How do Tourists Travel with Public Transportation?

Tobias Johannesson
tjohann@kth.se

May 19, 2024

Abstract

Travel data captured using automated systems allow service providers the opportunity to explore and optimize their offerings. By looking at temporal patterns for adult, single, and tourist ticket categories based on the ticket type it could showcase how ticket offerings can be optimized. User profiles consisting of day of the week, and hour slice of the day are created based on a week from May, and then a week from July. A RandomForest model is trained which classify users into one of these three categories with an accuracy of roughly 80% where tourists hold a F1-score of around 50% indicating either under representation in the data, overlap in temporal travel patterns, or need for additional data in classification such as spatial. A seasonal change from May to July show how adult users going on vacation can be seen, and a temporary reparation of trains show users utilizing replacement buses. The research show distinct behavior for each group and 40% points improved performance over random guessing. These findings could be used to further tailor offerings to users going forward, and to potentially evaluate user segmentation.

Contents

1 Introduction

Public transportation systems are essential for urban mobility, enabling millions of people to commute daily worldwide [1]. The rise of automated data collection systems, such as smart card and digital ticketing platforms, has transformed the way we capture and analyze transit patterns [2]. These systems not only streamline fare collection but also generate extensive datasets that detail the origin, destination, and timing of individual journeys. This wealth of data offers new opportunities to explore and optimize the services being offerings.

Understanding how the public transportation service is utilized today can provide valuable insights into user needs for the future. One such strategy involves segmenting users based on their travel patterns, allowing for targeted improvements based on the segment's size, impact, or cost [1]. However, smart card data is often anonymized, making it challenging to link data directly to individual users with the potential for multiple users to utilize the same card. User segmentation can therefore be complicated by such unknowns, resulting in analyses that do not accurately reflect reality [3].

To address these challenges, my project will leverage domain knowledge to categorize users and then train a classification model to assign users to one of these predefined categories. This classification model help utilize knowledge already acquired, and provides robust evaluation metrics that will help evaluate the current system. By using a combination of manual classification and machine learning, we aim to both improve the accuracy of user segmentation and to improve current understanding of user groups. This approach will provide deeper insights into user behavior which can help segment users more distinctly into groups based on temporal patterns building on the research done on the same dataset in [4]. These findings can in turn enable more effective planning and optimization of public transportation services with potentially more tailored offerings.

1.1 Problem Statement

In public transportation systems, accurately grouping users based on their travel patterns is challenging. Moreover, understanding how these user groups behave and even change over time can provide service providers with valuable metrics on current system performance. Traditional user segmentation methods are often unsupervised, lacking true labels or clear evaluation metrics. This makes it difficult to assess the accuracy and relevance of the segments currently identified. Relying solely on domain knowledge may not fully capture the complexity of user behavior. Additionally, user behavior is complex with a dynamic system offering can and sometimes need to change. For this it is important to understand how current offerings are utilized to see how the impact is received and by whom.

1.2 Research Question

This project aims to investigate how domain knowledge can be leveraged to identify and evaluate users based on their travel behavior. Specifically, it will explore the following research question: **"How can domain knowledge be utilized to accurately segment public transportation users based on their travel behavior, and how do these segments evolve over time?"** By addressing this question, the project seeks to test the feasibility of using domain knowledge for clear customer segmentation and to identify gaps where domain knowledge may fall short in encapsulating user behavior. The outcomes will provide insights into the dynamics of user segments and inform strategies for improving public transportation services.

2 Related Works

Previous research has extensively explored the analysis of public transportation data, particularly focusing on smart card data and other digital footprints of commuter behavior. These studies have illuminated patterns of user mobility, and the dynamics of transit networks [5, 6, 2, 7]. However, a common assumption in many of these analyses is the static nature of user behavior over time, potentially overlooking the fluidity and adaptability of commuter patterns in response to network changes, seasonal shifts, or infrastructural updates [1, 8].

2.1 Clustering and Community Detection in Public Transportation

Clustering within public transportation networks has been a focal point in identifying communities of stations that share similar usage patterns or serve similar user demographics. Representing station data as graph structures allows for the application of network analysis techniques to identify critical nodes and pathways. This approach provides insights into peak usage times, preferred routes, and the impact of network changes on travel behavior. For example, studies employing k-means and hierarchical clustering have successfully segmented users based on their travel regularity and patterns, revealing distinct commuter profiles that can inform service planning and targeted marketing strategies. [4][3]

2.2 Temporal Analysis of User Behavior

Understanding temporal variations in user behavior is crucial for optimizing public transportation services. Time-series data mining has been employed to analyze daily and weekly travel patterns, offering detailed insights into how commuter behavior changes over different timescales. For instance, [9] used time-series data mining to uncover significant travel patterns in Singapore's public train system, demonstrating the effectiveness of integrating traditional time-series analysis with data mining techniques. Random forest was shown to be performing well on travel data with around 100,000 unique users [10]. RandomForest classified between commuter and non-commuter with precision and recall rates of up to around 90-95%.

2.3 Impact of External Factors on User Behavior

External factors such as fare changes and weather conditions significantly influence public transportation usage. Research has shown that fare elasticity varies with socioeconomic characteristics and travel modes, affecting how different user groups respond to fare adjustments. For instance, [11] analyzed smart card data from Stockholm's public transport system to determine fare elasticities, revealing that lower socioeconomic groups are less sensitive to fare changes compared to higher-income users. Additionally, [12] explored the impact of weather on public transport usage in Berlin, finding that factors like temperature and precipitation can substantially alter travel behavior.

