Predicting Public Transport Adoption with Machine Learning: Regional Mode Choice Patterns

Guglielmo Fadiga[2], Alexandre Ferreira[2], Marta Norte[1], Ricardo Rocha[1], and Albert Wood[2]

[1] Faculty of Aerospace Engineering, Delft University of Technology
Kluyverweg 1, 2629 HS Delft, The Netherlands
[2] Faculty of Mechanical Engineering, Delf University of Technology
Mekelweg 2, 2628 CD Delft, The Netherlands

Abstract. Increasing the adoption of public transport, and thereby reducing the use of private transportation modes, is a key step in the reduction of emissions in urban environments. Given that transportation behaviour differs between regions, this project aimed to investigate in which regions of Switzerland, increasing the quality of service of public transport would lead to a higher increase in the adoption of this means of transport.

For this purpose, it was important to understand what motivates mode choice behaviour. Machine learning was used for this task as it has a strong ability to explore the correlation between different features and can therefore be a powerful behaviour modelling tool. A logistic regression model and a neural network were built. The neural network outperformed the logistic regression, with a final precision of 92.7%, an accuracy of 89.5%, and an F1-score of 92.92%.

This study found that walking time is the most influential parameter and that all regions could benefit from its reduction, which can be done by increasing the network coverage. Investement in quality of service should focus on Valais, Zürich, and Eastern-Switzerland, with Zürich being the most promising. Possible extensions to this study can include the estimation of the amount of people shifting from car to public transport in each region, followed by a CO_2 reduction estimation.

Keywords: Mode Choice · Machine Learning · Travel Behavior · Logistic Regression · Neural Networks

1 Introduction

1.1 Motivation

In the face of growing environmental concerns and increasing urban mobility demands, sustainable transport has become a key focus in transport policy. Shifting travellers toward greener modes, such as public transportation and active travel, is seen as a way to reduce emissions and improve congestion. However, for this to happen effectively, the reason why people choose one travel mode over another must be understood. Traditional models can be useful, but recent advancements in machine learning provide powerful new tools for exploring such decisions. This project aims to apply these tools to identify which regions in Switzerland may be falling behind in public transport adoption and could therefore benefit from targeted investment.

Switzerland is known for its well-developed public transportation system, but it is not clear whether all regions benefit from or make use of it to the same extent. Some regions might rely more on cars, while others might use public transport more often. This brings up interesting questions: are there patterns in how people choose their mode of transport? And do factors like income, location, or travel distance play a role in these decisions? This project explores these questions using machine learning. The goal is to find out where public transport might be underused and what could be influencing that, so that future investments or policies can be more targeted.

1.2 Research Questions

This project aims to answer the following main research question:

In which region of Switzerland could there be a larger increase in the choice of public transport as a means of transportation if its quality was improved?

To explore this, several related questions will be investigated:

– How does modifying public transport service quality, through changes in, for example, waiting time, number of transfers, cost, and journey time, affect individuals' travel mode choice across Swiss regions?

– Can a machine learning model based on fixed socio-economic and demographic characteristics accurately and precisely predict mode choice?

– Can machine learning be used to identify which regions have the highest potential for increased public transport usage under improved service conditions?

By answering these questions, the study seeks to provide actionable insights for improving public transport infrastructure, particularly in regions where current usage is low but the potential for mode shift is high.

1.3 Brief Literature Review

Commuting behavior has also been shown to depend on both individual motivations and the built environment. Charreire et al. (2021) found that walking, cycling, and public transport use were influenced not only by socio-demographics but also by urban form and trip purpose, with factors like access to infrastructure and perceived convenience playing key roles [1]. Several other studies have emphasized the importance of incorporating latent attitudes and lifestyle preferences into mode choice modeling [2] [3]. Atasoy et al. (2010) integrated qualitative and quantitative methods to develop latent variable models that explain behavioral factors such as environmental concern [2]. Hillel et al. (2021) provided a comprehensive review of machine learning approaches in mode choice modeling, highlighting their potential and common methodological pitfalls [4]. Kashif et al. (2022) demonstrated how interpretable ML techniques can shed light on the feature importance behind travel decisions [5].

In parallel, existing environmental transport studies have established models that estimate CO_2 emissions based on travel time by car. Reports from the European Environment Agency provide average emissions per kilometer [6]. These approaches support assessing the environmental impact of predicted shifts in travel mode.

This work builds on these contributions by using machine learning not only to predict transport mode choice, but also to identify regions with the highest potential for public transport uptake.

2 Data set preparation

This chapter describes how the dataset was prepared for modeling. In Subsection 2.1, the original survey data is introduced, including demographic, travel, and attitude-related information. In Subsection 2.2, several preprocessing steps were applied to make the dataset suitable for building predictive models.

2.1 Dataset Description

The data set is composed of 2265 entries from a survey conducted on transportation in Switzerland. Data points include socio-economic information such as age, language, region and salary. It also details travel patterns for a specific day. Each entry represents a linked trip from and returning to the starting location, with details including travel time, distance, and choice of transport method included. If a person made multiple trips within the specified day, then the demographic information is duplicated.
The data also includes a set of questions to judge attitude towards different modes of transportation, such as how likely an individual is to support public transit or active modes. Due to the nature of data collected through a survey, there are missing data points where an answer was not provided. This poses a challenge for prediction models as all data points must be present to properly make predictions.

2.2 Data preprocessing

To perform a sensible preprocessing of the data, it should be noted that all our research questions can be answered by building and training a prediction model for the choice of transportation. Because of this, the following steps have been implemented:

1. Deleting all the row entries that had invalid responses for the *Choice* question. Doing this, the data set already reduces to 1906 total entries.

2. Part of the data present in the original data set gets deleted, as it would only make the model heavier without any positive impact on our prediction objectives. In particular, the following columns get deleted: *ID*, *Weight*, *CoderegionCAR* (because it's the same as the *Region* column). Removing more data is explored later in the report to determine if it gives a better result.

3. Non ordinal variables are transformed to one hot encoding. Specifically the columns of: *DestAct, HouseType, Mothertongue, FamilSitu, OccupStat, SocioProfCat, TripPurpose, TypeCommune, ClassifCodeLine, ResidChild, Region and ModeToSchool.* This means that these values are converted from a single variable where each specific value has a meaning to multiple binary variables.

4. All the variables that didn't get converted to one hot encoding get scaled between between -1 and 1 using `StandardScaler()`. Note that also some categorical ordinal variables (non numerical) are included here like *Income* or *Education*.

5. The problem of missing responses and unavailable data points needs to be addressed. Several solutions have been implemented.
   One hot encoded variables get treated by either deleting all the rows that contain one or more missing data points, or by adding an extra variable, which indicates if the original variable was missing.
   For scaled variables five different imputation methods have been tested:

   - deletion of rows with one or more missing data points.
   - data imputation by forward fill.
   - data imputation by most frequent entry.
   - data imputation by mean value.
   - data imputation with `IterativeImputer()`, which models each feature with missing values as a function of other features.

Our data set is large, but by simply deleting all rows that contain missing data points, less than 1000 entries are left, which, in our testing, proves to be not sufficient to train an accurate prediction model. Data imputation is thus required to preserve as much data as possible. This also implies that some "noise" and slight inaccuracies might be generated in our data set; this risk needs to be assessed and managed. In this report the specific combinations of input methods, including fill method, will be explored to determine the most optimal combinations for the models used.

3 Data Analysis

This chapter analyzes the key factors influencing mode choice. In Subsection 3.1, regional differences in travel behavior and transit quality are explored. In Subsection 3.2, logistic regression is applied to identify the most important predictors of mode choice. Finally, in Subsection 3.3, PCA is used to group related survey questions into categories for analysis.

3.1 Exploratory data analysis

In the initial data analysis, factors affecting mode choice were compared between the regions of Switzerland. The overall mode choice metric in Figure 1, which is what the models will be trained to predict, indicates a higher value closer to 1 if a higher percentage of trips are taken by private mode. Vaud has, on average, a higher share of trips taken by a private mode than other regions, and Graubunden has a higher share of trips taken by public transport.

In the data, differences in transit quality, including cost, time, and convenience, between the regions were looked for, which can explain the differences in mode share. Vaud, which has the highest share of private mode use, exhibits a lower cost to drive than other regions, slower public transit trips and longer walking time to the transit stop, shown in Figures 5, 3, and 2. Graubunden, which has the highest public transit mode share, has the fastest public transit [3] and the shortest walking times. In Figure 4 it is interesting to note that

[3] In this report, "public transport speed" is defined as the average speed over the trip, calculated as the distance in kilometers (distance_km) divided by the total time spent on public transport (TimePT), resulting in a speed metric with units of km/min. This was used instead of raw travel time because in larger or more urbanized regions, longer travel times may reflect longer distances rather than lower service quality. Speed provides a more comparable and meaningful indicator of network efficiency across regions.

in areas with a higher public transit mode share, transit costs are also higher, and the inverse is true as well. This leads to the hypothesis that the most important factors in mode choice are speed and convenience and that lower cost of transit alone may not be enough to entice more people to use it. There are other, more complex factors, that also play a role in mode choice, including attitudes and specific living conditions. More factors will be explored in the next phase using logistic regression.

