BUSINESS UNDERSTANDING

Business Problem

The current standby-duty plan for Berlin's red-cross rescue service struggles with inefficiencies, leading to situations where there are not enough standby drivers or too many standby drivers are kept on hold. The HR planning department wants to improve the current planning logic by incorporating predictive models to estimate the daily number of standby rescue drivers more accurately.

Project Objectives

This project aims to develop a predictive model to estimate the daily number of standby rescue drivers required, improve the efficiency of the standby-duty plan by increasing the percentage of standby drivers being activated compared to the current approach of keeping 90 drivers on hold, reduce situations where there are not enough standby drivers or excess standby drivers, ensure that the new planning model can incorporate seasonal patterns and other factors affecting the number of required standby drivers, and meet the deadline for finishing the duty plan on the 15th of the current month for the upcoming month.

Success Criteria

The success criteria of this project include increasing the percentage of standby drivers being activated compared to the current approach, reducing the number of situations where there are not enough standby drivers or excess standby drivers, developing a predictive model with a reasonable level of accuracy in estimating the daily number of standby drivers required, and meeting the deadline for finishing the duty plan on the 15th of the current month for the upcoming month consistently.

Project Plan

The execution of this project followed the following steps:

1. Understand the current standby-duty plan and its shortcomings.
2. Understand and analyze the data provided.
3. Develop and evaluate predictive models to estimate the daily number of standby drivers required.
4. Implement the selected predictive model into the planning process.
5. Monitor the performance of the new planning model and make necessary adjustments.

Resources:

A Windows laptop with RAM and processor of 8.00 GB and RAM Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz 2.20 GHz, respectively, was used to run Python and its libraries and frameworks (as shown in the appendix) that were used to process the data and train the machine learning models. After training, the selected model was deployed as a web app using Flask. In addition, Brisqi was used as the project management tool.

DATA UNDERSTANDING

The available data was stored in a csv file. As shown below, pandas was used to load and understand the data. There were eight columns. None of the columns had any nun values nor outliers. One of the columns has object-type data stored in it, while the remaining columns store numerical data. The following description defines the nature of each column in the dataset.

• date: entry date

• n\_sick: number of drivers called sick on duty

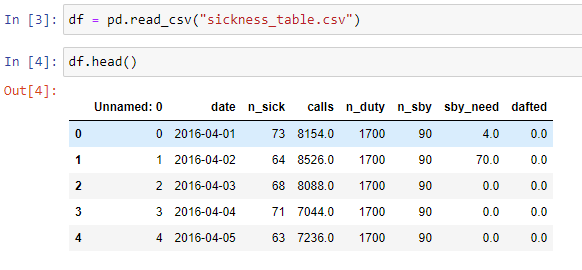
• calls: number of emergency calls

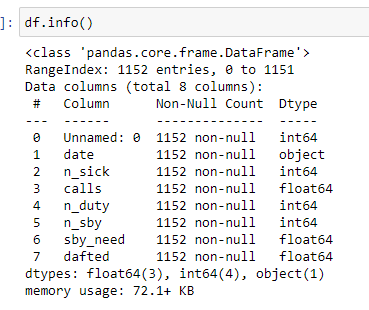
• n\_duty: number of drivers on duty available

• n\_sby: number of standby resources available

• sby\_need: number of standbys, which are activated on a given day

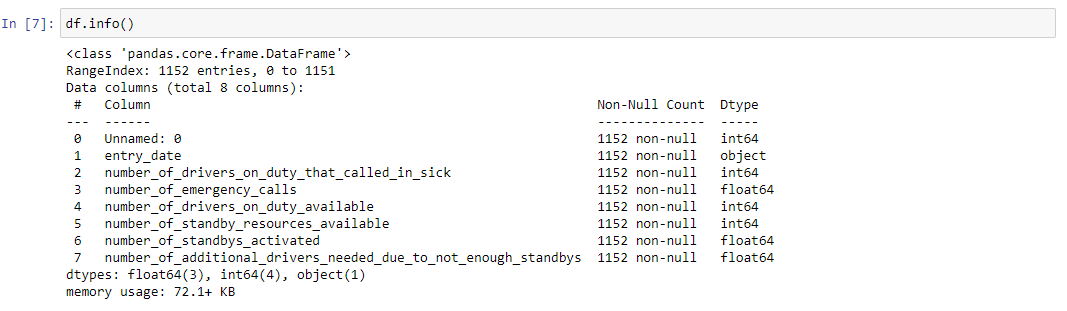
• dafted: number of additional drivers needed due to not enough standbys