2.4 Summary

The reviewed works provide a comprehensive understanding of user behavior in public transportation and the impact of various external factors. My research builds on previous research and utilize their findings to identify user categories based on their travel patterns. Users are defined by an aggregate of their temporal travel patterns utilizing the findings of [4] which showed statistically significant groups based on solely temporal data. The utilization of RandomForest proved to be efficient and effective at detecting pre-defined user categories in [10] with a similar amount of data showing the same two peak patterns as Stockholm's public transportation.

As we transition to the methodology section, the research approaches and tools utilized in this study will build on these foundational concepts. Our aim is to develop a nuanced understanding of public transportation as living ecosystems, responsive to both external changes and internal dynamics. Through the integration of smart card data, and user behavior modeling, this research aims to contribute to the development of more resilient, efficient, and user-centered public transportation systems.

3 The Dataset

3.1 Trafikförvaltningen and Stockholm

Trafikförvaltningen, also known as Stockholm Public Transport (SL), is responsible for managing public transportation in the Stockholm region of Sweden. Their services includes buses, trains, trams, and ferries that facilitate the daily commutes of millions of residents and visitors.

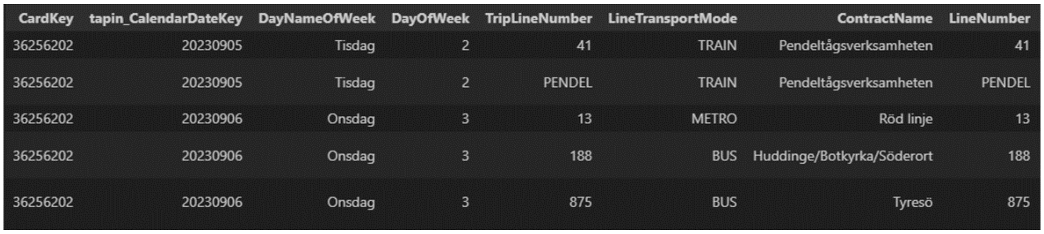
As of April 2024, Trafikförvaltningen's system does not track tap-out locations directly. Instead, these locations are inferred based on the next tap-in location. The algorithm used for this inference has demonstrated an accuracy of approximately 80% when applied to systems that do track tap-out locations. Individual trips are linked into single journeys if they occur close enough in both time and space. This assumption considers an extra tap-in within a short period and distance to be a transfer rather than a separate trip, reflecting the user's continuous travel to a final destination. For those interested in more specifics of the region and SLL [11] is recommended for further reading.

3.2 The Dataset

The dataset used in this project is provided by Trafikförvaltningen and encompasses data from the period of October 2022, up to February 2024. This dataset captures an average of 2 million trips daily, recorded with the following details:

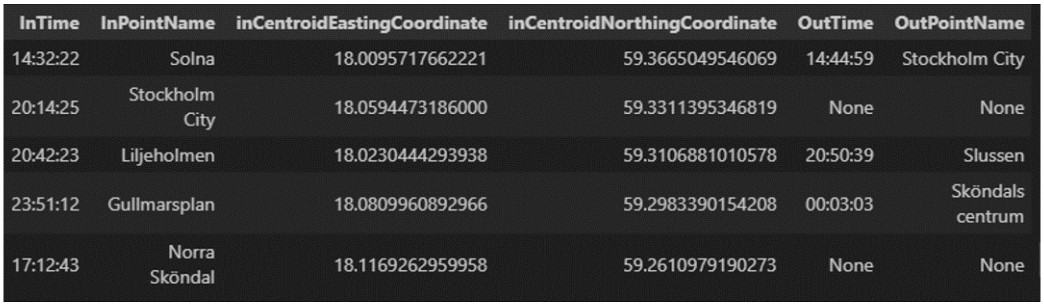
- Tap-in/Tap-out Location: The entry and exit points of each trip.
- Time of Day: The timestamp for each tap-in and tap-out.
- Ticket Type: The type of ticket used for the trip.
- Card ID: A unique identifier for each ticket, simplified in this report to represent a single traveler.

A single trip of a set of users could look something like:



CardKey	tapin_CalendarDateKey	DayNameOfWeek	DayOfWeek	TripLineNumber	LineTransportMode	ContractName	LineNumber
36256202	20230905	Tisdag	2	41	TRAIN	Pendeltågsverksamheten	41
36256202	20230905	Tisdag	2	PENDEL	TRAIN	Pendeltågsverksamheten	PENDEL
36256202	20230906	Onsdag	3	13	METRO	Röd linje	13
36256202	20230906	Onsdag	3	188	BUS	Huddinge/Botkyrka/Söderort	188
36256202	20230906	Onsdag	3	875	BUS	Tyresö	875

Figure 1: Temporal Data of User Trips



InTime	InPointName	inCentroidEastingCoordinate	inCentroidNorthingCoordinate	OutTime	OutPointName
14:32:22	Solna	18.0095717662221	59.3665049546069	14:44:59	Stockholm City
20:14:25	Stockholm City	18.0594473186000	59.3311395346819	None	None
20:42:23	Liljeholmen	18.0230444293938	59.3106881010578	20:50:39	Slussen
23:51:12	Gullmarsplan	18.0809960892966	59.2983390154208	00:03:03	Sköndals centrum
17:12:43	Norra Sköndal	18.1169262959958	59.2610979190273	None	None

Figure 2: Spatial Data of User Trips

Stockholm has on average 1-2 million daily trips recorded, with fewer trips on weekends compared to weekdays, and fewer in the vacations months such as July compared to May or November. The dataset shows two distinct peaks on most days, typically corresponding to the morning and evening rush hours, reflecting the behavior of users commonly referred to as commuters (as done in [4]. Each week, the dataset records roughly 2 million unique users, with approximately 800.000 unique people traveling every day responsible or roughly 30% of all journeys made. This indicates that some users make multiple trips within a month, but that a handful of travelers utilize the system heavily. The dataset is organized into multiple rows, each representing an individual trip linked to a unique card key/ID.