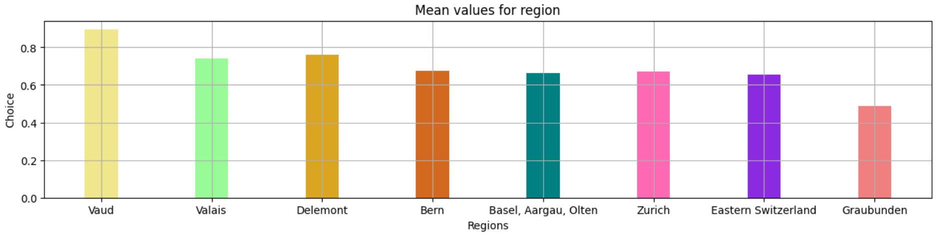


Fig. 1: Mean Choice per Region. Higher indicates private modes are used more frequently

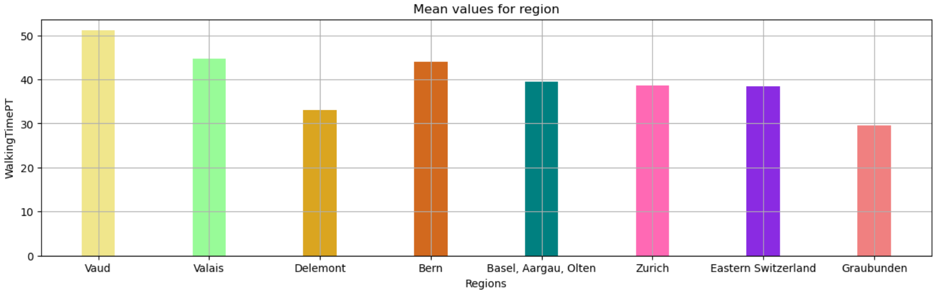


Fig. 2: Mean Walking Time to Public Transit Stop

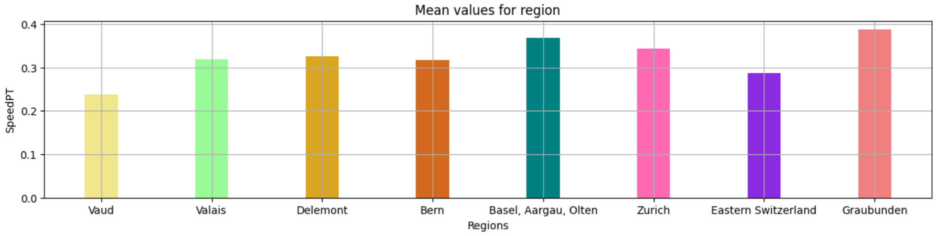


Fig. 3: Mean Speed for Public Transit

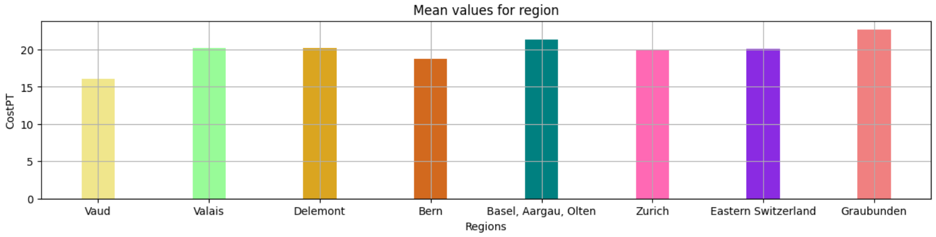


Fig. 4: Mean Cost of Public Transit

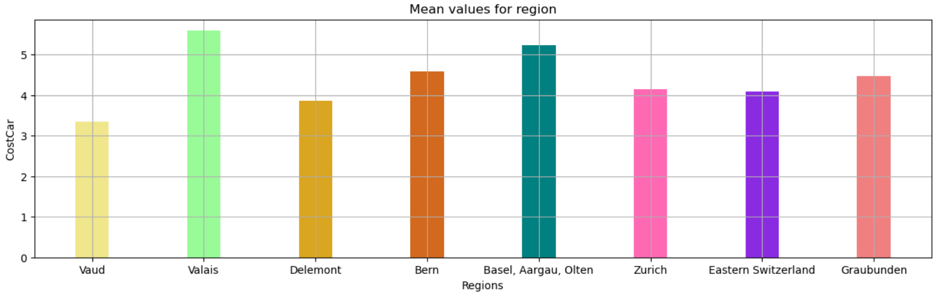


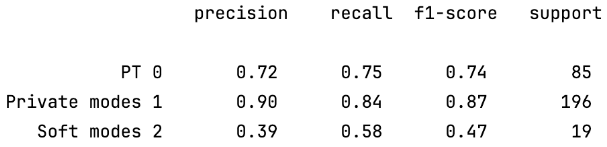
Fig. 5: Mean Cost of driving (mainly gas) per region

3.2 Logistic Regression Model

A Logistic Regression model was trained to predict the mode choice. This model was primarily used to compare its performance against a neural network (NN) model (more on that in Section 4), and secondly, to extract the different weights and determine which factors are most influential in the mode choice prediction.

Soft modes (cycling and walking) were initially included in the analysis. However, as shown in Figure 6, the model performed poorly for these categories, with F1 scores below 0.5, indicating that the predictions were neither accurate nor reliable. This is because the number of entries in the dataset for which the mode choice is soft modes is very small, so there is not enough support for the model to make these predictions correctly.

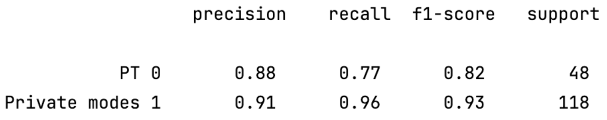
In addition to their poor predictive performance, soft modes were excluded because people who use them often do so for enjoyment or health-related reasons, making them less responsive to changes in transport service quality [1]. When soft modes are removed, as shown in Figure 7, the model's overall performance improves, not only for the excluded categories but also for the remaining ones. The final logistic regression model, which does not include soft modes, achieves significantly better results and is therefore used as the baseline for comparison with the Neural Network.



```
                 precision    recall  f1-score   support

            PT 0      0.72      0.75      0.74        85
 Private modes 1      0.90      0.84      0.87       196
    Soft modes 2      0.39      0.58      0.47        19
```

Fig. 6: Performance of Logistic Regression with Soft Modes



```
                 precision    recall  f1-score   support

            PT 0      0.88      0.77      0.82        48
 Private modes 1      0.91      0.96      0.93       118
```

Fig. 7: Performance of Logistic Regression without Soft Modes

In order to achieve the best performance from the logistic regression, an analysis was conducted to determine which combination of data preprocessing characteristics yielded the best results. A full breakdown of all the characteristics can be seen in Subsection 4.1, and the results for each configuration can be viewed in Appendix A, Table 2. These characteristics include how missing data is treated, whether the data is augmented, and whether attitude questions are included. Precision was prioritized by setting a precision threshold of 0.92, ensuring that the least amount of potential public transit users were missed. The F1 scores of the models that exceeded this threshold are plotted against accuracy in Figure 8. For logistic regression, a combination of imputation of data by mean value or iterative imputation and including all data with augmentation leads to the highest F1 score, which is representd by point 9 in Figure 8.

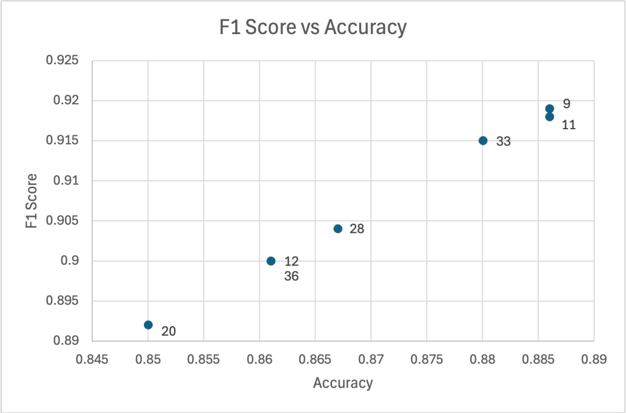


Fig. 8: F1 Score vs Accuracy for Logistic Regression

Each feature inputted into a logistic regression is weighted, which can provide insight into which factors were most influential on mode choice. Figure 9 shows the top 20 most influential factors in the logistic regression. A positive weight value indicates that as this factor increases, the likelihood of choosing a private mode increases as well. Transit quality indicators, including waiting time and walking time, are highly important to mode choice, with an increase in these factors understandably leading to lower likelihood of using public transit. Other factors that are also highly influential to choice include destination activity and socio-economic situation, which are beyond the focus of this report.

