University of St. Gallen

Group Project

Hardbrücke – Zürich

Joachim Nadir Lalou

Michael Moser

Kaito Pross

Benedikt Lukas Thoma

Project Report

Autumn Semester 2024

Certificate Programme Data Science Fundamentals

Workshop Fundamentals of Data Science

3,580

Prof. Ph.D. Lyudmila Grigoryeva, Ph.D. Jonathan Chassot, M.Sc. Hannah Busshoff

03.12.2024

Table of contents

1. Introduction

This report presents the methods and main results of our DSF group project. Using data from the Hardbrücke station in the city of Zürich about the number of passengers arriving at the station's bus stops every 5 minutes, we set out to answer the following question: "What will be the predicted/forecasted number of passengers on the next day?". To do so, we successively develop a time-series forecast and a random forest.

The results provided by this project can be very useful for the development of a dynamic pricing policy. We consider a business case, where the public transport operator of the city of Zürich (VBZ) considers imitating the model of the Swiss Federal Railways and adopting a pricing system depending on demand for transportation to smooth passengers' inflow during peak hours.

2. Data Processing

Data preprocessing is one of the most important steps to ensure that a model can work efficiently and even more importantly, with the correct data. The analysis of passenger frequencies at Hardbdrücke relies on three datasets. The most important one is the Hardbrücke dataset which contains passenger in- and outflows for the west and east stop at Hardbrücke in Zürich. It additionally contains data of how many people passed certain crosspoints at the train station. To be able to feed the machine learning algorithm with independent variables, another dataset of past weather observations in Zürich was introduced. The third and last one necessary contains data on the workdays in Zürich, which is crucial information to estimate the number of passengers on a given day. It was created by ourselves by looking up information on the local government's website.

The first step for all datasets is to import them and get rid of the variables which are not needed. For the Hardbrücke dataset this was the information about the different crosspoints and the outflow frequencies, as we are solely interested in passenger inflows. Then all variables are being transformed to the correct data type to ensure proper handling later on. The only issue was that both, the weather and passenger data time variables changed from winter to summertime and the Hardbrücke dataset additionally was timezone sensitive. The issue was resolved by converting both to the same timezone and afterwards dealing with the missing data from the time shift.

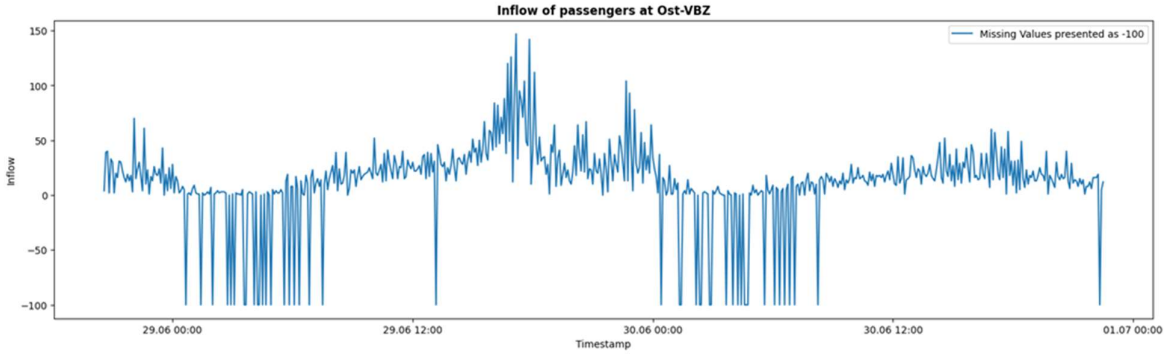

Figure 1: Missing values for Ost bus stop (labeled as negative numbers)

Fortunately, as can be seen in Figure 1 the missing data seems to adhere to some kind of repeating structure. By plotting the data over a narrow timespan, it can be seen that the data is mostly missing throughout the night and with many observed values in between the missing points. This allowed for the use of interpolation to deal with missing data. This technique takes the value of the last- and next healthy data point in the time series to construct a mean and fill in the missing values with this. The weather data's missing values were handled in the same exact manner. If there would have been missing data for perhaps a whole day, this method would have distorted the results as the last and next healthy points would have been far apart. However, with this not being, interpolation is a valid approach.
To be able to keep a time structure upright in the random forest model, additional variables were created in the X matrix. 288 lagged observations of the "In east/west" variable to be precise, so that the current prediction can be based on the data of the whole past day. A dummy variable was also created to identify holidays (weekends included) in order to ensure that models would be able to interpret it correctly.