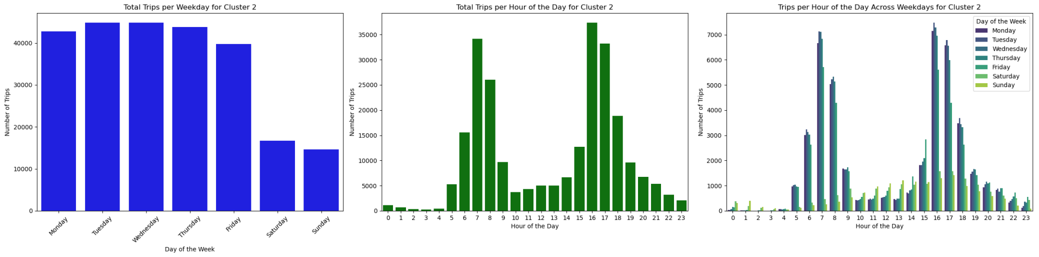


Figure 3: Commuters Trip Patterns

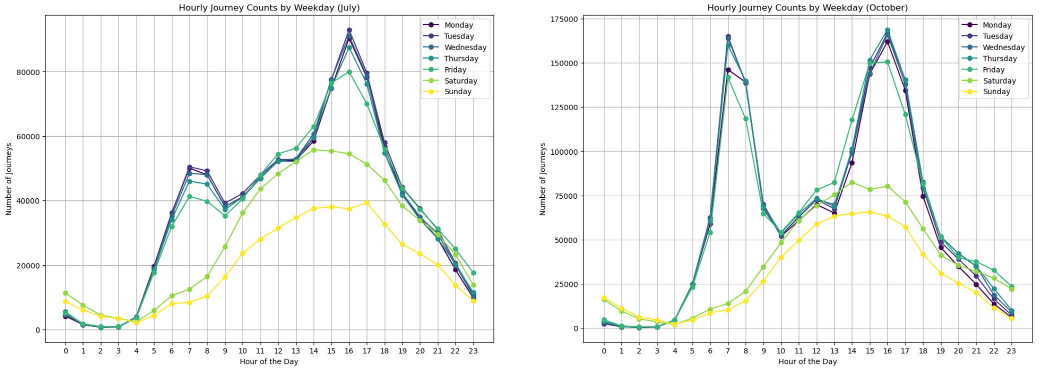


Figure 4: Hourly Travel Patterns - July 2024

4 Method

This project will utilize the fact that train stations could be grouped into a network connected by trips made by users. Utilizing spectral clustering a set of groups will be identified and analyzed. The temporal patterns will be analyzed by looking at the hourly incoming and outgoing trips made for stations. The data comes Trafikförvaltningen which is charge of public transportation in Stockholm, Sweden. The final evaluation will be made by linking to station groupings to previous knowledge with domain experts to evaluate if the groups found exhibit similar characteristics. The full code for this project can be found here: `https://github.com/Tobias-Johannesson/ID2211`.

4.1 Experimental Setup

The experiments will be run in an Azure notebook running on a dllsmall cluster with 3 cores. The data will be available in a data lake also stored in the same Azure account, and will be accessed directly from the notebook. The following versions of software were used in this study:

- **Python**: 3.10.6 (packaged by conda-forge)
- **PySpark**: 3.3.1
- **Pandas**: 1.5.1
- **NumPy**: 1.23.4
- **Matplotlib**: 3.6.2
- **Scikit-learn**: 1.1.3

4.2 Data Preparation

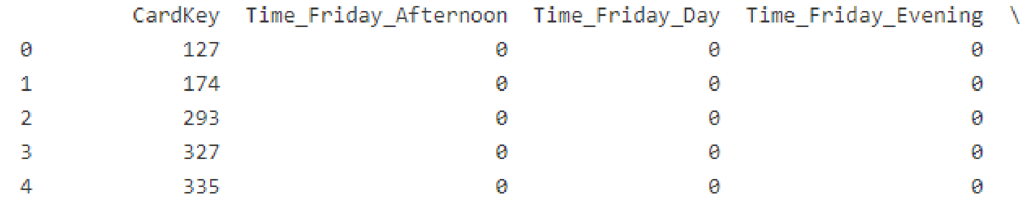
4.2.1 Data Selection

The dataset is filtered to include only: Adult-, Tourist-, and single- tickets. The idea is to capture users identified as residents with frequent travel patterns, and users identified as non-residents. This is done by looking at ticket types after removing reduced fares such as that of school children, students and elderly to reduce variability. Two distinct time periods are explored, a May week (2023-05-08 to 2023-05-14), and a July week (2023-07-10 to 2023-07-16).

4.2.2 User Representation

Each user's travel data is aggregated into weekday profiles including day of week, and time of day. This involves aggregating the number of trips each user makes within specific hourly intervals across the days and the weeks of this period.

The week-hour-trip profile for each user is a single row per user (CardKey) where the columns represent a specific day and time slice such as Monday_Morning, or Tuesday_Evening. The time of day is simplified from seconds after midnight into Morning, Day, Afternoon, Evening, or Night. The hour mapping turns: 0-4 into night, 5-9 into morning, 10-15 into day, 16-18 into afternoon, and 19-23 into evening. The most commonly used ticket type for that user during that period is selected in case of multiple tickets being used.



```
   CardKey  Time_Friday_Afternoon  Time_Friday_Day  Time_Friday_Evening  \
0      127                      0                0                    0
1      174                      0                0                    0
2      293                      0                0                    0
3      327                      0                0                    0
4      335                      0                0                    0
```

Figure 5: Weekday Profile

4.3 Model Training

The dataset is split into training and testing sets to evaluate the performance of the models. Typically, 70% of the data is used for training, and 30% is reserved for testing. A Random Forest model with the default value of 100 trees was set up using sklearn. This is an ensemble learning method that uses multiple decision trees to improve prediction accuracy.

