Short-term prediction of platform overloads using supervised machine learning

Jessica Weibel
Institute for Transport Planning and Systems
Swiss Federal Institute of Technology, ETH
Zurich, Switzerland
weibelj@student.ethz.ch

***Abstract*—This paper aims at creating short-term predictions of platform overloads on train station platforms using supervised machine learning. As this is one of the first attempts at predicting passenger flows for a small space and time resolution, the focus is kept on basic algorithms. More specifically, the paper compares the results of the *oneR*, *J48*, *RandomForest* and *NaiveBayes* algorithms using Matthew's Correlation Coefficient (MCC) and the area under the Precision-Recall curve (AuPRC) as performance indicators. The input data for the algorithms, i.e. the set of classified instances, describes the occurrence of overloads on the platform by combining operational, platform, and descriptive data. The overload class is defined using a theoretical definition based on the widely used Level of Service concept for pedestrian facilities. This methodology is tested by creating a prediction model for the Amsterdam Zuid train station in the Netherlands where new and detailed data and evaluation possibilities were available provided by ASE AG and NS Stations. The *J48* algorithm outperforms the others in the final prediction model. However, the results show that accurately predicting overload situations in the short-term future is difficult using basic supervised machine learning algorithms.**

Keywords—Platform overload, Passenger flows, prediction, supervised machine learning.

I. Introduction

Travel demand is constantly increasing around the globe at high rates forcing stations to operate at near, or even over capacity. The capacity at a train station is defined and measured by the number of passengers that can comfortably be present in a given space [1]. However, due to the cyclic nature of pedestrian flows, these issues are often only relevant during the morning and evening peak hours when the flows are the highest. It is crucial for stations to find ways to maximize the usage of their capacities to provide an attractive and safe service for the users with efficient planning and operation [2]. To be able to effectively run a station at its capacity limits at these crucial times, operators need to be able to closely monitor the current situations in the station and reliably predict future scenarios.

With the usage of the newly developed Pedestrian Analytics System (PAS) by ASE AG [3], it is now possible to anonymously track and extensively evaluate pedestrian movement allowing operators to gain valuable information on the usage of the station. The increasing concern around train stations reaching their capacities therefore benefits from the rise and availability of big data to gain a better understanding of the pedestrian distribution to maximize their usage. As the technology of the PAS has emerged and established recently, there is a research gap discovered regarding this detailed space-time knowledge and availability of pedestrian flow data [3]. This immense amount of data lends itself to the application of a machine learning approach, where large data inputs can be analyzed using appropriate algorithms.

II. Literature Review

A. Dimensioning Train Stations

A common assessment scheme used for dimensioning railway stations is the Level of Service (LoS) scheme, where six service levels, ranging from levels A to F, are used to quantify the quality, safety, and comfort of pedestrian facilities. The LoS scheme dates back to 1965 where it was introduced in the Highway Capacity Manual by the Transportation Research Board [4] for the dimensioning of roadways, and was later adapted for pedestrian facilities by Fruin [5]. Small changes and adaptions have been made to these densities over the years. The threshold densities defined in paper, based on [2] are shown in Table 1. LoS C is the standard level of service that should be maintained on the platform when no train is present, while higher densities, up to LoS E, are permitted during the passenger exchange taking place on the side of the train's arrival. LoS F should be avoided because the safety of pedestrians can no longer be guaranteed. Despite common adaptations, there is a lack of information regarding the time component of the LoS schemes, i.e. defining how long that higher densities can be tolerated. This can be linked back to the reduced amount of pedestrian data available at a fine time and spatial scale [6].

TABLE I. Threshold Densities [P/M^2] LoS

LoS	***Walking Areas***	***Waiting and Congestion***
A	0.30	0.55
B	0.45	0.75
C	**0.60**	**0.95**
D	0.75	1.25
E	1.50	2.50
F	>1.50	>2.50

B. Forecasting Passenger flows

Often, forecasts have been done at a larger temporal and spatial aggregation. However, to predict overloads on the platform, a forecast at a much smaller resolution is required. Previously, this smaller aggregation was difficult to achieve due to the lack of data, especially pedestrian data, now provided by PAS [3]. These new pedestrian sensing technologies allow the recording of the walking paths of passengers at a granularity of 0.25 seconds and a high spatial accuracy. These tracks can also be aggregated in space and time to create occupancy readings of pre-defined zones. Data availability, various queries in space and time, and analysis are provided by the PAS in high efficiency, integrated in an encompassing and sophisticated software environment [3]. Considering the newly available data, the relation between the demand and the capacity of the given infrastructure, or a direct prediction of the occurrence of dangerous situations on the train station platform now becomes feasible, however, has not been pursued so far.

