HSL: Satisfaction Analysis Technical Report

by Anita Braida, Helene Ilvonen and Chao Wang

18 Oct 2023

Summary

HSL: Satisfaction Analysis is a project that tries to discover the dominant feature, i.e., the key factor among Helsinki Regional Transport Authority (HSL) transportation services. In the project, we analyzed public data from the HSL satisfaction survey since 2011 and used the data to train multiple machine learning models. The result shows that, among all the features that are influencing the HSL service, the top five features are the crowdedness of the vehicle, the effectiveness of changing vehicles, fast and smooth traveling, information availability and punctuality.

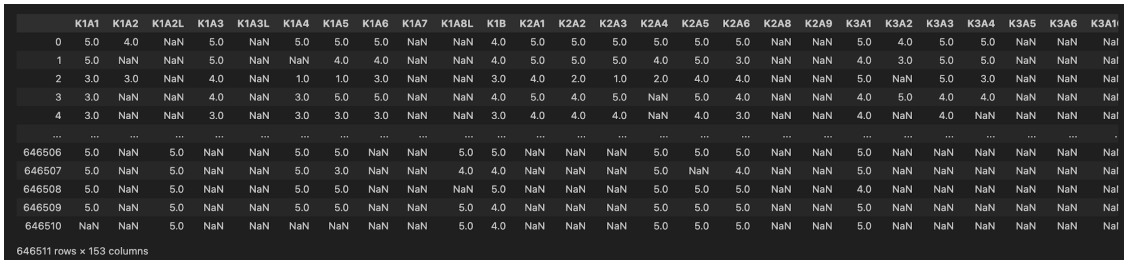
Introduction

This project aims to provide data-driven recommendations for HSL to enhance their service quality and efficiency by focusing on the top features. Our findings will help HSL increase the number of boardings, which aligns with their 2022-2025 development strategy. And by increasing the use of public transportation, we hope to accelerate our process of transforming towards a carbon-neutral society. Studies have shown that an efficient and widely used public transportation network is associated with lower pollutants emission, improving urban air quality and reducing the carbon footprint (Gonzalez et al., 2021; Jimenez & Roman, 2021).

Data & Methodology

In the project, we used data from multiple sources. The primary data source is the results of HSL customer satisfaction surveys, which are publicly available through HSL official API. The dataset consists of custom ratings from various aspects on a scale of 1-5 of HSL service (e.g, punctuality, cleanliness..) from 2011 to present. Additionally, we realized weather can be an important feature impacting the public transportation service rating, and the weather labels in the HSL survey table are not detailed enough (1 for "Rainy" and 2 for "Not rainy"). Therefore, we decided to integrate a historical weather data published by Finnish Meteorological Institute, which offers more detailed weather data for our model to better estimate the relationships.

We obtained the raw data (646511 rows x 153 features by Sept. 2023) from the HSL's public transport customer satisfaction survey online database Asty Web which provides a public free Data API to get the data (HSL, 2023).



	K1A1	K1A2	K1A2L	K1A3	K1A3L	K1A4	K1A5	K1A6	K1A7	K1A8L	K1B	K2A1	K2A2	K2A3	K2A4	K2A5	K2A6	K2A8	K2A9	K3A1	K3A2	K3A3	K3A4	K3A5	K3A6	K3A1
0	5.0	4.0	NaN	5.0	NaN	5.0	5.0	5.0	NaN	NaN	4.0	5.0	5.0	5.0	5.0	5.0	5.0	NaN	NaN	5.0	4.0	5.0	5.0	NaN	NaN	Nal
1	5.0	NaN	NaN	5.0	NaN	NaN	4.0	4.0	NaN	NaN	4.0	5.0	5.0	5.0	4.0	5.0	3.0	NaN	NaN	4.0	3.0	5.0	5.0	NaN	NaN	Nal
2	3.0	3.0	NaN	4.0	NaN	1.0	1.0	3.0	NaN	NaN	3.0	4.0	2.0	1.0	2.0	4.0	4.0	NaN	NaN	5.0	NaN	5.0	3.0	NaN	NaN	Nal
3	3.0	NaN	NaN	4.0	NaN	3.0	5.0	5.0	NaN	NaN	4.0	5.0	4.0	5.0	NaN	5.0	4.0	NaN	NaN	4.0	5.0	4.0	4.0	NaN	NaN	Nal
4	3.0	NaN	NaN	3.0	NaN	3.0	3.0	3.0	NaN	NaN	3.0	4.0	4.0	4.0	NaN	4.0	3.0	NaN	NaN	4.0	NaN	4.0	NaN	NaN	NaN	Nal
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
646506	5.0	NaN	5.0	NaN	NaN	5.0	5.0	NaN	NaN	5.0	5.0	NaN	NaN	NaN	5.0	5.0	5.0	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	Nal
646507	5.0	NaN	5.0	NaN	NaN	5.0	3.0	NaN	NaN	4.0	4.0	NaN	NaN	NaN	5.0	NaN	4.0	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	Nal
646508	5.0	NaN	5.0	NaN	NaN	5.0	5.0	NaN	NaN	NaN	5.0	NaN	NaN	NaN	5.0	5.0	5.0	NaN	NaN	4.0	NaN	NaN	NaN	NaN	NaN	Nal
646509	5.0	NaN	5.0	NaN	NaN	5.0	5.0	NaN	NaN	5.0	4.0	NaN	NaN	NaN	5.0	5.0	5.0	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	Nal
646510	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	NaN	5.0	4.0	NaN	NaN	NaN	5.0	5.0	5.0	NaN	NaN	5.0	NaN	NaN	NaN	NaN	NaN	Nal