4.4 Model Evaluation

The trained model is evaluated on the testing set with a detailed error analysis to understand where the models perform well and where they struggle. This includes:

- **Accuracy:** The proportion of correctly classified instances out of the total instances.
- **Precision:** The ratio of correctly predicted positive observations to the total predicted positives.
- **Recall:** The ratio of correctly predicted positive observations to all the observations in the actual class.

- **F1 Score:** The weighted average of precision and recall, providing a balance between the two.
- **Confusion Matrix:** A matrix that shows the number of correct and incorrect predictions for each class.
- **Misclassification Analysis:** Examination of the instances where the model predictions were incorrect to identify common patterns or characteristics that lead to errors.

5 Results and Analysis

In this section, we present the evaluation of our RandomForest model's performance in classifying users into three categories: Adult tickets, Single tickets, and Tourist tickets. We begin by discussing the model evaluation metrics, including accuracy, confusion matrix, and classification report. Following this, we provide detailed visualizations to illustrate the classification patterns for each category over two different periods, May and July which covers both a "normal" and a "vacation" week. These visualizations help to identify correctly and incorrectly classified user patterns across various dimensions such as day of the week, hour of the day, mode of transport, time traveled, and distance traveled. We then delve into a detailed analysis of the results, offering insights into the strengths and weaknesses of the model. Finally, we discuss the limitations of our approach and suggest potential areas for future improvement.

Model Performance

The RandomForest model was trained and evaluated on the dataset to classify users into one out of three categories: Adult, Single, and Tourist. The model's performance metrics, including accuracy, confusion matrix, and classification report, are presented below. Where the overall accuracy of the RandomForest model is **83%** for May and **77%** July.

Confusion Matrix

The confusion matrix provides a detailed breakdown of the model's performance across each class. For the month of May we have a total of 112,845, and for July we have a total of 78,961 unique users that have been assigned to a single category based on their most commonly used ticket type during that period. The confusion matrices has been rounded to the nearest 100. These show the number of true positives, false positives, and false negatives for each class:

Table 1: Confusion Matrices for May and July

	May		
	Predicted Adult	**Predicted Single**	**Predicted Tourist**
Actual Adult	42,100	7,300	500
Actual Single	5,800	47,600	1,019
Actual Tourist	1,400	3,200	3,900
	July		
	Predicted Adult	**Predicted Single**	**Predicted Tourist**
Actual Adult	23,000	6,400	1,200
Actual Single	3,900	32,000	1,300
Actual Tourist	2,200	3,400	5,800

Classification Report

The classification report provides precision, recall, and F1-score for each class, offering a comprehensive view of the model's performance. For both May and July the accuracy is around 80% showing a higher F1-score for single tickets in July, and F1-scores around 50-60% for tourist tickets in both May and July. Notice that there are more instances in total during May, but more tourists in July.

Table 2: Classification Report for May

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Adult	0.85	0.84	0.85	50,000
Single	0.82	0.88	0.85	54,400
Tourist	0.72	0.46	0.57	8,500
Accuracy	0.83 (112,800 instances)			
Macro avg	0.80	0.73	0.75	112,800
Weighted avg	0.83	0.83	0.83	112,800

Table 3: Classification Report for July

Class	Precision	Recall	F1-Score	Support
Adult	0.79	0.75	0.77	30,500
Single	0.77	0.86	0.81	37,100
Tourist	0.70	0.51	0.59	11,300
Accuracy	0.77 (79,000 instances)			
Macro avg	0.75	0.72	0.72	79,000
Weighted avg	0.77	0.77	0.76	79,000

5.1 Compared to Random Classification

The expected performance of a random classifier based on the class distribution for the period in July without any data for making these decisions would look something like the following:

$$\text{Expected Accuracy} = p_{\text{Adult}}^2 + p_{\text{Single}}^2 + p_{\text{Tourist}}^2$$

$$p_{\text{Adult}} = \frac{30,475}{78,961}, \quad p_{\text{Single}} = \frac{37,144}{78,961}, \quad p_{\text{Tourist}} = \frac{11,342}{78,961}$$

$$\text{Expected Accuracy} = \left(\frac{30,475}{78,961}\right)^2 + \left(\frac{37,144}{78,961}\right)^2 + \left(\frac{11,342}{78,961}\right)^2$$

$$\text{Expected Accuracy} = (0.386)^2 + (0.471)^2 + (0.144)^2 = 0.149 + 0.222 + 0.021 \approx 0.392$$

To compare the performance of the RandomForest model to the expected performance of a random classifier shows an improvement of around 40 percentage points across all three groups for precision, recall, and F1-Score.

Visualizing Classification Patterns

To analyze the classification patterns of the RandomForest model, we present visualizations for each user category: Adult, Single, and Tourist. Separate figures show the temporal travel patterns for May and July illustrating in each the patterns of correctly vs. incorrectly users for that class. These will be presented in the following order: Adult tickets May, Adult tickets July, Tourist tickets May, and then Tourist tickets July. Single tickets for May and July can be found in the appendix for figures 10 and 11.

The figures show the incorrectly classified users in the top row, and the correctly classified users in the bottom. Each row shows the percentage of users correctly classified, counts for the day of the week traveled, counts of the hour of each day traveled, counts for each mode of transportation, count for time and distance travelled measured in minutes and kilometers.

