourhood characteristics can influence social participation. This included the proportion of people in intermediate occupations, proportion of education level, proportion of small company owners, and percentage of households with children. Perhaps this can play a role in cultural and sport participation as well. This makes an interesting case to include such features from a publicly available neighbourhood data, retrieved from Onderzoek, Informatie en Statistiek (2018). The influence of the environment must be taken into account, since humans do not operate in 'vacuum'. This, in addition to City Pass user data, can all aid predicting and understanding City Pass unique use rate of current users, as well as cultural and sport participation. If the models generalize well, it would be possible to make predictions for any new user. This can be useful for estimating if someone will use the City Pass, and what they will use it for. Considering the possible influence of accessibility, especially since various activities are scattered across the city, travel distance will be taken into account as well. Other data which can be included are characteristics about the City Pass locations themselves. This can be the popularity of a location, which can be reflected by the number of ratings Jannach et al. (2013) Powell et al. (2017), be it from Google, Facebook, or any other directly available review site. It could be that people are more likely to visit a location because of higher perceived popularity. Overall, there is a decent amount of data to work with. For this study, machine learning is considered as method of choice. Previously discussed studies on cultural and sport participation primarily employed either statistical methods, or only logistic regression for prediction. These studies have not extensively looked into prediction, evaluation of various machine learning methods, and understanding predictions, using a wider range of data. This study aims to provide a different angle of approach.

1.4 Theoretical and practical relevance

Researching the application of machine learning on the City Pass case has a social character, and will not only help with trying to predict user behaviour, but also potentially create more understanding of low-income citizen behaviour. This can help with improving the Poverty Reduction programmes. The goal is not to only achieve reasonable model performance, but also learn something about the underlying structures that determine behaviour. We can make inferences about the relationships of various predictors by determining which features contribute to the best model fit. To help understand City pass use, it is also valuable to try characterize users who use or do not the City Pass, and users who visit or do no visit a type of location. Clustering can offer extra insight in this aspect, and can be implemented in order to understand which subgroups might require more focus. Various social factors can be at play in

determining participation, on a user-level as well as neighbourhood level. The number of City Pass locations available in a neighbourhood might influence whether people will use the City pass. Something simple such as travel distance might influence the likelihood of visiting a location, or the popularity of a location. The contribution lies in determining how well machine learning can model City Pass-user level behaviour, and enabling the inference of which factors are at play. This study forms a bridge to sociological knowledge, and can hopefully help improve the City Pass programme.

1.5 Goal and research questions

The aim of this study is to investigate how well City Pass use can be predicted and understood using machine learning techniques. This is done in combination with demographic, neighbourhood, geographical, location based data such as travel distances, and additional meta data of locations, such as online ratings. The objective is to predict unique use of the City Pass, as well as cultural and sport participation. Unique use of City Pass is defined as whether a person has actively used the City Pass, in this case during the time period of 2018. Cultural and sport participation are defined as whether an active City Pass user has visited any City Pass location belonging to these categories.

The main research question for this study can be formulated as:

How well can we predict and understand the use of City Pass among low-income citizens from Amsterdam with machine learning, using demographic, neighbourhood, geographical and meta data of locations?

The three sub questions are as follows:

1. Which supervised machine learning methods yield the best performance for predicting City Pass unique use, cultural participation and sport participation?
2. What are the top features for predicting City Pass unique use, cultural participation and sport participation?
    - What can we infer from these top features about behaviour and background of users?
3. Can subgroups be identified among users and non-users of the City Pass, and visitors and non-visitors of culture and sport locations?

1.6 Machine learning

For this thesis, multiple machine learning models for predicting City Pass use will be evaluated. Machine learning is a field of research that falls under the umbrella

of Artificial Intelligence, and has a wide range of scientific and practical applications (Russell and Norvig, 2010). In essence, it concerns various self-learning algorithms that can be trained and applied to predict labels of new data, based on previously learned patterns (Bishop, 2006). Such applications notably include natural language processing techniques such as speech recognition, computer vision in medical research and autonomous vehicles, fraud detection, sentiment analysis, but also modeling the brain architecture with deep learning (Kriegeskorte, 2015). It can also be applied to predict human behaviour, such as buying behaviour (Stubseid and Arandjelovic, 2018) and movie preferences (Bennett et al., 2007).

In the social-cultural domain, machine learning applications have been employed as well. Some recent examples include suicide risk prediction (Walsh et al., 2017), understanding and predicting rumour spreading behaviour during the 2016 US elections (Cao et al., 2018), and modeling trust (Nelson et al., 2016). Nelson et al. (2016) mentions that machine learning can be used to identify important features in data that can serve as an indicator of behaviour and help with further theorizing on a phenomenon and check for compliance with existing literature and theories. Walsh et al. (2017) emphasizes the use of machine learning over 'traditional' statistical methods, because they generally test predictors in isolation and perform additive or multiplicative operations, as opposed to complex combinations of different factors. Both Nelson et al. (2016) and Walsh et al. (2017) opted for decision tree based methods and analysed feature importance as a way to relate to domain knowledge.

Instances of predicting use of services are commonly conducted in the e-commerce sector, where researchers tried to predict consumer buying behaviour. Similarly, it has also been applied to predict participation in activities such as health check-up schemes (Shimoda et al., 2018) and medical studies (Linden and Yarnold, 2016). As for social participation in the community, Piscopo et al. (2017) looked into predicting social participation in various neighbourhoods using open government data, and Byeon (2019) worked on predicting social participation among South-Korean elderly. Both looked into the importance of various factors such as age, gender, educational level, income, but also neighbourhood characteristics.

This study aims to predict and cast light on human behaviour using machine learning. This means that one important aspect to look out for is the interpretability of the models. Therefore the choice firstly goes to several commonly used supervised machine learning techniques, which are easily interpretable on a modular level (Molnar et al., 2018). This means insights such as feature importance can be extracted from these models. This is critical to understanding how certain results were achieved, the role of each feature (Doshi-Velez and Kim, 2017), and any possible biases in the data. In addition to supervised models, predicting City Pass use means it would also be insightful to analyse what characterizes different groups of users, in respect to unique use, cultural and sport participation. This means it would be valuable to try and identify different subgroups among users who use, and do not use the City Pass, as well as visit or not visit a certain type of location. Cluster analysis would be best suited in this case. This

is a common method used for performing consumer segmentation in the e-commerce sector (Brito et al., 2015).

For this study, data has to be retrieved from several databases, merged, and pre-processed. The data is divided into multiple main sets, one for predicting unique use, and sets for predicting participation in the categories within culture and sport locations. These are tested with three supervised machine learning models, specifically classification models. This means that the models use classification to predict binary outcomes. Two off-the-shelf classifiers from the Skikit library by Pedregosa et al. (2011) will be used. This includes Logistic Regression, Random Forest. XGBoost from Chen and Guestrin (2016) will also be used. These models are tested using demographic, neighbourhood, geographic data (i.e. travel distance) and location data. K-prototypes will be used for clustering users based on demographic data.

2 Methodology

Implementation for this project involved multiple steps. The first part consisted of data collection, exploration and preprocessing in order to build the data sets for supervised and unsupervised learning, namely prediction and clustering. This is illustrated by figure 1. The second part consisted of application of the models, validation, assessing feature importance scores and evaluation. The programming language of choice for this project was Python (3.6).

Three classifiers were implemented for prediction, two are part of the Scikit-learn library from Pedregosa et al. (2011). These include Logistic Regression, Random Forest. The Scikit-learn wrapper API for XGBoost was also used (Chen and Guestrin, 2016). A simple zero rule classifier by Brownlee (2016) was implemented for baseline comparison. It classifies test data based on the label distribution of the train data, basically performing classification using the most frequently occurring label. Furthermore, K-prototypes based on work from Huang (1997) was used to perform clustering on mixed categorical and numerical demographic data in order to find subgroups of users.

2.1 Data and procedure

The data was delivered by the Municipality of Amsterdam. Parts of the data were already present in their databases, and had to be collected for this study. Two main data sets were provided: one containing data of all City Pass users, and one containing transaction data of the City Pass. For predicting unique use of the City Pass, a data set containing all users and non-users was created. Non-users were defined as users who own a City Pass, but have not used it in the defined period of time. This period constituted 2018, since there was data available over this entire year. For predicting participation in different types of culture and sport locations, only locations with offers available for all ages were used. This meant excluding for example child specific offers.

For predicting cultural and sport participation, the data was split into separate sets for predicting the binary outcomes *visit* or *no visit* of museum, cinema, theatre and sport locations. The outcome was based on whether a user visited a location at least once or not. The subdivision into different categories was done because interest lies primarily in finding the determining factors of participation for each type or category of activity. The assumption was that these factors might vary per type. Similarly, this process was also repeated for clustering. For clustering, only a subset of demographic data was used because the goal was to identify subgroups based on user level characteristics.

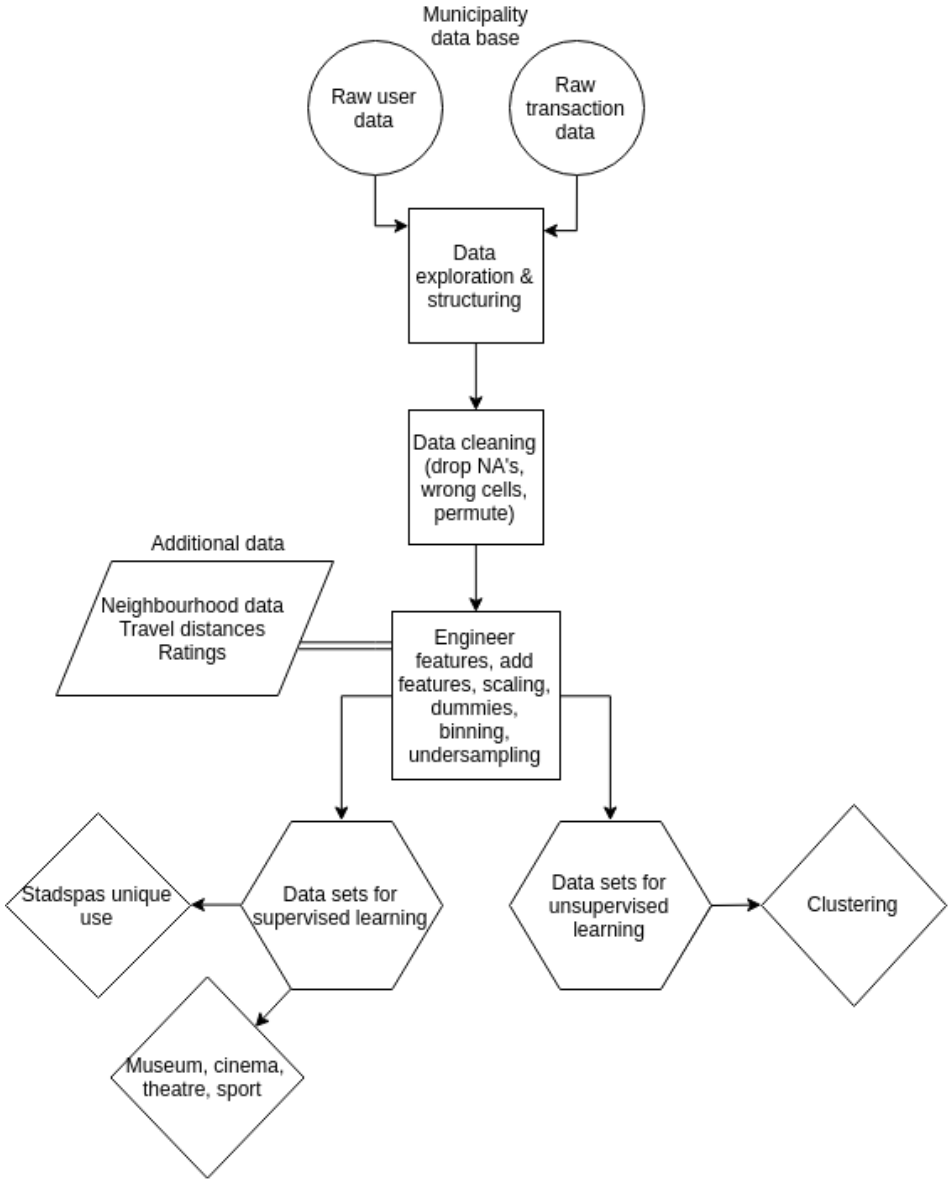


Figure 1: Data preprocessing steps.

2.1.1 Privacy

Before handling the data, a Privacy Impact Assessment was performed to assess the goal, nature and implications of this study. This study was regarded as a research pilot, and as such covered under the privacy statement of the Municipality of Amsterdam, which states that personal information can be used for research if it serves a common interest for its citizens. The data has been minimized and pseudonymised conform to the GDPR, and will only be used for the purpose of this project, as stated in the Privacy Impact Assessment. Minimized means that only data that was necessary for this study

was used. The pseudonymisation steps involved removing personal information such as names, house numbers and citizen service numbers. A virtual Safe Work Environment was set up to perform analysis on the data. This is a closed system that allows only specific network traffic and functions as a safety framework for the data. Any results in this study were featured on an aggregated level, no data of individuals was published.

2.2 Preprocessing

The following sections describe the different parts of data and steps used for creating the data sets for prediction and clustering.

2.2.1 User data

The delivered data set encompasses all City Pass users, 201.657 in total. Only data from low-income users from Amsterdam was used. Regular 65+ users were excluded, because they form a separate smaller group for which only a small subset of City Pass locations are available. After cleaning up the data, the data totalled 117.962 low-income City Pass users, of which in 2018 a total of 55.498 used the City Pass. This number was based on a match with unique users from the transaction data. Socioeconomic variables such as *income level* and *type of income* were also included, however since both were only present for 6% of the data, these were omitted. For the variable *gender*, the category 'unknown' was present for a small subset (n=10) of users. These users were omitted in order to keep the level of categories minimal. Non-existing or assumed incorrect zip codes were also omitted. NA's for *educational level*, based on highest completed level, were replaced with "Not (yet) known". These were present mostly for children, and missing at random for adults. There were 'Adult' labels assigned to children due to lack of registered guardian according to the data, these were corrected for all users below age 18. The data also includes whether a user applied for any Poverty Reduction programme. Detailed in table 1 is the derived set of features from this data set.

There were initially 25 levels within the educational level category. To decrease the space and improve interpretability, educational levels were categorised according to the CBS grouping (Centraal Bureau voor de Statistiek, 2019), which distinguishes Low, Middle, and High. There was also a 'Not (yet) known' level and 'Not applicable', these were left as unique levels, yielding a total of 5 levels. The originally present variable *household type* turned out not to be representative, so a new one was created based on encoding of partners and children in the data, together with household size. A user had a separate encoding for a partner included in the partner column, and children did so with a caretaker column. Caretaker was not necessarily equal to the parent label found in the *relation type* variable from the original data set. Nor do 'children' have to be below or equal to the age of 18. The relation type variable was encoded into separate *is parent* and *is child* features. Use of four different Poverty Reduction programmes part of the municipality social services was also included in predicting unique City Pass use.

Variable	Type	Description	Range
Age	Numerical	Age of user (years)	0 - 102
Educational level	Categorical	Highest completed education level	5
Gender	Binary	Gender of user: 1: f, 0: m	
Is parent	Binary	User is a parent	
Is child	Binary	User is a child	
Household type	Categorical	Type of household user belongs to	4
Household size	Numerical	Size of household	1 - 12
Schoolchildren's Allowance	Binary	Yearly extra allowance for children	
PC Provision	Binary	Free laptop or tablet for children	
Free Public Transport	Binary	Free public transport for seniors	
Individual Income Support	Binary	Extra income support	

Table 1: User data variables. Range and levels included for all variables.

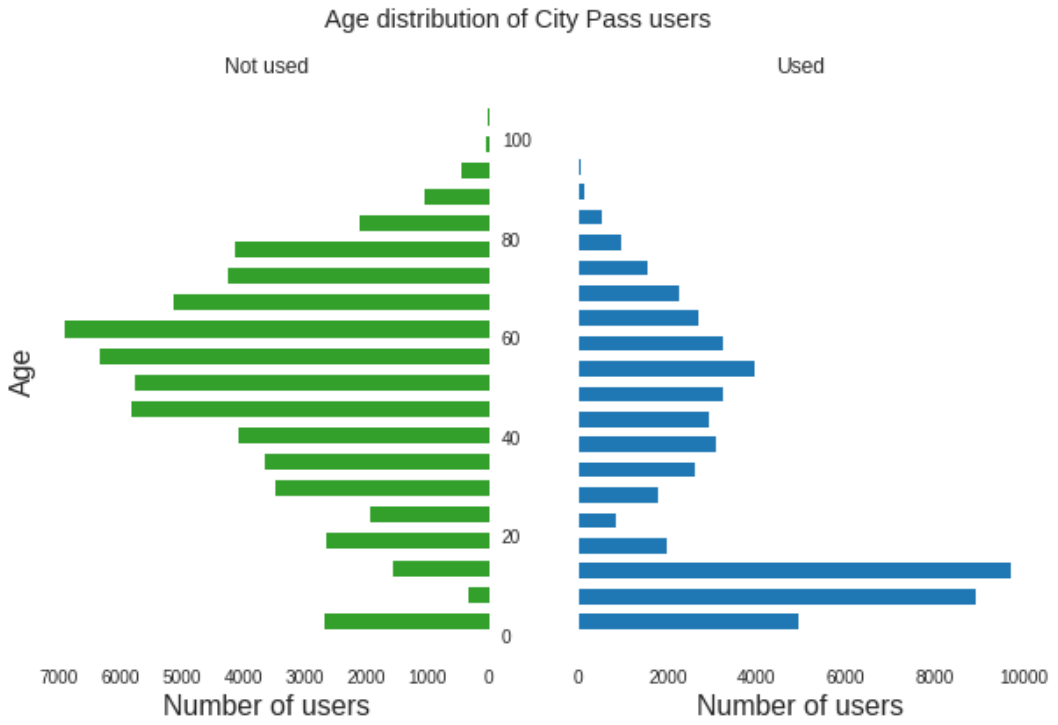


Figure 2: City Pass user population pyramid distribution by age and use.

Figure 3 provides a small summary on the categorical demographic variables of City Pass users, featuring use and no use in 2018. Percentages are in proportion in to all users instead of per group of users and non-users. The distribution of all City Pass users by use and non-use with age is shown in figure 2.

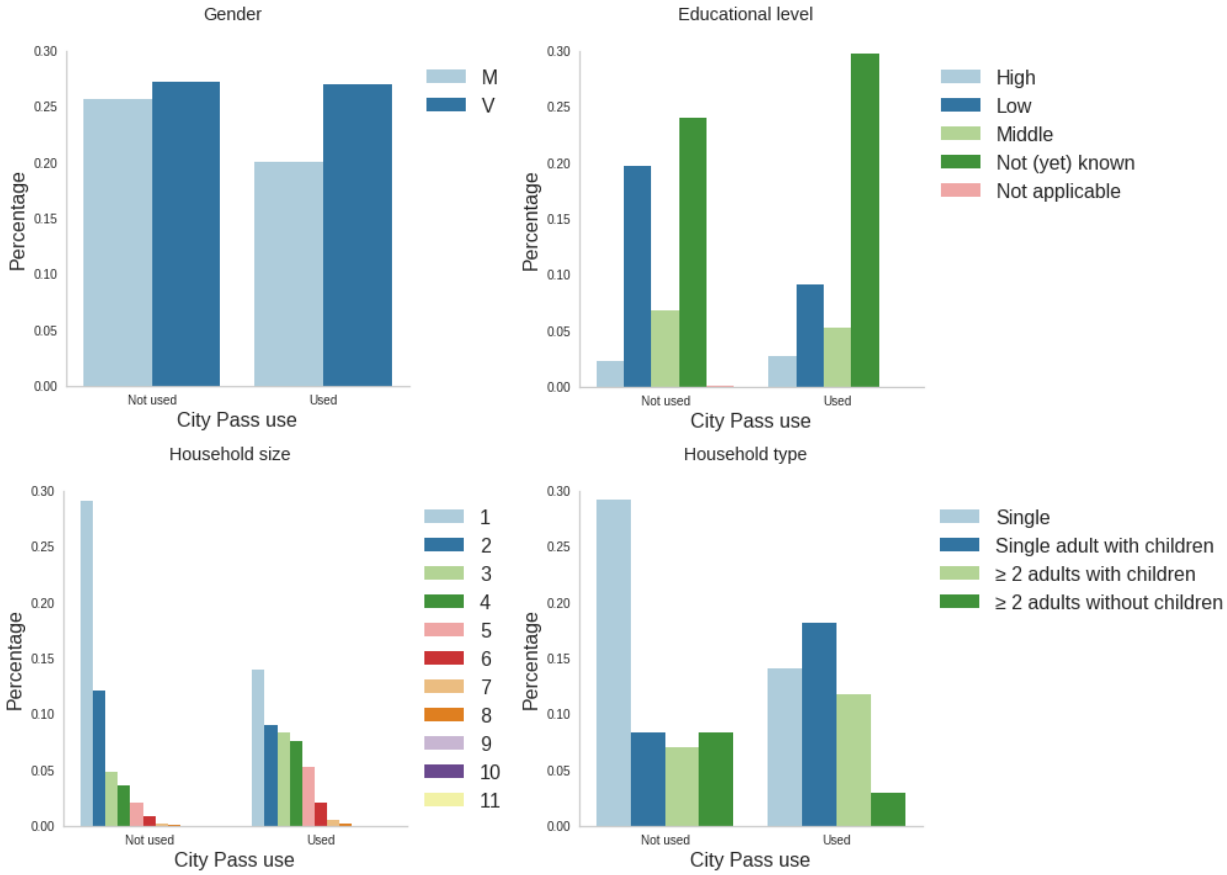


Figure 3: Summary of demographic variables for users and non-users of the City Pass in 2018. Percentages are per group in proportion in to all users.