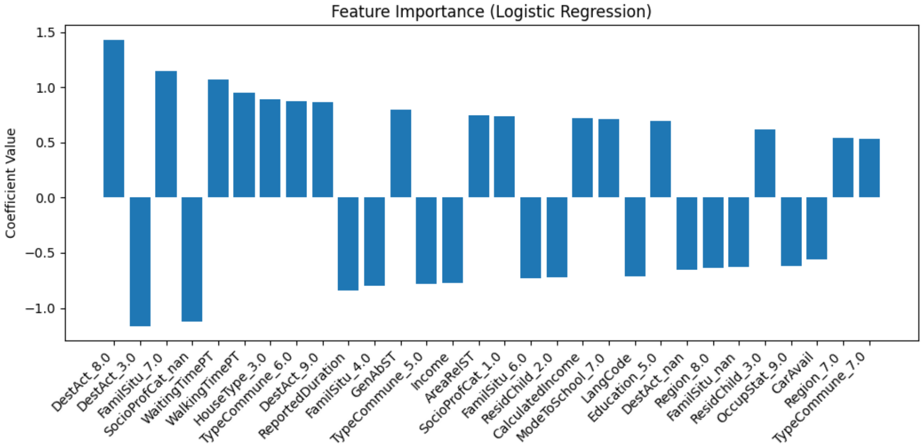


Fig. 9: Factors most influential to Logistic Regression

3.3 Principal Component Analysis for Question Clustering

A Principal Component Analysis (PCA) was initially applied to our data set with the intent of reducing feature dimensionality and thus removing some noise of less relevant features. In fact, training a prediction model with principal components instead of the singular features might help avoid overfitting problems and enhance the most influential feature for the desired output.

However, training the prediction models with principal components resulted in a notable loss of overall performance. It was thus decided to use the PCA to divide the data into "question categories" based on the correlation of their responses. The core idea of this procedure follows the work in [7]. These question categories provided some interesting insights into which ones might be more relevant in addressing the research questions. This consists of unsupervised learning, where questions are clustered by their response correlations. After performing the PCA, we perform a normalization using a SoftMax Function and then we filter out the most predominant questions on each component.

PCA with softmax normalization

PCA is an unsupervised method, and thus, it blindly tries to see correlations in features only by looking at their responses. The consequence of reducing dimensionality by PCA is that we also lose some variance in the data. Choosing an appropriate value for the cumulative explained variance of the output was not trivial, but a common rule use is to keep 90% as described in [8], but since our focus is to reduce the number of question categories, a trade-off was made. The chosen value of the cumulative explained variance ratio was 80% which resulted in 48 classes of questions. The PCA component values per category are shown in Figure 10.

In each of the resulting components from PCA, some of the questions have similar weights. Since we want to separate them, and get a clear distinction between questions, a softmax normalization is initially applied. After that, the principal components are filtered with a threshold of 0.06 to gather a binary representation of each feature contribution to the different principal components. This value was chosen as the minimal value for which every question falls in a specific component. A too high threshold will lead to some questions not falling in any of the components, a too low threshold will lead to questions falling in a big set of components at the same time. The end result can be seen in Figure 11.

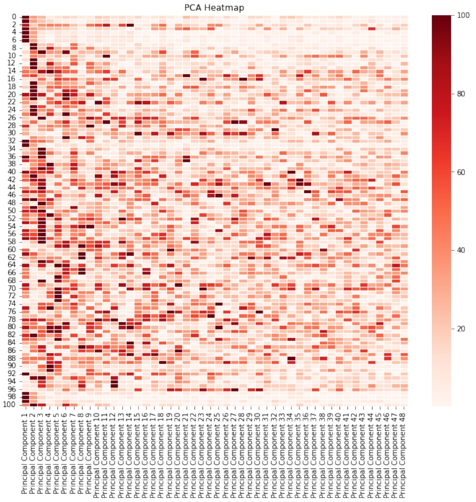


Fig. 10: PCA heatmap

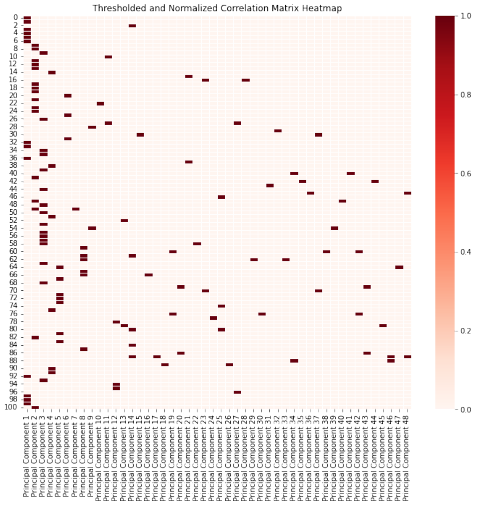


Fig. 11: PCA after softmax normalization and binary representation

Each of the columns in Figure 11 represents a set of questions that belong to a specific category. Since each question is a point in a large feature space, our goal would be to understand were in that feature space would our categories fall in. Lets think on this way, if each of our question is a vector in a feature space, then if we add all questions of a category together we will end up in a different point on our feature space, and thus that point is the representation of the category. With that we can understand what would be the name of the category.

Since this is the core principal behind large language models such as chatgpt, we will use it to get a summary of what each category encapsulates. The results are presented in Appendix B, Table 3.

An additional step was to see if there was any relevant connection between different categories. For that, a linear regression was performed between categories and the r^2 value was registered. A value of 30% was selected as a threshold since it represents at least a weak correlation as described by [9]. The heatmap in Figure 12 shows the r^2 of each of the calculated correlations.

This analysis is really interesting in the way that it shows some correlation between categories that highlights more information regarding the data we are dealing with. We see a significant correlation between people considering the car as a practical tool and Material Assets and Practicality. We also see a strong correlation between Car Freedom and Housing, Routine Travel Habits and Digital Tools for Mobility. Although the variables are quite different, it's fair to admit that both examples are related to how socioeconomic status affects transportation means and possession of goods.

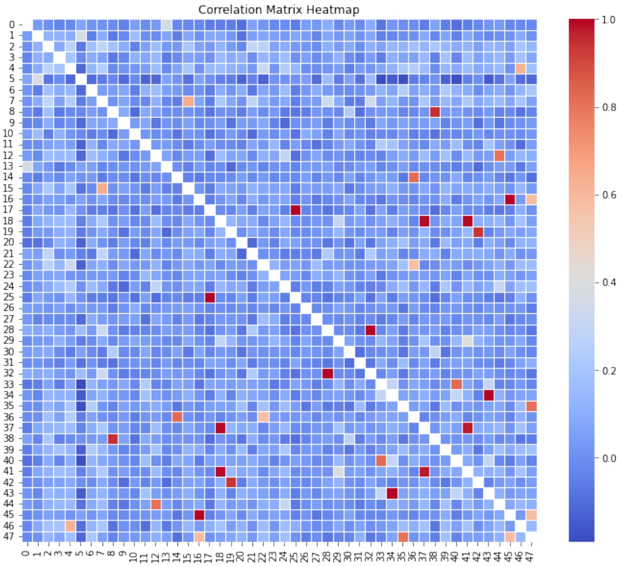


Fig. 12: PCA category correlation heatmap (r^2 values)

Other, more evident relations are also shown, such as Parking Discomfort and Home Ownership and Parking Issues.

In conclusion, this analysis shows that transportation behavior is a multifaceted phenomenon, strongly influenced by a combination of socioeconomic conditions, household structure, environmental awareness, and more.

4 Neural Network Methodology

Logistic regression and Neural Network techniques are both commonly applied in classification problems. Logistic regression was explored in Section 3, where it provided decent results and helped highlight which factors were most important in predicting mode choice. However, logistic regression is a much simpler model compared to neural network, which, depending on their architecture, are often better suited to capturing nonlinear relationships in the data.

For that reason, this chapter presents the neural network methodology used to predict mode choice. In Subsection 4.1, different input configurations were tested by varying preprocessing strategies such as imputation methods, categorical feature handling, inclusion of attitudinal variables, and data augmentation. In Subsection 4.2, multiple neural network architectures were evaluated to identify the best-performing design. Finally, in Subsection 4.3, the selected model is compared to the logistic regression benchmark and existing literature to validate its performance.

4.1 Feature Selection and Preprocessing

A range of preprocessing strategies was systematically applied to assess their impact on model performance. These strategies concerned how missing data was handled, how categorical features were treated, whether attitudinal variables were included, and whether data augmentation was employed. The aim was to isolate the contribution of each factor to the model's predictive capability, while maintaining a consistent neural network architecture across all experiments.

Input Configurations

The following preprocessing dimensions were varied:

– Missing Data Handling (Filltype): Five different approaches were used to deal with missing values.