Adult Category

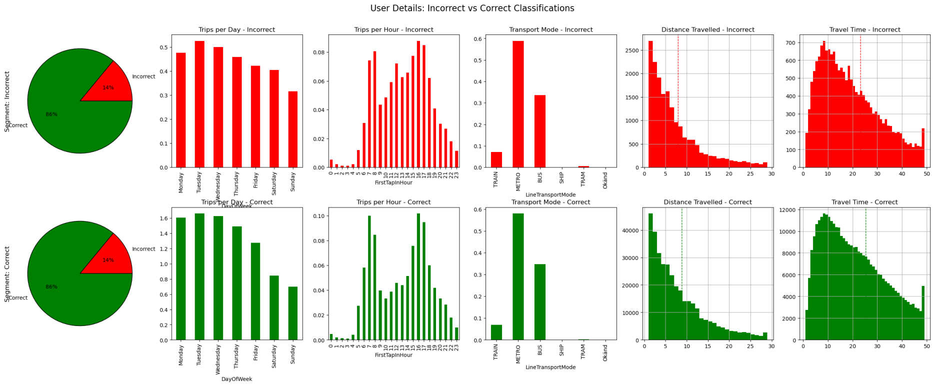


Figure 6: Classifications Over May for Adult Tickets

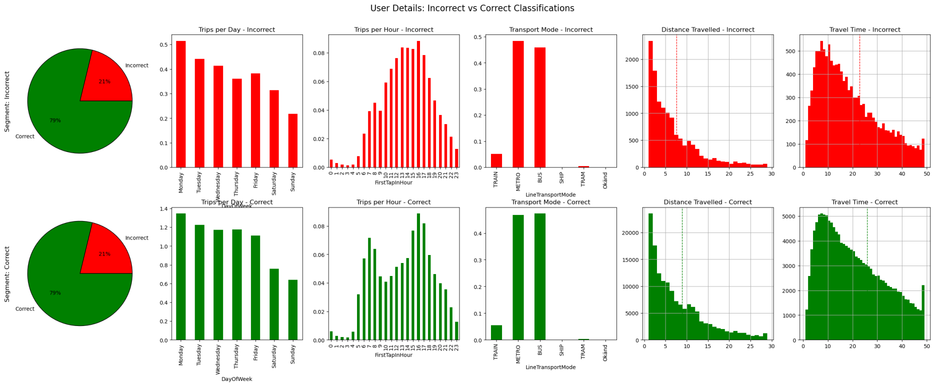


Figure 7: Classifications Over July for Adult Tickets

Tourist Category

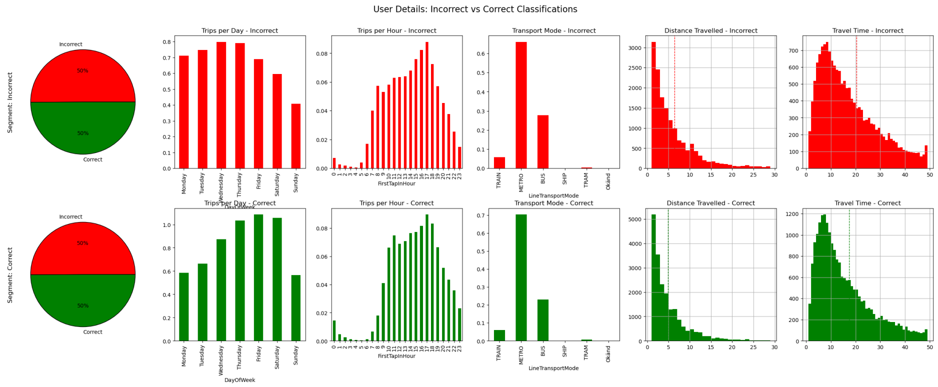


Figure 8: Classifications Over May for Tourist Tickets

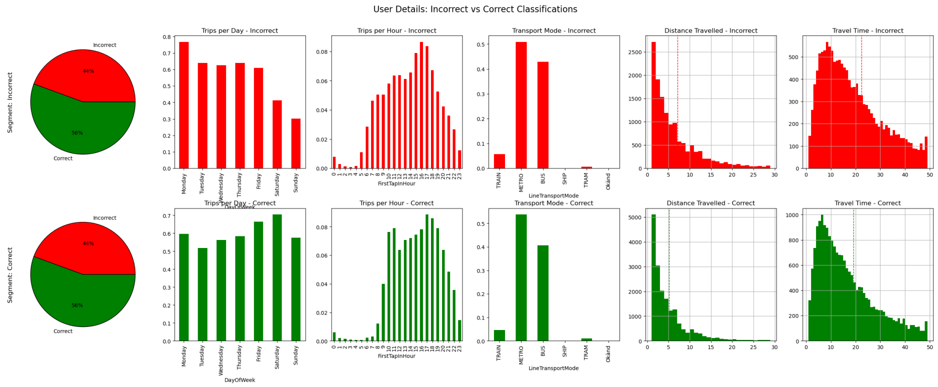


Figure 9: Classifications Over July for Tourist Tickets

5.2 Analysis of Category Characteristics and Classifications

Our model demonstrates a potential to predict user types based on their temporal travel patterns with a high degree of accuracy. The RandomForest model achieved an overall accuracy of roughly 80%, indicating robust performance in classifying users into the three categories: Adult, Single, and Tourist. However, a deeper analysis of the recall values reveals more nuanced insights into the model's performance for each user type.

