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

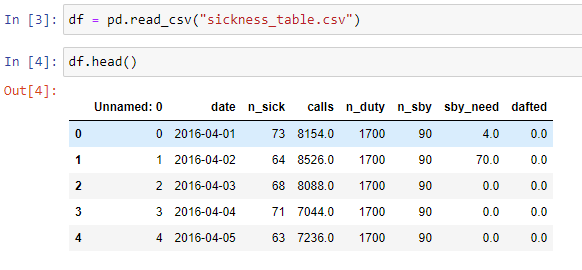
• calls: number of emergency calls

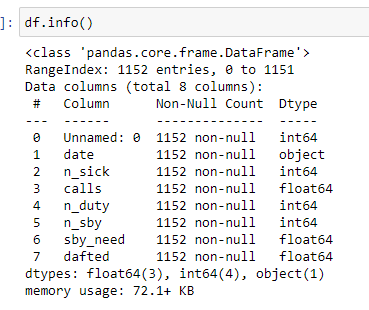
• n\_duty: number of drivers on duty available

• n\_sby: number of standby resources available

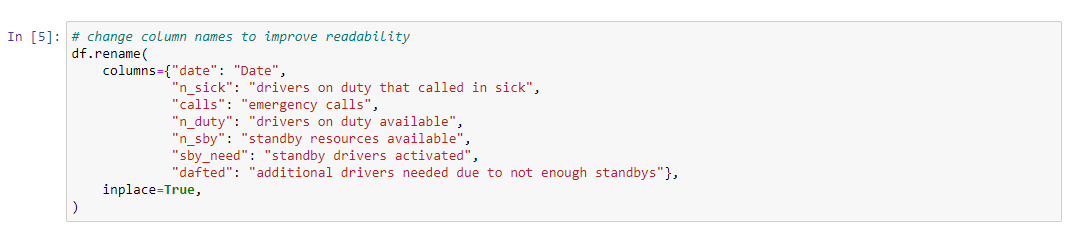
• sby\_need: number of standbys, which are activated on a given day

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





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



A screenshot of a computer

Description automatically generated

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

The 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 close-up of a computer screen

Description automatically generated

DATA PREPARATION

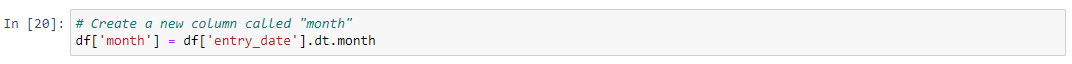
The HR department claims a seasonal pattern in the data; thus, we wanted to understand this seasonal pattern.

Figure XXX – Activated Standby Drivers

A graph of blue lines

Description automatically generated

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. The data also suggest that there are days in which no standby driver is activated. To further understand the data, Figure XXX plots all features over time.