2.2.2 Transaction data

Data from calendar year 2018 was requested, this meant data containing all transactions between September 2017 and April 2019 was delivered. This is because City Pass subscriptions start in September. For prediction, this set was subsetted to instances from 2018. This was then subsetted to data containing City Pass partners which were available to all age ranges. For the prediction data sets, duplicate rows were removed, as well as rows of users who are currently not residents of Amsterdam, since the City Pass is officially targeted at citizens of Amsterdam and the available neighbourhood data does not extend beyond Amsterdam. Rows with City Pass partners without an address were removed because no travel distances can be calculated for these instances. This yields a total of 295.002 transactions, 774 uniquely used offers, from 300 different City Pass partners. There were 6 main categories according to the data: Culture, Recreation, Shops & Restaurants, Sport, Education and Budget Third Parties. Only Culture and Sport was used. Detailed in table 2 is the complete set of features derived from this data set, together with additional features, explained in the next sections.

The main category 'Culture' was split into subcategories Museum, Cinema and Theatre. Sport was kept as single main category. Missing ratings or number of ratings were imputed using the mean. The mean discount, price before discount and after discount were calculated per City Pass location, based on their available offers.

Variable	Type	Description	Range
Mean discount	Numerical	Mean discount (euro)	0 - 96,06
Mean price before discount	Numerical	Mean price before discount (euro)	0 - 99,5
Mean price after discount	Numerical	Mean price after discount (euro)	0 - 20
Travel distance	Numerical	Distance from user to location (m)	
Rating	Numerical	Total online rating	0 - 5
Number of ratings	Numerical	Total online ratings	1 - 41.845

Table 2: Location related transaction data variables and additional variables, with range and levels. Locations: museum, cinema, theatre and sport.

2.2.3 Neighbourhood data

This additional set of variables for prediction consisted of neighbourhood data. Zip codes of users were matched with neighbourhood data from the Basisbestand Gebieden Amsterdam (Onderzoek, Informatie en Statistiek, 2019). The pre-selected variables are detailed below in table 3. These were selected based on availability and assumed effects. Variables were taken from either 2017 or 2016 depending on availability. Some variables were registered on zip code level, while others were on neighbourhood level. If possible, the first attempt was retrieval at neighbourhood level, otherwise at zipcode level. In case of missing data, the data was imputed with the mean value. One additional variable, the number of City Pass locations, was created by grouping the City Pass locations by neighbourhood, based on the available locations from the transaction data set.

Variable	Type	Description	Range
Safety index	Numerical	Measure of safety	53 - 186
Social cohesion	Numerical	Social cohesion	4.6 - 7.4
Population density	Numerical	Number of citizens per km^2	10 - 28.236
Western population %	Numerical	Population of Western origin (percentage)	8.9 - 53.7
65+ population %	Numerical	Senior population (percentage)	1.4 - 29.1
Families with children %	Numerical	Households with children (percentage)	2.5 - 51.7
Average residence duration	Numerical	Residence at address (years)	1.1 - 15.6
Working population	Numerical	People in workforce per 1000	79 - 158.517
Highly educated %	Numerical	Highly educated population (percentage)	13 - 68
Low income population %	Numerical	Population with $<$ 19.800 income (percentage)	5 - 45
Number of City Pass locations	Numerical	Total City Pass locations per neighbourhood	0 - 19

Table 3: Neighbourhood variables, with range within data set according to the users.

2.2.4 Travel distance

Additional features for predicting cultural and sport participation included travel distance. This refers to the distance between home addresses of users and various City Pass locations. Addresses were encoded as zip codes: the first 4 numbers and 2 letters. The first step was to convert zip codes to latitude and longitude coordinates. This was done with a dictionary of all zip codes from Amsterdam and their approximate corresponding coordinates, created by using the Basisregistratie adressen en gebouwen (BAG) Amsterdam (2018). This was done instead of using an API due to request constraints. A dictionary proved more suitable for fast conversion. It must be noted that the dictionary contained approximate coordinates, because coordinates were registered per unique address in the BAG, so the first corresponding coordinate per unique zip code was used. In case of City Pass partners with multiple locations, a manually calculated coordinate midpoint was used.

The second step was applying the Python package *OSMnx* by Boeing (2017), which produces a directed graph object of the car network of Amsterdam. Using this, travel distances between two points could be calculated with the Dijkstra algorithm, part of the *NetworkX* python package by Hagberg et al. (2008). Dijkstra provides the optimal travel distance. This method is more computationally intensive than using Euclidean distances, but yields more precise distance measures. Since the directed graph contains only coordinates of edges and nodes, travel distance was calculated by first determining the nearest nodes to the points between which the actual distance was to be calculated. The car network was used because various modalities such as bicycle, walking, public transport, were likely not matter much in general lines. Most travel usually occurs along car roads. For distances to locations outside Amsterdam, and distances within Amsterdam which could not be computed using the Networkx method, Euclidean distance was used.

2.2.5 Online ratings

The last set of features for predicting participation in activities consists of online ratings of visited locations. These are based on a 5 star scale. Additionally, the number of ratings were included as well, used as reflection of popularity of a location. If a location was lacking Google ratings, ratings from Facebook or any immediately available website was used as replacement. This was always taken from one source to account for possible double user reviews. If no ratings could be found at all, missing values were imputed with mean values.

2.2.6 Final prediction data

For prediction, two groups of data were formed: one for predicting unique use of the City Pass, and one for predicting cultural and sport participation. For predicting unique use, a combination of user data and neighbourhood data was used, as featured in tables

1 and 9. For predicting cultural and sport participation, the transaction data had to be merged with the user data, and neighbourhood data, travel distance and online ratings were added. These formed the sets for prediction of cultural and sport participation. The data was grouped by unique visit of City Pass partner, so on location level. This was done in order to decrease the sample size, because grouping by offers would have lead to a considerably larger space and therefore longer computation times. The data showed a strong class imbalance, i.e. the target class usually consisted of approximately 1% of the data. Therefore random undersampling of the majority class was performed. A 1.2:1 instead of 1:1 ratio was used so not too much information from the majority class was lost. This ratio denotes the proportion between the two binary classes 0 and 1. Another motivation behind undersampling was decreasing run-times, mostly for calculating travel distances between users and locations. This can be quite costly when calculating Dijkstra distances over a directed graph. Each row in this data set represents an unique visit of a City Pass location by an user. This finally resulted in four different data sets: museum, cinema, theatre and sport.

Before final input into the classification models, numerical data was standardized in order to create a similar range, and n levels of categorical variables were one-hot-encoded into n separate binary variables.

2.2.7 Final clustering data

For clustering, only demographic data was used. Similar steps were performed, on both user and transaction data, except age was binned. This was done because K-prototypes also takes categorical variables as input, and it leads to easier interpretation for this study. First three bins were 0-4, 4-12, 12-18, based on eligibility to go to primary (Rijksoverheid) and high school (Nationale Onderwijsgids). For adults these were 18-25, 25-35, 35-45, 55-65 and 65+, binned until pensioning age. For the users who visited or did not visit any location category, rows containing partners with no addresses were not dropped since travel distance is not part of the data. This part was grouped on category level, meaning if any category was visited, instead of location. The transaction data was matched with the user data, therefore distinguishing users who did or did not visit any location in the category. The complete overview of used variables per set can be found in table 4.

	Prediction		**Clustering**	
Variable	Unique use	Categories	Unique use	Categories
Age	✓	✓	✓	✓
Educational level	✓	✓	✓	✓
Gender	✓	✓	✓	✓
Is parent	✓	✓	✓	✓
Is child	✓	✓	✓	✓
Household type	✓	✓	✓	✓
Household size	✓	✓	✓	✓
Schoolchildren's Allowance	✓			
PC Provision	✓			
Free Public Transport	✓			
Individual Income Support	✓			
Safety index	✓	✓		
Social cohesion	✓	✓		
Population density	✓	✓		
Western population %	✓	✓		
65+ population %	✓	✓		
Families with children %	✓	✓		
Average residence duration	✓	✓		
Working population	✓	✓		
Highly educated %	✓	✓		
Low income population %	✓	✓		
Number of City Pass locations	✓	✓		
Mean discount		✓		
Mean price before discount		✓		
Mean price after discount		✓		
Travel distance		✓		
Rating		✓		
Number of ratings		✓		

Table 4: Complete overview of used variables per task and data set. Categories refers to museum, cinema, theatre and sport participation.

2.3 Models

This section elaborates on the modeling part of the implementation. This includes model validation, checking feature importance scores and evaluation. The classification models were chosen due easy interpretability. Data was split into a train and hold out test set according to a 80:20 ratio. Performance was assessed on the test split. N-fold cross validation was not used because this would mean each fold yields a newly trained model with different feature importance scores. Validation consists of hyperparameter tuning, this was done with 5-fold cross validation on the train set. This means multiple splits of the train and validation sets were generated to tune the parameters on. The models were validated and compared using standard performance metrics. To detect possible problems in fit, learning curves with accuracy as metric were used as a diagnostic tool. Feature importance retrieval was based on the model native scores, detailed in the each model specific section. Permutation importance as used as second metric. Permutation importance checks for decrease in accuracy when a variable is not available as measure of importance (Mikhail and Konstantin, 2017). This metric was used because it provides additional insight into the model's behaviour, and makes it easier to compare between models.

Evaluation metrics included the commonly used accuracy measure, in addition to precision, recall, F1, ROC AUC score. The first four are simply defined as following:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Accuracy measures the proportion of correctly identified labels.

$$Precision = \frac{TP}{TP+FP}$$

Precision measures the proportion of truly identified labels being correct.

$$Recall = \frac{TP}{TP+FN}$$

Recall measures the proportion of actual true labels being correct.

$$F1 = \frac{2 \cdot precision \cdot recall}{precision+recall}$$

F1 is the harmonic mean of precision and recall (Powers, 2011).

With TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives. Precision and recall were reported besides F1, because any large variations between these two scores can reflect specific performance problems. The ROC AUC is the Area Under the Receiver Operating Characteristic Curve, which indicates how well a model distinguishes between two classes, independently of class distribution. F1 and

ROC AUC are useful metrics when dealing with class imbalance (Sokolova et al., 2006).

McNemar's test was used to assess significance of the difference between the predictions of two classifiers. This is a non-parametric test that checks whether disagreements between two groups match and shows the significance of the relative difference in the proportion of error between two groups on the same test data (Dietterich, 1998). This test is well suited when predictions were made on the same test set, and can be applied to binary classification problems (McCrum-Gardner, 2008).

For the interpretation of feature importance, multicolinearity among features had to be taken into account. This is a phenomenon where two or more variables correlate with each other, meaning one feature might be redundant. This does not have to affect performance of the models, but can skew feature importance scores. One feature might turn out more important while the other does not, even if they are highly correlated. It can distort the coefficients in Logistic Regression. Ridge Regression regularization in Logistic Regression can minimize with this problem according to Duzan and Shariff (2015). Decision tree based ensemble classifiers such as Random Forest and XGBoost do not make relational assumptions between features like Logistic Regression. However, when looking at the feature importance scores of decision tree classifiers, it can occur that only one of two correlated features gets chosen during the learning process, causing the other feature to become less important. It can be difficult to gauge the relative importance of different features if they have a similar effect.

To detect multicolinearity, the variance inflation factor, also known as VIF, was used. It indicates the effect on the variance of the estimated regression coefficient due to multicolinearity. The VIF score is calculated by performing regression of one feature against another. For the particular data sets of this study, it only works well on continuous features, because there are one-hot-encoded binary features, and regression requires one encoded categorical feature to be dropped. A commonly used threshold is a VIF score of 10, larger scores imply high multicolinearity (O'brien, 2007). Correlation matrices using Pearson correlation were used to identify specific colinear features if a score larger than 10 was detected.

Partial dependence plots were used for Random Forest and XGBoost to check for the direction of the predictions. They show the relationship between the predicted target and given values of a feature, under the assumption that the feature is not correlated with others (Molnar et al., 2018). No partial dependence plots were used for Logistic Regression because the direction of its effect is already indicated by the sign of logits, and a PDP for Logistic Regression will only show linear relationship because it is constructed based on the logits.

For all classification models, the *class_weight* parameter was set at 'balanced' to minimize the impact of imbalance in the data. Table 5 shows the different data set sizes used as input.

		Size
	Unique use	117.962
Categories	Museum	30.472
	Cinema	20.713
	Theatre	12.434
	Sport	45.179

Table 5: Data set sizes for classifier input.

2.3.1 Logistic Regression

Logistic regression was developed by Walker and Duncan (1967) and is an extension of linear regression, but instead performs classification and uses a sigmoid function. The assumption is that in the input space, the data can be separated by a function. This function is defined below as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Logistic regression works by estimating the *coefficients*, also known as weights of features. This is done using maximum-likelihood estimation. Furthermore, corresponding log odds, also known as logits, and the standard errors of these weights can be retrieved per feature. The ratios of these logits refer to the effect of a feature on the probability of an event, specifically the probability of the event divided by the probability of $\neg event$. Probability outputs are between 0 and 1, calculated by inputting the estimated weights in the sigmoid function. The value cutoff point for classification as 1 is ≥ 0.5 (James et al., 2014). Logistic Regression logits can be positive or negative. For continuous features, a positive sign implies a positive relationship, meaning that when every feature remains the same, there is an increase in chance of an event occurring if this features increases in value. The logit value corresponds to the effect of one unit of change in a feature. In case of binary features, the probability increases in presence ($= 1$) of this feature, when everything else remains the same. The other way around applies for negative signs.

This model was implemented with built-in 5-fold cross validation to find the best hyperparameter C, which is the regularization parameter used to prevent overfitting of the model. For maximum-likelihood, Stochastic Average Gradient descent was used as solver. L2, also know as Ridge Regression, was used for regularization. The logits were used to gauge feature importance scores for this model. Permutation importance part of the ELI5 package was used as a second measure, this works by measurement of the accuracy score by leaving out features Mikhail and Konstantin (2017). This was used in order to know how much a model relies on a feature for making a prediction. Since

the scikit-learn implementation lacked an option for outputting standard errors and P values together with the logits, additional code was added to achieve these results.

2.3.2 Random Forest

Random Forest is an ensemble learning variant of Decision Tree learning. It was developed by Breiman (2001). In Decision Tree learning, a single 'tree' is constructed from the data, and splits on features which leads to subsets of data, also know as *nodes*. These nodes predominantly contain cases of one class. This happens by measuring the *impurity* of each 'current' node and calculating the quality of potential nodes, also known as impurity reduction, compared to the 'previous' node. Impurity refers to the fraction or relative frequencies of the classes in that node. It is the reduction of variance in a node. If a node is completely 'pure', it becomes a leaf Russell and Norvig (2010).

Contrary to single trees, Random Forest creates multiple trees and does the following: when determining the best split in a tree, it first randomly selects a subset of the features and then determines the best split on this random subset of features. The best split is based on the highest impurity reduction. This is done until the predetermined maximum depth is met, which finally leads to a binary label as output. Any set number of random trees can be created and averaged out, which makes it a more robust learner compared to a single tree.

The standard feature importance function for the Scikit-learn implementation is based on degree of impurity reduction compared to parent node. Permutation importance was also used to assess feature importance scores. The best hyper parameters for this model were found using random search with 5-fold cross validation through a grid of parameter settings, detailed in table 6.

Since standard Random Forest feature importance scores do not provide a direction of a feature in relation to its predictions, whereas Logistic Regression can, we use partial dependence plots to assess the relationship between the features and predictions.

Hyperparameter	Description
n_estimators	number of random trees
max_depth	maximum depth of random tree
min_samples_split	minimum samples required for split
min_samples_leaf	minimum samples required to become leaf node
max_features	maximum features required for split
bootstrap	use of random sampling with replacement

Table 6: Hyperparameters for Random Forest

2.3.3 XGBoost

Like Random Forest, XGBoost is also an ensemble machine learning method using decision trees. It was created by Chen and Guestrin (2016), and works by sequentially

creating trees, where each new tree is build in a way to reduce the error of previous trees. This error is minimized using gradient descent.

By default, XGBoost uses the complete set of all features for each split, then determines the best split based on gain compared the previous tree. Gain is a measurement of how good a tree is, and takes tree complexity into account. This metric is used for determining feature importance. Permutation importance is used as additional comparative metric. The best hyper parameters for this model were found using random search with 5-fold cross validation through a grid of parameter settings. These parameters are detailed in table 7.

Partial independence plots were used to assess the prediction direction of the variables.

Hyperparameter	Description
n_estimators	number of random trees
max_depth	maximum depth of random tree
min_child_weight	minimum sum of instance weight required for child node
gamma	minimum loss required for split
sub_sample	sub sample ratio of training set
colsample_bytree	subsample ratio of columns when creating tree
learning_rate	boosting learning rate

Table 7: Hyperparameters for XGBoost

2.3.4 K-prototypes

K-prototypes is a combination of K-modes and K-means, developed by Huang (1998). It can cluster both categorical and numerical data. This means it can accept and treat binary features as categorical ones. K-modes is an extension of K-means. Like K-means, K-modes clusters data into groups which contain a mean also know as *centroid*. Whereas K-means uses euclidean distance, K-modes uses dissimilarity between data points as measurement. Modes are denoted as vectors of points that minimize the dissimilarities between them and other data points. This is basically the difference in columns between two rows of data. K-modes attempts to minimize the sum of this distance within each cluster.

Clustering was used to identify subgroups among users and non-users of the City Pass and different categories in culture and sport. This was done using descriptive analysis, by providing the top 3 most frequent variable combinations based on the used features. This was done instead of reporting the centroids, with the idea this might provide a better way to characterize each cluster. The optimal number of clusters was evaluated by using an elbow plot. Below in table 8 is the size of each data set used for clustering.

		Used or visited	Not used or visited
	Unique use	62.464	55.498
Categories	Museum	9.768	45.730
	Cinema	6.131	49.367
	Theatre	4.218	51.280
	Sport	15.915	39.583

Table 8: Data set sizes for clustering.

3 Results

The following sections deal with the prediction and clustering results. Classification was used for prediction. Results included evaluation, feature importance scores, identification of multicolinearity and partial dependence. Two different feature importance metrics, feature importance native to the models, and permutation importance. Permutation importance was used to provide more general comparison between models. For clustering, most frequent variable combinations were reported.

During evaluation of each classification model, learning curves based on the training data were plotted. This was done to detect potential problems with regards to fit. The learning curves can be found in appendix B. Logistic Regression showed a minimal gap between train and validation accuracy. Random Forest and XGBoost showed a notable gap between training and validation accuracy at the end of the plot, varying in size, sometimes more than 10%. The direction of validation accuracy did not show convergence yet.

3.1 Prediction: unique use of City Pass

This section shows the results for predicting unique use of the City Pass in 2018 with the three selected models. Featured in table 9 is the performance of all three models. These all scored higher than the Zero Rule baseline. The baseline performed classification using the most frequently occurring label. The scores between accuracy, precision, recall and F1 do not vary much because the imbalance was minor. Too much imbalance can lead to a large number of negative samples, affecting the accuracy and recall score negatively. Precision is not affected by this large number of negative samples, and F1 would not reflect this as strongly either, since it is a harmonic mean of both precision and recall.

Overall, XGBoost had the highest performance across accuracy, precision, recall, F1 and ROC AUC. Random Forest comes in second place, and Logistic Regression third place. McNemar's test was used to show the significance of the relative difference in error proportion between pairs of models. The differences between all possible pairs of models were very significant as shown in table 10.

	Accuracy	Precision	Recall	F1	ROC AUC
0R (baseline)	0.525	0.262	0.5	0.344	0.5
LR	0.701	0.704	0.697	0.697	0.753
RF	0.725	0.734	0.719	0.719	0.791
XGB	0.731	0.755	0.723	0.719	0.799

0R = Zero Rule, LR = Logistic Regression, RF = Random Forest, XGB = XGBoost.

Table 9: Performance metrics for predicting City Pass unique use.

	0R	LR	RF	XGB
0R				
LR	0.0000^{****}			
RF	0.0000^{****}	0.0000^{****}		
XGB	0.0000^{****}	0.0000^{****}	0.0000^{****}	

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, $^{****}p < .0001$

Table 10: Mcnemar's test P values between models.

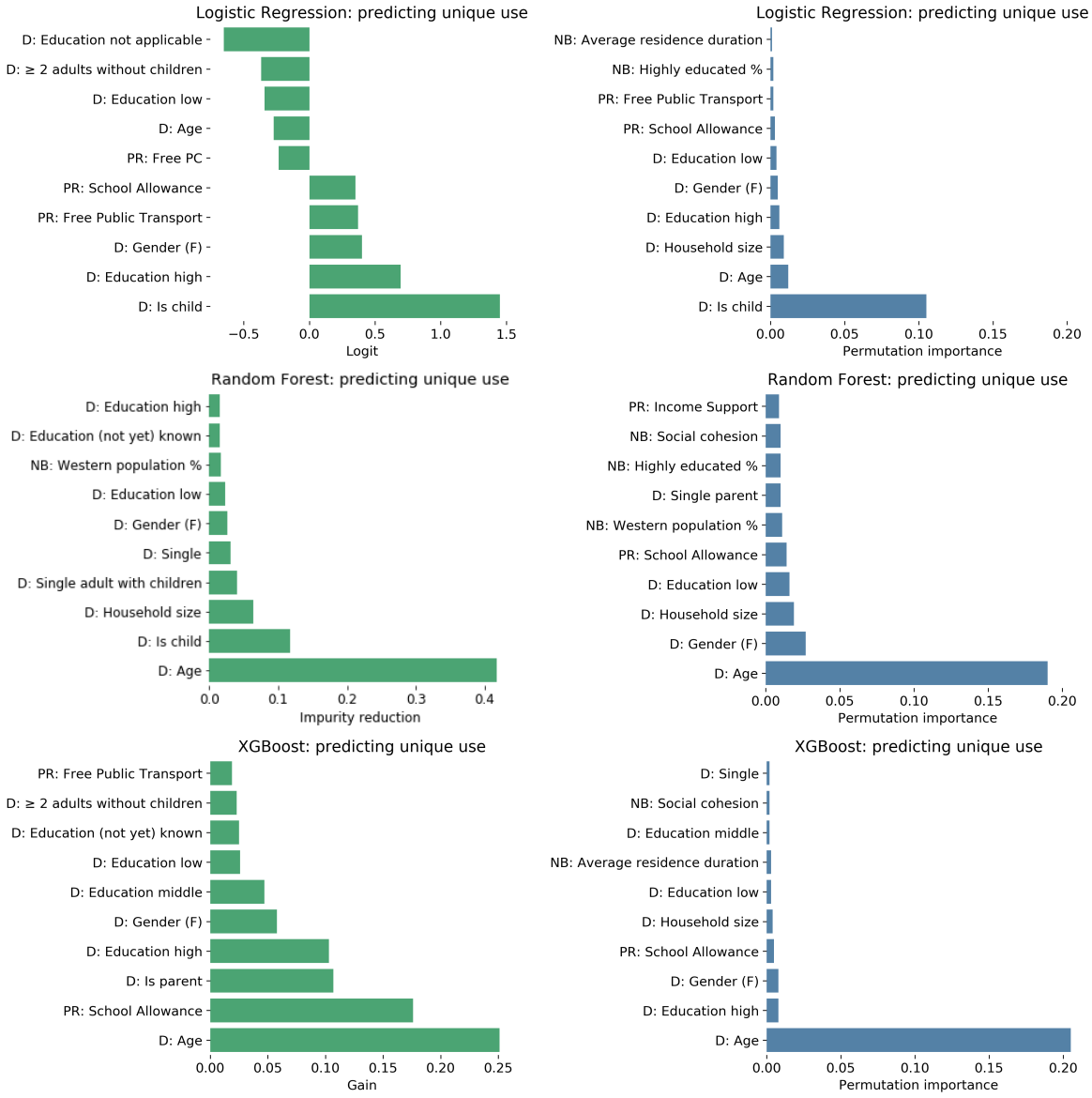


Figure 4: Top 10 feature importance scores for predicting unique City Pass use. Left side: model native feature importance, right side: permutation importance. D = demographic feature, NB: neighbourhood feature, PR: poverty reduction programme use.