III. Methodology

In order to apply the method of supervised machine learning, an input data set as set of classified instances must be given to the chosen algorithms. Here, this set of classified instances must contain features relating to the situation on the platform. These will then be classified by a chosen definition of platform overloads. The exact nature of the input data depends on the information available, however at minimum, the following three categories of data should be included:

- Descriptive data: information about the current situation.
- Operational data: information describing the scheduled timetable and actual train operations.
- Platform data: information describing the movements of pedestrians on the platform from a pedestrian tracking system. An important measure that should be derivable from the output is the density on the platform, since this is often the quantity used for the dimensioning of railway stations for the LoS scheme.

These individual data sources are merged together, depending on what should be predicted. The individual instances in the dataset are then each given their respective classification according to the chosen overload definition. Various appropriate machine learning algorithms were compared to see which model performs best at correctly identifying the platform overloads.

A. Classification

A theoretical definition based on the LoS concept is used for the definition of platform overloads, with the platform functions shown in "Fig 1." The safety area should be avoided, the circulation is used for walking, and the rest is used for waiting and congestion. The missing time component in this dimensioning scheme is set to 30s to avoid creating warnings for short, tolerable spikes in densities created by the fluctuating nature of pedestrian behavior.

The following classes were created:

- No warning
- Precautionary Warning: when the threshold density above LoS C has been reached in one zone for more than 30s.
- Overload Warning: threshold density above LoS C has been in the zone in question and at least one adjacent zone for more than 30 seconds.

Higher densities are permitted on the plaform during passenger exchange. These warnings can be ignored if a train is present.

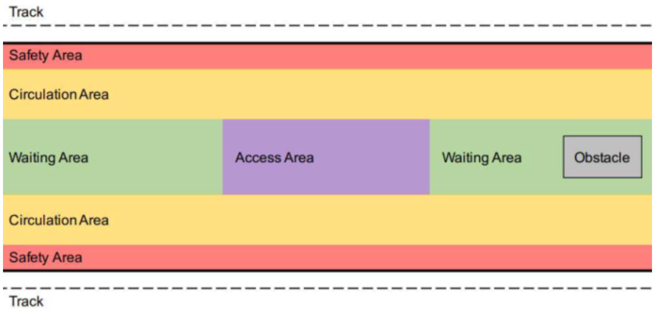

Fig. 1. Platform Functions

B. Model Set-up

The data is stored in a matrix form, where every row in the dataset is an instance of the dataset and every column is a feature, except for the last column which is the class. The dataset was created in a time-based approach, where every instance describes a time on the platform. To provide an easily adaptable model, the model was set-up so that the following variables can be given as an input to data-preprocessing script.

- Time aggregation (*agg*): The duration of time to be represented for every instance.
- Prediction window (*pw*): time difference between the given information and the time for which the prediction should be made.
- Historic window (*hw*): the amount of information from the past minutes the model is given.

C. Machine Learning

In a first attempt to predict platform overloads, basic algorithms will be tested to examine their performance. These basic algorithms are favored since they provide insight to their mapping from input to output. The positive class, here the overload class, is the class of interest. The results of the predictions can be summarized in a confusion matrix, as shown in "Table II", including the assumed cost for its occurrence. The cost of under classifying an instance being twice as bad as an overclassification. The model is trained using cost-sensitive learning, where the goal is to minimize the overall cost.

TABLE II. Confusion Matrix and Associated Cost

Actual Class	*Predicted Class* Overload	No Overload
Overload	true positive (TP) – 0	false positive (FP) - 10
No Overload	false negative (FN) - 5	true negative (TN) - 0

The performance of the algorithms will be judged based on AuPrC (eq. 1) and MCC (eq. 2) where both values have a maximum desired value of 1.0.

$$Precision = \frac{TP}{TP + FP} \quad , \quad Recall = \frac{TP}{TP + FN} \quad . \tag{1}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad . \tag{2}$$

Statistical modelling The *NaivesBayes* algorithm will be tested, where the output is based on the Bayes' Thereom of conditional probability. The assumption is made that all the features are of equal importance and independent of one another. Although this is often not the case in real world data, this algorithm has been shown to perform surprisingly well [8].