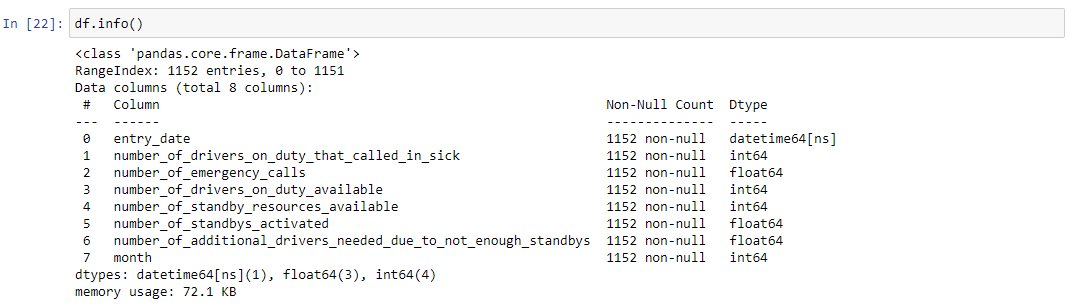


Figure XXX – Standby Drivers by Month

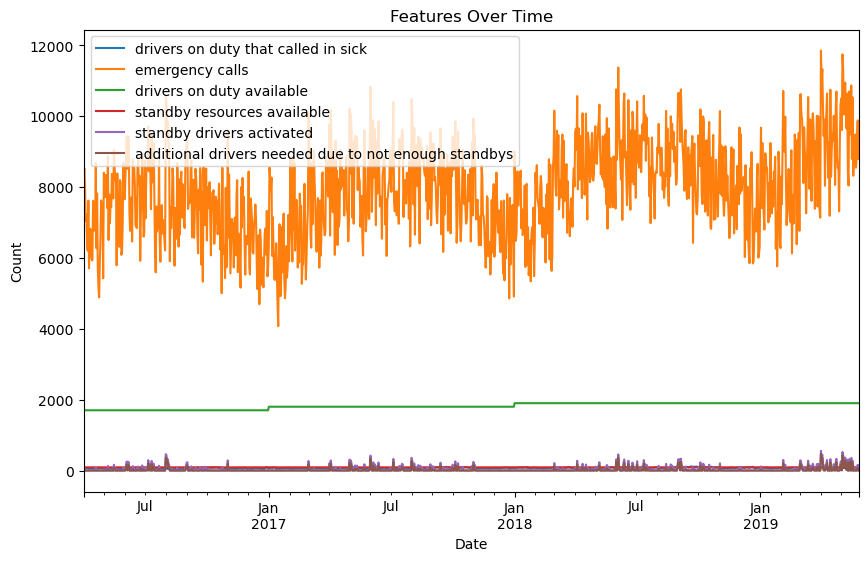


Figure XXX shows a clear seasonal pattern on most variables. In the following images, we present the relationship between the number of activated standby drivers and the remaining features of the dataset to help us further understand the data.

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

A graph of blue dots

Description automatically generated

Figure XXX – Relationship Between Activated Standby Drivers and Emergency Calls

A graph of a number of numbers

Description automatically generated

Figure XXX – Relationship between activated standby drivers and on-duty drivers that are available

A graph of drivers on duty

Description automatically generated

Figure XXX – Relationship Between Activated Standby Drivers and Standby Drivers Available

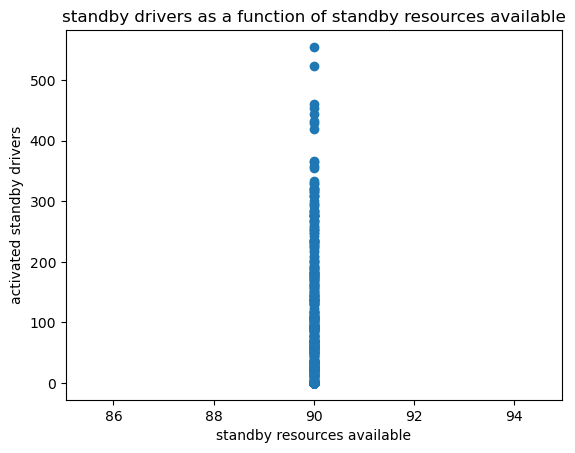


Figure XXX shows that the number of standby resources available is always 90. Thus, this column should not be used in the training of the machine learning models.

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

A graph with numbers and lines

Description automatically generated

Figure XXX – Relationship Between Activated Standby Drivers and Month

The following figures show the seasonal pattern of each feature.