- Type 0: Row deletion of entries containing any missing values;
- Type 1: Forward fill imputation (propagating the last valid value forward);

  - Type 2: Imputation with the most frequent value in each feature;
  - Type 3: Iterative imputation using `IterativeImputer`, where each feature with missing values was modeled as a function of the others;
  - Type 4: Mean value imputation for numerical features.
- Categorical Feature Handling (CatDealer): When enabled, entries with missing values in categorical variables were removed. When disabled, categorical and numerical features were processed jointly during imputation.
- Inclusion of Attitudinal Variables: Some configurations excluded variables related to environmental views, lifestyle preferences, and mobility habits in order to examine their influence on prediction accuracy.
- Data Augmentation: For selected configurations, Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generates new synthetic samples of the minority class based on nearest-neighbor interpolation, aiming to address class imbalance and improve generalization.

Neural Network Architecture

To ensure that the comparison across preprocessing pipelines remained fair, the neural network architecture was held constant throughout all experiments. The chosen model consisted of:

- An input layer matching the number of features in the dataset;
- Three fully connected hidden layers with 64, 32, and 16 neurons, respectively, all using ReLU activation;
- An output layer with 2 neurons using softmax activation for binary classification;
- The Adam optimizer was used, with categorical cross-entropy as the loss function.

This architecture was selected due to its balance between simplicity and representational power. Its consistency ensured that any performance differences observed could be attributed to preprocessing choices rather than model architecture.

Experimental Design

A total of 40 unique preprocessing configurations were constructed by combining the options above. Each configuration was evaluated using five separate training runs with different random seeds to account for variability in data splits. The mean accuracy, precision, and F1-score over the five runs were computed to provide a robust estimate of model performance.

A summary of the results for all 40 configurations is provided in Table 4 in Appendix C. Each row corresponds to a distinct preprocessing method defined by its treatment of missing values, use of categorical cleaning, inclusion of attitudes, and whether data augmentation was applied.

Choice of Input Data Configuration

While all 40 neural network configurations performed relatively well (with accuracy values never falling below 87% and precision scores consistently around or above 90%) it was still necessary to select one as the basis for further analysis. For this selection, precision was prioritized over accuracy. This is because precision reflects how often the model is correct when it predicts the positive class—in our case, public transport users. Since the goal of this study is to identify potential increases in public transport usage, we care most about ensuring that predictions of mode shift to public transport are indeed correct. Although overall accuracy remains relevant (as it measures how often the model is right in general), it can be misleading in imbalanced datasets or when false positives are more costly than false negatives.

To this end, a threshold of 92.5% was applied to the precision score. This filtering resulted in five configurations that exceeded this threshold. From these, the final configuration was chosen based on the highest mean F1-score. The F1-score represents a harmonic mean between precision and recall, the latter of which is also important because it measures how many actual positive cases (public transport users) were correctly identified. A high recall ensures that true opportunities for increased public transport use are not overlooked.

The final decision was supported by the results shown in Figure 13. Configuration ID 23 was selected as the final input configuration, achieving the highest F1-score among the high-precision candidates. [4]

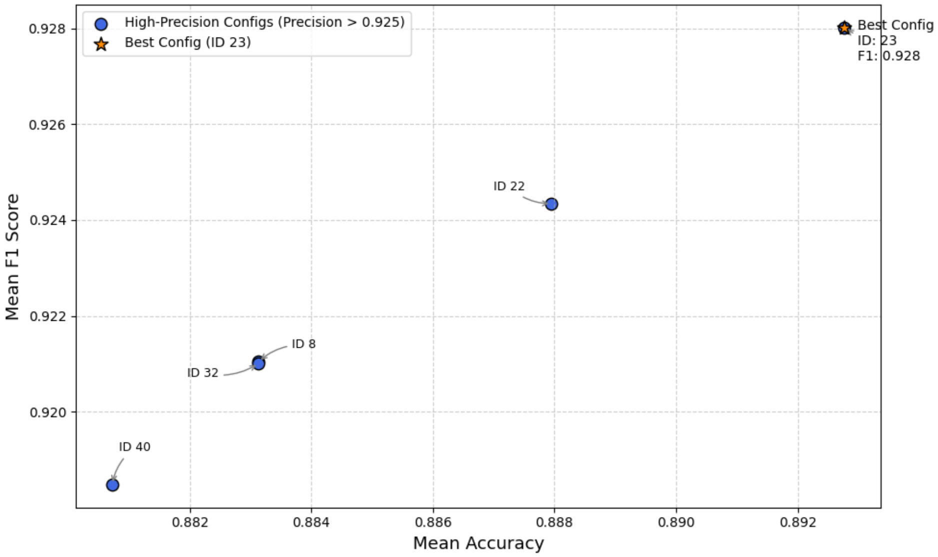


Fig. 13: High-precision configurations (Precision $>$ 0.925) plotted by accuracy and F1-score. Configuration ID 23 achieved the best overall balance.

4.2 Model Architecture

With the input configuration set to ID 23, consisting of mean imputation (Filltype 2), no data augmentation, categorical cleaning enabled (CatDealer 1), and the inclusion of attitudinal variables, attention was turned to the neural network architecture.

A total of 18 distinct architectures were tested, each varying in the number of hidden layers, number of neurons, and batch size used during training. The goal was to explore different levels of model complexity, while keeping activation functions fixed to eliminate unnecessary variables. Specifically:

- All models used `ReLU` activation in hidden layers and `softmax` in the output layer, as these are standard in classification tasks and had already produced satisfactory results in earlier experiments;
- Each layer configuration was tested using three batch sizes (32, 44, and 64) to evaluate training dynamics;
- Five random seeds were used across all architectures to ensure fairness and reproducibility.

The architectures tested ranged from shallow (a single hidden layer with 32 neurons) to deep networks with ten layers. The full list of models, with average accuracy precision, and F1 score over five runs, is shown in Table 1.

[4] It should also be noted that validation and training loss curves were not analyzed at this stage to assess potential overfitting. This evaluation is more appropriate when tuning or selecting the model architecture itself and was therefore deferred to a later stage. Loss curves will be used to evaluate model convergence and overfitting in future analysis when architectural changes are introduced.

Table 1: Performance of 18 neural network architectures using fixed input configuration (ID 23).

Architecture	**Hidden Layers**	**Batch Size**	**Accuracy**	**Precision**	**F1 Score**
1.1	[32]	32	0.8940	0.9235	0.9292
1.2	[32]	44	0.8843	0.9130	0.9231
1.3	[32]	64	0.8928	0.9165	0.9289
2.1	[64, 32]	32	0.8940	0.9172	0.9298
2.2	[64, 32]	44	0.8904	0.9233	0.9267
2.3	[64, 32]	64	0.8867	0.9108	0.9250
3.1	[64, 32, 16]	32	0.8976	0.9186	0.9323
3.2 - Baseline	[64, 32, 16]	44	0.8952	0.9297	0.9297
3.3	[64, 32, 16]	64	0.8807	0.9064	0.9213
4.1	[128, 64, 32, 16]	32	0.8952	0.9187	0.9307
4.2	[128, 64, 32, 16]	44	0.8916	0.9210	0.9278
4.3	[128, 64, 32, 16]	64	0.8819	0.9105	0.9218
5.1	[64, 32, 16, 32, 64, 32, 16]	32	0.8892	0.9188	0.9262
5.2	[64, 32, 16, 32, 64, 32, 16]	44	0.8831	0.9151	0.9224
5.3	[64, 32, 16, 32, 64, 32, 16]	64	0.8855	0.9175	0.9237
6.1	[128, 64, 32, 16, 32, 64, 128, 64, 32, 16]	32	0.8867	0.9140	0.9250
6.2	[128, 64, 32, 16, 32, 64, 128, 64, 32, 16]	44	0.8831	0.9103	0.9225
6.3	[128, 64, 32, 16, 32, 64, 128, 64, 32, 16]	64	0.9012	0.9191	0.9348

As before, a precision threshold of 92.5% was applied to filter models that have high values of precision. Among those that passed the threshold, the best F1-score was once again used as the deciding metric, for the same reasons discussed in Section 4. Interestingly, this process led back to the architecture initially used in our earlier experiments—Architecture 3.2 (Baseline)—as the optimal one.

Although it might be tempting to believe that more complex architectures (featuring deeper or broader layer structures) will always surpass simpler ones because of their greater representational ability, this isn't necessarily true. In reality, training extensive models with comparatively small datasets frequently results in overfitting and weak generalization. This emphasizes the importance of choosing architectures that are both expressive and suitably aligned with the data size and complexity of the task

To validate the choice further, training and validation accuracy curves for Architecture 3.2 are shown in Figure 14. These curves suggest that the model generalizes well: both metrics converge, and validation accuracy stabilizes around 92–93%, with no major divergence between training and validation over 100 epochs. This provides evidence that the model does not suffer from overfitting, and is well-calibrated to the problem at hand.