646511 rows × 153 columns

Both the original dataset and documentation is in Finnish with no English option. So we used Google translation to convert the page into English. In the documentation, it shows that the feature "K3B" is "The general rating for public transport in the HSL area". Therefore, we identified "K3B" as our target label of our training model and rest of the features can be our independent variables.

We processed the data primarily in Python using Pandas. In the dataset, each row is one survey result with each column being one question on the survey with its name in code. For example, "K1A1" stands for "Drivers/Staff serve customers in a friendly manner" and "K1A2" being "The driver knows how/ The train staff can give travel-related advice when asked". The values are users' scores for each question on a scale of 1-5, for 1 being the worst and 5 the best.

After we loaded the data, we conducted feature reduction. We investigated all the 153 features then decided to reduce the feature size in three stages.

In the first stage we removed features that are unrelated to transportation. For example, we removed "K3A6" which is "I would buy a ticket online if possible", "K3A8", "The timetable book is an important source of information..." and "K3A15" "I would like to buy and use a mobile ticket instead of a travel card" etc. Then we removed some features that are too specific on certain lines or certain times such as " K3A19", "I have previously received information about the ring road (Tikkurila-Lentoasema-Vantaankoski) that opened in summer 2015". Additionally, there are some demographic features that also get removed like "Gender", "Year of Birth" and "Home address". However, after discussion, the feature "T71": "Age group" is reserved.

In the second stage we combined some one-hot features such as "T9","Profession" having nine features and "T21","Transportation Vehicle" with four features.

In the third stage, we decided to remove the features with over 50% missing values. As a result, we effectively reduced the number of features from 153 to 26.

As mentioned above, during the data processing we found that the weather data in the HSL survey is not clear enough as our hypothesis is that temperature and precipitation (both rain and snow) have a significant impact on customer satisfaction. Therefore, we supplemented the survey data with external data published by the Finnish Meteorological Institute (Institute, 2023).

	Year	m	d	Time	Time zone	Precipitation amount (mm)	Snow depth (cm)	Air temperature (degC)	Ground minimum temperature (degC)	Maximum temperature (degC)	Minimum temperature (degC)
0	2010	1	1	00:00	UTC	-1	27	-12.9	-	-10.1	-15.3
1	2010	1	2	00:00	UTC	-1	26	-17.4	-	-15.2	-19.3
2	2010	1	3	00:00	UTC	1	24	-12.6	-	-3.2	-21.8
3	2010	1	4	00:00	UTC	3.5	26	-3.8	-	-2	-8.2
4	2010	1	5	00:00	UTC	0.5	28	-9.5	-	-0.9	-14.7
...	...	...	...	...	...	...	...	...	...	...	...
8703	2023	9	14	06:00	UTC	-	-	-	10.6	-	-
8704	2023	9	15	00:00	UTC	-1	-1	13.2	-	16.9	7
8705	2023	9	15	06:00	UTC	-	-	-	4.1	-	-
8706	2023	9	16	00:00	UTC	-1	-1	17	-	18.8	13.8
8707	2023	9	16	06:00	UTC	-	-	-	10.4	-	-

8708 rows × 11 columns

For the weather dataset, we also conducted feature reduction and then merged with the HSL data. Following this, we filled missing values with the mode of the column (if they are ratings).