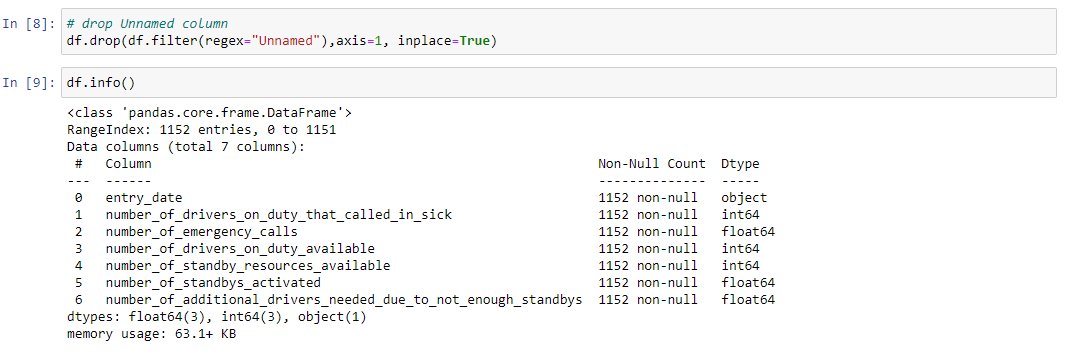


To improve readability and facilitate analysis, the remaining columns were renamed based on the description/definition given above; the new names are given below.





The Unnamed column is simply an index column and is not relevant for modeling. Therefore, it was removed from the data.



The entry\_date column stores object data. However, the information stored in it is date data. Therefore, the format of this column was changed from object to datetime.

A screenshot of a computer

Description automatically generated

DATA PREPARATION

The HR department claims a seasonal pattern; thus, we wanted to understand this seasonal pattern.

Figure XXX – Seasonal Pattern of Activated Standby Drivers

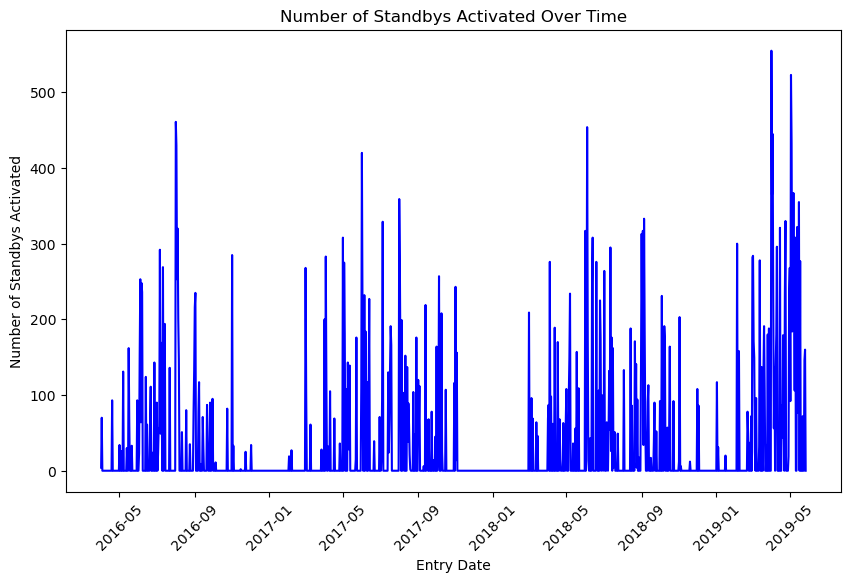
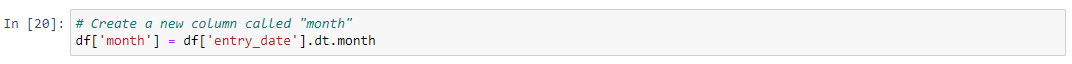


Figure XXX indicates a seasonal pattern in the number of activated standby drivers. The peak season is between May and September/October, while the first and last quarters register the lowest numbers of activated standby drivers. It means that month has an impact on the number of activated standby drivers. Therefore, we extracted the month from entry\_date and stored it in a new column called month.



