

Project 7: Passenger Satisfaction on SBB

Does the first class influence the satisfaction of the client?

Professor : Christos Dimitrakakis

Students: [Boris Verdecia Echarte](https://moodle.unine.ch/user/view.php?id=8436&course=11241), [Iago Baumann](https://moodle.unine.ch/user/view.php?id=7541&course=11241) and Allizha Theiventhiram

**Contents**

[Introduction 5](#_Toc187413781)

[Motivation 5](#_Toc187413782)

[Objective 5](#_Toc187413783)

[Simulation Model 5](#_Toc187413784)

[Data Categories 5](#_Toc187413785)

[Dependencies Between Variables 7](#_Toc187413786)

[Graphical Model 8](#_Toc187413787)

[Scenarios for Analysis 8](#_Toc187413788)

[Data preparation 9](#_Toc187413789)

[Simulator Modeling 9](#_Toc187413790)

[Data generation and Processing 9](#_Toc187413791)

[Methodology 9](#_Toc187413792)

[Simulation Approach 9](#_Toc187413793)

[Satisfaction Calculation 9](#_Toc187413794)

[Statistical Testing 9](#_Toc187413795)

[Results and Analysis 9](#_Toc187413796)

[Summary of Key Results 9](#_Toc187413797)

[Statistical Results 9](#_Toc187413798)

[Interpretation of Results 9](#_Toc187413799)

[Visualizations 9](#_Toc187413800)

[Discussion and Conclusion 9](#_Toc187413801)

[Answer to the Scientific Question 9](#_Toc187413802)

[Ethical Issues 9](#_Toc187413803)

[Consulted Resources 9](#_Toc187413804)

[References Used 9](#_Toc187413805)

[Appendices 10](#_Toc187413806)

[Code 10](#_Toc187413807)

# Introduction

## Motivation

Understanding customer satisfaction is vital for the success of train service providers. One aspect often debated is whether premium services, such as first-class seating, truly enhance the customer experience.

Our focus is into the potential impact of first-class offerings on passenger satisfaction, taking into account variables such as ticket price, punctuality, overcrowding and frequency of train usage. The goal is to explore the value these services bring to customers and their influence on overall satisfaction.

Due to the challenges of accessing detailed customer satisfaction data, and for privacy concerns, we opted for a simulation approach to explore these relationships.

## Objective

This study aims to determine the extent to which first-class service impacts passenger satisfaction. By simulating customer behavior, the analysis seeks to uncover the key factors that contribute to satisfaction levels and evaluate whether passengers opting for first-class services report a noticeably better experience compared to those using regular seating.

# Simulation Model

All the data used in this work is generated by ourselves. The variables, their weights, and the interdependencies used in our simulation are not derived from real-world data but are based on common sense and general assumptions about customer behavior. This approach was chosen partly because some of the relevant real-world data is proprietary or confidential to railway companies, and partly because we value privacy and aim to avoid gathering unnecessary data about real individuals. While these assumptions create a plausible framework for the simulation, they may reflect biases or preconceptions. Future work could validate these assumptions with access to anonymized real-world data, ensuring greater accuracy and realism.

## Data Categories

The simulated dataset comprises a range of variables designed to capture demographic, behavioral, and travel-related characteristics of train passengers. These variables are categorized as follows:

### Demographic Variables

#### Age

The age of each customer, categorized into discrete groups: 15, 26, 40, 60, and 90 years.

This reflects common age brackets relevant to travel habits and income levels.

#### Gender

A binary variable indicating the customer’s gender, represented as:

* + - 'M': Male
    - 'F': Female.

#### Income

The annual income of the customer, expressed in discrete ranges:

* + - Very Low: <30,000
    - Low: 30,000–60,000
    - Medium: 60,000–85,000
    - High: 85,000–100,000
    - Very High: >100,000.

Income is influenced by age and gender in the simulation.

### Behavioral Variables

#### Remote Working Days

The number of days per week the customer works remotely, ranging from 0 to 5.

This variable impacts travel frequency and overall satisfaction.

#### Has Car

A binary variable indicating car ownership:

* + - 'yes': The customer owns a car.
    - 'no': The customer does not own a car.

Car ownership reduces train usage frequency in the simulation.

#### Frequency

How often the customer uses trains, rated on a 5-point scale:

* + - 1: Very low usage.
    - 5: Very frequent usage.

Frequency is influenced by remote working days and car ownership.

### Travel Variables

#### Price

The ticket price paid by the customer, rated on a 5-point scale:

* + - 1: Very low price.
    - 5: Very high price.

Price is influenced by income, journey duration, and whether the customer travels first class.