Figure XXXX – On-duty sick drivers over time

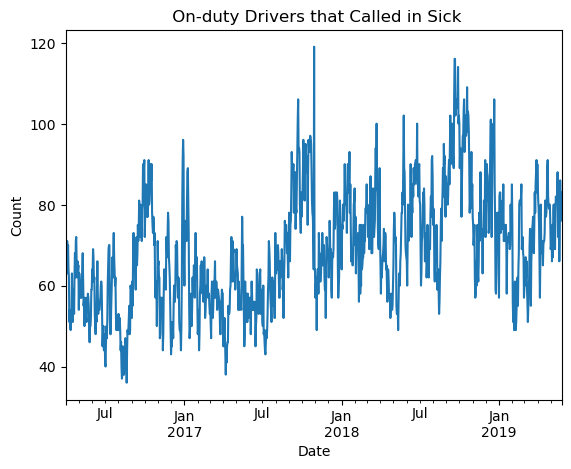


Figure XXX – Emergency Calls Over Time

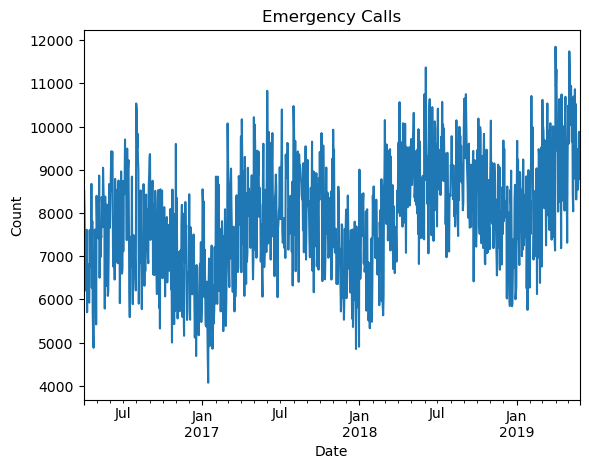


Figure XXXX – Available Drivers

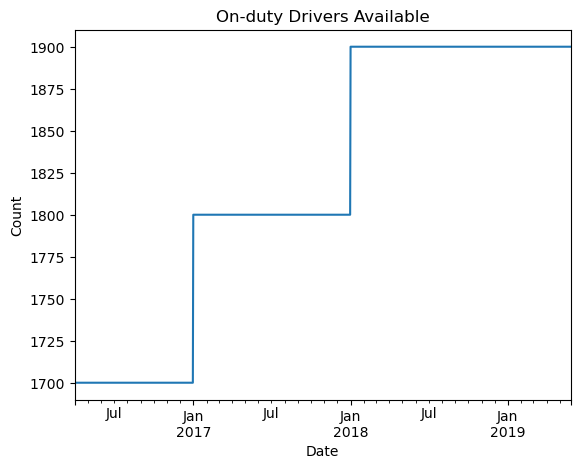


Figure XXXX – Standby Drivers Available

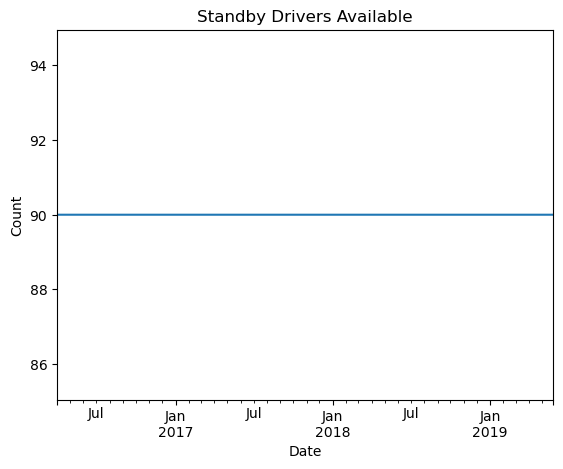
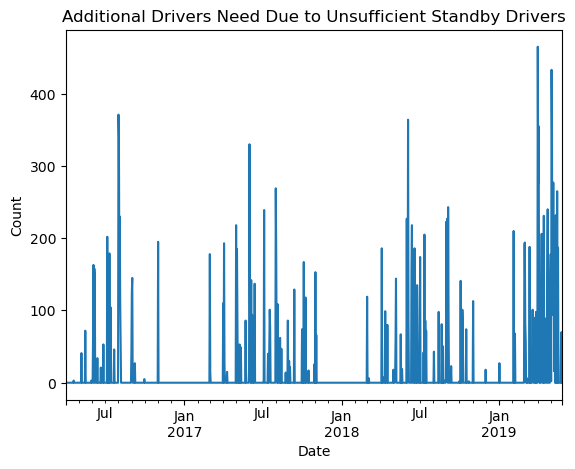