At last, we encoded all the values in the dataset as the last step of our data cleaning (with the data size being 623476 x 28).

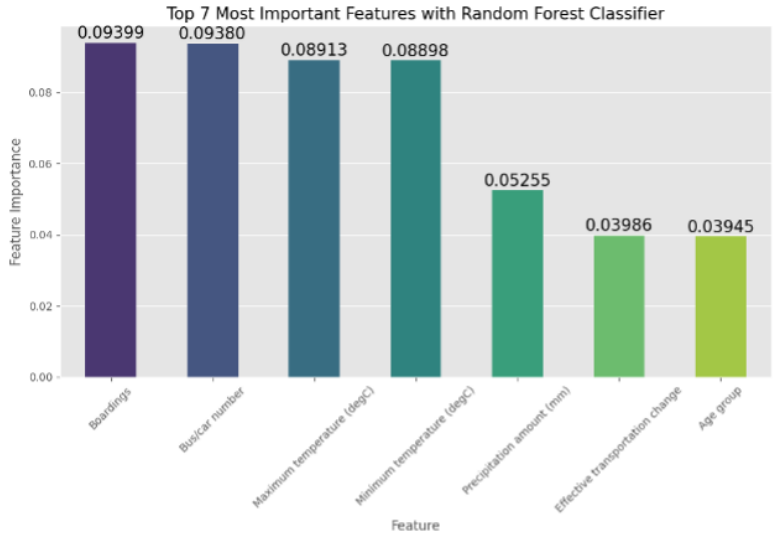
	K1A1	K1A2	K1A3	K1A4	K1A5	K1A6	K2A1	K2A2	K2A3	K2A4	K2A5	K2A6	K3A1	K3A2	T71	Transport mode	Region	Bus/car number	Bus/car filling	Weather	Season	Boardings
0	5.0	4.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	4.0	2.0	2	2010.0	58.0	1.0	2.0	1	3470.0
1	5.0	4.0	5.0	4.0	4.0	4.0	5.0	5.0	5.0	4.0	5.0	3.0	4.0	3.0	2.0	2	2010.0	58.0	1.0	2.0	1	3470.0
2	3.0	4.0	3.0	3.0	3.0	3.0	4.0	4.0	4.0	4.0	4.0	3.0	4.0	5.0	3.0	2	91.0	131.0	4.0	2.0	1	11422.0
3	5.0	4.0	4.0	4.0	3.0	3.0	5.0	3.0	4.0	2.0	5.0	5.0	4.0	5.0	2.0	2	2010.0	7.0	1.0	2.0	1	3470.0
4	4.0	4.0	4.0	2.0	4.0	4.0	2.0	3.0	4.0	4.0	5.0	4.0	5.0	5.0	1.0	2	2010.0	7.0	1.0	2.0	1	3470.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
646506	5.0	4.0	4.0	5.0	4.0	4.0	5.0	5.0	4.0	5.0	5.0	5.0	5.0	5.0	2.0	11	91.0	1.0	1.0	2.0	2	8311.0
646507	5.0	4.0	4.0	5.0	5.0	4.0	5.0	5.0	4.0	5.0	5.0	5.0	5.0	5.0	2.0	11	91.0	1.0	1.0	2.0	2	8311.0
646508	4.0	4.0	4.0	5.0	5.0	4.0	5.0	5.0	4.0	5.0	4.0	3.0	5.0	5.0	2.0	11	91.0	1.0	1.0	2.0	2	8311.0
646509	4.0	4.0	4.0	5.0	5.0	4.0	5.0	5.0	4.0	5.0	5.0	4.0	3.0	5.0	5.0	11	91.0	1.0	2.0	2.0	2	8311.0
646510	4.0	4.0	4.0	5.0	4.0	4.0	5.0	5.0	4.0	4.0	4.0	5.0	2.0	5.0	3.0	11	91.0	1.0	1.0	2.0	2	8311.0

623476 rows × 28 columns

Before training our model, first we adopt `train_test_split` function to split our data into a training set (80%) and test set (20%) then normalized the data.

As our target value is in category 1-5 (encoded as 0-4 in the model), we believe this is a multi-classification task. So the first model we used is multi-classification logistic regression from scikit-learn.