5.2.1 Class Specific Performance

The recall for single ticket users is high where it even increase in July with fewer samples, suggesting that the model effectively identifies most users in this category. In contrast, the recall for tourists is around 50% (compared to the 80-90% for single), indicating that the model misses many true tourist instances. This discrepancy could be attributed to several factors:

- **Class Imbalance:** The Tourist category is underrepresented in the dataset, with a ratio of around 4:25 compared to other classes. This imbalance can be addressed by employing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or by adjusting the sampling strategy (under-sampling or over-sampling). SMOTE did not perform well on the experimental set up and was therefore not further explored, and the adult and single tickets were downsampled by half for adult tickets and one fourth for single tickets.

- **Overlapping User Behaviors:** The travel behaviors of different user groups may not be distinctly defined using solely temporal patterns. For instance, tourists might use single tickets frequently, or some adults might purchase short-term or single tickets, leading to overlapping behaviors and making it difficult for the model to distinguish between these groups. It might also be that tourists happen to travel during the same hours and days as other users where another feature such as the stations visited need to be analyzed. It might also be that we capture the travel pattern where users head towards the office as well as their activities on weekends where there are multiple types of activities involved. This seemingly becomes more apparent during the week in July where some adult users show no distinct peaks.

- **Incorrectly labelled groups:** The tickets are labelled into one of the three categories manually based on a range of tickets, such as the 24 hour, 3 day, and 7 days ticket types are marked as tourists. This might capture users that are more likely to belong to the other groups such as single tickets might also belong with the 24 hour tickets.

- **Too few samples:** The performance might have increased with more samples to train on where outliers have less of an impact.

- **Too narrow user representation:** The weekday profile is simplified to day, morning, etc from the exact second a ticket is used and this might mean that users on the edge between morning and day are put in the same grouping.

- **Users hold the wrong ticket or multiple:** It is not certain that someone utilizing the single tickets ”should” use the single ticket. It might make more sense from a cost perspective to buy a period card, but there is nothing stopping users from purchasing any other type even if it would cost more in the long run. Since the aggregation of all user data only represent a single ticket type, users which wield multiple types are put into the one most frequently wielded by them which could capture their behavior from both types into one causing some noise.

All these considerations in mind it shows that it is possible to group users based on temporal patterns. It shows how more data provides better representations, and that it might be difficult capturing all users based on only this representation.

5.2.2 Temporal Patterns and User Behavior

Analyzing the variations between July and May for adult tickets reveals interesting patterns. In July especially, the model detects changes in travel patterns among adults, suggesting that more users might be on vacation or at least that there is a shift during this period. This is inferred from the disappearance of distinct rush hour patterns among incorrectly classified users where correctly classified adults still exhibit this temporal pattern. Additionally, both correctly and incorrectly classified adult users show a slightly larger volume of medium distance travel compared to the other groups, but the number of these trips decrease in July compared to May which furthers the theory that these long commuters are on vacation. This shift in travel behavior could be due to seasonal changes or vacation periods which aligns with common trends of Stockholm, Sweden.

Incorrectly classified single ticket users for May show a similar pattern to those of adult tickets with two distinct peaks. This could indicate that these tickets are utilized for the same types of travels, and that a longer period needs to be considered to fully capture commuter patterns. Mostly these users show a sporadic behavior over the day with more trips during the day and afternoon.

For tourist tickets an even spread over the day with a shift towards being later in the day showing rush hour peaks neither in May nor July. The incorrectly classified tourists are those with more trips earlier in the day which could indicate how more detailed features for the timing component could avoid confusing these users where the cutoff between morning and day is hour 9 and hour 10 of the day (starting at 0).

Examining the ticket types held by users categorized as a tourist in July compared to the two months prior and the two following months, the following findings reveal how these users hold other tickets. This could indicate how these users specifically change their way of behaving from already living in Stockholm, compared to those coming in from outside the region.

This distribution is based on 14,800 unique tourist users identified in May. Out of these 1,500 users held a 30-day card, which ranked eighth among the most used ticket types by tourists from tourists identified in May. Single tickets remained in the third position with approximately 3,500 users. The number of tourists increased significantly in July, with a total of 28,200 users marked as tourists, compared to 14,800 in May. From the tourists in July the 30-day ticket is held by more of these in other months indicating that some period users change their ticket for this period in July. The following shows the top 5 most used tickets by those using a tourist ticket during the week in July in the full period of 2023-05-10 up to 2023-09-16:

- 7-day tickets: 9,900
- 3-day tickets: 5,500
- Single tickets: 5,200
- 24-hour tickets: 4,400
- 30-day tickets: 3,900

5.2.3 Seasonal Variations

July has fewer trips made in total, and fewer unique users which can be explained by this being a common period for vacation in Stockholm, Sweden. The increase in bus usage in July compared to trains might indicate a shift in user activities, such as more recreational trips. This switch could also be attributed to disruptions in metro services, where buses were used as substitutes during renovations. Parts of the metro system was closed during this period where replacements had to be taken [13] [14]. Interestingly, tourists seemed least affected by this switch from trains to buses, possibly due to their flexible travel plans, greater use of diverse transport modes, or simply that regions most tourists visit were not as affected.

To fully capture seasonal variation it might be important to classify users based on their behavior over both of these periods where a user can change their group belonging over time. This behavior might also be different depending on if a user changes from a adult ticket compared to just visiting Stockholm as a tourist for the first time.

5.3 Discussion

This exploration is limited by several factors, including the type of data used, the methodology for grouping users, and the scope of the data:

5.3.1 Limitations

- **Temporal Data Only:** The analysis is restricted to temporal data, which captures the timing of trips but lacks spatial information. Temporal data provides initial insights into user behavior, but it may not fully capture the complexity of travel patterns. For example, spatial data could reveal important details about trip origins and destinations, which could improve the accuracy of user classification.