Figure XXXX – Additional Drivers Needed Due to Insufficient Standby Drivers

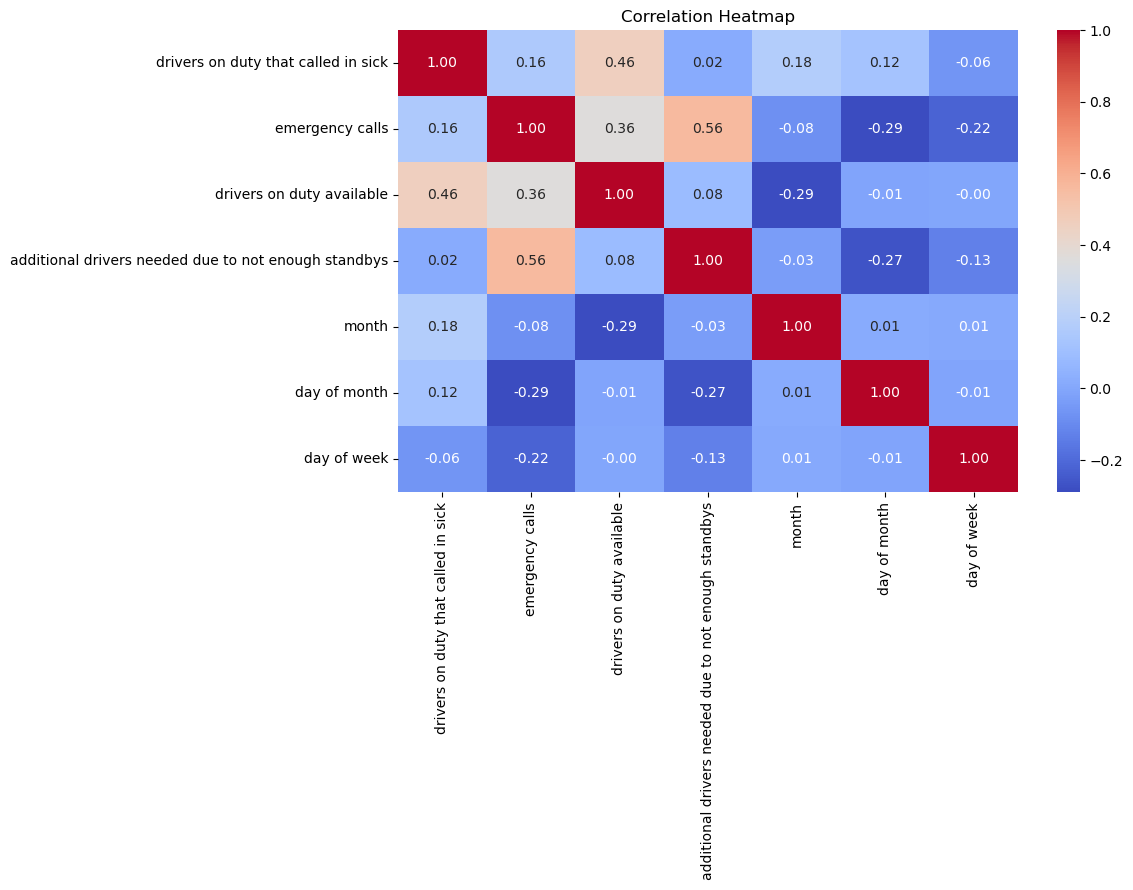


The fact that, regardless of the number of on-duty drivers calling in sick, the number of on-duty drivers available is almost always constant raises concern about the quality and veracity of the data. However, we will assume that when a number of on-duty drivers call in sick, the same number of standby drivers are put on duty to replace those who called in sick, and the same number of non-standby drivers are put on standby to replace those that are now on duty.

A date is a data point that does not repeat itself; it only appears once in the dataset. Because we are interested in building a model that can grasp the seasonal aspect of the data, three more variables were created from the date column; they are month, day of month, and day of week. The values of these variables can appear more than once and can be extracted from any given month. Thus, given the date whose activated standby drivers one wants to predict, the values of these columns are extracted and fed into the predicting system.

Although the data plots shown above show the nature of each variable, it is not possible to have a clear understanding of the relationship between each of them. Thus, a heat map expressing the correlation between the features is shown below.

