Motorbike Ambulance Calls Analysis
Text Report

Zhanat Shengelbayeva

December 2021

Contents

1 DATA AND OBJECTIVES

1.1 Motorbike Ambulance Calls

Motorbike accidents and corresponded ambulance calls are highly depended on environmental and seasonal settings, like weather conditions, precipitation, day of week, season, hour of the day, etc. The data was aggregated over the course of two years, on hourly basis, and then extracted and extended with the corresponding weather and seasonal information

1.2 Associated tasks

1.2.1 Regression

Prediction of the hourly ambulance calls count based on the environmental and seasonal settings. Prediction model should provide monotonic in the terms of some features, if it is proved by data. It is desirable to use and compare at least two Regression techniques

1.2.2 Event and Anomaly Detection

Define the ambulance calls patterns with respect of special events and define the cases that might be considered as abnormal behavior

1.3 Dataset characteristics

provided motorbike_ambulance_calls.csv file contains the following fields:

- index: record index
- date: date
- season: season (1:springer, 2:summer, 3:fall, 4:winter)
- yr: year (0: 2011, 1:2012)
- mnth: month (1 to 12)
- hr: hour (0 to 23)
- holiday: whether day is holiday or not
- weekday: day of the week
- workingday: if day is neither weekend nor holiday is 1, otherwise is 0
- weathersit:
    1. Clear, Few clouds, Partly cloudy, Partly cloudy
    2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- temp: Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- cnt: count of total ambulance calls

2 DATASET ANALYSIS

Provided dataset in the file *motorbike_ambulance_calls.csv* contains 17379 rows and 14 columns. The 'date' column has not been used since the 'yr', 'mnth', 'hr', 'holiday', 'weekday', 'workingday' columns provide this. Winter, spring, summer, and fall were the four distinct values in the 'season' column. These values were mapped to the set of natural numbers.
Figure 1 represents the distribution of predictor variables. It's done so to find influential outliers or concentrated values if there is any.
Except for the 'cnt' and 'holiday' columns,mostly, feature values are evenly distributed. There are significantly more feature vectors where 'cnt' (emergency calls) falls between 1 and 250, compared to the whole input range of 1 to 977. The logarithmic transformation can be used on this column to make the distribution more even and bell-shaped. The data providers normalized the input range of certain selected features.
The number of motorcycle ambulance calls is our target variable, which is located in the 'cnt' column. The correlation coefficients between the target variable and the others were calculated to choose the most relevant variables and can be seen in Figure 2. The 'temp', 'atemp', 'hr', 'hum', 'year', 'weathersit','mnth', and 'windspeed' columns have the highest correlations. The values of 'weathersit' are categorical, while the others are numerical. A subset of these predictors should produce good results while lowering the number of predictors in our data. As a result, all other predictors were eliminated from the data.
"yr" column has a strong positive association with the number of calls, indicating that in the second year there were 1.65x more calls. Unfortunately, the model trained using this variable cannot be used to forecast ambulance calls in the future. Therefore it's better to remove it from the dataset. The 'mnth' and 'windspeed' are numerical variables, their correlation coefficients are significantly lower than the ones for the 'temp', 'atemp', 'hr', and 'hum' numerical variables, so those two should be deleted from the dataset as well. Finally, btained dataset contains four numerical predictors ('temp', 'atemp', 'hr', and 'hum') and one categorical feature ('weathersit'). The numerical columns were additionally standardized, while the categorical one was further transformed using one-hot encoding.
For the possible redundant information analysis scatter plot of numerical variables is presented in Figure 3. It is clear that the predictors 'atemp' and 'temp' are highly correlated, being correlation coefficient of 0.988. As a result, to reduce redundancy, the 'atemp' variable is removed. The other factors are not connected with one another, therefore are left in the dataset.