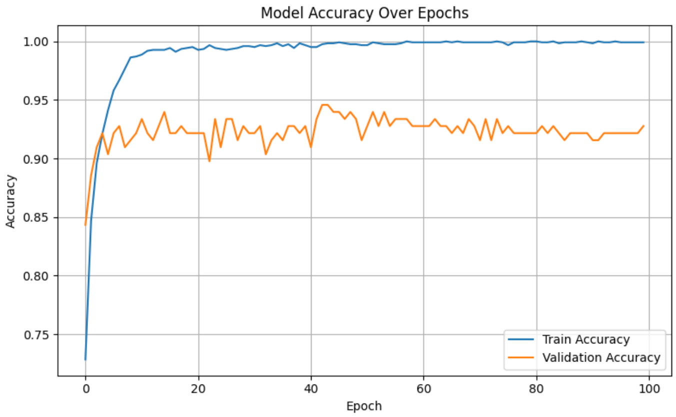


Fig. 14: Training and validation accuracy over 100 epochs for the selected neural network architecture (3.2).

4.3 Comparison with Benchmarks and Literature

To ensure that the results obtained from the model are meaningful and reliable, it is important to compare them with both a baseline model and with findings in existing literature.

The logistic regression model discussed in Subsection 3.2 served as a useful benchmark. It achieved solid predictive performance, with an average precision of 88% for public transport users and 91% for private mode users. Additionally, this model allowed for an understanding of which variables were most influential in determining mode choice, offering interpretability that is often useful in behavioral transport studies. While these results were strong, slightly better performance was obtained using the neural network model, which achieved an overall accuracy of 89.5%, a precision of 92.97%, and an F1-score of 92.97%. This suggests that the neural network was better able to capture more complex patterns in the data, particularly when incorporating a larger number of features, including attitudinal variables.

This difference in performance is consistent with what has been found in literature. Hillel et al. (2021) conducted a systematic review on machine learning methods for mode choice modeling and found that ML models (particularly neural networks) tend to outperform traditional models such as logistic regression in terms of predictive accuracy [4].

In addition, Atasoy et al. (2010) emphasize the importance of understanding the behavioral drivers behind mode choice decisions. Their work focuses on discrete choice models that incorporate latent variables and qualitative information to model decision-making more realistically [2]. While this paper does not attempt to model latent attitudes explicitly, the variables derived from PCA and attitudinal survey responses contribute to capturing some of this behavior in the neural network framework.

Taken together, the performance achieved by the models in this study is in line with expectations based on recent literature. The higher performance of the neural network, compared to logistic regression, reinforces its suitability for mode choice prediction in this context. At the same time, the consistency between model outputs and previous findings in the field provides a degree of validation that the results are meaningful and aligned with established knowledge.

5 Results and Discussion

This chapter presents the results of the final prediction model and discusses its implications. In Subsection 5.1, the public transport mode share in the original dataset is compared with the model's prediction. In Subsection 5.2, the impact of varying public transport service parameters on mode choice is evaluated both nationally and regionally. In Subsection 5.3, the potential effect of awarding different public transport season tickets to the population is analyzed. Finally, in Subsection 5.4, the influence of increasing car usage costs (e.g., through fuel price changes) is explored as a push policy to encourage public transport use.

Throughout this section, the regions of Switzerland are referred to by numerical identifiers (e.g., Region 1, Region 2, etc.). This convention was chosen by the authors to improve readability and ensure clarity when comparing results across regions. The regional mapping used is as follows: 1 = Vaud, 2 = Valais, 3 = Delemont, 4 = Bern, 5 = Basel–Aargau–Olten, 6 = Zurich, 7 = Eastern Switzerland, 8 = Graubünden.

5.1 Comparison of model results with original dataset

To compare the public transport mode share of the initial dataset and that predicted by the model, the dataset had to undergo some pre-processing such that it had the same shape as that required by the model. The pre-processed dataset obtained before the model training is reused here and its content is explained in Subsection 2.2. This dataset consists of 830 rows and 185 columns. Based on the 'Choice' column present in the dataset, the mode share of public transport is 30.28%. When the same dataset is fed into the final model, without the information on the mode choice, the model predicts a mode share of 27.71% for public transport. This is a small deviation but it is however, no guarantee that the model is making correct predictions, as the model was trained using part of this data. The accuracy and precision of the model that can be used for its evaluation are the ones presented in Subsection 4.3. These values will be used only in this section to evaluate how the change in certain parameters will impact the mode choice based on our model. Even though this results in some bias as the model was trained with part of this data, it is still expected that it provides an appropriate first estimate of the impact of some parameters on the mode choice.

5.2 Effect of varying public transport quality of service

The analysis of the available data through reasoning and PCA, showed that the following factors related to the quality of service of the public transport system have a impact on mode choice: number of transfers performed for all trips in the loop (NbTransf), the duration of the loop performed in public transport (TimePT), the total walking time in a loop performed in public transport (WalkingTimePT), the waiting time in a loop performed in public transport (WaitingTimePT), and the full cost for public transports to perform the loop (CostPT). In addition, the analysis showed that a lower value for these parameters was related to a higher mode share of public transport. A high quality of service is thus defined in this study as having low cost, low walking time, low time spent on transport, low waiting time, and a low number of transfers.

To evaluate whether improving the quality of service would increase public transport adoption, a sensitivity analysis was conducted. To simulate improvements or deterioration in service, the values of these parameters were varied from a 50% decrease to a 50% increase. This analysis was performed both globally (Figure 15) and by region (Figure 16).

Both graphs include two baselines: one taken directly from the dataset (the actual percentage of public transport users), and one from the model's prediction using the original data. These two baseline values are similar but not exactly the same, as discussed in Section 5. If a parameter is not changed (0% variation), the predicted share of public transport users lies on this baseline. Any deviation from the baseline reflects the effect of the parameter change.

In Figure 15, it is clear that walking time has the strongest influence on public transport mode share. This matches the results from the logistic regression analysis in Subsection 3.2, where walking time was also identified as a key factor. As expected, increasing the value of each parameter (worsening the service) led to a decrease in public transport use, while decreasing the values (improving service) led to an increase. All parameters followed this trend except for cost, which had a minimal effect—less than 0.25% variation. This suggests that cost may not be a major factor influencing mode choice in this context.

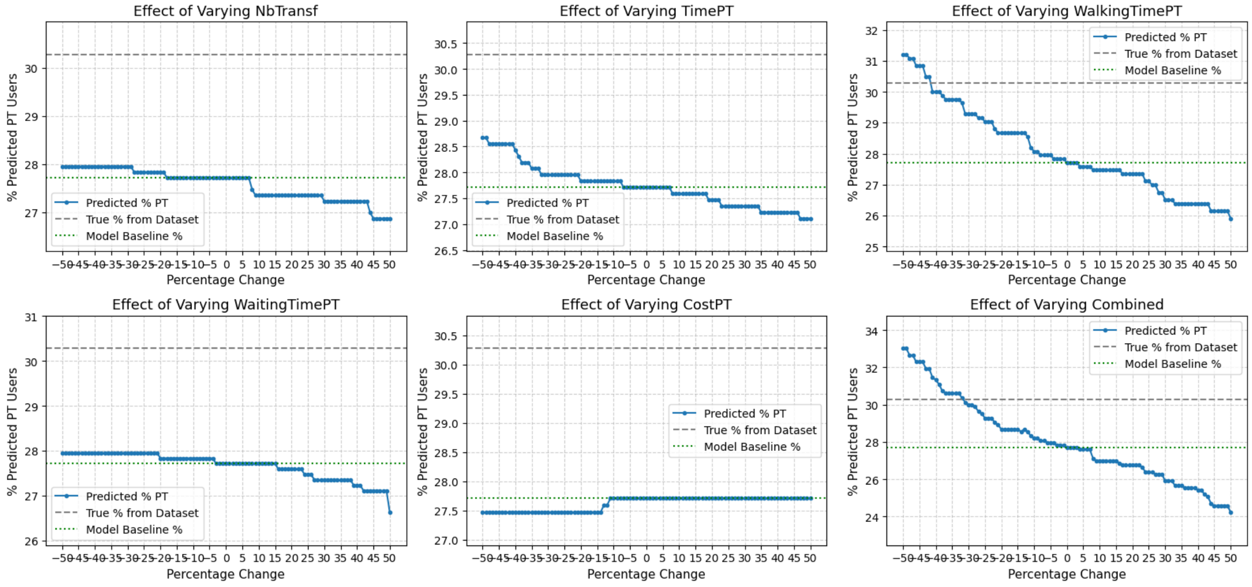


Fig. 15: Global Variation of Quality of Service

At the regional level, similar patterns were observed. As shown in Figure 16, walking time remains the most influential parameter. For instance, in Region 2, halving the walking time leads to a 3.25% increase in public transport share. This could be achieved by increasing the number of stations or stops in that region, improving accessibility. However, more stops can also increase the time spent on public transport, as the vehicle takes longer to complete its route. Therefore, it is important to consider how reducing walking time might simultaneously increase travel time, and to balance these effects when planning interventions.