3. AR Model

Our plan for the AR model was to create a simple time-series model which relies only on past data. First, it is meant to be a benchmark for a more sophisticated model afterwards, namely the random forest. We decided to build the same model for the eastern and western sides of the Hardbrücke. Our methods were applied for each side. This way, we can make sure that the observations of one side do not belong to the forecast of the other. We checked our dataset for stationarity using the augmented dickey fuller method as learnt in the lectures. Both sides returned a p-value of 0. This runs counter to our intuition, because the plot of our decomposed dataset showed a seasonal pattern. However, we decided to follow the lecture slides, which stated that a dataset is stationary if the p-value of the test is below 0,05. Therefore, we concluded that our data has a stationary behaviour. Had our dataset been stationary, we would have continued to develop our people by differencing the non-stationary data. Next. we calculated and plotted the Autocorrelation as well as the Partial Autocorrelation Function. The first showed a sinus function, the second converged to zero. Based on that, we chose to build an AR Model, because the ACF looked like a sinus function and the PACF converges towards zero. Importantly, we decide to plot only the eastern side of the bridge to save space (see Figure 2 & 3).

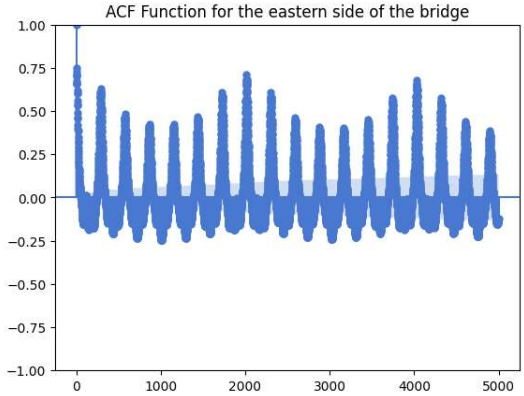

Figure 2: ACF of Ost bus stop

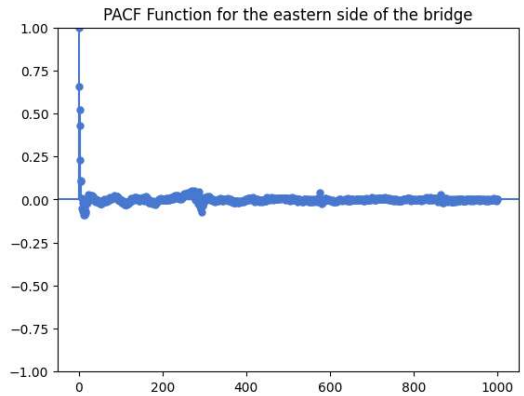

Figure 3: PACF of Ost bus stop

Subsequently, we created a training and testing set with approximately 80.58% of the data for the training and with the rest for the testing set. This way, we obtained 245 days for the training set, and 58 for the testing set.. To avoid data leakage, we made sure to split after sorting the data by time. We chose a value of 288 for our lags, which corresponds to the number of 5-minutes intervals in one day. We also decide to cross-validate our data using 4 splits with an expanding window, meaning that our training set is always equal to 20% of our dataset. Importantly, the percentage of the training and testing sets in the forecast and cross-validation diverge. We assumed that it might be an adequate choice because it allows us to to predict future values using complete days and weeks.