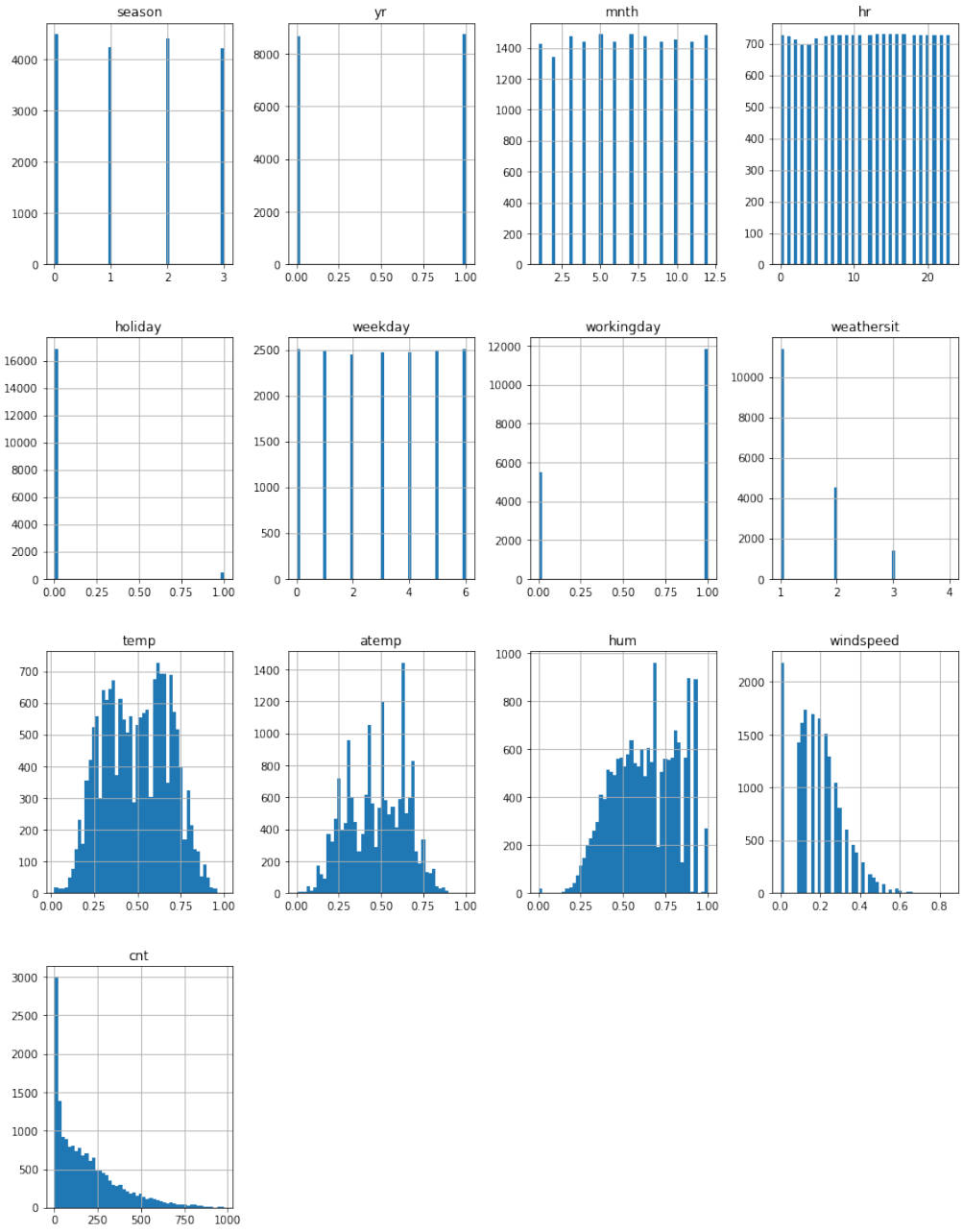

Figure 1: Distribution of predictor variables

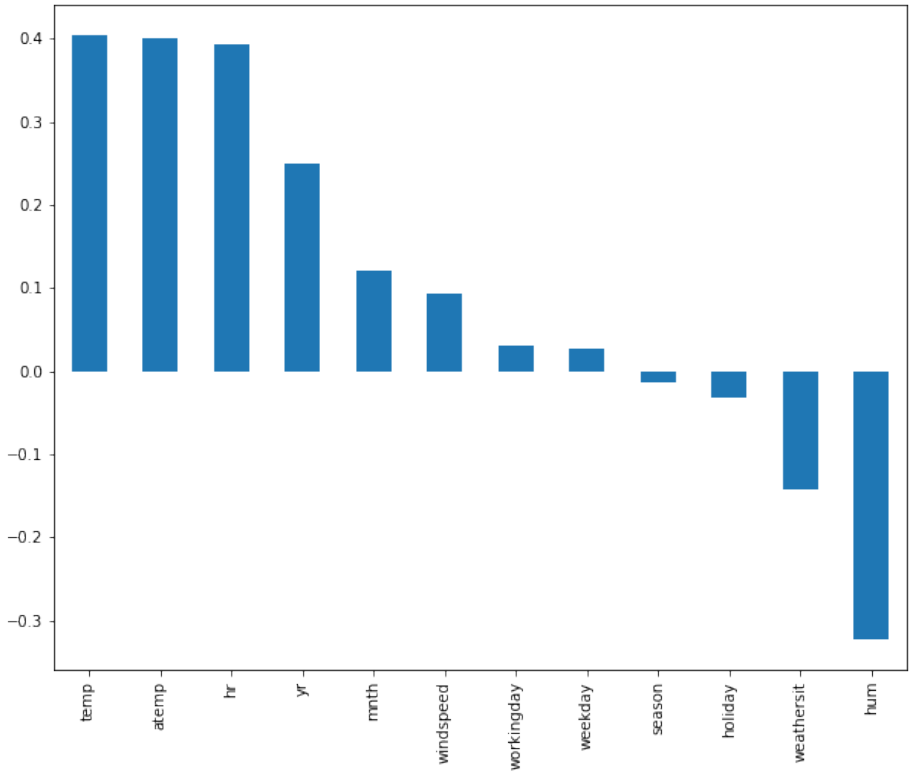

Figure 2: "cnt" ambulance calls