Figure 4 displays the top 10 highest ranked feature importance scores from all three models. A top 10 was chosen as a succinct reporting method and arbitrary cut off point. Most scores become smaller and similar around that point. The left side shows the model native feature importance scores. For Logistic Regression, this includes a negative or positive sign, indicating the relationship. Higher feature importance scores indicate that the feature is more closely related to the outcome. On the right side permutation importance scores are shown, which indicate how much a model relies on a feature to make a prediction.

In appendix A1 the complete overview of all features importance scores can be found. A mix of these type of features were present in the top 10 feature importance rankings for the model native feature importance scores. As was the case for permutation importance as metric, although there is a different magnitude of impact observed among features. Different features were also present among the top 10 permutation importance scores. Only 3 out of 10 neighbourhood features used for this study occur in the top 10 of feature importance scores, namely *average residence duration*, *population density*, *Western population %*, and *highly educated %*. All Logistic Regression scores in the top 10 were significant, meaning that these features were significantly associated with the outcome.

Table 11 gives the percentage per feature type across all three model top 10 feature importance scores, grouped per feature importance type. Among the model native and permutation importance scores *age* occurred four times as top feature (i.e. highest score).

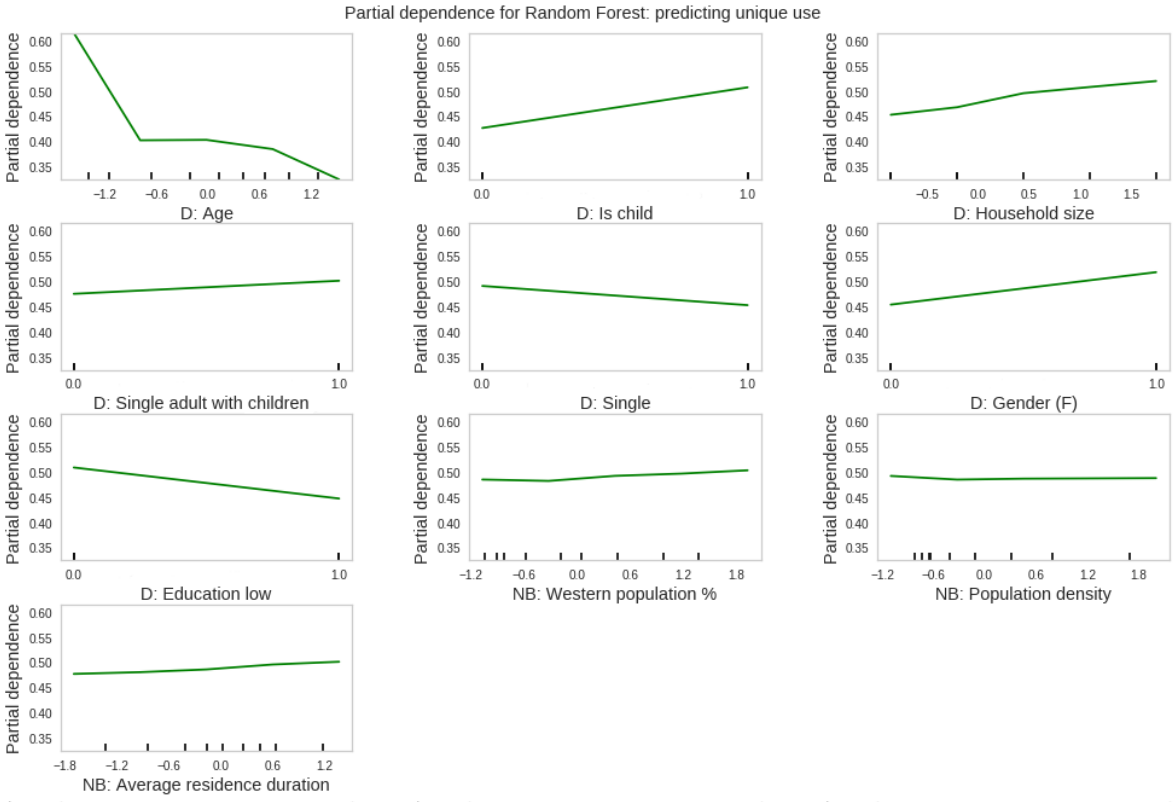
	Demographic	Neighbourhood	Poverty Reduction
Model native feature importance	73.3 %	10%	16.7%
Permutation importance	60%	23.3%	16.7%

Table 11: Percentage share per feature type in all three model top 10 feature importance scores.

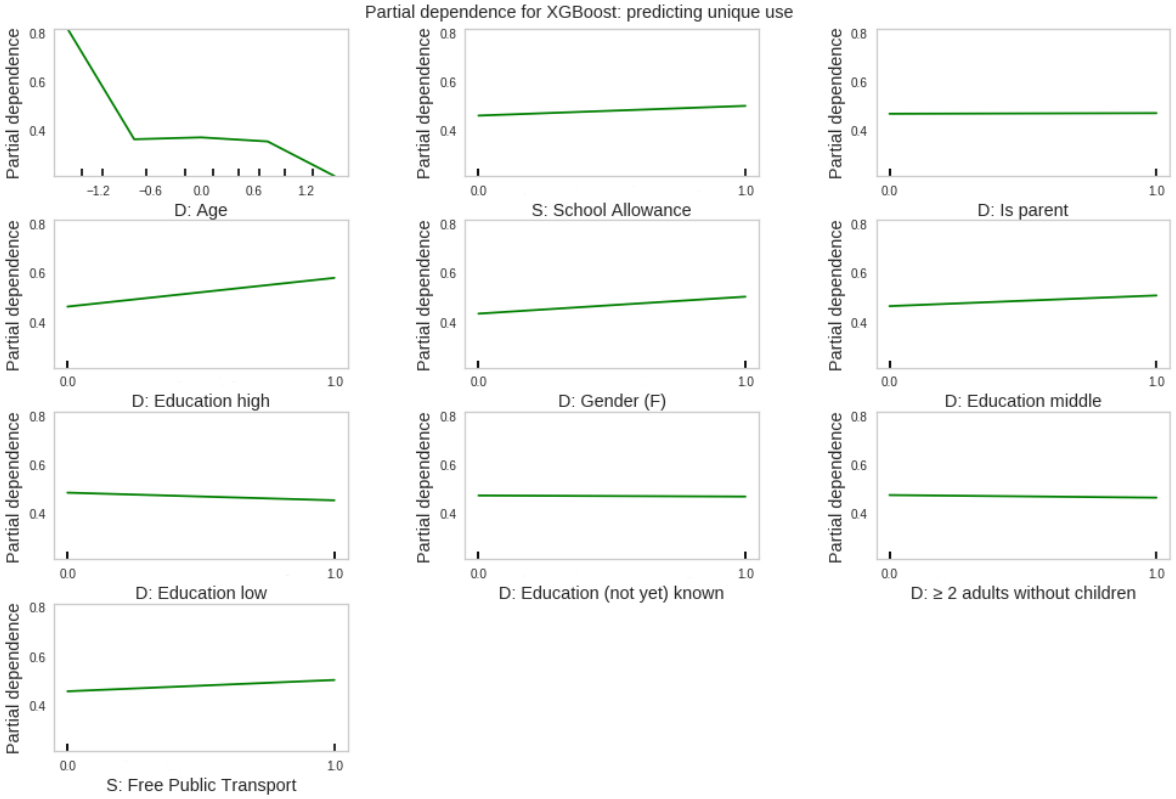
Of the top 10, only the neighbourhood feature *Western population %* has a variance inflation score (VIF) score of > 10. This score was used to detect multicolinearity. Among features with correlation coefficients > 0.7, one was found with a correlation of 0.87 between *Western population %* and *Highly educated %*. Both features can be observed in the top 10 of permutation importance for Random Forest, showing similar magnitudes. Model native and permutation importance scores correspond on the highest ranking feature, being *Is child* for Logistic Regression, and *Age* for Random Forest and XGBoost. Is *child* and *age* have a absolute correlation coefficient of > 0.7. High (multi)colinearity does not have to affect model performance, but it can affect the feature importance scores. One feature might appear to be a stronger predictor compared to other correlated features.

Partial dependence plots (PDP) were plotted based on the top 10 features according to the model native importance metrics. They show the relationship between the

predicted target given the values of a feature. The PDP for Random Forest in 5a shows the standardized values on the X-axis for continuous features. The tick marks on the x-axis represent deciles of the feature values. *Age* was the most common top feature in the top 10, as it had the most notably shaped PDP. The deciles for age were fairly evenly distributed, the PDP first shows a strong decline on the left side of the plot, after which the probability slowly decreases with increasing age. Minor positive trends were present for neighbourhood features *western population %* and *average residence duration*. The strongest relationship for a binary feature was observed for demographic feature *is child*. For *age*, the PDP for XGBoost in table 5b shows a similar continuation as Random Forest. The strongest relationship was found for *education high* in the positive direction.



(a)



(b)

Figure 5: Partial dependence plots for the top 10 features for (a) Random Forest and (b) XGBoost: predicting unique use.

3.2 Prediction: cultural and sport participation

This section details the results of predicting visit of locations across all ages, for City Pass locations available to all ages. These include museum, cinema, theatre and sport visit. The complete feature importance scores can be found in appendix A. Table 12 shows the performance across all models. All models score higher than the Zero Rule baseline. Accuracy, precision, recall and F1 show minor variation for the different models and categories. XGBoost generally scores highest for all metrics, with Random Forest as second, and Logistic Regression as third. The highest score was achieved for predicting sport visit.

As shown in table 13 for museum and cinema visit, the differences between all possible pairs of models were significant. For theatre, only the difference between Logistic Regression and XGBoost was not significant. For sport, the difference between Random Forest and XGBoost was not significant.

		Accuracy	Precision	Recall	F1	ROC AUC
Museum	0R (baseline)	0.554	0.277	0.5	0.357	0.5
	LR	0.697	0.694	0.694	0.694	0.757
	RF	0.811	0.81	0.813	0.81	0.89
	XGB	0.813	0.812	0.815	0.812	0.894
Cinema	0R (baseline)	0.537	0.268	0.5	0.349	0.5
	LR	0.705	0.704	0.705	0.704	0.779
	RF	0.756	0.756	0.757	0.756	0.845
	XGB	0.773	0.771	0.772	0.772	0.857
Theatre	0R (baseline)	0.557	0.278	0.5	0.358	0.5
	LR	0.657	0.655	0.657	0.655	0.715
	RF	0.768	0.77	0.773	0.768	0.846
	XGB	0.766	0.763	0.766	0.764	0.848
Sport	0R (baseline)	0.546	0.273	0.5	0.353	0.5
	LR	0.741	0.742	0.744	0.741	0.802
	RF	0.832	0.831	0.832	0.831	0.909
	XGB	0.834	0.832	0.833	0.833	0.912

0R = Zero Rule, LR = Logistic Regression, RF = Random Forest, XGB = XGBoost.

Table 12: Performance metrics for predicting location visit across all ages.

		OR	LR	RF	XGB
Museum	0R				
	LR	0.0000^{****}			
	RF	0.0000^{****}	0.0000^{****}		
	XGB	0.0000^{****}	0.0000^{****}	0.0022^{**}	
Cinema	0R				
	LR	0.0000^{****}			
	RF	0.0000^{****}	0.0288^{*}		
	XGB	0.0000^{****}	0.0083^{**}	0.0000^{****}	
Theatre	0R				
	LR	0.0000^{****}			
	RF	0.0000^{****}	$.0001^{****}$		
	XGB	0.0000^{****}	.8468	0.0000^{****}	
Sport	0R				
	LR	0.0000^{****}			
	RF	0.0000^{****}	0.0000^{****}		
	XGB	0.0000^{****}	0.0000^{****}	0.0868	

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, $^{****}p < .0001$

Table 13: Mcnemar's test P values between models.

3.2.1 Museum

Figure 6 shows a mix of demographic, neighbourhood, location specific features, and travel distance, were included in the top 10 ranking for the model native feature importance scores. A mix was also found for permutation importance scores. Of the neighbourhood features used in this study, only neighbourhood features *Western population %*, *families with children %* and *highly educated %* occur in the top 10.

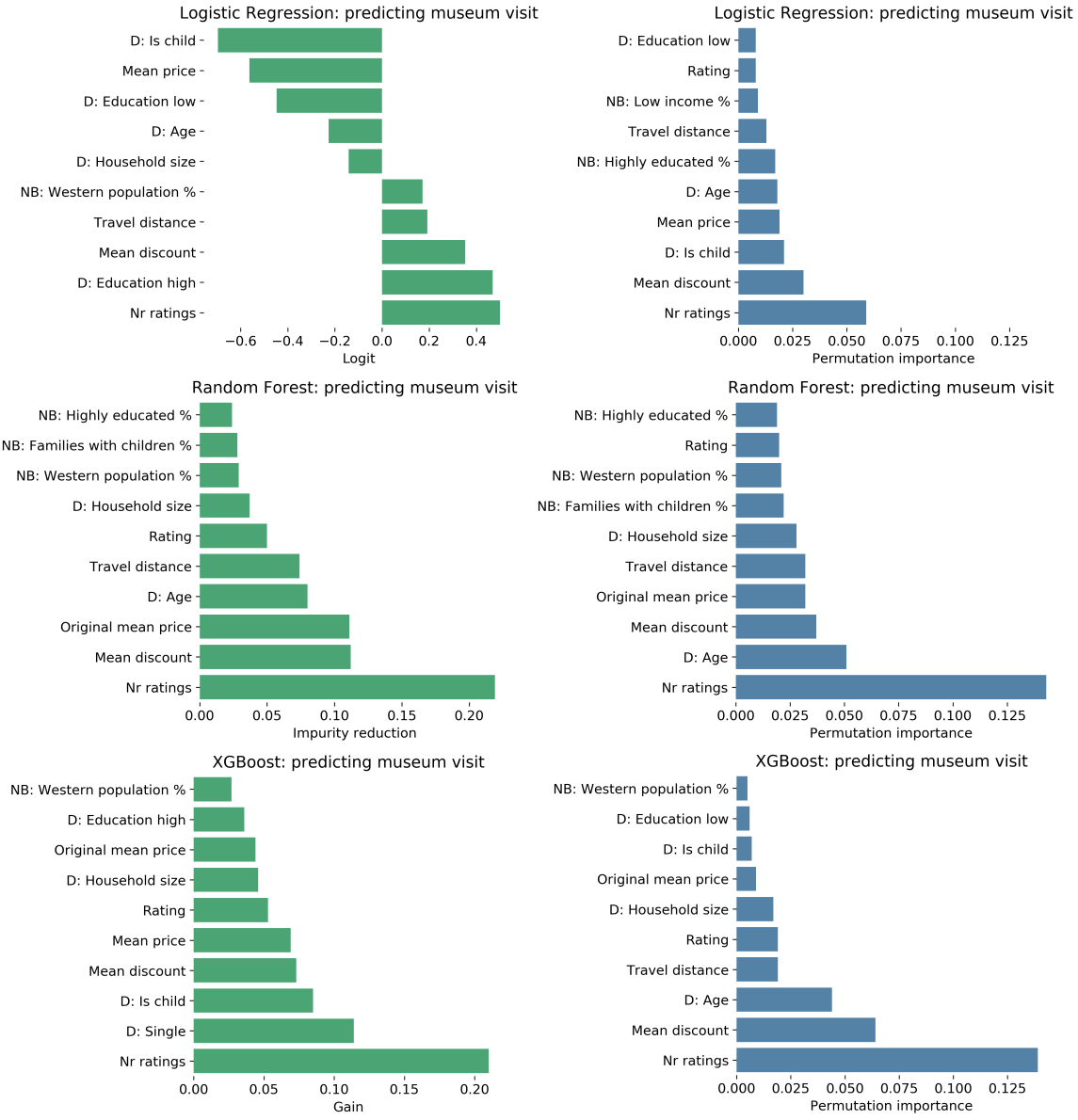


Figure 6: Top 10 feature importance scores for predicting museum visit. Left side: model native feature importance, right side: permutation importance. D = demographic feature, NB: neighbourhood feature.

For the Logistic Regression, all features were significant with $p < .05$. Table 14 gives the percentage per feature type across all three model top 10 feature importance scores, grouped per feature importance type. *Number of ratings* was generally the top feature among both feature importance metrics.

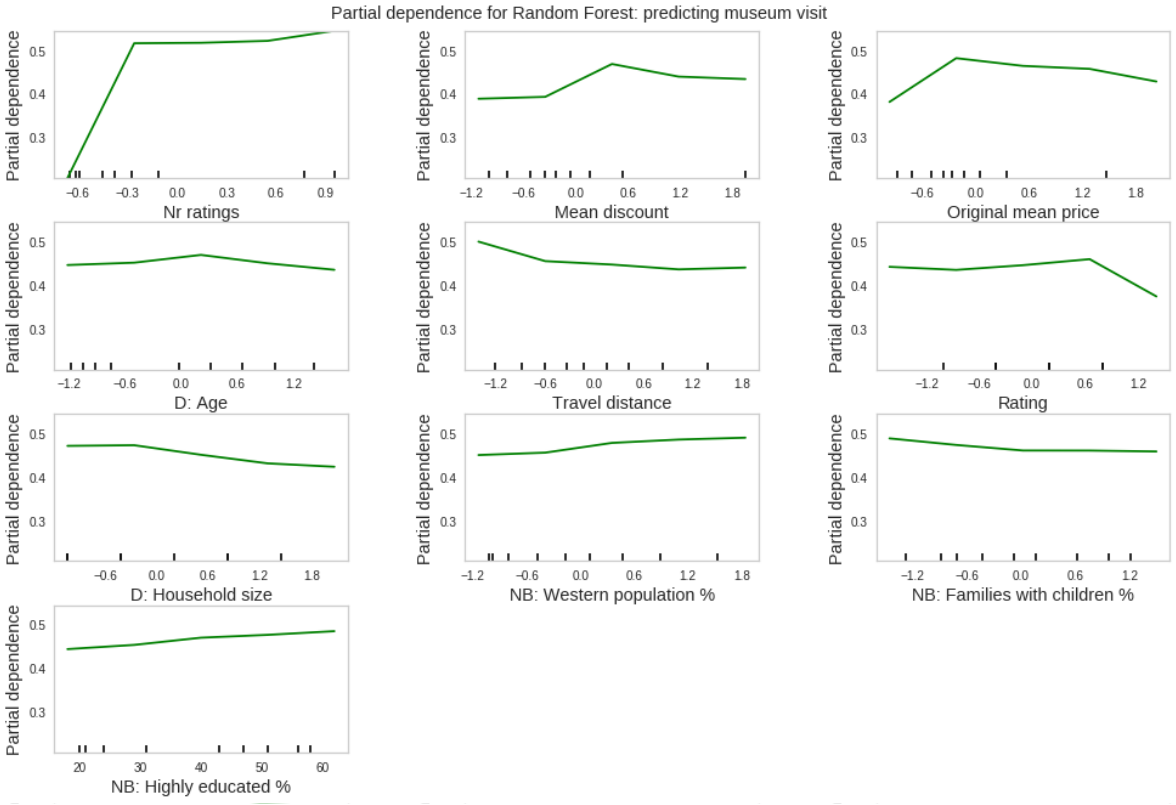
	Demographic	Neighbourhood	Location (including travel distance)
Model native feature importance	40%	13.3%	46.7%
Permutation importance	30%	20%	50%

Table 14: Percentage share per feature type in all three model top 10 feature importance scores.