Decision Trees Decision trees represent a sequence of decisions that need to be made in order to reach a certain outcome. Three types will be tested. The *oneR* algorithm is a basic decision tree that is limited to only making one rule for classification. The *J48* algorithm is a basic decision tree, which employs the method of pruning to increase its performance. The *RandomForest* is a combination of decision trees, where decision trees are created for random combinations of features and instances, and the final class is predicted using the majority class [8].

Artificial neural network (ann): An ann was tested as a comparison to the basic learning algorithms. In an ann there is always an input layer and output layer, with *x* number of hidden layers in-between. The output is dependent on the input, where the mapping is from the input to the output is done by multiplied the features by arbitrary weights in the number of hidden layers. These weights are optimized through an iterative process [8].

IV. Case Study Amsterdam Zuid

The methodology was tested on the train station Amsterdam Zuid in the Netherlands, with data provided by [3,9] for two train station platforms. Each of these platforms is equipped with the PAS system provided by ASE AG that records the occupancy for 47 zones on the platform on very high time and space accuracy and enabled evaluations and queries. These zones are located near the access areas to the platform. No readings are provided for the rest of the platform. The resulting input set has the following features:

- Description data: time of day, rain, temperature
- Platform data: zone overload
- Operational data: train series, rolling stock, dwell time, delayed arrival, previous train cancelled, phase.

The phase describes the current operational phase of the platform, where 4 phases are defined. The time directly before a train arrival is phase 1, the time between the train arrival and train departure is phase 2, and the time right after the train departure is phase 3. Phase 1 and 3 both last 1 minute. The rest of the time is defined by phase 0, where boarders are slowly making their way to the platform. Phase 2 is a non-safety critical phase, and can thus be neglected in terms of overloads.

The pedestrian flows at the station are similar to those typically seen in train station, with much higher frequencies experienced during the weekdays than on the weekends. This is further induced by the location of the station in a business district. At a time-resolution of 1 hour, the typical morning and evening peaks can be seen, with a maximum peak of around 3'000 passenger per platform per hour. The magnitude of two peaks is comparable. When these flows are separated between boarder and alighters, the location of the business district is distinctly seen with more alighters using the station in the morning and boarders in the evening, representing the people coming to work in the morning and leaving in the evening. No peak is seen towards noon.
The total passengers per hour per platform is show in "Fig. 2".

During the weekend, no peaks are seen, instead a relatively constant hourly flow of around 700 passengers per platform per hour can be observed. At a smaller time scale resolution of 1 min, the oscillating nature of these pedestrian flows is very relevant.

As expected, the class imbalance was very large, with 96.5% of the instances being in class 1, 3.2% in class 2, and 0.3% in class 3. A distribution of these overload classifications throughout the day in 15 minute intervals can be seen "Fig. 3". Class 0, i.e. the negative class, has been removed from the figure to have a better view of the positive class. The figure shows a similar pattern to the pedestrian flows with a morning and evening hour peak, although the passenger flow peaks were comparable during the morning and evening hour peak, this is not the case for the distribution of the overloads since a much larger frequency of overloads is seen in the evening. This is due to the differing behaviour of boarder and alighters, where alighters simply exit the platform as soon as they arrive at the station and boarders linger on the platform waiting for train to arrive. Therefore, this increase in overloads in the evening peak can be explained by the larger proportion of boarders at the station.

During the weekend, no peaks are seen, instead a relatively constant hourly flow of around 700 passengers per platform per hour can be observed. At a smaller time scale resolution of 1 min, the oscillating nature of these pedestrian flows is very relevant.