and other predictors correlation

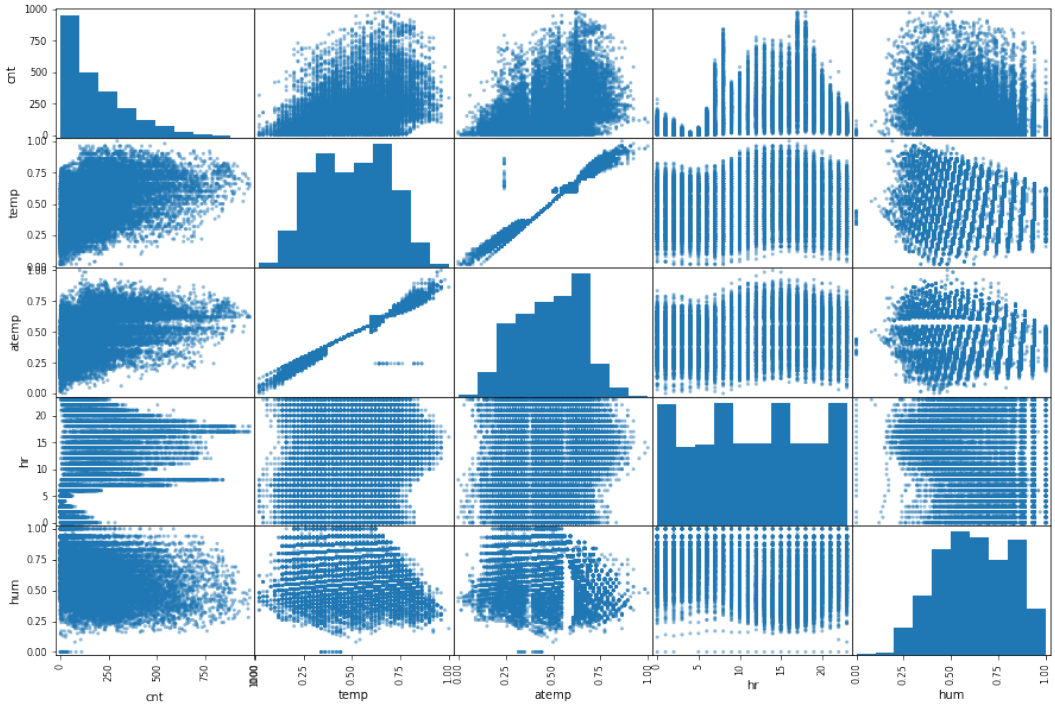

Figure 3: Scatter matrix of selected features: ”cnt”, ”temp”, ”atemp”, ”hr”, ”hum”