4. Random Forest

Taking the performance of the Autoregressive model as a baseline, we tried to achieve a more reliable estimate by developing two Random Forest models, one for each bus stops at the Zürich Hardbrücke stations. The motivation behind the development of these models was to outperform the baseline autoregressive models by integrating two "dimensions" of inputs:

- **Temporal component/dimension**
  Each one of the Random Forest models takes the inflow of passengers at the correspondent bus station over the last 24 hours with a frequency of five minutes as inputs, so as to let the models learn about daily developments of passengers' inflow. Additionally, a dummy variable indicating whether it is a holiday (weekends included) or not was included.
- **Meteorological component/dimension**
  Each one of the Random Forest models takes weather conditions as inputs as well, which are described by features recording air temperature, precipitation level, global radiation level and several other indicators at the time each observation was recorded.

Training the models on all these features resulted in Mean Squared Error (MSE) on the Test set of ~172.3 for the Ost bus stop' s model and of ~59.9 for the West bus stop' s model. To counteract possible overfitting in both models, meteorological features that showed low importance levels were removed in an attempt to decrease noise and reduced versions of the models were retrained. Only global radiation was kept as meteorological feature, as other variables did not show a sufficiently high influence so one could justify keeping them. Dropping these features resulted in no alteration of models' performance as measured by MSE.

The relatively high importance of global radiation is realistic, as people tend to go outside more often if the sun is shining. On the flipside, it is surprising that precipitation has such a low importance, but this could be due to the fact that people may have no real alternative to public transportation when it comes to commuting. All temporal features (t- features) were left untouched to upkeep the seasonality and equidistance of the lagged values.

At this point, the developed models could only predict passengers' inflow for the next time period. Since the models were trained on data with a 5 minutes frequency and take past values of passengers' inflow for the previous 288 periods as inputs (corresponding to the last 24 hours), both models could only produce a forecast for the next 5 minutes, which might be considered too short of a timespan for the prediction to be useful in any way. In order to have models able to predict values for periods beyond that one immediately following that of the given inputs, for loops that iterates through the models have been put in place, so that a prediction for the next period can be computed at each iteration, that is then used as input for the computation of the following period's prediction. At each iteration, the models also verify whether the next prediction is within a day that can be considered a Holiday or not (by using the dummy variable Holiday). Assuming we don't have access to input values for weather conditions in the forecasting period, an autoregressive model for this feature was developed to produce a forecast of global radiation over the same forecasting horizon, thereby providing the models with the necessary inputs to compute their prediction through the iterative procedure.

5. Results

The AR Model provided a baseline forecast and registered an overall poor performance over the test sets. The Mean Squared Error (MSE) for the Ost bus stop was 657.69 and for the West bus stop about 199.57. Down at a daily level, the forecast gets more precise. Forecasting the passengers' inflow on the first day of the test set resulted in a MSE of 236.4 for the Ost bus stop and a MSE of 100.22 for the West bus stop.

We forecasted the next day after the end of the Test sets (16th of November) and plotted it, but the estimation seemed to be smaller than the previous day. After plotting the next few days, we realized that the first forecast day was a Saturday and that the flow of passengers for a weekend is lower than for a weekday, which seems plausible (see Figure 4).