#### Punctuality

A measure of the train’s timeliness, rated on a 5-point scale:

* + - 1: Almost never punctual.
    - 5: Always punctual.

Punctuality impacts overcrowding perceptions.

#### Duration

The length of the customer’s journey, rated on a 5-point scale:

* + - 1: Very short duration.
    - 5: Very long duration.

#### Overcrowding

The customer’s perception of crowd levels, rated on a 5-point scale:

* + - 1: Very low overcrowding.
    - 5: Very high overcrowding.

Overcrowding is affected by punctuality and the customer’s travel class.

#### First-Class

A binary variable (only present in specific simulation scenarios) indicating whether the customer traveled first class:

* + - 1: First-class travel.
    - 0: Non-first-class travel.

First-class travel reduces perceived overcrowding and affects ticket price.

### Target Variable

#### Satisfaction

A binary variable representing the customer’s satisfaction:

* + - 1: Satisfied.
    - 0: Unsatisfied.

Satisfaction depends on multiple factors, including ticket price, punctuality, overcrowding, and first-class travel, with varying weights depending on the simulation scenario.

## Dependencies Between Variables

The simulation model incorporates a variety of interdependencies between variables to reflect realistic relationships observed in train passenger satisfaction. Below is a detailed explanation of these dependencies:

### Income and Price

The price a customer pays for their ticket depends on their income level.

Customers with higher incomes are more likely to purchase higher-priced tickets and opt for first-class travel. For instance, customers earning above 100,000 units are highly likely to select first-class travel and pay higher ticket prices. Conversely, lower-income groups tend to select lower-priced tickets.

### Price, Duration, and Satisfaction

Longer journey durations often result in slightly higher ticket prices, especially when the customer opts for lower-priced categories. This reflects a dependency where prolonged travel times are factored into pricing strategies.

Satisfaction is indirectly influenced by this relationship, as higher ticket prices can increase expectations, impacting satisfaction outcomes.

### First-Class Travel and Overcrowding

First-class travel significantly impacts perceived overcrowding levels. Customers traveling in first-class experience reduced overcrowding (assigned a value of 1 in the overcrowding variable).

Overcrowding is modeled to reflect higher levels in non-first-class travel, especially during peak times or when punctuality is low.

### Punctuality and Overcrowding

Punctuality directly influences the perception of overcrowding. When punctuality is rated low (e.g., less than 3 on a 5-point scale), customers are more likely to perceive higher overcrowding levels.

### Frequency of Travel

Customers with fewer remote working days are more frequent train users. However, owning a car reduces the frequency of train usage, even among customers with fewer remote working days.

This relationship demonstrates how lifestyle factors affect travel behavior.

### Satisfaction and Weighted Variable Impacts

Satisfaction is determined using multiple variables, such as price, punctuality, duration, frequency, and overcrowding. In the complex satisfaction model:

* Ticket price exerts a negative influence on satisfaction, with lower prices generally correlating to higher satisfaction.
* Punctuality and reduced overcrowding have positive impacts on satisfaction.
* Longer durations and higher travel frequencies tend to diminish satisfaction levels.

The combined effect of these variables, weighted appropriately, determines the satisfaction score.

### Age and Income

Income levels are adjusted based on age groups:

* Customers below 20 years of age are primarily in lower income brackets.
* Older customers (above 35) are more likely to be in higher income groups, with adjustments made to account for professional growth over time.

Gender introduces additional variability, with female customers experiencing a slight reduction in income to simulate potential gender-based pay disparities observed in real-world data.

### Interplay Between Remote Work and Travel Behavior

Remote working days influence both train frequency and perceptions of overcrowding. Customers with more remote workdays tend to travel less frequently, reducing the likelihood of experiencing overcrowding.

### Income and First-Class

The dependency between Income and First-Class indicates that higher income levels increase the likelihood of opting for first-class journeys.

These interdependencies are central to the simulation’s ability to model realistic customer satisfaction scenarios, allowing us to explore the nuanced impacts of factors like first-class travel, price, and service quality on overall satisfaction.

## Graphical Model

The graphical models presented below illustrate the relationships between various factors influencing satisfaction. These include direct factors, such as price, punctuality, and overcrowding, as well as indirect factors, such as age, income, and remote working days.

In the first model, the relationships are shown without incorporating the First-Class variable. This representation focuses solely on the interplay of factors affecting satisfaction without considering the additional dimension of service class differentiation.