- **Manual Grouping Based on Ticket Types:** The groups used for classification were manually defined based on ticket types, with input from domain experts at Trafikförvaltningen. While this expert knowledge is valuable, manually set groups might overlap in user behavior, leading to ambiguity in classification.

- **Limited Data Scope:** The study only explores data from two specific weeks, one in May and one in July. This limited temporal scope may not capture the full range of seasonal or monthly variations in travel behavior. Furthermore, the dataset only covers the Stockholm region, which may not be representative of travel patterns in other cities or regions.

- **Exclusion of Certain Ticket Types:** Tickets for students and the elderly were excluded to focus on tourists vs. non-tourists. However, these groups might exhibit travel behaviors similar to those of tourists or regular users, and their exclusion may omit relevant patterns from the analysis.

5.3.2 Mitigations

To address these limitations, several mitigations were implemented, and further steps are recommended:

- **Incorporation of Spatial Data:** Future studies should incorporate spatial data to complement the temporal data. This could involve analyzing trip origins and destinations, route preferences, and geographic clustering of travel patterns.
- **Automated Grouping with Clustering Algorithms:** Instead of relying on manual groupings, applying clustering algorithms can help identify distinct and meaningful groups based on travel behavior. Techniques such as k-means clustering or hierarchical clustering could uncover natural patterns in the data that are not apparent through manual classification.
- **Extended Data Periods:** Expanding the dataset to cover longer periods and multiple months can provide a more comprehensive understanding of travel behavior. This would help in capturing seasonal variations, holiday impacts, and long-term trends.
- **Broader Geographic Scope:** Repeating the analysis in different regions or cities can validate the findings and reveal how travel behaviors vary across different transportation systems. This would enhance the generalizability of the results.
- **Detailed Ticket Type Analysis:** A deeper investigation into different ticket types, such as 24-hour tickets vs. 7-day or single tickets, can provide finer-grained insights into user preferences and behaviors. This could involve analyzing the frequency of use, trip lengths, and the times of day these tickets are most commonly used.
- **Inclusion of Additional User Groups:** Including other user groups, such as students and the elderly, in future analyses can provide a more complete picture of travel behavior. This can help in understanding

how different demographics use the public transportation system and identifying unique patterns within these groups.

5.3.3 Future Work

To build on this work and address the identified limitations, future research could take several directions. Integrating spatial data with temporal analysis to provide a more holistic view of travel behavior. This can help in identifying spatial patterns and their correlation with temporal travel habits. Use a longer period and more fine-grained data to capture more detailed and comprehensive travel patterns. This can help in understanding seasonal trends and anomalies over extended periods. Conduct similar studies in other regions or cities to compare travel behaviors across different public transportation systems. This can help in understanding the generalizability of the findings and identifying region-specific patterns. Continuously refine the user grouping, representation, and modeling techniques. This includes exploring advanced machine learning algorithms and feature engineering to improve classification accuracy. Focus on how different user groups exhibit distinct temporal patterns. This involves analyzing travel behavior at different times of the day, days of the week, and during special events or holidays to identify unique temporal signatures for each group.

By addressing these limitations and following the recommended directions for future research, the analysis of public transportation data can be significantly enhanced, leading to better-informed decisions for transportation planning and management.

6 Conclusion

This study aimed to explore how temporal patterns in public transportation usage can be leveraged to classify users into predefined categories: Adult, Single, and Tourist. By utilizing a RandomForest model, we achieved a classification accuracy of approximately 80%.

6.1 Summary of Findings

The RandomForest model demonstrated robust performance with an overall accuracy of 80%. The model's precision, recall, and F1-scores indicate its effectiveness in classifying Single ticket users, while performance for Tourist tickets was comparatively lower, suggesting room for improvements.

These classes exhibit preferences such as correctly classified Adult users generally exhibited clear rush hour travel patterns, which diminished during the July vacation period, reflecting seasonal changes in travel behavior. The model effectively identified Single ticket users, with high recall values suggesting that these users' travel patterns are distinct and consistent. Tourist classification was less accurate, with a recall around 50%. This indicates that the model struggled to distinguish tourists, likely due to the overlap in travel behaviors with other user categories and the under-representation of tourists in the dataset.

The data highlighted significant changes in travel patterns between May and July. Where in July, increased recreational travel and the use of replacement buses due to metro renovations were evident. These findings suggest that user behavior is highly dynamic and influenced by external factors such as holidays and infrastructure changes.

6.2 Implications

The findings of this study have practical implications for public transportation planning and management. Defining distinct travel patterns allows for more targeted service improvements, such as adjusting schedules or routes to better accommodate different user groups. Insights into seasonal variations and user behavior can inform policies aimed at enhancing user satisfaction and system efficiency, particularly during peak and off-peak periods.

This study faced several limitations mainly due to choices and assumptions made for this project. The analysis was limited to two specific weeks, one in May and one in July. This may not capture the full range of seasonal variations and long-term trends. The reliance on temporal data without incorporating spatial information may have limited the model's ability to fully understand travel patterns which could be further investigated in future research to improve the metrics from this research. The manually defined groups and exclusion of reduced tickets could introduce bias and overlap, affecting classification accuracy. These topics I leave for future research.