Figure XXX – Correlation Heatmap

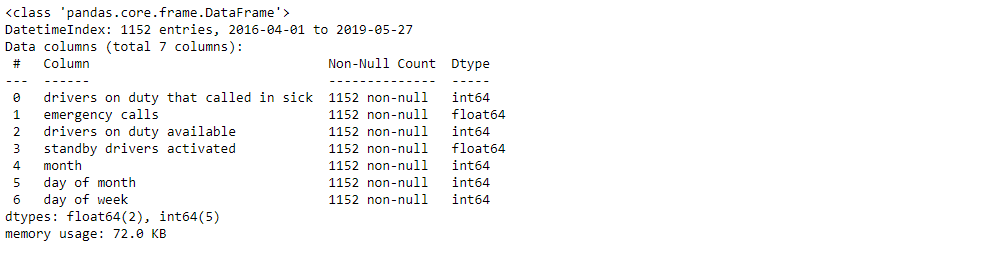


There is a moderate correlation between the number of emergency calls and the number of additional drivers needed due to not enough standby drivers. This correlation is expected because the higher the volume of calls, the higher the number of drivers activated and the higher the number of additional drivers needed due to not enough standby drivers. Considering that the variable standby resources available is always 90, the higher the number of calls, the higher the number of additional standby drivers are needed (once the threshold of calls that the on-duty and the 90 standby drives can be bear is reached). After the threshold of 90, these two variables are expected to be directly proportional. Nevertheless, we want to predict the number of standby drivers activated so that we can set the predicted number of drivers on standby and ideally keep the additional drivers required to zero.

In addition, this variable (additional drivers needed due to not enough standby drivers) only exists because the number of standby drivers available is always 90. If we set the number of standby drivers available based on the prediction for activated standby drivers, then we will be able to keep additional drivers to zero (or close to zero). Therefore, for modeling purposes, one of the columns had to be removed because keeping both variables would add unnecessary complexity. Because the additional drivers depend on the number of calls received, the variable additional drivers needed due to not enough standby drivers was dropped from the analysis.

Based on the understanding we have of the data, we expect that the number of additional drivers needed due to not enough standbys + the number of standby resources available to be equal to the number of standbys activated, when the number of standbys activated is equal to or greater than 90. When the number of standby activated is less than 90 (or the number of standby resources available), the number of additional drivers needed due to not enough standby drivers is expected to be zero. Because these expectations-assumptions are true (as shown in the appendix), added to the fact that it is constant, the variable standby resources available was dropped from the analysis.

After data preprocessing was finished, the variables that transitioned into the modeling stage of the project as shown below.

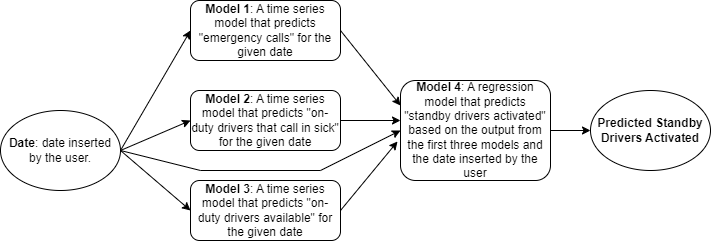


MODELING

We are interested in predicting the number of standby drivers activated. As demonstrated, there are temporal patterns in the data and there are independent features that do have some impact on the target feature. For example, the number of calls on a given day does have an impact on the number of standby drivers activated. Therefore, to predict the number of standby drivers activated, we need to first predict all the variables that have an impact on our target feature (standby drivers activated).

Based on Figure XXX, to predict our target feature, we need the values of six independent variable. Three of those variables are obtained from the current date, that is the date whose standby drivers activated one would want to predict. The variables obtained from the date are month, day of month, and day of week. The variables on-duty drivers that called in sick, emergency calls, and on-duty drivers that are available show temporal patterns based on Figures XXX, XXX, and XXX, respectively. Therefore, a regression model was built to predict standby drivers activated based on the variables shown in Figure XXX. In addition, the three variables that cannot be extracted from the date are predicted separately using different time series models. This means that the entire prediction system will work based on various machine learning models. Figure XXX illustrates how the system works.