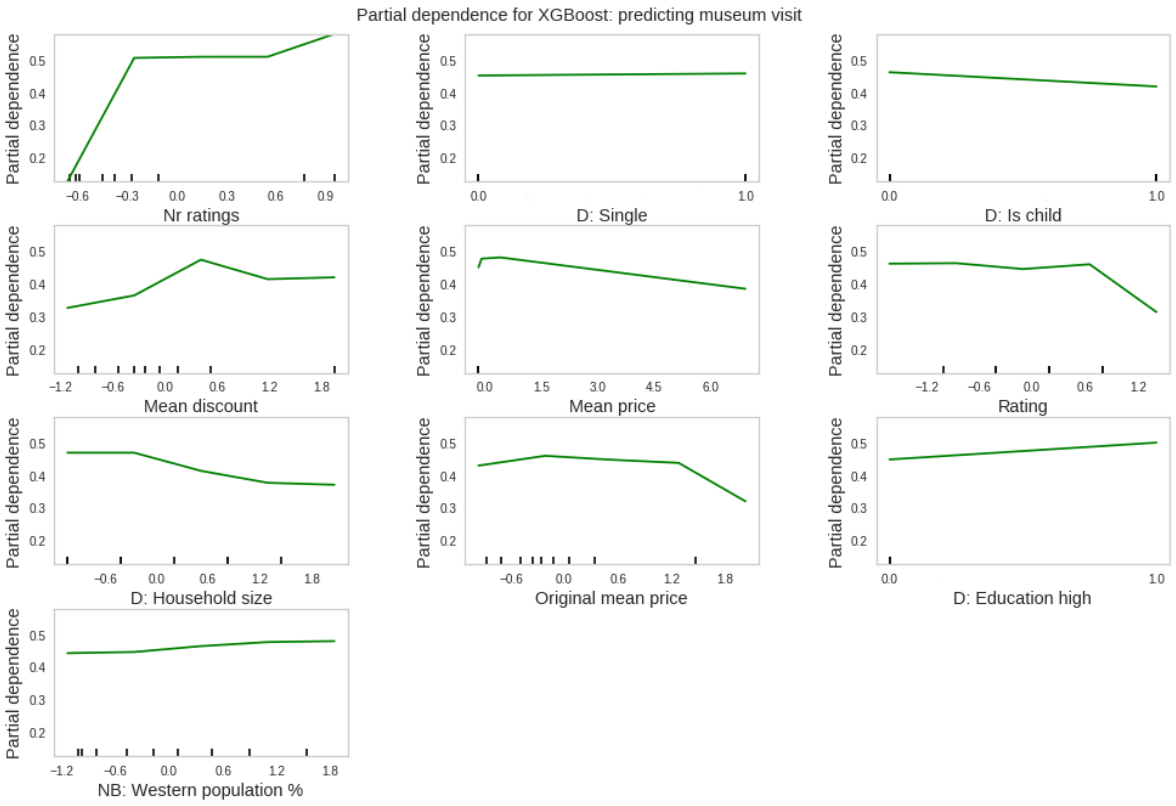
Similar to the predicting unique use, the neighbourhood feature *Western population %* had a VIF score of > 10, indicating a high colinearity. Original mean price, mean discount and mean price also have VIF scores of > 10. Among features with correlation coefficients > 0.7, one was found with coefficient of 0.9 between *mean discount* and *original mean price*. *Western population %* has a correlation coefficient of 0.86 with the neighbourhood feature *Highly educated %*. Both features co-occur in the model native importance and permutation importance top 10 for Random Forest. Model native and permutation importance scores correspond on the highest ranking feature, being *Number of ratings* for all three models.

The PDP for Random Forest in figure 7a shows an uneven distribution for *number of ratings*, *mean discount* and *original price*. These features had fewer data points available in one decile, meaning that the partial dependence estimates can be less reliable in that region. *Number of ratings* was the top feature among the top 10, and had the most notable shape. It can be seen that the probability at first increases as the value increases, but sharply starts to flatten out. *Travel distance* and *household size* show the clearest defined relationships, both in the negative direction. Demographic feature *age* does not show a clear relationship. The neighbourhood features show minor relationships.

Similar observations can be made from 7b. *Mean price* shows one decile tick mark because largely all locations were free to visit. *Household size* shows a more clearly defined relationship in the negative direction.



(a)



(b)

Figure 7: Partial dependence plots for the top 10 features for (a) Random Forest and (b) XGBoost: predicting museum visit.

3.2.2 Cinema

Figure 8 shows a mix of demographic, neighbourhood, location specific features, and travel distance, were included in the top 10 ranking for the model native feature importance scores. A mix was also found for permutation importance scores. For the Logistic Regression scores, all features except for demographic features $\leq$ *2 adults without children* and *single*, were significant.

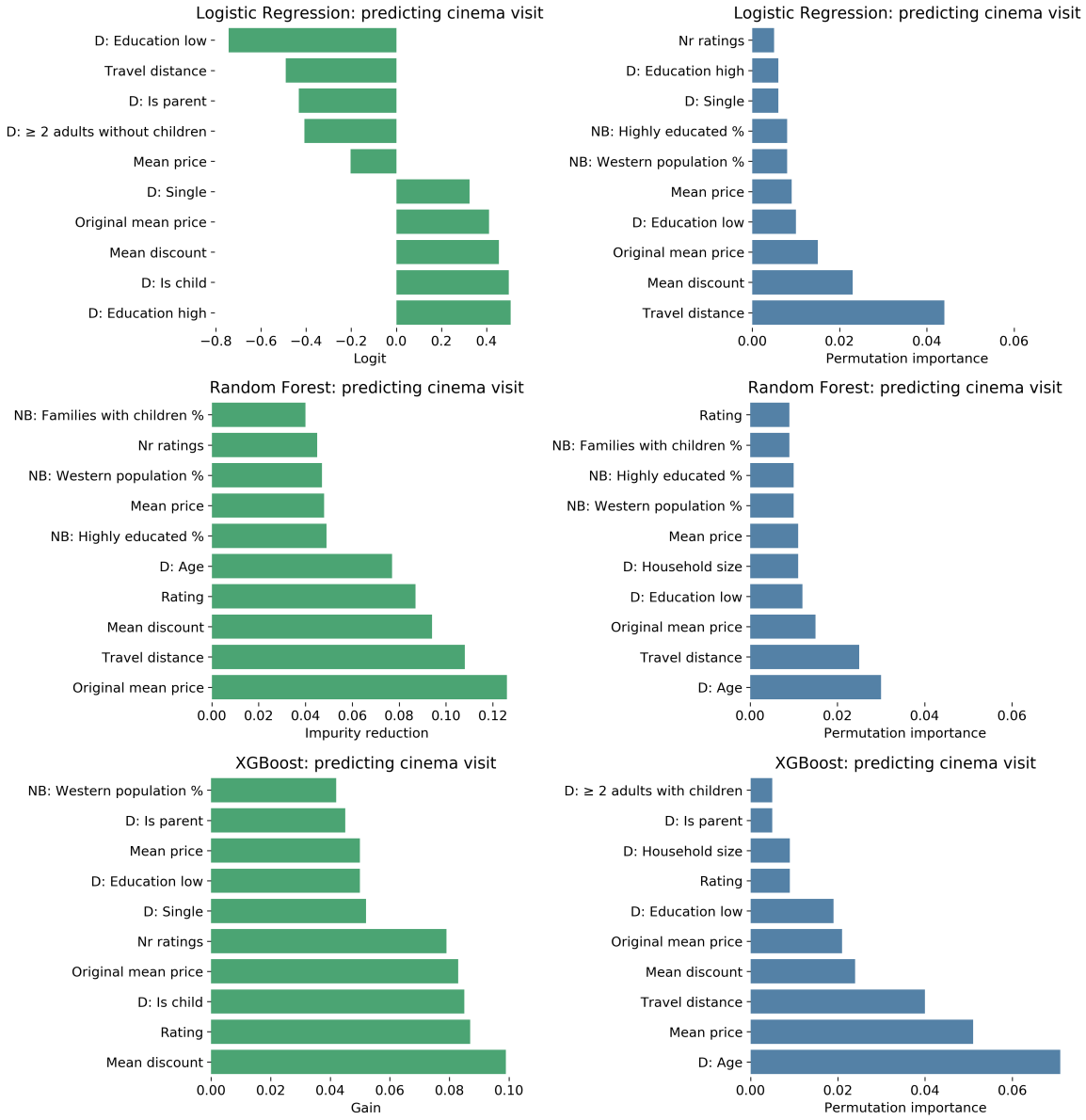


Figure 8: Top 10 feature importance scores for predicting cinema visit. Left side: model native feature importance, right side: permutation importance. D = demographic feature, NB: neighbourhood feature.

Table 15 gives the percentage per feature type across all three model top 10 feature importance scores, grouped per feature importance type. There was no specific top feature among the model native scores. *Age* occurs twice as top feature in the permutation importance rankings.

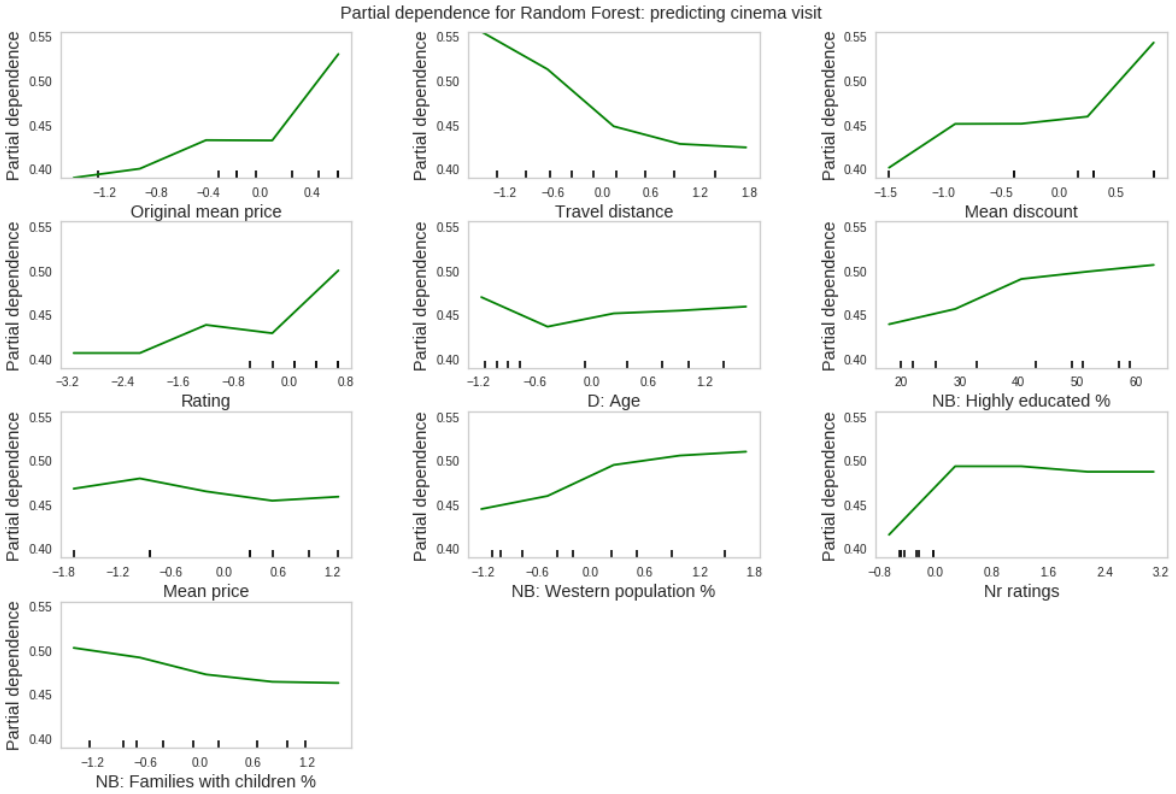
	Demographic	Neighbourhood	Location (including travel distance)
Model native feature importance	36.6%	13.3%	50%
Permutation importance	30%	20%	50%

Table 15: Percentage share per feature type in all three model top 10 feature importance scores.

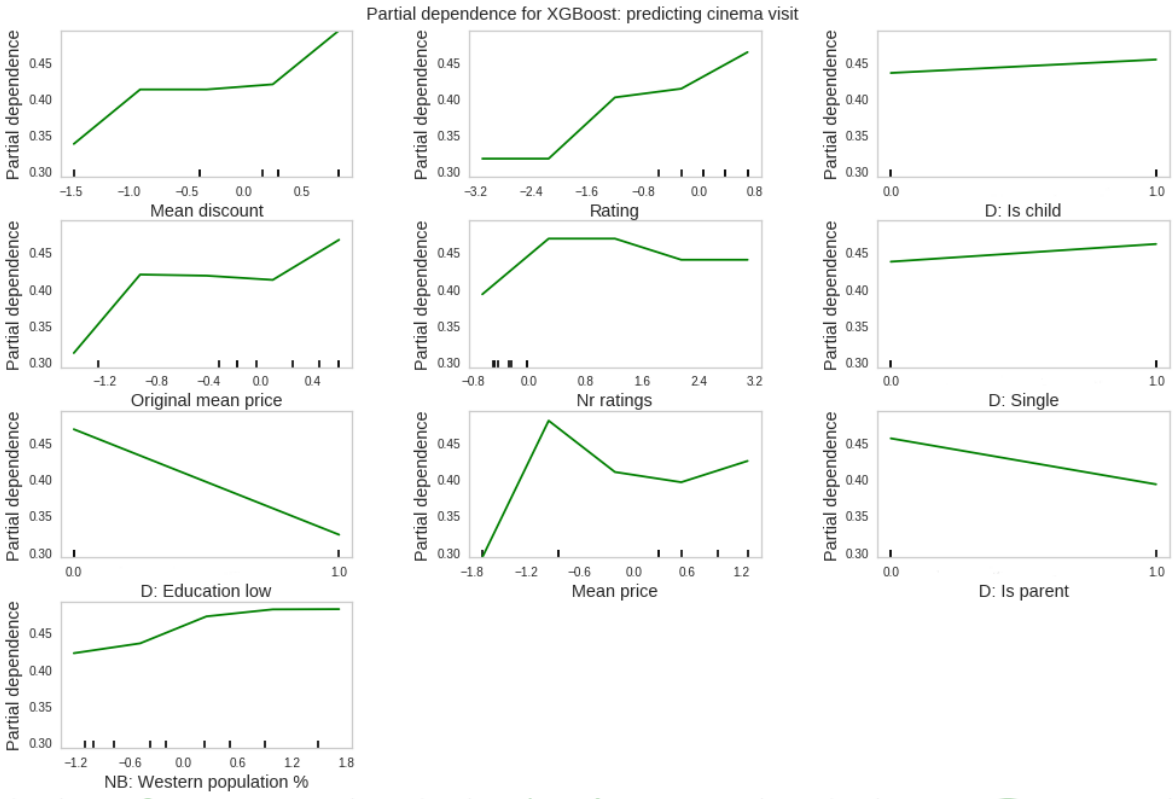
Similar to the predicting museum visit, neighbourhood feature *Western population %* had a VIF score of > 10, indicating high colinearity. Original mean price, mean discount and mean price also have VIF scores of > 10. Among features with correlation coefficients > 0.7, one was found with coefficient of 0.96 between *mean discount* and *original mean price*. *Western population %* has a correlation coefficient of 0.88 with the neighbourhood feature *Highly educated %*. Both features co-occur in the top 10 of model native Random Forest scores, and permutation importance scores from Logistic Regression and Random Forest. Model native and permutation importance scores do not correspond for any of the top features. Of the 10 neighbourhood features used for this study, the same features as predicting museum visit occur in the top 10, i.e. *Western population %*, *families with children %* and *highly educated %*.

The PDP for Random Forest in figure 9a shows an uneven distribution for *number of ratings*, *rating*, *mean discount* and *original price*. These features had fewer data points available in one decile, meaning that the partial dependence estimates can be less reliable in that region. *Mean price* shows only one tick mark at 0 because museums were largely free to visit. The PDP of *travel distance* shows a noticeably stronger trend compared to the one found in for predicting museum visit, with increasing travel distance leading to decreased probability of a visit. Demographic feature *age* does not show a clear relationship. The neighbourhood features show clear relationships.

The PDP's for XGBoost in 9b shows a similar uneven distribution for the earlier mentioned features, and shows a strong PDP for binary feature *education low* in the negative direction.



(a)



(b)

Figure 9: Partial dependence plots for the top 10 features for (a) Random Forest and (b) XGBoost: predicting cinema visit.

3.2.3 Theatre

Figure 10 shows a mix of demographic, neighbourhood, location specific features, and travel distance, were included in the top 10 ranking for the model native feature importance scores. A mix was also found for permutation importance scores. For the Logistic Regression scores, all features except for demographic features $\leq$ *2 adults without children* and *education high*, were significant.

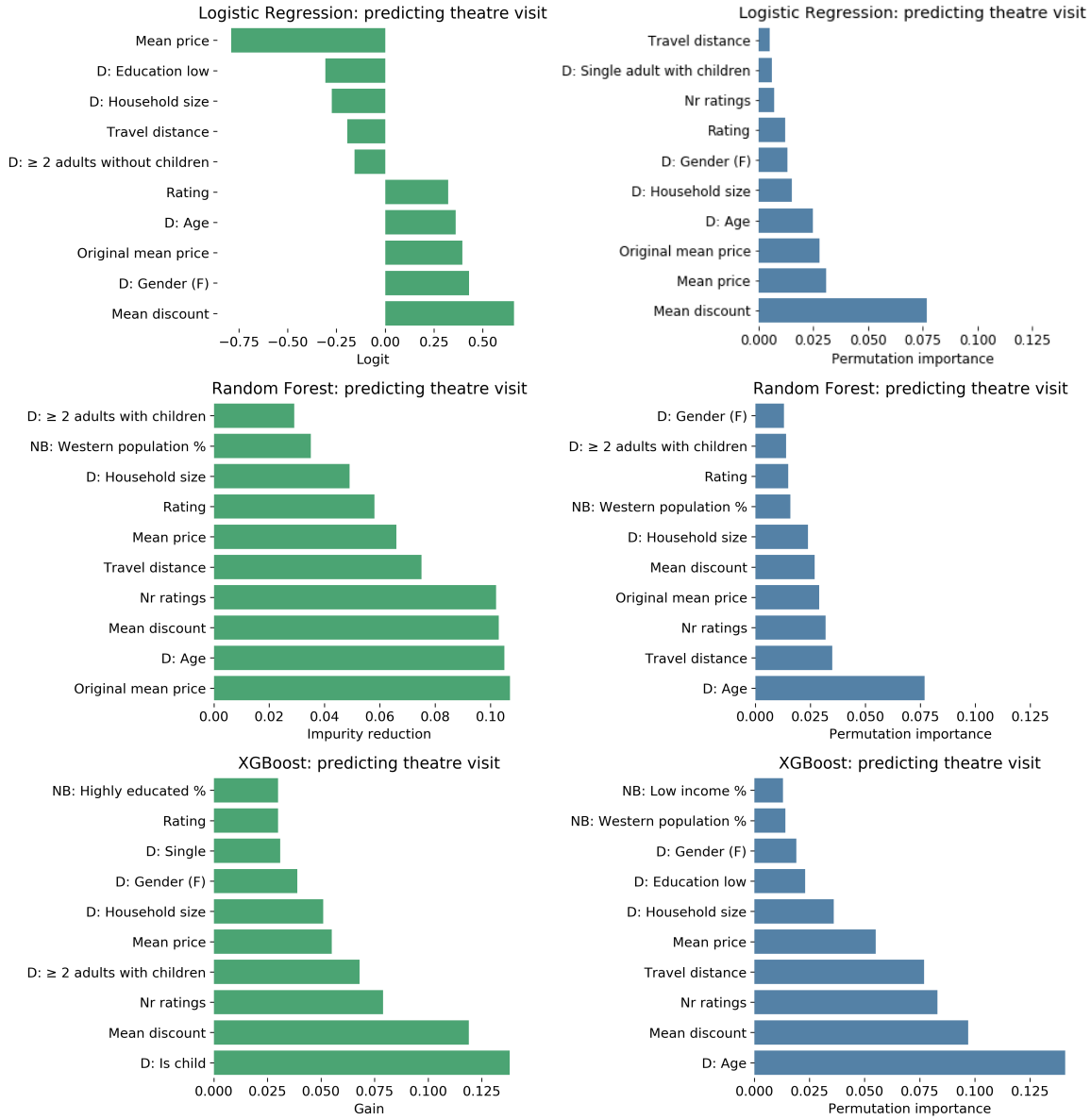


Figure 10: Top 10 feature importance scores for predicting theatre visit. Left side: model native feature importance, right side: permutation importance. D = demographic feature, NB: neighbourhood feature.

Table 16 gives the percentage per feature type across all three model top 10 feature importance scores, grouped per feature importance type. There was no specific top feature among the model native feature importance scores. *Travel distance* occurs four times as top feature in the permutation importance rankings.

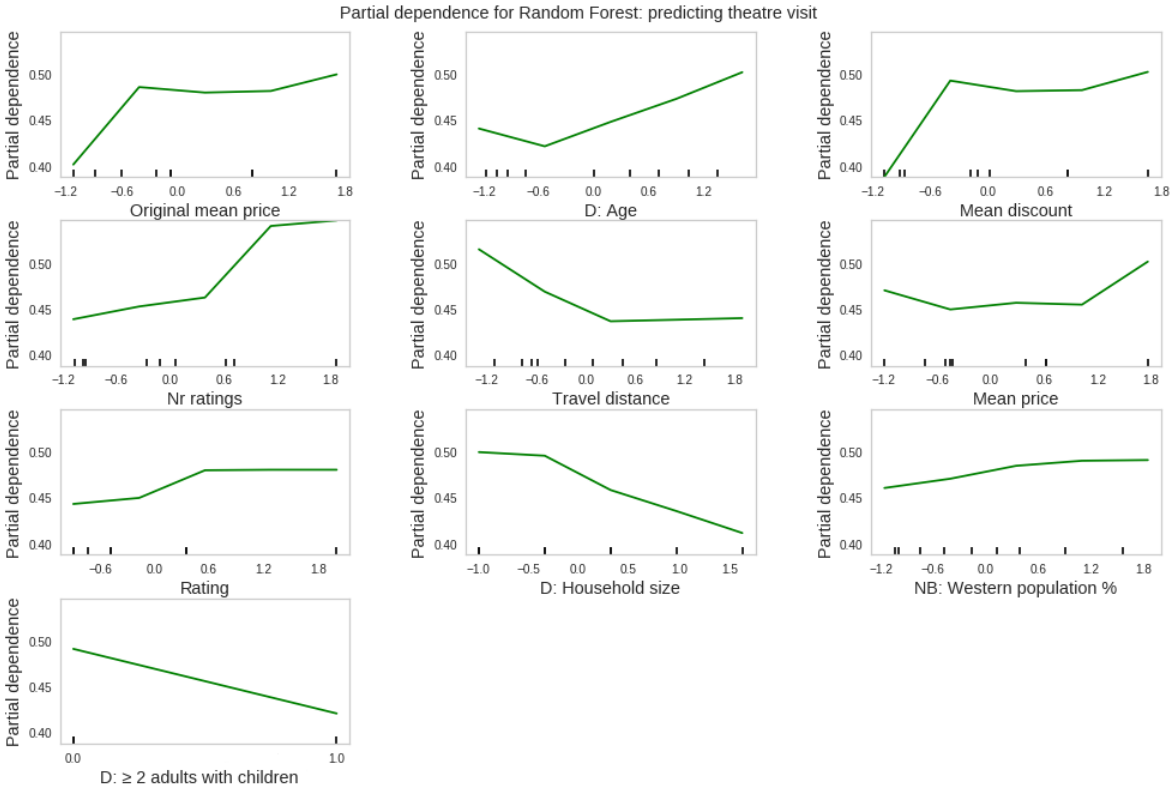
	Demographic	Neighbourhood	Location (including travel distance)
Model native feature importance	43.3%	6.7%	50%
Permutation importance	40%	10%	50%

Table 16: Percentage share per feature type in all three model top 10 feature importance scores.