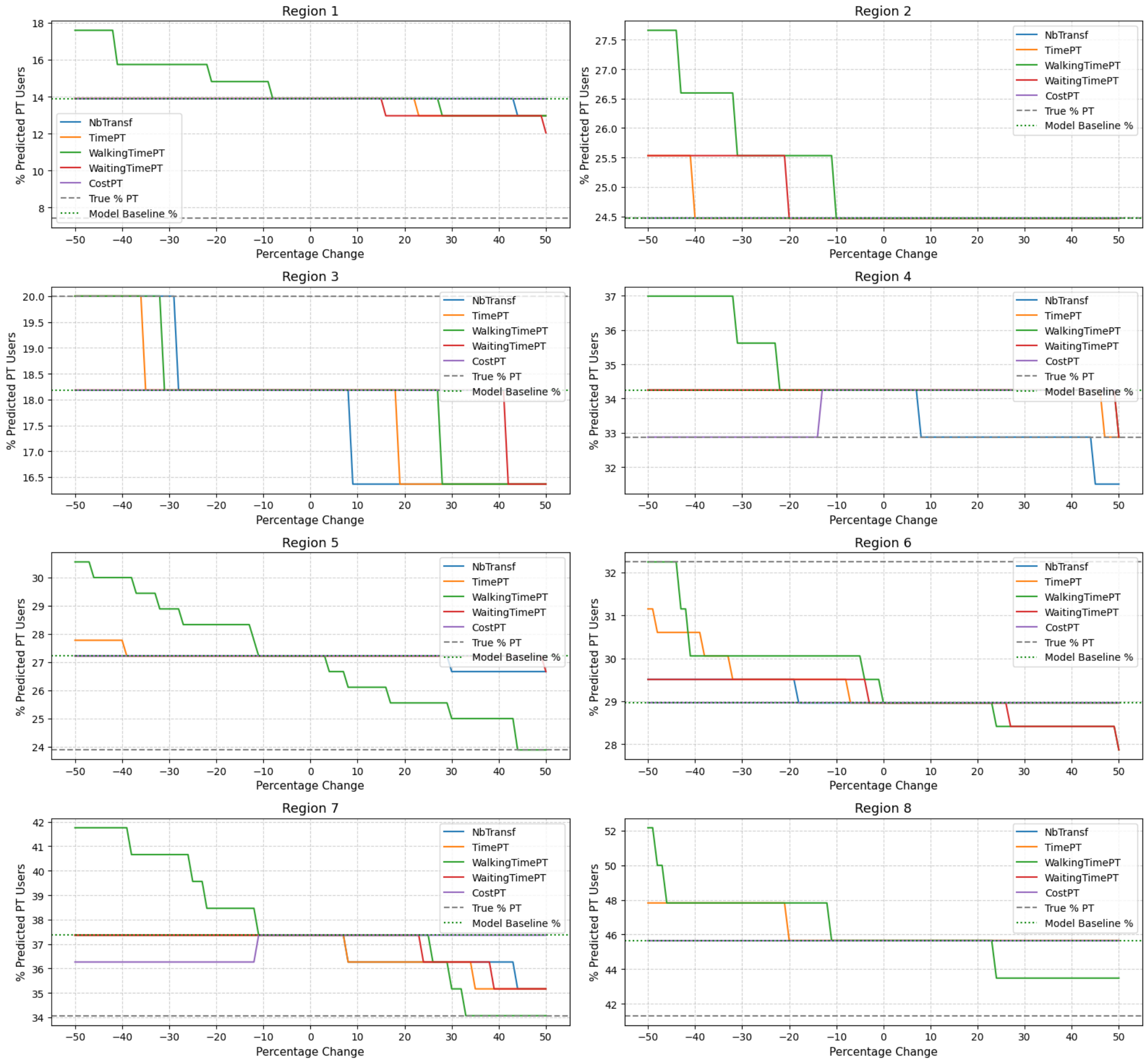


Fig. 16: Regional Variation of Quality of Service

For example, in Region 3, if the time on public transport increases by 20%, the share of users drops by around 2%, which may cancel out the benefits of reducing walking time. One potential solution is to increase the frequency of vehicles so that each one makes fewer stops, thus reducing total travel time without sacrificing accessibility. In contrast, Region 1 shows little to no change in response to increased travel time, suggesting that improving accessibility there may have more direct benefits, without incurring in any drawbacks.

The number of transfers is also important in some areas. For example, in Region 3, reducing the number of transfers by 30% increases public transport use by 1.8%. However, in regions 1, 2, 5, 6, and 8, this parameter has little effect, which may suggest that in those places, transfers are not a major factor in people's choice of transport.

Cost shows very little influence overall. This could be attributed to the relatively high income levels in Switzerland, where individuals may be less sensitive to public transport prices. This observation is consistent with the global results in Figure 15, where variations in cost produced minimal shifts in predicted mode choice. Additionally, as noted in Section 3, Region 8 exhibits the highest usage of public transport despite also having the highest average cost for public transit. This further supports the notion that factors such as convenience and accessibility may play a more decisive role in travel mode decisions than price alone.

Overall, it is difficult to say definitively which region would benefit the most from improved public transport service. As discussed, improving one parameter (e.g., walking time) may negatively affect another (e.g., travel time), depending on how the network is adjusted. Still, Region 3 seems to be the most sensitive to changes across multiple parameters, showing large shifts in mode share in response to modifications. Region 6 also appears promising, as improvements across several factors result in a clear increase in public transport share. Since Region 6 is Zürich, the most populated region in Switzerland, changes here would impact a larger number of people.

5.3 Effect of attributing season tickets to the population

Another possible measure to increase the share of public transport is to award season tickets to the population. In this analysis, it will be seen how the different season tickets impact the mode choice share.

First, an overview of the season ticket distribution over the population was obtained. As can be seen in Figure 17, most of the population already has a half-fare season ticket, but the share of people with the other season tickets is very low.

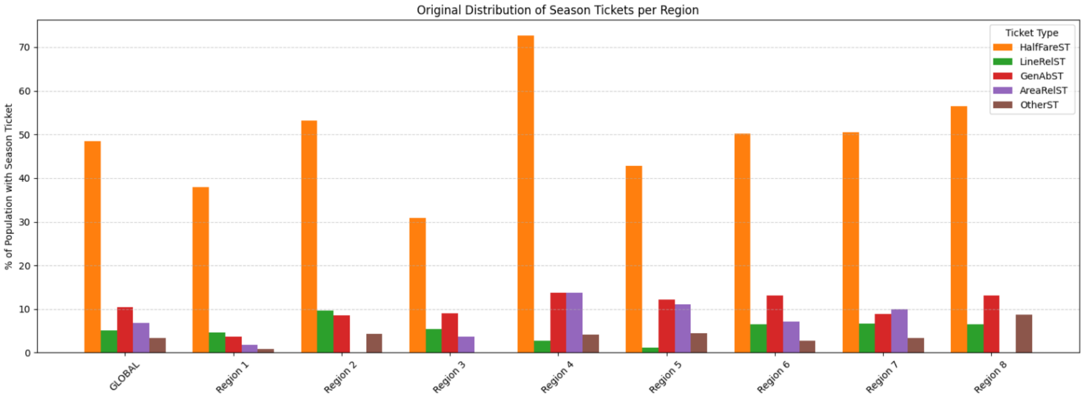


Fig. 17: Season Ticket Distribution

Figure 18 shows the effect of awarding people each type of ticket per region. This was done by awarding every person with a specific season ticket and inputting that dataset into the model. The model prediction was then compared to the original mode share. The impact of awarding half-fare season tickets is almost negligible which was expected as most of the population already has one. The general and area season tickets have the largest impact, especially in Region 8, where an increase of over 30% is observed. This is likely because this region has the lowest share of people with these tickets.