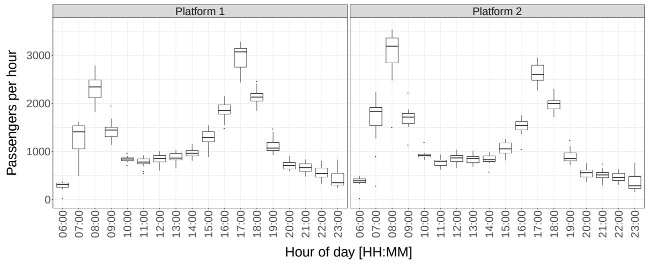

Fig. 2. Total passengers per platform per hour

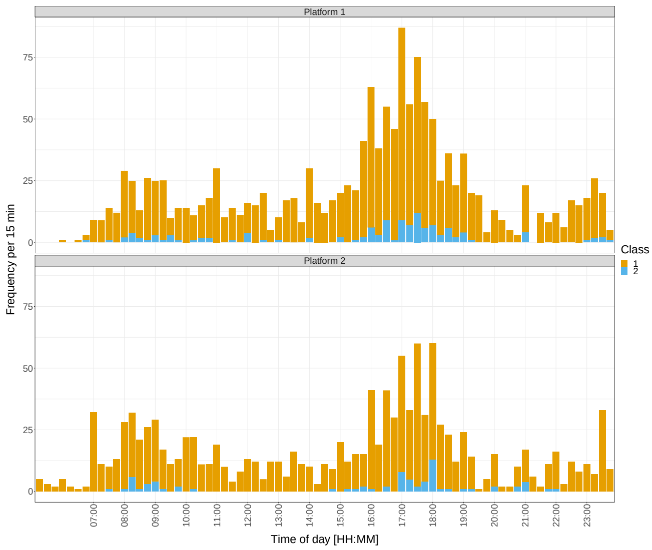

Fig. 3. Distribution of overloads per platform per 15 minutes

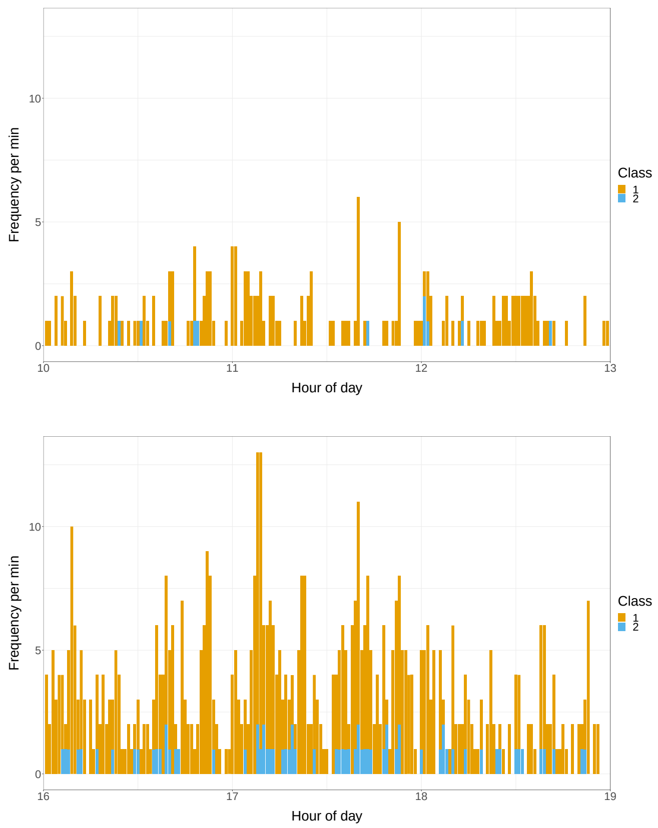

Fig. 4. Distribution of overloads per minute in regular operating hours from 10:00 to 13:00 (above) and in the evening peak hour (below) for platform 1.

As expected, the class imbalance was very large, with 96.5% of the instances being in class 1, 3.2% in class 2, and 0.3% in class 3. A distribution of these overload classifications throughout the day in 15 minute intervals can be seen “Fig. 3”. Class 0, i.e. the negative class, has been removed from the figure to have a better view of the positive class. The figure shows a similar pattern to the pedestrian flows with a morning and evening hour peak, although the passenger flow peaks were comparable during the morning and evening hour peak, this is not the case for the distribution of the overloads since a much larger frequency of overloads is seen in the evening. This is due to the differing behaviour of boarder and alighters, where alighters simply exit the platform as soon as they arrive at the station and boarders linger on the platform waiting for train to arrive. Therefore, this increase in overloads in the evening peak can be explained by the larger proportion of boarders at the station.