Figure XXX – “Standby Drivers Activated” Prediction System



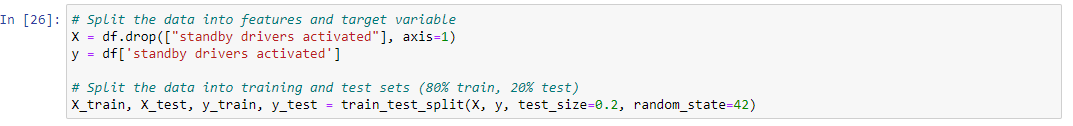
Three time-series models were built to predict on-duty drivers that called in sick, emergency calls, and on-duty drivers that are available. The output from these models and the intended month, day of the week, and day of the month shall be fed into a regression model that will predict the number of standby drivers activated. The available data can be used to predict the three models needed to complete the entire prediction system. In this report, model 4 is presented first. However, during operation, the order presented in Figure XXX is kept.

**Building Model 4**

As shown in Figure XXX, model 4 is the main model of the prediction system. In this chapter, the training and evaluation of model 4 are presented. Because the target variable, which is standby drivers activated, is a numerical continuous variable, regression is the appropriate machine learning method to be used in this project. In addition, the target variable data present a temporal pattern. Thus, a multivariate time series model could be used. However, due to extremely poor performance, multivariate time series models were put aside.

**Splitting Data Vertically and Horizontally**

The data was split into the target variable and independent variables. All the variables, except standby drivers activated, were selected as independent variables, while the variable 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.



A baseline model was developed, as shown below, against which all subsequent models were compared. Seven regression models were trained using the same data. The objective was to evaluate various models and select the best one. The metrics used to evaluate the models were mean-squared error and r-squared. Any model with values worse than those of the baseline model was considered unsuitable. Models performing better than the baseline were compared against each other. Figure XXX shows the models and their metrics.

Figure XXX – Model Evaluation

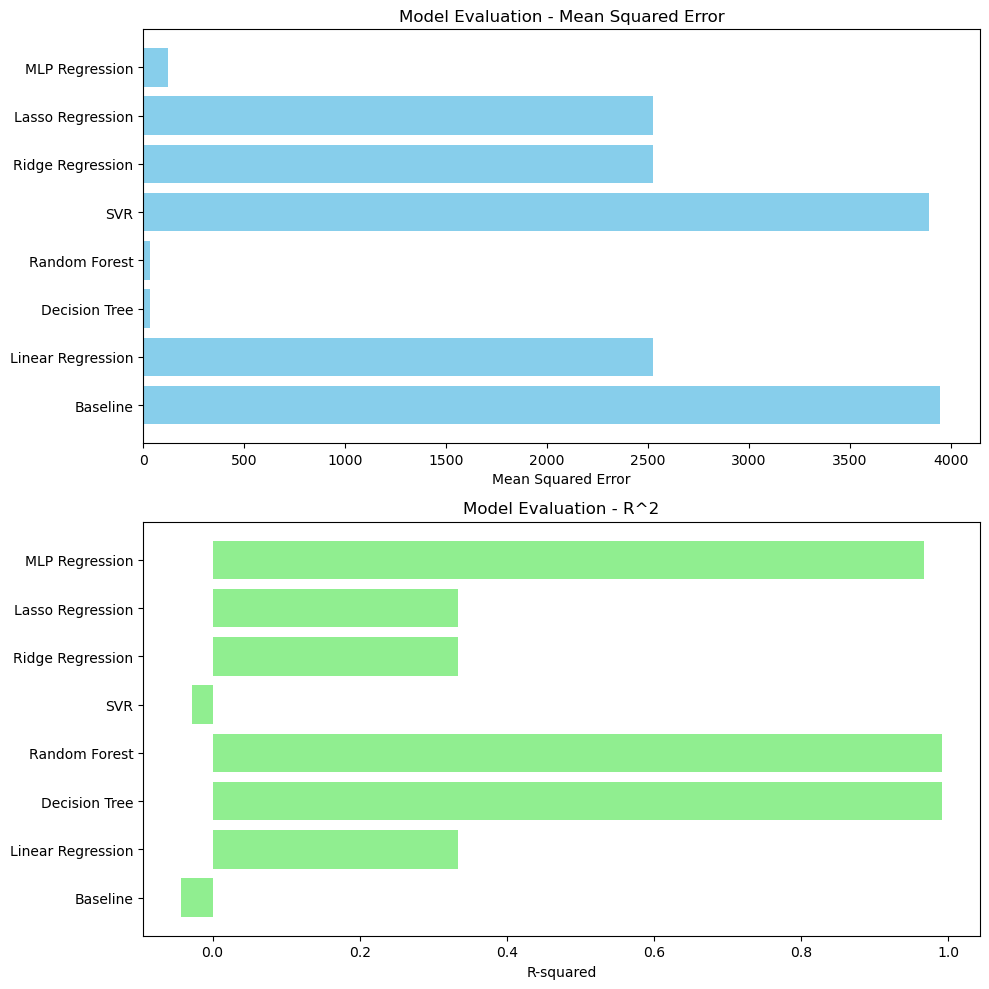
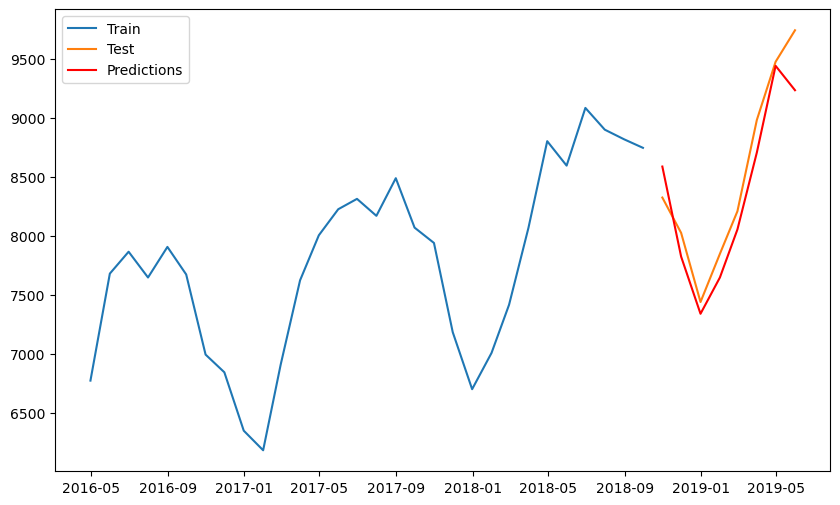