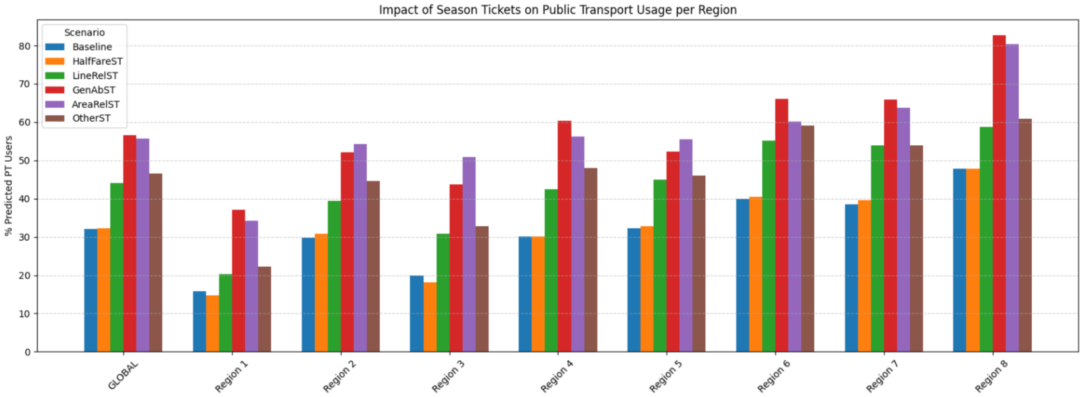


Fig. 18: Effect of Awarding Different Season Tickets

However, it is important to note that this analysis may be subject to reverse causality. That is, it is unclear whether owning a season ticket leads people to use public transport more, or whether people who already use public transport are more likely to purchase such tickets. It is probably a mix of both. Given this, we cannot confidently say that simply giving out public transport tickets would cause an increase in usage to the extent shown in Figure 18. While the effect might be positive, it is unlikely to be as large as the model suggests. This uncertainty should be taken into account when interpreting these results.

5.4 Effect of climate impact measures

Another possible way to increase the number of people using public transport is by increasing the cost of using a car through higher fuel prices. This is known as a push policy, because instead of making public transport more attractive, it discourages the use of alternatives like private modes.

Figure 19 shows how changing the cost of car usage affects the predicted percentage of public transport users.[5] We varied the cost from a 100% decrease to a 100% increase. According to the model, increasing the cost of using a car leads to a gradual rise in public transport usage. For example, doubling the car cost results in a predicted public transport share of around 29.2%, compared to the baseline of roughly 27.7Ṡimilarly, decreasing the cost reduces the share.

This trend can also be seen in the data analysis previously done, in Section 3, where Region 1, which shows the highest reliance on private modes, also has the lowest average cost of driving.

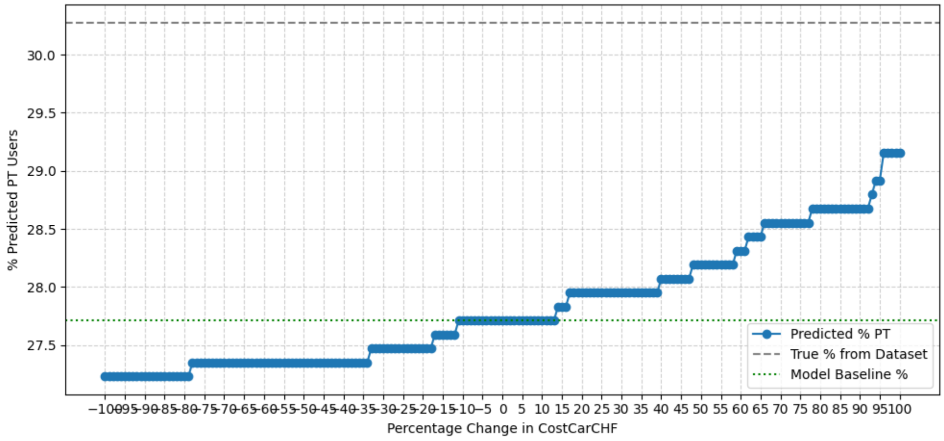


Fig. 19: Effect of Varying Cost of Car in CHF

At first glance, this might seem to contradict the earlier findings where the cost of public transport didn't show much influence. One possible explanation is that people in Switzerland, given the generally high income levels, may not be very sensitive to public transport prices. However, they might still be discouraged by high fuel prices, which could be seen as unnecessary spending. It's also possible that factors like environmental concern or social responsibility make people more willing to spend money on public transport rather than on fuel. Although, of course, this is just a hypothesis and would need more detailed research to confirm.

6 Conclusion

This project aimed to answer the following research question:

In which region of Switzerland could there be a larger increase in the choice of public transport as a means of transportation if its quality was improved?

The first step was to analyse the dataset. Regional differences in mode choice, walking time to public transit, mean speed of public transport and cost both of public transport and car were looked into. The regions which

[5] Since fuel prices are expected to stay the same across the country (to avoid people refueling in cheaper neighboring regions), this analysis was done at a national level only, and not per region as in previous cases.

exhibited higher car use had lower car costs, slower public transit trips, and longer walking times to public transport. On the other hand, the regions with higher public transport mode share, showcased faster public transport, shorter walking times and more expensive public transport cost. This showed that the cost of public transport might not be a key factor in mode choice but that speed and walking times are very relevant factors.

Afterwards, a regression model was built to use as a benchmark for the neural network. This logistic regression model had a precision of 88% when predicting public transport mode choice. This model also allowed the extrapolation of the importance of different factors for mode choice, with waiting and walking times standing out as quality of service parameters.

To further investigate the correlation between the features in the dataset, Principal Component Analysis (PCA) was applied to the dataset. It showed that mode choice is strongly influenced by socioeconomic conditions, environmental awareness and household structure, among others.

The final model developed was a neural network. To obtain the best data input configuration and neural network architecture, experiments with the different possible combinations were made, and a decision was reached based on accuracy and precision. The final data input configuration made use of mean imputation for missing values, did not make use of data augmentation, used categorical cleaning, and included attitude questions. The chosen neural network architecture has 3 hidden layers with ReLU activation functions, a softmax activation function applied to the output layer, and a batch size of 44. This model had a precision of 92.7%, an accuracy of 89.5% and an F1-score of 92.97%, which shows a better performance than the one obtained with the logistic regression.

Finally, the initial dataset was used as a baseline for the analysis of the effect of varying different parameters on the mode choice. It was observed that when it comes to quality of service, the walking time was the most influential parameter, while the cost of public transport was the least relevant. An increase in the public transport network coverage, balanced with fleet reinforcement might thus be beneficial. The relationship between being a season ticket holder and taking public transport was also investigated, showing that general and area season tickets have the largest impact, and the half-fare the least as most people already have this subscription. Awarding season tickets could thus lead to an increase in public transport adoption but the reverse causality present in this analysis introduces a degree of uncertainty in this prediction, overestimating the benefits. In terms of car cost, an increase in gas prices revealed an increase in public transport use, which suggests that push policies could be further investigated as they might discourage the use of private modes.

Regarding the main research question, it is difficult to define in which region there could be a larger increase in public transport adoption if there was an increase in its quality due to the complexity and dependence of mode choice on a wide range of factors. However, several recommendations can be made. Overall, all regions could benefit from a reduced walking time, which shows that the network coverage could be improved. Vaud, Bern, Basel-Aargau-Olten, and Graubünden are sensitive only to the decrease of a smaller number of parameters and therefore should probably not be prioritised. The investment in quality of service should then focus on Valais, Zürich, and Eastern-Switzerland, with Zürich being the most promising region for investment in quality of service, as it is sensitive to several factors, and is the most populated region, impacting a larger number of people.

There are several directions in which this work could be extended. One idea would be to estimate how many people would actually shift to public transport in each region, based on the predicted percentage changes and the known population. Combined with the average trip distance and using emission models discussed in Subsection 1.3—which estimate CO_2 emissions per kilometer [6]—this could give a rough idea of the environmental benefits of each intervention. However, due to the page limit and scope of the assignment, this analysis was not included.

7 Data and code availability

All the data and code used in this paper are openly available in the following GitHub Repository.

8 Contribution statement

- Guglielmo Fadiga: data set preparation and preprocessing, logistic regression implementation

– Alexandre Ferreira: initial NN implementation, PCA

– Marta Norte: neural network methodology, result discussion

– Ricardo Rocha: neural network methodology, result discussion

– Albert Wood: exploratory data analysis, logistic regression model tuning

All group members contributed to the writing of this report, with Ricardo and Marta playing a particularly significant role.

9 Use of generative AI/Chatbots

The use of generative AI was limited because the assignment required reasoning and critical thinking. Generative AI was mostly used to plot results and not to reason or develop the methodology. In the beginning, ChatGPT was used as an aid to research of concepts and get familiar with the topic. Apart from that, in the PCA analysis, it was used to generate the table containing the names of the components based on the description of the dataset.