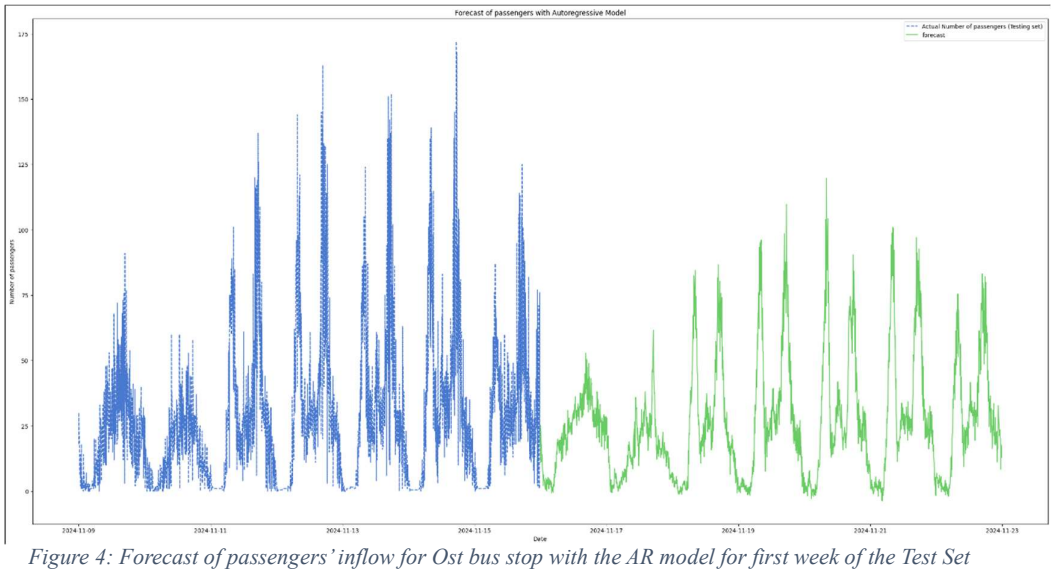

Figure 4: Forecast of passengers' inflow for Ost bus stop with the AR model for first week of the Test Set

As for the random forest, the final models for the two bus stops manage to provide a fairly reliable prediction for the desired forecasting horizon (one day). Unsurprisingly, the longer the forecasting horizon, the more inaccurate the predictions become, with the loss increasing the more iteration the models have to go through.

Producing a forecast for the first 288 periods of the test sets, i.e for the day following the last day of the training sets, results in a MSE of approximately 204.43 for the Ost bus stop and a MSE of 81.03 for the West bus stop. The following plot (see Figure 5) shows both the forecasted passengers' inflow (green) and the actual inflow (blue) throughout the first day of the test set, the 19th of September 2024.

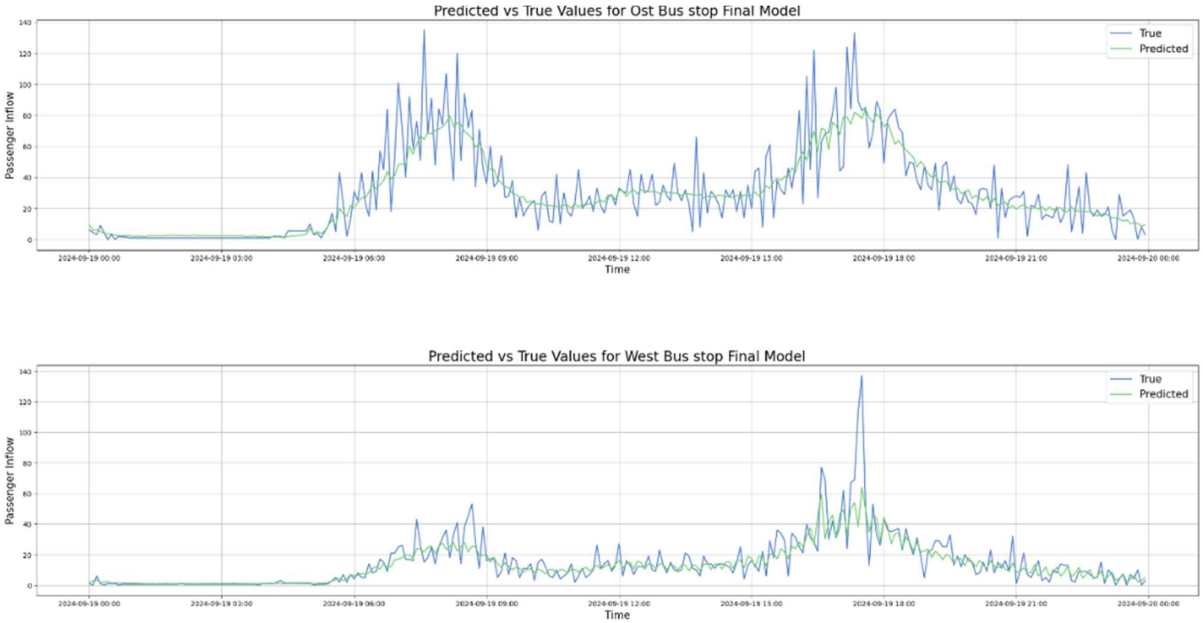

Figure 5: Predicted vs true values of Test sets' first day for Random Forest by iterative procedure