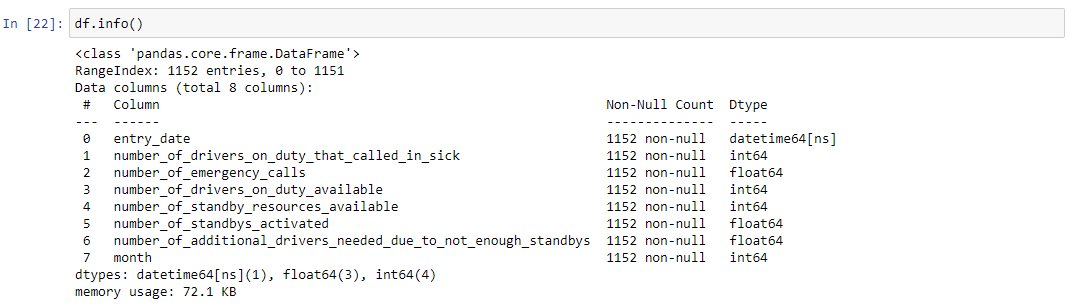


Figure XXX – Standby Drivers by Month

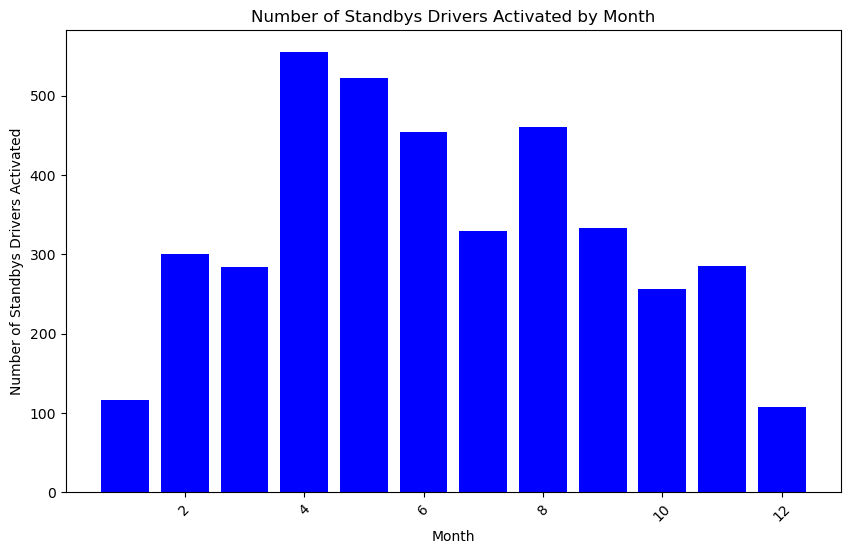


Figure XXX shows that indeed the number of activated standby drivers varies with month. The end and beginning of the year have the lowest numbers of activated standby drivers, while April and May have the highest numbers. In addition, for modeling purposes, it is best to use month instead of entry\_date because in entry\_date, each row has a unique value; this would not allow the model to generalize. On the other hand, in month there are only 12 readings and present enough repetition that could allow the model to learn something.

In the following images, we present the relationship between the number of activated standby drivers and the remaining features of the dataset.

Figure XXX – Relationship Between Activated Standby Drivers and Drivers on Duty that Called in Sick

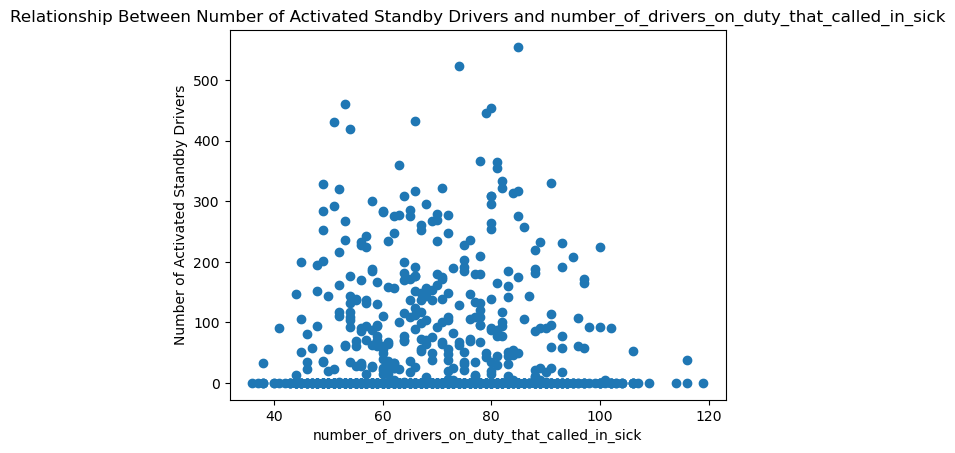


Figure XXX – Relationship Between Activated Standby Drivers and Drivers on Duty that Called in Sick