References

1. H. Charreire, C. Roda, T. Feuillet, A. Piombini, H. Bardos, H. Rutter, S. Compernolle, J.D. Mackenbach, J. Lakerveld, and J.M. Oppert. Walking, cycling, and public transport for commuting and non-commuting travels across 5 european urban regions: Modal choice correlates and motivations. *Journal of Transport Geography*, 96:103196, 2021.
2. B. Atasoy, A. Glerum, R. Hurtubia, and M. Bierlaire. Demand for public transport services: Integrating qualitative and quantitative methods. In *Swiss Transport Research Conference (STRC 2010)*, Monte Verità, Switzerland, 2010.
3. B. Atasoy, A. Glerum, and M. Bierlaire. Attitudes towards mode choice in switzerland. *disP - The Planning Review*, 49(3):101–117, 2013.
4. T. Hillel, M. Bierlaire, M.Z.E.B. Elshafie, and Y. Jin. A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*, 38:100221, 2021.
5. M.T. Kashif, A. Jamal, M.S. Kashef, M. Almoshaogeh, and S.M. Rahman. Predicting the travel mode choice with interpretable machine learning techniques: A comparative study. *Travel Behaviour and Society*, 29:279–296, 2022.
6. European Environment Agency. Monitoring co_2 emissions from passenger cars and vans in 2022, 2023. Available at: `https://www.eea.europa.eu/en/newsroom/news/co2-emissions-of-new-cars`.
7. E. Vigneau and E.M. Qannari. Clustering of variables around latent components. *Computers & Chemical Engineering*, 28(1–2):317–326, 2003.
8. X. Yu, P. Chum, and K.-B. Sim. Analysis of the effect of pca for feature reduction in non-stationary eeg-based motor imagery of bci system. *Optik*, 125(3):1498–1502, 2014.
9. J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

A Logistic Regression Configuration Results

Table 2: Performance of the 40 logistic regresion experiments using different input data preprocessing and selection strategies

ID	Fill	Augmentation	CatDealer	Attitude	Accuracy	Precision	F1 Score
1	0	TRUE	0	TRUE	0.855	0.892	0.899
2	0	TRUE	0	FALSE	0.867	0.92	0.906
3	0	TRUE	1	TRUE	0.873	0.894	0.913
4	0	TRUE	1	FALSE	0.867	0.920	0.906
5	0	FALSE	0	TRUE	0.867	0.887	0.909
6	0	FALSE	0	FALSE	0.833	0.867	0.886
7	0	FALSE	1	TRUE	0.880	0.889	0.918
8	0	FALSE	1	FALSE	0.839	0.873	0.890
9	1	TRUE	0	TRUE	0.886	0.930	0.918
10	1	TRUE	0	FALSE	0.867	0.92	0.906
11	1	TRUE	1	TRUE	0.886	0.923	0.919
12	1	TRUE	1	FALSE	0.861	0.926	0.900
13	1	FALSE	0	TRUE	0.898	0.911	0.929
14	1	FALSE	0	FALSE	0.861	0.894	0.904
15	1	FALSE	1	TRUE	0.873	0.894	0.913
16	1	FALSE	1	FALSE	0.85	0.911	0.893
17	2	TRUE	0	TRUE	0.873	0.915	0.911
18	2	TRUE	0	FALSE	0.861	0.913	0.902
19	2	TRUE	1	TRUE	0.873	0.915	0.911
20	2	TRUE	1	FALSE	0.85	0.925	0.892
21	2	FALSE	0	TRUE	0.898	0.911	0.929
22	2	FALSE	0	FALSE	0.856	0.887	0.901
23	2	FALSE	1	TRUE	0.88	0.902	0.917
24	2	FALSE	1	FALSE	0.844	0.904	0.89
25	3	TRUE	0	TRUE	0.880	0.915	0.915
26	3	TRUE	0	FALSE	0.867	0.920	0.906
27	3	TRUE	1	TRUE	0.88	0.915	0.915
28	3	TRUE	1	FALSE	0.867	0.934	0.904
29	3	FALSE	0	TRUE	0.904	0.911	0.934
30	3	FALSE	0	FALSE	0.867	0.901	0.908
31	3	FALSE	1	TRUE	0.886	0.896	0.922
32	3	FALSE	1	FALSE	0.861	0.906	0.903
33	4	TRUE	0	TRUE	0.88	0.922	0.915
34	4	TRUE	0	FALSE	0.861	0.913	0.902
35	4	TRUE	1	TRUE	0.873	0.915	0.911
36	4	TRUE	1	FALSE	0.861	0.933	0.900
37	4	FALSE	0	TRUE	0.904	0.911	0.934
38	4	FALSE	0	FALSE	0.856	0.887	0.901
39	4	FALSE	1	TRUE	0.88	0.895	0.917
40	4	FALSE	1	FALSE	0.861	0.913	0.902

B Principal Component Suggested Names

Table 3: Principal Component Suggested Names.

Principal Components 1–24

PC	Suggested Name
PC1	Public Transport & Socioeconomic Profile
PC2	Car Ownership & Household Assets
PC3	Mobility Behavior & Social Perception
PC4	Lifestyle Preferences
PC5	Public Transport Familiarity
PC6	Household Income & Socio-Professional Status
PC7	Activity Interference
PC8	Car Comfort & Time Reliability
PC9	Mobility Routine & Language
PC10	Residence Ownership
PC11	Demographics
PC12	Social Openness & Routine
PC13	Mode Choice Flexibility
PC14	PT Transfers & Social Perception
PC15	Occupational Status
PC16	PT Navigation Knowledge
PC17	Social Status Perception
PC18	Car as Practical Tool
PC19	Car Freedom & Housing Mobility
PC20	Disorientation & Need for Breaks
PC21	Household Digitalization & PT Ownership
PC22	PT Difficulty with Luggage
PC23	Digital Tools for PT
PC24	Local Social Network

Principal Components 25–48

PC	Suggested Name
PC25	Environmental Concerns & PT Interaction
PC26	Material Assets & Practicality
PC27	Trip Purpose
PC28	Home Entertainment
PC29	Parking Discomfort
PC30	Housing & Mobility Conditions
PC31	PT Investment Support
PC32	Family Situation
PC33	Home Ownership & Parking Issues
PC34	Materialistic Display
PC35	Environmental Taxes & PT Support
PC36	Employment Priority over Environment
PC37	Flexibility in Mode Choice
PC38	Routine Travel Habits
PC39	Social Trust & Environmental Trade-offs
PC40	Climate Action Support
PC41	PT Cost Consideration
PC42	Digital Tools for Mobility
PC43	Urban Living Preference
PC44	Environmental Justice Concerns
PC45	PT Perception
PC46	PT Information & Interaction
PC47	PT Limitations on Activities
PC48	Outskirts Living Preference

C Neural Network Configuration Results

Table 4: Performance of the 40 neural network experiments using different input data preprocessing and selection strategies, averaged over 5 runs with varying random seeds.

ID	Fill	Augmentation	CatDealer	Attitude	Accuracy	Precision	F1 Score
1	0	True	0	True	0.8807	0.9217	0.9194
2	0	True	0	False	0.8880	0.9188	0.9251
3	0	True	1	True	0.8795	0.9150	0.9192
4	0	True	1	False	0.8771	0.9204	0.9171
5	0	False	0	True	0.8747	0.9187	0.9152
6	0	False	0	False	0.8843	0.9246	0.9222
7	0	False	1	True	0.8747	0.9214	0.9150
8	0	False	1	False	0.8831	0.9269	0.9210
9	1	True	0	True	0.8831	0.9155	0.9219
10	1	True	0	False	0.8771	0.9191	0.9172
11	1	True	1	True	0.8855	0.9196	0.9231
12	1	True	1	False	0.8819	0.9234	0.9203
13	1	False	0	True	0.8795	0.9162	0.9191
14	1	False	0	False	0.8843	0.9224	0.9221
15	1	False	1	True	0.8940	0.9232	0.9290
16	1	False	1	False	0.8771	0.9177	0.9170
17	2	True	0	True	0.8735	0.9198	0.9143
18	2	True	0	False	0.8867	0.9240	0.9237
19	2	True	1	True	0.8928	0.9224	0.9284
20	2	True	1	False	0.8771	0.9151	0.9175
21	2	False	0	True	0.8795	0.9211	0.9187
22	2	False	0	False	0.8880	0.9267	0.9243
23	2	False	1	True	0.8928	0.9260	0.9280
24	2	False	1	False	0.8892	0.9204	0.9259
25	3	True	0	True	0.8831	0.9158	0.9217
26	3	True	0	False	0.8723	0.9131	0.9141
27	3	True	1	True	0.8867	0.9239	0.9239
28	3	True	1	False	0.8735	0.9171	0.9147
29	3	False	0	True	0.8819	0.9168	0.9209
30	3	False	0	False	0.8783	0.9233	0.9177
31	3	False	1	True	0.8831	0.9194	0.9216
32	3	False	1	False	0.8831	0.9252	0.9210
33	4	True	0	True	0.8783	0.9187	0.9180
34	4	True	0	False	0.8916	0.9231	0.9272
35	4	True	1	True	0.8747	0.9200	0.9152
36	4	True	1	False	0.8867	0.9161	0.9245
37	4	False	0	True	0.8880	0.9216	0.9250
38	4	False	0	False	0.8807	0.9137	0.9204
39	4	False	1	True	0.8855	0.9215	0.9232
40	4	False	1	False	0.8807	0.9287	0.9185