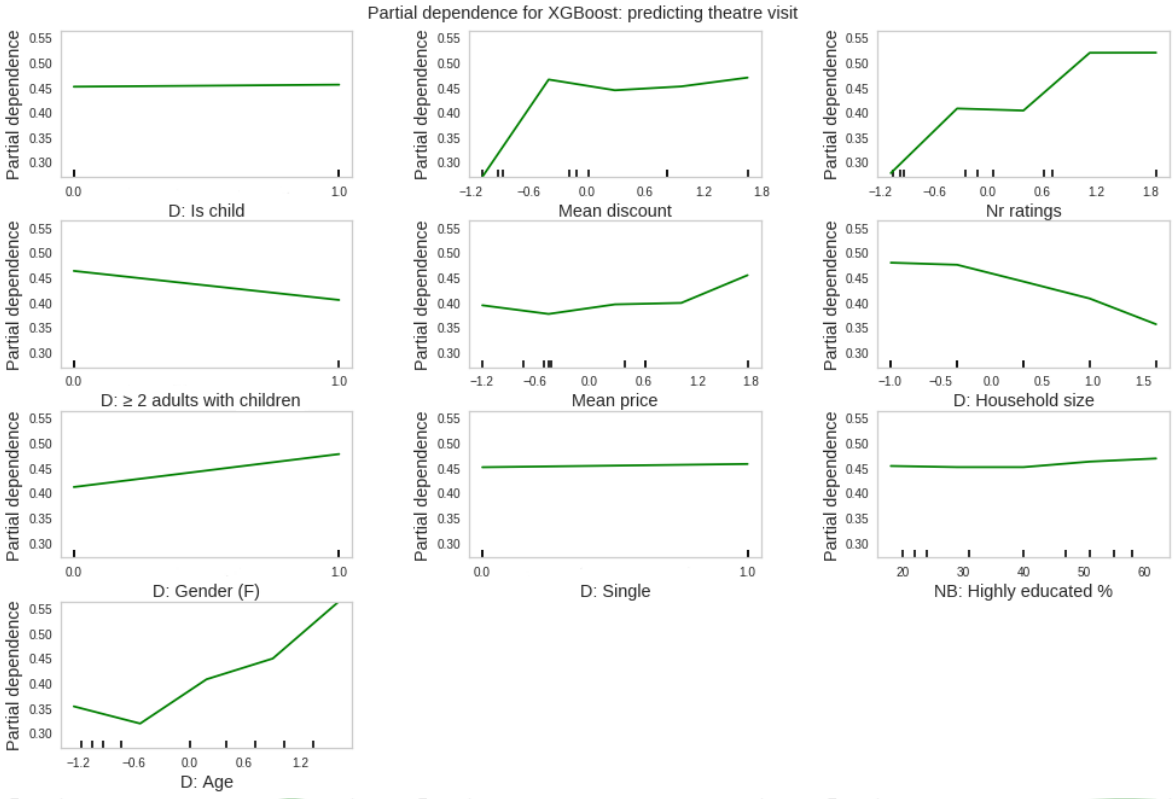
Similar to previous results, neighbourhood feature *Western population %* had a VIF score of > 10, indicating high colinearity. Original mean price, mean discount and mean price also have VIF scores of > 10. Among features with correlation coefficients > 0.7, a correlation of 0.99 was found between *mean discount* and *original mean price*. *Western population %* has a correlation coefficient of 0.87 with the neighbourhood feature *Highly educated %*. Contrary to previous results, these two features do not co-occur in the top 10 scores. Model native and permutation importance scores correspond for Logistic Regression. Of the 10 neighbourhood features used for this study, the following features were present in the feature importance top 10: *Western population %*, *families with children %* and *highly educated %*, together with *low income %*.

The PDP for Random Forest in figure 11a shows an uneven distribution for *number of ratings*, *rating*, *mean discount* and *original price*. These features had fewer data points available in one decile, meaning that the partial dependence estimates can be less reliable in that region. *Travel distance* was considered to be the top feature among the different metrics, its PDP shows a clear trend at first but then flattens out half way through. Demographic features *age* and *household size* show clear relationships, generally positive and negative respectively. Age shows a stronger effect compared to the PDP's from museum and cinema.

The PDP for XGBoost in 11b shows similar uneven distribution. There was a strong relationship for *age*, generally in the positive direction.



(a)



(b)

Figure 11: Partial dependence plots for the top 10 features for (a) Random Forest and (b) XGBoost: predicting theatre visit.

3.2.4 Sport

Figure 12 shows a mix of demographic, neighbourhood, location specific features, and travel distance, were included in the top 10 ranking for the model native feature importance scores. A mix was also found for permutation importance scores. For the Logistic Regression scores, only 3 out of 10 features were significant. These were *number of ratings*, *travel distance* and *age*.

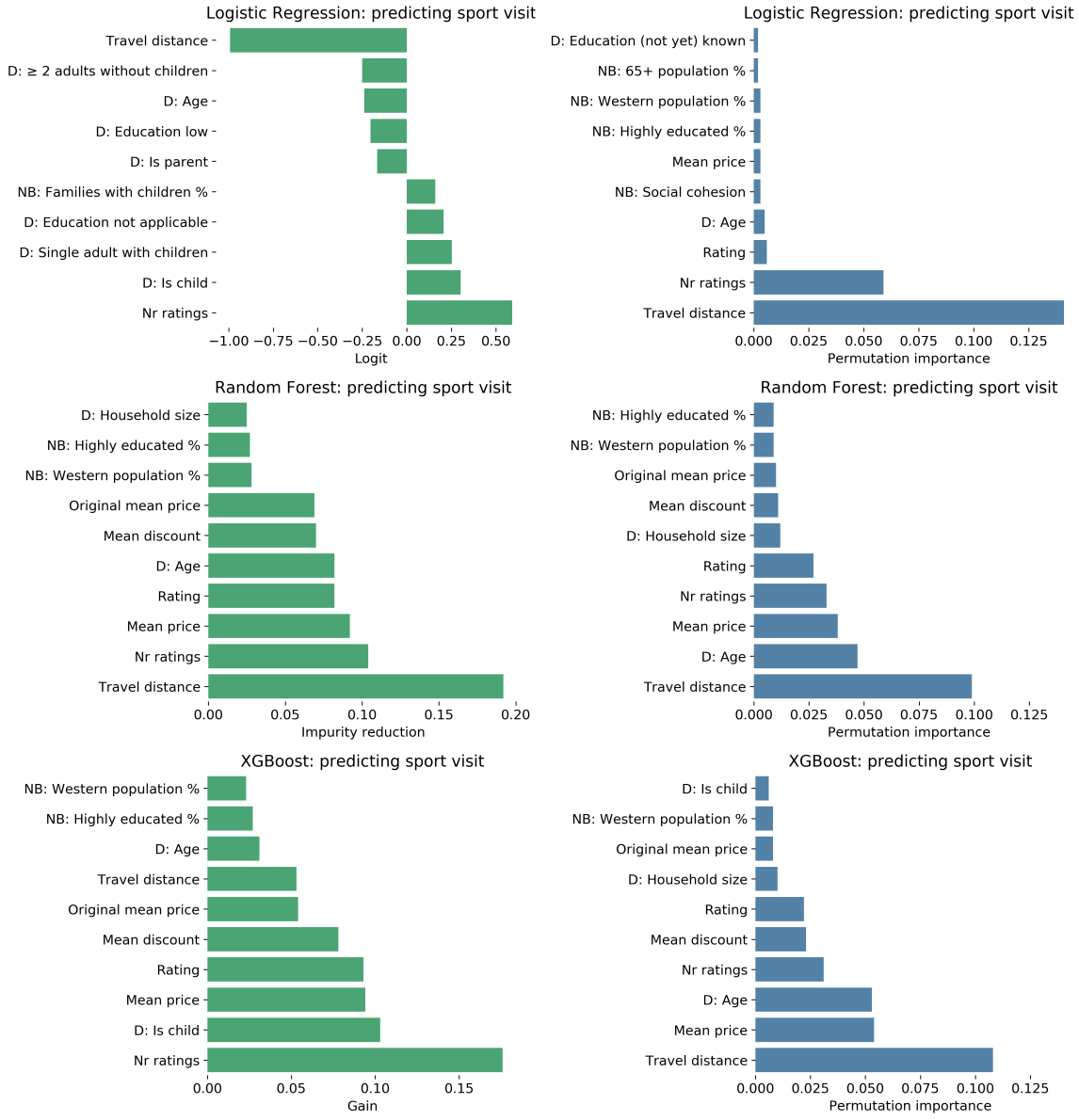


Figure 12: Top 10 feature importance scores for predicting sport visit. Left side: model native feature importance, right side: permutation importance. D = demographic feature, NB: neighbourhood feature.

Table 17 gives the percentage per feature type across all three model top 10 feature importance scores, grouped per feature importance type. *Number of ratings* occurs twice as top feature in the model native rankings. *Travel distance* is the top feature in the permutation importance rankings.

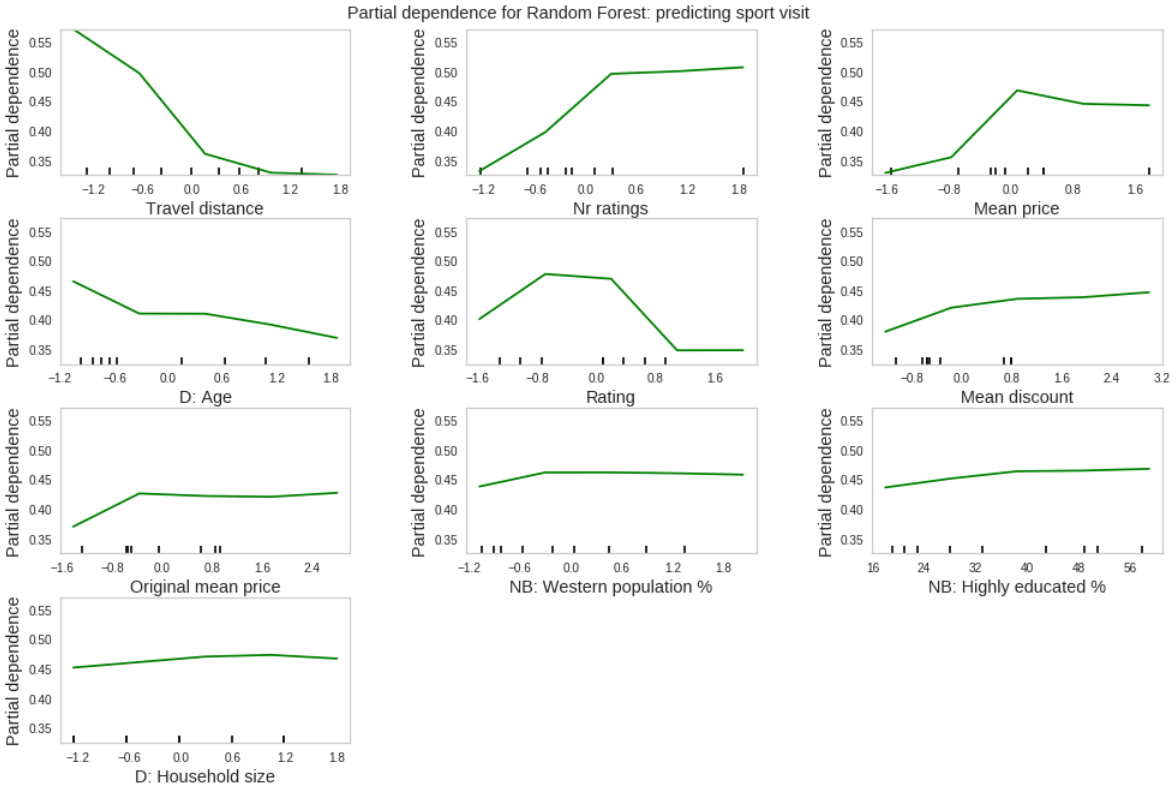
	Demographic	Neighbourhood	Location (including travel distance)
Model native feature importance	36.6%	16.6%	46.7%
Permutation importance	23.3%	20%	50%

Table 17: Percentage share per feature type in all three model top 10 feature importance scores.

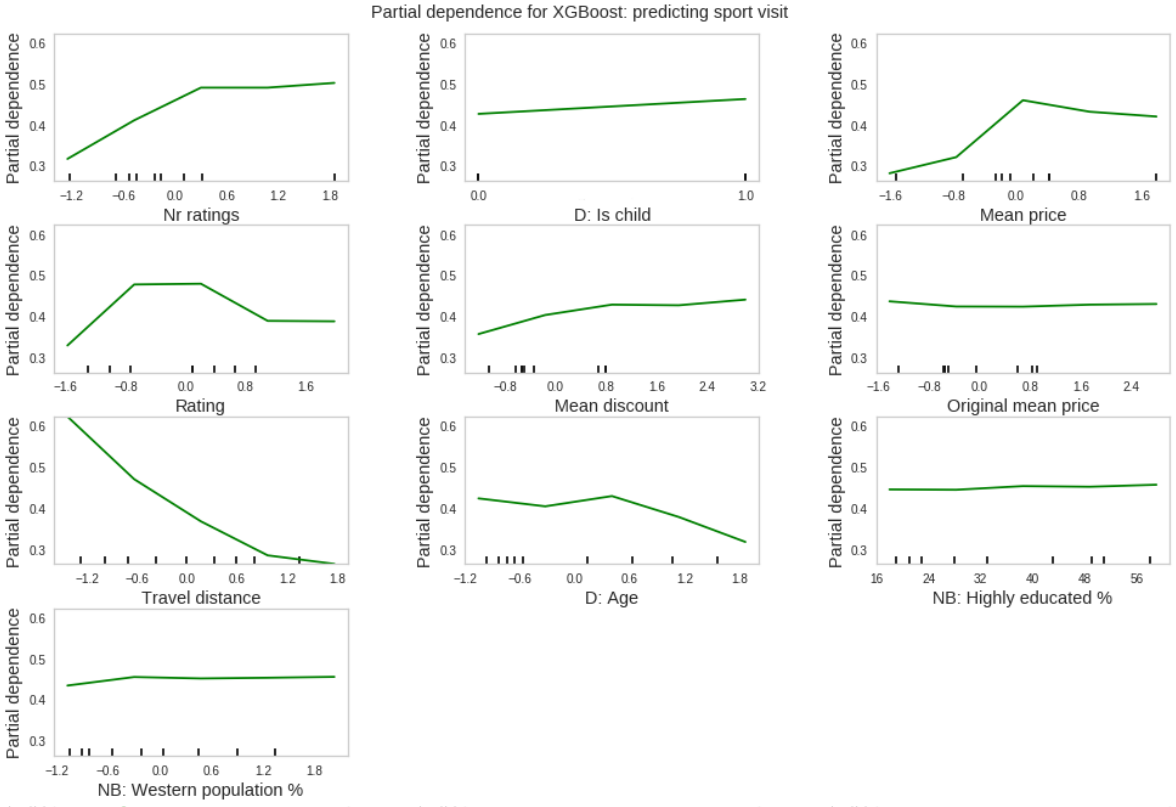
Similar to previous results, neighbourhood feature *Western population %* had a VIF score of > 10, indicating high colinearity. Original mean price, mean discount and mean price also have VIF scores of > 10. Among features with correlation coefficients > 0.7, a correlation of 0.99 was present between *mean discount* and *original mean price*. *Western population %* has a correlation coefficient of 0.87 with the neighbourhood feature *Highly educated %*. These two features co-occur in the model native top 10 of Random Forest and XGBoost, and permutation top 10 for Logistic Regression and Random Forest. The top model native and permutation importance score correspond for Random Forest and Logistic Regression, being *travel distance*. Of the 10 neighbourhood features used for this study the following features occur in the feature importance top 10: *Western population %*, *families with children %* and *highly educated %*, *low income %* and *65+ population %*.

The PDP for Random Forest in figure 13a shows an uneven distribution for *number of ratings*, *mean discount* and *original price* and *mean price*. These features had fewer data points available in one decile, meaning that the partial dependence estimates can be less reliable in that region. The PDP of *travel distance* shows a clear and strong relationship in the negative direction. Higher distances mean lower probability of a visit. Demographic feature *age* generally moves into a negative direction.

The PDP for XGBoost in 13b shows a similar uneven distribution. Similarly, there is a strong relationship for *travel distance*, generally in the negative direction. This was the same for *age*.



(a)



(b)

Figure 13: Partial dependence plots for the top 10 features for (a) Random Forest and (b) XGBoost: predicting sport visit.

3.3 Clustering

Results for clustering were reported using descriptive analysis, as top 3 most frequently occurring variable combinations. This was done to keep reporting and interpretation succinct. The goal was to identify specific subgroups among users and non-users of the City Pass, and visitors and non-visitors of culture and sport locations. It must be noted that 53.8 % of all users belonged to the educational level 'Not (yet) known', of which children formed the largest proportion. The optimal number of clusters was evaluated using an elbow plot. n = 2 was picked as optimum value, a small value which enables clear interpretation of the results.

3.3.1 Subgroups unique use of City Pass

The data was split into two groups: those who have used the City Pass, and those who have not. These groups were subsequently clustered separately to identify subgroups. This was set at n = 2 for each group. Cluster ID denotes which cluster the results belong to. Percentage was reported as proportion of users with a given combination of features, divided by total number of users within the cluster that combination belongs to.

Table 18 shows that *Not used: Cluster ID = 1* has three different type of users which were very similar, all within the same age group (65+) and educational level. *Not used: Cluster ID = 2* has a wider age range and were all from Single adult with children households of size = 3. The percentages for the second cluster were however relatively low compared to the first one.

Used: Cluster ID = 1 has a different frequent feature combinations, most notably different in age. *Used: Cluster ID = 2* is a group similar in age (4-12), household type, educational level.

Cluster ID	Gender	Educational level	Household size	Household type	Age	Is child	Is parent	%
Not used								
1	V	Not (yet) known	1	Single	65+	0	0	10.9%
1	M	Not (yet) known	2	$\geq$ 2 adults without children	65+	0	0	5.3%
1	M	Not (yet) known	1	Single	65+	0	0	5.0%
2	V	Low	3	Single adult with children	35-45	0	1	0.7%
2	M	Not (yet) known	3	Single adult with children	12-18	1	0	0.6%
2	M	Not (yet) known	3	Single adult with children	0-4	1	0	0.6%
Used								
1	V	Not (yet) known	1	Single	65+	0	0	8.0%
1	V	Low	1	Single	55-65	0	0	4.4%
1	M	Not (yet) known	2	Single adult with children	4-12	1	0	4.1%
2	M	Not (yet) known	3	Single adult with children	4-12	1	0	5.8%
2	V	Not (yet) known	3	Single adult with children	4-12	1	0	5.5%
2	M	Not (yet) known	4	Single adult with children	4-12	1	0	4.0%

Table 18: Clustering results for City Pass unique use. Percentage denotes proportion of users with a given combination of features, divided by total number of users within the cluster it belongs to.

3.3.2 Subgroups cultural and sport participation

The data was split into two groups: those who had visited a location in a category, and those who had not.

Table 19 shows the results for clustering the groups that had visited and not visited a museum. *Not used: Cluster ID = 1* most frequent combinations included children of two different age groups, from single adult with children households. *Not used: Cluster ID = 2* shows a wider age range.

Used: Cluster ID = 1 is characterized by children from the same age (4-12), all from single adult with children house holds. *Used: Cluster ID = 2* is a group which is not similar in age.

Cluster ID	Gender	Educational level	Household size	Household type	Age	Is child	Is parent	%
Not visited								
1	M	Not (yet) known	3	Single adult with children	4-12	1	0	5.3%
1	V	Not (yet) known	3	Single adult with children	4-12	1	0	5.0%
1	M	Not (yet) known	3	Single adult with children	12-18	1	0	4.0%
2	V	Not (yet) known	1	Single	65+	0	0	7.6%
2	V	Low	1	Single	55-65	0	0	4.4%
2	M	Not (yet) known	2	Single adult with children	12-18	1	0	4.2%
Visited								
1	M	Not (yet) known	3	Single adult with children	4-12	1	0	7.1%
1	V	Not (yet) known	3	Single adult with children	4-12	1	0	6.8%
1	V	Not (yet) known	4	Single adult with children	4-12	1	0	4.8%
2	V	Not (yet) known	1	Single	65+	0	0	8.0%
2	V	Not (yet) known	2	Single adult with children	4-12	1	0	5.7%
2	M	Not (yet) known	2	Single adult with children	4-12	1	0	5.6%

Table 19: Clustering results for museum visit. Percentage denotes proportion of users with a given combination of features, divided by total number of users within the cluster it belongs to.