References

[1] E. Deschaintres, C. Morency, and M. Trépanier, “Analyzing Transit User Behavior with 51 Weeks of Smart Card Data,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 6, pp. 33–45, Jun. 2019. doi: 10.1177/0361198119834917. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0361198119834917

[2] H. Jiao, S. Huang, and Y. Zhou, “Understanding the land use function of station areas based on spatiotemporal similarity in rail transit ridership: A case study in Shanghai, China,” *Journal of Transport Geography*, vol. 109, p. 103568, May 2023. doi: 10.1016/j.jtrangeo.2023.103568. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0966692323000406

[3] L. M. Kieu, A. Bhaskar, and E. Chung, “Passenger segmentation using smart card data,” vol. 16, no. 3, pp. 1537–1548. doi: 10.1109/TITS.2014.2368998. [Online]. Available: http://ieeexplore.ieee.org/document/6981952/

[4] O. Cats and F. Ferranti, “Unravelling individual mobility temporal patterns using longitudinal smart card data,” vol. 43, p. 100816. doi: 10.1016/j.rtbm.2022.100816. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2210539522000372

[5] C. C. Aggarwal and H. Wang, “A Survey of Clustering Algorithms for Graph Data,” in *Managing and Mining Graph Data*, C. C. Aggarwal and H. Wang, Eds. Boston, MA: Springer US, 2010, vol. 40, pp. 275–301. ISBN 978-1-4419-6044-3 978-1-4419-6045-0 Series Title: Advances in Database Systems. [Online]. Available: https://link.springer.com/10.1007/978-1-4419-6045-0_9

[6] K. A. Seaton and L. M. Hackett, “Stations, trains and small-world networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 339, no. 3-4, pp. 635–644, Aug. 2004. doi: 10.1016/j.physa.2004.03.019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0378437104003036

[7] Q. Liang, J. Weng, W. Zhou, S. B. Santamaria, J. Ma, and J. Rong, “Individual Travel Behavior Modeling of Public Transport Passenger Based on Graph Construction,” *Journal of Advanced Transportation*,

vol. 2018, pp. 1–13, 2018. doi: 10.1155/2018/3859830. [Online]. Available: https://www.hindawi.com/journals/jat/2018/3859830/

[8] J. Idrais, Y. E. Moudene, and A. Sabour, "Characterizing user behavior in online social networks: Study of seasonal changes in the moroccan community on facebook," in *2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS)*, 2019. doi: 10.1109/ICDS47004.2019.8942365 pp. 1–5.

[9] R. K.-W. Lee and T. S. Kam, "Time-series data mining in transportation: A case study on singapore public train commuter travel patterns," *International journal of engineering and technology*, vol. 6, pp. 431–438, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:51508700

[10] Z. Mei, W. Ding, C. Feng, and L. Shen, "Identifying commuters based on random forest of smartcard data," vol. 14, no. 4, pp. 207–212. doi: 10.1049/iet-its.2019.0414. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1049/iet-its.2019.0414

[11] Y. Kholodov, E. Jenelius, O. Cats, N. Van Oort, N. Mouter, M. Cebecauer, and A. Vermeulen, "Public transport fare elasticities from smartcard data: Evidence from a natural experiment," vol. 105, pp. 35–43. doi: 10.1016/j.tranpol.2021.03.001. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0967070X2100055X

[12] K. M. Nissen, N. Becker, O. Dähne, M. Rabe, J. Scheffler, M. Solle, and U. Ulbrich, "How does weather affect the use of public transport in berlin?" vol. 15, no. 8, p. 085001. doi: 10.1088/1748-9326/ab8ec3. [Online]. Available: https://iopscience.iop.org/article/10.1088/1748-9326/ab8ec3

[13] "Så förändras SL-trafiken 2023 — mitti.se," https://www.mitti.se/nyheter/sa-forandras-sltrafiken-2023-6.27.40054.873bc8ebd7, [Accessed 19-05-2024].

[14] "Därför stängs tunnelbanan av i sommar — mitti.se," https://www.mitti.se/nyheter/darfor-stangs-tunnelbanan-av-i-sommar-6.3.80738.4c468e1287, [Accessed 19-05-2024].

Single Category

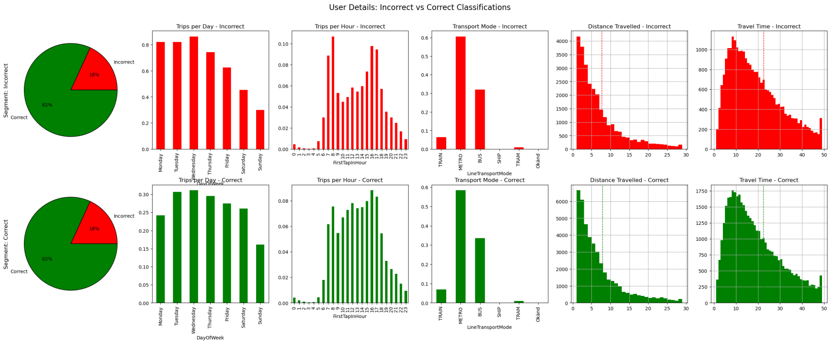


Figure 10: Classifications Over May for Single Tickets

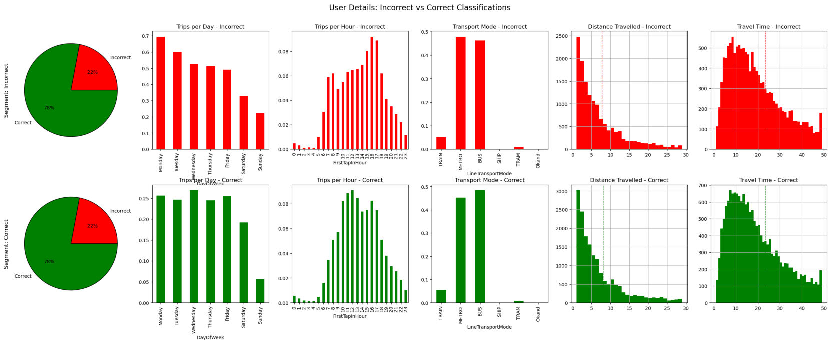


Figure 11: Classifications Over July for Single Tickets