3 REGRESSION

To select the best regression model, several regression models provided by scikit-learn library: Linear Model trained with L1 prior as regularizer (aka the Lasso), Linear least squares with l2 regularization, random forest regression (RFR), and support vector regression (SVR).
First, the performance of abovementioned models by mean absolute error (MAE). The MAE estimation was done via 5-fold cross validation for models' default hyper-parameters. The best accuracy was demonstrated by RFR with MAE = 87.81 and STD = 13.17, though the score obtained by SVR (MAE = 92.58 and STD = 27.2) was quite close. The worst performance cases were the linear models with MAE = 116.9 and STD = 22.5 (STD stands for standard deviation).
Next, the best hyper-parameters for the most relevant regression model, i.e. random forest regression, was selected using grid search and are follwing:

- the number of trees n estimator = 50
- the maximum depth of a tree max depth = 5
- the number of features to consider when looking for the best split was set to the number of features in our pre-processed data max features = 'auto'
- default values for the remaining hyper-parameters

This model was evaluated in the same way as previous, and MAE = 85.24, STD = 12.4. Unfortunately, it is not a significant improvement over the model trained using default parameters. A boxplot of the error distribution (Figure 4) can be used to compare the performance of the various models. The fine-tuned random forest regression model has fewer outliers than the other models, in addition to having a lower mean absolute error.

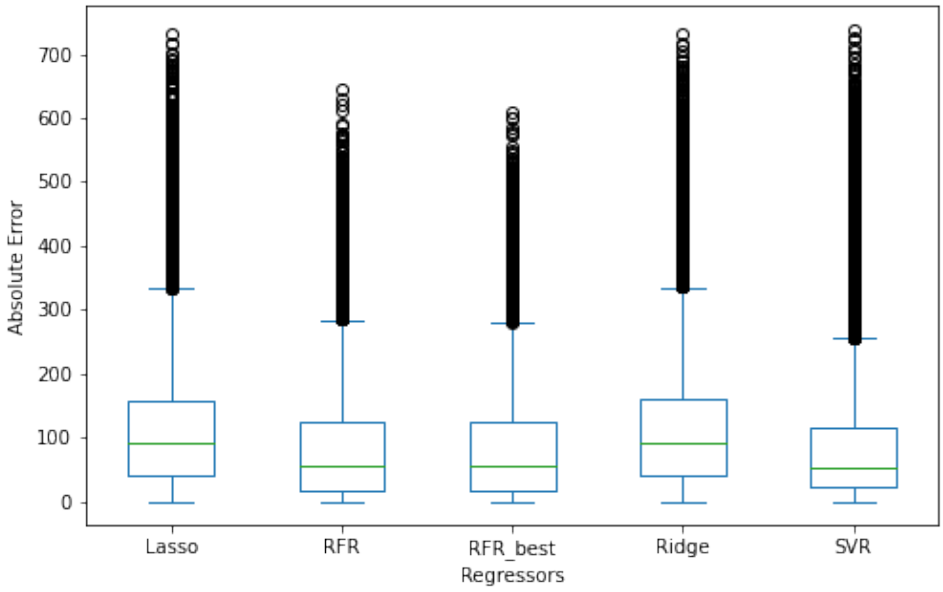

Figure 4: Distribution of absolute error values for the selected predictors: linear model with L1 regularization (Lasso), linear model with l2 regularization (Ridge), random forest regression (RFR), and support vector regression (SVR), trained using default parameters. The RFR best model parameters were fine-tuned using grid search