Table 20 shows the results for groups who had visited and not visited a cinema. *Not used: Cluster ID = 1* shows once again frequent combinations of children (4-12) from single adult with children households. The second cluster *Not used: Cluster ID = 2* shows the same frequent combinations as in the previous results from table 19. This was the same case for *Used: Cluster ID = 1* and *Used: Cluster ID = 2*.

Cluster ID	Gender	Educational level	Household size	Household type	Age	Is child	Is parent	%
Not visited								
1	M	Not (yet) known	3	Single adult with children	4-12	1	0	5.3%
1	V	Not (yet) known	3	Single adult with children	4-12	1	0	5.0%
1	M	Not (yet) known	4	Single adult with children	4-12	1	0	3.8%
2	V	Not (yet) known	1	Single	65+	0	0	7.5%
2	V	Low	1	Single	55-65	0	0	4.4%
2	M	Not (yet) known	2	Single adult with children	12-18	1	0	3.9%
Visited								
1	M	Not (yet) known	3	Single adult with children	4-12	1	0	8.2%
1	V	Not (yet) known	3	Single adult with children	4-12	1	0	7.8%
1	V	Not (yet) known	5	$\geq$ 2 adults with children	4-12	1	0	5.7%
2	V	Not (yet) known	3	Single adult with children	4-12	1	0	9.6%
2	V	Not (yet) known	1	Single	65+	0	0	5.6%
2	M	Not (yet) known	2	Single adult with children	4-12	1	0	5.2%

Table 20: Clustering results for cinema visit. Percentage denotes proportion of users with a given combination of features, divided by total number of users within the cluster it belongs to.

Table 21 shows the results for groups who had visited and did visited a theatre. *Not used: Cluster ID = 1* is and *Not used: Cluster ID = 2* were similar to the previously reported frequent combinations for museum and cinema visit.

Used: Cluster ID = 1 is contains two different age ranges, with a notably large percentage for single female 65+, encompassing 14.4% within that cluster. *Used: Cluster ID = 2* is again characterized by children (age = 4-12) from single adult with children households.

Cluster ID	Gender	Educational level	Household size	Household type	Age	Is child	Is parent	%
Not visited								
1	M	Not (yet) known	3	Single adult with children	4-12	1	0	5.4%
1	V	Not (yet) known	3	Single adult with children	4-12	1	0	4.9%
1	M	Not (yet) known	4	Single adult with children	4-12	1	0	3.9%
2	V	Not (yet) known	1	Single	65+	0	0	6.7%
2	V	Low	1	Single	55-65	0	0	4.0%
2	M	Not (yet) known	2	Single adult with children	12-18	1	0	3.7%
Visited								
1	V	Not (yet) known	1	Single	65+	0	0	14.4%
1	V	Not (yet) known	2	Single adult with children	4-12	1	0	5.2%
1	M	Not (yet) known	2	Single adult with children	4-12	1	0	4.5%
2	V	Not (yet) known	3	Single adult with children	4-12	1	0	6.2%
2	M	Not (yet) known	3	Single adult with children	4-12	1	0	5.3%
2	V	Not (yet) known	4	Single adult with children	4-12	1	0	2.6%

Table 21: Clustering results for theatre visit. Percentage denotes proportion of users with a given combination of features, divided by total number of users within the cluster it belongs to.

Finally, Table 22 shows the results for groups who had visited and not visited a sport location. Clusters *Not used: Cluster ID = 1* and *Not used: Cluster ID = 2* contain similar frequent user feature combinations to the previously reported results, however *Not used: Cluster ID = 2* contains notably a group of single 65+ female users, encompassing 10.7 % of the cluster.

Both *Used: Cluster ID = 1* and *Used: Cluster ID = 2* were characterized entirely by children (age = 4-12), all from single adult with children households.

Cluster ID	Gender	Educational level	Household size	Household type	Age	Is child	Is parent	%
Not visited								
1	M	Not (yet) known	3	Single adult with children	4-12	1	0	4.7%
1	V	Not (yet) known	3	Single adult with children	4-12	1	0	4.5%
1	M	Not (yet) known	3	Single adult with children	12-18	1	0	3.5%
2	V	Not (yet) known	1	Single	65+	0	0	10.7%
2	V	Low	1	Single	55-65	0	0	5.7%
2	M	Not (yet) known	2	Single adult with children	12-18	1	0	4.0%
Visited								
1	M	Not (yet) known	3	Single adult with children	4-12	1	0	8.7%
1	V	Not (yet) known	3	Single adult with children	4-12	1	0	8.1%
1	M	Not (yet) known	2	Single adult with children	4-12	1	0	5.3%
2	V	Not (yet) known	4	Single adult with children	4-12	1	0	5.8%
2	M	Not (yet) known	4	Single adult with children	4-12	1	0	5.5%
2	M	Not (yet) known	5	$\geq$ 2 adults with children	4-12	1	0	5.1%

Table 22: Clustering results for sport visit. Percentage denotes proportion of users with a given combination of features, divided by total number of users within the cluster it belongs to.

4 Discussion

The aim of this study was to investigate how well City Pass use can be predicted and understood using machine learning techniques, with a focus on interpretability of the results. Interpretability refers to insights into why a model came to certain conclusions, and can be achieved using measures such as feature importance scores. City Pass use included unique use as well as cultural and sport participation.

The main research question was formulated as:

How well can we predict and understand the use of City Pass among low-income citizens from Amsterdam with machine learning, using demographic, neighbourhood, geographical and meta data of locations?

The three sub questions were as follows:

1. Which supervised machine learning methods yield the best performance for predicting City Pass unique use, cultural participation and sport participation?
2. What are the top features for predicting City Pass unique use, cultural participation and sport participation?
    - What can we infer from these top features about behaviour and background of users?
3. Can subgroups be identified among users and non-users of the City Pass, and visitors and non-visitors of culture and sport locations?

First, the performance of the prediction part will be discussed, which includes interpretation and implications of the results. These focus on what the results mean and why they matter. Secondly, the clustering results will also be discussed, and, finally, several limitations will be mentioned, as well as recommendations for future research.

4.1 Prediction performance

In general, the results show that the performance of the classification models were always higher than the Zero Rule baseline classifier, with statistically significant differences between baseline and classification models. XGBoost showed the highest performance for prediction of unique use, museum, cinema and sport visit. Random Forest scored the highest for theatre. In the case of theatre and sport, the difference was not always significant.

Varying magnitudes of performance scores can be observed across different models, and across the prediction for different category outcomes. The different outcomes can be explained by the different data sizes used for each prediction task, as shown in table

5, as well by the ability of the models themselves. XGBoost is a classifier that commonly outperforms other models in competitions (Chen and Guestrin, 2016). The size of the data sets varies depending on the number of locations part of a category. This difference seems to be reflected in the differing performance scores between models, meaning that more training examples might have lead to these differences. Based on the performance scores and significance, one could say that these models perform decently at predicting City Pass use.

During evaluation of each supervised learning model on each data set, learning curves based on the train data were plotted to detect possible problems in regards to the fit. These can be found in appendix B. There was always a gap between training and validation accuracy for the Random Forest and XGBoost models. If the training score is much greater than the validation score for the maximum number of training samples, adding more training samples would have been beneficial. This might have improved generalization. A gap between the two scores can, but does not necessarily indicate overfitting. There is no straight answer to how much of a gap is allowed, but it is good to keep this aspect in mind. Overfitting can mean that the feature importance scores are not trustworthy.

4.2 Feature importance

This study used two different metrics for measuring feature importance. Metrics native to the model, and permutation importance. Some caution has to be taken when interpreting the features individually or independently. Observations were made firstly about predictive contribution, and secondly about the direction of the prediction. No causal inferences can be made. Logistic Regression assumes a linear combination between features, and decision trees work by splitting based on variance reduction. When interpreting the direction of the logits, or partial dependence plots, one must not assume that an outcome is more likely for all users that have a increased value of a particular feature. Interpretation is in terms of associations with increased or decreased likelihoods, but the relationship with other features always have to be kept in mind, since the input was an series of features.

4.2.1 Top features

The findings for predicting unique City Pass use show that demographic features generally contribute by far the most, as reflected by table 11, with either *age* or *is child* being important candidates for unique use according to both feature importance metrics. This observation is likely related to the fact that children are heavily over-represented in the group that used the City Pass. This effect might become less pronounced if children were removed from the data set. Both the Logistic Regression score and partial dependence plots of *age* indicate that an increase in age is associated with lower probability of the outcome 'used'. The use of Poverty Reduction programmes apparently also offer

some predictive power, however it is not easy to derive further implications from since they are age-bound, and each limited to a specific user group.

For the different categories within cultural and sport participation, varying proportions of contribution for demographic, neighbourhood and location based features can be observed in the top 10 feature importance scores. This was the case for both model native importance metrics and permutation importance. In some categories, demographic features formed a bigger proportion than others. These differences were generally minor. A more important comparison to look at are the different magnitudes of importance scores, as well as the direction between the features and outcome. This can be interpreted from the signs of the logits, and the partial dependence plots for Random Forest and XGBoost. For predicting museum visit with Logistic Regression, the highest feature importance score was found for *is child*, with a logit of -0.696, while this feature has a logit of 0.302 for predicting sport visit. Note the sign of both logits, while one has a negative relationship, the other is positive. Similarly, *travel distance* scores highest in the Random Forest model for predicting sport visit, with an impurity reduction of 0.192, but, 0.74 was observed for the same feature and model when predicting museum visit. Different categories yield different importance scores for various features. This suggest some features might play a larger role in participation than others. For example, the different scores found for travel distance can suggest something about the willingness of people to travel for certain locations and type of activities. This was often the case for other features as well. Looking per category within cultural and sport participation, differences in feature contribution can be observed in the permutation importance scores as well. The different models output different permutation importance scores per feature, which suggest that some models rely more heavily on certain features than other models for prediction accuracy. Looking back at the feature importance scores for predicting unique use in figure 4, it can be seen that both Random Forest and XGBoost rely by far more on *age* than Logistic Regression does.

Some categories such as cinema and theatre did not show an obvious presentation of top features in the overall top 10. These were different per model. The top features are the ones with the highest score in their respective feature importance ranking. One way of determining a top feature can be based on the best performing model. This was XGBoost in most cases. The model native feature importance scores provide the following best features: *age* for unique use, *number of ratings* for museum visit, *mean discount* for cinema visit, *original mean price* for theatre and *number ratings* for sport.

However, it is more interesting to look at the majority vote for all models to get a broader scope. For the final interpretation of which features are the best per category, permutation score can be considered as primary measurement. This offers a more general way to look at feature importance and makes it easier to compare between models. Looking at which top feature had a majority presence, meaning occurred at least twice in the top permutation importance scores for all three models, this was *age* for unique use. For museum it was *number of ratings*, cinema was *age*, theatre was *age* and for sport it was *travel distance*.

4.2.2 Neighbourhood

Out of the 10 neighbourhood features used for this study, few were present in the top 10 ranking. One particular observation is the low contribution of *number of City Pass locations* for predicting unique use. One would have expected availability to play a bigger role because users might be more likely to use the City Pass if there are enough locations in their vicinity (i.e. neighbourhood) to visit. The total feature importance score overview in the appendix A shows that neighbourhood features have generally low feature importance according to the model native and permutation scores. Features such as percentage low income population, working population, safety index and social cohesion did not show up high in the ranking. This might imply they play less of a role. Especially considering the fact that the City Pass population is already homogeneous until a certain degree, meaning they form a similar socioeconomic group, in this case low-income. Another aspect to take into account is the degree in which neighbourhoods are mixed on a population level, which can also affect how fine-grained the neighbourhood features are.

The results for predicting unique use show that the number of City Pass locations in a users neighbourhood was not a strong predictor for City Pass use. What this can mean is that accessibility to enough facilities does not play a strong role in whether people will use or not use the City Pass.

Features such as percentage *western population* and *high educated population* generally scored highest. Both can be related to socioeconomic status of a neighbourhood, which was mentioned as an important predictor in sport participation by Eime et al. (2015) and Allen and Vella (2015). Shenassa et al. (2006) even stated that perceived safety can play a role. Hoekman et al. (2016) mentioned travel distance as relevant feature, with larger travel distances being positively related to, at least, monthly sport participation. This is not reflected in the data, and it has to be taken into account that City Pass data might not be representative or complete for overall sport participation of a person. However, clear literature about neighbourhood predictors for cultural participation is lacking for this study, as were extensive comparisons between different categories. Therefore, there were not well defined expectations for this study as to which features would specifically contribute more. Brook (2016) did mention accessibility as neighbourhood characteristic. He used this as a feature for predicting museum visitation. This study encoded accessibility by travel distance, which indeed often showed up in the top 10 feature importance scores and had some effect.

4.2.3 Demographic

Based on work by Notten et al. (2015) and Nagel (2009), there was an expectation that educational level would play a role in cultural participation. The same was for gender. Toepoel (2011) mentioned that older adults participate more often in high brow activities such as museums and theatres. The features *education high* and *education low* both show up in different top 10 rankings, *education not (yet) known*, the most com-

mon category, does not. There might be some confounding effects with other variables such as age, since children make up the largest proportion. One important note is that because the City Pass grants discounted or sometimes free access to locations, this can diminish the effect of educational level. It can be observed among only a few top 10 scores that museum and cinema *education high* relate to increased likelihood, and *education low* to decreased likelihood. So this leaves the interpretation of educational level undecided. Notten et al. (2015) did mention that cultural participation was affected more so by educational level than income level. As for gender, this also does not seem to have a noticeable effect on cultural participation. Gender only occurred in the top 10 for predicting theatre visit. Having the female gender seems to be associated with increased probability of going to the theatre.

Ruseski et al. (2011) found that household type, notably containing children, reduces the likelihood that individuals participate in sport activities. However, the findings of this study showed that for predicting sport participation, household type features to be either insignificant in the case of Logistic Regression, or just not present among the top features. Therefore it is hard to draw any conclusions on the effect. A possible explanation for this is that children, and therefore households with children, were over represented in the data.

Careful interpretation of the *is child* feature is required. This appeared several times in the top 10, but *is parent* did not. One would expect this to co-occur more often, but it does not because the transaction data often often did not register the companion of a child when visiting a location. A companion could be the actual parent, but also some other type of guardian. This means there is some hidden use present, which makes the data not entirely trustworthy. This hidden use was estimated to be around 8 % according to an internal report by the Municipality. This probably lead to a skewed distribution for the *age*, *is child* and *is parent* features, which the latter feature has less of an effect than it probably should.

Regarding observations on *household type* and *household size* for unique use, there was an indication that being from a household without children is associated with decreased likelihood of using the City Pass, which is also related to household size. Household type *single* and *household size* were strongly correlated (> 0.7), and household type $\geq$ *2 adults with children* with *household size* moderately (> 0.6). For cultural participation, based on the top 10, it can be observed that a larger household size is associated with decreased likelihood in the outcome of visiting a museum. Being from a single household seems associated with slightly increased likelihood of visiting a cinema. Being from a household with at least two adults with children, together with increasing household size, seems to be associated with decreased likelihood of going to the theatre.

4.2.4 Location

Concerning location based features, the feature *number of ratings* was most prominent. This feature was used as a reflection of popularity. The idea was to tell something

about the effect of popularity on visiting a type of location. However, the number of ratings can be strongly related to number of visitors, which means that the independent variable number or ratings is not entirely independent since it can also be regarded as inherent to the outcome. Popular locations can be considered popular because they receive more visits. This makes it harder to interpret whether the importance and partial dependence relationship of this feature means that something that is perceived as having higher popularity is associated with increased chances of the outcome. So the only inference that can be made as it is: popularity of a location is associated with increased likelihood of visiting a place, but it is not necessarily related to perceived popularity, i.e. 'word by mouth'.

4.2.5 Multicolinearity

There were some co-occurring features with high colinearity within the top 10 feature importance rankings. Multicolinearity can affect feature importance scores and make a feature appear to be a stronger predictor compared to other correlated features. Most notable colinear features included *mean price*, *original mean price* and *mean discount*. Mean discount and original mean price were in particular highly correlated, however in many instances not equal. When looking at the permutation importance scores, original mean price and mean discount both do seem to add new information, as seen from different scores. This suggests there was not necessarily a case of redundancy. When features are correlated, their importance can be shared by roughly per their correlation. *Is child* and *age* are also highly correlated features, which means they cannot be interpreted entirely independently.

4.3 Clustering for subgroups

Clustering with k-prototypes was used in an attempt to identify different subgroups, also known as clusters, among the two different outcome groups, i.e. users and non-users, visitors and non-visitors. The clustering results show that there was not always a clear difference present between clusters found for the two different groups, since there was a overlap between some combinations. Nor were there always distinct unique clusters: clusters characterized by an obvious set of user types. Intuitively, it is more easy to look at the differences in age groups. For unique use in table 18, there seems to be an observable difference between the two groups, with different type of clusters within each group. The same can be said for the museum, cinema, theatre and sport groups. One common cluster among the non-visitors in culture and sport, is one characterized by three different age groups: 12-18, 55-65 and 65+, from single adult with children, and adult households respectively. Another common cluster in the visitors group, is characterized by users between ages 4-12 from single adult with children households. This cluster sometimes also occurs in the non-visitors group. What the results suggest is that there are some distinct clusters, but this is not always clearly the case. Sometimes the clusters in both groups seem quite similar, which was to be expected when the

same specific type of users occur in both. Perhaps demographic features can not always provide enough nuance to find specific clusters, and uncover clear differences between the two groups.

4.4 Limitations

There were several limitations that could have potentially affected this study. This section discusses several limitations that could have had the biggest potential impact on the quality of the findings of this study and ability to effectively answer the research questions.

4.4.1 Data

There were several limitations to the data and preprocessing. The first limitation was incomplete, missing or incorrect data, which meant a small portion of data had to be discarded. The second limitation is the unbalanced nature of the classification data. Random undersampling of the majority class was done, therefore data had to be discarded as well. Different ratio's for undersampling could have had different impacts on the results. On hindsight, this should have been tested to evaluate its effects. It is likely that a larger majority class size would have suited this study as well, and lead to clearer results. A third possible limitation is the fact that educational level was eventually downsized to three levels, but perhaps too much granularity might have been lost in this case. Educational level as indicated in the original data did not reflect current education, but highest completed. Which meant that teenagers often had the 'not (yet) known' label assigned. Missing data on parental participation could have also lead to slightly skewed results.

The City Pass includes an extensive list of partners within Amsterdam, especially cultural locations. However, there a more potential locations not yet partnered with the City Pass, and this study only used data from 2018. Nothing can be said about citizens outside this specific City Pass demographic. This study covers a specific target group that includes only low-income citizens who have the City Pass, it does not say anything about for example the regular income 65+ senior citizens who have a City Pass.

4.4.2 Models and analysis

For this part, one limitation might concern one-hot-encoding, which was performed as preprocessing step. This involves converting categorical variables of n levels into n separate binary features. This meant that the reported percentages of each type of feature contribution, as shown in for example 14, can be a bit skewed. This is because one-hot-encoding basically increases the feature space for each type of feature. Furthermore, this study did not create $n-1$ features from categorical variables, which could have been a problem for regression based models. However, this did not form an

obstacle for this study, as previously explained in the Methods. It did make calculating the VIF scores not suitable for one-hot-encoded features when looking for possible multicolinearity. This is because regular regression requires $n - 1$ instead of n for categorical encoding.

Feature selection was not used for this study, mainly because the focus was on the contribution of all features, also in relation to each other. Feature selection might have offered some minor improvements on model performance, but the feature space was not very large to begin with, and feature selection might have eliminated some unique effects of supposedly redundant features.

Another potential limitation is the descriptive method used to describe each cluster, which does not give a clear enough reflection of each different cluster or subgroup. Thus it is hard to use to gain insight and target specific groups. However, if a different method was used, such as summarizing per category, it would lead to even more vague interpretation of what kind of subgroups and specific type of users are present. Perhaps scaling down to less features would have been better, for example age, gender and household type.

Lastly, sport was kept as a single category, but contains different types of physical activities that users could participate in, with locations for swimming forming the majority. A more fine grained distinction would have perhaps yielded different results. This can make the interpretation of feature importance less straight-forward, because now it is assumed that these features are relevant for all types of sport, while it might still differ per type of activity.

4.5 Future research

Several recommendations can be made for future research and practical actions to follow, with regards to the methodology or type of data to be included for future studies. For example a more geography or accessibility based approach using other features such as wheelchair accessibility of locations, or how well the location is connected with public transport. For cultural participation, more additional features, especially individual level data from transaction data could have been used to enrich the data set with features such as 'literacy', which could have been created since use of free library membership was registered in the data. Notten et al. (2015) mentioned that literacy skills are strongly associated with cultural participation. The number of specific type of locations in a neighbourhood could have also been included in predicting cultural and sport participation. This study did not use that, as travel distance indirectly and partially encoded that aspect.

The clustering results did not yield entirely distinct ways to identify subgroups, and while feature importance scores can provide some insights in factors that affect behaviour, due to lack of substantial background knowledge, it is hard to infer actual motivations of users, and really understand why they choose to use or not use the City Pass, and visit or not visit a certain location. More qualitative research would

be required to understand users and their motivations. One internal report, which questioned some users, did mention personal attractiveness of a location for example. While general attractiveness based on popularity can be captured by the data, the data cannot easily capture personal taste and psychological factors, except those that are available from the data. One way to capture this could have been done by encoding the attendance of other locations as features, to reflect 'taste'.

A more general recommendation is to involve more sociological expertise and a broader involvement of sociological literature in this type of study. A cross-disciplinary study would perhaps yield better model performance and more elaborate insights. Sociological expertise can help place everything in a broader context, and understand which features can be best used, together with more in-depth interpretation of the results.

If there is a motive to increase City Pass use or attendance at certain locations, one practical recommendation to make is taking some of these top features into consideration. While the directional effect of these features on the outcome is not literally causal, it can give an idea of why some locations are for example less likely to be visited.