Furthermore, it can be seen that Platform 1 has slightly higher pedestrian flows, and thus also a slightly higher number of overloads. Class 0 has been removed from the figure to have a better view of the positive class. The distribution of the overload per minute for platform 1 is shown in “Fig. 4”, where the sporadic nature of the classifications can be seen. In the evening peak hour there is a clearer relation between a minutes’ class and the class before can be seen. Although, large spikes are often followed by low peaks. In the regular operating hours, seemingly random appearances of overloads occur. This could be due to the theoretical approach of overloads applied, where even though an increase in density is seen, and thus an overload is incurred, no dangerous situation happened on the platform.

V. Results

An important step in developing a machine learning model is finding the optimal feature set [10]. Due to the very small representation of class 2, first a binary classification model was attempted. The initial set of classified instances was input to the chosen algorithms, yielding the results shown in “Fig. 5”. It can be seen that all models perform better than a random prediction model but *J48* performs significantly better than the rest of the other algorithms, at a 95% confidence level using a paired t-test, with an MCC = 0.39 and AuPRC = 0.19. Since these metrics maximize at 1, there is some room for improvement. It can be seen that the ann performs very similarly to the *randomForest*, and *NaivesBayes* model, therefore it will not be pursued further for the feature manipulations due to its’ much longer computational times, with computation times over 180x as long as the other models.

The confusion matrix for the initial J48 matrix can be seen in “Table III”, where as expected the number of true negatives is very high, while the number of true positive is relatively low.

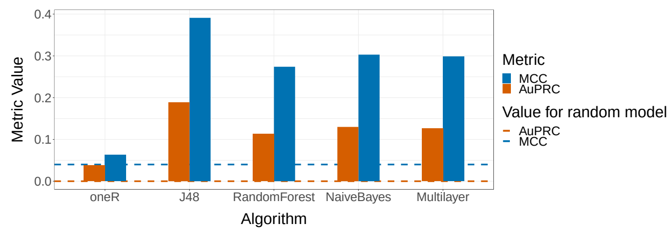

Fig. 5. Initial Results for all the algorithms.

TABLE III. Confusion Matrix J48 initial

Actual Class	*Predicted Class* Overload	No Overload
Overload	26.6 %	73.4 %
No Overload	0.06 %	99.4 %

A. Feature Manipulation

In order to further improve the metrics of the model, different, the features were manipulated to create different model variations. 13 different models were created were the following manipulations were done:

Summarizing platform features: Instead of giving the algorithm the information to all 47 zones, a feature describing the total number of overload zones on one platform side is given. This increased the model the model performance. The J48 model performs best with MCC = 0.402 and AuPRC = 0.197.

Categorizing features: First, the time of day feature given as morning peak, evening peak or regular operating hour. Second, the weekday was given as a weekend or workday. Both these variations decreased the performance of the model.

Model set-up: the model set-parameters were adjusted. First, the historic window was set to 5, then it was reduced to 0. Second, the prediction window was raised to 5. All these variations decreased the performance of the model. This shows that due to the sporadic distribution of the classes the model does not benefit from including too much past, information, however, having information to the direct past improves the model

Focus on overload times: the goal was to reduce the amount of negative class instances in the dataset. This was done by focusing on situations were a high relative frequency of overloads occurred. First, only the phase 1 instances were taken, where phase 1 is the phase just before the train arrives, therefore, high densities are to be expected since it can be assumed that almost all the boarder are on the platform waiting for their train arrival [1]. Secondly, an evening peak hour model was created. Both variations decreased the performance of the model. The peak hour slightly increased the performance of the *NaivesBayes* model, however, due to the significant loss in information by only predicting the evening peak, the performace would have to be substantially higher to make up for this loss worthwhile.

Removal of features: Two extremes were tested. First, the removal of seemingly useless features was tested. Second, the removal of the most important features. The results of the decision tree algorithm were used to decide which information was important based on their information gain. The only feature removal that has a positive effect on the model was removing the features related to the weather, where the *J48* algorithm still performs best, with MCC = 0.403, and AuPRC = 0.198.

Multi-class classification: The multi-class (3 class) classification problem was also attempted to see the effects of the results. Like in the binary classification, all of the models were able to predict more accurately than a random precition model, except for *oneR.* The *J48* algorithm performs best in respect to the preliminary warning class (class 1) but it is outperformed by the *NaivesBayes* model for the (class 2). Therefore, if only the prediction of the extreme class is of interest, then there lies potential within the *NaivesBayes mode*. However, as these models do not perform identical to their binary counterparts, another set of feature manipulation would have to be conducted to be able to confirm that *NaivesBayes* algorithm is the best at predicting the class 2 overloads.