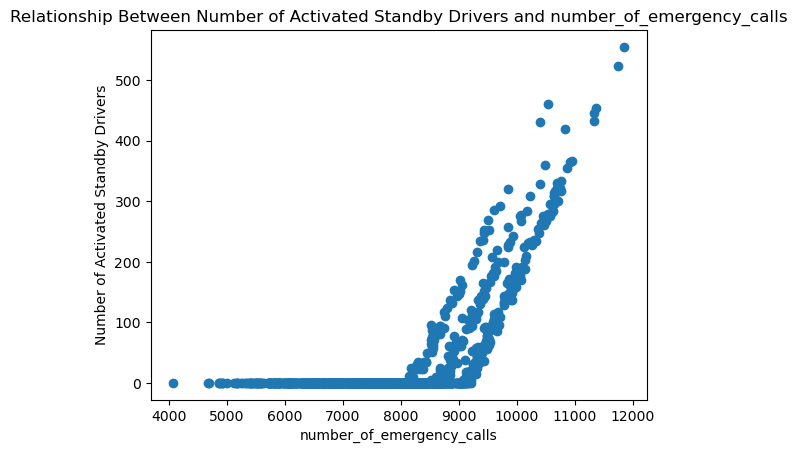


Figure XXX – Relationship Between Activated Standby Drivers and Drivers on Duty that Called in Sick

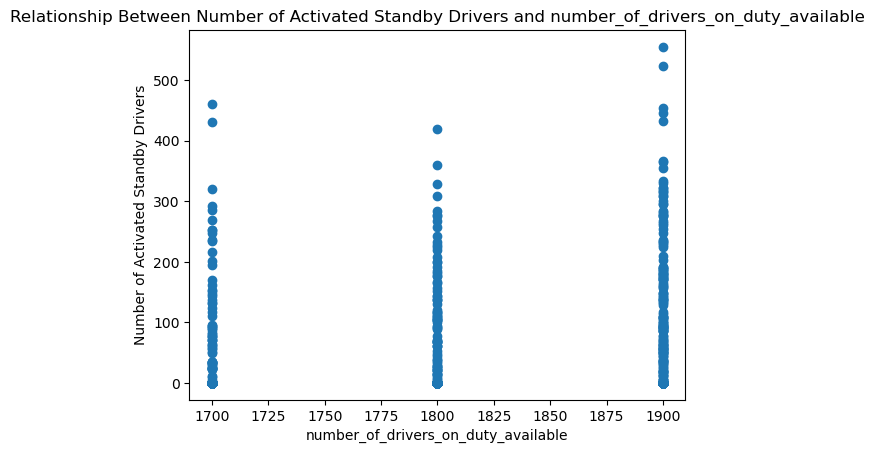


Figure XXX – Relationship Between Activated Standby Drivers and Drivers on Duty that Called in Sick

A graph of a number of activated standardby drivers and number of resources available

Description automatically generated

Figure XXX shows that the number of standby resources available is always 90. Thus, for modeling purposes, this column was not used in the training of the machine learning models.

Figure XXX – Relationship Between Activated Standby Drivers and Drivers on Duty that Called in

A graph showing a number of drivers

Description automatically generated with medium confidence

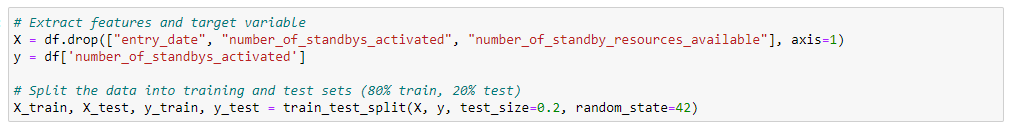
Figure XXX – Relationship Between Activated Standby Drivers and Month

A graph of blue dots

Description automatically generated

Training and Test Data

The data was split into the target variable and independent variables. All the variables, except entry\_date, number\_of\_standby\_resources\_available, and the number\_of\_standby\_drivers\_activated, were selected as independent variables, while the number\_of\_standby\_drivers\_activated was selected as the target feature. Furthermore, the data was also split into training and test sets, with 80% of the data being used for training and the remaining being used for testing.



MODELING

The target or dependent variable is the number of standbys activated, which is a numerical continuous variable. Thus, the appropriate machine learning type to be used is regression. There are various regression algorithms. In this project, we trained various regression algorithms, evaluated all of them, and selected the best-performing algorithm.

Fkvm;fm;am

Zkkmv;m;vfm

Zm;sm;dm