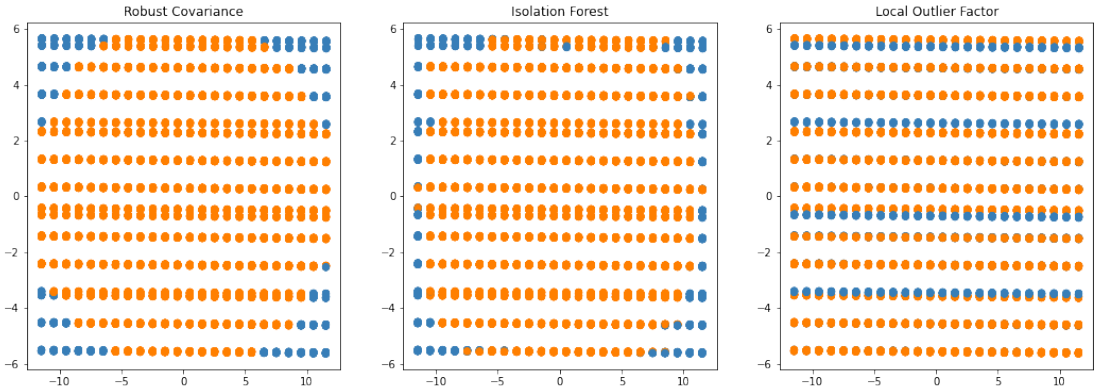

Figure 5: Outlier detection on the 2D. Dimentionality reduction was implemented using PCA. The 'cnt', 'date' columns were excluded. The orange points correspond to inliers, while the blue ones are outliers.

4 EVENT AND ANOMALY DETECTION

After the data exploration step, dataset is left with 17379 rows and 7 columns ('temp', 'hr', 'hum', and 'weathersit'). To be able to visualize the distribution of out data entries it needs to reduce the dimensionality to 3 or less. To do so, the principal component analysis algorithm (PCA) was selected, while keeping either 2 or 3 most significant components. Prior to running PCA on the data, several pre-processing steps were performed. More detailed, information in the 'season' column was transformed from text to natural numbers. Along with it, the 'cnt' and 'date' columns are dropped, further obtained data was projected into low-dimensional feature spaces.

To test three outlier detection algorithms, namely, robust covariance, isolation forest, and local outlier factor transformed data were used. The outlier fraction was set to 15 percent for every algorithm. The results for 2 principal components can be seen in Figure 5. Though, the relative explained variance was around 90.7, the transformed data is distributed almost uniformly across the feature space. Consequently, it is hard to visually spot any anomaly based on the data visualization. The robust covariance and isolation forest algorithms showed similar results by trying to eliminate the data on the boundaries of the reduced feature space. The local outlier factor algorithm produced completely different output. As it was mentioned above, there is no obvious inherent structure to the data at hand or any ground truth result to test against, so it is impossible to compare the produced results. Although, the rejection of the outer data entries made by first two algorithms is quite reasonable. Experiments for 3 dimensional feature space (PCA with 3 principal components) did not display any improvement over 2D feature space.

Furthermore, it was investigated if using outlier detection techniques for data pre-processing could increase regression accuracy. On the inliers, the random forest regression was trained using the best parameters established in Section 3. The obtained results are the folowing: robust covariance MAE = 92.84, standard deviation = 11.36; isolation forest MAE = 92.88, standard deviation = 11.0; and local outlier factor MAE = 87.6, standard deviation = 14.36. As a result, the new pre-processing pipeline did not outperform the one without outlier rejection. It is worth mentioning that the pipeline integrating the local outlier factor algorithm yielded the greatest results when compared to the others.

5 CONCLUSIONS

The provided data set contained records of motorbike ambulance calls. To built a robust regression model several pre-processing steps were performed:

- irrelevant and redundant information removed from the data
- standardization was applied on the numerical columns
- one-hot encoding was applied on the categorical one

The best accuracy was demonstrated by the random forest regression model with MAE = 85.24 and STD = 12.4. Regarding the input data distribution (Figure 1) and relatively low correlation between separate features and the target column (Figure 2), it can be concluded that the regression models performs quite well. Just to compare the scale, the best obtained MAE = 85.24 is well below the average value of calls ¡ Ncalls ¿= 189.5, while the input range was from 1 to 977 calls. PCA and outlier detection algorithms were used to spot outliers in the provided data. Incorporation of outlier rejection algorithms did not give us any significant improvement over the initial pipeline.