Figure XXX shows that decision tree is the best-performing model, closely followed by random forest, while SVR (support vector regressor) is the worst-performing model. Therefore, the decision tree model was used as model 4 of Figure XXX.

Building Model 1

The data on the column emergency calls was split into 80% and 20% for training and testing, respectively, while the Date column was used as the index. Then a SARIMAX model was trained and tested; it had a root-mean-squared error of 254.5966. Figure XXX shows the predictions of this model compared with the actual values.

Figure XXX – Emergency Call Training, Testing, and Prediction Values



Building Models 2 & 3

Models 2 and 3 were built using a convolutional neural network designed based on Keras with a TensorFlow backend.

916 data points were used as training data, while 115 and 114 data points were used as validation and testing data. In both models, 7 previous readings are used to predict the next reading. Models 2 and 3 had mean-squared errors of 45.35 and 0.641, respectively. Figures XXX and XXX show the results and performance of models 2 and 3, respectively.

Figure XXX Model 2: Predicting On-duty Drivers That Call in Sick

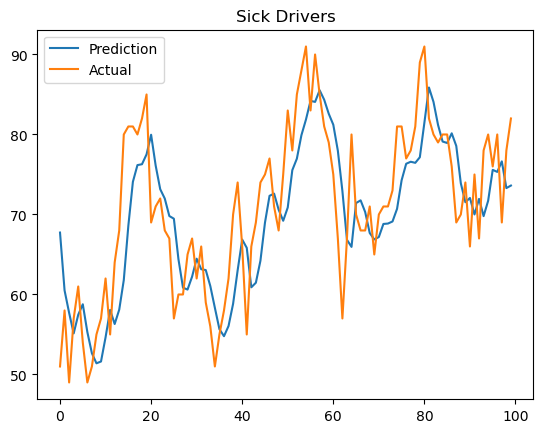


Figure XXX Model 3: Predict On-duty Drivers Available

A graph with a blue line and a blue line

Description automatically generated

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