5 Conclusion

In this study, the application of machine learning was investigated for predicting the use of City Pass. This included unique use, cultural and sport participation. The main findings show that the classification models predict City Pass use fairly well, with XGBoost proving to be overall the best classifier. While the method used for clustering leaves some things to be desired in regards to identifying subgroups, the results of the classification models provide potentially new insights into factors which determine user behaviour. Each model performs slightly different from each other, and yield different feature importance scores. However, the general observation is that a mix of demographic, neighbourhood and location related features can be used for predicting unique use, cultural and sport participation. The top features based on model native feature importance scores tended to vary. Using the majority of top permutation importance scores across all three classification models, the following top features can be observed: *age* for unique use, *number of ratings* for museum visit, *age* for cinema visit, *age* for theatre, and *travel distance* for sport.

Some neighbourhood features were not always as important, and most generally showed lower contribution to the models. The use of permutation importance in addition to model native importance has shown that both metrics can be used to assess importance. Together with partial dependence plots and checking for multicolinearity, it has proven to be useful in understanding how the models produce certain results. What can be inferred from these results is that there is not one single factor that functions as a top predictor. There are multiple possible factors at play, and while this study has uncovered some and its associated importance scores and effects, there are many more which require further investigation.

To conclude, a data-driven approach with machine learning has shown to be a promising method for predicting and understanding City Pass use. It helped with understanding, for example, how factors such as age, travel distance and number of ratings of a location, relates to the likelihood of using the City Pass, or visiting a location. This also provided a clear view of how features weights are distributed in relation to each other. Such insights might be useful for recommending certain City Pass locations or offers to users.

References

Allen, M. S. and Vella, S. A. (2015). Are the correlates of sport participation similar to those of screen time? *Preventive medicine reports*, 2:114–117.

Amsterdam, G. (2018). Basisregistratie adressen en gebouwen (bag). https://data.amsterdam.nl/datasets/CxSRcN9AhiPipQ/. Visited on 2019-1-7.

Bennett, J., Lanning, S., and Netflix, N. (2007). The netflix prize. In *In KDD Cup and Workshop in conjunction with KDD*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brito, P. Q., Soares, C., Almeida, S., Monte, A., and Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing*, 36:93 – 100. Sustaining Resilience in Today's Demanding Environments.

Brook, O. (2016). Spatial equity and cultural participation: how access influences attendance at museums and galleries in london. *Cultural Trends*, 25.

Brownlee, J. (2016). How to implement baseline machine learning algorithms from scratch with python. https://machinelearningmastery.com/implement-baseline-machine-learning-algorithms-scratch-python/. Visited on 2019-4-24.

Byeon, H. (2019). Developing a model to predict the social activity participation of the senior citizens living in south korea by combining artificial neural network and quest algorithm. *International Journal of Engineering & Technology*, 8(1.4):214–221.

Cao, J., Guo, J., Li, X., Jin, Z., Guo, H., and Li, J. (2018). Automatic rumor detection on microblogs: A survey. *arXiv preprint arXiv:1807.03505*.

Centraal Bureau voor de Statistiek (2019). Standaard onderwijsindeling 2016. https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/onderwijs-en-beroepen/standaard-onderwijsindeling--soi--/standaard-onderwijsindeling-2016. Visited on 2019-4-24.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Duzan, H. and Shariff, N. S. B. M. (2015). Ridge regression for solving the multicollinearity problem: review of methods and models. *Journal of Applied Sciences*, 15(3):392.

Eime, R. M., Charity, M. J., Harvey, J. T., and Payne, W. R. (2015). Participation in sport and physical activity: associations with socio-economic status and geographical remoteness. In *BMC public health*.

European Commission (2014). The europe 2020 poverty target: lessons learned and the way forward.

Eurostat (2017). Culture statistics - cultural participation by socioeconomic background. https://ec.europa.eu/eurostat/statistics-explained/index.php/Culture_statistics_-_cultural_participation_by_socioeconomic_background. Visited on 2019-2-4.

Gemeente Amsterdam (2018). Stadspas update.

Gemeente Amsterdam (2019). Stadspas. https://www.amsterdam.nl/toerisme-vrije-tijd/stadspas. Visited on 2019-1-7.

Hagberg, A. A. H., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In *The 7th Python in Science Conference (SciPy2008)*, volume 65, page 1–15. Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds).

Hoekman, R., Breedveld, K., and Kraaykamp, G. (2016). Sport participation and the social and physical environment: explaining differences between urban and rural areas in the netherlands. *Leisure Studies*, 36:1–14.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Jannach, D., Lerche, L., Gedikli, F., and Bonnin, G. (2013). What recommenders recommend–an analysis of accuracy, popularity, and sales diversity effects. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 25–37. Springer.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446.

Linden, A. and Yarnold, P. R. (2016). Using data mining techniques to characterize participation in observational studies. *Journal of evaluation in clinical practice*, 22(6):839–847.

McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery*, 46(1):38–41.

Michon, L. and Slot, J. (2014). Armoede in amsterdam een stadsbrede aanpak van hardnekkige armoede. https://www.ois.amsterdam.nl/nieuws/armoede-in-amsterdam-een-stadsbrede-aanpak-van-hardnekkige-armoede. Visited on 2019-1-7.

Mikhail, K. and Konstantin, L. (2017). ELI5: Permutation importance. https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html. Visited on 2019-4-24.

Molnar, C. et al. (2018). Interpretable machine learning: A guide for making black box models explainable. *E-book at< https://christophm. github. io/interpretable-ml-book/>, version dated*, 10.

Nagel, I. (2009). Cultural participation between the ages of 14 and 24: Intergenerational transmission or cultural mobility? *European Sociological Review - EUR SOCIOL REV*, 25.

Nationale Onderwijsgids. Wanneer mag mijn kind naar de basisschool? https://www.nationaleonderwijsgids.nl/voortgezet-onderwijs/paginas/wat-is-voortgezet-onderwijs.html. Visited on 2019-7-15.

Nelson, J. B., Kennedy, W. G., and Krueger, F. (2016). Exploratory models of trust with empirically-inferred decision trees. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 42–50. Springer.

Notten, N., Lancee, B., van de Werfhorst, H. G., and Ganzeboom, H. B. G. (2015). Educational stratification in cultural participation: cognitive competence or status motivation? *Journal of Cultural Economics*, 39(2):177–203.

Onderzoek, Informatie en Statistiek (2017). Amsterdamse armoedemonitor 2017. https://www.ois.amsterdam.nl/downloads/nieuws/2018_armoedemonitor%202017.pdf. Visited on 2019-1-7.

Onderzoek, Informatie en Statistiek (2018). Amsterdam in cijfers 2018. https://www.ois.amsterdam.nl/downloads/pdf/2018%20jaarboek%20amsterdam%20in%20cijfers.pdf. Visited on 2019-1-7.

Onderzoek, Informatie en Statistiek (2019). Basisbestand gebieden amsterdam (bbga). https://data.amsterdam.nl/datasets/G5JpqNbhweXZSw/. Visited on 2019-1-7.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality Quantity*, 41:673–690.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Piscopo, A., Siebes, R., and Hardman, L. (2017). Predicting sense of community and participation by applying machine learning to open government data. *Policy and Internet*, 9:55–75.

Powell, D., Yu, J., DeWolf, M., and Holyoak, K. J. (2017). The love of large numbers: a popularity bias in consumer choice. *Psychological science*, 28(10):1432–1442.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Rijksoverheid. Wanneer mag mijn kind naar de basisschool? https://www.rijksoverheid.nl/onderwerpen/basisonderwijs/vraag-en-antwoord/wanneer-mag-mijn-kind-naar-de-basisschool. Visited on 2019-7-15.

Rijksoverheid (2019). Armoede verminderen. https://www.rijksoverheid.nl/onderwerpen/armoedebestrijding/armoede-verminderen. Visited on 2019-1-7.

Ruseski, J. E., Humphreys, B. R., Hallmann, K., and Breuer, C. (2011). Family structure, time constraints, and sport participation. *European review of aging and physical activity*, 8(2):57.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Series in Artificial Intelligence. Prentice Hall, Upper Saddle River, NJ, third edition.

Shenassa, E. D., Liebhaber, A., and Ezeamama, A. (2006). Perceived Safety of Area of Residence and Exercise: A Pan-European Study. *American Journal of Epidemiology*, 163(11):1012–1017.

Shimoda, A., Ichikawa, D., and Oyama, H. (2018). Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. *Computer methods and programs in biomedicine*, 163:39–46.

Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

Stubseid, S. and Arandjelovic, O. (2018). Machine learning based prediction of consumer purchasing decisions: The evidence and its significance.

Toepoel, V. (2011). Cultural participation of older adults: Investigating the contribution of lowbrow and highbrow activities to social integration and satisfaction with life. *American Economic Review - AER*, 10:123–129.

United Nations (2019). Goal 1: End poverty in all its forms everywhere. https://www.un.org/sustainabledevelopment/poverty/. Visited on 2019-1-7.

Van Wel, F., Couwenbergh-Soeterboek, N., Couwenbergh, C., Ter Bogt, T., and Raaijmakers, Q. (2006). Ethnicity, youth cultural participation, and cultural reproduction in the netherlands. *Poetics*, 34(1):65–82.

Walker, S. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1/2):167–179.

Walsh, C. G., Ribeiro, J. D., and Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, 5(3):457–469.

Yoshikawa, H., Lawrence Aber, J., and Beardslee, W. (2012). The effects of poverty on the mental, emotional, and behavioral health of children and youth implications for prevention. *The American psychologist*, 67:272–84.

Appendices

A Feature importance scores

The following tables show the complete feature importance scores for all features. Logit denotes the log odds of the Logistic Regression model. *Gini* denotes impurity reduction in the Random Forest model. *Gain* is a measurement of how good a tree is in XGBoost. *Permutation* is the permutation importance.

	LR				RF		XGB	
Feature	Logit	*SE*	*P*	Permutation	Gini	Permutation	Gain	Permutation
Age	-0.271	0.016	0.0000	0.012	0.417	0.19	0.251	0.205
Education (not yet) known	-0.224	0.055	0.0001	0.0000	0.016	0.008	0.025	0.0000
Education high	0.695	0.06	0.0000	0.006	0.016	0.009	0.103	0.008
Education low	-0.342	0.055	0.0000	0.004	0.024	0.016	0.026	0.003
Education middle	0.152	0.056	0.0069	0.0000	0.009	0.007	0.047	0.002
Education not applicable	-0.651	0.238	0.0063	0.0000	0.0000	0.0000	0.004	0.0000
Gender (F)	0.4	0.015	0.0000	0.005	0.026	0.027	0.058	0.008
Is child	1.451	0.073	0.0000	0.105	0.118	0.005	0.0000	0.0000
Is parent	-0.158	0.065	0.0155	0.001	0.013	0.007	0.107	0.0000
$\geq$ 2 adults with children	-0.06	0.08	0.4563	0.0000	0.012	0.004	0.007	0.0000
$\geq$ 2 adults without children	-0.368	0.065	0.0000	0.0000	0.013	0.003	0.023	0.001
Single	-0.043	0.064	0.5007	0.0000	0.031	0.008	0.009	0.002
Single adult with children	0.101	0.08	0.2059	0.0000	0.041	0.01	0.013	0.001
Household size	0.236	0.016	0.0000	0.009	0.064	0.019	0.011	0.004
School Allowance	0.351	0.035	0.0000	0.003	0.013	0.014	0.176	0.005
Free PC	-0.234	0.057	0.0000	0.0000	0.002	0.001	0.007	0.001
Free Public Transport	0.371	0.028	0.0000	0.002	0.012	0.005	0.019	0.001
Income Support	0.176	0.02	0.0000	0.0000	0.012	0.009	0.007	0.0000
Western population %	0.087	0.022	0.0001	0.001	0.017	0.011	0.013	0.001
Safety index	0.021	0.011	0.0692	0.0000	0.015	0.008	0.008	0.0000
Population density	-0.032	0.011	0.005	0.0000	0.016	0.008	0.007	0.001
65+ population %	-0.005	0.013	0.6791	0.0000	0.015	0.007	0.007	0.0000
Families with children %	-0.005	0.016	0.7324	0.0000	0.015	0.008	0.008	0.0000
Average residence duration	0.09	0.011	0.0000	0.001	0.016	0.009	0.013	0.003
Working population	-0.012	0.008	0.1441	0.0000	0.015	0.007	0.008	0.001
Social cohesion	0.086	0.01	0.0000	0.001	0.015	0.01	0.016	0.002
Low income %	0.009	0.002	0.0002	0.0000	0.012	0.006	0.007	0.0000
Highly educated %	0.006	0.002	0.0002	0.002	0.015	0.01	0.011	0.002
Number of City Pass locations	0.025	0.01	0.0133	0.0000	0.009	0.005	0.008	0.0000

LR = Logistic Regression, RF = Random Forest, XGB = XGBoost.

Table A1: Feature importance scores for all models and metrics, for predicting City Pass unique use.

	LR				RF		XGB	
Feature	Logit	*SE*	*P*	Permutation	Gini	Permutation	Gain	Permutation
Age	-0.226	0.031	0.0000	0.018	0.08	0.051	0.023	0.044
Education (not yet) known	-0.094	0.124	0.4506	0.0000	0.009	0.006	0.008	0.0000
Education high	0.47	0.128	0.0002	0.005	0.012	0.01	0.036	0.004
Education low	-0.447	0.125	0.0003	0.008	0.009	0.009	0.027	0.006
Education middle	0.001	0.125	0.9962	0.0000	0.006	0.004	0.012	0.001
Education not applicable	0.045	0.623	0.9425	0.0000	0.0000	0.0000	0.0000	0.0000
Gender (F)	0.078	0.03	0.0087	0.0000	0.009	0.007	0.009	0.001
Is child	-0.696	0.175	0.0001	0.021	0.021	0.011	0.085	0.007
Is parent	0.0000	0.163	0.9989	0.0000	0.006	0.003	0.012	0.0000
$\geq$ 2 adults with children	0.089	0.183	0.629	0.002	0.007	0.003	0.009	0.001
$\geq$ 2 adults without children	-0.109	0.145	0.4516	0.001	0.004	0.002	0.008	0.001
Single	-0.046	0.137	0.7365	0.001	0.014	0.012	0.114	0.002
Single adult with children	0.041	0.182	0.8222	0.0000	0.006	0.005	0.008	0.001
Household size	-0.142	0.027	0.0000	0.001	0.037	0.028	0.046	0.017
Western population %	0.172	0.044	0.0001	-0.001	0.029	0.021	0.027	0.005
Safety index	0.044	0.021	0.0377	0.002	0.018	0.013	0.01	0.003
Population density	0.021	0.021	0.3328	-0.001	0.02	0.013	0.009	0.003
65+ population %	0.048	0.024	0.0437	0.001	0.016	0.011	0.009	0.003
Families with children %	-0.014	0.03	0.6345	-0.001	0.028	0.022	0.021	0.004
Average residence duration	0.083	0.021	0.0001	0.004	0.015	0.009	0.009	0.002
Working population	0.008	0.019	0.6749	0.0000	0.019	0.013	0.01	0.002
Social cohesion	0.1	0.018	0.0000	0.0000	0.015	0.011	0.012	0.003
Low income %	0.024	0.005	0.0000	0.009	0.015	0.009	0.007	0.002
Highly educated %	0.022	0.003	0.0000	0.017	0.024	0.019	0.024	0.003
Mean price	-0.563	0.028	0.0000	0.019	0.015	0.003	0.069	0.002
Mean discount	0.353	0.013	0.0000	0.03	0.112	0.037	0.073	0.064
Original mean price	0.025	0.009	0.0078	0.0000	0.111	0.032	0.044	0.009
Rating	0.171	0.02	0.0000	0.008	0.05	0.02	0.053	0.019
Number of ratings	0.5	0.022	0.0000	0.059	0.219	0.143	0.21	0.139
Travel distance	0.192	0.017	0.0000	0.013	0.074	0.032	0.015	0.019

LR = Logistic Regression, RF = Random Forest, XGB = XGBoost.

Table A2: Feature importance scores for all models and metrics, for predicting museum visit.

	LR				RF		XGB	
Feature	Logit	*SE*	*P*	Permutation	Gini	Permutation	Gain	Permutation
Age	0.138	0.043	0.0014	0.004	0.077	0.03	0.023	0.071
Education (not yet) known	-0.106	0.19	0.5791	0.001	0.009	0.002	0.014	0.002
Education high	0.507	0.195	0.0093	0.006	0.008	0.002	0.03	0.003
Education low	-0.743	0.192	0.0001	0.01	0.021	0.012	0.05	0.019
Education middle	0.122	0.192	0.5243	0.001	0.004	0.001	0.008	0.001
Education not applicable	0.035	1.122	0.9755	0.0000	0.0000	0.0000	0.0000	0.0000
Gender (F)	0.173	0.038	0.0000	0.001	0.005	0.001	0.009	0.004
Is child	0.498	0.236	0.0349	0.005	0.023	0.005	0.085	0.004
Is parent	-0.432	0.219	0.0489	0.001	0.011	0.003	0.045	0.005
$\geq$ 2 adults with children	-0.142	0.262	0.5883	0.001	0.007	0.002	0.016	0.005
$\geq$ 2 adults without children	-0.408	0.214	0.0569	0.001	0.005	0.001	0.013	0.002
Single	0.325	0.205	0.1134	0.006	0.02	0.006	0.052	0.002
Single adult with children	0.04	0.26	0.8777	0.0000	0.007	0.002	0.012	0.002
Household size	0.011	0.036	0.766	0.0000	0.037	0.011	0.022	0.009
Western population %	0.172	0.056	0.0022	0.008	0.047	0.01	0.042	0.005
Safety index	-0.02	0.027	0.4663	0.0000	0.024	0.005	0.019	0.003
Population density	0.124	0.028	0.0000	0.001	0.022	0.006	0.011	0.005
65+ population %	0.163	0.031	0.0000	0.004	0.013	0.004	0.013	0.004
Families with children %	0.162	0.039	0.0000	0.003	0.04	0.009	0.026	0.005
Average residence duration	0.009	0.026	0.7209	0.0000	0.013	0.004	0.01	0.004
Working population	-0.011	0.018	0.5501	0.0000	0.023	0.005	0.012	0.003
Social cohesion	0.004	0.024	0.8766	0.0000	0.012	0.004	0.009	0.002
Low income %	-0.011	0.006	0.0841	0.001	0.015	0.005	0.012	0.003
Highly educated %	0.013	0.004	0.0008	0.008	0.049	0.01	0.041	0.004
Mean price	-0.203	0.022	0.0000	0.009	0.048	0.011	0.05	0.051
Mean discount	0.455	0.033	0.0000	0.023	0.094	0.008	0.099	0.024
Original mean price	0.411	0.037	0.0000	0.015	0.126	0.015	0.083	0.021
Rating	0.199	0.046	0.0000	0.003	0.087	0.009	0.087	0.009
Number of ratings	0.284	0.032	0.0000	0.005	0.045	0.002	0.079	0.002
Travel distance	-0.49	0.026	0.0000	0.044	0.108	0.025	0.027	0.04

LR = Logistic Regression, RF = Random Forest, XGB = XGBoost.

Table A3: Feature importance scores for all models and metrics, for predicting cinema visit.

	LR				RF		XGB	
Feature	Logit	*SE*	*P*	Permutation	Gini	Permutation	Gain	Permutation
Age	0.361	0.053	0.0000	0.025	0.105	0.077	0.03	0.141
Education (not yet) known	0.178	0.222	0.4213	0.001	0.01	0.004	0.019	0.002
Education high	0.03	0.229	0.8976	0.0000	0.002	0.001	0.014	0.002
Education low	-0.307	0.221	0.165	0.004	0.009	0.006	0.028	0.023
Education middle	0.039	0.224	0.8616	-0.001	0.003	0.001	0.013	0.003
Education not applicable	0.011	1.337	0.9937	0.0000	0.0000	0.0000	0.0000	0.0000
Gender (F)	0.43	0.048	0.0000	0.013	0.017	0.013	0.039	0.019
Is child	0.321	0.27	0.234	0.002	0.025	0.007	0.138	0.001
Is parent	-0.157	0.248	0.5255	0.001	0.009	0.002	0.014	0.002
$\geq$ 2 adults with children	-0.135	0.304	0.657	-0.001	0.029	0.014	0.068	0.006
$\geq$ 2 adults without children	-0.158	0.25	0.5265	0.001	0.003	0.002	0.016	0.003
Single	-0.034	0.241	0.8893	0.0000	0.014	0.006	0.031	0.003
Single adult with children	0.276	0.301	0.3578	0.006	0.011	0.005	0.021	0.008
Household size	-0.274	0.043	0.0000	0.015	0.049	0.024	0.051	0.036
Western population %	0.105	0.068	0.123	0.004	0.035	0.016	0.026	0.014
Safety index	0.001	0.034	0.9798	0.0000	0.02	0.009	0.02	0.011
Population density	-0.048	0.033	0.1462	-0.001	0.018	0.008	0.017	0.01
65+ population %	-0.027	0.037	0.4635	0.0000	0.016	0.006	0.016	0.011
Families with children %	-0.041	0.048	0.3891	0.0000	0.022	0.008	0.019	0.012
Average residence duration	0.149	0.032	0.0000	0.003	0.018	0.009	0.02	0.013
Working population	0.018	0.031	0.5555	-0.001	0.019	0.007	0.017	0.009
Social cohesion	0.073	0.029	0.0119	0.001	0.014	0.006	0.019	0.008
Low income %	-0.001	0.007	0.8439	0.0000	0.015	0.005	0.016	0.013
Highly educated %	0.006	0.005	0.2233	0.003	0.027	0.01	0.03	0.005
Mean price	-0.789	0.061	0.0000	0.031	0.066	0.007	0.055	0.055
Mean discount	0.66	0.033	0.0000	0.077	0.103	0.027	0.119	0.097
Original mean price	0.395	0.018	0.0000	0.028	0.107	0.029	0.014	0.004
Rating	0.322	0.029	0.0000	0.012	0.058	0.015	0.03	0.008
Number of ratings	0.24	0.039	0.0000	0.007	0.102	0.032	0.079	0.083
Travel distance	-0.194	0.028	0.0000	0.005	0.075	0.035	0.022	0.077

LR = Logistic Regression, RF = Random Forest, XGB = XGBoost.