B. Summary

The *J48* algorithm tends to perform best on exactly half of the variations (7/14), with 5 exceptions where it is outperformed by the *NaivesBayes* model and 2 exceptions where it is outperformed by the *RandomForest* algorithm. The *oneR* algorithm always performs significantly worse than the rest of the tested algorithms, meaning that the data cannot be adequately be classified using one simple rule. The performance of the *oneR* algorithm mostly remains the same, since it is often using the same rule to split its root node. This rule is based on the feature total overload south. Since this is the feature with the highest information gain for a tree, it is also expected to be used as the root node for the decision trees constructed in the *J48* algorithm, which is indeed the case. It seems the added randomness in the *RandomForest* model does not help the algorithm outperform its' simpler predecessor in most cases. This could be due to the imbalance in the importance of the variables. Although, one variable is not enough to describe the data well, as shown by the *oneR* algorithm, there is a strong tendency, for the *J48* to favor the "total overload" features and the train information. This shows that these features contain a lot of information for the model. It is assumed that when a random subset is made for the individual trees in the algorithm, and these features are not included, this individual tree will reduce the performance of the algorithm. Since no exact definition of the trees is given in the Weka output, this cannot be confirmed for certain. Finally, the *NaiveBayes* model, outperforms the other algorithms in 5 of the cases, however, these models are all significantly worse than the best *J48* model, except for the evening peak model.

C. Choice

The best performing variation is the *J48* model with the summarized weather data and the weather data removed, with an MCC = 0.403 and AuPRC = 0.198. The resulting confusion matrix can be seen in "Table IV".

TABLE IV. CONFUSION MATRIX J48 FINAL

Actual Class	*Predicted Class* Overload	No Overload
Overload	29.6 %	70.4 %
No Overload	0.07 %	99.3 %

TABLE V. CONFUSION MATRIX J48 FINAL

Actual Class	*Predicted Class* Overload	No Overload
Overload	40.4 %	59.6 %
No Overload	1.9 %	98.1 %

To further increase the number of true positives, the cost difference was increase, by decreasing the cost of misclassifying an overload to 1, resulting in a 10:1 cost matrix. The resulting confusion matrix is shown in "Table V". The most effective way to affect the performance of the model was discovered to be adjusting the cost matrix. Based on this an economically justified cost matrix could be beneficial to the interpretability of the results. Finding the best balance between the four metrics shown in the confusion matrix is a highly practically oriented problem. This depends on what the model is to be used for. The resulting MCC and AuPRC for these models are lower, i.e. their theoretical performance might be worse, but their practical performance or usability is higher. This practical performance or usability depends highly on the exact use of the model, i.e. what action will be taken based on the model's prediction.

VI. Conclusion

The new availability of large amount of precise pedestrian data by ASE's PAS [3] combined with additional data sources provides input data for new and detailed evaluations. The PAS provides input data on precise spatial and temporal scale, enabling machine learning algorithms specifically suitable for "big data". The challenge of predicting train station platform overloads in the short-term remains pertinent. This paper has shown that a basic set of machine learning algorithms, more specifically those based on the concepts of statistical modeling and decision trees, have trouble learning what leads to a platform overload at a high enough accuracy to be implemented in practice. The practical implications of not being able to predict high densities on the platform effects the operations of a train station are directly relevant for the safety and comfort of railway service users and for capacity of railway operators. Due to the cyclic patterns of pedestrian flows, these high peaks are often localized in the morning and evening peaks. In order to try and minimize the difference between the required dimensions between peak hours and non-peak hours, stations need to be able to safely operate near their capacity limits. In the extreme case, where platform or station overloads cannot be predicted with a high enough accuracy, higher infrastructure costs would result since stations would need to be constantly upgraded/extended to be able to handle the large pedestrian flows experienced during the peak hours. Therefore, the benefit in achieving these accurate predictions, lies in the maximal use of a station's capacity, and thus the reduced need for infrastructure expansions. The more accurately these overloads can be predicted, the closer to capacity station can be operated without the risk of dangerous situations occurring.

This paper also showed the significant effect the cost matrix for cost-sensitive learning had on the prediction models. Instead of applying arbitrary cost ratios between the possible misclassification, an in-depth analysis of the economic cost of a misclassification would be an interesting component to include in such a model. This cost would again be very dependent on which actions are taken as a result of the given prediction.