The following plots (Figure 6), on the other hand, show the forecast for the day following the last day of the test sets, i.e. for the 16th of November 2024. The models seem to correctly predict a lower and smoother inflow of passengers than that of the days preceding it, as the former is a Saturday, which always tends to have smoother and lower passengers' inflow than weekdays across the data.

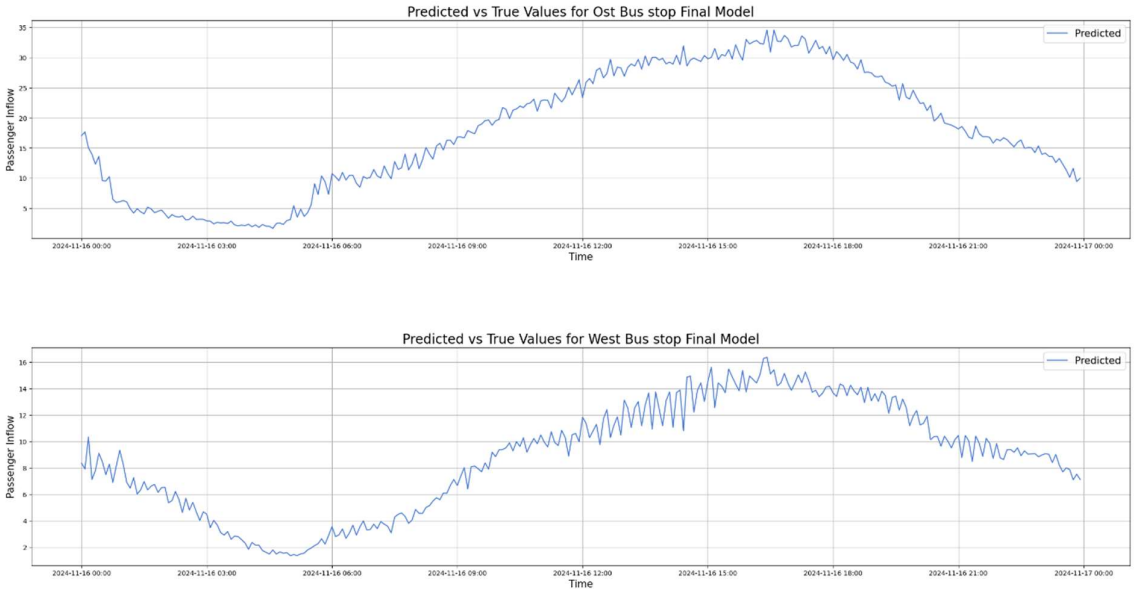

Figure 6: Forecast for first day after the end of the Test set for Random Forest by iterative procedure

Comparing the random forest and the AR models, we see that the former perform better, as shown by the lower values of the MSE for both bus stops Ost and West. In addition, the western side seems to have a lower MSE for every model. Presumably, this difference may be due to the fact that the bus stop continuously register a lower and smoother inflow of passengers than the eastern counterpart. Bus lines serving this bus stop go towards the west side of Zürich, away from the city center, and it thus makes sense that passengers inflow is lower. On the other hand, the Ost bus stop serves bus lines with direction Schiffbau, going towards the city center of Zürich, which draws a higher amount of people.

Overall, the Random Forest model used to predict passengers' inflow for the West bus stop seems to be the most reliable one, although for the implementation of a dynamic pricing strategy, the model for the Ost bus stop is the most impactful one, as demand for transportation in that direction is much stronger and less smooth overall.