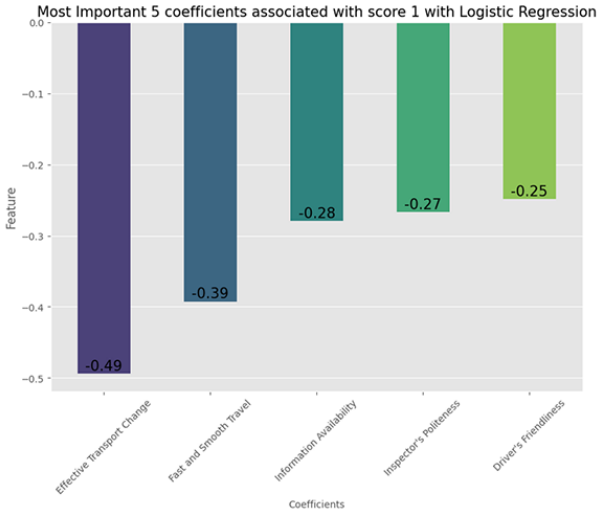
After training, the model has an accuracy of 69.178% on the test set. The logistic regression model shows us the five most important features that are related to score 5. They are "Effective Transport Change", "Fast and Smooth Travel", "Information Availability", "Punctuality" and "Inspector's Politeness".

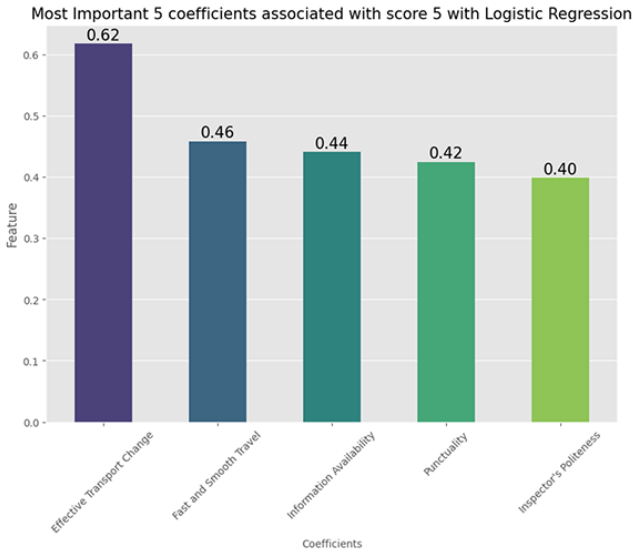


The model also shows the top five features related to the lowest score 1. They are: "Effective Transport Change", "Fast and Smooth Travel", "Information Availability", "Inspector's Politeness" and "Drivers' Friendliness".

The second model we used is random forest multi-classification from scikit-learn. After training, the model's accuracy is 69% on the test set. It gives us the top seven most important features (as top 3-5 are weather features, we decided to show two more).

They are "Boardings", "Bus/car numbers", "Weather", "Effective transportation change" and "Age group".





Discussion and Limitations

From both models we can find that "Effective transportation change" sits as one of the top features. This means that the ability to smoothly switch between vehicles is of primary importance for the users;

Although the models provided informative features which align with the goal of the project, we found a few limitations in our project, which could be addressed in future analysis. First, there are many features which contain a high number (>50%) of missing values. This means that the dataset might not be as comprehensive as initially thought.

In a future project, it might be valuable to focus on the accessibility of the transportation network for different categories of users. Using data such as age, gender and profession in combination with the satisfaction scores might provide additional insight into the weaknesses and strengths of the service.

Conclusion

Therefore we strongly recommend HSL can focus on optimizing their transportation changing system as it will improve users' experience most effectively. Then HSL needs to maintain their basic transportation service in "Fast and Smooth Travel" and "Punctuality". They also need to push transportation information to the public in a timely and accessible fashion. HSL needs to remind their staff's attitude towards customers especially for drivers and inspectors.

Appendix:

Full code:

https://colab.research.google.com/drive/1BJFdGe3eYM62DjuEF4LbadDL7SKVMvPI?usp=sharing

Works Cited

González, L., Perdiguero, J., & Sanz, À. (2021). Impact of public transport strikes on traffic and pollution in the city of Barcelona. Transportation Research Part D: Transport and Environment, 98. https://doi.org/10.1016/j.trd.2021.102952

Jiménez, F., & Román, A. (2016). Urban bus fleet-to-route assignment for pollutant emissions minimization. Transportation Research Part E: Logistics and Transportation Review, 85, 120-131. https://doi.org/10.1016/j.tre.2015.11.003

HSL. (2023). Data API. Retrieved from Asty Web: https://hsl.louhin.com/asty/help

Institute, F. M. (2023). Weather Observations. Retrieved from Finnish Meteorological Institute: https://en.ilmatieteenlaitos.fi/download-observations