In the second model, the First-Class variable is introduced. This addition highlights its potential influence on satisfaction and its interaction with other factors, such as price and overcrowding, but also the influence of income on that variable.

|  |  |
| --- | --- |
|  |  |
| *Fig. 1: Dependencies without first-class variable* | *Fig. 2: Dependencies with first-class variable* |

## Scenarios for Analysis

### Scenario 1: Baseline Scenario

This scenario excludes the first-class variable and focuses on standard travel factors such as price, punctuality, overcrowding, and journey duration. The aim is to establish a baseline understanding of how these factors independently influence satisfaction.

### Scenario 2: First-Class Inclusion

In this scenario, the first-class variable is introduced. This allows us to assess its impact on satisfaction by comparing first-class and non-first-class travelers under identical conditions. Key aspects include changes in overcrowding perceptions and ticket price sensitivity.

### OTHER SCENARIOS?

# Data preparation

## The data preparation process involved generating synthetic datasets, preprocessing them, and structuring them for analysis. This section outlines the mechanics of simulator modeling, data generation, and processing steps.

## Simulator Modeling

The study employed two simulators:

* ComplexDependentSatisfaction: Focuses on satisfaction without first-class travel, modeling dependencies like price, punctuality, and overcrowding.
* ImpactOnOvercrowding: Includes first-class travel and models its specific impact on perceived overcrowding.

These simulators generate realistic datasets by incorporating dependencies between variables, ensuring alignment with the research objectives.

## Data generation and Processing

### Synthetic Data Generation

### Datasets of varying sizes (10 to 50,000 samples) were generated to simulate diverse customer behaviors and allow robust testing. Each dataset included key demographic, behavioral, and travel-related variables.

### Preprocessing Pipelines

The numerical features are scaled with MinMaxScaler to normalize feature ranges. These include variables like income, price, and punctuality. The categorical features are instead encoded using OneHotEncoder. The features varied by simulator:

* ComplexDependentSatisfaction: Included "Has Car" and "Gender."
* ImpactOnOvercrowding: Added "First-Class" to capture its unique effects.

Pipelines combined numerical and categorical preprocessing to standardize data preparation for machine learning models.

# Methodology

The methodology outlines the approach to simulation, satisfaction calculation, and statistical testing, explaining the rationale for choices in model design and analysis.

## Simulation Approach

The choice of two simulators was driven by the study’s focus:

1. ComplexDependentSatisfaction: Provides a baseline for understanding core satisfaction dynamics without first-class travel.
2. ImpactOnOvercrowding: Introduces first-class travel to assess its impact, particularly on overcrowding and satisfaction.

This dual-simulator approach balances complexity and interpretability, ensuring that the study addresses the research question comprehensively.

## Satisfaction Calculation

Satisfaction was modeled as a binary variable (1 for satisfied, 0 for unsatisfied), influenced by multiple factors:

* Direct Factors: Ticket price, punctuality, and overcrowding.
* Indirect Factors: Age, income, remote working days, and car ownership.

The variables were weighted to reflect their real-world importance, for example with first-class travel reducing overcrowding and influencing satisfaction in the second simulator.

## Feature Selection Methodology

To identify the most important variables contributing to satisfaction, three feature selection methods were employed:

### Filter Method (Correlation Analysis):

* Measures the correlation between each feature and the target variable (satisfaction).
* Features with high absolute correlations were considered significant.

### Recursive Feature Elimination (RFE):

* Iteratively removes the least important features based on model performance, ultimately retaining the top-ranked features.
* A Logistic Regression model was used as the estimator to rank features.

### Random Forest Feature Importance:

* Uses a Random Forest classifier to assign importance scores to features based on their contribution to reducing prediction error.

Feature selection was applied to datasets of varying sizes to observe how feature importance stabilizes as the dataset size increases. The selected features were later used to train predictive models, and their performance was compared against models trained on all features.

## Statistical Testing

The study evaluated the scalability and performance of machine learning models using datasets of increasing sizes. Four classifiers were chosen for their complementary strengths:

* Random Forest: For capturing non-linear relationships.
* Logistic Regression: As a baseline for linear separability.
* Single-Layer Perceptron: To test the dataset’s separability with a simple neural network.
* K-Nearest Neighbors: For capturing local relationships.

Model performance was assessed using accuracy scores, with preprocessing integrated into pipelines to ensure consistency and reproducibility.

# Results and Analysis

## Summary of Key Results

## Statistical Results

## Interpretation of Results

## Visualizations

# Discussion and Conclusion

## Answer to the Scientific Question

FUTURE WORK COULD TRY OUR METHODS WITH REAL-WORLD DATA

## Ethical Issues

Good because of privacy.

Questionable because synthetic data, maybe reinforcing false assumptions and biases.

# Consulted Resources

## References Used

## Use of other tools

We used ChatGPT from OpenAI, to help us rephrasing some sentences, as English is neither of us first language.

# Appendices

## Code