Table A4: Feature importance scores for all models and metrics, for predicting theatre visit.

	LR				RF		XGB	
Feature	Logit	*SE*	*P*	Permutation	Gini	Permutation	Gain	Permutation
Age	-0.239	0.031	0.0000	0.005	0.082	0.047	0.031	0.053
Education (not yet) known	-0.124	0.113	0.2737	0.002	0.007	0.004	0.01	0.001
Education high	0.153	0.117	0.1938	0.001	0.003	0.002	0.015	0.001
Education low	-0.204	0.112	0.067	-0.001	0.004	0.002	0.011	0.001
Education middle	-0.049	0.114	0.6629	0.0000	0.003	0.001	0.01	0.001
Education not applicable	0.206	0.563	0.7142	0.0000	0.0000	0.0000	0.0000	0.0000
Gender (F)	-0.019	0.026	0.4674	0.0000	0.007	0.005	0.011	0.003
Is child	0.302	0.175	0.0843	0.002	0.022	0.008	0.103	0.006
Is parent	-0.167	0.162	0.3042	0.0000	0.005	0.002	0.017	0.0000
$\geq$ 2 adults with children	0.079	0.172	0.6471	-0.001	0.005	0.003	0.012	0.002
$\geq$ 2 adults without children	-0.25	0.133	0.0597	-0.001	0.003	0.001	0.012	0.0000
Single	-0.099	0.124	0.4239	0.0000	0.01	0.005	0.016	0.002
Single adult with children	0.252	0.171	0.141	0.001	0.008	0.006	0.015	0.003
Household size	0.109	0.022	0.0000	0.0000	0.025	0.012	0.013	0.01
Western population %	-0.121	0.037	0.0013	0.003	0.028	0.009	0.023	0.008
Safety index	0.082	0.02	0.0000	0.001	0.019	0.006	0.014	0.005
Population density	0.076	0.019	0.0000	0.002	0.021	0.007	0.017	0.003
65+ population %	0.139	0.021	0.0000	0.002	0.019	0.006	0.015	0.005
Families with children %	0.16	0.026	0.0000	0.002	0.023	0.008	0.02	0.005
Average residence duration	-0.026	0.018	0.1615	0.0000	0.018	0.007	0.01	0.003
Working population	0.039	0.013	0.0024	0.0000	0.02	0.006	0.018	0.004
Social cohesion	0.108	0.017	0.0000	0.003	0.017	0.006	0.018	0.004
Low income %	-0.017	0.004	0.0001	0.001	0.016	0.005	0.013	0.003
Highly educated %	0.008	0.003	0.0029	0.003	0.027	0.009	0.027	0.004
Mean price	0.113	0.013	0.0000	0.003	0.092	0.038	0.094	0.054
Mean discount	0.081	0.007	0.0000	0.0000	0.07	0.011	0.078	0.023
Original mean price	0.097	0.006	0.0000	0.002	0.069	0.01	0.054	0.008
Rating	-0.145	0.013	0.0000	0.006	0.082	0.027	0.093	0.022
Number of ratings	0.591	0.013	0.0000	0.059	0.104	0.033	0.176	0.031
Travel distance	-0.992	0.015	0.0000	0.141	0.192	0.099	0.053	0.108

LR = Logistic Regression, RF = Random Forest, XGB = XGBoost.

Table A5: Feature importance scores for all models and metrics, for predicting sport visit.

B Learning curves

This section includes all the learning curves per category and model. Blue and green shaded areas represent the standard deviations above and below the means of the 5-fold cross validations.

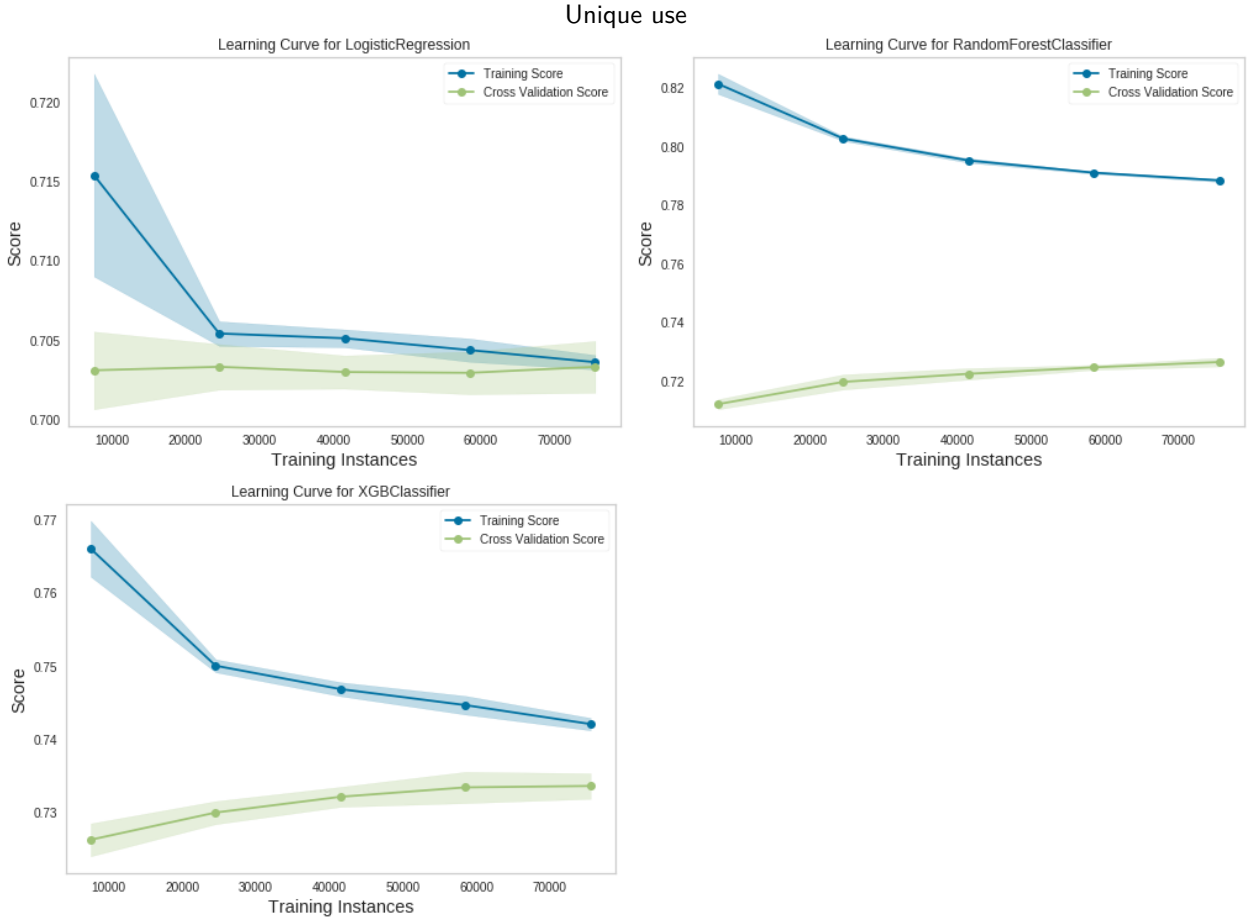


Figure B1: Learning curves for predicting unique use.

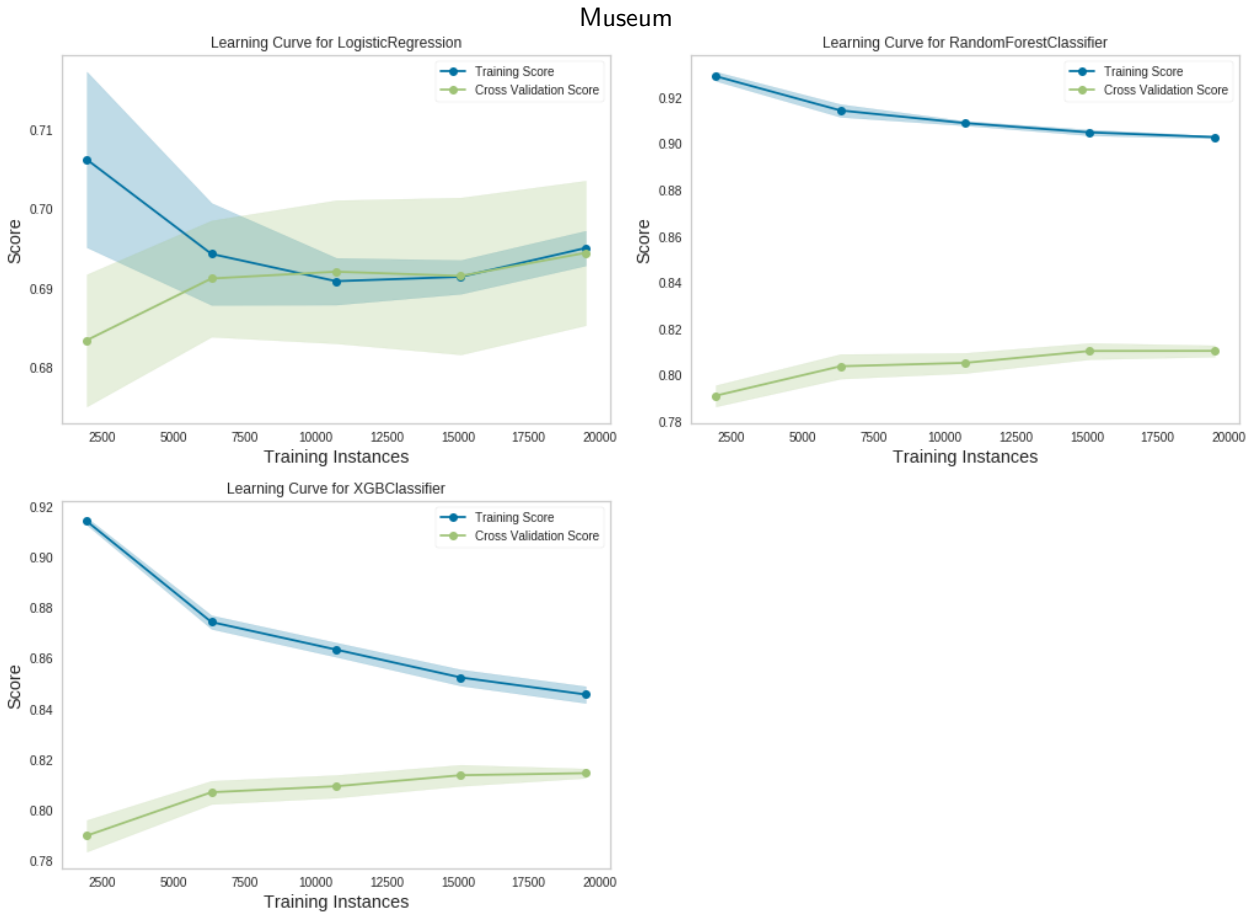


Figure B2: Learning curves for predicting museum visit.

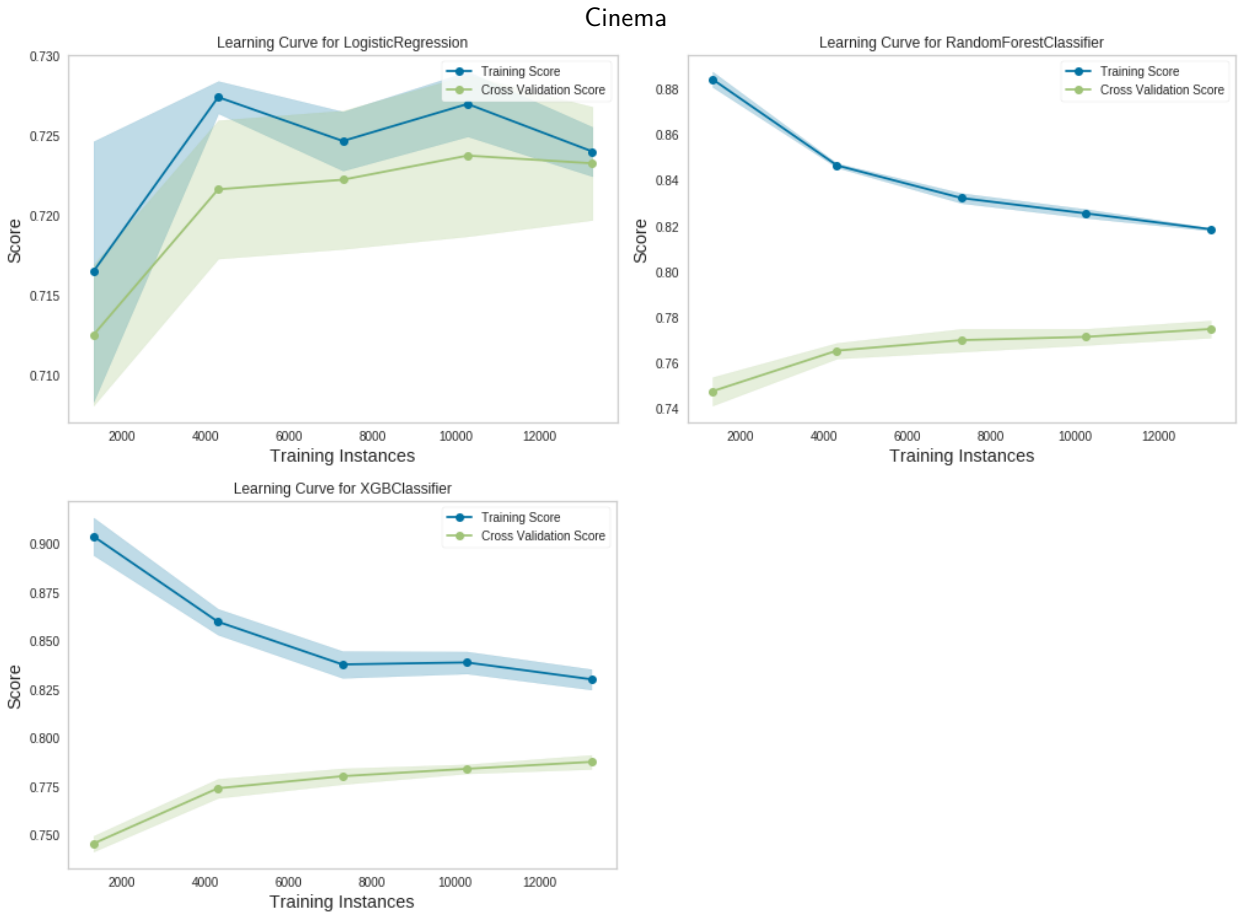


Figure B3: Learning curves for predicting cinema visit.

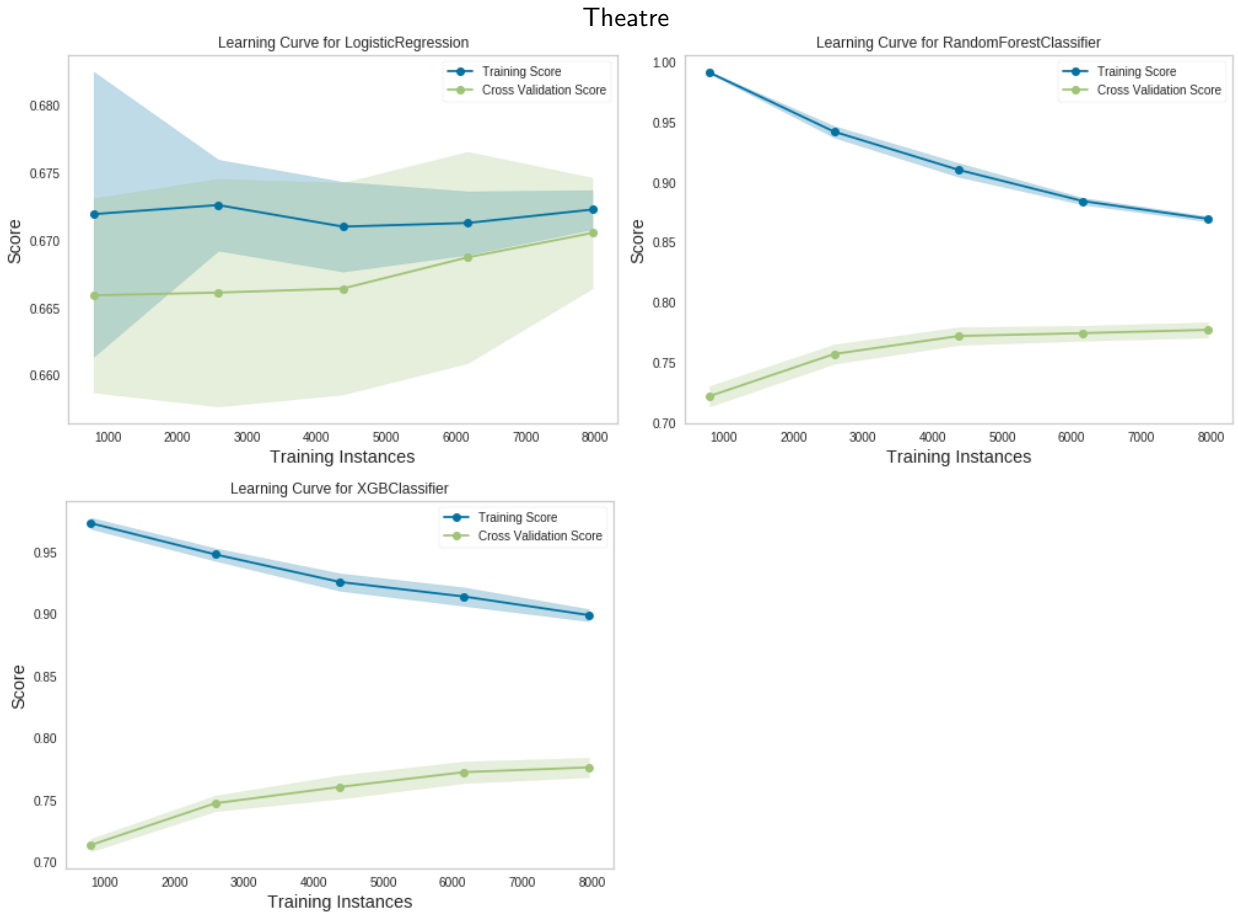


Figure B4: Learning curves for predicting theatre visit.

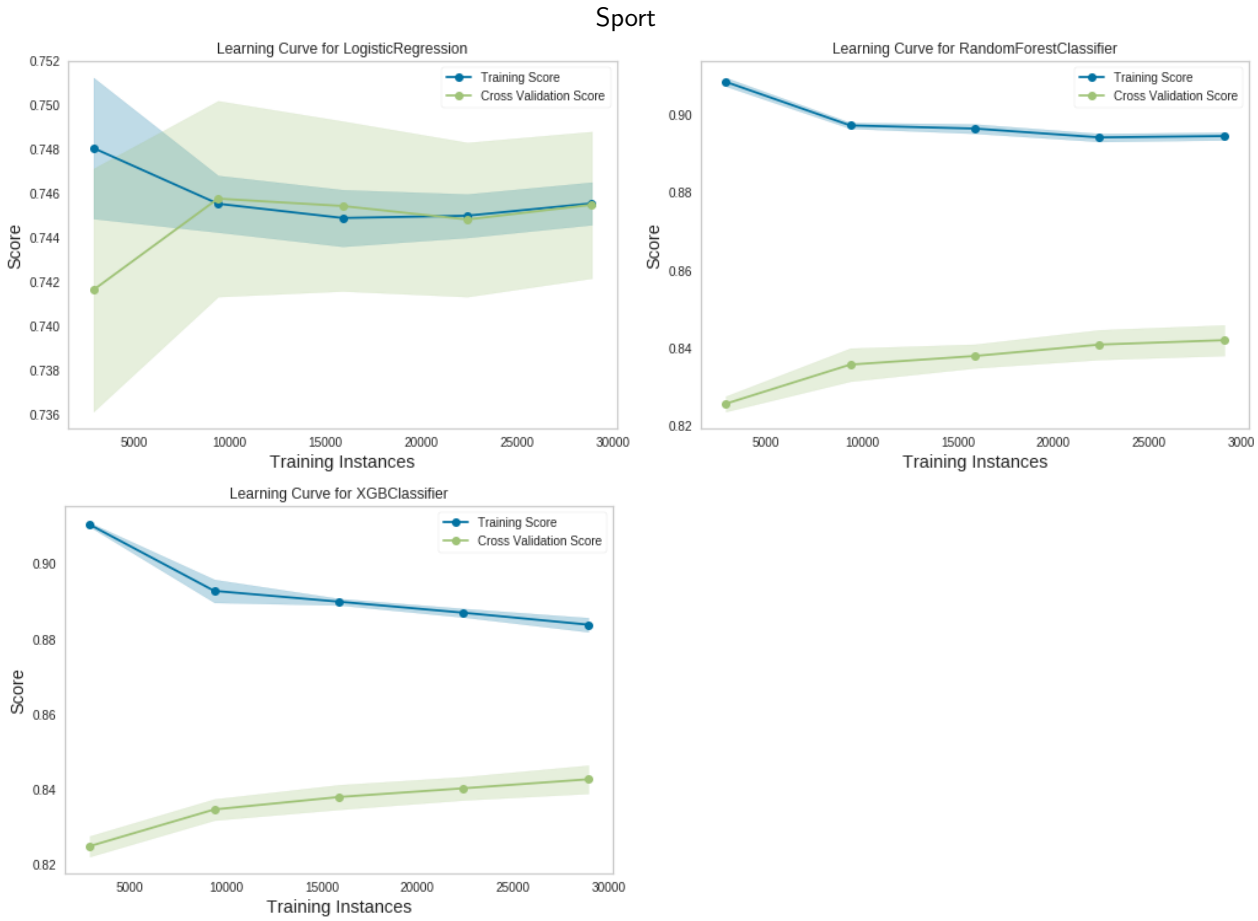


Figure B5: Learning curves for predicting sport visit.