VII. Outlook

A. *Defining platform overloads*

Defining the platform overloads was done using the widely used LoS scheme for dimensioning railway stations. This is a very theoretical approach. The benefit of using a theoretical approach lies in the possibility to create a universal model that could be applied for any given train station, so long as operation data, platform data, and descriptive data are available. Most importantly, pedestrian tracking data on the platform which can be converted into density readings are required using sophisticated technologies such as the PAS developed by ASE AG [3]. Another important part of this theoretical platform overload definition is its time component. This time should potentially be defined in correlation with the zone size, where smaller zones could have higher densities since the pedestrian will experience them for shorter periods of time. If data of the individual trajectories are available, it would be beneficial to try and use the average density experienced by one pedestrian during their time on the platform, instead of the density of a zone, since the high zone densities are likely experienced by different people.

This theoretical definition of platform overloads could potentially lead to many random overloads where, although high densities in a zone were noted, no dangerous situation occurred. An example that could lead to this is a small group standing in a zone together. An alternative approach to attempt to create a more practical definition of overloads would be to have an operator closely monitor the situation on a platform and define the platform overload based on his expert opinion. This would mean that no direct link between the features and the output and no specific classification rule could be given. Instead, the expert opinion of a train station operator would be used, creating a more practical definition of a platform overload. Ideally, these experts would judge the situation in a similar manner, however, this cannot be guaranteed. There would be an increased bias due to the human factor added, where not every train operator would necessarily have the same expert opinion. This could either be done with video footage of historical events or the collection of this information could begin with new situations. This would, however, be a very station-specific, or even platform-specific approach, and would not allow for an general model for all train station platforms.

B. *Machine Learning*

As this was one of the first attempts at using machine learning in a platform overload prediction, the focus was set on basic algorithms. It has often been shown that these basic algorithms tend to be overlooked even though they can produce comparable results. However, with the best model a true positive rate of 40.4%, can be achieved. It remains unclear if an improvement can be achieved using more complex algorithms, or even ensembles of complex algorithms. However, it is believed that these should be explored further, especially with the introduction of larger quantities of data. No guarantee can be made that more data will lead to better results in this specific case but either way it would allow the model to have more examples of the concept it is trying to learn.

Finally, the ability to accurately predict dangerous situations on train stations is important in order to provide a safe and attractive railway service. As the availability of data continues to increase, machine learning algorithms can and should be utilized to take advantage of these large quantities of data and evaluation possibilities provided through new technologies such as the PAS developed by ASE AG [3].

VIII. ACKNOWLEDGEMENT

I would like to thank Prof. Dr. Francesco Corman, Beda Büchel, and Dr. Basil Vitins for their support and guidance throughout my thesis. For providing the required data, I would like to thank NS Stations, specifically Jeroen van den Heuvel, but also for his prompt responses to all of my questions. In addition, I would like to give a big thank you to the whole team at ASE AG for providing access to PAS, its underlying structures and software, and evaluation possibilities. Additionally, I thank the team of ASE AG for the support, for taking the time to answer any of my questions, and for the constructive discussions over a round of billiard. And last but certainly not least, to all my friends and family, thank you for providing me with the much-appreciated support, and the necessary distractions to clear my mind.

[1] Verband öffentlicher Verkehr (VöV) (2017) *Planungshilfe Publikumsanlagen*.

[2] Weidmann, U., E. Kirsch, U. annd Pue, D. Jacobs, C. Pestalozzi and V. Conrad (2013) Verkehrsqualität und Leistungsfähigkeit von Anlagen des leichten Zweirad- und des Fussgängerverkehrs, Eidgenössisches Departement für Umwelt, Verkehr, Energie und Kommu- nikation (UVEK).

[3] ASE AG, https://www.ase.ch/

[4] Transportation Research Board (1965) *Highway Capacity Manual*, Washington, D.C, HRB Special Report 87 edn.

[5] Fruin, J. J. (1971) *Pedestrian planning and design*, Metropolitan Association of Urban Designers and Environmental Planners, New York.

[6] Bosina, E., M. Meeder and U. Weidmann (2018) Defining the time component of the pedestrian los concept, Washington, D.C.

[7] Witten, I. H., E. Frank and M. A. Hall (2011) *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann

[8] NS Stations (2019) Zug daten amsterdam zuid march 2019, *Operation Data*.